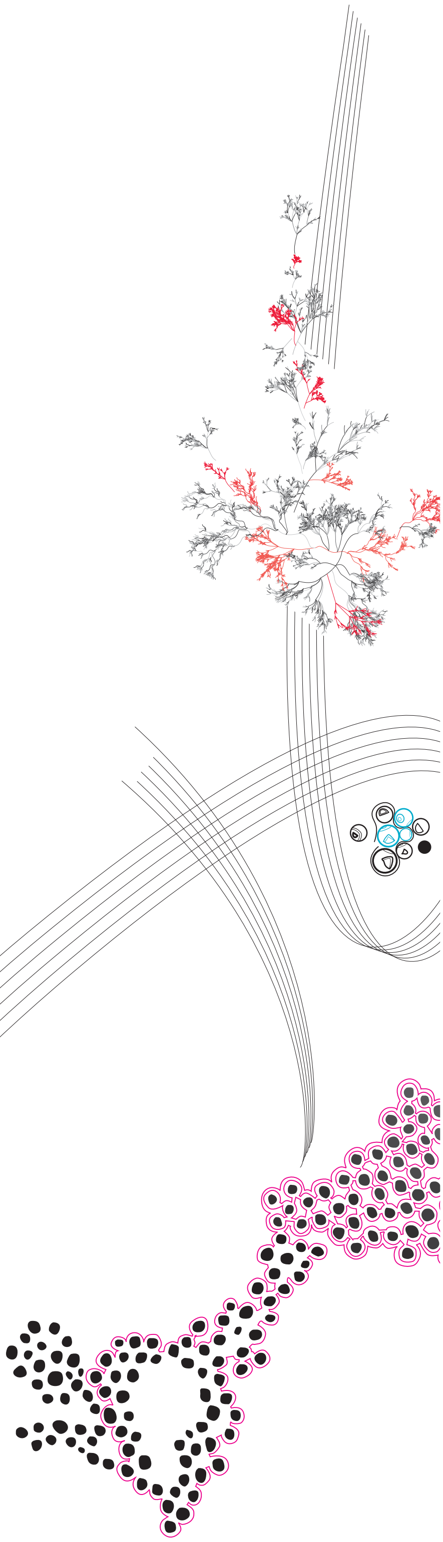
MSc Thesis



Optimizing geriatric care capacity from a system perspective

Anouk E. Beursgens

Graduation committee:	Role:
dr. ing. A.B. Zander	Daily supervisor
prof. dr. R.J. Boucherie	Chair graduation committe
prof. dr. M. Vlasiou	External member graduation committee

April 26, 2024

Department of Applied Mathematics
Faculty of Electrical Engineering, Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Acknowledgement

I would like to thank Anne Zander for her role as my daily supervisor. Your quick responses to my emails, even to just say ‘Thanks for the update!’, and availability to plan meetings on short notice, were much appreciated. Moreover, you were an inexhaustible source of ideas and further research options.

I would like to thank Richard Boucherie for his help concerning the theoretical body of this research. It was nice to have you as an ‘intellectual sparring partner’ for my ideas regarding proving the bounds on the mean queue length of a call packing system and regarding the theoretical substantiation of the Adapted Service Time Approach.

I would like to thank Maria Vlasiou for taking place as the third member of my graduation committee and making time to read this report.

Last, but certainly not least, I would like to thank my parents and friends for their support and interest in my research. Specifically, I would like to thank Gavin for making me tea when I was too occupied writing, or distracting me at moments when I should not think (or stress) about my thesis.

Anouk Beursgens,
April 26, 2024

Optimizing geriatric care capacity from a system perspective

By A. E. Beursgens
April 26, 2024

Abstract

Driven by the enormous costs of bed-blocking days in hospitals, this research develops mathematical models to determine the optimal capacity of the *Verpleeg-, Verzorgingshuizen en Thuiszorg (VVT)* organisations in the region of Enschede. Bed-blocking occurs when a patient is ready to leave the hospital, but appropriate care in the geriatric aftercare institutions is not immediately available. The optimal capacity minimizes the total costs of bed-blocking days, unused VVT beds and used VVT beds, in which the financial objectives of the hospital and the VVT are combined. The situation is modelled as an overflow system with immediate call packing. In addition to the formula for the exponential call packing system by Van Dijk and Schilstra (2022), approximations for a call packing system with general service times are studied. In addition to a stationary analysis, a time dependent analysis based on the Modified Offered Load approximation is performed as well. The performance measures obtained from the queueing analysis are the input to the optimization phase. A Mixed Integer Program is introduced which allows the time dependent behaviour of the capacity to be constrained. Moreover, possible interactions between the capacities of multiple VVT types are considered, like shared capacity. A simulation is used to evaluate the performance of the outcomes of the mathematical models. A numerical analysis is performed based on the data of the geriatric rehabilitation care division of a VVT organisation. The optimal capacity is shown to decrease when the time in overflow contributes more to a reduction in the residual length of stay in the VVT organisation. Moreover, it is influenced by the ratio between the cost of an overflow bed and the cost of an unused bed. The mathematical models and choice of financial objective function allow for a comparison between two situations: the situation in which the VVT organisations individually decide on their capacity and the situation in which the financial incentives of the hospital are taken into account as well. The results suggest that cooperation between the hospital and the VVT would be beneficial and that it would be reasonable to compensate the VVT organisations for increasing their capacity.

Management summary

Driven by the enormous costs of bed-blocking days in hospitals, this research has developed mathematical models to determine the optimal capacity of the *Verpleeg-, Verzorgingshuizen en Thuiszorg (VVT)* organisations in the region of Enschede.

Context

Elderly people who are ready to leave the hospital often need a form of additional (after) care. Depending on their situation, they can go home to receive care there or they should move to, for example, a nursing home. In the Netherlands, the collective name for all care concerning elderly people who cannot live on their own anymore is the VVT (Verpleeg- en Verzorgingshuizen en Thuiszorg). When appropriate care in the VVT is not immediately available, the elderly people need to stay at the hospital until the appropriate care is available. During this time, they block a hospital bed that could otherwise be used to treat new patients. The (limited) capacity of VVT organisations thus influences the number of so-called *bed-blocking days*.

Modelling approach

In this research, mathematical models are developed to determine the optimal capacity of aftercare institutions. Optimal is here defined as minimizing the total costs of bed-blocking days, unused VVT beds and used VVT beds. The financial objectives of the hospital and the VVT are combined, that is, a system perspective is used. This research is performed on a strategic and tactical level. The mathematical analysis is based on average performance measures. The performances of the outcomes of these analytical models are evaluated using a simulation model. The results can serve as guidelines for the average capacity that should be available and show possible daily adaptations to the capacity level based on the predictable fluctuations in the arrival process.

Results

A numerical analysis is performed based on the data of the geriatric rehabilitation care division of a VVT organisation. The optimal stationary capacity ranges between 30 and 27 for this care division under the used parameter values. A capacity of 30 would be optimal if the time waiting in the hospital would not contribute to a reduction in the residual length of stay in the VVT organisation, whereas a capacity of 27 would be optimal if it would make no difference to the patients' length of stay whether they are in a hospital bed or a VVT bed. The optimal capacity is influenced by the ratio between the cost of an overflow bed and the cost of an unused bed: a relative increase in the cost of an overflow bed means a higher optimal capacity, while a relative increase in the cost of an unused bed results in a lower optimal capacity.

The data showed that the number of patients arriving on weekdays is around three to six times higher than on Saturdays and Sundays, respectively. To benefit from this predictable variation in the daily number of arrivals, a time dependent capacity is considered. The fewer arrivals on Saturday and Sunday relate to a lower capacity on Wednesday, so there is a shift between the dips and peaks in the arrival process and those in the daily capacity level. Due to the quickly varying number of arrivals per day, varying the capacity over the week is only beneficial when the cost of changing capacity is rather low. The downside of a time dependent capacity is that a decrease in capacity may be planned while all beds are occupied. Depending on how undesired the resulting overbeds are, the monetary benefits of allowing the capacity to vary over the week might be outweighed by the robustness of a stationary capacity.

The mathematical models and choice of financial objective function allowed for a comparison between two situations: the situation in which the VVT organisations individually decide on their capacity and the situation in which the financial incentives of the hospital are taken into account as well. For the used parameter settings, the optimal stationary capacity in the first situation is 19, while the optimal stationary capacity from a system perspective would be 29. The costs for the system would decrease from 6446 euros per day in the first situation to 812 euros per day in the second situation and the expected number of bed-blocking days would decrease from 17.71 days to 0.81 days. On the other hand, the costs for the VVT are expected to increase by 688 euros. Although the precise values heavily depend on the parameters that are chosen, they serve to illustrate the difference between the two situations. These results clearly suggest that cooperation between the hospital and the VVT would be beneficial and that it would be reasonable to compensate the VVT organisations for increasing their capacity.

Contents

1 Introduction

The title of an analysis by a Dutch national newspaper *de Volkskrant* from October 2023 is clear: Elderly care can no longer be ignored [71]. The analysis shows that the costs of elderly care are skyrocketing and the waiting list for nursing homes continues to grow. Meanwhile, the *minister voor Langdurige Zorg en Sport* (minister of Long Term Care and Sport) published in 2022 a vision in which they want to freeze the number of beds in the nursing homes [54]. In a report of the *Nederlandse Zorgautoriteit* (2023), an independent Dutch governing body that monitors the Dutch healthcare system, also from October 2023, it was concluded that the way the long-term care is currently organized can not be maintained much longer, partly in response to this freeze. A very recent article in the *Tubantia*, the regional newspaper of the eastern part of the Netherlands, revives the discussion about the decision to freeze the number of beds as two of the main hospitals of region Twente appear to miss out on almost nine thousand euros daily due to patients needing to wait in the hospital for a place in geriatric care [45]. Only in these two hospitals, there are on average sixteen patients per day unnecessarily in the hospital since they are waiting for their geriatric care. In a succeeding article, a politician says that the decision to freeze the number of nursing home beds will lead to higher healthcare costs and he urges to find out how many beds there should be to prevent blocking the outflow from the hospital [69].

This research aims to develop mathematical models to determine the optimal capacity of the *Verpleeg-, Verzorgingshuizen en Thuiszorg (VVT)* organisations in the region of Enschede, but the models could be applied to all kinds of aftercare situations. The VVT organisations provide care for the elderly, both in the form of nursing homes or geriatric facilities as well as home care for elderly people who can still live on their own. Since this research is applied to the Dutch health care system, we will use the corresponding Dutch naming, or abbreviation, in the remainder of this report. The use of the mathematical models is illustrated by a numerical analysis of one type of VVT care in one organisation, due to limited data available. This research is performed on a strategic and tactical level. The analysis is based on the long-run behaviour of the system and the results could therefore serve as guidelines for the average capacity that should be available and the predictable possible weekly fluctuations. This is in contrast to operational planning, which concerns day-to-day decisions on the staffing of the beds which is highly determined by short-term fluctuations.

As the news article [45] mentioned, elderly people need to wait in the hospital until they can go to a bed in the VVT. These *bed blocking days* cost the hospital a lot of money since they can not use the blocked bed for new patients. The hospital receives a fixed amount of money per care request [2] and thus benefits financially from optimizing its throughput. The VVT, on the other hand, gets money for the occupancy level [6], which is particularly high if the capacity is close to the demand which in turn causes high waiting times. From a financial perspective, the VVT has little incentive to change the current situation, while the hospital might be willing to pay money to solve the bed-blocking problem. Therefore, this research will act as if the hospital and VVT are shared problem owners with their combined financial situation. Since not only bed blocking but also unused capacity can cost a lot of money, the desired optimal capacity will be the capacity that minimizes the total combined costs of bed blocking and unused VVT beds. Although a financial objective is used to determine the optimal capacity, the resulting system will also be an improvement from the perspective of care provision since a decrease in the number of bed-blocking days is expected as hospital beds are more expensive than VVT beds.

To compare the performance of several capacity levels in the VVT, an objective function will be set up and expressions for the necessary performance measures will be derived. The derivation of the performance measures, being the mean number of patients in the queue for the VVT and the mean number of unused VVT beds, will be based on an overflow system with call packing. Since the steady state distribution of this system is only in a convenient form for exponentially distributed service times, and since the infinite server system is a special boundary case of the call packing system, an Adapted Service Time Approach is introduced for the infinite server system. The adapted service time is the fixed point of a system of two equations, for which existence and uniqueness are proven. We conjecture, and numerically substantiate, that an infinite server system with this adapted service time provides a stricter insensitive lower bound on the mean number of patients in an overflow system with call packing than the unadapted infinite server system. No such insensitive bound on the call packing system could be found in literature before.

Next to a stationary analysis, time dependent capacity is considered in the case of a time-dependent arrival process. The performance measures for the infinite server system are made time dependent by using the time dependent offered load. This time dependent offered load is also used to replace the stationary load in the formulas for the exponential call packing system. Such a Modified Offered Load approximation simplifies the time dependent capacity decision significantly as the values of the performance measures, and thus the objective function, only depend on the choice of capacity at that time point. To determine the optimal time dependent capacity, a Mixed Integer Program will be introduced. Optimization via a Mixed Integer Program enables us both to incorporate constraints on the time dependent behaviour of the capacity as well as to study shared capacity or relabeling of beds between different types of patients. Using a Mixed Integer Program to solve the time dependent staffing problem adds to the theory on this topic as it offers more possibilities than the commonly used squared staffing rule.

With data from a VVT organisation, a numerical analysis based on these analytical models is performed. Since the real situation could not be captured exactly in the analytical analysis, the capacities obtained with the analytical analysis are verified with a simulation and compared to the costs of neighbouring capacity solutions.

This report is structured as follows. First, results from the field of queueing theory are explained in Section 2, both for the stationary as well as the time dependent case. Theory on specifically time dependent staffing is discussed and the contributions of this research to the existing theory are stated. Moreover, the structure of the VVT and its financials are explained in more depth. Then, the models for the analytic analysis are explained in Section 3. This section consists of two parts. In Section 3.1, the situation is modelled as an overflow system and the necessary performance measures are derived. Moreover, the Adapted Service Time approach is introduced and theoretically founded. Section 3.2 covers the optimization part by constructing an objective function and introducing the Mixed Integer Program and several possible constraints. The data for the numerical analysis and the parameter estimation from the data is explained in Section 4. The results from the numerical analysis are discussed in Section 5. This section also contains a simulation study and compares the simulation outcomes to the analytic results. Limitations and possibilities for further research are finally stated in Section 6. Section 7 concludes this report by giving the highlights of this research and recommendations based on the results of this research.

2 Theoretical Background

In this section several concepts and results from the field of queueing theory are gathered that will form the basis for the discussions and derivations in Section 3.1. The first part of this section discusses the well-known infinite server model (Section 2.1.1) and finite server model (Section 2.1.2). The formulas for these models can be found in most textbooks on queuing theory, for example, Shortle et al. (2018) and Adan and Resing (2015). Then, results for the less well-known overflow model are given (Section 2.1.3). The topic of the second part of this section is the concept of time dependence of the arrival process. An overview of results for a time dependent arrival process is given, both for the case of infinite servers (Section 2.2.1) and a finite number of servers (Section 2.2.2). The application to staffing is kept in mind while writing the latter but is also explicitly discussed in Section 2.3. The contribution of this research to the existing theory is discussed in Sections 2.1.4 and 2.3.5 for the stationary case and the time dependent case, respectively. The last part of this section (Section 2.4) covers different types of the VVT and its financial incentives.

2.1 Stationary models

2.1.1 Infinite server models

Infinite server models are considered easy models to analyse because the infinite availability of resources leads to independence between all individuals in the system. It is well-known that the steady-state distribution of the number of customers in an infinite server model with Poisson arrivals happening at rate λ and general service distribution ($M/G/\infty$) with mean $E[S] = \frac{1}{\mu}$ is Poisson distributed with mean $\rho = \lambda E[S] = \frac{\lambda}{\mu}$, that is,

$$\pi(n) = \frac{\rho^n}{n!}e^{-\rho}. \tag{1}$$

A derivation of this formula can be found in textbooks on Queueing Theory, among which Section 6.2.2. of Shortle et al. (2018).

Although most real-life systems have a finite number of resources, which leads to interesting concepts such as waiting and blocking, infinite server models can be of great use when analysing these systems. Several useful applications of the infinite server model are listed in Whitt (2012) and Whitt (2017). In general, infinite server models can serve as remarkably good approximations of multiserver queueing systems. For systems with time dependent arrival rates, infinite server models can furthermore give important insights into the global behaviour or physics of the corresponding multiserver models. In such systems, infinite server models form the basis for the analysis of the offered load, which is the load of the system when no waiting or blocking would take place. The latter, Whitt (2017) states to be ‘arguably of greatest importance’ and will be explained in Section 2.2.2.

A recent review of the applications of infinite server models in healthcare is given in Worthington et al. (2020) for both time-homogeneous and time-inhomogeneous models. Time-homogeneous, or stationary, infinite server models are often used to analyse capacity issues. A common approach to approximate the steady-state blocking probability, that is, the probability a patient is blocked due to all beds being full, is by simply truncating the steady-state distribution of the number of patients in the system for the equivalent infinite server system.

The strength and popularity of the infinite server model lie in its simplicity. By using infinite server models, we deliberately do not model what might happen after exceeding capacity which could overcomplicate the model and hinder the understanding of the global process, as argued by Gallivan et al. (2002). Exactly this property is the reason why Worthington et al. (2020) sees great usefulness of the infinite server model as the low fidelity model for multi-fidelity modelling. Multi-fidelity modelling exploits the benefit of the low computational cost of a low fidelity model by narrowing down the solution space of interest such that the higher computational costs of a high fidelity model are partly circumvented, while still being able to obtain a solution with high accuracy. Multi-fidelity modelling can be applied in queueing theory by having an analytical (infinite server) model as low fidelity model to obtain a rough guess for the optimal capacity, and a simulation model as high fidelity model to analyse the system for capacities ranging around the capacity obtained by the analytical model. One of the most important challenges of multi-fidelity modelling is that well (and bad) performing solutions in the low fidelity model should also perform well (and bad) in the high fidelity model. Since infinite server models behave the same as their corresponding finite server model until capacity is reached, and since when looking for optimal capacity the number of times capacity is exceeded should be limited, the required relation between the infinite model and the more complex simulation is likely to hold.

2.1.2 Finite server model

As stated above, most real systems do not have infinite capacity, but a finite capacity c. The offered load ρ is then commonly defined as $\rho = \frac{\lambda}{c\mu}$. Unfortunately, the insensitivity property of the infinite server system does not hold for a system with a finite number of servers and infinite waiting room, that is, the distribution of the service time influences its steady state distribution. An exact formula for the steady-state distribution is only available for the case of an exponential service distribution ($M/M/c/\infty$),

$$\pi(n) = \begin{cases} \frac{(c\rho)^n}{n!}\pi(0), & n \leq c \\ \rho^{(n-c)}\frac{(c\rho)^c}{c!}\pi(0), & n > c, \end{cases} \tag{2}$$

where $\pi(0)$ follows from normalization, $\pi(0) = \left(\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{1}{1-\rho}\frac{(c\rho)^c}{c!}\right)^{-1}$.

The blocking probability in a $M/G/c/c$ queue, $B(c, c\rho)$, defined as

$$B(c, c\rho) = \frac{(c\rho)^c/c!}{\sum_{n=0}^{c-1}(c\rho)^n/n! + (c\rho)^c/c!},$$

for which the recurrence relation

$$B(c, c\rho) = \frac{\rho B(c-1, c\rho)}{c + \rho B(c-1, c\rho)}, \tag{3}$$

holds with initialization $B(0, \rho) = 1$, can be used to obtain a recursive equation for the delay probability in a $M/M/c/\infty$ queue,

$$\Pi_W = \sum_{k=0}^{\infty} \pi(c+k) = \frac{\rho B(c-1, c\rho)}{1 - \rho + \rho B(c-1, c\rho)}, \tag{4}$$

which appears to be quite useful when determining the mean queue length in Section 3.1.1.

The results for this special case of exponentially distributed service times can be used to approximate the mean waiting time and the mean number of people in the queue for general service time distributions by multiplying the value of the corresponding performance measure for $M/M/c/\infty$ with $\frac{1+C_S^2}{2}$, where $C_S^2 = \frac{\sigma^2}{\mu^2}$ is the squared coefficient of variation of the service time distribution,

$$E[L_q^{(M/G/c)}] \approx \frac{1 + C_S^2}{2} E[L_q^{(M/M/c)}]. \tag{5}$$

Whitt (2009) states that this use of the Kingsman's formula or this Allen-Culleen approximation usually yields an excellent heavy-traffic approximation and is exact for $c = 1$.

2.1.3 Overflow system with call packing

The study of overflow models has its roots in telecommunication and more specifically circuit switched networks. In this application area, the overflow process known as *grading* is important: there are two independent arrival streams to two finite server queues, where arrivals to the first queue who find all servers busy are offered to the second queue (overflow). According to Hordijk and Ridder (1987), no closed form expressions exist for general service time distribution. Several papers, e.g. Hordijk and Ridder (1987), van Dijk (1987) and Van Dijk and Van Der Sluis (2009), therefore study insensitive lower and upper bounds to this system. One of the main product form modifications, i.e. modifying the system in such a way that the steady-state distribution has a product form by 'repairing' the initial violation of local balance, is known in teletraffic literature as *call packing*. In a call packing system, an overflowed type 1 call is switched back (repacked) to the first queue once a server at that queue becomes available. Although the main references for call packing used in this paper, use call packing as an approximation or bound, call packing on its own is in this research of practical use.

Van Dijk and Schilstra (2022) overcome an important shortcoming of earlier research by enabling the service rate of overflow jobs and type two jobs at the second station to be different. They prove the product form formula for the steady-state distribution of a two-station overflow model with immediate call packing,

$$\pi(n_1, n_2, m) = CF(m) \prod_{i=1,2} \frac{1}{n_i!}\Big(\frac{\lambda_i}{\mu_i}\Big)^{n_i}, \text{ with} \tag{6}$$

$$F(m) = \begin{cases} \lambda_1^m / \prod_{k=1}^m (N_1\mu_1 + k\gamma) & m > 0 \\ 1 & m = 0 \end{cases}$$

where $n_i \leq N_i$ are the number of (non-overflow) jobs in station i, m the number of overflow jobs, μ_i and γ the service rates of the type i jobs at station i and overflow jobs at station 2 respectively, and C the normalization constant. For non-exponential service times with unequal service rates, their simulation results show that there is no strict insensitivity and thus no product form solution. This formula thus does not hold for general service times.

Van Dijk and Schilstra (2022) note that in case of non-exponential service times, the (residual) service time at the first station in the case of repacking can be determined in two ways: the service can be resumed or resampled. In the case of the latter the service completely starts over, while in the case of the first, the 'residual service time' at the first station is the service time that was left at the moment of repacking multiplied with a factor $\frac{\gamma}{\mu}$ to account for the difference in service speed.

Even though overflow models were developed for telecommunication, the idea that a person is (temporarily) served by an auxiliary facility when the primary service facility is congested and the corresponding mathematical modelling can also be applied to health care as pointed out by Litvak et al. (2008). Although the use of the terms 'call packing' and 'repack' give a detached feeling when used for the transfer of patients in health care in the authors' opinion, we will keep using them throughout this paper as it is the common term in the mathematical theory of overflow system.

2.1.4 Contribution of this research

Since the overflow system with call packing has a product form solution, its main appearance in literature is its use to approximate the *grading* overflow system. Little research could be found that has the callpacking system itself as the main focus. This research concerns a real-life situation that can be modelled directly as a call packing system, showing a practical use of the call packing system.

Van Dijk and Schilstra (2022) showed that in the case of distinct service rates for overflow jobs and the other jobs, the product form solution only holds for exponential service times. To the best of our knowledge, no literature has been published on how this call packing system with distinct service time itself can be approximated or bounded analytically for general service times, other than of course by the exponential product form formula. This research makes the first step to fill this literature gap.

2.2 Time dependence

A lot of the research on queueing theory, and especially textbooks, (only) treats the case of a stationary arrival process. However, in reality, it is quite common that the arrival intensity varies over time; think of rush hours on the highway or seasonal influences on ice cream sales. In various hospital departments, the demand is shown to depend on the time-of-day and the day-of-week, with in general a difference between weekdays and weekends, as remarked by, among others, Green (2006) and de Bruin et al. (2007). Since time dependency in the arrival process causes a predictable variability in the arrival pattern, it is beneficial, even crucial, to take this time dependency into account when determining the capacity, especially when flexible capacity is considered. Several studies have already been performed on the impact of time dependency of the arrival pattern on the system performance and capacity as listed by Bekker and de Bruin (2010). According to Massey (2002), one of the only textbooks that devote a complete chapter to non-stationary queues is Hall (1990). A reader interested in more extensive reading on the basics of the topic of non-stationary arrivals is therefore referred to Section 6 of that book. An extensive survey and classification of the literature on single-stage queueing systems with time dependent parameters is given by Schwarz et al. (2016).

2.2.1 Infinite server system

If the arrival rate at time t is a function of t, $\lambda(t)$, a non-homogeneous Poisson process is obtained. In Section 5.4 of Ross (2007) it is proven that every non-homogeneous Poisson process, with a bounded intensity function, can be thought of as being a time sampling of a Poisson process. That is, an arrival at time t occurring in a Poisson process with rate λ is counted with probability $p(t)$. With this characterisation in mind, the reader might not be surprised that the number of customers in a time dependent $M(t)/G/\infty$ system, $N(t)$, follows a Poisson distribution just like the stationary variant. Already around the middle of the previous century, it was shown by Palm (1943) and A.Ya.Khinchine (1955) that $N(t)$ has a Poisson distribution with mean $m(t)$, where $m(t)$ depends on the arrival rate function $\lambda(t)$ and the service time distribution. These and more basic results for the $M(t)/G/\infty$ are summarized in Theorem 1 of Eick et al. (1993), among which the fact that, for each t, $N(t)$ has a Poisson distribution with mean

$$m(t) = E[\int_{t-S}^{t} \lambda(u)du] = E[\lambda(t - S_e)]E[S], \tag{7}$$

where S is a random variable for the service time and S_e is a random variable for the associated stationary-excess or equilibrium-residual lifetime. The underlying assumptions are that the $M(t)/G/\infty$ system started empty at $t = -\infty$ and that the arrival rate function is non-negative, measurable, and integrable over any bounded interval. These assumptions will hold throughout this report as well. Given the cumulative distribution function $G(t)$ of the service times, the equilibrium-residual lifetime has a cumulative distribution function

$$G_e(t) = P(S_e \leq t) = \frac{1}{E[S]} \int_0^t 1 - G(u)du, \quad t \geq 0.$$

Using this definition in formula (7) yields an often given third expression for $m(t)$, as in formula (2.2) in Feldman et al. (2008) and in formula (4.1) in Green et al. (2007), or stated differently in formula (3) in Bekker and de Bruin (2010),

$$m(t) = \int_{-\infty}^{t} (1 - G(t-u))\lambda(u)du = \int_0^\infty (1 - G(v))\lambda(t - v)dv. \tag{8}$$

Each of the three expressions for $m(t)$ has its own interpretation as clearly described in Section 4.2 of Green et al. (2007).

Moreover, the moments of S_e can be expressed via the moments of S via

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}, \quad k \geq 1. \tag{9}$$

The mean residual service time is thus given by $E[S_e] = \frac{E[S^2]}{2E[S]} = E[S]\frac{c_S^2+1}{2}$, where $c_S^2 = \frac{Var(S)}{E[S]^2}$ is the squared coefficient of variation of the service time. So, for a squared coefficient close to 1, the mean residual service time is close to the mean service time. This observation is in line with the theory of the exponential distribution; the exponential distribution has a c_S^2 of one and due to the memoryless property the residual service time indeed equals the complete service time.

It should be clear to the reader that the time-dependent infinite server queue is not insensitive anymore, since the offered load $m(t)$ depends on the (residual) service time through more than only its mean.

Simplified expressions for offered load for specific arrival rate functions

For specific arrival rate functions, the offered load $m(t)$ can be calculated more straight-forward, avoiding the integral definition (8). Eick et al. (1993, 2000) derive several simple formulas for $m(t)$ in case the arrival rate function is quadratic, the step function or a spike function and a sinusoidal function, respectively. The spike function may not be interesting on its own, but in combination with theorem 8 of Eick et al. (1993), which broadly says that a linear decomposition of $\lambda(t)$ is inherited by $m(t)$, it enables us to model traffic surges and piecewise constant (approximations of general) arrival rate functions. Analysis with piecewise constant arrival rate functions appears to be quite useful since $\lambda(t)$ is often estimated through counting the arrivals over short intervals as noted by Hall (1990). Due to the flexibility and tractability that this choice of arrival rate function offers, Bekker and de Bruin (2010) assume a piecewise constant arrival rate function. They note that a sinusoidal function fails to cover cyclic periods that do not have some kind of symmetry, which makes it hard to cover the weekday-weekend pattern often observed in healthcare applications. To model such a weekday-weekend pattern, the cycle length of one periodic cycle is seven days which could be divided into seven intervals (a daily arrival rate) or two intervals (working days and weekends) in which the arrival rate is said to be constant. Bekker and de Bruin derive a convenient closed-form expression for the time-dependent offered load in case of a hyper exponential service time distribution (a mixture of k exponentials with rates μ_k).

Approximations for offered load

In case no simple expression for the time-dependent offered load exists and the integral definition 8 is not desired or can not be solved satisfactorily, several approximations for the offered load are available. First of all, it is important to realise that the (time dependent) offered load $m(t)$ is defined by formula 7 and does in general not equal the instantaneous offered load $\rho(t) = \lambda(t)E[S]$. Equality would only hold if the arrival process would be homogeneous with a constant arrival rate, and therefore the expression $\lambda(t)E[S]$ is well-known as the *pointwise stationary approximation (PSA)* for $m(t)$. Feldman et al. (2008) quantifies the difference between the PSA offered load and the time dependent offered load $m(t)$,

$$m(t) - \lambda(t)E[S] = \frac{1}{2}E[\lambda'(t - (S_e)_e)]E[S^2],$$

in which $(S_e)_e$ is a random variable with the twofold stationary-excess cdf $(G_e)_e$. This expression shows that the PSA offered load $\lambda(t)E[S]$ will not be a good approximation when the arrival rate varies rapidly in time (since λ' will be large) or when the service time has a large variance.

Comparing the PSA to formula 7, the difference is the occurrence of a random time lag in $\lambda(t)$. As the time shift is especially important according to Feldman et al. (2008), a better approximation is obtained by the *lagged PSA*,

$$m(t) \approx \lambda(t - E[S_e])E[S] = \lambda(t)E[S] - \frac{\lambda'(t)E[S^2]}{2}, \tag{10}$$

where the equality follows from relation (9) [24]. Since the time shift depends on the second moment of the service distribution, Massey (2002) warns that especially service distributions with heavy tails can have reasonable means but enormous time lag between the peaks in arrivals and in offered load. Whitt (2007) relates the PSA and the lagged PSA to the three defining equations of the offered load $m(t)$.

The lagged PSA originates from the first-order Taylor-series approximation for the arrival rate function centred at t. If instead a second-order approximation is used, a more refined approximation is obtained,

$$m(t) \approx \lambda(t - E[S_e])E[S] + \frac{\lambda^{''}(t)}{2}Var(S_e)E[S] = \lambda(t)E[S] - \frac{\lambda^{'}(t)E[S^2]}{2} + \frac{\lambda^{''}(t)E[S^3]}{6}, \tag{11}$$

where the equality again follows from relation (9) [24]. This *quadratic approximation* contains next to the time shift also a space shift. As noted by Feldman et al. (2008), the space shift acts as a smoothing operator since the second derivative $\lambda^{''}$ is negative at a peak, and thus the behaviour of $m(t)$ will be less extreme than that of $\lambda(t)$. This damping property is also noted by Massey (2002) but then concluded from the Fourier transform of $E[N(t)]$. According to Green et al. (2007), experience shows that the space shift does not matter much for service times that are not too long.

In simulation

According to Hall (1990), the interarrival times of a time dependent Poisson process are, in contrast to the stationary variant, not exponentially distributed anymore. So, one can not simply draw at time t the time to the next arrival from an exponential distribution with mean $\lambda(t)$. Instead, we draw candidate interarrival times from an exponential distribution with mean $\lambda^* = sup_t\lambda(t)$ and accept these at time t with probability $\frac{\lambda(t)}{\lambda^*}$. The working of this method is substantiated by Ross (2007) by the fact that a non-homogeneous Poisson process actually is a time-sampled Poisson process.

2.2.2 Finite server system

Time dependent analysis of finite server systems is in general extremely more difficult than the infinite server case since the number of people in the system is not Poisson distributed anymore. As mentioned before, Schwarz et al. (2016) contains a clear extensive survey of the literature on single-stage queueing systems with time dependent parameters. They classify the approaches in three main categories of which the last two are relevant for this research. One of these two categories contains approaches that assume piecewise constant parameters. The time horizon is divided into intervals which are studied separately. The two simplest subgroups of methods assume constant parameters during each interval and apply a steady-state formula to these parameters; the piecewise stationary models. The first subgroup of models furthermore assumes that there is no dependence between the intervals. The earlier described PSA is among these methods, just like the Simple Stationary Approximation (SSA) which used the average arrival rate taken over the complete time horizon. Jennings et al. (1996) show a specific $M(t)/M/s(t)$ scenario for which these methods perform badly. They note that the shortcoming of these models is the way the effective arrival rate at time t is determined. Although one naturally would like to average the arrival rate over an appropriate interval before time t, the PSA takes the arrival rate at time t, meaning an interval length of zero, while the Simple Stationary Approximation (SSA) uses the long-run average arrival rate, meaning an interval of infinite length. A small improvement is therefore the Stationary Independent Period-by-Period approximation (SIPP) of Green et al. (2001) in which the arrival rate is averaged over the according interval. Intuitively, these methods can only work satisfactorily if the change in the arrival rate function is limited and if there is little build-up of the queue preventing carry-over of congestion between consecutive intervals. Furthermore, Green et al. (2007) emphasize the importance of the length of the service times. For middle to long service times, it is important to take into account the time lag of the offered load by using the lagged refinements of PSA and SIPP in which the arrival rate is shifted to the right by the mean service time. However, for reasonably long service time, they advocate using the Modified Offered Load approximation, which will be discussed in more detail below.

The other subgroup of methods is still piecewise stationary but does take into account dependencies between intervals. In the Stationary Backlog-Carryover approximation (SBC) introduced by Stolletz (2008, 2011), the queue at the end of one interval is carried over to the succeeding interval by using an appropriately calculated modified arrival rate.

The last subgroup with piecewise constant parameters studies the transient behaviour within each interval. In the approaches based on transient models (BOT), as named by Schwarz et al. (2016), the system state at the end of an interval serves as the initial condition for the transient analysis of the subsequent interval. It is evident that few publications are available on this topic, and it strikes that the mentioned literature only considers single-server systems. Although Yoo uses a transient analysis per period to determine the optimal flexible staffing, as summarized and lowerbounded by Assad et al. (1997) and theoretically substantiated by Michael C. Fu and Wang (2000), this research is not mentioned by Schwarz et al. (2016). Uniformization/randomization is a common approximation technique for the transient analysis of a queuing system but is unfortunately characterized by high computation times [70].

The other category concerns approaches in which the number of servers (or properties of processed jobs) are modified to obtain approximations for the desired performance measures. From this category, we will focus on the Modified Offered Load approximation.

Modified Offered Load approximation

In 2.2.1, it was stated that the number of people in an infinite server system is Poisson distributed with a time dependent mean $m(t)$, for which an exact integral definition (and several simplifications in special cases) is known. Jagerman (1975) decided in 1974 to use the solution to such an infinite server system to obtain approximations for several performances of the finite server variant, laying the basis for the Modified Offered Load (MOL) approximation. With the MOL approach, the time dependent offered load in a non-stationary system with a finite number of servers is approximated by that of the same system but with an infinite number of servers, that is, the offered load in a finite server system is 'modified' to that of the corresponding infinite server system. Using this offered load of the infinite server system, an estimation for the desired performance measures of the finite server system can be calculated. After Massey and Whitt (1994) showed that such substitution of the infinite server offered load performed well when calculating the blocking probability for a $M(t)/G/c/c$, MOL has become a common approach to determine the impact of a time dependent arrival process, see for example Bekker and de Bruin (2010). According to Whitt (2007), this approach furthermore generalizes well to $M(t)/GI/s_t/r_t + GI$.

2.3 Staffing

So far, several approaches to analyse the performance of (time dependent) queueing systems are discussed. Although this discussion was done with the application of (time dependent) staffing in mind, this section will consider staffing in specific. A review of the use of time dependent queueing models for staffing decisions in service systems can be found in e.g. Feldman et al. (2008) Green et al. (2007) and Whitt (2007).

2.3.1 Staffing based on steady-state analyis

First of all, Green et al. (2001) mentions several scenarios in which stationary staffing over the time horizon could be inappropriate, among which healthcare settings with long service times. The convenience of then using a piecewise stationary approach, like SIPP, to determine staffing levels is that a time dependent staffing problem is reduced to a series of independent staffing problems for stationary models, according to Green et al. (2007). However, as mentioned before, several authors showed that MOL outperforms piecewise stationary methods for time dependent staffing, among which Jennings et al. (1996) and Green et al. (2007). For situations in which SSA, PSA and lagged PSA worked well, the MOL approximation was shown to behave approximately the same as those methods, while it performed significantly better in situations with longer service times. This should not be surprising as Thompson (1993) points out that with the piecewise stationary methods patients arriving towards the end of an interval are not taken into account in the next interval, while it is likely that their service is still continuing (especially with relatively long service times). Since MOL accounts for the time shift between the arrival process and the offered load, it is taken into account that an arrival can have an influence on multiple periods.

Square root staffing rule

Halfin and Whitt (1981), and later Jennings et al. (1996) for time dependent staffing, continue on the idea to approximate a finite server system by an infinite server system by introducing the so-called *square root staffing rule*. This square root staffing rule follows directly from the normal distribution by which a Poisson distribution can be approximated. The rule therefore originates from the observation that the mean number of busy servers in an infinite server system is Poisson distributed. The time dependent staffing rule is commonly given as

$$c(t) = m(t) + \beta\sqrt{m(t)}, \tag{12}$$

where $m(t)$ is the offered load of the infinite server system and β is a parameter reflecting the quality of service. According to Whitt (2007), experience suggests that an offered load of at least five already suffices for appropriate use of the square root staffing rule. The Quality of Service parameter β is in most literature determined based on a target delay probability. To deal with the trade-off between service level and efficiency, Bekker and de Bruin (2010) mention that β could also be determined with

$$\beta = \frac{c^* - \rho}{\sqrt{\rho}}, \tag{13}$$

where c^* is the desired capacity level and ρ is the average offered load over the complete period. Green et al. (2007) advises $\beta = 2$ as a simple rule of thumb to avoid congestion while limiting excessive capacity since $\beta = 2$ corresponds to a probability of delay of approximately 0.02. A more involved term to replace β is suggested by Borst et al. (2004). To obtain from (12) an actual staffing policy, most literature, e.g. Jennings et al. (1996) and Green et al. (2007), takes at every time t the least integer greater than or equal to the obtained $s(t)$, while Bekker and de Bruin (2010) round the obtained $s(t)$ to the nearest integer.

Interesting to mention is that Whitt (2007) explains how the time-dependent square root staffing rule can be useful within simulations. The search over staffing functions, as needed in every iteration of the simulation based Infinite Server Algorithm (ISA), is improved by limiting the candidate staffing functions to the ones satisfying the square root staffing rule. Using the service parameter β to range over, results in a one-dimensional search. The simulation based ISA is introduced by Feldman et al. (2008) for a $M(t)/G/s(t)+G$ system in their attempt to achieve targeted time-stable performance. They remark that the staffing and performance obtained with ISA closely agree with the MOL approach in case of no customer abandonment. Starting with an infinite server system, they estimate in every iteration for a given staffing function the distribution of the number of customers in the system at time t by performing multiple independent realisations over the full planning period. They then use the estimated distribution to determine the staffing function for the next iteration such that the desired delay probability is met. This iterative process stops when the change in consecutive staffing function is negligible.

2.3.2 Transient staffing

During the discussion of transient models, we already briefly mentioned that Assad et al. (1997) applies a transient model to solve a flexible staffing problem. His research is, to the best of our knowledge, one of the only that perform an analytic transient analysis per period to tackle a time dependent staffing problem. Buyukkaramikli et al. (2013) note this as well and speculate that the unpopularity of periodic capacity control might be due to the complexity of the necessary transient queueing behaviour. Nevertheless, Yoo sets up a finite horizon dynamic program in which the cost function is evaluated every period using a transient analysis. Due to the transient analysis, it can be studied what the influence of fixing a capacity in a certain period is on the build-up and spillover of congestion to the next period and thus on the capacity decision.

2.3.3 Using a cost function to determine the staffing level

The best staffing levels are often determined based on a specific performance measure, like the blocking or delay probability. For example Bekker and de Bruin (2010) and Li et al. (2016) select the minimum capacity for which a target delay probability is met. Less research focuses on the capacity for which minimal costs are obtained. Examples that do have the minimization of a cost function as main objective are Jennings et al. (1997) and Zychlinski et al. (2020). Zychlinski et al. (2020) also consider the topic of bed blocking in hospitals due to scarce capacity in geriatric facilities, especially geriatric institutions. In their analysis, they assume that there is a single organization that operates both the hospitals as well as the geriatric institutions and they aim to find the optimal number of beds for each geriatric ward such that the long-term total cost of care is minimized. This total cost is simply the sum of the underage costs multiplied by the bed shortage and overage cost multiplied by the bed surplus at every time t integrated over the planning horizon. They describe the underage cost as 'the amount that could have been saved if the level of geriatric beds had been increased by one unit in the event of an underage' (Section 4, p. 403) and is thus the difference in cost of hospitalization in the hospital compared to a bed in the geriatric institution, where hospitalization costs also include risk costs like expected costs for medical deterioration of the patient. Similarly, the overage costs are 'the amount that could have been saved if the level of geriatric beds had been reduced by one unit in the event of an overage' (Section 4, p. 403) and set to the daily cost of having staffed a geriatric bed. The cost components in this research will be similar to the underage and overage costs of Zychlinski et al. (2020). In addition to these two cost components, this research will consider profit (negative costs) on used beds to make the financial incentives of the VVT more apparent. The main difference between this research and Zychlinski et al. (2020) is how the proxy for the bed shortage, or the number of overflow beds, and the bed surplus, or the number of unused beds, that appear in the objective function are determined. Zychlinski et al. use a fluid model in which the hospital is explicitly modelled to capture the blocking effects. We refrain from explicit inclusion of the hospital, but instead use an overflow model with call packing to incorporate the blocking. The overflow model allows the time spent in overflow to influence the residual service time in the VVT organisation, which relation was not present in Zychlinski et al. (2020). Furthermore, Zychlinski et al. (2020) focuse on stationary capacities that can be reallocated periodically, whereas this research considers time dependent capacity for which, among other things, the constraint that capacity can only be changed at predetermined time points is implemented as well.

2.3.4 Practical remarks on the application of staffing results

Most papers on flexible staffing refrain from comments on the practicable feasibility of switching, especially decreasing, capacity, mostly because practical issues when decreasing capacity are part of operational decisions while setting capacity guidelines belongs to the tactical, or even strategical, level. Bekker and de Bruin (2010) do remark that opening and closing beds multiple times during the week may not be desirable so their staffing results should merely serve as guidelines. They have some more practical relevant observations, like that the range between the minimum and the maximum offered load gives an indication of the variability in the required number of beds. Furthermore, they give the rule of thumb that the fluctuations in arrival rate have limited impact on the staffing if the average service time exceeds five times the cycle length of the arrival rate function.

2.3.5 Applicability to and contribution of this research

Due to the lack of literature on transient analysis, the high computational time of transient analysis, and the fact that there are good alternatives for staffing available, we will address the time dependent staffing problem in this research by approaches based on steady-state analysis. Because the analytical model should serve as low fidelity model for multi-fidelity modelling, an approximation in low computational time is preferred. Since the Modified Offered Load approximation is shown to outperform piecewise stationary methods, the idea of approximating the offered load at time t by the time dependent offered load of an infinite server system, $m(t)$, will form the basis for the time dependent staffing analysis. Most research that considers time dependent staffing based on an infinite server approach, falls back on the time dependent staffing rule. The downside of the time dependent staffing rule is that the variety of constraints on the time dependent behaviour of the capacity that can be taken into account is limited. In this research, the infinite server approximation will serve as input for a Mixed Integer Program such that capacity constraints can be taken into account. Furthermore, the Modified Offered Load approximation will be applied to the call packing system. The resulting values of the performance measures will be used in combination with the Mixed Integer Program as well.

2.4 Structure of the VVT

The *Verpleeg-, Verzorgingshuizen en Thuiszorg (VVT)* organisations provide care for the elderly in the Netherlands, both in the form of nursing homes or geriatric facilities as well as home care for elderly people who can still live on their own. There are several types of care provided by VVT organisations. However, there is no clear general list of these types of care, and their possible subtypes. The types used in this research are based on a categorisation made in association with *Medisch Spectrum Twente* (Medical Spectrum Twente). This led to the following types: Crisis, ELV High Complexity, ELV Low Complexity, ELV Palliative, GRZ, Hospice, Psychogeriatric, Somatic, *Thuiszorg* (Home care). We will restrict ourselves to the three types of ELV, the GRZ and the three types of WLZ (Hospice, Psychogeriatric and Somatic). This will be discussed in more detail in the remainder of this section. The information given in this section is mostly based on information that can be found on the government site of the Dutch National Health Care Institute (*Zorginstituut Nederland*) [7], and a meeting that we have had with a finance director of the VVT organisation that provided the data for the numerical analysis in this research [6].

ELV (*eerstelijnsverblijf*, a stay in primary care) provides care and housing for patients who can temporarily not live at home for medical reasons. These patients do not need medical specialistic care. The main focus of ELV is providing time and space to recover in contrast to intensive therapy [1]. A stay via ELV is in general short. Care in ELV can either be of low complexity, of high complexity or palliative. The latter focuses on providing relief from the pain of other symptoms caused by a serious illness.

Just like ELV, GRZ (*geriatrische revalidatiezorg*, geriatric rehabilitation) provides care and housing intending to help the patient return to their home and participate in society as well as possible. However, the type of care that GRZ provides differs strongly from the care ELV provides. GRZ provides more complex care to vulnerable patients who deal with multiple illnesses or handicaps (multimorbidity) and have decreased learnability. Most patients receive GRZ after they undergo surgery or have been hospitalized for a stroke.

Both the ELV and GRZ are financed from the *Zorgverzekeringswet* (Health Insurance Act), which everyone in the Netherlands is legally obliged to take out. However, the coverages are determined in different ways. For the ELV, there is a fixed maximum coverage per day during the complete stay of the patient, which depends on the type of ELV care (LC, HC, palliative). In the GRZ, the coverage is specified for each care process. The realised length of stay of the patient only partially influences the amount of coverage: there is a fixed coverage for a range of days. For example, for two patients who received GRZ for two and seven days for the same care request, a VVT organisation gets the same coverage, namely the one corresponding to that care request with a length of stay of 1-7 days. So, depending on whether the realised length of stay of a patient is in the lower or upper part of such range, the VVT organisation can make a profit or a loss, respectively. The maximum coverages by the *Zorgverzekeringswet* are quite close to the costs that the care organisations make, meaning that the profit margin on care that is financed from the *Zorgverzekeringswet* is rather small.

The three other categories covered in this Section (Hospice, Psychogeriatric and Somatic) are all financed by the WLZ (*Wet Langdurige Zorg*, Long-Term Care Act). In this research, we will only refer to the collection of these types by their financing system WLZ. Patients in the WLZ need 24-hour supervision or care close-by and they will need this until the end of their lives. The coverage for providing care that is part of the WLZ is per day. The profit margins on WLZ care are larger than those on ELV and GRZ.

In contrast to most VVT care, hospital care is financed based on performance-based financing (*prestatiebekostiging*): a hospital gets a fixed amount of money for all care that they provide within one care request. The amount of money depends on the care request. The billing of such care activities is done via so-called *DBCs* (*diagnosebehandelcombinaties*, diagnosis-treatment combinations). A DBC contains all relevant care activities for a specific care request, like the diagnosis, treatments and possible follow-ups, as explained in [2]. Hospital care is in general significantly more costly than VVT care due to, for example, the presence of more advanced medical equipment and staff and a more complex infrastructure. In case a patient has to stay in the hospital perforce due to insufficient spots at a WLZ institution, the hospital can get a maximum coverage of 526.01 or 384.20 for each bed blocking day (*verkeerde beddag*), depending on the type of aftercare, according to Article 7 of [3] by the *Nederlandse Zorgautoriteit (NZa)*. Interesting to note is that this compensation per bed blocking day is only applicable when the patient is waiting for a spot care financed by the WLZ, and thus not when they are waiting for a spot in GRZ or ELV. Nevertheless, the use of the NZa prices or DBC prices for valuing care activities is advised against in Section 3.3 of a report commissioned by the National Health Care Institute (2024) due to possible influences of macrobudgetting and incomes policy.

3 Models

In this section, the mathematical models developed to solve the staffing problem at the VVT are discussed. The first part of this section concerns the derivation of the relevant performance measures for the objective function, being the mean number of patients in overflow and the mean number of unused beds. This part (Section 3.1) starts with a detailed description of the situation that is modelled in this research. After that, several possibilities for stationary models are discussed (Section 3.1.1) and an approximation method for the call packing model with general service times is introduced (Section 3.1.3). The validity of this method is theoretically analysed. The second part of this section (Section 3.2) covers the minimization of the cost function to obtain the optimal capacity. The cost function is explained for both the stationary as well as time dependent case (Section 3.2.1). To take into account constraints on the time dependent behaviour of the capacity in the time dependent staffing problem, a Mixed Integer Program is set up and explained (Section 3.2.2).

3.1 Queueing models

After treatment in the hospital, some patients have to go to a VVT institute. However, there may not be a place in the VVT institute immediately. In that case, the patient stays in the hospital until a spot at the VVT comes free. During this period of waiting, they block a hospital bed.

Remark 1. *We would like to make clear that there is an important difference between the patient blocking a hospital bed due to limited capacity at the VVT, and the patient being blocked from the VVT from a queueing theoretical perspective. In queueing theory, the term blocking is namely used to indicate that the patient is not able to enter the system and therefore leaves the system or, at least, does not wait for their service in the queue. The latter is not the case in this research, since we assume that all patients who can not enter the VVT immediately, will keep waiting (queueing) for VVT care, where the queue is situated in the hospital in reality. To avoid possible confusion about the use of blocking, we will refer to patients not being able to enter VVT as overflow and the beds they occupy in the hospital in that case as overflow beds.*

Since we are only interested in the overflow caused by the VVT and not in the characteristics of the hospital, and to avoid unnecessary complexity, the hospital will not explicitly be modelled. Although the service and the waiting take place at two distinct locations (VVT and hospital respectively), the complete ‘VVT model’ will be said to have infinite waiting room where the reader should keep in mind that all patients in the queue for the VVT in the model occupy in reality a bed in the hospital. An arrival to the VVT refers to a patient at the hospital being ready to go to the VVT, and it thus does not refer to a patient physically arriving at the VVT. Although arrivals to the VVT come from the hospital which can influence the arrival process, we will assume that the arrival process to the VVT is a Poisson process for tractability reasons. This assumption is checked with the available data in Section 4.

The capacity of all VVT organisations in the neighbourhood is considered to be pooled, where only a distinction in beds between different types of VVT care is made. The VVT organisations can therefore be modelled as R parallel VVT stations, one for each type of care. The capacity of a VVT station r will be denoted by c_r being the number of staffed beds. Patients that find all c_r beds occupied ‘on arrival’, will wait in the hospital if they just ended treatment in the hospital or will wait at home if they did not come from the hospital, i.e. if they were an external arrival. Under the assumption that external arrivals follow the same arrival and service distribution as arrivals from the hospital, the distinction between an arrival being from the hospital or external is only relevant for the optimization model. The inclusion of external arrivals will therefore be discussed more elaborately in Section 3.2.1. We will comment on the validity of the assumptions in the Discussion.

Due to the assumption of infinite capacity at the hospital, underlying the infinite waiting room for the ‘VVT model’, and the beds for each type of VVT being distinct, different types of VVT patients do not notice the existence of other types. The system therefore consists of R independent stations with infinite waiting room. During the remainder of this subsection, the analysis can therefore be limited to a single station.

Remark 2. *When stating that the system that we are concerned with has infinite waiting room, we implicitly assume that the influence of overflow patients on the working of the hospital is negligible. This is approximately the case if the hospital is sufficiently large and the number of overflow days is small. The latter is reasonable* *when studying the optimal capacity for the VVT organisations. Furthermore, the implicit assumption is made that the hospital itself is never at its full capacity. In the real world, the finiteness of capacity may come into play by the fact that elective patients are rejected when there is a lot of overflow from the VVT, which may in turn result in fewer arrivals later on to the VVT.*

In case the mainstream of patients goes to VVT after their current care or if the overflow time is substantial, the arrival rate to the VVT can not be assumed to be independent of the number of patients in overflow anymore. In that case, one might want to consider a state-dependent arrival rate to the VVT, or more precisely, an arrival rate that depends on the total number of patients in the queue. A short overview of an initial analysis of this case is given in Appendix A.7. Insights from this analysis can be of interest either in practice for a situation as described above, which might be realistic for flow from one specific VVT type to another, or just for theoretical purpose on its own, namely conditions for the existence of a product form distribution.

So far, we have narrowed down the modelling of the situation to the analysis of a station with a finite number of servers and infinite waiting room. However, due to the application area, the time in overflow will likely have an influence on the 'service time' in the VVT. Over time, especially under hospital supervision, most health statuses will naturally improve, although not as fast as in the presence of specialised VVT care. One could say that the patient thus receives some sort of service while waiting with a service rate lower than the service rate at the VVT, but higher than zero.

3.1.1 Overflow model and boundary cases

From a mathematical view, the situation just described is an overflow system with call packing with distinct service rates between overflow and 'normal service'. As elaborated on in the theoretical background, Van Dijk and Schilstra (2022) give a product form solution for the case of exponential service times, see formula (6). We should remark that in this analysis the call packing is assumed to be immediate, that is, at the moment a server comes free at the service station, an overflow job, if present, is instantaneously moved from the overflow station to the service station. As one can imagine, this is not entirely the case in reality due to administrative and logistical reasons.

Van Dijk and Schilstra (2022) analyse a two-station overflow system. This means that both stations provide regular service to jobs of their own type (station 1 serves jobs of type I and station 2 serves jobs of type II) and jobs of type I that arrive when all servers at station 1 are occupied are overflowed to station 2. Thus, the station to which jobs from the first station overflow serves regular (type II) jobs as well. In our research, no 'regular' jobs are served at the overflow station. So, formula (6) for a two-station overflow is simplified to a single-station overflow by simply removing the term belonging to the second station and noting that $N_1 = c$,

$$\pi(n,m) = \pi(0,0)F(m)\frac{1}{n!}\Big(\frac{\lambda}{\mu}\Big)^n, \text{ with} \tag{14}$$
$$F(m) = \begin{cases} \lambda^m / \prod_{k=1}^m (c\mu + k\gamma) & m > 0 \\ 1 & m = 0, \end{cases}$$

where the normalization constant $\pi(0,0)$ can be obtained by summing (14) over all states,

$$\begin{aligned}\pi(0,0)^{-1} &= \frac{1}{c!}\Big(ce^{\frac{\lambda}{\mu}}\Gamma(c,\frac{\lambda}{\mu}) + \Big(\frac{\lambda}{\mu}\Big)^c\frac{c\mu}{\gamma}e^{\frac{\lambda}{\gamma}}\Big(\frac{\lambda}{\gamma}\Big)^{-\frac{c\mu}{\gamma}}(\Gamma(\frac{c\mu}{\gamma}) - \Gamma(\frac{c\mu}{\gamma},\frac{\lambda}{\gamma}))\Big) \\ &= \frac{1}{c!}\Big(ce^{\rho}\Gamma(c,\rho) + \rho^c\frac{c\mu}{\gamma}e^{\sigma}\sigma^{-\frac{c\mu}{\gamma}}(\Gamma(\frac{c\mu}{\gamma}) - \Gamma(\frac{c\mu}{\gamma},\sigma))\Big),\end{aligned} \tag{15}$$

where we set $\rho = \frac{\lambda}{\mu}$ and $\sigma = \frac{\lambda}{\gamma}$.

It is important to realize that in the general theory about overflow, a service rate in overflow γ is used, as also appears in formula 14. However, for the application to this research, it does not make sense to have a known service rate in overflow, since patients in overflow are located in the hospital and receive no specific care there. It will depend on the patient and the type of VVT care they need, how much this 'normal' hospital care will contribute to their (natural) recovery. So, instead of a service rate in overflow γ, it is more natural to talk about a fraction $\kappa \leq 1$ that represents how much the time spent in overflow contributes to a reduction in service time in the VVT. For example, if every three days in overflow would reduce the expected time in the VVT by one day, we will say $\kappa = \frac{1}{3}$. From another view, the fraction κ represents how the service rates in VVT and overflow relate to each other. In the same example, the amount of service that is provided in three days in overflow is the same as in one day in the VVT and thus the service rate in VVT is three times the service rate in overflow. This illustrates the relation $\kappa = \frac{\gamma}{\mu}$, which is not surprisingly the factor Van Dijk and Schilstra give in the case of 'resume of service' to account for the difference in service speed. Due to this relation, the theory of overflow can be applied to this research with $\gamma = \kappa\mu$. The steady-state distribution for exponential service times with this notation becomes

$$\pi(n,m) = \pi(0,0)F(m)\frac{1}{n!}\rho^n, \quad \text{with} \tag{16}$$
$$F(m) = \begin{cases} \lambda^m / \prod_{k=1}^{m}((c+k\kappa)\mu) & m > 0 \\ 1 & m = 0, \end{cases}$$

where

$$\pi(0,0)^{-1} = \frac{1}{c!}\Big(ce^{\rho}\Gamma(c,\rho) + \rho^c\frac{c}{\kappa}e^{\frac{1}{\kappa}\rho}(\frac{1}{\kappa}\rho)^{-\frac{c}{\kappa}}(\Gamma(\frac{c}{\kappa}) - \Gamma(\frac{c}{\kappa},\frac{1}{\kappa}\rho))\Big). \tag{17}$$

In the case of general service distribution, we stated in the theoretical background that there is a difference between resuming and resampling the service in the case of repacking. Since the time in overflow is assumed to contribute to the improvement of the patient's health, and thus to cause a reduction in residual service time, the service time is resumed when the patient is repackaged. Consistency can be shown between the mathematical overflow analysis with overflow rate γ and resume of service and the intuitive reasoning with κ. In general call packing analysis, a job in overflow has a service time with mean $\frac{1}{\gamma}$. Given a time in overflow x, the residual service time in VVT after being repacked would (in expectation) be $\frac{\gamma}{\mu}(\frac{1}{\gamma} - x) = \frac{1}{\mu} - \frac{\gamma}{\mu}x = \frac{1}{\mu} - \kappa x$. The same residual service time is obtained when an arriving patient gets a service time drawn from a distribution with mean $\frac{1}{\mu}$, to which their time in overflow W_i contributes a fraction κ.

From the steady-state distribution (16) with normalization constant (17), an exact formula for the average number of patients in overflow can be found,

$$\begin{aligned} E(L_{\text{q}}^{\text{overflow}}) &= \sum_{k=0}^{\infty} k\pi(c+k) \\ &= \pi(0,0)\frac{1}{c!}\rho^c\sum_{k=0}^{\infty} k\frac{\lambda^k}{\prod_{i=1}^{k}((c+i\kappa)\mu)} \\ &= \pi(0,0)\frac{1}{c!}\rho^c\Big[\frac{1}{\kappa}\rho + \frac{1}{\kappa}\Big(\frac{1}{\kappa}\rho\Big)^{-\frac{c}{\kappa}}e^{\frac{1}{\kappa}\rho}(c-\rho)\Big(\Gamma(\frac{c}{\kappa}+1,\frac{1}{\kappa}\rho) - \Gamma(\frac{c}{\kappa}+1)\Big)\Big], \end{aligned} \tag{18}$$

where the complete gamma function is defined as $\Gamma(x) = (x-1)!$ and the upper incomplete gamma function $\Gamma(c+1,\rho)$ satisfies the recurrence relation

$$\Gamma(c+1,\rho) = c\Gamma(c,\rho) + \rho^c e^{-\rho} \qquad \text{with} \quad \Gamma(1,\rho) = e^{-\rho}. \tag{19}$$

In case $\frac{1}{\kappa}$ is integer, the recurrence of the incomplete gamma function can be exploited. Otherwise, the built-in function calculation of the gamma function in programming languages can be used.

The derivation of the expected number of unused beds is simpler since it only depends on the steady-state probabilities of the states $n \leq c$,

$$\begin{aligned} E(U^{\text{overflow}}) &= \sum_{n=0}^{c}(c-n)\pi(n) \\ &= c\sum_{n=0}^{c}\pi(n) - \sum_{n=0}^{c}n\pi(n) \\ &= \pi(0,0)\frac{\rho^{c+1} + e^{\rho}(c-\rho)\Gamma(c+1,\rho)}{c!}. \end{aligned} \tag{20}$$

In the explanation of the model, we argued that the time in overflow will contribute to the patient's health, leading to a reduction in residual service time when the patient is repacked, i.e., the fraction κ is between zero and 1. It is interesting what happens at the boundary cases: $\kappa = 0$ and $\kappa = 1$.

Special case: $M/M/c/\infty$**.** When κ equals zero, the time in overflow does not contribute to the patient's health at all. This could be the case for very specialised VVT care in which only time without the presence of this specialised care, e.g. in a hospital bed, does not cause significant improvements in the condition of the patient. The system then simplifies to a normal finite server system with infinite waiting room: $M/G/c/\infty$. For this system, a product form solution exists only in the case of exponentially distributed service times given by formula (2). Formula (5) gives an approximate relation between the queue length of the exponential case and in case of a general service distribution.

The expected number of patients in overflow can easily be derived from the steady-state distribution (2), and put in a more convenient form using the recurrence relation for the delay probability in a $M/M/c/\infty$ queue (4),

$$\begin{aligned} E(L_{\mathrm{q}}^{(M/M/c)}) &= \sum_{k=0}^{\infty} k\pi(c+k) = \frac{(c\rho)^c}{c!}\pi(0)\sum_{k=0}^{\infty} k\rho^k = \frac{\rho}{(1-\rho)^2}\frac{(c\rho)^c}{c!}\pi(0) = \frac{\rho}{(1-\rho)}\Pi_W \\ &= \frac{\rho}{(1-\rho)}\frac{\rho B(c-1,c\rho)}{1-\rho+\rho B(c-1,c\rho)}. \end{aligned} \tag{21}$$

Note that although ρ is defined as $\frac{\lambda}{c\mu}$, which depends on c, the recursion in the function $B(c, c\rho)$ has fixed term $c\rho = \frac{\lambda}{\mu}$. The latter is very convenient when calculating $E(L_{\mathrm{q}})$ for a range of different capacities c, since the computation $E(L_{\mathrm{q}})$ for $c+1$ only requires a single computation of $B(c, c\rho)$ via formula (3) instead of a complete summation or recursion from start.

Since ρ is the fraction of time each server is working, each server is expected to be empty $(1-\rho)$ fraction of time. The expected number of unused beds is thus simply

$$c(1-\rho) = c - \frac{\lambda}{\mu}. \tag{22}$$

That is, the difference between the capacity and the mean number of busy servers in a finite server system.

Special case: $M/G/\infty$**.** If the time in overflow has the same influence on the service time as the time in the VVT, the fraction $\kappa = 1$. This might sound remarkable, but it is approximately the case for VVT type WLZ (*Wet Langdurige Zorg*, Long-Term Care Act) in which almost all patients only end service when they die. For their service time, it does not matter if they lie in an overflow bed in the hospital or in a VVT bed. The indifference between being in service in the VVT or in the queue yields a $M/G/\infty$ queue. Besides a practical application of this special case, this special case is relevant due to its general applicability. First of all, the infinite server system is insensitive, i.e. holds for general distributed service times, while the call packing system only holds for exponential service times. Secondly, the analysis of infinite server systems with time dependent arrival rates is relatively straightforward, in contrast to that of a call packing system.

Let the offered load be denoted by $\rho = \frac{\lambda}{\mu}$. We will use the steady state distribution given in formula (1), the result that its expectation equals ρ and the fact that a probability distribution sums up to one, to derive a convenient form for the expected number of patients in overflow $E(L_{\mathrm{q}}^{(M/G/\infty)})$.

$$\begin{aligned} E(L_{\mathrm{q}}^{(M/G/\infty)}) &= \sum_{k=0}^{\infty} k\pi(c+k) \\ &= \sum_{k=0}^{\infty} [(c+k)\pi(c+k) - c\pi(c+k)] \\ &= \sum_{k=0}^{\infty}(c+k)\pi(c+k) - c\sum_{k=0}^{\infty}\pi(c+k) \\ &= \sum_{n=0}^{\infty} n\pi(n) - \sum_{i=0}^{c} i\pi(i) - c\Big(\sum_{n=0}^{\infty}\pi(n) - \sum_{i=0}^{c}\pi(i)\Big) \\ &= \rho - \sum_{i=0}^{c} i\pi(i) - c + c\sum_{i=0}^{c}\pi(i) \end{aligned}$$

$$
\begin{aligned}
&= \rho - c - e^{-\rho}\sum_{i=0}^{c} i\frac{\rho^i}{i!} + ce^{-\rho}\sum_{i=0}^{c}\frac{\rho^i}{i!} \\
&= \rho - c - e^{-\rho}\frac{\rho(e^{\rho}\Gamma(c+1,\rho) - \rho^c)}{c!} + ce^{-\rho}\frac{e^{\rho}\Gamma(c+1,\rho)}{c!} \\
&= \rho - c + \frac{\rho^{c+1}e^{-\rho} + (c-\rho)\Gamma(c+1,\rho)}{c!}, \qquad (23)
\end{aligned}
$$

where the incomplete gamma function $\Gamma(c+1,\rho)$ satisfies the recurrence relation (19). Again the recurrence eases the computational complexity of calculating the expected number of patients in overflow for a range of capacities c, as $\Gamma(c+1,\rho)$ follows from $\Gamma(c,\rho)$ for a fixed arrival rate and mean service time.

For the expected number of unused beds, note that the steady state probabilities of the states $n \leq c$ in the call packing system and infinite server system are the same up to the normalization constant. Since the normalization constant for an infinite server system is $e^{-\rho}$, it follows from formula (20) that

$$
E(U^{(M/G/\infty)}) = \frac{\rho^{c+1}e^{-\rho} + (c-\rho)\Gamma(c+1,\rho)}{c!}. \qquad (24)
$$

Remark 3. *A close reader might have recognized that the expression for $E(U^{(M/G/\infty)})$ in (24) appears in the expression for $E(L_q^{(M/G/\infty)})$ in (23). The latter can thus be rewritten to*

$$
E(L_q^{(M/G/\infty)}) = \rho - (c - E(U^{(M/G/\infty)})).
$$

Noting that subtracting the mean number of unused beds from the capacity yields the mean number of used beds, we see that the mean number of patients in overflow equals the offered load ρ minus the mean number of used beds. This result holds specifically for the $M/G/\infty$ queue since for this queuing type the offered load equals the mean number of patients in the system. The general relation between $E(L_q)$ and $E(U)$ is

$$
E(L) = E(L_q) + (c - E(U)),
$$

that is, the mean number of patients in the system is simply the sum of the mean number of patients in the queue and the mean number of unused beds, which can be derived from the definitions of the expectations,

$$
\begin{aligned}
E(L) &= \sum_{n=0}^{\infty} n\pi(n) = \sum_{n=0}^{c} n\pi(n) + \sum_{n=c+1}^{\infty} n\pi(n) \\
&= \sum_{n=0}^{c} n\pi(n) + \sum_{n=c+1}^{\infty}(n-c)\pi(n) + \sum_{n=c+1}^{\infty} c\pi(n) \\
&= \sum_{n=c+1}^{\infty}(n-c)\pi(n) + \sum_{n=0}^{c} n\pi(n) + c(1 - \sum_{n=0}^{c}\pi(n)) \\
&= \sum_{k=1}^{\infty} k\pi(c+k) + c - \sum_{n=0}^{c}(c-n)\pi(n) \\
&= E(L_q) + c - E(U).
\end{aligned}
$$

Remark 4. *We argued that the $M/M/c/\infty$ and $M/G/\infty$ are special cases of the call packing problem for $\kappa = 0$ and $\kappa = 1$, respectively. Indeed, it can be shown that for $\kappa = 0$, formula (14) simplifies to the steady state distribution of an $M/M/c/\infty$ queue. Setting $\kappa = 1$ in that same formula yields the steady state distribution of an $M/G/\infty$ queue. Moreover, the formula for average number of patients in the queue $E(L_q^{overflow})$, (18), can be shown to simplify to formula (23) for $E(L_q^{(M/G/\infty)})$ when $\kappa = 1$ by first noting that (15) simplifies to $\pi(0,0) = e^{-\rho}$.*
Interestingly, the product form solution for the call packing system is said to only hold for exponential service times, while the formulae for the infinite server system hold for general service times. This observation is in line with the remark by Van Dijk and Schilstra (2022) that strict insensitivity only holds for the special case $\gamma = \mu_1$.

3.1.2 Bounds on the call packing system

The $M/M/c$ and $M/G/\infty$ systems are special cases of the overflow system with call packing resume. The advantage of the infinite server system is that it is easy to compute and holds for general service distributions. Since the service times are often not exponential in reality, such insensitive approximations or bounds are of great importance. However, the expressions obtained for a $M/M/c$ or a $M/G/\infty$ might not be accurate for values of κ other than zero and 1, respectively.

Approximating the call packing system by a system with a finite number of servers ($M/M/c$) neglects the fact that the patient's health improves during their time in overflow, which would in expectation result in a shorter residual service time in the VVT than the modelled $E[S] = \frac{1}{\mu}$. The computed expected number of patients in overflow will therefore intuitively yield an upper bound for the true mean number of patients in overflow. On the other hand, approximating the call packing system by a $M/G/\infty$ system means that the improvement in health during the time in overflow is overestimated, resulting in an underestimation of the residual service time at the VVT. Thus, this will yield a lower bound on the expected number of patients in overflow. So far, intuitive arguments have been provided as to why the mean number of patients in overflow in the corresponding $M/G/c$ and a $M/G/\infty$ system are an upper bound and lower bound, respectively, on the mean number of patients in an overflow system. Below, formal proofs for these statements are provided.

Remark 5. *In the paragraph above, we switched from $M/M/c$ to $M/G/c$ to denote the finite server system. The computations for the finite server system, and the call packing system, in this report are limited to the case of exponential service distribution since product form solutions are only available for those. The proof of Theorem 1, however, holds for general service times.*

Let the call packing system with resume be denoted by Σ_C and the $M/G/\infty$ system and the $M/G/c$ system with the same arrival and service distribution as this call packing system be denoted by Σ_A and Σ_E, respectively. Comparison of the systems will be done via a sample path approach in a coupling argument, which is a common method to compare two different queueing systems, for example in Jouini and Dallery (2007). In this method, a sample path is studied in two coupled queueing systems, meaning that we look at what happens if patients arrive at the same moments with the same service requirement to each queueing system. In some cases, the queueing systems will behave the same, while for other patients a difference between the systems can be observed. The idea is that a general conclusion can be drawn from this different behaviour, for example, that the mean number of patients in one system is always larger than (or equal to) the mean number of patients in the other system. Formally, a sample path in coupled systems means that identical successive arrival epochs are created and identical realisations of service times are used. The latter is possible since the service rate at the VVT is equal to μ in all three systems.

Throughout this section, the term 'residual service time' will be used to refer to the time that it will cost to serve the residual service requirement of a patient at the service rate of the VVT, μ, determined at the moment the patient goes into service at the VVT. That is, the residual service time of patient i is $(S_i - \kappa W_i)^+$, where S_i denotes their initially drawn service time (when service would take place at rate μ) and W_i the time patient i had to wait in overflow.

Theorem 1. $E(L_q)^{(\Sigma_C)} \leq E(L_q)^{(\Sigma_E)}$

Proof. A sample path will be used to make sure that comparisons of service times are well-defined and not subject to the stochasticity of drawing service times from a distribution. So, both systems have identical successive arrival epochs and identical realisations of initial service times $S_1, S_2, \cdots$. Let $\Sigma_{\tilde{C}}$ be a slightly adapted version of Σ_C in which patients can not leave from overflow. That is, if a patient i is waiting in overflow for more than $\frac{1}{\kappa}S_i$, they stay there (instead of leaving) until it is their turn to go to a server in the VVT, from where they then can immediately leave since their residual service time is 0. Clearly, $E(L_\text{q})^{(\Sigma_C)} \leq E(L_\text{q})^{(\Sigma_{\tilde{C}})}$. This auxiliary system $\Sigma_{\tilde{C}}$ is more convenient to compare to Σ_E, since in both systems no patient can now leave from overflow. The residual service times of patients in both systems will be compared.

Since Σ_E is a normal $M/G/c$ system, the residual service times of each patient i is simply their initial service time S_i. In $\Sigma_{\tilde{C}}$, the initial service time S_i is reduced by the positive fraction κ of the time in overflow to obtain the residual service time of patient i, $(S_i - \kappa W_i)^+$. Since the time in overflow $W_i \geq 0$, each patient in $\Sigma_{\tilde{C}}$ has a (not necessarily strictly) smaller residual service time than its coupled patient in Σ_E. So, the mean amount of work that arrives per unit time to the VVT (ρ) in $\Sigma_{\tilde{C}}$ is smaller than or equal to that in Σ_E. Since the mean queue length is increasing in ρ, $E(L_\text{q})^{(\Sigma_{\tilde{C}})} \leq E(L_\text{q})^{(\Sigma_E)}$. By construction of $\Sigma_{\tilde{C}}$, $E(L_\text{q})^{(\Sigma_C)} \leq E(L_\text{q})^{(\Sigma_{\tilde{C}})}$, and thus $E(L_\text{q})^{(\Sigma_C)} \leq E(L_\text{q})^{(\Sigma_E)}$. □

Unfortunately, the same line of reasoning does not hold for the comparison of Σ_A and Σ_C, since the residual service time depends in both systems on the waiting time. Where we could conclude that the residual service time of a specific patient in Σ_C is always smaller than or equal to their residual service time in Σ_E independent of the value of their waiting time, such an ordering of the residual waiting time is not possible for Σ_A and Σ_C. If after a waiting time $W_i > 0$, a patient goes into service in both systems at the same time, their residual service time is strictly smaller in Σ_A than in Σ_C. However, if patient i can go into service in Σ_A upon arrival, while they have to wait in Σ_C, their residual service time in Σ_A (when going into service) is strictly larger than in Σ_C.

Intuitively, the fact that this patient could enter the VVT in Σ_A earlier than in Σ_C compensates for the smaller reduction in residual service time by the waiting time, but it would make the approach of the above proof invalid. Instead, the proof for $E(L_\text{q})^{(\Sigma_A)} \leq E(L_\text{q})^{(\Sigma_C)}$ will revolve around the observation that a patient arriving to both systems at the same time can never go into service in system Σ_C before they went into service in system Σ_A.

Theorem 2. $E(L_q)^{(\Sigma_A)} \leq E(L_q)^{(\Sigma_C)}$

Proof. A sample path approach in coupled systems will be used, i.e. both systems have identical successive arrival epochs, resulting in a sequence of predetermined arrival times $t_1, t_2, \cdots$ and identical realisations of initial service times $S_1, S_2, \cdots$. The subscript i is used to refer to the characteristics of the ith patient, for example, t_i is their arrival time and W_i their waiting time.

We will prove in a moment that the service in VVT of each patient starts in Σ_C only after (or at the same time as) it has started in Σ_A. Since each patient i arrived at the same time t_i to both systems due to the sample path approach, they have spent equal or more time in overflow. Taking the average of these individual times in overflow, it is obvious that the average time spent in overflow is then bigger in Σ_C than in Σ_A. It now follows from Little's law that the mean number of patients in overflow is bigger in Σ_C than in Σ_A, i.e. $E(L_\text{q})^{(\Sigma_A)} \leq E(L_\text{q})^{(\Sigma_C)}$.

Left to prove is the statement that the service in VVT of each patient starts in Σ_C only after or at the same time as it has started in Σ_A. First, it is important to realize that since Σ_A is an infinite server model, a patient i will always spend S_i time in system Σ_A which is clearly independent of the amount of time the patient was in overflow. In Σ_C, a patient spends after W_i time in overflow still $S_i - \kappa W_i$ in the VVT, under the condition that $W_i < \frac{1}{\kappa} S_i$, else they can leave from overflow after $\frac{1}{\kappa} S_i$ time in overflow. A patient spends thus in total $\min(S_i + (1-\kappa)W_i, \frac{1}{\kappa} S_i)$ time in system Σ_C. Note that both expression are larger than S_i, the time spent in Σ_A, since $0 < \kappa < 1$.

Let us now look at a specific patient which we call patient $i = X$, who arrives at time t_X to both systems. If patient X arrives when there are less than c patients in the system, they can immediately go into service in the VVT. Thus, the service in the VVT clearly starts in Σ_C at the same time as in Σ_A.

If upon arrival of patient X, there are $n \geq c$ patients in the system, all c VVT spots are occupied and thus patient X is in overflow. Let us refer to the n patients that were already present upon arrival of patient X by $i_1, i_2, \cdots, i_n$. Let it be clear to the reader that these n patients do not have to be patients that arrived consecutively in the sample path, but that it can very well be that patient $i = 6$ and patient $i = 8$ are still in the system while patient $i = 7$ already left, due to the stochasticity of the service time, i.e. $i_j + 1$ is not necessarily equal to i_{j+1}. Since patients are served in order of arrival, patient X can go into service in the VVT if $n - c$ patients of the patient set $\{i_1, i_2, \cdots, i_n\}$ have left the system. Each patient j of this set will leave system Σ_A at time $t_j + S_j$ and system Σ_C at time $t_j + \min(S_i + (1-\kappa)W_j, \frac{1}{\kappa} S_j) \geq t_j + S_j$, which holds independently of how long patient j has to wait as argued earlier. Denote the departure time of patient j in Σ_A and in Σ_C by $\tilde{t}_j$ and $\bar{t}_j$ respectively. Let $[\tilde{t}_{\tilde{i}_1}, \tilde{t}_{\tilde{i}_2}, \cdots, \tilde{t}_{\tilde{i}_n}]_{\Sigma_A}$ be a list of the departure times of $\{i_1, i_2, \cdots, i_n\}$ in Σ_A ordered from low to high, and similarly $[\bar{t}_{\bar{i}_1}, \bar{t}_{\bar{i}_2}, \cdots, \bar{t}_{\bar{i}_n}]_{\Sigma_C}$ for Σ_C. So, patient X can go into service in the VVT at time $\tilde{t}_{n-c}$ in Σ_A and at time $\bar{t}_{n-c}$ in Σ_C. Due to the ordering of the list of departure times and the individual relation $\bar{t}_j \geq \tilde{t}_j \forall j$, we obtain $\bar{t}_{n-c} \geq \tilde{t}_{n-c}$. That is, patient X goes into service in the VVT in Σ_A earlier than or at the same time as in Σ_C.

This completes the proof of $E(L_\text{q})^{(\Sigma_A)} \leq E(L_\text{q})^{(\Sigma_C)}$.

□

3.1.3 Adapted Service Time Approach

Since it is clear why the $M/G/c$ system and the $M/G/\infty$ system respectively overestimates and underestimates the mean number of patients in overflow, it is interesting if this knowledge could be used to adapt those systems in such a way that the resulting systems better mimic the call packing system. To approximate or bound the call packing system more accurately by a $M/G/c$ system or a $M/G/\infty$ system, we want to capture the influence of the overflow service in the service time of the VVT. This will be done through an effective, or adapted, service time.

Effective service times are common in literature on queueing theory to deal with the effect of blocking, for example Hunt (1956); Hillier and Boling (1967); Perros and Altiok (1986); Koizumi et al. (2005); Osorio and Bierlaire (2009). Blocking after service occurs when the next station on the route is full. The additional time a job spends at station i due to being blocked by station $i+1$ is added to the normal service time at station i to constitute the effective service time of station i. However, another form of the concept of effective service is necessary for application to the situation dealt with in this research, since the influence of overflow/blocking at the hospital (station i) due to reached capacity limit at the VVT (station $i+1$) should now be incorporated in the service time at the VVT (station $i+1$), instead of the hospital (station i). To avoid confusion, the adapted service time will be referred to as such, and not as an effective service time.

Recall that μ is the service rate in the VVT and that a fraction κ captures the influence of the time in overflow on the reduction in service time in the VVT. Suppose that, in steady-state, a patient spends on average some known time x in overflow. By definition of κ, they thus still need to spend on average

$$\frac{1}{\mu} - \kappa x \tag{25}$$

time in the VVT, given an overflow time x. Using (25), the patient is expected to have spent in total

$$x + \frac{1}{\mu} - \kappa x = \frac{1}{\mu} + (1-\kappa)x \tag{26}$$

time in the system, defined as the moment of wanting to transfer to the VVT up until leaving the VVT.

To understand what a suitable adapted service time for the $M/M/c$ and $M/G/\infty$ systems would be, it is important to once again point out which simplifying assumptions these systems make compared to the overflow system. With the $M/M/c$ system, patients do not receive any service in overflow. If it is known that a patient has spent x days in overflow before 'starting' to service at the VVT, their expected service time is reduced to $\frac{1}{\mu} - \kappa x$. So, in the $M/M/c$ system it is better to use an adapted service time of (25). For $\kappa = 0$, the adapted service time simplifies to $\frac{1}{\mu}$, as it should since a call packing system without service in overflow is simply a $M/M/c$ system.

A $M/G/\infty$ system, on the other hand, acts as if patients receive the same service in overflow as they get in the VVT. The expected total time in the system should match with the call packing case, and therefore an adapted service time of (26) should be used. The second expression immediately shows that the patient expected adapted service time, which is assumed to happen completely at a rate μ in an $M/G/\infty$, is a normal service time lasting $\frac{1}{\mu}$ in expectation and a compensation factor for the fact that in reality a time x is executed at a lower rate in overflow. The $M/G/\infty$ system is a special case of the call packing when $\kappa = 1$, so when we take $\kappa = 1$ the effective service time indeed simplifies to $\frac{1}{\mu}$.

So far, we assumed that x, the time that a patient spends on average in overflow in steady-state, was known. By Little's Law [84], the mean time in overflow (waiting time) is the mean queue length divided by the arrival rate λ,

$$E(W_\text{q}) = \frac{1}{\lambda}E(L_\text{q}). \tag{27}$$

Remark 6. *One could argue that when applying Little's Law to overflow, the proper arrival rate is the arrival rate to overflow which is $\lambda P(overflow)$. In this way, the mean overflow time given that someone is in overflow is obtained, namely $\frac{E[L_q]}{P(overflow)\lambda}$. However, for x we are interested in the mean time in overflow of a general patient. Conditioning on the fact that this patient is in overflow or not gives an overflow time of $\frac{E[L_q]}{P(overflow)\lambda}$ or zero, respectively. Since the probability of overflow is $P(overflow)$, deconditioning yields a mean overflow time of $P(overflow)\frac{E[L_q]}{P(overflow)\lambda} + (1 - P(overflow))0 = \frac{E[L_q]}{\lambda}$. Thus, conditioning and deconditioning show that it is not necessary to take $P(overflow))$ explicitly into account since it cancels out.*

The problem we are left with is determining the mean number of patients in overflow $E(L_\text{q})$, after which $x = \frac{1}{\lambda}E(L_\text{q})$. The unknown mean number of patients in overflow could simply be estimated by the number of patients in overflow in the $M/M/c$ or $M/G/\infty$ system, given by formulae (21) and (23) respectively. However, we can go somewhat further by repetitively applying the formula for the mean number of patients in an $M/M/c$ ($M/G/\infty$) system to the resulting system until the value does not change anymore. Formally, we are looking for the fixed point to the set of equations which consists of the formula for the mean queue length given an offered load (and capacity) and the formula for the adapted offered load given a queue length. That is, we are interested in the solution $E(L_\text{q})$ of the fixed point equation $E(L_\text{q}) = T(E(L_\text{q}))$, which we will write shorthand as $L = T(L)$, where $T(L) = (Q \circ R)(L) = Q(R(L))$ with $Q(\rho)$ the function to calculate the mean queue length given an offered load ρ and $R(L)$ a function to calculate the adapted offered load based on the mean queue length L. The precise formulas for $Q(\rho)$ and $R(L)$ depend on the type of queueing system. We will use subscripts 1 and 2 to refer to the $M/G/\infty$ system and $M/M/c$ system respectively. The formulas for the adapted offered loads were not stated explicitly yet, but are for the $M/G/\infty$ system,

$$R_1(L) = \lambda_\text{VVT}(\frac{1}{\mu} + (1-\kappa)\frac{1}{\lambda}L) = \frac{\lambda_\text{VVT}}{\mu} + (1-\kappa)\frac{\lambda_\text{VVT}}{\lambda}L, \tag{28}$$

and for the $M/M/c/\infty$ system,

$$R_2(L) = \lambda_\text{VVT}(\frac{1}{\mu} - \kappa\frac{1}{\lambda}L) = \frac{\lambda_\text{VVT}}{\mu} - \kappa\frac{\lambda_\text{VVT}}{\lambda}L. \tag{29}$$

Please note that the offered load for the $M/M/c/\infty$ with this definition should be lower than c for stability. This is in contrast to the earlier definition in which it is divided by c and therefore had to be below one. Since the offered load ρ in formula (21) for the mean queue length in the $M/M/c$ system was defined as $\frac{\lambda}{c\mu} < 1$, while we would like to use $R_2(L) < c$ as input, we reformulate formula (21) to

$$Q_2(\rho) = E(L_\text{q}^{(M/M/c)}) = \frac{\rho}{(c-\rho)}\frac{\rho B(c-1, c\rho)}{c - \rho + \rho B(c-1, c\rho)}, \tag{30}$$

where $\rho = \frac{\lambda}{\mu}$. The precise formulas for $Q(\rho)$ and $R(L)$ are given by formulas (23) and (28) for the $M/G/\infty$ queue and by formulas (30) and (29) for the $M/G/c$ queue.

To calculate the offered load, an arrival rate of λ_VVT is used, which is the arrival rate to the VVT. Note that this is not necessarily the same as λ, since λ is defined as the arrival rate of patients that want to go to the VVT (regardless if they can enter immediately or should go in overflow) and it can happen that a patient already ends their service in overflow after which they leave without entering the VVT itself. Remarkable, Van Dijk and Schilstra (2022) do not mention this possible deviation in arrival rates. Since for most applications, the probability that a patient can leave from overflow is small, we assume for tractability and simplicity that $\lambda_\text{VVT} \approx \lambda$ during the numerical application, see also Remark 7.

Remark 7. *An exact expression for λ_{VVT} can be derived, from the relation to several factors appears. A fraction $1 - P(\textit{in overflow})$ of the arriving patients can go in service in the VVT immediately while the rest will go into overflow. From the latter group, only the patients that do not finish their service while being in overflow, flow through to VVT. This yields the following formula for the true arrival rate to VVT,*

$$\begin{aligned} \lambda_{VVT} &= (1 - P(\textit{in overflow}))\lambda + P(\textit{in overflow})(1 - P(\textit{finish service in overflow}))\lambda \\ &= \lambda(1 - P(\textit{in overflow})P(\textit{finish service in overflow}))) \\ &= \lambda(1 - P(n > c)P(\textit{service time} < \kappa\ \textit{waiting time})). \end{aligned}$$

It is clear that if the probability of ending service in overflow is reasonably small, i.e. if both the probability of going to overflow as well as the time in overflow are small, it is reasonable to assume $\lambda_{VVT} \approx \lambda$. Since the goal of this study is to find the optimal capacity, we are most interested in situations in which the capacity is close to optimal meaning little overflow and short overflow times. Moreover, the defining formula for λ_{VVT} shows that a small value of κ can also help to have $\lambda_{VVT} \approx \lambda$.

We must remark that in reality, it does happen that patients leave from overflow, mostly by dying or transferring them to another hospital. We did not account for this in the formula above. This could affect the validity of $\lambda_{VVT} \approx \lambda$, and more specifically, of the assumption that the fraction of patients leaving from overflow is small. Suggestions for further research are made in the Discussion.

Intending to approximation the mean number of patients in overflow in a call packing system, we studied the adaptation of service times in a $M/G/\infty$ and $M/M/c$ system and derived a fixed point equation $L = T(L)$ yielding the mean number of patients in the adapted systems. For this approach to make sense, it is important to establish the existence and uniqueness of these fixed points. After that, we will comment on the usefulness of this approximation.

Existence and uniqueness of fixed point for the $M/G/\infty$ queue

An important observation in order to prove the existence and uniqueness of a solution to the fixed point equation $L = T_1(L)$, for the $M/G/\infty$, is that the mean queue length is increasing in the load. Although we can imagine that this is intuitively straightforward, a formal proof is given using the derivative.

Lemma 1. *The function $Q_1(\rho)$, as given in 23, is monotone increasing in ρ.*

Proof. The derivative of $Q_1(\rho)$ with respect to ρ is strictly positive for all ρ,

$$Q_1'(r) = 1 + \frac{r^c e^{-r} - \Gamma(c+1,r)}{c!} = 1 - \frac{\Gamma(c,r)}{\Gamma(c)} > 0, \tag{31}$$

where it is used that $\frac{\partial \Gamma(c,r)}{\partial r} = -r^{c-1}e^{-r}$, $\Gamma(c+1,r) = c\Gamma(c,r) + r^c e^{-r}$ and $(c-1)! = \Gamma(c)$. The last inequality follows from the fact that the gamma function $\Gamma(c) = \Gamma(c+1,r) + \gamma(c+1,r)$, where $\gamma(c+1,r)$ is the lower incomplete gamma function which is positive for $c, r > 0$, and thus $\Gamma(c) > \Gamma(c+1,r)$. □

Remark 8. *Since the probability of going into overflow $\sum_{n=c}^{\infty} \pi(n) = \sum_{n=c}^{\infty} e^{-\rho}\frac{\rho^n}{n!}$ can be rewritten to $1 - \frac{\Gamma(c,r)}{\Gamma(c)}$, we see that the derivative of the mean number of patients in overflow in a $M/G/\infty$ queue equals the probability of being in overflow. A directer proof starting from $\sum_{n=c}^{\infty} \pi(n)$ can be found in Appendix A.6.*

The existence and uniqueness of a fixed point $L = T_1(L)$ will now be proven by Banach's fixed point theorem, based on the observation that T_1 is a contraction mapping. First, the necessary definitions and theorem will be stated which can be found in almost all basic textbooks on Scientific Computing, for example, Granas and Dugundji (2003).

Definition. *Let $F : (X,d) \to (X,d)$ be a mapping of metric spaces. The smallest fixed constant M for which $d(F(x), F(y)) \leq M d(x,y)$, $\forall x, y$ is the Lipschitz constant K_F of F. If $K_F < 1$, the mapping F is called a contraction, and if $K_F \leq 1$, it is said to be nonexpansive. For any given $x \in X$, define the nth iterative of x under F, $F^n(x)$, inductively by $F^0(x) = x$ and $F^{n+1}(x) = F(F^n(x))$.*

The Banach contraction principle, or Banach fixed point theorem, now states

Theorem (Banach fixed point theorem). *Let (X,d) be a complete metric space and $F : X \to X$ be a contraction. Then F has a unique fixed point u, and $F^n(x) \to u$ for each $x \in X$.*

Note that for this application the space X is the space of real numbers and the metric d is the absolute value. Since T_1 is a composition of two mappings, we will study the Lipschitz constant separately and combine the results using the following proposition.

Proposition 1. *On the metric space (X,d), the composition $(B \circ A) : X \to X$ of a contraction mapping $A : X \to X$ and a nonexpansive mapping $B : X \to X$ is a contraction mapping.*

Proof. Since A is a contraction mapping, there exists a real number $0 \leq K_A < 1$ such that for all $x, y \in X$, $d(A(x), A(y)) \leq K_A d(x,y)$. Since B is a nonexpansive mapping, it holds for all $x, y \in X$ that $d(B(x), B(y)) \leq d(x,y)$. All $x, y \in X$ thus satisfy $d(B(A(x)), B(A(y))) \leq d(A(x), A(y)) \leq K_A d(x,y)$, proving that $B \circ A$ is a contraction mapping. □

Theorem 3. *The fixed point equation $L = T_1(L)$ has a unique solution. Moreover, the sequence of iterates $L^{(n)} = T_1(L^{(n-1)})$ resulting from repeated substitution of the operator T_1 converge to this unique fixed point L^*, independent of the starting point $L^{(0)}$.*

Proof. Recall that $T_1(L) = Q_1(R_1(L))$. Clearly, R_1 is a contraction mapping as

$$\begin{aligned} |R_1(x) - R_1(y)| &= |\frac{\lambda_{\text{VVT}}}{\mu} + \frac{\lambda_{\text{VVT}}(1-\kappa)}{\lambda}x - \frac{\lambda_{\text{VVT}}}{\mu} + \frac{\lambda_{\text{VVT}}(1-\kappa)}{\lambda}y| \\ &= |(1-\kappa)\frac{\lambda_{\text{VVT}}}{\lambda}(x-y)| \\ &\le (1-\kappa)|x-y|, \end{aligned}$$

where $K_{R_1} = (1-\kappa)$ satisfies $0 \le K_{R_1} < 1$ since $0 < \kappa \le 1$.

To show that Q_1 is a nonexpansive mapping, the Mean Value Theorem is used, which states that $\left|\frac{Q_1(x)-Q_1(y)}{x-y}\right| = |Q_1'(\eta)|$ for some point η in the domain. From equation (31) it can be concluded that $Q_1'(\rho) < 1$ for all $\rho > 0$, since $\Gamma(c, r) > 0$ for $c, \rho > 0$. Therefore, $|Q_1'(\eta)| < 1$ for all η in the domain. By the Mean Value Theorem $\left|\frac{Q_1(x)-Q_1(y)}{x-y}\right| < 1$, i.e. $|Q_1(x) - Q_1(y)| < |x-y|$. So, Q_1 is a nonexpansive mapping.

It now follows from Proposition 1 that T_1 is a contraction mapping. Since the metric space $(\mathbb{R}, |.|)$ is complete, the conditions for Banach fixed point theorem are met. So, the fixed point $L_1^* = T(L_1^*)$ exists and is unique. Moreover, convergence to this fixed point is guaranteed when iteratively applying the mapping T_1 starting from any point in the domain. □

Existence and uniqueness of the fixed point for the $M/M/c$ system

The proof for the existence and uniqueness of a solution to the fixed point equation $L = T_2(L)$, for the $M/M/c$ system, is inspired by the proof of Ross (2011) for a fixed point blocking probability in the reduced load approximation for single-service networks. Ross first shows existence by Brouwer's fixed point theorem for the Euclidean space.

Theorem. *(Brouwer's fixed-point theorem) Every continuous function from a nonempty convex compact subset X of a Euclidean space to X itself has a fixed point*

Uniqueness is then proven by showing that all solutions to the fixed-point equation minimize a convex optimization problem. By uniqueness of the minimum of a convex optimization problem, all solutions are the same.

Since the quantity of interest here, the mean queue length, has no clear upper bound, the domain of the mapping T_2 is not a compact set and thus Brouwers theorem does not apply directly to the mapping T_2. Note that the offered load $\rho = \frac{\lambda}{\mu}$, on the other hand, does have a clear bounded interval, namely $[0, c]$, assuming stability. Since T_2 is a composition of $Q_2(\rho)$ and $R_2(L)$, we will define another composite mapping $U_2(\rho) = R_2(Q_2(\rho))$. In Lemma 2, this latter mapping is proven to have a fixed point. In Lemma 3, a fixed point of this latter mapping is shown to imply a fixed point of the original mapping T_2. The existence of a fixed point $L = T_2(L)$ in Theorem 4 then follows naturally from these Lemmas, after which the uniqueness of this fixed point will be proven.

Lemma 2. *The equation $\rho = U_2(\rho)$ has a solution, where $U_2(\rho) := R_2(Q_2(\rho))$*

Proof. Since the mean queue length $Q_2(\rho)$ and the fraction $\frac{\kappa\lambda_{\text{VVT}}}{\lambda}$ are non-negative, $U_2(\rho) \le \frac{\lambda_{\text{VVT}}}{\mu}$. Since we assumed that the system is stable, the initial offered load $\rho = \frac{\lambda}{\mu}$ is smaller than c. So, the function $U_2(\rho)$ maps a convex set $[0, c]$ to itself. Furthermore, $U_2(\rho)$ is continuous by continuity of the two composed functions. It therefore follows from Brouwer's fixed point theorem that there exists a solution to $\rho = U_2(\rho)$. □

Lemma 3. *The existence of a fixed point $\rho^* = U_2(\rho^*)$ implies the existence of a fixed point $L = T_2(L)$*

Proof. Let ρ^* be the solution to the fixed point equation $\rho^* = U_2(\rho^*)$. Let L^* be the queue length resulting from the fixed point ρ^*, i.e. $L^* = Q_2(\rho^*)$. We then have,

$$L^* = Q_2(\rho^*) = Q_2(U_2(\rho^*)) = Q_2(R_2(Q_2(\rho^*))) = T_2(Q_2(\rho^*)) = T_2(L^*),$$

where the second equality follows from the fact that ρ^* is a fixed point of U_2 and the fourth equality from the associativity of the composite function. So, the mean queue length $L^* = Q_2(\rho^*)$ satisfies the fixed-point equation $L = T_2(L)$. The existence of ρ^* implies the existence of L^*, and thus of the existence of a fixed point of $L = T_2(L)$. □

Theorem 4. *The fixed point equation $L = T_2(L)$ has a unique solution.*

Proof. The existence of a solution follows directly from Lemmas 2 and 3.

To prove uniqueness, define the function $Q_2^{-1}(L)$ as the inverse of formula (21) (for fixed capacity c). That is, $Q_2^{-1}(L)$ is the value of the load ρ such that $L = Q_2(\rho)$. Note that since (21) is strictly increasing in ρ, the function $Q_2^{-1}(L)$ is a strictly increasing function of L. The integral $\int_0^L Q_2^{-1}(z)dz$ is thus a strictly convex function of L, based on the fact that the derivative of a strictly increasing function is positive, and the fact that a twice differentiable function is convex if its second derivative is positive.

Applying $Q_2^{-1}(L)$ to both sides of the fixed point equation $L = T_2(L)$ yields

$$Q_2^{-1}(L) = Q_2^{-1}(Q_2(R_2(L))) = R_2(L) = \lambda_{\text{VVT}}(\frac{1}{\mu} - \kappa\frac{1}{\lambda}z). \tag{32}$$

So, all solutions to the fixed point equation will also satisfy (32). Define for all $L \geq 0$, the auxiliary function

$$\psi(L) = \int_0^L Q_2^{-1}(z)dz + (-\lambda_{\text{VVT}}(\frac{1}{\mu}z - \kappa\frac{1}{2\lambda}z^2)).$$

Note that $\psi(L)$ is strictly convex since it is the sum of two strictly convex functions. The first term is strictly convex by monotonicity of $Q_2^{-1}(L)$, and the second term has strictly positive second derivative $\frac{\kappa\lambda_{\text{VVT}}}{\lambda}$. Let L^* be a solution to the stationary condition $\frac{d\psi(L)}{dL} = 0$. Since $\psi(L)$ is strictly convex, the minimizer L^* is unique. By construction of $\psi(L)$, all points satisfying the stationary condition $\frac{d\psi(L)}{dL} = 0$ will satisfy (32), and thus will be a fixed point, and vice versa. Since the minimizer L^* of $\psi(L)$ is unique, the solution to the fixed point equation is unique as well. □

Computation of the fixed point

The existence and uniqueness of the fixed points are shown, both for $M/G/\infty$ and $M/M/c$ system. However, this does not tell us how to numerically compute the values of the fixed points. Unfortunately, no closed-form expression for these fixed points could be derived, so another approach is necessary.

A common method to find a (unique) fixed point is repeated substitution. In repeated substitution, as the name suggests, the outcome of a function is repeatedly substituted in the same function until consecutive outcomes do not change too much. It was precisely this intuition that led to the definition of the adapted service times through a system of fixed point equations. For the $M/G/\infty$ system, the fact that the sequence converges to the unique fixed point already followed from the Banach fixed point theorem. For the $M/M/c$ system, the sequence is shown to have two convergent subsequences, where one sequence converges to L^- and the other to L^+ with $L^- \leq L^* \leq L^+$. That is, the complete mapping of which we want to find the fixed point has a periodic orbit of size 2. The proof of this convergence can be found in Appendix A.1. Numerical analysis shows that these two subsequences converge to the same value (L^*) for some parameter settings while for other parameter settings the points of convergence L^- and L^+ are different, as illustrated for ranging over κ in Figure 1. We hypothesize that a decrease in κ leads to $L^- = L^* = L^+$ due to the changes in adapted service time being less extreme. Moreover, an increase in μ also results in $L^- = L^* = L^+$ more often. This is illustrated by the fact that when μ is increased from $\frac{1}{20}$ to $\frac{1}{15}$ for the parameter settings of Figure 1, we observe $L^- = L^+$ for every $0 < \kappa < 1$. Due to limited time for this research and limited practical added value, no further research into this 2-cycle behaviour is done. Further research could be done to determine what constraints on the set of parameters would guarantee the two convergent sub-sequences to converge to the same value.

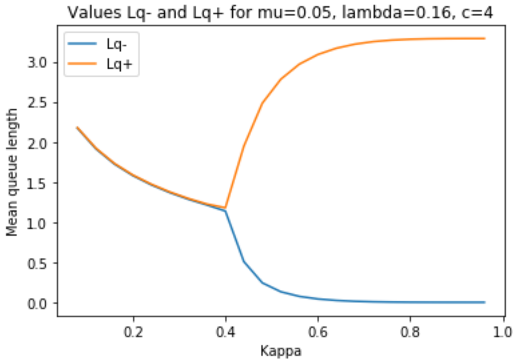


FIGURE 1: The values of the two convergence points L^- and L^+ with respect to κ. We see that for $\kappa < 0.35$, $L^- = L^+$

.

Bounds of adapted service time system

In this section, we already provided the intuition and formally proved that analysis with the normal $M/G/\infty$ and $M/G/c$ will yield a lower and upper bound, respectively, on the mean number of patients in overflow in the call packing system. To substantiate that the Adapted Service Time Approach yields a better bound on the call packing system, a necessary condition is that its mean queue length is respectively higher and lower than the mean queue length obtained by the corresponding $M/G/\infty$ and $M/M/c$ system. The comparison of the mean number of patients in overflow in the normal $M/G/\infty$ and $M/M/c$ systems and in their respective variants with adapted service times is rather straightforward. In line with Theorems 1 and 2, we will refer to the normal $M/G/\infty$ and $M/M/c$ system by Σ_A and Σ_E respectively and to their corresponding systems with adapted service time by Σ_B and Σ_D.

Theorem 5. $E(L_q)^{(\Sigma_A)} \leq E(L_q)^{(\Sigma_B)}$

Proof. Observe that Σ_A and Σ_B are both $M/G/\infty$ queues that only differ in their mean service rate. The mean service rate in Σ_B, $\tilde{\mu_1}^* = \frac{1}{\frac{1}{\mu}+\frac{1-\kappa}{\lambda}L^*}$, is smaller than the service rate in Σ_A, $\mu = \frac{1}{\frac{1}{\mu}}$, since $\frac{1-\kappa}{\lambda}L^* \geq 0$. The desired inequality now follows from the fact that the mean queue length is increasing in the load ρ and thus decreasing in the service rate. □

Theorem 6. $E(L_q)^{(\Sigma_E)} \geq E(L_q)^{(\Sigma_D)}$

Proof. Observe that Σ_D and Σ_E are both $M/M/c$ queues that only differ in their mean service rate. The mean service rate in Σ_D, $\tilde{\mu_2}^* = \frac{1}{\frac{1}{\mu}-\frac{\kappa}{\lambda}L^*}$, is bigger than the service rate in Σ_E, $\mu = \frac{1}{\frac{1}{\mu}}$, since $\frac{\kappa}{\lambda}L^* \geq 0$. The desired inequality now follows from the fact that the mean queue length is increasing in the load ρ and thus decreasing in the service rate. □

Referring to the call packing system by Σ_C, we showed so far that $E(L_\text{q})^{(\Sigma_A)} \leq E(L_\text{q})^{(\Sigma_C)} \leq E(L_\text{q})^{(\Sigma_E)}$ and that $E(L_\text{q})^{(\Sigma_A)} \leq E(L_\text{q})^{(\Sigma_B)}$ and $E(L_\text{q})^{(\Sigma_D)} \leq E(L_\text{q})^{(\Sigma_E)}$. Of course, the most interesting is the comparison of the mean number of patients in overflow in the call packing system Σ_C and the mean number of patients in the adapted service systems. Based on a numerical study that can be found in Appendix A.2, we hypothesize that the infinite server system with adapted service time Σ_B gives a lower bound to the call packing system for every service distribution and that the adapted $M/M/c$ system Σ_D yields an upper bound in case of an exponential service time distribution. Unfortunately, the sample path approach in coupled systems as used in Section 3.1.2 does not work for the comparison of Σ_B (and Σ_D) with Σ_C. Due to the adapted service rate in Σ_B, the two systems have different mean service times implying that we can not have the same sample path with identical service time realisations for the two systems. The fact that the service time in Σ_B (and Σ_D) depends on the fixed point L^*, for which no closed form expression is available and for which the relation to (the mean of) individual waiting times in the call packing system is not known, makes it impossible to split the draw of service time in a part which has mean equal to that in the call packing system and a part that has the (possibly negative) additional service time. Although several attempts were made, we were not able to provide a formal proof for the relation between the mean number of patients in overflow in Σ_B, Σ_D and Σ_C within the period of this research. Therefore, we will state our hypotheses based on the numerical analysis as conjectures.

Conjecture 1. $E(L_q)^{(\Sigma_B)} \leq E(L_q)^{(\Sigma_C)}$ *for general service distributions*

Conjecture 2. $E(L_q)^{(\Sigma_C)} \leq E(L_q)^{(\Sigma_D)}$ *for exponential service distributions*

Especially, Conjecture 1 can have great importance since it provides an insensitive lower bound to the call packing system which is tighter than the one obtained from the normal $M/G/\infty$ system by Theorem 5.

Mean number of unused beds

So far, we only focussed on the mean number of patients in overflow. The mean number of unused beds in both a $M/M/c$ system as well as in a $M/G/\infty$ system can be expressed solely in terms of the mean queue length, the offered load and the capacity, as given by formulas (22) and (3), respectively. The offered load belonging to the adapted system is given by $\rho^* = R(L^*)$, in line with Lemma 3. The mean number of unused beds of the adapted systems is thus $E(U)^{(\Sigma_B)} = E(L_\text{q})^{(\Sigma_B)} + c - \rho_1^*$ for the adapted $M/G/\infty$ system and $E(U)^{(\Sigma_D)} = c - \rho_2^*$ for the adapted $M/M/c$ system.

3.1.4 Time dependence

As explained in Section 2.3.5, the time dependent analysis in this research will be based on the Modified Offered Load approximation. The MOL approximation uses the time dependent offered load $m(t)$ as defined in Section 2.2.1. Since the offered load in the infinite server system loses its insensitivity property when made time-dependent, $m(t)$ depends on the complete service time distribution. Since only the mean is changed in the Adapted Service Time Approach, this approach can not naturally be extended to time dependent analysis without rigorous assumptions, as will be pointed out in the Discussion.

We have already proven that in steady state the mean number of patients in overflow in the call packing system is bounded below by that in the corresponding infinite server system. Since both the number of patients in an $M(t)/G/\infty$ with time dependent arrival rate as well as the number of patients in the stationary case are Poisson distributed with mean $m(t)$ and ρ, respectively, the stationary derivation for the expected number of patients in overflow and number of unused beds remains valid at each time point t when ρ is replaced by $m(t)$. The expected number of patients in overflow at time t and the expected number of unused beds at time t are therefore obtained by simply replacing ρ by the time dependent offered load $m(t)$ in formulas (23) and (24), respectively,

$$E(L_\text{q}^{(M/G/\infty)}(t)) = m(t) - c + \frac{(m(t))^{c+1}e^{-m(t)} + (c - m(t))\Gamma(c+1, m(t))}{c!}, \tag{33}$$

$$E(U^{(M/G/\infty)}(t)) = \frac{(m(t))^{c+1}e^{-m(t)} + (c - m(t))\Gamma(c+1, m(t))}{c!}. \tag{34}$$

where $m(t)$ is defined by formula (7). That is, the MOL approach is exact for the infinite server system. Note that Remark 3 also holds in the case of time dependence.

Thanks to the introduction of κ, the results for the call packing system can straightforwardly be extended to time dependence through the MOL approximation. Since both the mean number of unused beds (formula (20)) as well as the mean number of people in overflow (formula (18)) do not depend directly on the arrival rate, but only through the offered load, a time-dependent approximation for these values can be obtained by substituting $m(t)$ for ρ. Although formulas (20) and (18) only hold for the case of the exponential call packing, we will determine $m(t)$ via formula (7) for the desired (non-exponential) distribution.

The time dependent case for the finite server system will not be considered in research, since no exact method is available and some preliminary numerical analyses showed limited added value of the finite server system.

3.2 Optimization model

In the previous subsection, expressions for the mean number of patients in overflow $E(L_\mathrm{q})$ and the mean number of unused beds $E(U)$ were derived for the different systems discussed. These performance measures, together with the time dependent offered load $m(t)$, will serve as the basis for determining the optimal capacity in this section. First, the objective function is introduced for the stationary case and its cost parameters are explained. The minimization of this stationary case is straightforward. The objective function is then extended to the time dependent case. To be able to incorporate constraints on the capacity function and facilitate interaction between the capacity of different patient types, the problem is formulated as a Mixed Integer Program (MIP). Then, several possibilities for constraints on the time dependent behaviour of the capacity are introduced, including their MIP formulation. Moreover, the MIP is extended to take into account multiple patient types enabling interesting concepts as shared capacity between patient types and the relabeling of beds from one patient type to the other.

It is important to realize that the values of the performance measures are obtained via an infinite server analysis, and thus that the underlying distribution of the number of patients in the system is independent of the capacity. The benefit of this is that the values of the performance measures at time t for capacity c are independent of the selected capacity before that time. In this way, there is a strict distinction between the results from the queueing analysis, which serves as input for the optimization model, and the optimization process itself. The downside is that neglecting the influence of having limited capacity on the performance measures means that inaccuracies can occur when switching capacities, as will be elaborated on in the Discussion. Nevertheless, the literature on the Modified Offered Load approximation indicates that performance measures based on the offered load of an infinite server system should still give a reasonable approximation for the case of a finite number of servers.

Before we solve the staffing problem, we should make clear that in this research we assume that the hospital and the VVT act as one decision-maker that wants to minimize their total costs. In this way, the total cost that gets spent on the capacity at the VVT and their overflow beds is minimized, and thus both parties could profit from this new situation. However, the costs and profit for implementing more capacity are not automatically distributed well over the two parties; the VVT will likely face an increase in costs due to facilitating more capacity while the hospitals benefit by having less costs for overflow. It is important that the two parties are aware of this and that agreements are made about a fair division of the costs and profits. To illustrate how substantial the individual loss and the shared profit could be, the optimal situation for the VVT on its own is also studied in Sections 5.1.3 and 5.4.

3.2.1 Objective function

When determining the optimal capacity, a lot of different factors come into play, but they can mostly be divided in costs due to having too few beds or too many beds, in line with Zychlinski et al. (2020). All the costs for having overflow (too few beds) and all the costs for having unused beds (too many beds) are gathered into two cost parameters per patient type r, $C_{\mathrm{o},r}$ and $C_{\mathrm{u},r}$, respectively. *Costs* include here actual monetary costs as well as things like inconvenience or deterioration in health, to which a certain cost reflecting its importance should be assigned. In line with the view of Zychlinski et al. (2020), $C_{\mathrm{o},r}$ and $C_{\mathrm{u},r}$ are the costs which could have been avoided if there would have been one bed more or less, respectively, in the VVT. With this view, it is clear that for the overflow cost $C_{\mathrm{o},r}$ the cost of having a VVT bed should be deducted from the cost of having a hospital bed, since this is the amount that could have been saved by having an extra VVT bed. Moreover, $C_{\mathrm{o},r}$ may contain costs to represent the missing out on the appropriate care, which may lead to a deterioration (or less improvement) in health. This cost factor can vary a lot depending on the patient type. For some patient types, like WLZ, it will not matter too much if they lie in the hospital or in the VVT, while other patient types, like GRZ, will miss out on specialized care. Other costs that can be included in $C_{\mathrm{o},r}$ are opportunity costs, which account for the loss of profit for the hospital caused by the fact that having an overflow patient in a hospital bed prevents a new patient being served, which would yield more money. In case no good estimate of the values of these extra cost components of $C_{\mathrm{o},r}$ can be made, it is good that the reader is aware that exclusion or underestimation of these cost components will lead to the found capacity being a lower bound.

In addition to these common cost parameters, $C_{\mathrm{o},r}$ and $C_{\mathrm{u},r}$, we choose to explicitly include the profit that the VVT makes by treating patients. As explained in Section 2.4, most VVT types receive money per patient per day. This will be modelled by a fixed profit (negative cost) per used bed for each patient type, $C_{\mathrm{b},r} < 0$. By including this profit on used beds explicitly, we want to make the trade-off of the VVT as an individual decision-maker apparent. The VVT on its own can neglect the costs of overflow in the hospital, but will not have an optimal capacity of zero to minimize the costs of unused beds due to the profit they can make on used beds. The individual decision of the VVT thus follows from setting the loss per overflow bed, $C_{\mathrm{o},r}$, to zero.

Remark 9. *Since the VVT would miss out on the profit for a used beds when there is one bed 'too few' in the VVT, i.e. one person in overflow, the profit on used beds could have been included implicitly in the opportunity costs of overflow, and thus in $C_{o,r}$. In that case, the VVT on its own does notice 'cost' of overflow by missing out on possible profit. When studying the optimal decision for the VVT on its own, the correct value for C_o would not be zero but exactly this lost profit. $C_{b,r}$ Since the profit for treating patients in the VVT is explicitly captured in $C_{b,r}$ in this research, it should not be included in (the opportunity cost component of) C_o.*

In this research, we assumed that the beds for different types of VVT care are distinct. In combination with the assumption of infinite waiting room, the different types of patients will not note the existence of the other types. For determining the optimal capacity, this means that each patient type can be optimized separately. The optimal solution is then simply the optimal capacity found for each patient type concatenated with as cost the sum of the optima for each patient type. Since the analysis for each patient type can be done separately, the objective function and mathematical program will be explained for a single patient type, thus dropping the subscript r referring to the patient type and the summation over the patient types. Only in section (3.2.2), the assumption of distinct capacity for each patient type will be dropped so that interesting options as shared capacity and relabeling of bed types can studied by implementation in the Mixed Integer Program.

At the start of Section 3.1, we already briefly mentioned the possibility of arrivals that do not come from the hospital, so-called external arrivals. With the application to the VVT type GRZ and the limited availability of data in mind, the external arrivals are assumed to follow the same arrival distribution and service distribution as the arrivals originating from the hospital, to which we will refer as internal patients in this section. In case all VVT beds are full upon an external arrival, the external patient will wait at home until it is their turn. There is just one (virtual) queue for the external and internal arrivals, with no priority for one type or the other. We will elaborate on the assumptions and their validity in the Discussion. Let ξ be the fraction of patients that come from the hospital. Since internal and external patients behave identically and are served in order of arrival, the performance measures can be determined from the system without taking into account the former location of a patient (hospital or external). The 'type' of the patient comes into play when evaluating the costs of a situation since only internal arrivals occupy an overflow bed in the hospital in case of overflow. Since external and internal patients have identical system characteristics, a fraction ξ of the number of patients in overflow is internal. So, we only charge the cost for overflow C_o for a fraction ξ of the expected number of patients in overflow. Although external patients do not occupy a hospital bed while waiting for VVT care, one might still want to charge a (lower) cost for each external patient waiting, for example, to reflect possible deterioration in health. Similar to the cost for overflow, the cost for external waiting should be charged to a fraction $1 - \xi$ of the expected number of patients in overflow.

The cost parameters C_o, C_u and $C_{\mathrm{b},r}$ are the cost per bed per day, so they should be multiplied by the number of patients in overflow, the number of unused beds and the number of used beds, respectively, weighted by the probability that this occurs (in steady state) and, in case of time dependence, these expressions should be summed over time. We assume that there is a fixed cost for each overflow (or unused) bed, that is, the cost of having ten overflow beds is exactly ten times the cost of having one overflow bed. This enables us to take C_o out of the summation,

$$\sum_{n=c}^{\infty} C_\mathrm{o}(n-c)\pi(n) = C_\mathrm{o}\sum_{k=0}^{\infty} k\pi(c+k) = C_\mathrm{o}[E(L_\mathrm{q})](c). \tag{35}$$

Similarly, for the costs for unused beds, we obtain $C_\mathrm{u}[E(U)](c)$, and for the costs of used beds $C_\mathrm{b}(c-[E(U)](c))$. The necessary expressions for $[E(L_\mathrm{q})](c)$ and $[E(U)](c)$ were derived in the previous section, where its dependence on the capacity c is now explicitly shown.

Remark 10. *In Remark 9, we already noted that the profit for used beds could have been implicitly included in C_o. We will now argue that the choice to explicitly include the profit on used beds, C_b, does not change the optimal capacity but that it does change the objective value. Since the offered load ρ as inputted to the MIP is based on the infinite server approach, its value is fixed for one parameter set, i.e. independent of the choice of capacity. From $C_o[E(L_q)](c) = C_o([E(U)](c) - c) + C_o\rho$, we then see that the cost (lost profit) on used beds is linear with gradient $([E(U)](c) - c)$ if the profit on used beds would be included in the cost of overflow. Since the profit for used beds is negative ($C_b \leq 0$), the 'cost' for used beds is also linear with gradient $([E(U)](c) - c)$. The difference in objective value between these two modelling scenarios comes from the nonzero term $C_o\rho$, causing the objective value of the implicit modelling to be larger than that of the explicit modelling. This matches the intuition that with the implicit modelling an opportunity cost (missed profit) is charged only in case of an overflow bed, while with the explicit modelling profit for every used bed is earned.*

The sensitivity of the objective value to the implicit or explicit modelling of profit on used beds, as explained in Remark 10, illustrates that caution is necessary when dealing with the objective value. One should not attach too much meaning to the exact objective value for a single capacity, but the objective values corresponding to several capacities can be compared to get an impression of their relation.

For the stationary analysis, the performance measures have a fixed value for each capacity c. Since there is no difference between days, minimization over one day will yield the same capacity result as minimization over a time period. The objective is thus simply minimizing the total cost per day over the possible capacities c,

$$TC(c) = \xi C_{\text{o}}[E(L_{\text{q}})](c) + C_{\text{u}}[E(U)](c) + C_{\text{b}}(c - [E(U)](c)). \tag{36}$$

Time dependence

When the arrival rate is inhomogeneous, the offered load and thus the mean number of overflow patients and unused beds differ per day. The total cost over the complete time period T, instead of a single day, should therefore be considered. We can simply sum (36) over all days and take the average, where the mean number of overflow patients and unused beds are now time dependent,

$$TC(\mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\xi C_{\text{o}}[E(L_{\text{q}}(t))](c(t)) + C_{\text{u}}[E(U(t))](c(t)) + C_{\text{b}}(c(t) - [E(U(t))](c(t)))\Big]. \tag{37}$$

By using the time dependent formulas (33) and (34) for $[E(L_{\text{q}}(t))](c(t))$ and $[E(U(t))](c(t))$, respectively, the system is approximated by an infinite server system. Since we showed in the previous subsection that the infinite server system yields a lower bound on the mean number of patients in overflow, the capacity results obtained in this way can serve as an indication of the minimal capacity. Alternatively, the MOL approximation for the time dependent callpacking can be used by using formulas (18) and (20) with ρ replaced by $m(t)$.

Since the offered load per day is not constant anymore, it is likely beneficial to be able to change the capacity to prevent peaks in overflow or unused beds when there are predictable peaks or low points in the offered load but constant capacity. Therefore, we will consider time dependent capacity. As mentioned at the beginning of this subsection, the performance measures are obtained via an infinite server analysis making their values measures at time t for capacity c independent of the selected capacity before that time. Therefore, we can simply change c to $c(t)$ in the expressions (33) and (34) for the mean number of patients in overflow and unused beds, respectively.

Of course, changing the capacity daily is not desirable as it is practically inconvenient and likely costly. This behaviour of $c(t)$ should therefore be discouraged by charging a cost for changes in capacity, C_{c}. This leads to the following objective function

$$TC(\mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\xi C_{\text{o}}[E(L_{\text{q}}(t))](c(t)) + C_{\text{u}}[E(U(t))](c(t)) + C_{\text{b}}(c(t) - [E(U(t))](c(t))) + C_{\text{c}} * |c(t) - c(t-1)|\Big], \tag{38}$$

where $|.|$ denotes the absolute value such that a cost $C_{\text{c},r}$ is charged for each bed that is either added or removed. Instead of an absolute value, an indicator function could be considered such that a fixed cost is incurred for changing capacity independent of the size of this change. One could also consider charging different costs for the increase and decrease in number of beds. Moreover, constraints on the behaviour of capacity over time could be set, for example, capacity may only change at set capacity-change points, as suggested by Green et al. (2007), or the capacity level should stay fixed for at least x days after a change, and the capacity can not change by too much.

In line with the observation made in Remark (3), objective (37) for the $M/G/\infty$ can be rewritten as

$$TC(\mathbf{c}) = \frac{1}{T}\sum_{t=1}^{T}\Big[\xi C_{\mathrm{o}}m(t) - (\xi C_{\mathrm{o}} - C_{\mathrm{b}})c(t) + (\xi C_{\mathrm{o}} + C_{\mathrm{u}} - C_{\mathrm{b}})[E(U(t))](c(t))\Big], \tag{39}$$

to which the objective term for changing capacity can of course also be included.

3.2.2 Mixed Integer Program

The stationary analysis can simply be solved by calculating for each possible capacity c the value of the objective function (36), after which the minimum value can just be selected. This approach can be extended to the time dependent case by repeating this procedure at every timepoint t, where the time dependent objective is the summand of (37), and concatenating the capacity solutions. Recall that this independent analysis of each timepoint is possible thanks to the infinite server analysis for determining of the values the performance measures. However, the time dependent capacity vector obtained with this approach will likely just mimic the behaviour of the offered load $m(t)$, changing capacity regularly, since it lacks the ability to take the days around t into account. For example, if the time dependent offered load increases only for a single day after which it drops again, it might be better not to change capacity (by too much) for that single day, depending on the costs for changing capacity C_c. Therefore, it is better to determine the capacity for all days at once based on the objective function (38), which can be done with a Mixed Integer Program. Of course, one could just range over all capacity vectors and select the one for which the objective function (38) is minimal, and even speed up the search over the capacity vectors by implementing some neighbourhood search or decent method initialised by the optimal solution obtained from the independent analysis at each time point. However, a Mixed Integer Program (MIP) is in general faster and can easily be extended to study more specific scenarios, like when there are certain constraints on the capacity or when capacity can be shared between multiple patient types. Again, the reader should be aware that also the MIP approach relies on the fact that the performance measures $[E(L_{\mathrm{q}}(t))](c(t))$ and $[E(U(t))](c(t))$ are determined from an infinite server system and therefore independent of the choice of capacity $c(\bar{t})$ for $\bar{t} \neq t$. Dependency between the capacity decision per day is only present due to the introduction of the cost for changing capacity in the objective function and, later on, the constraints on the behaviour of the time dependent capacity.

So, a Mixed Integer Program will be set up to determine the optimal capacity. First, the framework of the MIP to solve the simplest case of unconstrained optimization of (38) is explained. Then, several improvements and extensions are treated which are divided into constraining the behaviour of the capacity over time and into the sharing of capacity between different patient types. For the latter, it is explained how the MIP can take into account multiple patient types and the assumption of independence between VVT types is dropped.

As mentioned before, expressions for $[E(L_{\mathrm{q}}(t))](c(t))$ and $[E(U(t))](c(t))$ are derived earlier in this report and their values are fixed given $c(t)$ and t. This means that their values can be calculated before the optimization process and stored in a look-up table. A reduction in both computational time in the preparation phase as well as storage in the optimization phase can be made by using objective function (39), based on the observation in Remark (3). In this way, the vector of $m(t)$ replaces the lookup table for $[E(L_{\mathrm{q}}(t))](c(t))$, making both the construction and storage of this second lookup table unnecessary. A matrix U will contain the values of $[E(U(t))](c(t))$ with rows corresponding to capacities and columns to time points, $U[c][t]$. In case $[E(L_{\mathrm{q}}(t))](c(t))$ and $[E(U(t))](c(t))$ are determined with the MOL approximation for the call packing system, the observation of Remark (3) does not hold and thus both a look up table for the mean number of unused beds, $U[c][t]$, as well as a look up table $L[c][t]$ for the mean number of patients in overflow are necessary. This influences only the objective function of the Mixed Integer Program derived below. By comparing objective functions (37) and (39) and implementing $[E(L_{\mathrm{q}}(t))](c(t))$ similar to $[E(U(t))](c(t))$, the necessary adaptation to the MIP should be straight forward.

Remark 11. *It might feel more intuitive to let time belong to rows and capacities to columns since for each time point (row) only one capacity should be selected. However, the construction of* $[E(U(t))](c(t))$ *happens* *per capacity over the complete time horizon due to time dependence via $m(t)$, leading to a format with rows corresponding to capacities.*

In order to optimize over an objective function with a Mixed Integer Program, decision variables need to be introduced. In this MIP, binary decision variables $x_{c,t}$ are used to indicate if capacity c is selected at time t. Naturally, only one capacity can be selected at each time point leading to the MIP constraint

$$\sum_c x_{c,t} = 1.$$

Due to this constraint, the decision variables uniquely define the capacity at time t via the relation

$$c(t) = \sum_c c x_{c,t}. \tag{40}$$

The basis MIP for one patient type therefore is

$$\begin{aligned} \min_x \quad & \frac{1}{T}\sum_t \Big[\xi C_\mathrm{o} m(t) - (\xi C_\mathrm{o} - C_\mathrm{b})c(t) + \sum_c (\xi C_\mathrm{o} + C_\mathrm{u} + C_\mathrm{b})U[c][t]x_{c,t}\Big] \\ \text{s.t.} \quad & \sum_c x_{c,t} = 1 \quad \forall t \in T \\ & c(t) = \sum_c c x_{c,t} \quad \forall t \in T \\ & x_{c,t} \in \{0,1\} \quad \forall t \in T, \forall c \in C \end{aligned}$$

where $T = \{0, 1, \cdots, t_\mathrm{end}\}$ is the set of time points within the time horizon and $C = \{c_\mathrm{min}, \cdots, c_\mathrm{max}\}$ is the set of all capacities considered at each day.

The next step is to include the cost of capacity change in the objective. It is important to realize that the absolute value of the capacity change in (38) is not linear and can thus not be put directly in the objective function. A standard technique, as described by for example Bertsimas and Tsitsiklis (1997), is to replace the expression in the absolute value $(c_r(t) - c_r(t-1))$ by an auxiliary variable $z(t)$ which should satisfy the constraints

$$\begin{aligned} c(t) - c(t-1) &\leq z(t) \quad \forall t \in T \setminus \{0\}, \\ -(c(t) - c(t-1)) &\leq z(t) \quad \forall t \in T \setminus \{0\}, \\ z(0) &= 0. \end{aligned} \tag{41}$$

The objective function then becomes

$$\frac{1}{T}\sum_t \Big[\xi C_\mathrm{o} m(t) - (\xi C_\mathrm{o} - C_\mathrm{b})c(t) + \sum_c [(\xi C_\mathrm{o} + C_\mathrm{u} + C_\mathrm{b})U[c][t]x_{c,t}] + C_\mathrm{c} z(t)\Big]. \tag{42}$$

In (38), the same cost for an increase and decrease in capacity is charged, namely C_c. One could argue that the cost of increasing capacity is higher than of decreasing capacity. To implement this in the model, auxiliary integer variables $i(t)$ and $d(t)$ are introduced to capture the increase and decrease, respectively, in the number of beds at time t. Both variables are thus non-negative and satisfy the constraints

$$\begin{aligned} c(t) - c(t-1) = i(t) - d(t) \quad & \forall t \in T \setminus \{0\}, \\ i(0) = 0, \quad d(0) = 0. & \end{aligned} \tag{43}$$

Interestingly, one could recognise this model implementation as the alternative approach for dealing with an absolute value as described by Bertsimas and Tsitsiklis (1997). The introduced auxiliary variables replace the auxiliary variable $z(t)$ and constraints (43) replace constraints (41). Furthermore, the term $C_\mathrm{c} z(t)$ in the objective function should be replaced by $C_\mathrm{i} i(t) + C_\mathrm{d} d(t)$. The non-negativity of the costs for increasing and decreasing capacity guarantees that either $i(t)$ or $d(t)$ is set to zero when satisfying constraints (43), which explains why constraints (43) suffice.

Constraints on behaviour $c(t)$

Being the big advantage of using a MIP for determining time dependent capacity, the behaviour of the capacity over time can be regulated in certain ways by introducing appropriate constraints. Here, three possible requirements are explained, of which only one of the first two should be selected, whichever fits best to the situation.

Capacity may only change at set time points. Having fixed staffing intervals is common in healthcare practices Feldman et al. (2008). The case in which capacity may only change at timepoints $(\tilde{t}_1, \tilde{t}_2, \cdots, \tilde{t}_n)$ is therefore considered. In other words, the capacity for the time between $\tilde{t}_{i-1}$ and $\tilde{t}_i$ is fixed. This is captured in the constraints

$$c(t) = c(\tilde{t}_{i-1}) \quad \forall \tilde{t}_{i-1} \leq t < \tilde{t}_i \quad \forall i \in [n],$$

where [.] denotes the set of first integers $\{1, 2, \cdots, n\}$ and $t_0 = 0$.

Capacity should stay fixed for a minimum time span. If instead the capacity is allowed to change at any time point, bounds on the frequency of change might be desirable: after a change in capacity, the capacity can not change for the next q time units. To model this constraint, an auxiliary variable $s(t)$ is introduced to indicate if there is a change in capacity at time t or not. It is sufficient to only model that $s(t)$ is binary and can not be zero when there has been a change in capacity at time t, which is achieved by lower bounding $s(t)$ by $\frac{|c(t)-c(t-1)|}{M}$. The denominator M is a number at least as large as the maximum capacity considered such that the fraction is always smaller than one. Since the absolute value is not linear, $s(t)$ is defined by the constraints

$$\begin{aligned} \frac{c(t)-c(t-1)}{M} &\leq s(t) \quad \forall t \in T \setminus \{0\}, \\ -\frac{c(t)-c(t-1)}{M} &\leq s(t) \quad \forall t \in T \setminus \{0\}, \\ s(0) &= 1, \\ s(t) &\in \{0,1\} \quad \forall t \in T \setminus \{0\}. \end{aligned}$$

Using this auxiliary variable $s(t)$, a new change in capacity within q time steps after t is then prevented with the constraint

$$\sum_{k=0}^{q} s(t+k) \leq 1 \quad \forall 0 \leq t \leq t_{\text{end}}.$$

Remark 12. *Only a lower bound on the binary variable $s(t)$ is necessary for $s(t)$ to behave as the desired indicator, since $s(t) = 1$ puts a restriction on the values of $s(t^*)$ for t^* in a radius q from t via above constraint, while $s(t) = 0$ does not. So, setting $s(t) = 1$ when no change in capacity takes place at t only limits the number of solutions and will therefore not happen at time points where it matters.*

Capacity can change only by a limited amount. Due to logistical reasons, it might not be possible to change the capacity by more than b beds per time. Since a bound on the possible change in capacity involves an absolute value, two simple constraints for each time step should be added to the MIP,

$$\begin{aligned} c(t) - c(t-1) &\leq b \quad \forall t \in T \setminus \{0\}, \\ -(c(t) - c(t-1)) &\leq b \quad \forall t \in T \setminus \{0\}. \end{aligned}$$

Different patient types

Until this point, we acted as if there is only one patient type. In the remainder of this section, we will consider multiple patient types and drop the assumptions of strictly distinct beds per type. The extension to multiple patient types is quite straightforward. All variables are made dependent on patient type r, for example, the decision variables become $x_{r,c,t}$, the look-up matrix becomes three dimensional, $U[r][c][t]$, and the implied capacities become $c_r(t)$. The objective function will be summed over all VVT types. Lastly, all constraints should hold $\forall r \in R$, where R is the set of all VVT types. Without additional constraints, the optimal solution of this model is simply the optimal capacity per VVT type found by optimizing for each VVT type separately. However, parallel optimization of multiple VVT types makes incorporation of interaction between capacities of different types possible. The simplest interaction is that the total capacity available for all VVT types is fixed at (or upper bounded by) N, which is implemented by introducing the constraints

$$\sum_r c_r(t) = N \quad \forall t \in T \quad (\sum_r c_r(t) \leq N \quad \forall t \in T).$$

Naturally, a bound on the available capacity for a subset $\tilde{R}$ of VVT types is set by only summing over $r \in \tilde{R}$.

Shared capacity. It can be the case that some beds can be used by multiple patient types $r \in \tilde{R}$. Besides this general or shared capacity, each type can have some extra specific beds for their type only. Suppose that the number of general beds is fixed at G. Let variables $g_r(t)$, $r \in \tilde{R}$, be the number of general beds that are used by type r at time t. Naturally,

$$\sum_{r \in \tilde{R}} g_r(t) \leq G \quad \forall t \in T.$$

The decision variables $x_{r,c,t}$ will still indicate the total number of beds c that are used by type r at time t. However, the total number of beds in use by type r consists now of the number of general beds occupied by type r, $g_r(t)$, and the number of (possibly additional) *type-specific* beds, for which the variable $c_r(t)$ are used. The dependence of the type-specific capacity $c(t)$, as initially given by (40), now thus changes to

$$c_r(t) = \sum_c c x_{r,c,t} - g_r(t),$$

where $g_r(t) = 0 \ \forall t \in T$ by default for all types that are not eligible for the general beds, that is, for $r \notin \tilde{R}$.

In the first term in objective function (42) the total capacity is of interest. This term $C_{\text{o},r}(m_r(t) - c_r(t))$ therefore changes to $C_{\text{o},r}(m_r(t) - (g_r(t) + c_r(t)))$. It is assumed that it does not matter how the general beds are distributed between patient types over time, and thus both the cost for changing capacity, given by $C_{\text{c},r} z_r(t)$ in the objective, as well as the constraints in Section 3.2.2 are still applicable to $c_r(t)$. If a cost for switching general beds between patient types is desired, the implementation will be similar to charging costs for a change in $c(t)$, but now for the variables $g_r(t)$. Furthermore, a cost for the total number of general beds in use could be charged by simply introducing a cost term depending on $\sum_r g_r(t)$ in the objective function.

Lower costs for relabeling existing beds than for new beds. It might be possible that VVT beds currently for a certain patient type $\bar{r}$ could be made suitable for patient type r^* by some easy actions. In such case, this *relabeling* of existing beds could be cheaper than the alternative of realizing extra (new) beds for patient type r^*. Let the integer variable $l_r(t)$ be the number of beds that are relabeled from patient type r to another patient type at time t. A negative value of $l_r(t)$ means that there are beds relabeled from other patient types to patient type r, which leads to more beds for patient type r. Clearly, there should be no net change in the number of beds due to relabeling, that is,

$$\sum_r l_r(t) = 0 \quad \forall t \in T.$$

Besides relabeling, the number of beds assigned to a certain type can still change due to an increase (or decrease) in total capacity, that is, by introducing (or removing) a completely new bed. Earlier the auxiliary variable $i_r(t)$ ($d_r(t)$) represented the increase (decrease) in number of beds of type r at time t. When relabeling is possible, $i_r(t)$ ($d_r(t)$) will represent the increase (decrease) in number of beds of type r *due to introduction (removal) of a new bed*, and will thus exclude a change in capacity due to relabeling. The following relation between the mentioned variables is obtained,

$$c_r(t) - c_r(t-1) = -l_r(t) + i_r(t) - d_r(t) \quad \forall r \in R \quad \forall t \in T \setminus \{0\},$$

and will replace constraints (43).

For relabeling capacity of type r, a cost C_{l} should be charged per relabeled bed instead of the regular cost C_{i} (C_{d}) for increase (decrease) of capacity. Since $i_r(t)$ ($d_r(t)$) now only represents the increase (decrease) in number of beds of type r due to introduction (removal) of a new bed, the objective term $C_{\mathrm{i}} i_r(t) + C_{\mathrm{d}} d_r(t)$ summed over all r and t is unchanged. An extra term to capture the cost of relabeling should be introduced. Since the relabeling of one bed from patient type $\bar{r}$ to r^*, causes the value of two auxiliary variables to change, $l_{\bar{r}}(t) = 1$ and $l_{r^*}(t) = -1$, while the relabeling cost should be charged only once, the costs for relabeling at time t are $\sum_r C_{\mathrm{l}} \max(0, l_r(t))$. To preserve linearity of the objective function, the term $\sum_t \sum_r C_{\mathrm{l}} \bar{l}_r(t)$ is added to the objective function where the auxiliary variables $\bar{l}_r(t)$ satisfy

$$\begin{aligned} \bar{l}_r(t) \geq l_r(t) \quad & \forall r \in R \quad \forall t \in T, \\ \bar{l}_r(t) \geq 0 \quad & \forall r \in R, \quad \forall t \in T. \end{aligned}$$

If only the beds within a limited set $\bar{R}$ of patient type are suitable for relabeling, the variables $l_r(t)$ for $r \notin \bar{R}$ are simply forced to equal zero $\forall t \in T$.

4 Data for numerical analysis

This section covers the estimation of the parameters needed for the numerical capacity analysis. First, we will comment on the available data. Then, the estimation of the different parameters is discussed per parameter. The time step in this research is days.

4.1 Available data

Although this research aims to study the optimal capacity per patient type for a collection of VVT organisations, the numerical analysis will only be performed for one specific patient type of one VVT organisation due to the lack of appropriate data available. Two distinct data files were needed to determine the parameter estimations: POINT data about the transfer of patients from hospitals to VVT organisations and data from the specific VVT organisation. Unfortunately, it was not possible to link information from these data sets in order to have both information on the arrival and overflow process as well as on the service process of an individual patient.

The data files. The first data file is the POINT data. POINT is a system in which the transfer of patients from hospitals to VVT organisation can be organised. It contains per patient a lot of information on the transfer. For example, from which hospital department to which VVT location the transfer took place, when the transfer request was created, the date the hospital initially thought the patient would be ready to transfer and the date on which the transfer was realised. The oldest transfer in this data file was created on 17 March 2021 and the most recent transfer was created on 30 December 2022. The file is filtered on the receiving department name to match the organisation and patient type analysed in this numerical research. This filtered file contains 384 patients. The second data file contains data about the length of stay in the VVT for a specific patient type, GRZ (*geriatrische revalidatiezorg*, geriatric rehabilitation), at a specific VVT organisation. The oldest starting date of care is 1 March 2021 and the most recent is 14 December 2023. The file contains 562 patients. We received both data files on 14 December 2023.

The POINT data file contains a lot of date entries on the (request for) transfer. In discussion with the main hospital involved in this research, the best estimate for the number of overflow days (bed-blocking days in the hospital) should be obtained by subtracting the *Initial discharge date* from the *Realized discharge date*. We will comment on possible shortcomings of the use of the *Initial discharge date* in the Discussion.

4.2 Parameter estimation

4.2.1 Arrival process

In line with the advice to use the *Initial discharge date* as the starting moment of the overflow period, this date is also the arrival moment of the patient to the VVT. Recall that the arrival to the VVT system is namely defined as the moment the patient is ready to transfer to the VVT.

Taking the average over the number of arrivals per day yields an average arrival rate of 0.58 arrivals per day. However, there is a significant difference between the number of arrivals per day, especially on weekdays versus weekends. Figure 2 shows the total number of all arrivals per day of the week over the period of the data set (18 March 2021 to 30 December 2022).

Since the period over which the sum of the arrivals is taken is 93 weeks (and 2 days), dividing the total number of arrivals per day by 93 (or 94) yields an estimate for the average arrival rate per day of the week as given in Table 1.

TABLE 1: The average arrival rate from the hospital per day of the week.

Weekday	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Arrival rate	0.74	0.80	0.77	0.65	0.73	0.26	0.13

Important to note is that the arrival rates obtained from the POINT data are only the arrivals from the hospital. Since we assumed that only a fraction ξ of the arrivals came from the hospital, the total arrival rate to the VVT is higher. Including external arrivals, who were assumed to have the same arrival process, the total arrival rate is $\frac{\lambda}{\xi}$.

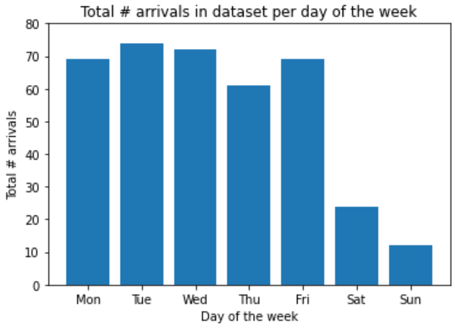


FIGURE 2: The total number of arrivals per day of the week over a period of 653 days.

Fraction ξ. The fraction ξ of patients that came from the hospital could not be estimated based on the available data. The data from the POINT system only contains requests for VVT care made by hospitals; we do not have data on VVT requests by external arrivals. We compared the number of patients that have the correct GRZ institution as the receiving organisation in the VVT file with the number of patients that arrived to that GRZ location in the same period to obtain a rough estimate for the fraction of internal arrivals. However, the number of patients in the POINT file was higher than the number of arrivals according to the data of the VVT institution. This inconsistency could be caused by patients ending up in a (temporary) different bed than planned. Regardless of the cause, a good estimate could therefore not be obtained from the data. The fact that the observed inconsistency could occur does indicate that the fraction of internal patients is likely to be rather high. It is also known from practice that most GRZ patients come from the hospital since recovery from surgery is the main reason for entering GRZ. We will therefore set the fraction of internal arrivals to $\xi = 0.95$.

Including external arrivals by setting $\xi = 0.95$, a total average arrival rate of $\frac{0.58}{0.95} = 0.61$ per day is obtained. The corrected total arrival rates per day can be found in Table 2.

TABLE 2: The average total arrival rate per day of the week for $\xi = 0.95$.

Weekday	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Arrival rate	0.78	0.84	0.81	0.68	0.77	0.27	0.14

Arrival distribution. In Section 3.1, the arrival process is assumed to be a Poisson process. This assumption is checked here. In a Poisson process, the number of arrivals per day is Poisson distributed. We count how many arrivals occur per day on each day between the first *Initial discharge date* and the last *Initial discharge date* in the data set. Then, we count the number of days that have 0 arrivals, 1 arrivals etcetera. This yields an observed frequency per number of arrivals on a day given in Table 3. Since the mean number of arrivals per day is 0.58, we hypothesize that the data is generated from a Poisson distribution with mean 0.58. The expected frequencies that belong to this distribution are given in Table 3 as well.

TABLE 3: The frequency table of the observed and the expected number of days on which a specific number of arrivals occurred. The expected frequency is based on a Poisson distribution with mean 0.58.

Number of arrivals on a day	0	1	2	3	4	5	6	≥ 7
Observed frequency	380	193	52	21	4	1	0	0
Expected frequency if $Poisson(0.58)$	362.58	212.20	62.10	12.11	1.77	0.21	0.02	≈ 0

Figure 3 visualises the comparison of the observed and expected frequency. The observed frequency of both 0 arrivals per day and 3 to 5 arrivals per day is higher than what would be expected from a Poisson distribution with mean 0.58, while the frequency of 1 and 2 arrivals per day is lower. A Chi-square test is used to determine the goodness of fit. The test statistic χ^2 has a value of 16.58. With $5 - 1 = 4$ degrees of freedom, we have $c = 9.49$ taken from the χ^2_4 table with significance level 0.05. Since $\chi^2 = 16.58 > 9.49 = c$, the null hypothesis is rejected at 5% significance level. So, we have to conclude from this goodness of fit test that it is likely that the number of arrivals per day is not Poisson distributed with mean 0.58. Since Figure 2 already showed that the number of arrivals per day depends on the day, this is not surprising. The assumption that the arrival process is a Poisson process is necessary for the analysis, but it is good that the reader is aware of the extent to which this assumption fits the data.

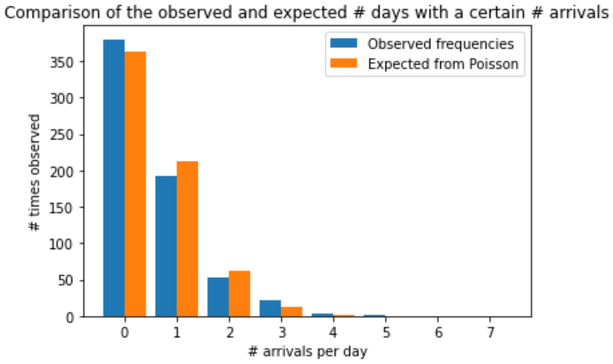


Figure 3: The observed and the expected frequency of each number of arrivals per day. The expected frequency is based on a Poisson distribution with mean 0.58.

In the time dependent analysis, a distinct arrival rate per day is used. The assumption of the number of arrivals on a specific day of the week being Poisson distributed with the corresponding arrival rate reported in Table 1, can be evaluated as well with the goodness of fit analysis described. For example, for the number of arrivals taking place on Tuesdays, we obtain a test statistic of $\chi^2_3 = 2.84$. Since $\chi^2_3 = 2.84 < 7.82 = c$, the null hypothesis that the number of arrivals on Tuesdays is Poisson distributed with mean 0.80 can not be rejected.

4.2.2 Service process

To determine the distribution of the service process, the length of stay (LOS) of the patients in the GRZ data file is determined. It is important to realize that we deal with so-called censored data here. Some patients were still receiving care at the moment the data files were generated and therefore their LOS is unknown. The observations of their LOS are right-censored, since only a lower bound on their LOS is known. The right-censored LOS of these patients is the number of days between their starting date and 14 December 2023, which is the day we received the data files. The LOS of the other patients is simply the number of days between their starting date and their end-of-care date. A survival analysis is performed on the censored LOS data to obtain an empirical cumulative distribution function (cdf) of the service times for which Kaplan-Meier is a common method [60]. To find an appropriate description of the service time distribution, several distributions are fitted to this cdf. The Python class *scipy.stats.CensoredData* in combination with the *scipy.stats.fit* function is used to deal with the fact that the data is censored. The best fits for the exponential, lognormal and gamma distribution can be found in Figure 4.

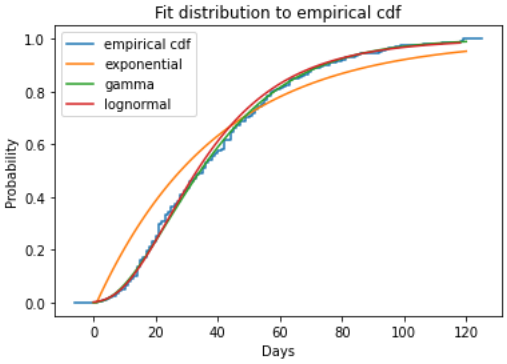


Figure 4: The fitted distributions to the empirical cumulative distribution function of the LOS.

The lognormal distribution shows a good fit to the empirical cdf, and is a common choice for modelling LOS in healthcare settings. The one-sample Kolmogorov-Smirnov test, which is a common goodness of fit test, gives a p-value of 0.070. Since this p-value is above the threshold of 0.05, the null hypothesis of the empirical cdf being lognormal with the fitted parameters can not be rejected with 95 percent confidence. The fitted lognormal has parameters (shape, location, scale) = $(0.50, -10.64, 44.51)$. A lognormal distribution with these parameters has mean 39.79 and variance 722.29. To these values corresponds a coefficient of variation of $\frac{\sigma}{\mu} = 0.68$. We note that the mean LOS of the fitted distribution is satisfactorily close to the average LOS of 39 days as reported in 2018 by ActiZ (2018).

4.2.3 Fraction κ

Since the information about the (time before) transfer and the time in the VVT are obtained from separate data files, between which there is no clear connection on individual patient level, we have no data to base an estimate for κ on. Moreover, since we can not see the influence of the time in overflow on the service time in VVT, we do not know how the service times that we use from the data are possibly already influenced by the time in overflow.

Remark 13. *An attempt was made to link the two data files to each other by a combination of the birth year of a patient and on one side the initial discharge date or date on which the transfer was closed in the POINT file and on the other side the starting date in the VVT data. Only a limited number of matches were found. Moreover, we can not be sure that found 'matches' indeed refer to the same person in reality. In general, (too) little information is known on the relation between the dates in the POINT data and those in the VVT data. Therefore, we can not match the two data files to obtain insight into the possible influence of time spent in the hospital and overflow on the length of stay in the VVT.*

4.2.4 Base set of parameter values

The parameters obtained from the data will be used to perform the numerical analysis in the next Section. Moreover, values for κ and the cost parameters will be selected that will be used to generate the results if not stated otherwise, like when ranging over another parameter. We will refer to this set of values of the parameters as the 'base set'.

The system characteristics in the 'base set' are a Poisson arrival process with either a stationary rate of 0.61 or time dependent rate given by Table 2, a fraction of internal arrivals $\eta = 0.95$ and a lognormal service distribution with a mean of 39.79 days and a variance of 722.29, which yields a coefficient of variation of 0.68. To this combination of arrival rate and mean service time corresponds an (infinite server) offered load of $\rho = \lambda E[S] = 24.27$. For the fraction κ a base value of 0.3 will be used.

Since the costs for a hospital and VVT bed are influenced by a lot of factors and are hard to find out, as illustrated by the list of different possible sources for valuing care units with several upsides and downsides as given in Section 3.3 in van Roijen et al. (2024), we have little information to base the cost parameters on. According to a recent news article, the hospitals in region Twente miss out on roughly 550 euros per day per bed due to patients in overflow [45]. With a cost of 150 euros per day in a VVT bed, as given in te Meerman (2017), we obtain $C_{\text{o}} = 550 - 150 = 400$. Interestingly, the ratio between $C_{\text{o}}, = 400$ and $C_{\text{u}} = 150$ is $\frac{400}{150} = 2.67$, while Zychlinski et al. reported a ratio of 2.67 between their two cost parameters for geriatric rehabilitation. Since the margins for care financed from the *Zorgverzekeringswet* are rather small, we use a profit of 10 euros on each used VVT bed. A close reader might have noticed that we did not try to include 'costs' for the unavailability of proper care that could result in the deterioration of the patients' health. As Section 3.2.1 points out, excluding such cost components will lead to the resulting capacity underestimating the true capacity. In the case of time dependent capacity, we additionally have cost parameters for the increase and decrease in capacity. These costs should be larger than zero to capture the fact that changing capacity both costs time as well as brings a certain inconvenience, but not too high since it should still allow for fluctuations in capacity so that there is an added value in analysing a time dependent capacity. Therefore, the cost of decreasing capacity is set at 5 euros per bed and the cost of increasing at 10 euros per bed. In summary, the cost parameters in the base set are set at $C_{\text{u}} = 150$, $C_{\text{o}} = 400$ and $C_{\text{b}} = -10$. For the time dependent analysis, we additionally have $C_{\text{d}} = 5$ and $C_{\text{i}} = 10$.

5 Numerical analysis

In this section, numerical results are given to illustrate the application of the mathematical models developed in this research. First, a stationary analysis is performed to study the difference in numerical outcomes between the different models of Section 3.1 and the influence of the fraction κ on this (Section 5.1.1). Moreover, the sensitivity of the optimal capacity to variations in the different cost parameters is visualised (Section 5.1.2). The stationary analysis is used to draw a preliminary conclusion on the difference in optimal capacity and associating cost between the situation in which the VVT is the individual decision maker and the situation in which the hospital and VVT have a common objective (Section 5.1.3). Second, the influence of a time dependent arrival rate is discussed. The time dependent analysis allows for fluctuations in capacity over time, so the influence of κ on the time dependent behaviour of the capacity is studied (Section 5.2.1). Since the time dependency of the capacity leads to the introduction of extra cost parameters, the influence of the change in these extra parameters on the optimal time dependent capacity is analysed (Section 5.2.2). The time dependent results are compared to the square root staffing rule (Section 5.2.3). It is shown how the constraints of the Mixed Integer Program affect the optimal solution (Section 5.2.4). Third, the performances of the analytic solutions in reality are mimicked through a simulation, both for the stationary capacity (Section 5.3.1) and the time dependent capacity (Section 5.3.2), and neighbouring solutions are analysed. Lastly, a conclusion is drawn on the difference between the optimal situation for the VVT and the system optimal.

Although the value of the objective function is used to determine which capacity is optimal, the use of the exact value of the objective function is limited. Its value is more sensitive to parameter changes than the corresponding optimal capacity. Since the values of some parameters are uncertain, this higher sensitivity is inconvenient. The objective value is also strongly influenced by certain modelling choices, like the explicit incorporation of the profit for used beds as argued in Remark 10. For these reasons, the influence of the varying parameters on the optimal capacity decision is studied and not their influence on the objective value.

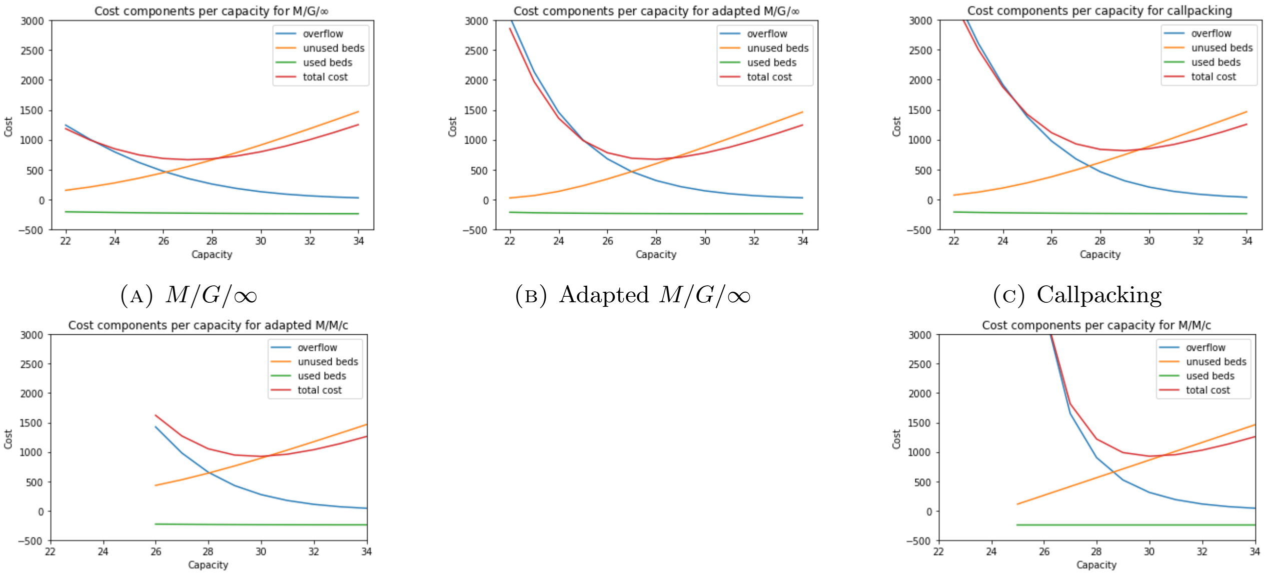
The numerical analysis in this research is limited to a specific VVT type GRZ at a specific institution. The data that is used as input for this numerical study is discussed in Section 4. If not stated otherwise, the results are generated with the parameter values from the base set as summarized in Section 4.2.4.

5.1 Stationary analysis

In the first analysis, the arrival rate is assumed to be stationary and therefore its value is 0.61 arrivals per day. We would like to point out that all the systems discussed in this research are either insensitive to the service time distribution in the stationary case or limited to the case of exponentially distributed service times. Therefore, only the mean service time of 39.79 days, corresponding to a service rate $\mu = \frac{1}{39.79}$, is taken into account from the service characteristics fitted in Section 4.2.2. Since the values of κ and the cost parameters are uncertain, sensitivity to these parameters is discussed later in this section. The influence of the arrival rate and service time are as expected and can be found in Appendix A.3.

The total cost comprises several cost components, as should be apparent from objective function (36). Figure 5 shows how these cost components are influenced by the capacity, and make up the total cost, for each system. The total cost of each system over the capacity is also plotted in Figure 6 so that the behaviour of the total costs with respect to the capacity can be compared between systems. The total cost curve of the adapted infinite server system behaves over the capacity similar to that of the call packing. Especially for lower capacities, the different behaviour of the unadapted infinite server system and the adapted infinite server system is noticeable. Figures 5a to 5c show that the overflow component is the main difference between the unadapted infinite server system and both the adapted infinite server system and the call packing system. A better approximation of the mean number of overflow in the call packing systems is thus the likely cause that the adapted infinite server system better mimics the call packing system than the unadapted version.

The computations for the $M/M/c$ are only given for a capacity of 25 and higher since the stability condition $\frac{\lambda}{c\mu} < 1$ is not met for lower capacities. For the adapted $M/M/c$, reasonable results could only be given for capacity 26 and higher since the repeated substitution to determine the fixed point mean number of patients in overflow did not converge to a single value for a capacity of 25. This inability to determine the appropriate cost for a capacity of 25 will also be apparent in some later results. The evaluation of the performance of the Adapted Service Time Approach for the $M/M/c$ in its current form compared to the call packing system is therefore limited. The available results do suggest that, just like for the infinite server systems, the adapted variant better mimics the total cost of the call packing system for decreasing capacities.



(A) $M/G/\infty$ (B) Adapted $M/G/\infty$ (C) Callpacking

(D) Adapted $M/M/c$ (E) $M/M/c$

FIGURE 5: The value of the cost components against the capacity, given for each system in this research.

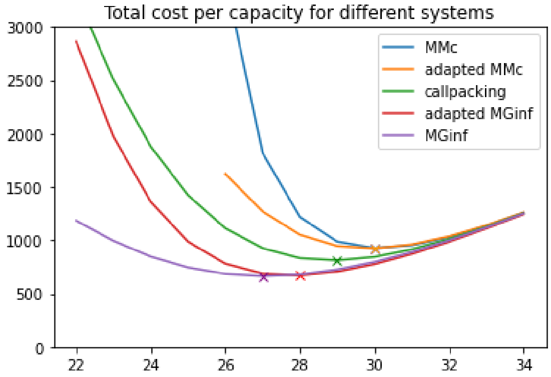


FIGURE 6: The behaviour of the total cost of each system over the capacity. The cross marks the minimum of the cost curve for each system.

Figure 6 shows that the total cost in each system becomes the same for increasing capacity. This is as expected since in the case of abundant capacity, the system is rarely saturated and thus the influence of the finiteness of the capacity is limited. In Sections 3.1.2 and 3.1.3, we either proved ($\leq^{\text{pr}}$) or conjectured ($\leq^{\text{conj}}$) the following ordering of the mean queue length in the distinct system: $E(L_{\text{q}})^{(\Sigma_A)} \leq^{\text{pr}} E(L_{\text{q}})^{(\Sigma_B)} \leq^{\text{conj}} E(L_{\text{q}})^{(\Sigma_C)} \leq^{\text{conj}} E(L_{\text{q}})^{(\Sigma_D)} \leq^{\text{pr}} E(L_{\text{q}})^{(\Sigma_E)}$, where systems Σ_A to Σ_E refer to the $M/G/\infty$, adapted $M/G/\infty$, call packing, adapted $M/M/c$ and $M/M/c$ system in that order. Figure 6 suggests that this ordering is almost preserved for the cost function. Remarkably, the total cost for the adapted infinite server system is slightly below that for the unadapted version for capacity 28 and higher. This does not undermine Theorem 5 as the mean number of patients in overflow in the unadapted $M/G/\infty$ system is still lower than that in the adapted system. The number of unused beds in the unadapted system is apparently sufficiently higher than in the adapted system for capacities above 28 such that the total cost in the unadapted system exceeds that in the adapted system.

In Figure 6 the minimum of the cost curve is marked with a cross for each system. The corresponding capacity (value on the x-axis) is the optimal capacity in that system. The minimum cost of each system satisfies the ordering of the systems as given above, and the optimal capacity for this ordered list of systems is non-decreasing.

5.1.1 Influence of κ

From Figure 6, the optimal capacity in each system could be determined for the base parameters, among which the fraction κ was set at 0.3. Since this choice for κ could not be based on the data, it is important to know how the value of κ influences the optimal capacity in each system, as is shown in Figure 7.

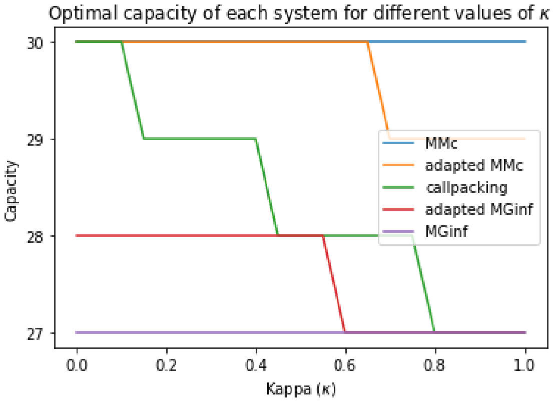


FIGURE 7: The optimal capacity of each system over different values of κ.

Since the unadapted systems do not depend on κ, the optimal capacity is a straight line over κ. The optimal capacity in the other systems decreases when κ increases. Since κ indicates how much the time in overflow reduces the residual service times, a higher value of κ means a shorter residual service time in the VVT and thus a decrease in offered load, which is generally known to lead to lower capacities. In line with the boundary cases, the optimal capacity for the call packing equals that for the $M/M/c$ and $M/G/\infty$ for $\kappa = 0$ and $\kappa = 1$, respectively. The optimal capacity satisfies the ordering of the systems given above for every value of κ. Thus, the optimal capacity in the adapted systems is at least as close to the optimal capacity of the call packing system as the optimal capacity in their respective unadapted versions. However, it does depend on the value of κ if the adapted finite server system (low values of κ) or adapted infinite server system (high values of κ) yields a capacity closer to that in the call packing system. We would like to stress that the optimal capacities of the call packing system, and finite server systems, given here hold for the case of exponentially distributed service times. Results for the call packing system with the fitted lognormally distributed service times are obtained through simulation and are given in Figure 20.

5.1.2 Influence of different cost parameters

Since the appropriate values for the cost parameters are not straightforward, it is important to study the sensitivity of the optimal capacity to the cost parameters.

Overflow cost C_o

First, the influence of the overflow cost C_o is studied. In Figure 8 we see that the optimal capacity increases as the cost for overflow increases which makes sense as the system aims to limit the amount of overflow. The influence of the overflow cost on the optimal capacity is apparent. Especially when the overflow cost drops below $C_\text{u} = 150$, the optimal capacity in most systems drops faster. For an overflow cost of zero, we note three things. Firstly, the capacity in the call packing system and infinite server systems drops, but not to zero. Since there is no cost for overflow while the cost for unused beds is significant, the main concern is minimizing the number of unused beds. However, the small profit on used beds prevents a drop to zero capacity, since guaranteed used beds will be beneficial. The $M/M/c$ system drops to the minimum capacity possible; a lower capacity yields an unstable system. The second thing we note for $C_\mathrm{o} = 0$ is that the adapted $M/M/c$ system stays at 26. This is due to the inability to calculate the appropriate cost of having capacity 25, as mentioned earlier. Lastly, the optimal capacity in the call packing system is lower than in the adapted $M/G/\infty$ system for $C_\mathrm{o} = 0$. This will be discussed later in relation to Figure 10b.

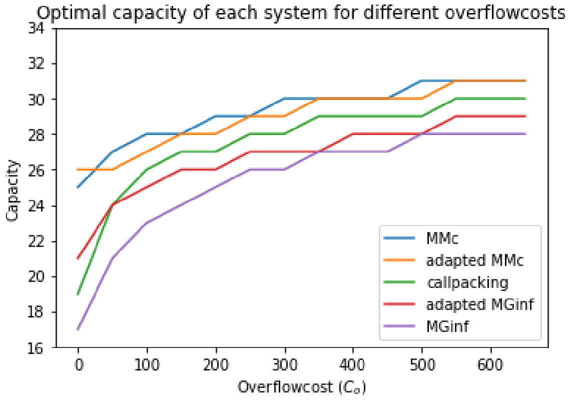
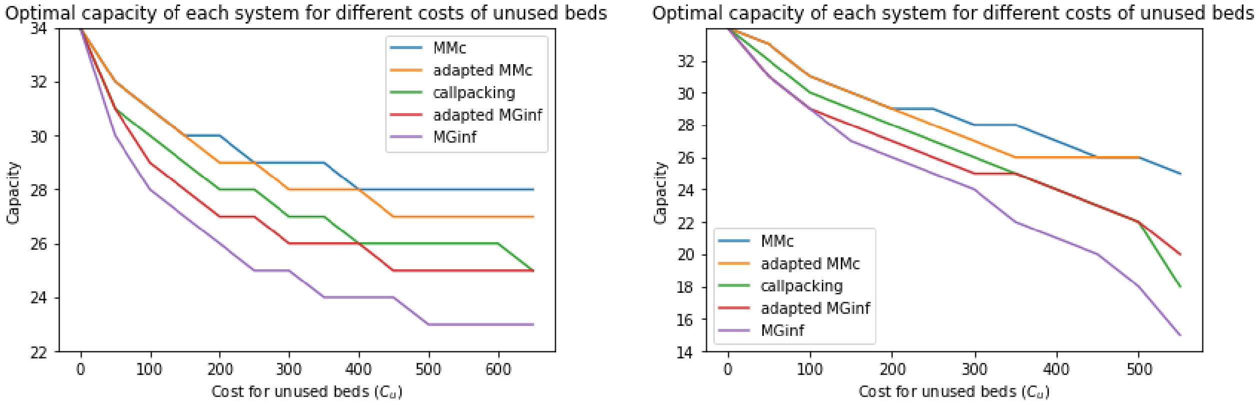


FIGURE 8: The optimal capacity over different values of the cost parameter for overflow, C_o.

Cost for unused beds C_u

We will now study the influence of the cost of unused beds C_u. Figure 9a shows how the optimal capacity behaves when C_u is ranged while C_o is kept fixed, where a capacity of 34 was the maximal capacity considered. The information contained in Figure 9a is similar to that in Figure 8, since the ratio between C_o and C_u determines mostly the optimal capacity. The optimal capacity for each system in Figure 8 for, for example, $C_\mathrm{o} = 200$ and $C_\mathrm{o} = 600$ is the same as in Figure 9a for, respectively, $C_\mathrm{u} = 300$ and $C_\mathrm{u} = 100$. Since C_u was fixed at 150 in Figure 8 and C_o was fixed at 400 in Figure 9a, these were the optimal capacities corresponding to a ratio of $C_\mathrm{o} : C_\mathrm{u}$ of 4:3 and 4:1, respectively. Although C_b was kept fixed at -10, and thus not scaled accordingly, it did not influence the capacity decision, as we will also see later. So, for the capacity decision only the ratio between the cost parameters is of importance. Please be aware that the objective value, on the contrary, is influenced by the exact values of the cost parameters.



(A) Overflow costs fixed at $C_\mathrm{o} = 400$.

(B) Costs for hospitalization fixed at C_o+$C_\mathrm{u} = 550$.

FIGURE 9: The optimal capacity over different values of the cost parameter for unused beds, C_u.

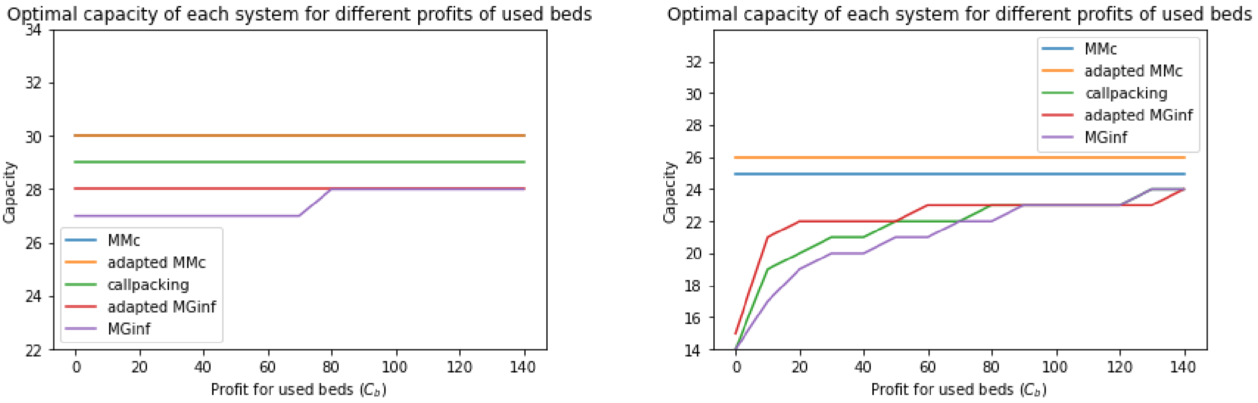
The reader should be aware that the analysis of variable C_u for fixed C_o, as in Figure 9a, can not be used to draw conclusions on the influence of the cost of unused beds on its own. Recall that the cost for overflow

C_o is the difference between the cost of hospitalization in the hospital and the cost of a VVT bed, C_u. That is, the cost of hospitalization is the sum of C_o and C_u. So, if we range C_u while keeping C_o fixed, not only the cost for unused beds, but also the cost for hospitalization in the hospital varies. To study the influence of different costs for unused beds compared to the cost of hospitalization better, the cost of hospitalization $C_o + C_u$ is kept fixed at 550 while ranging C_u in Figure 9b. Since $C_o + C_u$ is fixed at 550, C_u can only be ranged from 0 to 550. The influence of a change in C_u is now bigger than in Figure 9a. When $C_u = 550$, and thus $C_o = 0$, we again see that the optimal capacity in the call packing system is lower than that in the adapted infinite server system and that the capacity in the systems does not drop to zero likely due to the small profit on used beds.

Although the cost parameters C_o and C_u clearly influence the optimal capacity in all systems, they have limited influence on the accuracy of each system as an approximation for the call packing system, for the used $\kappa = 0.3$. The difference between the optimal capacity in the call packing system and the adapted systems is mostly one, and one to three between the call packing system and the unadapted systems, for most reasonable parameter combinations. Only when the cost for overflow is smaller than the cost for an unused bed, the difference in optimal capacity in each system becomes worse. This is especially the case when the cost for overflow is set at zero, since then the adapted systems perform badly.

Profit on used beds $-C_b$

We already noted that the ratio between C_u and C_o, even when C_b is fixed at -10, is (most) decisive for the optimal capacity. Figure 10 confirms the impression that the profit for used beds does not strongly influence the capacity decision. Although the almost complete absence of influence of C_b on the optimal capacity might be striking at first, it can be explained using the dependencies of the cost components in the objective function on the used beds on $[E(L_q)](c)$. If we study the value $[E(U)](c)$ for each c, we note that, roughly, for little capacity it is close to zero while for reasonable capacity it increases with approximately 0.9 per increase in the capacity, and for abundant capacity it increases with almost 1. That is, $(c - [E(U)](c))$ either increases by almost one per capacity for relatively low c and is almost constant for reasonable capacities. In case of relatively little capacity, the cost of overflow is so significant that the objective value for those capacities will be too far from optimal for the specific value of C_b to make a difference. For capacities close to the optimal capacity, the cost component for used beds, $C_b(c-[E(U)](c))$, decreases with slope close to zero, while the cost component for unused beds, $C_u[E(U)](c)$, increases with slope close to one. So, even for C_b close to C_u, the objective function is not likely to be significantly decreased by the profit on used beds to change the optimal capacity.



(A) Base parameters ($C_u = 150$ and $C_o = 400$).

(B) Overflow costs are set to zero ($C_o = 0$).

FIGURE 10: The optimal capacity over different profits for used beds, -C_b.

So far, we reasoned that the possible influence of the value of C_b on the objective value for low capacities is nullified by the high costs for overflow. The insensitivity of the optimal capacity to C_b as observed for the base parameters might therefore not hold for a significantly low value of C_o. The situation with zero costs for overflow is interesting when studying what happens if the VVT would decide on their own on their optimal capacity. The influence of the profit per used VVT bed on the optimal capacity when $C_o = 0$ is therefore studied in Figure 10b. We note that both the $M/M/c$ system and the adapted $M/M/c$ system stay at the minimum capacity for which it was reasonable to calculate the objective function. For the other systems, the optimal capacity is now more sensitive to C_b, as expected. Remarkably, the optimal capacity in the adapted $M/G/\infty$ system is larger than that in the call packing system for all values of C_b. We already observed this behaviour at the other moments the overflow cost C_o equalled zero, in Figures (8) and Figures (9b). Since the objective function, and thus the decision on the optimal capacity, depends only on the performance measure number of unused beds when $C_o = 0$, the result is caused by the calculated mean number of unused beds in the $M/G/\infty$ system being smaller than that in the call packing system for a given capacity. In Figure 6, we observed that the line for the adapted $M/G/\infty$ lay below the unadapted version for which we concluded that the mean number of unused beds in the normal $M/G/\infty$ system was sufficiently higher than that in the adapted system. In other words, the mean number of unused beds in the adapted system was rather low compared to that in the normal $M/G/\infty$ system. So, based on the numerical analysis, we conclude that the mean number of unused beds in the adapted $M/G/\infty$ system is rather low compared to both the normal $M/G/\infty$ system as well as the call packing system, but so far no reasonable explanation could be found.

5.1.3 Difference in optimal for VVT and optimal for system

We will use the stationary results for the call packing system to illustrate the difference in optimal capacity from a system perspective, in which the hospital and the VVT have a shared objective, and the optimal capacity if the VVT would decide on their own. Although the objective values should not be taken as absolute values, as pointed out earlier in this section, significant differences in their value for similar parameter combinations can indicate a reasonable difference in costs in reality. The main purpose of giving the objective value is thus to illustrate the influence of monetary incentives in the two scenarios.

For the set of base parameters, the optimal capacity from a system perspective is according to the call packing system 29 with an objective value of 812.32, as could also be concluded from Figure 6. The scenario in which the VVT decides on its own is best modelled by setting $C_o = 0$. The costs for overflow are, after all, on the hospital's account and not that of the VVT. With $C_o = 0$, the optimal capacity is 19, see Figure 8, with a corresponding objective value of -183.48 (a profit). The optimal capacity if the VVT decides is lower than if the hospital and the VVT decide together, as expected. From a financial perspective for the VVT, limited capacity will only limit the possible profit that could be made but will not bring losses as the VVT does not bear the overflow costs. More capacity, on the other hand, increases the risk of having a lot of unused beds that does result in losses for the VVT. The VVT on its own is therefore expected to be more conservative in their capacity than the system as a whole. The objective values provided so far can not be compared fairly, since one is the cost for the complete system (812.32 for a capacity of 29) while the other is the cost for the VVT (-183.48 for a capacity of 19). For a proper comparison of the costs of each scenario, both the costs for the system as well as the costs for the VVT are needed for each scenario. The cost for the VVT for a capacity of 29 is the objective value for $c = 29$ when $C_o = 0$: 505.28. The cost for the system for a capacity of 19 is the objective value for $c = 19$: 6445.79. The optimal capacity decision and costs per scenario are summarized in Table 4. The calculated mean number of patients in overflow per scenario is included as well.

Table 4: The optimal capacity and associated cost from both the system perspective and the VVT on its own. The situation in which only the VVT is considered is modelled by setting $C_o = 0$.

	Hospital & VVT decide	VVT decides
Optimal capacity	29	19
Cost for system	812.32	6445.79
Cost for VVT	505.28	-183.48
Mean number of overflow	0.81	17.71

So, the VVT on its own can minimize their expected costs to -183 euros by having a capacity of 19. This decision is expected to cost the hospital and VVT together 6446 euros. From a system perspective, this total cost could be reduced to 812 by having a VVT capacity of 29. However, that would increase the costs for the VVT to 505 euros, which is 688 euros more than if the VVT would decide individually on their capacity. This numerical analysis illustrates that it would be reasonable (and necessary) to compensate the VVT for increasing their capacity since it is expected to pay off for the system as a whole, both financially as well as in the availability of proper care.

The analysis performed here for the base parameters is expected to extend to other parameter combinations as well, since the drop in capacity when $C_\text{o} = 0$ in Figures 8 and 9b is significant compared to all other values of C_o and C_u, respectively. Furthermore, Figure 10 shows the influence of C_b is limited as the optimal capacity when $C_\text{o} = 0$ is reasonably smaller than when $C_\text{o} = 400$ even for a significant increase in the profit for a used bed. Moreover, the optimal capacity as a function of κ in Figure 7 ranges between 30 and 27 while the optimal capacity when $C_\text{o} = 0$ ranges between only 23 and 17. Indeed, the results for $\kappa = 0.8$ as given in Table 5 are similar to that for $\kappa = 0.3$.

Table 5: The optimal capacity and associated cost from both the system perspective and the VVT on its own when $\kappa = 0.8$.

	Hospital & VVT decide	VVT decides
Optimal capacity	27	17
Cost for system	702.09	3232.00
Cost for VVT	301.45	-157.69
Mean number of overflow	1.05	9.18

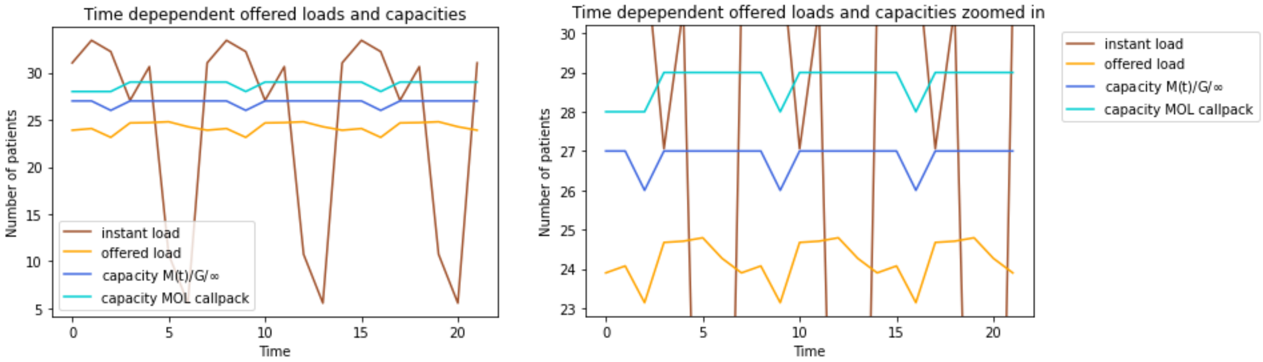
5.2 Time dependent analysis

In the numerical analysis so far, the arrival process was assumed to be stationary. Since Figure 2 shows that the number of arrivals per day is not steady over the week, an analysis with a time dependent arrival rate is performed as well. Since the time dependent offered load $m(t)$ is not insensitive to the service distribution, the time dependent analysis takes into account the fitted lognormal distribution. The time dependent analysis is performed for the time dependent infinite server system and with the Modified Offered Load approximation for the call packing system, as discussed in Section 3.1.4. Please note that for the time dependent analysis, a cost for increasing capacity ($C_\text{i} = 10$) and a cost for decreasing capacity ($C_\text{d} = 5$) are included in the base parameters, as explained in Section 4.2.4.

Since the arrival rate, and thus the time dependent offered load, has a period of 7 days and we perform an analysis using the steady-state expected values of performance measures, we expect that a time horizon of 21 days suffices to find the optimal time dependent capacity. The effect of freely choosing a capacity towards the boundaries of the timeframe is observed to stay limited to only the first few days and the last few days, see for example the green lines in Figures 13c and 13b respectively. The middle week of the three weeks, therefore, yields the optimal time dependent capacity for the parameter combination under consideration.

Stationary capacity. Firstly, if an arrival rate of 0.61 per day is used, the offered load $m(t) = 24.27\ \forall t$ equals the stationary offered load and thus the results match those in the stationary analysis. Before allowing the capacity to change over time, it is good to know how the time dependency of the arrival rate influences the optimal stationary capacity. We observe that the optimal (stationary) capacity found in the case of stationary arrival rates is also the optimal stationary capacity in the case of a time dependent arrival rate. The graph with the cost of each stationary capacity in case of a time dependent arrival rate looks the same as Figure 6 for both the infinite server system and (the MOL approximation of) the call packing system. The minimum objective value for the time dependent infinite server system is 668.85 compared to the stationary 663.61 and for the time dependent MOL approximation of the call packing system 821.02 compared to the stationary 812.32. Furthermore, the graph with the optimal stationary capacity in the MOL approximation of the call packing system per value of κ is the same as the stationary call packing system in Figure 7. So, when looking for a stationary capacity, there is little added value of a time dependent analysis compared to the stationary analysis for the base set of parameters. This conclusion is in line with the rule of thumb stated by Bekker and de Bruin (2010) that the impact of fluctuations in the arrival rate is relatively small if the average service time exceeds five times the cycle length at which the fluctuations occur. The mean service time is 39.79 days while the cycle length of the arrival rate is 7 days and thus we have $39.79 > 5 * 7 = 35$.

Shift in offered load The analysis with time dependent arrival rates allows for the evaluation of more advanced capacity vectors. By allowing the capacity to change over the week, knowledge about peaks and dips in the arrival process can be used to our advantage. The influence of the time dependency of the arrival rate is taken into account via the time dependent offered load $m(t)$. Figure 11a shows a clear time shift of the time dependent offered load with respect to the instantaneous offered load $\frac{\lambda(t)}{\mu}$ at time t. Furthermore, the fluctuation in the time dependent offered load are less extreme than in the instantaneous offered load. Both the observed time shift as well as the damping are in line with the literature discussed in 2.2.1.



(A) For the full range of the load.

(B) Zoomed in on the time dependent offered load.

FIGURE 11: The time dependent offered load and capacity compared to the instantaneous offered load.

Figure 11a also contains the optimal capacity for the base parameters as obtained with the infinite server approach and with the MOL approximation on the call packing system. Figure 11b is zoomed in on the offered load and the capacity vectors. Except for start-up behaviour, the time dependent behaviour of the optimal capacity for the infinite server approach and for the MOL approximation on the call packing system is the same. Analogous to the stationary results, the main capacity in the two system are 27 and 29, respectively. We observe a single drop in the optimal capacity on the same day as in the modified offered load (Wednesday). This drop in capacity is thus shift compared to the drop in the arrival rate (Saturday and Sunday). The offered load is plotted in the upcoming figures in this section as a reference line. It is important to keep in mind that this time dependent offered load is shifted with respect to the instantaneous offered load, and thus arrival rate. So, behaviour of the capacity that makes perfect sense when compared to this time dependent offered load might feel counterintuitive in reality when compared to the daily arrival rates.

5.2.1 Influence of κ

In the stationary analysis we observed that the optimal capacity in the (exponential) call packing system depends on κ. Using the time dependent MOL approximation of the call packing system, Figure 12 shows the influence of κ on the time dependent behaviour of the optimal capacity. The time dependent behaviour for $\kappa = 0.3$ and $\kappa = 1$ is the same, which we already saw in Figure 11 as the case $\kappa = 1$ corresponds to the infinite server system. Morover, the shapes for $\kappa = 0.2$, $\kappa = 0.5$ and $\kappa = 0.9$ are the same and have more fluctuations than $\kappa = 0.3$. In addition to the decrease in capacity at Wednesday, on which the drop of the modified offered load takes place, there is an increase in capacity on Thursday to Saturday. These capacity vectors thus mimic the time dependent offered load more closely. Concluding, κ influences the time dependent behaviour to a certain extent, where we observe reoccurring shapes over different values of κ. No analysis is performed based on the infinite server approach because the infinite server system does not depend on κ.

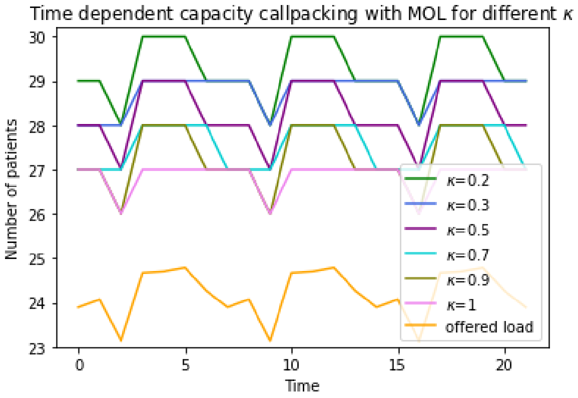


FIGURE 12: The optimal time dependent capacity for different values of κ.

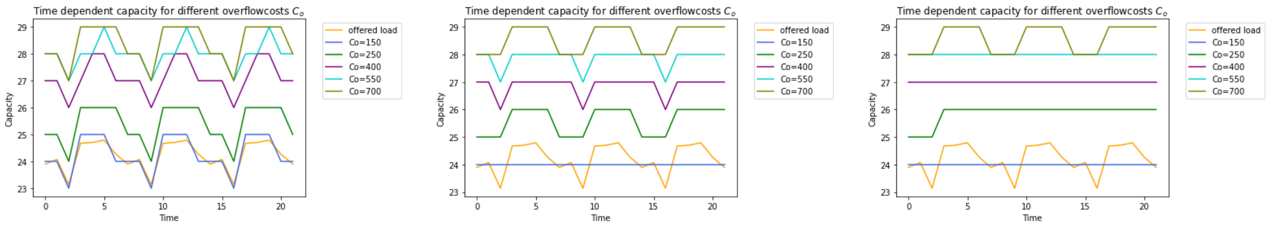
5.2.2 Influence of different cost parameters

The influence of the cost parameters C_o, C_u and C_b on the stationary capacity level was already studied in detail in the stationary analysis. We observed that the ratio between C_o and C_u was leading in the capacity decision. For the time dependent analysis, we will therefore only study the influence of C_o on the time dependent behaviour of the optimal capacity. In addition to these three cost parameters, the time dependency of the capacity asks for the introduction of a cost for changing capacity. After the analysis concerning C_o, the influence of the cost of changing capacity is studied in more detail by studying distinct values of the cost of increasing capacity C_i and of decreasing capacity C_d.

Although a Modified Offered Load approximation for the time dependent call packing system was introduced, the influence of the cost parameters is studied on the time dependent infinite server system. Based on our numerical experiments, similar results are expected for the MOL approximation of the call packing system.

Overflow cost C_o

Since we expect the time dependent behaviour of the capacity to be influenced by the cost of changing capacity, the influence of the cost of overflow C_o on the capacity vector is studied for different costs of changing capacity in Figure 13. The different lines in each subfigure correspond to different values of C_o for a fixed cost of increasing and decreasing capacity, $C_\mathrm{i} = C_\mathrm{d}$. There is no clear relation between the value of C_o and the time dependent behaviour of the optimal capacity for a given cost of changing capacity since the same shape of capacity vector occurs for different values of C_o. More specifically, for each cost of changing capacity the shape of the line corresponding to the base value $C_\mathrm{o} = 400$ matches that of other values of C_o. It is therefore likely that the patterns that we will observe from now for $C_\mathrm{o} = 400$ will hold for other overflow costs as well. The most noticeable differences in time dependent behaviour are between the subfigures 13a to 13c, that is, for different costs of changing capacity. This will be discussed in the next paragraph.



(A) $C_\mathrm{i} = C_\mathrm{d} = 0$.

(B) $C_\mathrm{i} = C_\mathrm{d} = 10$.

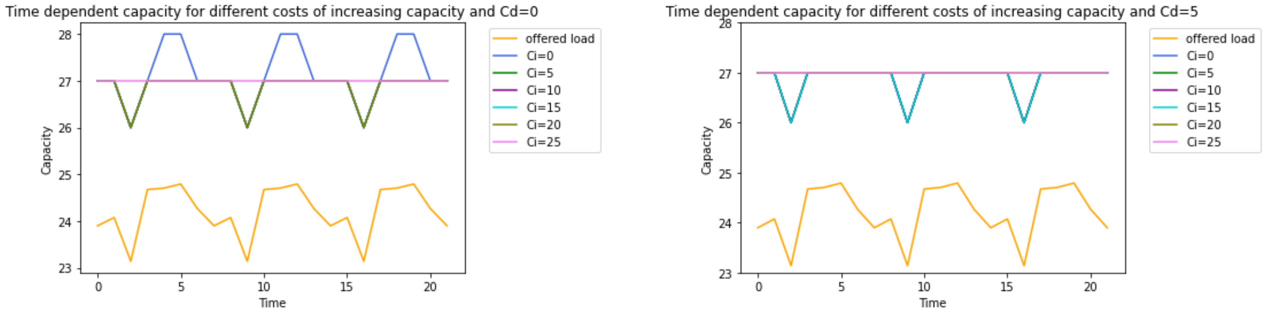
(C) $C_\mathrm{i} = C_\mathrm{d} = 20$.

FIGURE 13: Optimal time dependent capacity for different overflowcosts C_o, for three different values of costs for changing capacity C_i and C_d.

Cost of increasing capacity C_{i} and decreasing capacity C_{d}

The difference in shape of the optimal capacity vectors in Figures 13a to 13c indicate that the cost of changing capacity influences the time dependent behaviour of the optimal capacity strongly. In Section 3.2.2, we argued that the naive time dependent approach of concatenating the separately computed optimal capacity at each time point would just mimic the behaviour of the offered load $m(t)$. This corresponds to having no cost for changing capacity, as in Figure 13a. Indeed, we see the same pattern in the optimal capacity as in the time dependent offered load. Figures 13b and 13c show that increasing the cost of changing capacity flattens out the change in capacity. This illustrates the benefits of decision-making in which the complete time horizon is taken into account over the naive approach of an independent decision per time point. The observation that the fluctuations in time dependent capacity are the largest when there are no costs for changing capacity and that it then mimics the time dependent offered load, are in line with the observation of Bekker and de Bruin (2010) that the difference between the minimum and the maximum offered load $m(t)$ gives a good indication of the extent to what the capacity may vary. Although the increase in the cost of changing capacity is relatively small compared to the cost of unused beds or overflow, its influence on the fluctuation in capacity is significant. Due to the rapidly varying offered load and its short cycle time, a change in capacity is only useful for one or a few days and is thus rather expensive.

For the subfigures of Figure 13, one cost of changing capacity was used, regardless of it being an increase or a decrease. Figure 14 allows us to study the case in which the cost of increasing capacity C_{i} and the cost of decreasing capacity C_{d} are different. The cost of decreasing capacity is fixed at a certain value (per subfigure), after which we range the cost of increasing capacity. Quite some lines overlap in Figure 14, that is, the optimal capacity vectors for multiple values of C_{i} are the same. In Figure 14a, that is, for $C_{\mathrm{d}} = 0$, the lines for $C_{\mathrm{i}} = 5$ to $C_{\mathrm{i}} = 20$ overlap. In Figure 14b, the lines for $C_{\mathrm{i}} = 0$ to $C_{\mathrm{i}} = 15$ overlap and the lines for $C_{\mathrm{i}} = 20$ and $C_{\mathrm{i}} = 25$ overlap. Since the plots for $C_{\mathrm{d}} = 10$ to $C_{\mathrm{d}} = 20$ look the same as Figure 14b, except that with every increase of C_{d} by 5, the optimal capacity of first $C_{\mathrm{i}} = 15$, then $C_{\mathrm{i}} = 10$ and lastly $C_{\mathrm{i}} = 5$ become stationary at 27. Since the lines in the graphs overlap, this is not apparent from the graphs and therefore they are not included in this report.



(A) $C_{\mathrm{d}} = 0$.

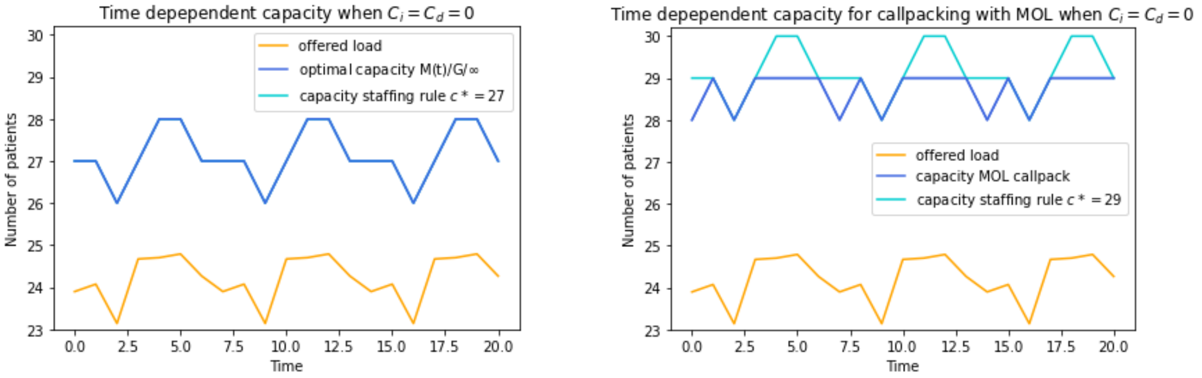
(B) $C_{\mathrm{d}} = 5$.

FIGURE 14: The optimal capacity for different costs of changing capacity

We observe the same decrease in variability of the capacity vectors as in Figure 13 when either ranging C_{i} (within one subfigure) or ranging C_{d} (between subfigures). Indeed, the parameter combinations ($C_{\mathrm{d}} = 0$, $C_{\mathrm{i}} = 20$), ($C_{\mathrm{d}} = 5$, $C_{\mathrm{i}} = 15$) and ($C_{\mathrm{d}} = 10$, $C_{\mathrm{i}} = 10$) yield the same optimal capacity vector that can be found in, respectively, Figure 14a (olive line), Figure 14b (cyan line) and Figure 13b (purple line). This observation indicates that the separate values of C_{d} and C_{i} do not matter, but only their sum. This makes sense as an increase in capacity always implies a decrease in capacity, and vice versa, for a cyclic capacity vector. So, in the case of cyclic offered loads, there is no added value of having separate costs of increasing and decreasing capacity compared to one cost for changing capacity.

5.2.3 Comparison to the square root staffing rule

In Section 2.3, we discussed the square root staffing rule $c(t) = m(t) + \beta\sqrt{m(t)}$ which is a well-known heuristic for time dependent staffing. Bekker and de Bruin (2010) determine the quality of service parameter β via $\beta = \frac{c^* - \rho}{\sqrt{\rho}}$, where c^* is the desired average number of beds and ρ is the average offered load over the complete period. The outcome of this staffing rule is compared to the analytic time dependent results of both the infinite server system as well as the MOL approximation of the call packing system. These results are given in Figure 11 for the base parameters and in Figure 15 when there are no costs for changing capacity.



(A) Infinite server system.

(B) MOL approximation of callpacking system.

FIGURE 15: The optimal capacity when $C_{\mathrm{i}} = C_{\mathrm{d}} = 0$ for the two time dependent systems. The capacity vector obtained from the square root staffing rule, when the desired average capacity c^* equals the optimal stationary capacity of the system under consideration, is plotted as well.

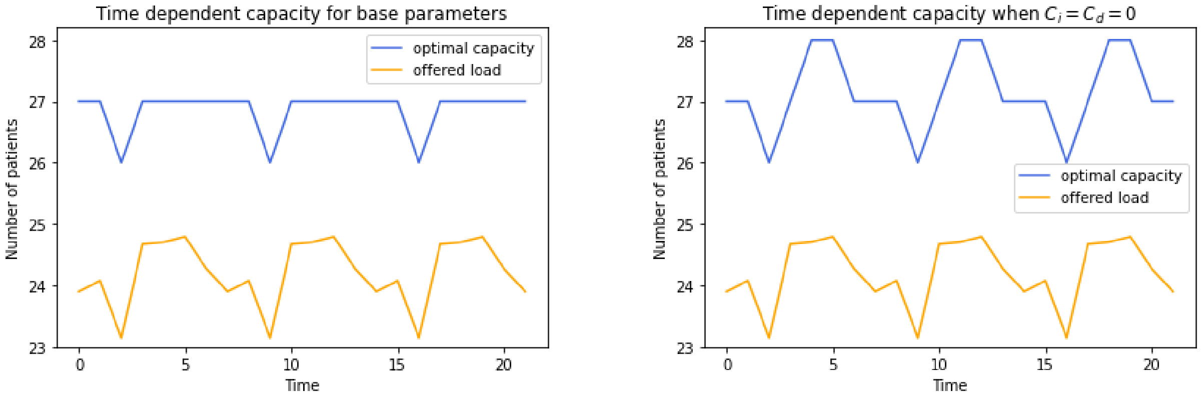
For c^*, we take the optimal capacity found in the stationary analysis of the corresponding system. The average offered load is $\rho = 0.61 * 39.79 = 24.27$. For the infinite server system, which has $c^* = 27$, we obtain $\beta = 0.55$. The square root staffing rule then gives a capacity vector [26.6, 26.8, 25.8, 27.4, 27.5, 27.5, 27.0]. If we round each $c(t)$ to the nearest integer, we obtain [27, 27, 26, 27, 28, 28, 27]. This is the same optimal capacity vector as in Figure 15a. Since the square root staffing rule is derived from the infinite server system and does not take into account costs for changing capacity, the reader might not be surprised by this result. However, the costs of the other parameters have been shown to influence the optimal capacity decision while the selected values of these parameters do not influence the outcome of the square root staffing rule.

The stationary analysis of call packing with base parameters gave an optimal capacity of 29. Setting $c^* = 29$, for which $\beta = 0.96$, gives a square root staffing vector of [28.6, 28.8, 27.8, 29.4, 29.5, 29.6, 29.0], which is just two higher than for the $M/G/\infty$ system. Interestingly, the optimal capacity vector resulting from the MOL approximation of the call packing system for $C_{\mathrm{i}} = C_{\mathrm{d}} = 0$ as given in Figure 15b fluctuates less.

In conclusion, the analytic results in this research align with the square root staffing rule. In addition, the analytic analysis in this paper can be customized by setting the appropriate parameter values.

5.2.4 Additional analysis with the Mixed Integer Program

This research introduced a Mixed Integer Program in the optimization phase. The MIP enables easy incorporation of constraints on the behaviour of the capacity. Moreover, parallel optimization over multiple patient types allows for interesting concepts such as shared capacity or relabeling of capacity. The influences of these additions to the optimization phase on the optimal capacity are shown here. The analysis is based on the infinite server system. The optimal capacity vector for the base parameters is given in Figure 16a.



(A) Base parameters, includes $C_\mathrm{d} = 5$ and $C_\mathrm{i} = 10$.

(B) Base parameters, but with $C_\mathrm{i} = C_\mathrm{d} = 0$.

FIGURE 16: A reference picture of the optimal solution in the time dependent infinite server system when no constraints on capacity and only one patient type.

Constraints on the behaviour of the time dependent capacity

Three possibilities for constraining the time dependent behaviour of the capacity have been introduced in Section 3.2.2. Every constraint will be analysed separately. Since the change in capacity over time is very limited, constraining the time dependent behaviour of the capacity vector in Figure 16a will yield either the same vector or a stationary capacity of 27. In order to demonstrate the working of the constraints, the case in which there is no cost for changing capacity is selected as 'standard' case instead. Figure 16b shows the optimal capacity for this scenario.

Capacity should stay fixed for a minimum number of days. Figure 17 shows the optimal capacity for different numbers of days on which the capacity should stay fixed. If the capacity should stay fixed for at least two days, both the peak as well as the drop in time dependent offered load are still visible in the optimal capacity. Having a fixed capacity for four or more days results in a stationary optimal capacity. Changing capacity under such a condition would conflict with the periodicity of 7 days and is thus not expected.

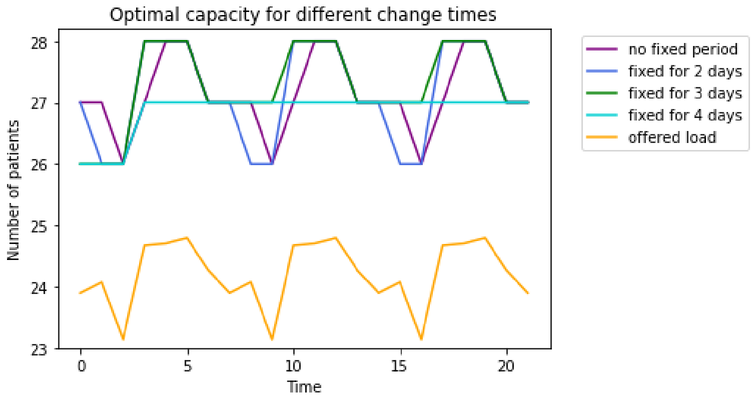
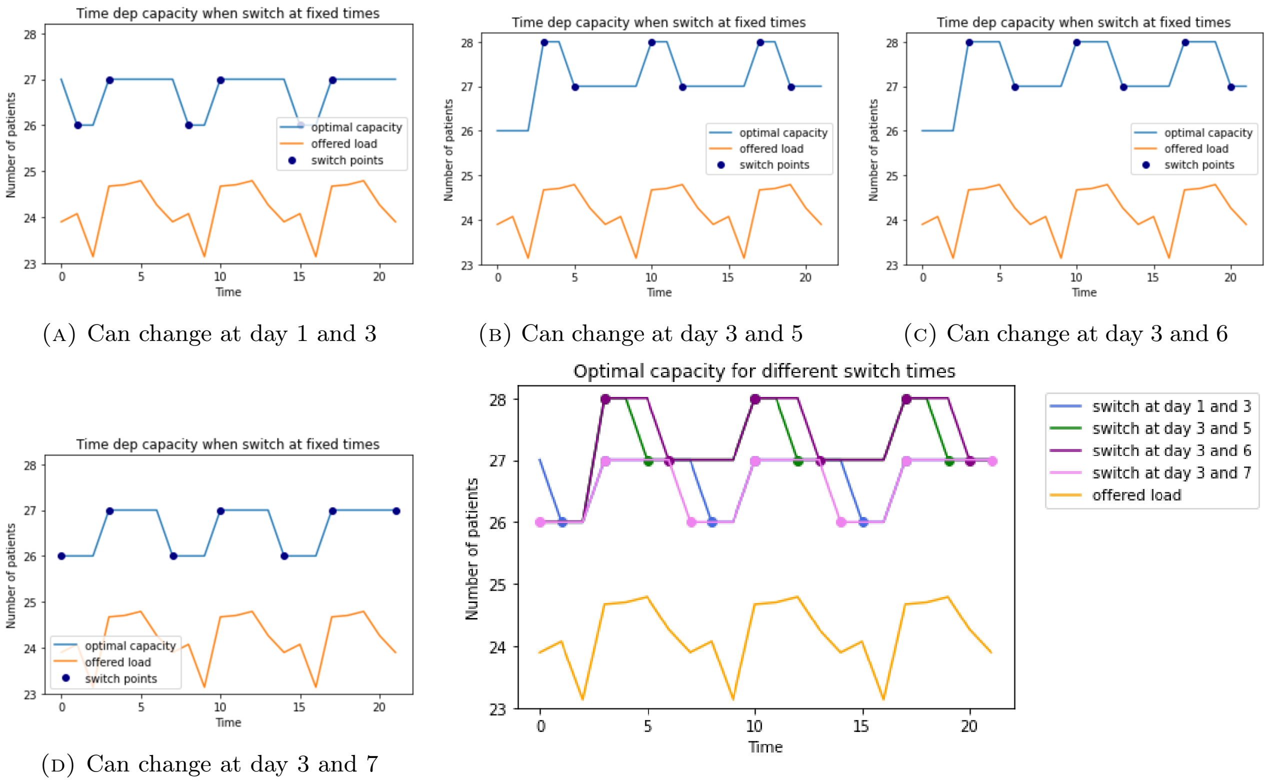


FIGURE 17: The optimal capacity when the capacity should stay fixed for a minimum number of days.

Capacity may only change at set time points. To study what happens if the capacity is only allowed to change at set time points, several sets of two days per week are selected on which the capacity may change. The optimal capacity for each set of days is given in Figures 18a to d separately and in Figure 18e combined. The dots indicate the days before which the capacity was allowed to switch. Please be aware that day 0 refers to Monday when interpreting the legend. Taking a capacity of 27 as reference, the optimal capacity adapts either to the peak in time dependent offered load at day 3 to 5 or to the drop in offered load at day 2. Depending on which days the capacity is allowed to switch, the option is selected that leads to the smallest number of days of capacity unequal to the reference capacity 27.



(A) Can change at day 1 and 3

(B) Can change at day 3 and 5

(C) Can change at day 3 and 6

(D) Can change at day 3 and 7 (or 0)

(E) Collection of the individual plots.

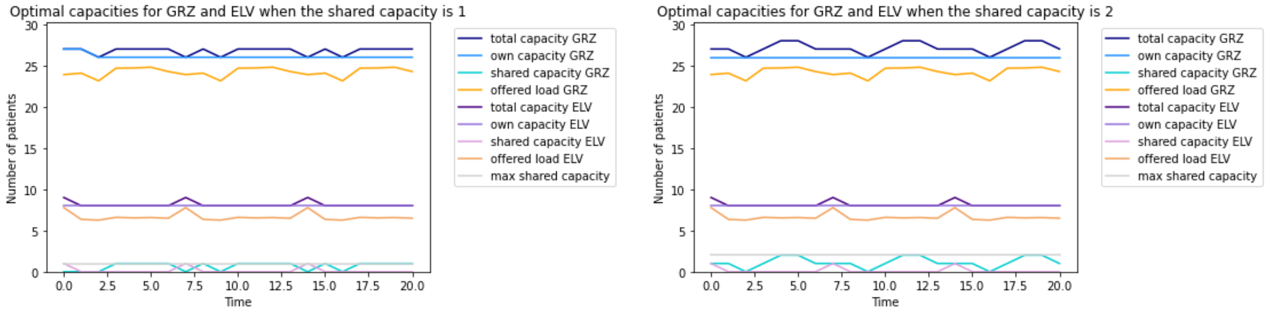
FIGURE 18: The optimal capacity when the capacity can only change at two fixed days of the week.

Capacity can change only by a limited amount. Since the capacity does not change by more than one bed per time, this constraint has no added value in this numerical analysis.

Different patient types

Since the Mixed Integer Program can decide on the optimal capacity for multiple VVT types at once, dependencies between the capacity of the different types could be incorporated. In order to demonstrate this, parameters for a second VVT type are needed. Based on the POINT data set, the arrival process to the ELV (*eerstelijnsverblijf*, a stay in primary care) of the VVT organisation from which the GRZ data originates is determined. With $\xi = 0.95$, a total arrival rate vector $[0.20, 0.25, 0.31, 0.15, 0.27, 0.05, 0.03]$ is obtained. Unfortunately, we have no data on the length of stay of patients in this ELV care. Therefore, a mean length of stay of 35.77 days is used which is the weighted average of the mean length of stay of low complexity ELV care and high complexity ELV care in 2018 as reported by ActiZ [9]. The coefficient of variation is set equal to that of the GRZ.

Shared capacity. Figure 19 shows the optimal capacity for both types when there is shared capacity available. The maximum shared capacity available is 1 or 2 in Figures 19a and 19b, respectively. We observe that a stationary type-specific capacity is optimal for both types. With a maximal shared capacity of 2, the capacity for both types can already freely accommodate to the offered load by only using the shared capacity. When there is only a shared capacity of 1 available, that capacity is used on days 7 and 14 for the ELV to accommodate the peak in the modified offered load of the ELV as a result of which the total capacity in the GRZ decreases on that day by one.



(A) The shared capacity is at most 1.

(B) The shared capacity is at most 2.

FIGURE 19: The optimal capacity for the GRZ and ELV when a shared capacity is available.

5.3 Simulation

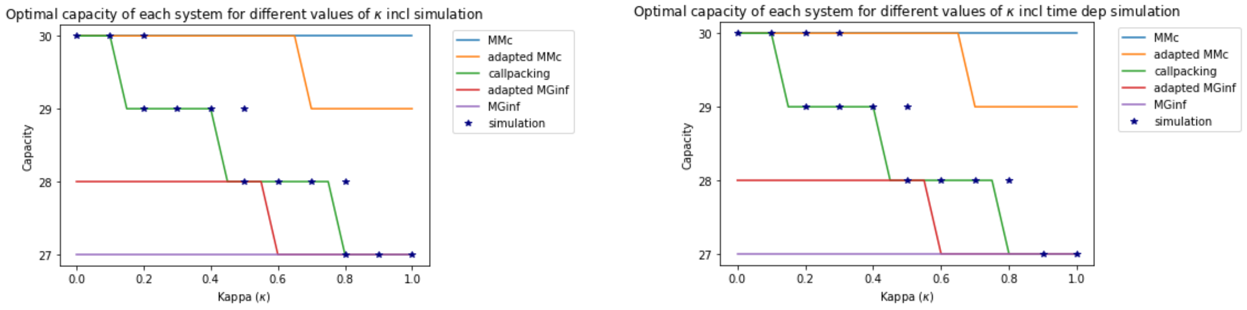
An analytical analysis generally requires some simplifications or assumptions that can influence the direct applicability of the result to reality. In this research, this is both the time dependency of the arrival process as well as the nonexponentiality of the service time distribution. If the analytic outcomes are appropriately close to the optimal solution in reality, the analytic models can be used to obtain an easy reference value or approximation of the optimal solution or as a low-fidelity model for multi-fidelity modelling, as discussed in Section 2.1.1. A simulation model can be used to study the validity of the analytic outcomes and as a high-fidelity model to find the optimal solution for the realistic scenario by a neighbourhood search on the analytic result.

In this section, a simulation is used to study how well the capacity (vectors) found with the analytical analysis would perform in reality. Moreover, a neighbourhood search on the time dependent capacity vector is performed.

The simulation model. A simulation should mimic reality as well as possible. As explained at the start of Section 3.1, each patient type has its own station with c_r servers and infinite waiting room to which patients arrive in a Poisson process with an arrival rate that depends on the day of the week. The service time of a patient is lognormally distributed with the fitted mean and variance, but is reduced by the time they spent in overflow weighted by a fraction κ. The realisation of the service time of a patient, S_i, is determined upon arrival since they may leave the system from overflow when they are in overflow for at least $\frac{1}{\kappa}S_i$ days. The simulation is a Discrete Event Simulation with three events: *Arrival*, *EndService* and *EndOverflowService*. The idea of a Discrete Event Simulation is that it suffices to study the system only when an event occurs since that is the only moment that the state of the system changes. In the case of time dependent capacity, an additional event at the start of each day guarantees that the capacity is increased or decreased if applicable. In case of an increase in capacity, someone from the queue is taken into service. However, if the capacity is decreased, it would not be realistic to move a patient in service back to the queue. Instead, a cost C_c is charged for every patient more in service than the current capacity, i.e. for every *overbed*. Based on a small study on varying C_c when the time dependent capacity vector is [29,28,29,29,29,29,29] as obtained from the analytic analysis, we conclude that there is no significant influence of the exact value of C_c. The simulation results in this section are obtained with $C_\text{c} = 300$ fixed, as an overbed is expected to be more expensive than the scheduled VVT capacity but less than an overflow bed in the hospital. By performing multiple simulation runs, a confidence interval for the average total cost per day is obtained from the simulation. More details on the collection of statistics are given in Appendix A.4.

5.3.1 Stationary capacity

Firstly, the simulation with exponential service times and a stationary arrival rate yields for a stationary capacity of 29 a cost of 811.61 [806.608, 816.617]. With these properties, the simulation matches an exponential call packing system for which the optimal stationary capacity of 29 had an objective value of 812.31. Then, the simulation is used to study how well each of the five systems in the stationary analysis approximates a call packing system with lognormal service times. This is done both for stationary arrivals as well as the specific time dependent arrival rates. Since it depends strongly on κ by which system the exponential call packing system is best approximated, as was shown in Figure 7, the optimal capacity is determined for different values of κ. To ease the comparison with the analytic results, the optimal stationary capacity as obtained from the simulation for each value of κ is added to the analytic plot using a star in Figure 20. This is first done for the case of stationary arrivals in Figure 20a and then for the case of time dependent arrivals in Figure 20b. For some values of κ, the confidence intervals of the total cost per day for two capacities overlap. In such cases, the number of simulation runs is increased to 1000. If the confidence intervals still overlap, both the capacities are marked with a star as the optimal capacity can not uniquely be determined.



(A) Stationary arrival rate.

(B) Time dependent arrival rate.

FIGURE 20: The optimal stationary capacity obtained with simulation for different values of κ plotted in combination with the analytic stationary results.

Figure 20a shows that the optimal capacity from simulation equals the analytic optimal capacity of the exponential call packing system. In case the optimal capacity cannot be uniquely determined, the possibilities for the optimal capacity are the optimal capacity of the exponential call packing system and one higher. This occurs every time at the first single-decimal value of κ after the analytic call packing result decreases by one. These observations imply that the lognormal service times likely require a little more capacity than the exponential service times. Remarkably, the lognormal service distribution has a coefficient of variation of 0.68 compared to the coefficient of variation of 1 of the exponential distribution. Based on the fact that a lower variation allows for less capacity and the observation in Figure 22 in Appendix A.2 that the overflow result for the lognormal distribution with a coefficient of variation of 0.5 is approximately in the middle between that of the call packing system and that of the adapted infinite server system, it is remarkable that the optimal capacity is possibly higher (instead of lower) than that of the exponential call packing system. We note that the cost corresponding to the optimal capacity from the simulation is significantly higher than that of the analytic exponential call packing system for most values of κ. For example, the simulation gives a cost of 892.13 ([887.01, 897.243]) for a capacity of 29 for $\kappa = 0.3$ while the exponential call packing has an objective value of 812.32. The difference between the cost values decreases over κ until the analytic objective value is within the confidence interval obtained by the simulation for $\kappa = 1$.

To study the influence of the time dependency of the arrival rate on the optimal capacity, we first look at the case $\kappa = 1$. Since the call packing system with $\kappa = 1$ is simply the $M/G/\infty$ system and is thus insensitive to the service distribution, the only difference between the analytic result and the call packing system is the time dependency of the arrival process. For $\kappa = 1$, the simulation gives an optimal capacity of 27 for cost 666.61 ([663.327, 669.894]). Although the analytic objective value of 663.61 is close to the lower bound of the confidence interval, there is no significant difference. Figure 20b is similar to Figure 20a with as main difference that an optimal capacity of 28 for $\kappa = 0.8$ is now significant, which is one higher than the optimal capacity of the exponential call packing. Furthermore, the confidence intervals belonging to a capacity of 29 and 30 overlap for $\kappa = 0.3$. Since $\kappa = 0.3$ is a base parameter, the lack of a unique result on the optimal capacity for this parameter value is unfortunate. In general, it seems like the influence of the time dependency of the arrival rate on the stationary capacity results is small.

Based on the results concerning the stationary capacity, the system is best approximated by the analytic exponential call packing system.

5.3.2 Time dependent capacity

For the time dependent analysis, we have only two models compared to the five in the stationary case. Since the capacity is allowed to change over time in the time dependent analysis, a planned decrease in capacity may not be possible in reality due to all initial capacity being occupied. Recall that the simulation takes into account the *overbeds* that result from this through the cost parameter C_c, as mentioned at the beginning of Section 5.3.

Evaluation of the analytic solutions

The simulation is first used to determine how the optimal time dependent capacity vectors of the two models, which were given in Figure 11, would perform in reality.

The first model is the infinite server approach. The analytically determined optimal capacity is 27 on each day, except on day 2 on which it is 26. An objective value of 667.83 belonged to this time dependent capacity vector. The capacity vector [27, 27, 26, 27, 27, 27, 27] brings in the simulation a cost of 1043.73 ([1035.587, 1051.879]). It is not surprising that the infinite server approach does not perform well in reality. From the stationary results in Figure 20, we already concluded that the optimal capacity based on an infinite server system underestimates the necessary capacity for $\kappa = 0.3$.

The second model is the Modified Offered Load approximation of the call packing system. The optimal analytic capacity vector [29, 29, 28, 29, 29, 29, 29] costs in the simulation 889.70 ([884.526, 894.869]) compared to its objective value of 816.99. The simulated cost of the capacity vector obtained via the MOL callpacking system is thus closer to the objective value. There are several differences between the analytic analysis and the simulation that could cause a difference in the analytic objective value and the cost obtained by the simulation. Although the lognormal service distribution is used to determine the time dependent offered load, the formulas (20) and (18) for the performance measures in the call packing system only hold for exponential service distributions. However, in the stationary analysis we concluded that the results of the exponential call packing system match the simulation results quite well, so we expect the influence of the nonexponential service times to be limited. The main inaccuracy is expected to be caused by the fact that the performance measures (of the finite server model) are based on the offered load of an infinite server system. Especially in the case of a time dependent capacity, this means that the values of the performance measures on the next days are not influenced by the choice of capacity on the previous days. Since limited capacity can lead to overflow and thus a higher load in the next period, the inability to take the limited capacity into account can lead to an underestimation of the optimal capacity (on some days), or put differently, to an underestimation of the cost of capacity vectors that have less capacity than the real optimal capacity vector. In addition to the shortcoming of the analytic model to take the influence of limited capacity into account, the analytic model can not account for overbeds since it is based on steady-state behaviour. The extra costs of having overbeds are present in the simulation which can lead to an increase in cost obtained by the simulation.

Comparison to neighbouring vectors

To judge the performance of the analytically found capacity vector [29, 29, 28, 29, 29, 29, 29] better, we will compare its costs to that of neighbouring vectors. Table 6 contains the simulated costs for some vectors in the neighbourhood of [29, 29, 28, 29, 29, 29, 29].

TABLE 6: The costs of several neighbouring vectors of the optimal analytic capacity vector obtained with the simulation.

Capacity vector	Cost per day (Confidence interval)
$[29, 29, 28, 29, 29, 29, 29]$	892.73 ([887.625, 897.835])
$[29, 29, 29, 29, 29, 29, 29]$	892.13 ([887.01, 897.243])
$[29, 29, 28, 30, 29, 29, 29]$	874.15 ([869.231, 879.077])
$[29, 29, 28, 29, 30, 29, 29]$	879.43 ([874.561, 884.298])
$[29, 29, 28, 30, 30, 29, 29]$	870.54 ([865.995, 875.075])
$[29, 29, 28, 29, 30, 30, 29]$	891.36 ([885.587, 895.13])
$[29, 29, 28, 30, 30, 30, 29]$	888.07 ([883.472, 892.662])
$[30, 30, 29, 30, 30, 30, 30]$	894.54 ([890.048, 899.034])

First of all, it performs the same as the stationary capacity of 29. This is not surprising as the simulation results for the stationary capacity already showed that the optimal capacity in the simulation might be higher than in the analytical system. Moreover, an overbed, which is likely to occur due to the single drop in capacity, is expensive. Based on the behaviour of the offered load, more capacity on day 3 or day 4 or both might be beneficial. The costs for these capacity vectors are indeed significantly lower. Among these vectors, there is no significantly cheaper one. Furthermore, the vectors [29, 29, 28, 29, 30, 30, 29] and [29, 29, 28, 30, 30, 30, 29] are considered as we notice that in Figure 15a, the increased capacity for the infinite server model happened on day 4 and 5, instead of day 3 and 4. Interstingly, these capacity vectors perform significantly worse than the previous ones. Lastly, based on the fact that the optimal stationary capacity could be 30 for the case of time dependent arrivals instead of 29, as shown in Figure 20b, the vector [30, 30, 29, 30, 30, 30, 30] is considered. The cost corresponding to this capacity is significantly higher than those for options like [29, 29, 28, 30, 30, 29, 29], but is comparable to the cost of [29, 29, 28, 29, 29, 29, 29]. Therefore, the optimal capacity vector is likely in the neighbourhood of these two vectors.

Earlier, we reasoned that the analytic time dependent result based on the MOL call packing system might underestimate the optimal capacity (on some days) due to the shortcoming of the analytic model to take the influence of limited capacity into account. Based on Table 6, we conclude that there is an underestimation of capacity on some days, but not extreme.

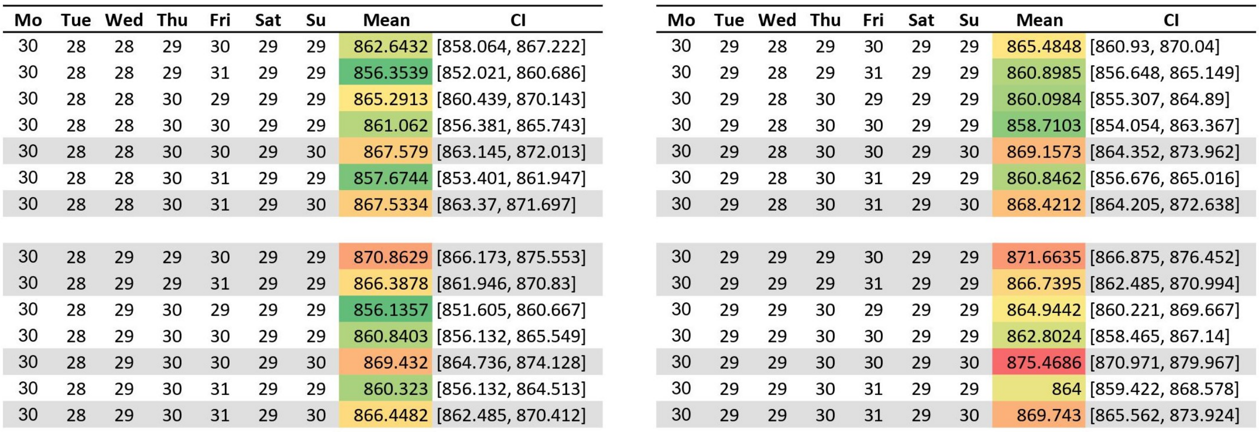
Extensive neighbourhood search

In addition to the comparison with some neighbouring vectors, we would like to know how the optimal analytic capacity performs compared to the real optimal capacity vector. To find this optimal capacity vector, we perform a neighbourhood search. The cost for a large set of capacity vectors is determined via simulation. Based on the stationary results, a reference capacity of 29 is used, which is tweaked based on the time dependent behaviour of the offered load.

As shown in Figure 11, the offered load is middle to low on days 0 to 2 and high on days 3 to 6. We will therefore consider all capacity vectors that have a capacity of 29 or 28 on days 0 to 2 and a capacity of 29 or 30 on days 4 to 6. Moreover, a capacity of 30 on day 0 and a capacity of 31 on day 4 is allowed as well. Due to the large amount of capacity vectors, the cost for each capacity vector will be calculated based on 100 simulation runs. Although this causes the width of the confidence intervals to increase, it allows the evaluation of a large amount of vectors in reasonable computation time. With these results, we can see if there is a pattern in vectors that perform well or badly.

An overview of the vectors and their costs as obtained from the simulation can be found in Appendix A.5. A colour scale is used on the mean cost, where dark green indicates the lowest cost, and thus the best-performing capacity vectors in their simulation runs. The vectors in the three columns have respectively a capacity of 28, 29 and 30 on Monday. In general, the vectors with a capacity of 30 on Monday perform the best, which is surprising as there is no increase in offered load around that day and the capacity is lower on the neighbouring days. We observe a repeating pattern of capacity vectors with low means. These vectors are outlined in black in Appendix A.5. These vectors have capacities on Thursday till Sunday corresponding to one of the following subvectors: $\{[29, 30, 29, 29], [29, 31, 29, 29], [30, 29, 29, 29], [30, 30, 29, 29], [30, 30, 29, 30], [30, 31, 29, 29], [30, 31, 29, 30]\}$. Note that all these vectors have a capacity of 29 on Saturday. This is in line with the comparison made in Table 6, but remarkable as the time dependent offered load is still high on Saturday. The capacity on Tuesday and Wednesday does not seem to matter. Since the size of the confidence intervals makes it hard to draw conclusions on which vector is optimal, the outlined vectors are analysed again using 500 simulation runs to construct the confidence interval of their cost. The results are given in Figure 21. No vector performs significantly better than the other vectors. However, a collection of vectors performs significantly worse than the vector with the lowest mean, $[30, 28, 29, 30, 29, 29, 29]$. The background of these vectors is made light grey.

We see again that the influence of the capacities on Tuesday and Wednesday is limited, while the combination of capacities on the remainder of the week is leading. The subvectors for Thursday till Sunday [30,30,29,30] and [30,31,29,30] perform badly for all combinations of capacities on Tuesday and Wednesday. This would indicate that a capacity of 29 on Sunday would be beneficial. If the capacity on Wednesday is 29, the subvectors [29,30,29,29] and [29,31,29,29] perform badly as well. Thus, we observe that increasing capacity from Wednesday to Thursday outperforms staying at a capacity level of 29.



Mo	Tue	Wed	Thu	Fri	Sat	Su	Mean	CI
30	28	28	29	30	29	29	862.6432	[858.064, 867.222]
30	28	28	29	31	29	29	856.3539	[852.021, 860.686]
30	28	28	30	29	29	29	865.2913	[860.439, 870.143]
30	28	28	30	30	29	29	861.062	[856.381, 865.743]
30	28	28	30	30	29	30	867.579	[863.145, 872.013]
30	28	28	30	31	29	29	857.6744	[853.401, 861.947]
30	28	28	30	31	29	30	867.5334	[863.37, 871.697]
30	28	29	29	30	29	29	870.8629	[866.173, 875.553]
30	28	29	29	31	29	29	866.3878	[861.946, 870.83]
30	28	29	30	29	29	29	856.1357	[851.605, 860.667]
30	28	29	30	30	29	29	860.8403	[856.132, 865.549]
30	28	29	30	30	29	30	869.432	[864.736, 874.128]
30	28	29	30	31	29	29	860.323	[856.132, 864.513]
30	28	29	30	31	29	30	866.4482	[862.485, 870.412]

Mo	Tue	Wed	Thu	Fri	Sat	Su	Mean	CI
30	29	28	29	30	29	29	865.4848	[860.93, 870.04]
30	29	28	29	31	29	29	860.8985	[856.648, 865.149]
30	29	28	30	29	29	29	860.0984	[855.307, 864.89]
30	29	28	30	30	29	29	858.7103	[854.054, 863.367]
30	29	28	30	30	29	30	869.1573	[864.352, 873.962]
30	29	28	30	31	29	29	860.8462	[856.676, 865.016]
30	29	28	30	31	29	30	868.4212	[864.205, 872.638]
30	29	29	29	30	29	29	871.6635	[866.875, 876.452]
30	29	29	29	31	29	29	866.7395	[862.485, 870.994]
30	29	29	30	29	29	29	864.9442	[860.221, 869.667]
30	29	29	30	30	29	29	862.8024	[858.465, 867.14]
30	29	29	30	30	29	30	875.4686	[870.971, 879.967]
30	29	29	30	31	29	29	864	[859.422, 868.578]
30	29	29	30	31	29	30	869.743	[865.562, 873.924]

FIGURE 21: The costs corresponding to the capacity vectors outlined in black in Appendix A.5, based on 500 simulation runs. The vectors that have significantly higher cost than $[30, 28, 29, 30, 29, 29, 29]$ are coloured light grey.

Since the extra analysis is only performed for the capacity vectors that satisfied the pattern of well-performing vectors, and since the confidence intervals of the costs of a reasonable number of vectors in this extra analysis still overlap, no guaranteed optimal time dependent capacity vector is found. The overlapping confidence intervals of the capacity vectors with low means suggest that multiple capacity vectors could perform well. One could choose a capacity vector from this collection that fits certain preferences, like changing capacity as little as possible or having less capacity on a certain day due to the availability of personnel.

Square root staffing rule

In the time dependent analytic analysis, we compared the results to the square root staffing rule. Most often the quality of service parameter β is determined based on the desired maximum probability of delay. Although Bekker and de Bruin (2010) gave a formula for β based on the average number of beds decided upon, this formula is not found in other literature. To validate the choice of β, which had for $c^* = 29$ a value of 0.96, the simulation is used to determine the cost of the vectors originating from different values of β. Simulations were performed for β ranging from 0.2 to 2.0 with steps of 0.2. Since the lowest costs were obtained when $\beta = 0.8$ and $\beta = 1.0$, the values 0.7, 0.9 and 1.1 were simulated as well and can be found in Table 7.

TABLE 7: The simulated cost for the best-performing values of β and the corresponding capacity vector as obtained by the square root staffing rule when rounding to the nearest integer.

β	Capacity vector	Cost per day (Confidence interval)
0.6	[27, 27, 26, 28, 28, 28, 27]	988.34 ([981.163, 995.514])
0.7	[27, 28, 27, 28, 28, 28, 28]	943.87 ([937.454, 950.288])
0.8	[28, 28, 27, 29, 29, 29, 28]	906.65 ([901.222, 912.087])
0.9	[28, 28, 27, 29, 29, 29, 29]	900.71 ([895.126, 906.298])
1.0	[29, 29, 28, 30, 30, 30, 29]	880.48 ([875.847, 885.112])
1.1	[29, 29, 28, 30, 30, 30, 30]	882.02 ([877.482, 886.555])
1.2	[30, 30, 29, 31, 31, 31, 30]	909.94 ([905.872, 914.008])

The capacity vectors corresponding to $\beta = 1.0$ and $\beta = 1.1$ have the lowest cost. This is rather close to $\beta = 0.96$, which yields capacity vector [29,29,28,29,30,30,29], that resulted from the formula of Bekker and de Bruin (2010) with $c^* = 29$ based on the stationary analysis. The performance of this capacity vector was already studied in Table 6 and had cost 891.36 ([885.587, 895.13]). The vectors obtained from the square root staffing rule are outperformed by the best-performing vectors from the simulation study in Figure 21. For example, the capacity vector $[30, 28, 29, 30, 29, 29, 29]$ had a cost of 856.14 ([851.605, 860.667]). The best-performing vectors in Figure 21 differ from the vectors obtained from the square root staffing rule mainly by having an increased capacity on Monday and not having an increased capacity on the weekends.

Conclusion

In conclusion, the results of the simulation study in Appendix A.5 and in Figure 21 show that well-performing vectors have a base capacity of 29, corresponding to the stationary result for the exponential call packing system, with a lower capacity on Tuesday or Wednesday or both and a higher capacity on Monday and Thursday or Friday or both. Except for the increase in capacity on Monday, the decrease and increase in capacity are in line with the time dependent behaviour of the offered load. The time dependent behaviour of the optimal capacity is thus shifted with respect to the arrival rate function, as Figure 11a already suggested. The objective value of the optimal time dependent capacity as found by the analytic analysis is significantly higher than the cost of the vectors found with the simulation study in Figure 21. Since we already reasoned that the analytic result would give an underestimation, increasing the capacity on some days based on the offered load as was done in Table 6 yields capacity vectors that are reasonably close to the expected optimum. The analytic optimal vector performs approximately the same as the optimal vector obtained by the square root staffing rule with $c^* = 29$. It is outperformed by the vector obtained by the square root staffing rule with β optimised through simulation. However, the vector with increased capacity on Thursday and Friday outperforms this optimal square root staffing rule vector. So, for the researched parameter combination, the approximations based on the call packing system performed best compared to the simulated reality. The result from the stationary analysis with the call packing system performs rather well with a cost of 892.13 ([887.01, 897.243]). No improvement in simulated cost (892.73 ([887.625, 897.835])) was obtained by the result from the time dependent analysis. However, the capacity vectors resulting from the simulation study significantly decreased the cost to, for example, 856.14 ([851.605, 860.667]) for $[30, 28, 29, 30, 29, 29, 29]$.

5.4 Difference in optimal for VVT and optimal for system

The situation in which the VVT decide their capacity on their own was compared to the optimal capacity from a system perspective for the stationary case in Section 5.1.3. Since the results show limited added value of the time dependent analysis, the conclusion based on the stationary analysis is expected to hold for the case of time dependent capacity as well. The time dependent MOL approximation of the call packing system is used to determine the optimal time dependent vectors and their costs are determined via the simulation. The results in Table 8 are comparable to those based on the stationary analysis in Table 4. Instead of the analytic results, a well-performing vector in the simulation study like [30,28,29,30,29,29,29] could have been used. This gives a cost for VVT of 526.16 ([521.887, 530.435]). One would like to compare that to the equivalent best-performing vector in simulation for the scenario in which the VVT would decide. However, that would require a complete simulation study, while the result would not add to the message that is already clear from Table 8. The time dependent results support the stationary results: it would be reasonable (and necessary) to compensate the VVT for increasing their capacity since it is expected to pay off for the system as a whole.

Table 8: The optimal analytic time dependent capacity and associated cost from both the system perspective and the VVT on its own. The situation in which only the VVT is considered is modelled by setting $C_o = 0$.

	Hospital & VVT decide	VVT decides
Analytic optimal capacity	$[29, 29, 28, 29, 29, 29, 29]$	$[19, 19, 19, 19, 19, 19, 19]$
Cost for system	892.73 ([887.625, 897.835])	6661.49 ([6629.89, 6693.097])
Cost for VVT	483.55 ([479.572, 487.536]))	-187.82 ([-188.034, -187.614])
Mean number of overflow	1.06 ([1.042, 1.076])	18.03 ([17.947, 18.116])

6 Discussion

Several assumptions made in this research are questioned or explained in this section. Directions for further research are mentioned and more ideas are proposed. The discussion topics can roughly be categorized as belonging to a section of this research: Sections 6.1.1 and 6.1.2 cover some theoretical gaps and relate the application to the theory for the queueing models and the optimization model, respectively, Section 6.2 discusses the challenges of the data and Section 6.3 adds to the discussion of the results. Section 6.4 concludes with some remarks about the application of this research.

6.1 Models

6.1.1 Queueing models

For the stationary case, we tried to approximate the call packing system using an infinite server system and a system with c servers by properly adapting the service time in those systems. Based on the numerical comparison in Appendix A.2, we formulated Conjectures 1 and 2 that state that the mean number of patients in overflow in the call packing system is bounded from below and above by that obtained from the Adapted Service Time Approach applied to the $M/G/\infty$ system and $M/M/c$ system, respectively. Unfortunately, we were not able to prove these bounds during this research. This could be a topic for future research. For another direction of future research, we note that the Adapted Service Time Approach was only introduced for the mean number of people in overflow. The other performance measure, the mean number of unused beds, was determined from its simple relation to the mean number of people in overflow for an infinite and finite server system. It would be interesting to study if the mean number of unused beds could be determined directly by a similar Adapted Service Time Approach. Since we concluded from the results that the number of unused beds in the adapted $M/G/\infty$ system is rather low, extra research into determining the number of unused beds is advisable. A third topic for future research arises from the Adapted Service Time Approach applied to the $M/M/c$ system. When determining the fixed point L^*, we observed the two-cyclic behaviour of the repeated substitution. This raised two questions: can we relate the occurrence of two-cyclic behaviour to the parameter values, so that we know when it converges to a single value, and how could a reasonable guess for L^* be made based on L^- and L^+ in case it does not converge to one value?

The Adapted Service Time Approach was performed on both boundary cases of the call packing system. Since the product form formula for the call packing system only holds for exponentially distributed service times, the insensitivity of the adapted infinite server system makes the conjectured lower bound very useful. The conjectured upper bound provided by the adapted finite server system is less useful since its applicability is, just like the call packing system, limited to the case of exponential service times. In Section 2.1.2, we mentioned that an accurate approximation for the mean queue length of a general $M/G/c$ can be obtained from the exponential case by Kingsman's formula. Due to the low coefficient of variation of the fitted lognormal and the high accuracy of the exponential call packing system itself, we did not look further into the possibilities that Kingsman's formula offers in this research. Further research could be done on how Kingsman's formula applied to the adapted finite server system performs as an approximation or a bound for call packing systems with general service times. Taking this idea even one step further, one could try to come up with a formula like Kingsman's formula to approximate the mean queue length in a call packing system with general service times based on the result for the exponential call packing system. To do this, one could first evaluate how Kingsman's formula applied to the call packing system performs and then adapt it, possibly by incorporating κ, based on the results.

In the Adapted Service Time Approach, λ_{VVT} was approximated by λ. The influence and validity of the approximation would be interesting for further research. Firstly, research could be done into the difference between the arrival rate of requests for the VVT (λ) and the true arrival rate to the VVT (λ_{VVT}) in real life, and how that can be found back in the data. We would like to know if there are factors that have a significant influence on the difference in arrival rates. For example, the care type, via the probability of dying, or the availability of other care in the region, via the probability of transfer. Secondly, more research should be done on how a better approximation for λ_{VVT} could be incorporated into the Adapted Service Time Approach, since λ_{VVT} and, for example, the probability of overflow influence each other. Thirdly, it would be important to know how sensitive the Adapted Service Time Approach, and call packing system in general, is to the value of λ_{VVT} compared to λ.

To make the call packing system, and its interpretation, more suitable for our application, we introduced a fraction κ which multiplied by the service rate in the VVT μ, replaced the rate of the overflow service γ. This faction κ represents how much the time spent in overflow contributes to a reduction in service time in the VVT, which can differ per type of VVT care. However, the contribution of the time in overflow on the residual service times might differ over the service period. After an operation, there might be little difference between spending the first day in overflow or in a VVT organisation as time is the most important factor for recovery. This would indicate a high value of κ. However, such a high value would be inappropriate if the patient would be in overflow for a long time, since later the presence of specialised care will be the most important factor for the speed of recovery. In this example, the contribution of the time in overflow decreases per day. Instead of the stationary parameter κ, a (decreasing) function $\kappa(t)$ of the time in overflow could be more accurate. Please note that this shortcoming is not caused by the introduction of κ, but was present in the call packing system of Van Dijk and Schilstra (2022) as well due to the rate of the overflow service γ being a fixed value. An interesting topic for future research would thus be a call packing system with a service rate in overflow that depends on the time in overflow.

The inclusion of external arrivals in this research was done based on limiting assumptions. External arrivals are assumed to follow the same arrival distribution and service distribution as arrivals originating from the hospital. The choice for such easy, but also limited, implementation of external arrivals was made based on the fact that this research is applied to the VVT type GRZ and the availability of data. Since GRZ has very few external arrivals, the influence of these external arrivals was expected to be small and thus the incorporation of external arrivals had a low priority. A simple implementation outweighed a very accurate one. Moreover, an implementation that allowed the external arrivals to have a distinct arrival or service distribution would not necessarily increase the accuracy of the model in the numerical analysis due to the necessary data on external VVT requests not being available to us. A more general inclusion of external arrival would, therefore, have mostly theoretical value and limited added practical value to this research. Future research could focus on how external arrivals could be implemented for more general cases. Before this, it would be good to analyse if there is a significant difference in arrival distribution or service distribution of external arrivals and internal arrivals. The reader should be aware that for a different arrival rate of the external arrivals, the convenient form in the optimization model of charging overflow cost for only a fraction ξ of the mean queue length will not apply anymore. The influence of a distinct arrival process could be studied via the simulation as well.

At first, the time dependent analysis was only performed for a $M(t)/G/\infty$ system. Since we proved that the infinite server system provides a general lower bound on the call packing system, and is exact for $\kappa = 1$, this time dependent analysis already gave an impression for the time dependent behaviour of the true system. Later a Modified Offered Load approximation based on the call packing formula was studied as well. Since the formulas for the call packing system only hold for exponentially distributed service times, it would be interesting to study the accuracy of this MOL approximation for a general time dependent call packing system in more detail. To the best of the authors' knowledge, no literature on the MOL approximation on the call packing system has been published. As stated in Section 3.1.4, the lognormality of the service distribution is used for determining the time dependent offered load, while the formulas (20) and (18) for the performance measures in the call packing system only hold for exponential service distributions. The choice was made to use the lognormal service distribution for $m(t)$ so that the time dependent (infinite server) offered load was as accurate as possible. We reasoned that a time dependent offered load that better reflected reality, could lead to better results. Further research could look into the difference in accuracy of the MOL approximation between using an exponential distribution when determining $m(t)$ or the true nonexponential distribution. In general, more research should be done on the MOL approximation of the call packing system.

Another approach for a time dependent analysis could arise from using the $M/G/\infty$ with adapted service time to approximate a call packing system and making it time dependent. The challenge for this approach is that the method of the fixed point adapted service should be extended to the time dependent scenario. Whereas adapting the mean service time works well for the stationary case since the system characteristics only depend on the service time through the mean, the time dependent offered load depends on the complete distribution of the service time. This means that assumptions about the influence of adapting the service time on the service distribution should be made and validated.

6.1.2 Optimization model

The optimization phase is based on pre-calculated performance measures. For the time dependent capacity, this entails that the chosen capacity on one day does not influence the value of the performance measures on the next days. Independent of whether a very low or high capacity is selected on day t, the value of the performance measures on day $t+1$ are the same. In reality, the low capacity will lead to a build-up of queue, that is, a higher number of patients in overflow, for which a higher capacity on day $t+1$ may be better. Since both the time dependent infinite server system and the Modified Offered Load approximation of the call packing system are based on the time dependent offered load of an infinite server system, the possible underestimation is linked to the approach of calculating the performance measures. In addition to possibly underestimating the optimal capacity, the independence between days might unintentionally 'reward' switching capacity more or more extremely by remitting accumulated overflow (unused beds) when increasing (decreasing) capacity. By incorporating a dependence between the choice of capacity on one day and the value of the performance measures on the next day, these consequences of an infinite server approach might be tackled. Our idea is to adapt the offered load at day $t+1$ based on the choice of capacity on day t. Suppose the capacity on day t is c, to which a mean number of patients in overflow of $[E(L_{\text{q}}(t))](c)$ corresponds. The offered load that is used to calculate the performance measures on day $t+1$ for a capacity $\tilde{c}$, will be increased by the difference between the overflow if capacity c would have been kept and if there is switched to capacity $\tilde{c}$, i.e. increased by $[E(L_{\text{q}}(t+1))](c) - [E(L_{\text{q}}(t+1))](\tilde{c})$. Note that this expression is positive when $\tilde{c} > c$ and negative when $c > \tilde{c}$, in which case the adapted offered load is lower than the offered load. The implementation and evaluation of this idea could be a topic for further research. Interestingly, this idea to adapt the offered load based on the queue shows resemblance to the Stationary Backlog-Carryover approximation (SBC) of Stolletz (2008, 2011) in which the blocked customers of a period are carried over to the next period through an artificial arrival rate. The applicability of the SBC approximation to the optimization phase of this research could be studied.

Although most types of VVT care receive money per day a patient stays in their organisation, the GRZ is the only VVT type for which the amount of received money depends mostly on the type of care request, as explained in Section 2.4. This only became clear to us during the meeting with the financial director of a VVT organisation [6], which took place rather late in this research. The optimization model is based on the fixed payment per patient per day. Further research is necessary to incorporate the financing system of the GRZ.

The introduction of a Mixed Integer Program for determining the optimal capacity enabled us to incorporate several constraints on the time dependent behaviour of the capacity and shared capacity between patient types. The list of constraints implemented in this research is of course not exhaustive. The MIP allows for easy implementation of additional constraints.

6.2 Data

The biggest discussion point regarding parameter estimation from the data is the lack of clarity on how the current overflow situation affected the data. First of all, the difference between the *Initial discharge date* and the *Realized discharge date* is defined as the number of overflow days. Besides *Initial discharge date*, the POINT file contains the entry *First Initial Discharge Date* that contains the initial guess for the discharge date. This initial guess can be changed to another date, the *Initial discharge date*, if the patient is not ready to be discharged at the estimated date. A reason for such delay could be complications in the recovery process of the patient. However, the reason for the change in date is not known and might as well be influenced by the unavailability of a bed in the VVT. In case of the latter, the limited capacity of the VVT influenced the *Initial discharge date* as a result of which the number of overflow days is underestimated. Moreover, the analysis of the arrival process is influenced as well since the *Initial discharge date* is also used as the arrival moment of the patient. Secondly, the limited capacity of other VVT organisations might influence the observed length of stay. For example, some patients go after their stay in GRZ to WLZ. Thus, the observed time that such a patient spent in the GRZ organisation consists of both their service time as well as their overflow time for the WLZ. Lastly, it is unknown how the current length of stay obtained from the data is already influenced by the time the patient has spent in overflow. Based on the idea that the time in overflow reduces the residual service time when entering the VVT, the length of stay taken from the data is officially the residual service time.

However, since it was not possible to match the two data sets to study the influence of the time in overflow, this is the best approximation for the service time that we could obtain. The underestimation of the service time leads to an underestimation of the optimal capacity. More research should be done into the influence of the time in overflow on the service time. In this way, a more accurate service time estimation can be obtained as well as an estimate for κ.

For the arrival process, it should be taken into account if a patient already had a bed in the VVT. Especially in WLZ care, a patient might need to go to the hospital after which they return to their 'own' WLZ bed. We should be able to recognize these patients in the data so that they are not counted double.

It is important to note that the numerical analysis in this research is performed based on historical data, while the number of vulnerable elderly is expected to grow [4]. The capacity obtained based on historical data is thus an underestimation of the necessary capacity.

Since the data mainly concerns 2021 and 2022, it is important to mention that, especially in 2021, the coronavirus Covid-19 was still very much present in Netherlands [5]. It is unknown how much the data is influenced by Covid-19.

6.3 Numerical analysis

In the numerical analysis, the influence of several parameters was studied. However, the sensitivity analysis with respect to one parameter was mostly executed for only one combination of values of the other parameters. More specifically, the influence of the cost parameters was only studied for $\kappa = 0.3$. A more extensive analysis could be done to study if κ affects the influence of the cost parameters. Such an analysis is advised because the value of κ is unknown in this research and, with more general applications in mind, it is the most characteristic parameter of the call packing system.

Earlier in this section we mentioned that the time dependent analysis was first performed for only the infinite server system, and that the Modified Offered Load approximation based on the call packing formula was added only later. Due to this, the study on the influence of the cost parameters on the time dependent behaviour was based on the infinite server system. Since the optimal time dependent capacity in the MOL call packing system is just shifted compared to the one of the infinite server system, the results obtained on the infinite server system are expected to generalize rather well. Nevertheless, the influence of the cost parameters on the time dependent behaviour of the capacity should be studied with the MOL call packing system as well, especially since the optimal capacity vector did look different for $C_\mathrm{i} = C_\mathrm{d} = 0$ in Figure 15.

The simulation introduced the concept of overbeds, i.e. more patients in service than the selected capacity level would indicate. At the start of Section 5.3, we noted the influence of the cost of an overbed C_c is limited based on a preliminary study with the analytic optimal vector, [29,29,28,29,29,29,29]. The value of C_c was therefore fixed at 300 for the remainder of the analytic analysis. In the neighbourhood search, capacity vectors with considerably more fluctuations than the analytic optimal vector were considered. By accident, the results were first obtained for $C_c = 30$ instead of $C_c = 300$. We noted that some vectors that performed well for $C_c = 30$ did not perform well for $C_c = 300$. Based on these observations, a sensitivity analysis with respect to C_c is still advised.

The neighbourhood search performed in this research can be considered brute-forced. Since a cyclic capacity of seven days was considered and the difference between the lowest and highest daily capacity was small, it was feasible to consider a large amount of vectors in the neighbourhood. However, we can not guarantee that the optimal capacity vector was found. For example, the best-performing capacity vectors had a capacity of 30 on Monday, but 30 was also the highest capacity level considered for that day. A more elaborate neighbourhood search would thus be necessary to guarantee the optimal solution. A more advanced neighbourhood search based on sophisticated heuristics would be advised for such case. Moreover, since the confidence intervals of the best-performing capacity vector overlap, more simulation runs would be necessary to exclude more vectors from being possibly optimal. Nevertheless, the current neighbourhood search sufficed for this research to give an impression of the near-optimal time dependent behaviour under the used parameter settings.

We could not explain all the observations made in the numerical analysis. For example, why capacity vectors with an increased capacity of 30 on Monday perform well. More research would be necessary to understand such outcomes. Moreover, an attentive reader might have noticed that no numerical results were provided for the case in which beds can be relabeled from one VVT type to another. We observed that the possibility of relabeling beds from one type to another for lower costs is not used, even when there were no costs for relabeling beds. The Mixed Integer Program simply outputted an optimal capacity for GRZ stationary at 27 and for ELV stationary at 8. Since the optimal capacity for GRZ is not stationary for the base parameters when analysed on its own, there is likely an implementation error in the additional constraints.

6.4 Application related

The data showed a significant difference in arrival rates on weekdays and weekends. To benefit from the predictable change in offered load throughout the week, the capacity was allowed to fluctuate over the week. The question of what to do when all beds are occupied on a day on which a decrease in capacity was planned, is normally dealt with on an operational level and not a tactical level. However, on a tactical level, one could question the reasonability of daily varying capacity when the mean length of stay is almost 40 days. Combined with the observation that the optimal capacity of 29 or 30 is lower than the length of stay, overbeds could occur frequently in practice. This may result in the VVT organisation working at a capacity close to the maximum capacity throughout the week. If overbeds are observed more often than desired, change in capacity could be discouraged by increasing the cost for changing capacity in the analytic analysis or the cost for overbeds in the simulation. We indeed observed that already for relatively low costs of changing capacity a stationary capacity was optimal. As advised earlier in this section, the influence of the cost of overbeds should still be studied.

In the numerical analysis, we studied the optimal capacity given a certain time dependent arrival rate. It would be interesting to study if a certain capacity performs better when the number of arrivals on certain days is slightly different. Future research could study if small tweaks in the arrival rate function can lead to substantially lower costs of the optimal capacity. Especially for GRZ, it might be possible to realize a small change in arrival pattern when the financial benefits are substantial. Most arrivals to the GRZ namely originate from the hospital where the patients received surgery. The scheduling of surgeries is thus expected to influence the arrival rate to the VVT. If substantial improvements can be shown, the hospital has a direct incentive to adapt the scheduling of the surgeries since they benefit from the decrease in costs for overflow.

In Section 2.3, we noted that Zychlinski et al. (2020) deals with the optimal capacity in geriatric facilities to prevent bed blocking in hospitals as well. Several similarities and differences between our research and the research of Zychlinski et al. were already mentioned in that section. It would be interesting to perform a numerical analysis with our models based on the data that Zychlinski et al. used and compare the results. Based on these results and the difference in what the models can account for, recommendations could be made on which model could best be used when. Perhaps certain aspects of the two studies could be combined to obtain a more sophisticated model.

The numerical analysis showed a clear difference in cost between the situation in which the VVT would decide on their own capacity and the situation from a system perspective. Based on this, it would make sense that the VVT receives money to increase their capacity to the systems' optimal capacity. It is not within the scope of this research to advise on who is paying for this and on what terms. The financing of the health care system is quite complex and the authors have currently too little knowledge or expertise to judge how much each involved party would profit from the optimal capacity. Clearly, the hospital benefits from the reduction in overflow days. However, the health insurance companies are expected to receive less claims for such overflow days, and thus save money. Moreover, there is likely a time difference between when the money is necessary to increase the capacity and when the financial benefits from the increased capacity are available. So, the roles of all involved parties and their financial incentives should be mapped out. At the same time, more information should be gathered on the costs of overflow beds and VVT beds for each party. In this way, more clarity is obtained about the values of the parameters for each involved party so that the optimal situation for different parties can be examined.

From a mathematical perspective, a game theoretic analysis could be considered to give better advice on the division of the increased VVT cost and the profit. For example, Kverndokk and Melberg (2021) studies how the introduction of a fee to reduce bed-blocking in hospitals influences the strategic decisions of the hospitals and the long-term care providers by using a Stackelberg game.

Due to the limited availability of appropriate data and the limited time of this research, the numerical analysis is only performed for one specific VVT organisation and type. It would give more information about the total capacity available per type in the region to perform the numerical analysis on all relevant VVT organisations. Firstly, pooling the arrivals and capacity of all VVT organisations of each type shows if the total available capacity would suffice. By also performing, per type, the analysis on each distinct organisation using the fraction of patients that currently go to that organisation, it can be studied for which specific organisation the current capacity would be insufficient. Based on these results, either the division of patients to the VVT organisations could be adapted, or the capacity of those specific organisations could be increased in case the total current capacity was shown to be insufficient. Thus, such analyses would help to locate demand and capacity best. Considering the enormous increase in the shortage of skilled nurses predicted [4], research on the best possible use of available personnel is key.

In this research, we wanted to reduce the bed-blocking days by determining the optimal capacity. The question remains what part of the bed-blocking days could indeed be reduced by the optimal VVT capacity, and what part is unrelated to the availability of beds in the VVT. For example, Armony et al. (2015) found that patients often have to wait even when there are beds available. This implies that there are more causes for delay in a transfer process than only bed unavailability. The financial director of a VVT organisation [6] indicated as well that the current overflow situation is caused by both limited VVT capacity as well as by logistical and administrative complications. Gaughan et al. (2015) devoted a complete paper to the extent to which the bed blocking in the hospital can be reduced by more capacity in nursing homes. So, adequate capacity of the VVT organisations on its own will not solve the complete problem of bed blocking. In addition, the transfer process needs to be improved, for which the introduction of Zorgschakel Twente is a step in the right direction [6].

During the meeting with the financial director of a VVT organisation [6] the idea of a transfer department (*schakelafdeling*) was discussed. If a patient is ready to leave the hospital, but there is no place in the VVT, the patient goes to the transfer department. The transfer department is less expensive than the hospital and the hospital bed comes free for a new patient. Introducing capacity in the transfer department is not the same as increasing the capacity of the VVT as studied in this research. VVT capacity is specific for a certain type, while the transfer department is not type-specific. So instead of increasing the VVT capacity of multiple types in the long term, which can lead to a high number of unused beds, (unrelated) accidental increase in demand for a certain VVT type can be accommodated on a higher level by the transfer department. A trade-off should be made between the capacity assigned to this transfer department and the normal VVT types. Since the transfer department is available to all VVT types, the introduction of shared capacity in the Mixed Integer Program could be used to help decide on this trade-off. However, depending on the level of care that is provided in this transfer department and the VVT type, the influence of the time that a patient spends in this department on their residual service time differs. If time in the transfer department does not contribute as much to the service time as time in the normal VVT, a complicated relation between the time in the transfer department, the residual service time and the capacity in VVT takes place. The phase in which the performance measures are determined and the optimization phase are not distinct anymore. Other research methods would be necessary to evaluate the idea of a transfer department. For this, one may start by looking at studies that look into the Step Down Unit, which is a department in the hospital between the Intensive Care and the general wards and which existence and optimal capacity are much discussed.

7 Conclusion

This research was motivated by the question of what the optimal capacity of the *Verpleeg-, Verzorgingshuizen en Thuiszorg (VVT)* organisations in region Enschede would be. This question arose from the desire to reduce the bed blocking days in the hospital, which costs two of the main hospitals in region Twente already almost nine thousand euros per day [45].

Determining the capacity was divided into two phases. In the first phase, the appropriate queueing models were analysed to determine the values of the necessary performance measures. The obtained mean number of patients in overflow and mean number of unused beds were the input to the second phase: the optimization phase. By separating the calculation of the performance measures and the optimization into two phases, the process of determining the time dependent capacity was simplified. The values of the performance measures on one day did not depend on the choice of capacity on other days.

In the first phase, discussed in Section 3.1, the system was modelled as an overflow system with immediate call packing, in which the fraction κ represented how much the time spent in overflow contributed to a reduction in service time in the VVT. The boundary cases corresponding to $\kappa = 1$ and $\kappa = 0$ were the $M/G/\infty$ system and $M/M/c$ system, respectively. The mean queue length in these systems was shown to provide a lower and upper bound, respectively, of that in the corresponding call packing system. By adapting the service time in these systems appropriately, we conjectured that stricter bounds on the call packing system are obtained. Since the infinite server system is insensitive to the service distribution, this Adapted Service Time Approach yielded an insensitive lower bound to the mean queue length in the call packing system, for which a convenient form exists only for exponentially distributed service times. In addition to the stationary analysis, a time dependent analysis of the models was discussed. By replacing the stationary offered load ρ in the formula for the performance measures by the time dependent offered load $m(t)$, a time dependent infinite server approach and a Modified Offered Load approximation for the call packing system were obtained. The latter could not be found in literature before. Thus, Section 3.1 contained important theoretical contributions to the approximation of the call packing system both in the stationary and the time dependent case.

Section 3.2 discussed the optimization phase. The financial incentives of both the hospital and the VVT were taken into account by having three components in the objective function: overflow costs, costs for unused beds and costs for used beds. In the case of a time dependent capacity, costs for changing capacity were included as well. A Mixed Integer Program was introduced through which the time dependent behaviour of the capacity could be constrained. The ability to choose the values of the cost parameters and the possibility to include constraints have an added value to the frequently used square root staffing rule. Moreover, the Mixed Integer Program allowed for interaction between the capacity of several VVT types by incorporating shared capacity and the relabeling of beds.

The derived models were used in a numerical analysis in Section 5 and the performances of the different queueing models were compared. The system characteristics and parameters used for this analysis were based on the data of a specific VVT type of one VVT organisation, as discussed in Section 4.

The stationary analysis in Section 5.1 showed that the proven and conjectured ordering of the mean queue length in the five queueing systems considered in this research almost holds for the objective value corresponding to each capacity as well. The adapted $M/G/\infty$ system was the only exception, for which we observed that the mean number of beds is low compared to that in the unadapted $M/G/\infty$ system and that in the call packing system. The capacity decision was observed to be influenced by the fraction κ and the ratio between the cost of overflow C_{o} and the cost of unused beds C_{u}. Comparing the stationary analytic results to the simulation results for a stationary capacity showed that the exponential call packing system best approximated the true system.

The time dependent analysis of the infinite server system and the Modified Offered Load approximation of the call packing system allowed for the consideration of a time dependent capacity in Section 5.2. The behaviour of the time dependent capacity was similar to that of the time dependent offered load. Since the time dependent offered load was shown to be shifted compared to the arrival rate function, peaks and dips in the capacity are shifted compared to those in the arrival rate as well. The biggest influence on the time dependent behaviour of the capacity was the cost of changing capacity. Even for a relatively small cost of changing capacity compared to the cost of unused beds or overflow, the optimal capacity did not change over time. The main reason for this was that a change in capacity was only useful for one or a few days due to the rapidly varying offered load and its short cycle time. The real cost of the analytic optimal solution was judged using a simulation and compared to that of neighbouring vectors in Section 5.3. The best-performing vectors from the simulation study had a base capacity of 29, corresponding to the stationary result for the exponential call packing system, around which they fluctuated to adapt for predictable changes in the time dependent offered load. The best-performing vectors in the simulation study outperformed the analytic optimal time dependent solution, for which we already reasoned that it would underestimate the capacity on some days. The result from the stationary analysis with the call packing system performed rather well in the simulation since no improvement in simulated cost was obtained by the result from the analytic time dependent analysis.

Due to conflicting monetary incentives, as outlined in Section 2.4, and the fact that the expensive overflow of the VVT takes place in the hospital, the VVT was expected to select a lower capacity than what would be optimal from a system perspective. The situation in which the VVT would decide on their own was studied by setting the cost for overflow C_o to zero. The results in Sections 5.1.3 and 5.4 confirmed that the optimal capacity when the VVT would decide on its own is lower than the optimal capacity from a system perspective. The decrease in total cost that is expected when the capacity at the VVT equals the optimal capacity from the systems perspective instead of the VVT's optimal, is significantly larger than the increase in costs for the VVT. These results suggest that it would be reasonable to compensate the VVT for increasing their capacity.

Several topics for future research were described in Section 6. Most importantly, the numerical study should be performed for other VVT types and other organisations since these are likely to have different system characteristics. The goal of this extra research is two-fold. On the one hand, it will become clear how well the theoretical results and sensitivity analyses hold in general. On the other hand, a more complete picture of the current status and the optimal situation in region Enschede is obtained and a better allocation of demand and capacity could be researched. Outside of this research, more insight into the involved parties and their financial contributions must be obtained.

References

[1] ELV at *de Posten*. URL https://www.deposten.nl/eerstelijnsverblijf#:~:text=Het%20verschil%20is%20vooral%20dat,en%20rust%20om%20te%20herstellen.

[2] Prestatiebekostiging ziekenhuizen. URL https://www.rijksoverheid.nl/onderwerpen/kwaliteit-van-de-zorg/prestatiebekostiging-ziekenhuizen.

[3] Beleidsregel prestatiebeschrijvingen en tarieven verkeerde bed Wlz 2024 - BR/REG-24119. URL https://puc.overheid.nl/nza/doc/PUC_743477_22/1/.

[4] ActiZ infographic Ontwikkeling Arbeidsmarkt VVT. URL https://www.actiz.nl/sites/default/files/2021-02/ActiZ_Infographic-Ontwikkeling-Arbeidsmarkt-VVT.pdf.

[5] Coronavirus tijdlijn. URL https://www.rijksoverheid.nl/onderwerpen/coronavirus-tijdlijn.

[6] Discussed in a meeting with the financial director of a VVT organisation.

[7] Website of the Dutch National Health Care Institute. URL https://www.zorginstituutnederland.nl/.

[8] Actiz infographic Geriatrische revalidatiezorg, 2018. URL https://leden.actiz.nl/cms/streambin.aspx?documentid=24421.

[9] Actiz infographic Eerstelijnsverblijf (ELV), 2019. URL https://www.google.nl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjOw_icp9uFAxV2gf0HHQKcACAQFnoECBgQAQ&url=https%3A%2F%2Fleden.actiz.nl%2Fcms%2Fstreambin.aspx%3Fdocumentid%3D24420&usg=AOvVaw3RUmeqNUL5KRHOROwSzv1A&opi=89978449.

[10] I. Adan and J. Resing. *Queueing Systems*. Eindhoven University of Technology, Mar. 2015. URL https://www.win.tue.nl/~iadan/queueing.pdf.

[11] M. Armony, S. Israelit, A. Mandelbaum, Y. N. Marmor, Y. Tseytlin, and G. B. Yom-Tov. On Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective. *Stochastic Systems*, 5(1):146–194, June 2015. ISSN 1946-5238, 1946-5238. doi: 10.1287/14-SSY153. URL https://pubsonline.informs.org/doi/10.1287/14-SSY153.

[12] A. Assad, M. Fu, and J.-S. Yoo. A lower bounding result for the optimal policy in an adaptive staffing problem. 1997.

[13] A.Ya.Khinchine. Mathematical methods in the theory of queueing. *Trudy Mat Inst. Steklov 49*, 1955. (in Russian).

[14] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, Closed, and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22(2):248–260, Apr. 1975. ISSN 0004-5411. doi: 10.1145/321879.321887. URL https://dl.acm.org/doi/10.1145/321879.321887.

[15] R. Bekker and A. M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, July 2010. ISSN 1572-9338. doi: 10.1007/s10479-009-0570-z. URL https://doi.org/10.1007/s10479-009-0570-z.

[16] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization*. Athena Scientific series in optimization and neural computation. Athena Scientific, Belmont, Mass, 1997. ISBN 978-1-886529-19-9.

[17] S. Borst, A. Mandelbaum, and M. I. Reiman. Dimensioning Large Call Centers. *Operations Research*, 52 (1):17–34, 2004. ISSN 0030-364X. URL https://www.jstor.org/stable/30036558.

[18] R. J. Boucherie and N. M. Van Dijk. Spatial Birth-Death Processes with Multiple Changes and Applications to Batch Service Networks and Clustering Processes. *Advances in Applied Probability*, 22(2): 433–455, 1990. ISSN 0001-8678. doi: 10.2307/1427544. URL https://www.jstor.org/stable/1427544.

[19] N. C. Buyukkaramikli, J. W. M. Bertrand, and H. P. G. van Ooijen. Flexible hiring in a make to order system with parallel processing units. *Annals of Operations Research*, 209(1):159–178, Oct. 2013. ISSN 1572-9338. doi: 10.1007/s10479-011-0958-4. URL `https://doi.org/10.1007/s10479-011-0958-4`.

[20] K. Chandy. Analysis and solutions for general queueing networks. Oct. 2002.

[21] Y. Dallery and Y. Frein. On Decomposition Methods for Tandem Queueing Networks with Blocking. *Operations Research*, 41(2):386–399, 1993. ISSN 0030-364X. URL `https://www.jstor.org/stable/171785`. Publisher: INFORMS.

[22] A. M. de Bruin, A. C. van Rossum, M. C. Visser, and G. M. Koole. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137, June 2007. ISSN 1572-9389. doi: 10.1007/s10729-007-9009-8. URL `https://doi.org/10.1007/s10729-007-9009-8`.

[23] S. Eick, W. Massey, and W. Whitt. Mt/G/\infty Queues with Sinusoidal Arrival Rates. *Management Science*, 39:241–252, Aug. 2000.

[24] S. G. Eick, W. A. Massey, and W. Whitt. The Physics of the Mt/G/ Queue. *Operations Research*, Aug. 1993. doi: 10.1287/opre.41.4.731. URL `https://pubsonline.informs.org/doi/abs/10.1287/opre.41.4.731`. Publisher: INFORMS.

[25] Z. Feldman, A. Mandelbaum, W. Massey, and W. Whitt. Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, 54:324–338, Feb. 2008. doi: 10.1287/mnsc.1070.0821.

[26] S. Gallivan, M. Utley, T. Treasure, and O. Valencia. Booked inpatient admissions and hospital capacity: mathematical modelling study. *BMJ*, 324(7332):280–282, Feb. 2002. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.324.7332.280. URL `https://www.bmj.com/content/324/7332/280`. Publisher: British Medical Journal Publishing Group Section: Information in practice.

[27] J. Gaughan, H. Gravelle, and L. Siciliani. Testing the Bed-Blocking Hypothesis: Does Nursing and Care Home Supply Reduce Delayed Hospital Discharges? *Health economics*, 24 Suppl 1:32–44, Mar. 2015. doi: 10.1002/hec.3150.

[28] W. J. Gordon and G. F. Newell. Closed queuing systems with exponential servers. *Operations Research*, 15(2):254–265, 1967. ISSN 0030364X, 15265463. URL `http://www.jstor.org/stable/168557`.

[29] A. Granas and J. Dugundji. *Elementary Fixed Point Theorems*, pages 9–84. Springer New York, New York, NY, 2003. ISBN 978-0-387-21593-8. doi: 10.1007/978-0-387-21593-8_2. URL `https://doi.org/10.1007/978-0-387-21593-8_2`.

[30] L. Green. *Queueing Analysis in Healthcare*, pages 281–307. Springer US, Boston, MA, 2006. ISBN 978-0-387-33636-7. doi: 10.1007/978-0-387-33636-7_10. URL `https://doi.org/10.1007/978-0-387-33636-7_10`.

[31] L. V. Green, P. J. Kolesar, and J. Soares. Improving the SIPP Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research*, 49(4):549–564, 2001. ISSN 0030-364X. URL `https://www.jstor.org/stable/3088586`. Publisher: INFORMS.

[32] L. V. Green, P. J. Kolesar, and W. Whitt. Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management*, 16(1):13–39, Jan. 2007. ISSN 1059-1478, 1937-5956. doi: 10.1111/j.1937-5956.2007.tb00164.x. URL `https://onlinelibrary.wiley.com/doi/10.1111/j.1937-5956.2007.tb00164.x`.

[33] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29:567–588, June 1981. doi: 10.1287/opre.29.3.567.

[34] R. Hall. *Queueing Methods—For Services and Manufacturing.* Oct. 1990. ISBN 978-0-13-744756-5.

[35] F. S. Hillier and R. W. Boling. Finite Queues in Series with Exponential or Erlang Service Times-A Numerical Approach. *Operations Research*, 15(2):286–303, 1967. ISSN 0030-364X. Publisher: INFORMS.

[36] A. Hordijk and A. Ridder. Stochastic Inequalities for an Overflow Model. *Journal of Applied Probability*, 24(3):696–708, 1987. ISSN 0021-9002. doi: 10.2307/3214100. URL `https://www.jstor.org/stable/3214100`. Publisher: Applied Probability Trust.

[37] G. C. Hunt. Sequential Arrays of Waiting Lines. *Operations Research*, 4(6):674–683, 1956. ISSN 0030-364X. Publisher: INFORMS.

[38] J. R. Jackson. Jobshop-like Queueing Systems. *Management Science*, 10(1):131–142, 1963. ISSN 0025-1909. URL `https://www.jstor.org/stable/2627213`. Publisher: INFORMS.

[39] D. L. Jagerman. Nonstationary Blocking in Telephone Traffic. *Bell System Technical Journal*, 54 (3):625–661, Mar. 1975. ISSN 00058580. doi: 10.1002/j.1538-7305.1975.tb02858.x. URL `https://ieeexplore.ieee.org/document/6774097`.

[40] O. Jennings, W. Massey, and C. Mccalla. Optimal profit for leased lines services. *Proceedings of the 15th International Teletraffic Congress - ITC 15*, pages 803–814, Aug. 1997.

[41] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt. Server Staffing to Meet Time-Varying Demand. *Management Science*, 42(10):1383–1394, 1996. ISSN 0025-1909. URL `https://www.jstor.org/stable/2634372`. Publisher: INFORMS.

[42] O. Jouini and Y. Dallery. Monotonicity properties for multiserver queues with reneging and finite waiting lines. *Probability in The Engineering and Informational Sciences*, 21:335–360, July 2007. doi: 10.1017/S0269964807000010.

[43] F. Kamoun and L. Kleinrock. Analysis of Shared Finite Storage in a Computer Network Node Environment Under General Traffic Conditions. *Communications, IEEE Transactions on*, 28:992–1003, Aug. 1980. doi: 10.1109/TCOM.1980.1094756.

[44] F. P. Kelly. Networks of Queues. *Advances in Applied Probability*, 8(2):416–432, 1976. ISSN 0001-8678. doi: 10.2307/1425912. URL `https://www.jstor.org/stable/1425912`. Publisher: Applied Probability Trust.

[45] J. Kodde. MST en ZGT raken hun patiënten niet kwijt en lopen daardoor miljoenen aan omzet mis. *Tubantia*, Jan. 2024. URL `https://www.tubantia.nl/enschede/mst-en-zgt-raken-hun-patienten-niet-kwijt-en-lopen-daardoor-miljoenen-aan-omzet-mis~a0e85487/`.

[46] N. Koizumi, E. Kuno, and T. E. Smith. Modeling Patient Flows Using a Queuing Network with Blocking. *Health Care Management Science*, 8(1):49–60, Feb. 2005. ISSN 1572-9389. doi: 10.1007/s10729-005-5216-3. URL `https://doi.org/10.1007/s10729-005-5216-3`.

[47] S. Kverndokk and H. O. Melberg. Using fees to reduce bed-blocking: a game between hospitals and long-term care providers. *The European Journal of Health Economics*, 22(6):931–949, Aug. 2021. ISSN 1618-7601. doi: 10.1007/s10198-021-01299-9. URL `https://doi.org/10.1007/s10198-021-01299-9`.

[48] S. S. Lam. Queuing networks with population size constraints. *IBM Journal of Research and Development*, 21(4):370–378, July 1977. doi: 10.1147/rd.214.0370.

[49] A. Li, W. Whitt, and J. Zhao. Staffing to stabilize blocking in loss model with time-varying arrival rates. *Probability in the Engineering and Informational Sciences*, 30(2):185–211, Apr. 2016. ISSN 0269-9648, 1469-8951. doi: 10.1017/S0269964815000340. URL `https://www.cambridge.org/core/journals/probability-in-the-engineering-and-informational-sciences/article/staffing-to-stabilize-blocking-in-loss-models-with-timevarying-arrival-rates/B22F61E8B47DE34664494AF87715932D`.

[50] N. Litvak, M. van Rijsbergen, R. J. Boucherie, and M. van Houdenhoven. Managing the overflow of intensive care patients. *European Journal of Operational Research*, 185(3):998–1010, Mar. 2008. ISSN 0377-2217. doi: 10.1016/j.ejor.2006.08.021. URL `https://www.sciencedirect.com/science/article/pii/S0377221706005819`.

[51] W. A. Massey. The Analysis of Queues with Time-Varying Rates for Telecommunication Models. *Telecommunication Systems*, page 173–204, 2002. doi: 10.1023/A:1020990313587. URL `https://doi.org/10.1023/A:1020990313587`.

[52] W. A. Massey and W. Whitt. An Analysis of the Modified Offered-Load Approximation for the Nonstationary Erlang Loss Model. *The Annals of Applied Probability*, 4(4):1145–1160, 1994. ISSN 1050-5164. URL `https://www.jstor.org/stable/2245085`. Publisher: Institute of Mathematical Statistics.

[53] S. I. M. Michael C. Fu and I.-J. Wang. Monotone optimal policies for a transient queueing staffing problem. *Operations Research*, 48(2):327–331, 2000.

[54] Ministerie van Volksgezondheid, Welzijn en Sport. Programma Wonen, Ondersteuning en Zorg voor Ouderen, July 2022.

[55] Nederlandse Zorgautoriteit. Stand van de zorg 2023, Oct. 2023.

[56] W. Oniszczuk. Tandem Models with Blocking in the Computer Subnetworks Performance Analysis. In K. Saeed, J. Pejaś, and R. Mosdorf, editors, *Biometrics, Computer Security Systems and Artificial Intelligence Applications*, pages 259–267, Boston, MA, 2006. Springer US. ISBN 978-0-387-36503-9. doi: 10.1007/978-0-387-36503-9_24.

[57] C. Osorio and M. Bierlaire. An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research*, 196(3):996–1007, Aug. 2009. ISSN 0377-2217. doi: 10.1016/j.ejor.2008.04.035.

[58] C. Palm. Intensity variations in telephone traffic. *Ericsson Technics*, 44:1–189, 1943. (in German).

[59] H. G. Perros and T. Altiok. Approximate analysis of open networks of queues with blocking: Tandem configurations. *IEEE Transactions on Software Engineering*, SE-12(3):450–461, Mar. 1986. ISSN 1939-3520. doi: 10.1109/TSE.1986.6312886. URL `https://ieeexplore.ieee.org/document/6312886`. Conference Name: IEEE Transactions on Software Engineering.

[60] J. T. Rich, N. J. Gail, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang. A practical guide to understanding Kaplan-Meier curves. *Otolaryngology–head and neck surgery: official journal of American Academy of Otolaryngology-Head and Neck Surgery*, 143(3):331–336, Sept. 2010. ISSN 0194-5998. doi: 10.1016/j.otohns.2010.05.007. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/`.

[61] K. W. Ross. *The Reduced Load Approximation for Single-Service Networks*, chapter 5.5, pages 181–187. Springer-Verlag, Dec. 2011. ISBN 978-1447121282.

[62] S. M. Ross. *Introduction to Probability Models.* Academic Press, 2007. ISBN 978-0-12-598062-3.

[63] J. A. Schwarz, G. Selinka, and R. Stolletz. Performance analysis of time-dependent queueing systems: Survey and classification. *Omega*, 63:170–189, Sept. 2016. ISSN 0305-0483. doi: 10.1016/j.omega.2015.10.013. URL `https://www.sciencedirect.com/science/article/pii/S0305048315002170`.

[64] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris. *Fundamentals of queueing theory.* John Wiley and Sons, Jan. 2018. ISBN 9781118943526. doi: 10.1002/9781119453765.

[65] R. Stolletz. Approximation of the non-stationary $m(t)/m(t)/c(t)$-queue using stationary queueing models: The stationary backlog-carryover approach. *European Journal of Operational Research*, 190(2):478—-493, 2008. doi: 10.1016/j.ejor.2007.06.036. URL `https://doi.org/10.1016/j.ejor.2007.06.036`.

[66] R. Stolletz. Analysis of passenger queues at airport terminals. *Research in Transportation Business Management*, 1(1):144–149, 2011. ISSN 2210-5395. doi: https://doi.org/10.1016/j.rtbm.2011.06.012. URL `https://www.sciencedirect.com/science/article/pii/S2210539511000198`.

[67] W. te Meerman. Verpleeghuis minder duur dan het lijkt. *Skipr*, Sept. 2017. URL `https://www.skipr.nl/blog/verpleeghuis-minder-duur-dan-het-lijkt/`.

[68] G. M. Thompson. Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. *Journal of Operations Management*, 11(3):269–287, Sept. 1993. ISSN 0272-6963, 1873-1317. doi: 10.1016/0272-6963(93)90004-9. URL `https://onlinelibrary.wiley.com/doi/10.1016/0272-6963%2893%2990004-9`.

[69] F. Timmer. Opheldering gevraagd over opstopping in Twentse ziekenhuizen: 'Toch meer bedden in verpleeghuizen?'. *Tubantia*, Jan. 2024. URL `https://www.tubantia.nl/enschede/opheldering-gevraagd-over-opstopping-in-twentse-ziekenhuizen-toch-meer-bedden-in-verpleeghuizen~aa6337b6/?cb=af9201e5c0fd065a9754b8410451efee&auth_rd=1`.

[70] S. van Brummelen, W. de Kort, and N. van Dijk. Queue length computation of time-dependent queueing networks and its application to blood collection. *Operations Research for Health Care*, 17:4–15, June 2018. ISSN 22116923. doi: 10.1016/j.orhc.2018.01.006. URL `https://linkinghub.elsevier.com/retrieve/pii/S2211692316301217`.

[71] M. van der Geest. De ouderenzorg is niet lang meer te negeren: de grootste problemen in drie grafieken. *de Volkskrant*, Oct. 2023. URL `https://www.volkskrant.nl/wetenschap/de-ouderenzorg-is-niet-lang-meer-te-negeren-de-grootste-problemen-in-drie-grafieken~b498ad47/?referrer=https://www.google.nl/`.

[72] N. Van Dijk and B. Schilstra. On two product form modifications for finite overflow systems. *Annals of Operations Research*, 310(2):519–549, Mar. 2022. ISSN 0254-5330, 1572-9338. doi: 10.1007/s10479-021-03940-5. URL `https://link.springer.com/10.1007/s10479-021-03940-5`.

[73] N. M. van Dijk. Simple and insensitive bounds for a grading and an overflow model. *Operations Research Letters*, 6(2):73–76, May 1987. ISSN 0167-6377. doi: 10.1016/0167-6377(87)90033-2. URL `https://www.sciencedirect.com/science/article/pii/0167637787900332`.

[74] N. M. Van Dijk and E. Van Der Sluis. Call packing bound for overflow loss systems. *Performance Evaluation*, 66(1):1–20, Jan. 2009. ISSN 01665316. doi: 10.1016/j.peva.2008.06.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0166531608000564`.

[75] N. M. Van Dijk and E. Van Der Sluis. Call packing bound for overflow loss systems. *Performance Evaluation*, 66(1):1–20, Jan. 2009. ISSN 01665316. doi: 10.1016/j.peva.2008.06.003. URL `https://linkinghub.elsevier.com/retrieve/pii/S0166531608000564`.

[76] J.-K. van Ommeren. *Stochastic Simulation.* June 2020.

[77] L. H. van Roijen, S. Peeters, and T. Kanters. Kostenhandleiding voor economische evaluaties in de gezondheidszorg: Methodologie en referentieprijzen. 2024. URL `https://www.zorginstituutnederland.nl/over-ons/publicaties/publicatie/2024/01/16/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg`.

[78] W. v. d. Weij, N. M. v. Dijk, and R. D. v. d. Mei. Product-form results for two-station networks with shared resources. *Performance evaluation*, 69(12):662–683, 2012. ISSN 0166-5316. doi: 10.1016/j.peva.2012.08.002. URL `https://research.utwente.nl/en/publications/product-form-results-for-two-station-networks-with-shared-resourc`. Publisher: Elsevier.

[79] W. Whitt. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)*, 54(5):476–484, Aug. 2007. ISSN 0894-069X, 1520-6750. doi: 10.1002/nav.20243. URL `https://onlinelibrary.wiley.com/doi/10.1002/nav.20243`.

[80] W. Whitt. Approximations for the gi/g/m queue. *Production and Operations Management*, 2(2), 2009.

[81] W. Whitt. The infinite-server queueing model: The center of the many-server queueing universe. *Operations Research Letters*, Feb. 2012. URL `https://www.columbia.edu/~ww2040/8100S12/ISqueue021412.pdf`.

[82] W. Whitt. Queues with Time-Varying Arrival Rates: A Bibliography. Apr. 2017. URL `https://www.columbia.edu/~ww2040/TV_bibliography_091016.pdf`.

[83] P. Whittle. Equilibrium distributions for an open migration process. *Journal of Applied Probability*, 5(3): 567–571, 1968. ISSN 00219002. URL `http://www.jstor.org/stable/3211921`.

[84] W. L. Winston and J. B. Goldberg. *Operations research: applications and algorithms.* Thomson/Brooks/Cole, Belmont, CA, 4th ed edition, 2004. ISBN 978-0-534-38058-8. OCLC: 52815313.

[85] D. Worthington, M. Utley, and D. Suen. Infinite-server queueing models of demand in healthcare: A review of applications and ideas for further work. *Journal of the Operational Research Society*, 71(8): 1145–1160, Aug. 2020. ISSN 0160-5682. doi: 10.1080/01605682.2019.1609878.

[86] N. Zychlinski, A. Mandelbaum, P. Momčilović, and I. Cohen. Bed Blocking in Hospitals Due to Scarce Capacity in Geriatric Institutions—Cost Minimization via Fluid Models. *Manufacturing & Service Operations Management*, 22(2):396–411, Mar. 2020. ISSN 1523-4614. doi: 10.1287/msom.2018.0745. URL `https://pubsonline.informs.org/doi/10.1287/msom.2018.0745`. Publisher: INFORMS.

Appendices

A.1 Proof of 2-periodic convergence of repeated substitution for $M/M/c$ system

In Section 3.1.3, we stated that the sequence of iterates obtained by repeated substitution of $T_2(L) := Q_2(R_2(L))$ has two convergent subsequences, where one sequence converges to L^- and the other to L^+ with $L^- \leq L^* \leq L^-$. The proof for this statement is provided in this section.

First, we note that adapted offered load $R_2(L)$ decreases in the queue length L and that $Q_2(\rho)$ increases in ρ, the adapted mean queue length switches each iteration between high and low values. Let $L^{(n)}$ denote the nth iterate defined iteratively by $L^{(n)} = T_2(L^{(n-1)})$ resulting from repeated substitution of the operator T_2 with initialization $L^{(0)} = 0$. Note that after one iteration $L^{(1)} = T_2(0)$, the mean queue length of the related $M/M/c$ is obtained. Since T_2 is a composition mapping, repeated substitution will also yield a sequence of intermediate adapted offered loads $\rho^{(n)}$. For convenience in the proof below, the offered load which corresponds to the nth mean queue length $L^{(n)}$ is named $\rho^{(n)}$, i.e. $L^{(n)} = Q_2(\rho^{(n)}) = Q_2(R_2(L^{(n-1)}))$. So, the non adapted offered load $\frac{\lambda}{\mu}$ is defined as $\rho^{(1)}$, which yields the queue length $L^{(1)}$.

We will show below that $L^{(0)} \leq L^{(2)} \leq \cdots \leq L^{(2n)} \leq L^* \leq L^{(2n+1)} \leq \cdots \leq L^{(3)} \leq L^{(1)}$. This would mean that $L^{(2n)} \to L^-$ and $L^{(2n+1)} \to L^+$ with $L^- \leq L^* \leq L^+$, due to the monotonocity of the sub sequences.

Theorem 7. $0 = L^{(0)} \leq L^{(2)} \leq \cdots \leq L^{(2n)} \leq L^* \leq L^{(2n+1)} \leq \cdots \leq L^{(3)} \leq L^{(1)}$

Proof. Clearly, $L^{(0)} = 0 \leq L^*$. Since $\rho^{(1)} = \frac{\lambda_{\text{VVT}}}{\mu}$, while the load corresponding to the fixed point ρ^* satisfies $\rho^* = \frac{\lambda_{\text{VVT}}}{\mu} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^* = \rho^{(1)} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^*$, we have $\rho^* \leq \rho^{(1)}$. Since the composite function for the mean queue length $Q_2(\rho)$ is increasing in ρ, $L^* \leq L^{(1)}$. Thus, $L^{(0)} \leq L^* \leq L^{(1)}$.

Left to show is $L^{(2(n+1))} \geq L^{(2n)}$ and $L^{(2(n+1)+1)} \leq L^{(2n+1)}$, which will be done by induction. Similar relations are shown to hold for the related offered loads, after which the desired relations follow by the fact that $Q_2(\rho)$ is increasing in ρ.

Base cases $k = 0$:
$L^{(2)} \geq L^{(0)} = 0$ is trivial by non-negativity of the mean queue length.
$L^{(3)} \leq L^{(1)}$ follows from $\tilde{r}^{(3)} = R_2(L^{(2)}) = \frac{\lambda_{\text{VVT}}}{\mu} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^{(2)} = \tilde{r}^{(1)} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^{(2)} \leq \tilde{r}^{(1)}$.

We now assume that the desired statements hold for some $n = k - 1$.

For $n = k$, it then follows from the induction hypothesis ($L^{(2k+1)} \leq L^{(2(k-1)+1)}$) that $\tilde{r}^{(2(k+1))} = R_2(L^{(2k+1)}) = \frac{\lambda_{\text{VVT}}}{\mu} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^{(2k+1)} \geq \frac{\lambda_{\text{VVT}}}{\mu} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^{(2k-1)} = R_2(L^{(2k-1)}) = \tilde{r}^{(2k)}$. Since $Q_2(\rho)$ is increasing in ρ, $L^{(2(k+1))} \geq L^{(2k)}$.

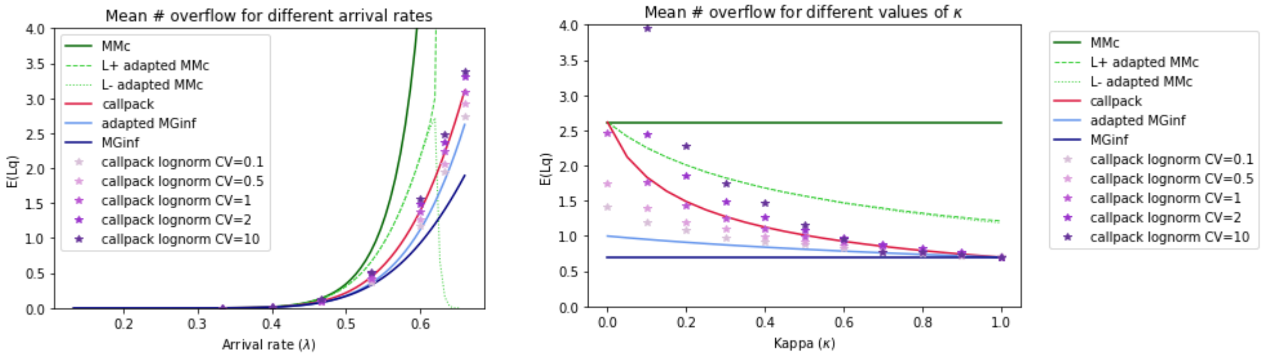
From this statement it follows that $\tilde{r}^{(2(k+1)+1)} = R_2(L^{(2(k+1))}) = \frac{\lambda_{\text{VVT}}}{\mu} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^{(2(k+1))} \leq \frac{\lambda_{\text{VVT}}}{\mu} - \frac{\kappa\lambda_{\text{VVT}}}{\lambda}L^{(2k)} = R_2(L^{(2k)}) = \tilde{r}^{(2k+1)}$. From $Q_2(\rho)$ being increasing in ρ, it follows $L^{(2(k+1)+1)} \geq L^{(2k+1)}$.

□

A.2 Numerical comparison of Adapted Service Time Approach

In Section 3.1.3, we hypothesized that the infinite server system with adapted service time gives a lower bound to mean number of patients in overflow in the corresponding call packing system for every service distribution (Conjecture 1) and that the adapted $M/M/c$ system yields an upper bound in case of an exponential service time distribution (Conjecture 2). The numerical analysis on which these conjectures were based is provided in this section.

Figure 22 contains the mean number of patients in overflow in a system with call packing, calculated with formula (18), in an unadapted $M/M/c$ system (21), in an unadapted $M/G/\infty$ system (23) and the fixed points from the Adapted Service Time Approach as explained in section (3.1.3), for a specific set of parameters. Note that the results for the call packing system and the finite server system only hold for exponentially distributed service times, while the results for the infinite server system hold for general service time distributions. Besides these analytic results, the mean number of patients in overflow in a call packing system is therefore also determined in the case of lognormal service times via a simulation. In line with Van Dijk and Schilstra (2022), the analysis for the lognormal service times is done for the coefficients of variations 0.1, 0.5, 1, 2 and 10. These are marked with stars. In each plot, one parameter is of interest while the rest are kept fixed at $\lambda = 0.58$, $\mu = \frac{1}{39}$, $\kappa = \frac{1}{2}$ and $c = 26$, if not stated otherwise.



(A) Ranging over the arrival rate λ.

(B) Ranging over κ.

FIGURE 22: The mean number of patients in overflow in the different systems over different parameters.

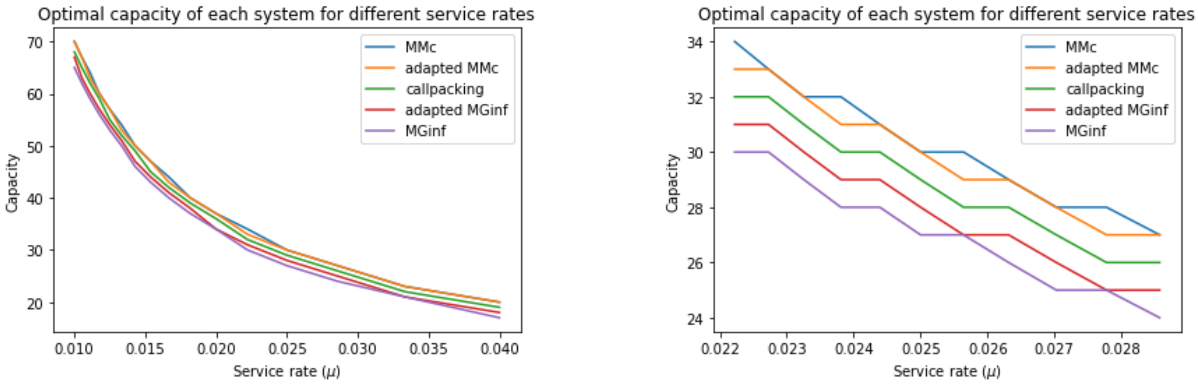
First of all, the numerical results are consistent with the bounds proven in Theorems 1, 2, 5 and 6. That is, the mean number of patients in overflow in the call packing system with exponential service times is between that of the $M/G/\infty$ system and that of $M/M/c$ system, and the Adapted Service Time Approaches are higher, resp. lower than that of their unadapted $M/G/\infty$ system, resp. $M/M/c$ system.

The influence of the offered load is studied in Figure 22a by ranging over the arrival rate λ while μ and c are kept fixed. The Adapted Service Time Approach of the infinite server system yields a stricter lower bound on the mean number of overflow patients in the call packing system than their respective unadapted systems. The lower bound remains valid for call packing systems with lognormal service time distributions, underpinning Conjecture (1). For high arrival rates, the repeated substitution to determine the fixed point for the $M/M/c$ system with adapted service times yielded two subsequences converging to two different values. For these parameter combinations, we can not determine the value of the fixed point. The results at least do not contradict Conjecture (2).

The dependency on κ is shown in Figure 22b. Since an increase in κ results in a shorter residual service time once admitted to the VVT, the throughput of the system will increase and thus a decrease in the mean number of patients in overflow is expected. It can be seen that the mean number of overflow patients in the call packing system approaches that in the $M/G/\infty$ system and in the $M/M/c$ system when κ approaches 1 and 0, respectively, as expected. The unadapted $M/G/\infty$ and $M/M/c$ systems do not depend on κ. The mean number of patients in overflow in these systems is therefore constant when ranging over κ. For every value of κ, the adapted infinite server system provides a lower bound on the mean number of patients in the call packing system for all researched service distributions, in line with Conjecture (1). Note that for large coefficients of variation, the mean number of patients in overflow in the call packing system with lognormal service times can be larger than in the $M/M/c$ system (with adapted service times). The reader should not be surprised by this as the latter only holds for exponential service time. It is therefore not in contradiction with Conjecture 2.

A.3 Sensitivity with respect to mean service rate and arrival rate

The capacity in each system behaves as expected with respect to a change in service rate and arrival rate. Figure 23 shows that the capacity increases rapidly for decreasing service rates, which should be the case as the capacity should go to infinity when μ tends to zero. Figure 24 shows that the optimal capacity increases almost linear with the arrival rate. The remarkable behaviour of the adapted $M/M/c$ at $\lambda = 0.3$ is due to the repeated substitution for determining L^* having the two subsequences not converge to the same value, and should thus be considered as not being able to produce a reliable result for $\lambda = 0.3$. Most importantly, the ordering of the systems concerning the optimal capacity is maintained both when ranging the service rate as well as when ranging the arrival rate.



(A) Mean service time ranges from 100 to 25 days.

(B) Mean service time ranges from 45 to 35 days.

FIGURE 23: The optimal capacity for different service rates.

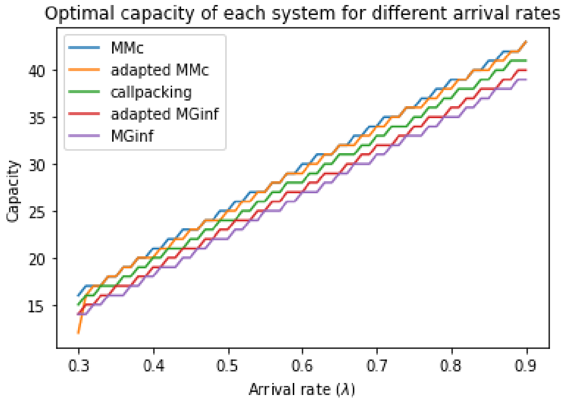


FIGURE 24: The optimal capacity when ranging arrival rate from 0.3 to 0.9 per day.

A.4 Collection of statistics in simulation

During the simulation, several performance measures are kept track of. In this way we collect for each simulation run the time-weighted average number of patients in overflow, number of unused beds, number of used beds and, in case of time dependent capacity, number of overbeds. These values are used to calculate the average total cost per day in that simulation run. The statistical analysis of the total cost per day of a certain situation (parameters and capacity) is based on the replication method as described in, among others, Section 9.3.4 of [64]. Denote the average total cost per day in simulation run j by z_j. The mean and variance of these independent observations are $\bar{z} = \frac{\sum_{j=1}^{m} z_j}{m}$ and $s_{\bar{z}} = \sqrt{\frac{\sum_{j=1}^{m}(z_j - \bar{z})^2}{m-1}}$ where m is the number of simulation runs. The $100(1-\alpha)\%$ confidence interval (CI) for the cost per day of a certain situation is then given by

$$[\bar{z} - \frac{t(m-1, 1-\alpha/2)s_{\bar{z}})}{\sqrt{m}}, \bar{z} + \frac{t(m-1, 1-\alpha/2)s_{\bar{z}})}{\sqrt{m}}], \tag{44}$$

where $t(m-1, 1-\alpha/2)$ is the inverse of the Student t-distribution with $(m-1)$ degrees of freedom at $1-\alpha/2$.

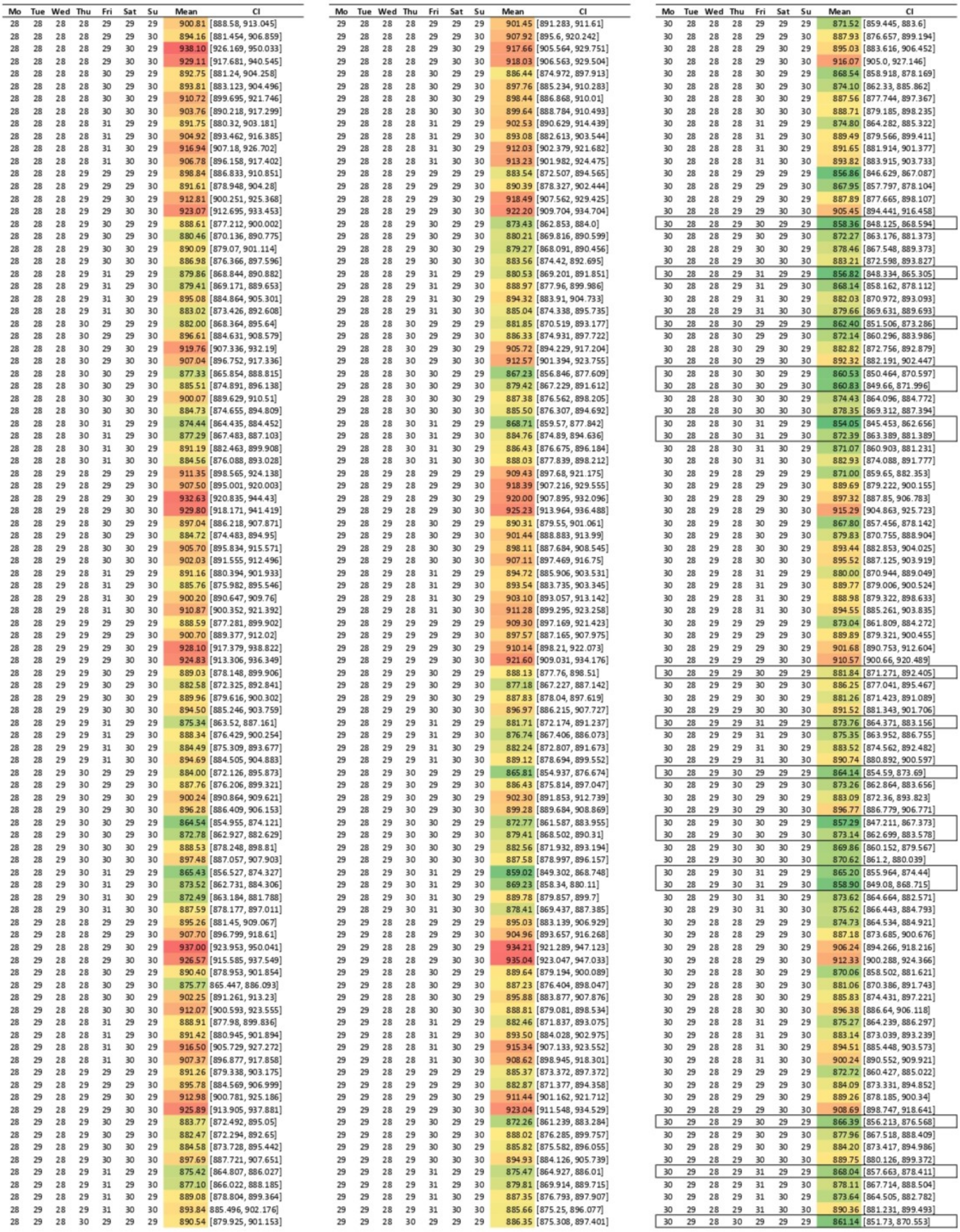
At the start of a simulation run, all beds are empty. To prevent an underestimation of the performance measure, this start-up behaviour should be neglected, i.e. we only want to measure the steady-state behaviour of the system. The relevant statistics are therefore only collected after a warm-up period of D days. Besides trial and error, several methods for determining an appropriate warm-up period can be found in literature. Based on [76], we study when the change in observations of a performance measure is sufficiently small. We are interested in the first integer B for which

$$\left| \frac{\frac{1}{2B}\sum_{i=1}^{2B} \bar{w}_{.i}}{\frac{1}{B}\sum_{i=1}^{B} \bar{w}_{.i}} - 1 \right| \leq 0.05,$$

where $\bar{w}_{.k} = \sum_{j}^{m} w_{ji}$ is the point estimator for the expectation of the i-th observation of the overflow time in an arbitrary run. This yields $B = 592$. Since the average arrival rate is 0.61 per day, this corresponds to $\frac{592}{0.61} = 970$ days. Therefore, a warm-up period of 1000 days is used. To determine when to terminate the simulation run, the rule of thumb to take a length of $4D = 4000$ is used as suggested in [76].

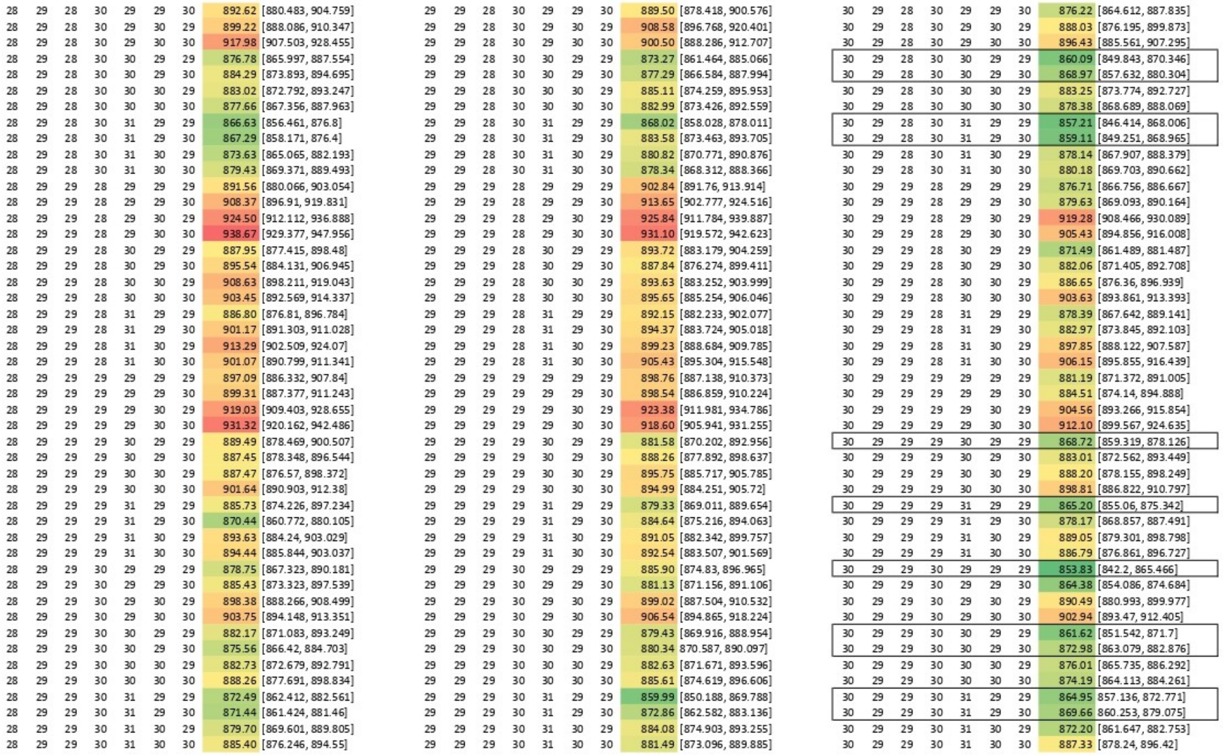
The confidence interval 44 relies on the central limit theorem. By using this, we implicitly assume that the number of simulation runs m is sufficiently large. Since the width of the confidence interval scales with one over the square root of m, a high number of simulation runs decreases the size of the confidence interval. However, a trade-off with the simulation time must be made. In this research, the confidence intervals are constructed based on 500 simulation runs if not stated differently. It is checked that 500 observations of z_j suffice to meet a relative accuracy of $\delta = 0.1$ by verifying that the half-width of the confidence intervals satisfies $\frac{t(m-1,1-\alpha/2)s_{\bar{z}})}{\sqrt{m}} \leq \frac{\delta \bar{x}}{(1+\delta)}$.

A.5 Results extensive neighbourhood search



Mo	Tue	Wed	Thu	Fri	Sat	Su	Mean	CI
28	28	28	28	29	29	29	900.81	[888.58, 913.045]
28	28	28	28	29	29	30	894.16	[881.454, 906.859]
28	28	28	28	29	30	29	938.10	[926.169, 950.033]
28	28	28	28	29	30	30	929.11	[917.681, 940.545]
28	28	28	28	30	29	29	892.75	[881.24, 904.258]
28	28	28	28	30	29	30	893.81	[883.123, 904.496]
28	28	28	28	30	30	29	910.72	[899.695, 921.746]
28	28	28	28	30	30	30	903.76	[890.218, 917.299]
28	28	28	28	31	29	29	891.75	[880.32, 903.181]
28	28	28	28	31	29	30	904.92	[893.462, 916.385]
28	28	28	28	31	30	29	916.94	[907.18, 926.702]
28	28	28	28	31	30	30	906.78	[896.158, 917.402]
28	28	28	29	29	29	29	898.84	[886.833, 910.851]
28	28	28	29	29	29	30	891.61	[878.948, 904.28]
28	28	28	29	29	30	29	912.81	[900.251, 925.368]
28	28	28	29	29	30	30	923.07	[912.695, 933.453]
28	28	28	29	30	29	29	888.61	[877.212, 900.002]
28	28	28	29	30	29	30	880.46	[870.136, 890.775]
28	28	28	29	30	30	29	890.09	[879.07, 901.114]
28	28	28	29	30	30	30	886.98	[876.366, 897.596]
28	28	28	29	31	29	29	879.86	[868.844, 890.882]
28	28	28	29	31	29	30	879.41	[869.171, 889.653]
28	28	28	29	31	30	29	895.08	[884.864, 905.301]
28	28	28	29	31	30	30	883.02	[873.426, 892.608]
28	28	28	30	29	29	29	882.00	[868.364, 895.64]
28	28	28	30	29	29	30	896.61	[884.631, 908.579]
28	28	28	30	29	30	29	919.76	[907.336, 932.19]
28	28	28	30	29	30	30	907.04	[896.752, 917.336]
28	28	28	30	30	29	29	877.33	[865.854, 888.815]
28	28	28	30	30	29	30	885.51	[874.891, 896.138]
28	28	28	30	30	30	29	900.07	[889.629, 910.51]
28	28	28	30	30	30	30	884.73	[874.655, 894.809]
28	28	28	30	31	29	29	874.44	[864.435, 884.452]
28	28	28	30	31	29	30	877.29	[867.483, 887.103]
28	28	28	30	31	30	29	891.19	[882.463, 899.908]
28	28	28	30	31	30	30	884.56	[876.088, 893.028]
28	28	29	28	29	29	29	911.35	[898.565, 924.138]
28	28	29	28	29	29	30	907.50	[895.001, 920.003]
28	28	29	28	29	30	29	932.63	[920.835, 944.43]
28	28	29	28	29	30	30	929.80	[918.171, 941.419]
28	28	29	28	30	29	29	897.04	[886.218, 907.871]
28	28	29	28	30	29	30	884.72	[874.483, 894.95]
28	28	29	28	30	30	29	905.70	[895.834, 915.571]
28	28	29	28	30	30	30	902.03	[891.555, 912.496]
28	28	29	28	31	29	29	891.16	[880.394, 901.933]
28	28	29	28	31	29	30	885.76	[875.982, 895.546]
28	28	29	28	31	30	29	900.20	[890.647, 909.76]
28	28	29	28	31	30	30	910.87	[900.352, 921.392]
28	28	29	29	29	29	29	888.59	[877.281, 899.902]
28	28	29	29	29	29	30	900.70	[889.377, 912.02]
28	28	29	29	29	30	29	928.10	[917.379, 938.822]
28	28	29	29	29	30	30	924.83	[913.306, 936.349]
28	28	29	29	30	29	29	889.03	[878.148, 899.906]
28	28	29	29	30	29	30	882.58	[872.325, 892.841]
28	28	29	29	30	30	29	889.96	[879.616, 900.302]
28	28	29	29	30	30	30	894.50	[885.246, 903.759]
28	28	29	29	31	29	29	875.34	[863.52, 887.161]
28	28	29	29	31	29	30	888.34	[876.429, 900.254]
28	28	29	29	31	30	29	884.49	[875.309, 893.677]
28	28	29	29	31	30	30	894.69	[884.505, 904.883]
28	28	29	30	29	29	29	884.00	[872.126, 895.873]
28	28	29	30	29	29	30	887.76	[876.206, 899.321]
28	28	29	30	29	30	29	900.24	[890.864, 909.621]
28	28	29	30	29	30	30	896.28	[886.409, 906.153]
28	28	29	30	30	29	29	864.54	[854.955, 874.121]
28	28	29	30	30	29	30	872.78	[862.927, 882.629]
28	28	29	30	30	30	29	888.53	[878.248, 898.81]
28	28	29	30	30	30	30	897.48	[887.057, 907.903]
28	28	29	30	31	29	29	865.43	[856.527, 874.327]
28	28	29	30	31	29	30	873.52	[862.731, 884.306]
28	28	29	30	31	30	29	872.49	[863.184, 881.788]
28	28	29	30	31	30	30	887.59	[878.177, 897.011]
28	29	28	28	29	29	29	895.26	[881.45, 909.067]
28	29	28	28	29	29	30	907.70	[896.799, 918.61]
28	29	28	28	29	30	29	937.00	[923.953, 950.041]
28	29	28	28	29	30	30	926.57	[915.585, 937.549]
28	29	28	28	30	29	29	890.40	[878.953, 901.854]
28	29	28	28	30	29	30	875.77	865.447, 886.093]
28	29	28	28	30	30	29	902.25	[891.261, 913.23]
28	29	28	28	30	30	30	912.07	[900.593, 923.555]
28	29	28	28	31	29	29	888.91	[877.98, 899.836]
28	29	28	28	31	29	30	891.42	[880.945, 901.894]
28	29	28	28	31	30	29	916.50	[905.729, 927.272]
28	29	28	28	31	30	30	907.37	[896.877, 917.858]
28	29	28	29	29	29	29	891.26	[879.338, 903.175]
28	29	28	29	29	29	30	895.78	[884.569, 906.999]
28	29	28	29	29	30	29	912.98	[900.781, 925.186]
28	29	28	29	29	30	30	925.89	[913.905, 937.881]
28	29	28	29	30	29	29	883.77	[872.492, 895.05]
28	29	28	29	30	29	30	882.47	[872.294, 892.65]
28	29	28	29	30	30	29	884.58	[873.728, 895.442]
28	29	28	29	30	30	30	897.69	[887.721, 907.651]
28	29	28	29	31	29	29	875.42	[864.807, 886.027]
28	29	28	29	31	29	30	877.10	[866.022, 888.185]
28	29	28	29	31	30	29	889.08	[878.804, 899.364]
28	29	28	29	31	30	30	893.84	885.496, 902.176]
28	29	28	30	29	29	29	890.54	[879.925, 901.153]
29	28	28	28	29	29	29	901.45	[891.283, 911.61]
29	28	28	28	29	29	30	907.92	[895.6, 920.242]
29	28	28	28	29	30	29	917.66	[905.564, 929.751]
29	28	28	28	29	30	30	918.03	[906.563, 929.504]
29	28	28	28	30	29	29	886.44	[874.972, 897.913]
29	28	28	28	30	29	30	897.76	[885.234, 910.283]
29	28	28	28	30	30	29	898.44	[886.868, 910.01]
29	28	28	28	30	30	30	899.64	[888.784, 910.493]
29	28	28	28	31	29	29	902.53	[890.629, 914.439]
29	28	28	28	31	29	30	893.08	[882.613, 903.544]
29	28	28	28	31	30	29	912.03	[902.379, 921.682]
29	28	28	28	31	30	30	913.23	[901.982, 924.475]
29	28	28	29	29	29	29	883.54	[872.507, 894.565]
29	28	28	29	29	29	30	890.39	[878.327, 902.444]
29	28	28	29	29	30	29	918.49	[907.562, 929.425]
29	28	28	29	29	30	30	922.20	[909.704, 934.704]
29	28	28	29	30	29	29	873.43	[862.853, 884.0]
29	28	28	29	30	29	30	880.21	[869.816, 890.599]
29	28	28	29	30	30	29	879.27	[868.091, 890.456]
29	28	28	29	30	30	30	883.56	[874.42, 892.695]
29	28	28	29	31	29	29	880.53	[869.201, 891.851]
29	28	28	29	31	29	30	888.97	[877.96, 899.986]
29	28	28	29	31	30	29	894.32	[883.91, 904.733]
29	28	28	29	31	30	30	885.04	[874.338, 895.735]
29	28	28	30	29	29	29	881.85	[870.519, 893.177]
29	28	28	30	29	29	30	886.33	[874.931, 897.722]
29	28	28	30	29	30	29	905.72	[894.229, 917.204]
29	28	28	30	29	30	30	912.57	[901.394, 923.755]
29	28	28	30	30	29	29	867.23	[856.846, 877.609]
29	28	28	30	30	29	30	879.42	[867.229, 891.612]
29	28	28	30	30	30	29	887.38	[876.562, 898.205]
29	28	28	30	30	30	30	885.50	[876.307, 894.692]
29	28	28	30	31	29	29	868.71	[859.57, 877.842]
29	28	28	30	31	29	30	884.76	[874.89, 894.636]
29	28	28	30	31	30	29	886.43	[876.675, 896.184]
29	28	28	30	31	30	30	888.03	[877.839, 898.212]
29	28	29	28	29	29	29	909.43	[897.68, 921.175]
29	28	29	28	29	29	30	918.39	[907.216, 929.555]
29	28	29	28	29	30	29	920.00	[907.895, 932.096]
29	28	29	28	29	30	30	925.23	[913.964, 936.488]
29	28	29	28	30	29	29	890.31	[879.55, 901.061]
29	28	29	28	30	29	30	901.44	[888.883, 913.99]
29	28	29	28	30	30	29	898.11	[887.684, 908.545]
29	28	29	28	30	30	30	907.11	[897.469, 916.75]
29	28	29	28	31	29	29	894.72	[885.906, 903.531]
29	28	29	28	31	29	30	893.54	[883.735, 903.345]
29	28	29	28	31	30	29	903.10	[893.057, 913.142]
29	28	29	28	31	30	30	911.28	[899.295, 923.258]
29	28	29	29	29	29	29	909.30	[897.169, 921.423]
29	28	29	29	29	29	30	897.57	[887.165, 907.975]
29	28	29	29	29	30	29	910.14	[898.21, 922.073]
29	28	29	29	29	30	30	921.60	[909.031, 934.176]
29	28	29	29	30	29	29	888.13	[877.76, 898.51]
29	28	29	29	30	29	30	877.18	[867.227, 887.142]
29	28	29	29	30	30	29	887.83	[878.04, 897.619]
29	28	29	29	30	30	30	896.97	[886.215, 907.727]
29	28	29	29	31	29	29	881.71	[872.174, 891.237]
29	28	29	29	31	29	30	876.74	[867.406, 886.073]
29	28	29	29	31	30	29	882.24	[872.807, 891.673]
29	28	29	29	31	30	30	889.12	[878.694, 899.552]
29	28	29	30	29	29	29	865.81	[854.937, 876.674]
29	28	29	30	29	29	30	886.43	[875.814, 897.047]
29	28	29	30	29	30	29	902.30	[891.853, 912.739]
29	28	29	30	29	30	30	899.28	[889.684, 908.869]
29	28	29	30	30	29	29	872.77	[861.587, 883.955]
29	28	29	30	30	29	30	879.41	[868.502, 890.31]
29	28	29	30	30	30	29	882.56	[871.932, 893.194]
29	28	29	30	30	30	30	887.58	[878.997, 896.157]
29	28	29	30	31	29	29	859.02	[849.302, 868.748]
29	28	29	30	31	29	30	869.23	[858.34, 880.11]
29	28	29	30	31	30	29	889.78	[879.857, 899.7]
29	28	29	30	31	30	30	878.41	[869.437, 887.385]
29	29	28	28	29	29	29	895.03	[883.139, 906.929]
29	29	28	28	29	29	30	904.96	[893.657, 916.268]
29	29	28	28	29	30	29	934.21	[921.289, 947.123]
29	29	28	28	29	30	30	935.04	[923.047, 947.033]
29	29	28	28	30	29	29	889.64	[879.194, 900.089]
29	29	28	28	30	29	30	887.23	[876.404, 898.047]
29	29	28	28	30	30	29	895.88	[883.877, 907.876]
29	29	28	28	30	30	30	888.81	[879.081, 898.534]
29	29	28	28	31	29	29	882.46	[871.837, 893.075]
29	29	28	28	31	29	30	893.50	[884.028, 902.975]
29	29	28	28	31	30	29	915.34	[907.133, 923.552]
29	29	28	28	31	30	30	908.62	[898.945, 918.301]
29	29	28	29	29	29	29	885.37	[873.372, 897.372]
29	29	28	29	29	29	30	882.87	[871.377, 894.358]
29	29	28	29	29	30	29	911.44	[901.162, 921.712]
29	29	28	29	29	30	30	923.04	[911.548, 934.529]
29	29	28	29	30	29	29	872.26	[861.239, 883.284]
29	29	28	29	30	29	30	888.02	[876.285, 899.757]
29	29	28	29	30	30	29	885.82	[875.582, 896.055]
29	29	28	29	30	30	30	894.93	[884.126, 905.739]
29	29	28	29	31	29	29	875.47	[864.927, 886.01]
29	29	28	29	31	29	30	879.81	[869.914, 889.715]
29	29	28	29	31	30	29	887.35	[876.793, 897.907]
29	29	28	29	31	30	30	885.66	[875.25, 896.077]
29	29	28	30	29	29	29	886.35	[875.308, 897.401]
30	28	28	28	29	29	29	871.52	[859.445, 883.6]
30	28	28	28	29	29	30	887.93	[876.657, 899.194]
30	28	28	28	29	30	29	895.03	[883.616, 906.452]
30	28	28	28	29	30	30	916.07	[905.0, 927.146]
30	28	28	28	30	29	29	868.54	[858.918, 878.169]
30	28	28	28	30	29	30	874.10	[862.33, 885.862]
30	28	28	28	30	30	29	887.56	[877.744, 897.367]
30	28	28	28	30	30	30	888.71	[879.185, 898.235]
30	28	28	28	31	29	29	874.80	[864.282, 885.322]
30	28	28	28	31	29	30	889.49	[879.566, 899.411]
30	28	28	28	31	30	29	891.65	[881.914, 901.377]
30	28	28	28	31	30	30	893.82	[883.915, 903.733]
30	28	28	29	29	29	29	856.86	[846.629, 867.087]
30	28	28	29	29	29	30	867.95	[857.797, 878.104]
30	28	28	29	29	30	29	887.89	[877.665, 898.107]
30	28	28	29	29	30	30	905.45	[894.441, 916.458]
30	28	28	29	30	29	29	858.36	[848.125, 868.594]
30	28	28	29	30	29	30	872.27	[863.176, 881.373]
30	28	28	29	30	30	29	878.46	[867.548, 889.373]
30	28	28	29	30	30	30	883.21	[872.598, 893.827]
30	28	28	29	31	29	29	856.82	[848.334, 865.305]
30	28	28	29	31	29	30	868.14	[858.162, 878.112]
30	28	28	29	31	30	29	882.03	[870.972, 893.093]
30	28	28	29	31	30	30	879.66	[869.631, 889.693]
30	28	28	30	29	29	29	862.40	[851.506, 873.286]
30	28	28	30	29	29	30	872.14	[860.296, 883.986]
30	28	28	30	29	30	29	882.82	[872.756, 892.879]
30	28	28	30	29	30	30	892.32	[882.191, 902.447]
30	28	28	30	30	29	29	860.53	[850.464, 870.597]
30	28	28	30	30	29	30	860.83	[849.66, 871.996]
30	28	28	30	30	30	29	874.43	[864.096, 884.772]
30	28	28	30	30	30	30	878.35	[869.312, 887.394]
30	28	28	30	31	29	29	854.05	[845.453, 862.656]
30	28	28	30	31	29	30	872.39	[863.389, 881.389]
30	28	28	30	31	30	29	871.07	[860.903, 881.231]
30	28	28	30	31	30	30	882.93	[874.088, 891.777]
30	28	29	28	29	29	29	871.00	[859.65, 882.353]
30	28	29	28	29	29	30	889.69	[879.222, 900.155]
30	28	29	28	29	30	29	897.32	[887.85, 906.783]
30	28	29	28	29	30	30	915.29	[904.863, 925.723]
30	28	29	28	30	29	29	867.80	[857.456, 878.142]
30	28	29	28	30	29	30	879.83	[870.755, 888.904]
30	28	29	28	30	30	29	893.44	[882.853, 904.025]
30	28	29	28	30	30	30	895.52	[887.125, 903.919]
30	28	29	28	31	29	29	880.00	[870.944, 889.049]
30	28	29	28	31	29	30	889.77	[879.006, 900.524]
30	28	29	28	31	30	29	888.98	[879.322, 898.633]
30	28	29	28	31	30	30	894.55	[885.261, 903.835]
30	28	29	29	29	29	29	873.04	[861.809, 884.272]
30	28	29	29	29	29	30	889.89	[879.321, 900.455]
30	28	29	29	29	30	29	901.68	[890.753, 912.604]
30	28	29	29	29	30	30	910.57	[900.66, 920.489]
30	28	29	29	30	29	29	881.84	[871.271, 892.405]
30	28	29	29	30	29	30	886.25	[877.041, 895.467]
30	28	29	29	30	30	29	881.26	[871.423, 891.089]
30	28	29	29	30	30	30	891.52	[881.343, 901.706]
30	28	29	29	31	29	29	873.76	[864.371, 883.156]
30	28	29	29	31	29	30	875.35	[863.952, 886.755]
30	28	29	29	31	30	29	883.52	[874.562, 892.482]
30	28	29	29	31	30	30	890.74	[880.892, 900.597]
30	28	29	30	29	29	29	864.14	[854.59, 873.69]
30	28	29	30	29	29	30	873.26	[862.864, 883.656]
30	28	29	30	29	30	29	883.09	[872.36, 893.823]
30	28	29	30	29	30	30	896.77	[886.779, 906.771]
30	28	29	30	30	29	29	857.29	[847.211, 867.373]
30	28	29	30	30	29	30	873.14	[862.699, 883.578]
30	28	29	30	30	30	29	869.86	[860.152, 879.567]
30	28	29	30	30	30	30	870.62	[861.2, 880.039]
30	28	29	30	31	29	29	865.20	[855.964, 874.44]
30	28	29	30	31	29	30	858.90	[849.08, 868.715]
30	28	29	30	31	30	29	873.62	[864.664, 882.571]
30	28	29	30	31	30	30	875.62	[866.443, 884.793]
30	29	28	28	29	29	29	874.73	[864.534, 884.921]
30	29	28	28	29	29	30	887.18	[873.685, 900.676]
30	29	28	28	29	30	29	906.24	[894.266, 918.216]
30	29	28	28	29	30	30	912.33	[900.288, 924.366]
30	29	28	28	30	29	29	870.06	[858.502, 881.621]
30	29	28	28	30	29	30	881.06	[870.386, 891.743]
30	29	28	28	30	30	29	885.83	[874.431, 897.221]
30	29	28	28	30	30	30	896.38	[886.64, 906.118]
30	29	28	28	31	29	29	875.27	[864.239, 886.297]
30	29	28	28	31	29	30	883.14	[873.039, 893.239]
30	29	28	28	31	30	29	894.51	[885.448, 903.573]
30	29	28	28	31	30	30	900.24	[890.552, 909.921]
30	29	28	29	29	29	29	872.72	[860.427, 885.022]
30	29	28	29	29	29	30	884.09	[873.331, 894.852]
30	29	28	29	29	30	29	889.26	[878.185, 900.34]
30	29	28	29	29	30	30	908.69	[898.747, 918.641]
30	29	28	29	30	29	29	866.39	[856.213, 876.568]
30	29	28	29	30	29	30	877.96	[867.518, 888.409]
30	29	28	29	30	30	29	884.20	[873.417, 894.986]
30	29	28	29	30	30	30	889.75	[880.126, 899.372]
30	29	28	29	31	29	29	868.04	[857.663, 878.411]
30	29	28	29	31	29	30	878.11	[867.714, 888.504]
30	29	28	29	31	30	29	873.64	[864.505, 882.782]
30	29	28	29	31	30	30	890.36	[881.231, 899.493]
30	29	28	30	29	29	29	861.14	[851.73, 870.553]

Figure 25: The cost per capacity vector of the extensive simulation study. The repeating pattern of well-performing capacity vectors is indicated by the bordered cells. The list continues on the next page.



28	29	28	30	29	29	30	892.62	[880.483, 904.759]
28	29	28	30	29	30	29	899.22	[888.086, 910.347]
28	29	28	30	29	30	30	917.98	[907.503, 928.455]
28	29	28	30	30	29	29	876.78	[865.997, 887.554]
28	29	28	30	30	29	30	884.29	[873.893, 894.695]
28	29	28	30	30	30	29	883.02	[872.792, 893.247]
28	29	28	30	30	30	30	877.66	[867.356, 887.963]
28	29	28	30	31	29	29	866.63	[856.461, 876.8]
28	29	28	30	31	29	30	867.29	[858.171, 876.4]
28	29	28	30	31	30	29	873.63	[865.065, 882.193]
28	29	28	30	31	30	30	879.43	[869.371, 889.493]
28	29	29	28	29	29	29	891.56	[880.066, 903.054]
28	29	29	28	29	29	30	908.37	[896.91, 919.831]
28	29	29	28	29	30	29	924.50	[912.112, 936.888]
28	29	29	28	29	30	30	938.67	[929.377, 947.956]
28	29	29	28	30	29	29	887.95	[877.415, 898.48]
28	29	29	28	30	29	30	895.54	[884.131, 906.945]
28	29	29	28	30	30	29	908.63	[898.211, 919.043]
28	29	29	28	30	30	30	903.45	[892.569, 914.337]
28	29	29	28	31	29	29	886.80	[876.81, 896.784]
28	29	29	28	31	29	30	901.17	[891.303, 911.028]
28	29	29	28	31	30	29	913.29	[902.509, 924.07]
28	29	29	28	31	30	30	901.07	[890.799, 911.341]
28	29	29	29	29	29	29	897.09	[886.332, 907.84]
28	29	29	29	29	29	30	899.31	[887.377, 911.243]
28	29	29	29	29	30	29	919.03	[909.403, 928.655]
28	29	29	29	29	30	30	931.32	[920.162, 942.486]
28	29	29	29	30	29	29	889.49	[878.469, 900.507]
28	29	29	29	30	29	30	887.45	[878.348, 896.544]
28	29	29	29	30	30	29	887.47	[876.57, 898.372]
28	29	29	29	30	30	30	901.64	[890.903, 912.38]
28	29	29	29	31	29	29	885.73	[874.226, 897.234]
28	29	29	29	31	29	30	870.44	[860.772, 880.105]
28	29	29	29	31	30	29	893.63	[884.24, 903.029]
28	29	29	29	31	30	30	894.44	[885.844, 903.037]
28	29	29	30	29	29	29	878.75	[867.323, 890.181]
28	29	29	30	29	29	30	885.43	[873.323, 897.539]
28	29	29	30	29	30	29	898.38	[888.266, 908.499]
28	29	29	30	29	30	30	903.75	[894.148, 913.351]
28	29	29	30	30	29	29	882.17	[871.083, 893.249]
28	29	29	30	30	29	30	875.56	[866.42, 884.703]
28	29	29	30	30	30	29	882.73	[872.679, 892.791]
28	29	29	30	30	30	30	888.26	[877.691, 898.834]
28	29	29	30	31	29	29	872.49	[862.412, 882.561]
28	29	29	30	31	29	30	871.44	[861.424, 881.46]
28	29	29	30	31	30	29	879.70	[869.601, 889.805]
28	29	29	30	31	30	30	885.40	[876.246, 894.55]

29	29	28	30	29	29	30	889.50	[878.418, 900.576]
29	29	28	30	29	30	29	908.58	[896.768, 920.401]
29	29	28	30	29	30	30	900.50	[888.286, 912.707]
29	29	28	30	30	29	29	873.27	[861.464, 885.066]
29	29	28	30	30	29	30	877.29	[866.584, 887.994]
29	29	28	30	30	30	29	885.11	[874.259, 895.953]
29	29	28	30	30	30	30	882.99	[873.426, 892.559]
29	29	28	30	31	29	29	868.02	[858.028, 878.011]
29	29	28	30	31	29	30	883.58	[873.463, 893.705]
29	29	28	30	31	30	29	880.82	[870.771, 890.876]
29	29	28	30	31	30	30	878.34	[868.312, 888.366]
29	29	29	28	29	29	29	902.84	[891.76, 913.914]
29	29	29	28	29	29	30	913.65	[902.777, 924.516]
29	29	29	28	29	30	29	925.84	[911.784, 939.887]
29	29	29	28	29	30	30	931.10	[919.572, 942.623]
29	29	29	28	30	29	29	893.72	[883.179, 904.259]
29	29	29	28	30	29	30	887.84	[876.274, 899.411]
29	29	29	28	30	30	29	893.63	[883.252, 903.999]
29	29	29	28	30	30	30	895.65	[885.254, 906.046]
29	29	29	28	31	29	29	892.15	[882.233, 902.077]
29	29	29	28	31	29	30	894.37	[883.724, 905.018]
29	29	29	28	31	30	29	899.23	[888.684, 909.785]
29	29	29	28	31	30	30	905.43	[895.304, 915.548]
29	29	29	29	29	29	29	898.76	[887.138, 910.373]
29	29	29	29	29	29	30	898.54	[886.859, 910.224]
29	29	29	29	29	30	29	923.38	[911.981, 934.786]
29	29	29	29	29	30	30	918.60	[905.941, 931.255]
29	29	29	29	30	29	29	881.58	[870.202, 892.956]
29	29	29	29	30	29	30	888.26	[877.892, 898.637]
29	29	29	29	30	30	29	895.75	[885.717, 905.785]
29	29	29	29	30	30	30	894.99	[884.251, 905.72]
29	29	29	29	31	29	29	879.33	[869.011, 889.654]
29	29	29	29	31	29	30	884.64	[875.216, 894.063]
29	29	29	29	31	30	29	891.05	[882.342, 899.757]
29	29	29	29	31	30	30	892.54	[883.507, 901.569]
29	29	29	30	29	29	29	885.90	[874.83, 896.965]
29	29	29	30	29	29	30	881.13	[871.156, 891.106]
29	29	29	30	29	30	29	899.02	[887.504, 910.532]
29	29	29	30	29	30	30	906.54	[894.865, 918.224]
29	29	29	30	30	29	29	879.43	[869.916, 888.954]
29	29	29	30	30	29	30	880.34	870.587, 890.097]
29	29	29	30	30	30	29	882.63	[871.671, 893.596]
29	29	29	30	30	30	30	885.61	[874.619, 896.606]
29	29	29	30	31	29	29	859.99	[850.188, 869.788]
29	29	29	30	31	29	30	872.86	[862.582, 883.136]
29	29	29	30	31	30	29	884.08	[874.903, 893.255]
29	29	29	30	31	30	30	881.49	[873.096, 889.885]

30	29	28	30	29	29	30	876.22	[864.612, 887.835]
30	29	28	30	29	30	29	888.03	[876.195, 899.873]
30	29	28	30	29	30	30	896.43	[885.561, 907.295]
30	29	28	30	30	29	29	860.09	[849.843, 870.346]
30	29	28	30	30	29	30	868.97	[857.632, 880.304]
30	29	28	30	30	30	29	883.25	[873.774, 892.727]
30	29	28	30	30	30	30	878.38	[868.689, 888.069]
30	29	28	30	31	29	29	857.21	[846.414, 868.006]
30	29	28	30	31	29	30	859.11	[849.251, 868.965]
30	29	28	30	31	30	29	878.14	[867.907, 888.379]
30	29	28	30	31	30	30	880.18	[869.703, 890.662]
30	29	29	28	29	29	29	876.71	[866.756, 886.667]
30	29	29	28	29	29	30	879.63	[869.093, 890.164]
30	29	29	28	29	30	29	919.28	[908.466, 930.089]
30	29	29	28	29	30	30	905.43	[894.856, 916.008]
30	29	29	28	30	29	29	871.49	[861.489, 881.487]
30	29	29	28	30	29	30	882.06	[871.405, 892.708]
30	29	29	28	30	30	29	886.65	[876.36, 896.939]
30	29	29	28	30	30	30	903.63	[893.861, 913.393]
30	29	29	28	31	29	29	878.39	[867.642, 889.141]
30	29	29	28	31	29	30	882.97	[873.845, 892.103]
30	29	29	28	31	30	29	897.85	[888.122, 907.587]
30	29	29	28	31	30	30	906.15	[895.855, 916.439]
30	29	29	29	29	29	29	881.19	[871.372, 891.005]
30	29	29	29	29	29	30	884.51	[874.14, 894.888]
30	29	29	29	29	30	29	904.56	[893.266, 915.854]
30	29	29	29	29	30	30	912.10	[899.567, 924.635]
30	29	29	29	30	29	29	868.72	[859.319, 878.126]
30	29	29	29	30	29	30	883.01	[872.562, 893.449]
30	29	29	29	30	30	29	888.20	[878.155, 898.249]
30	29	29	29	30	30	30	898.81	[886.822, 910.797]
30	29	29	29	31	29	29	865.20	[855.06, 875.342]
30	29	29	29	31	29	30	878.17	[868.857, 887.491]
30	29	29	29	31	30	29	889.05	[879.301, 898.798]
30	29	29	29	31	30	30	886.79	[876.861, 896.727]
30	29	29	30	29	29	29	853.83	[842.2, 865.466]
30	29	29	30	29	29	30	864.38	[854.086, 874.684]
30	29	29	30	29	30	29	890.49	[880.993, 899.977]
30	29	29	30	29	30	30	902.94	[893.47, 912.405]
30	29	29	30	30	29	29	861.62	[851.542, 871.7]
30	29	29	30	30	29	30	872.98	[863.079, 882.876]
30	29	29	30	30	30	29	876.01	[865.735, 886.292]
30	29	29	30	30	30	30	874.19	[864.113, 884.261]
30	29	29	30	31	29	29	864.95	857.136, 872.771]
30	29	29	30	31	29	30	869.66	860.253, 879.075]
30	29	29	30	31	30	29	872.20	[861.647, 882.753]
30	29	29	30	31	30	30	887.33	[878.24, 896.42]

Figure 26: The cost per capacity vector of the extensive simulation study. The repeating pattern of well-performing capacity vectors is indicated by the bordered cells. This is a continuation of the list on the previous page.

A.6 Alternative proof that $Q_1(\rho)$ equals probability of overflow

In Section 3.1.3, we noted that the derivative of the mean number of people in overflow in an infinite server system equals the probability of being in overflow. If we take the derivative directly from the summation originally defining $Q_1(r) = \sum_{n=c+1}^{\infty}(n-c)\pi(c)$, and assume we can switch derivative and summation (series converges), we see that the derivative of the mean number of people in overflow in a $M/G/\infty$ queue equals the probability of being in overflow,

$$\begin{aligned}
Q_1'(r) = \frac{d}{dr}\sum_{n=c+1}^{\infty}(n-c)\pi(n) &= \frac{d}{dr}\sum_{n=c+1}^{\infty}(n-c)e^{-r}\frac{r^n}{n!} \\
&= \sum_{n=c+1}^{\infty}(n-c)\frac{d}{dr}(e^{-r}\frac{r^n}{n!}) \\
&= \sum_{n=c+1}^{\infty}(n-c)(-e^{-r}\frac{r^n}{n!}+e^{-r}\frac{nr^{n-1}}{n!}) \\
&= \sum_{n=c+1}^{\infty}(n-c)e^{-r}\frac{r^{n-1}}{(n-1)!} - \sum_{n=c+1}^{\infty}(n-c)e^{-r}\frac{r^n}{n!} \\
&= \sum_{n=c}^{\infty}(n+1-c)\pi(n) - \sum_{n=c+1}^{\infty}(n-c)\pi(n) \\
&= \sum_{k=1}^{\infty}k\pi(c+k-1) - \sum_{k=1}^{\infty}(k-1)\pi(c+k-1) \\
&= \sum_{k=1}^{\infty}\pi(c+k-1) \\
&= \sum_{n=c}^{\infty}\pi(n).
\end{aligned}$$

A.7 Queue length dependent arrival process

In this research, we assumed that the influence of overflow patients on the working of the hospital is negligible. As noted in Remark 2, if the influence of overflow is substantial, the number of patients in overflow will influence the arrival rate to the VVT. This can be modelled by a state-dependent arrival rate to the VVT, or more precisely, an arrival rate that depends on the total number of patients in the queue. At the beginning of the period in which this research took place, we have studied this scenario. Since there are situations in which such state-dependent arrival rates are applicable and since theoretical insights were obtained, we share our analysis here. Our analysis aimed for a product form solution for the steady-state distribution of the number of patients in a system with parallel queues with a state-dependent arrival process and shared buffer space.

A.7.1 Literature

The concept of product form solutions was first introduced by Jackson (1963) and Gordon and Newell (1967) for First Come First Served stations with exponential service times and one customer class. Kelly (1976) extended this to multiple customer classes. Baskett et al. (1975) extended the general product form solution as to include multiple customer classes and (a specific type of) general service time distributions. The well-known collection of networks that satisfy that product form formulation is named after them: BCMP networks. The proof of the product form is based on the technique of 'independent balance equations' as introduced by Whittle (1968), or 'the principle of local balance' as formulated by Chandy (2002). Unfortunately, the service time distribution in the BCMP network is restricted to have a rational Laplace transform for the service types Last Come First Serve and Processor Sharing and is even further restricted to negative exponential in case of FCFS.

Baskett et al. (1975) consider two types of state-dependent arrival processes. In the first type, the total arrival rate to the network is Poisson with a mean that depends on the total number of customers in the network. In the second type, each subchain has a Poisson arrival stream that depends on the number of customers in that subchain.

Lam (1977) proves that the product formula for the BCMP network holds for more advanced population size constraints than treated by Baskett et al. (1975). The choice of Lam to base his proof on local balance as introduced by Chandy (2002) enabled him to show that the product form also holds in case of state-dependent loss and trigger functions. In these state-dependent loss and trigger functions, population constraints on the total population as well as on each individual class can be captured. Interestingly, Kamoun and Kleinrock (1980) also derived the product form solution for the same application, the store-and-forward networks for packet switching, around the same year. However, Kamoun and Kleinrock consider specifically a system of R parallel $M/M/1$ queues with shared finite storage space, while Lam looks at a broader class of networks. Kamoun and Kleinrock explicitly state some possible constraints on the shared buffer space, among which the *sharing with a minimum allocation* and the *sharing with maximum queue length*. The latter matches the population constraints posed by Lam (1977) in his Example 1. Although most literature on product form networks seems to consider single server stations, they can be applied to stations with multiple servers by observing that the working of a multiserver station is captured by a single server with state-dependent service rate $\min(c, n)\mu$.

A.7.2 Model design

We are interested in the situation in which the patients waiting for a place in a *Verpleeg-, Verzorgingshuizen en Thuiszorg (VVT)* organisation block a bed in their prior care organisation. Contrary to our main research described in the rest of this report, this bed blocking now has a substantial influence on the working of the prior care organisation.

The prior care organisation is not explicitly modelled, but the influence of bed blocking on the prior care organisation is captured in the arrival process to the VVT. One can picture that as follows. If a substantial part of the beds in the prior care organisation are blocked by patients waiting for VVT care, the prior care organisation has less room to accept new patients. If the organisation schedules fewer people for surgery, there will be fewer people who need rehabilitation afterwards, and thus less demand for the VVT. Since the patients that block a bed are the patients waiting for the VVT, the arrival rate to the VVT depends on the number of patients in the queue for the VVT. Please note that the situation described mainly applies to elective patients. The arrival rate of emergency patients is not expected to be influenced strongly by the queue length of the

VVT. In Section 2.4, several types of VVT care were discussed. Let R denote the number of VVT types. These R VVT types are assumed to be independent in the VVT organisations. Several possibilities for their interaction in the prior care organisation are considered. The easiest case is when they are independent in the prior care organisation as well, for example when distinct prior care is applicable. The most interesting case is when patients for different VVT types come from the same department of the prior care organisation, since then the queue length of one VVT type influences the arrival rate of another type. These different interactions in the prior care organisation correspond to different constraints on the shared *buffer space* of Kamoun and Kleinrock (1980). In summary, the situation of interest is modelled as a system with R parallel queues with a state-dependent arrival rate and shared buffer space. The goal is to find the steady-state distribution of the number of patients in this system, for which a product form expression is desired. Since a product form expression is desired, both the interarrival times as well as the service times are assumed to be exponentially distributed.

A.7.3 Model formulation

Let c_r be the capacity of VVT type $r \in [R]$ and c_0 be the capacity of the prior care organisation. The states are denoted by $\mathbf{n} = (n_1, n_2, .., n_R)$ where n_r is the number of patients that want VVT care of type r. Since the capacity of VVT type r is c_r, there are $n_r - c_r$ patients in the queue when $n_r > c_r$. So, the number of patients in service at VVT type r is $\min(n_r, c_r)$ and the number of patients in the queue for type r is $(n_r - c_r)^+$. Since the capacity in the prior care organisation limits the possible number of patients in the queue and is shared by all types, the state space contains all states that satisfy the total capacity available,

$$S = \{\mathbf{n} = (n_1, n_2, .., n_R) |\ 0 < n_r < c_r + c_0 \ \ \forall r \in [R], \sum_{r=1}^{R} (n_r - c_r)^+ \leq c_0\}.$$

The most interesting case in which all patients for the VVT organisations come from the same department of the prior care organisation is considered. The arrival rate to each type r is assumed to be a fixed fraction p_r of the total arrival rate from the prior care organisation, λ_0. Let λ denote the arrival rate from the prior care organisation when no bed blocking takes place. The total arrival rate from the prior care organisation is assumed to decrease linearly in the number of occupied beds in that organisation. Thus, we have a state dependent arrival rate $\lambda_0(\mathbf{n}) = (1 - \frac{\sum_{r=1}^{R}(n_r - c_r)^+}{c_0})\lambda$ from the prior care organisation. The state-dependent arrival rate λ_r to VVT type r is thus given by

$$\lambda_r(\mathbf{n}) = p_r \lambda_0(\mathbf{n}) = p_r (1 - \frac{\sum_{r=1}^{R}(n_r - c_r)^+}{c_0})\lambda. \tag{45}$$

The service rate for one patient of type r is μ_r. The departure rate in state $\mathbf{n}$ is thus $\min(n_r, c_r)\mu_r$.

A.7.4 Two VVT types

Figure 27 shows the transition state diagram for the case $R = 2$. Let $\pi(\mathbf{n})$ be the probability that the system is in state $\mathbf{n}$ in steady state. Per state, the sum of the rates into the state can be set equal to the rates out of the state. The system of these balance equations, supplemented with $\sum_{\mathbf{n} \in S} \pi(\mathbf{n}) = 1$, can be solved numerically. However, the number of states grows significantly when the number of VVT types increases. A product form for the steady state probabilities is therefore desired.

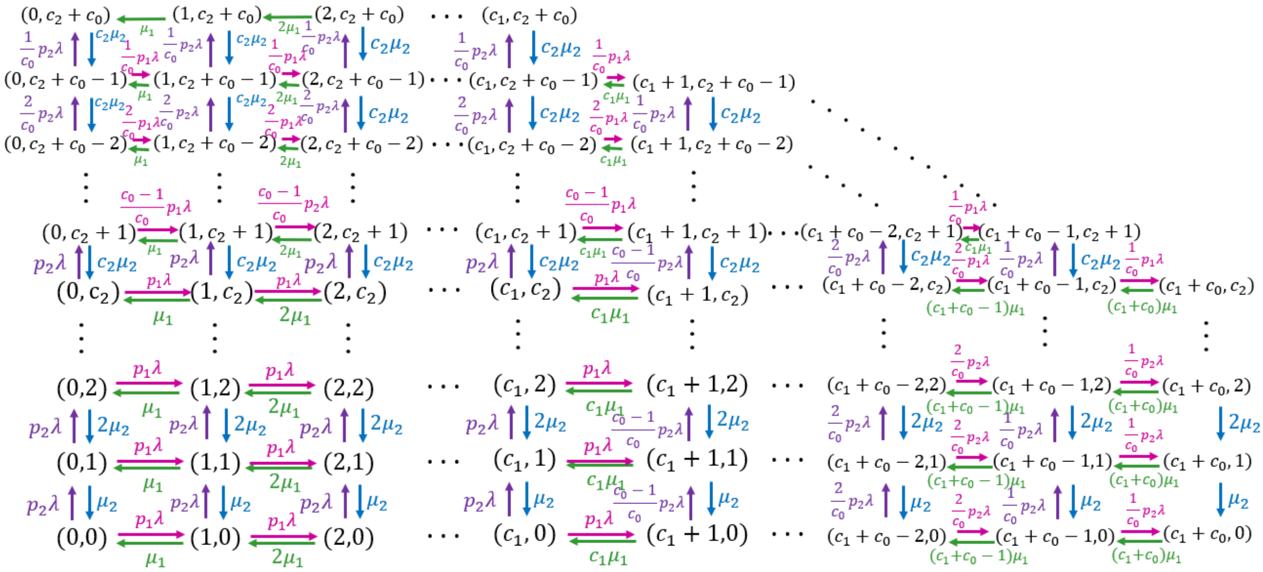


FIGURE 27: The transition state diagram for the case $R = 2$.

A.7.5 Product form solution

First, we note that the state space S corresponds to the *sharing with a minimum allocation (SMA)* scheme of Kamoun and Kleinrock (1980). Since this differs from their *sharing with maximum queue length (SMXQ)* only in the determination of the normalization constant, and the SMXQ is a special case of the systems considered by Lam (1977), we will focus on the results from Lam. To limit the material in this Appendix, the reader is advised to familiarise themselves with Lam (1977) first.

The introduction of loss and trigger functions allows Lam to derive a sufficient condition for the existence of a product form in their *Theorem* on page 373. They note that the condition in this theorem means that in the graphical representation neighbouring points in an irreducible subset of states must either be "doubly connected" or not connected at all. We note that in Figure 27, the states $(n_1, c_2 + c_0)$ with $n_1 \leq c_1$ are connected by a single edge. Similarly for $(c_1 + c_0, n_2)$ with $n_2 \leq c_2$. It is thus likely that no product form exists for the system depicted in Figure 27.

Interestingly, results that a certain system does not have a product form do not seem to appear often in literature. Weij et al. (2012) note this as well. They stress in Remark 3.10 the importance of the observation that a model does not have a product form. Our proof that a product form can not hold for the system with arrival rates given by (45) is based on Lemma 2.2 *(Equivalent Adjoint Reversibility Conditions)* of Weij et al. (2012). We will show that there exists a cycle $p := \mathbf{n}_0 \to \mathbf{n}_1 \to \cdots \to \mathbf{n}_t \to \mathbf{n}_{t+1} = \mathbf{n}_0$, with reverse cycle $\bar{p} := \mathbf{n}_0 = \mathbf{n}_{t+1} \to \mathbf{n}_t \to \cdots \to \mathbf{n}_1 \to \mathbf{n}_0$, for which $\theta(p) \neq \theta(\bar{p})$, where $\theta(p)$ is defined as the products of the transitions rates $\theta(p) := q(\mathbf{n}_0, \mathbf{n}_1)q(\mathbf{n}_1, \mathbf{n}_2) \cdots q(\mathbf{n}_t, \mathbf{n}_0)$.

In the system of our interest, $q(\mathbf{n}, \mathbf{n} + e_r) = \lambda_r(\mathbf{n})$ as given by (45) and $q(\mathbf{n}, \mathbf{n} - e_r) = \min(n_r, c_r)\mu_r$. We will take a path containing states in which $n_r = c_r$ and $n_r = c_r + 1$ for some types r. Without loss of generality, we will show the existence of such a path for the case of $R = 2$. The path can be extended to the general case by keeping $n_{\tilde{r}}$ fixed for $\tilde{r} > 2$. Consider path $p = (c_1, c_2 - 1) \to (c_1, c_2) \to (c_1, c_2 + 1) \to (c_1 + 1, c_2 + 1) \to (c_1 + 1, c_2) \to (c_1 + 1, c_2 - 1) \to (c_1, c_2 - 1)$. This path is shown in red in the state diagram in Figure 28. Its reverse cycle $\bar{p}$ is shown in blue. Calculating $\theta(p)$ and $\theta(\bar{p})$, or comparing the red and blue transitions in Figure 28, yields that $\theta(p)(1 - \frac{1}{c_0}) = \theta(\bar{p})$, and thus $\theta(p) \neq \theta(\bar{p})$. Thus, a product form does not exist for the system in which the arrival rate to the VVT depends on the queue length of all VVT types.

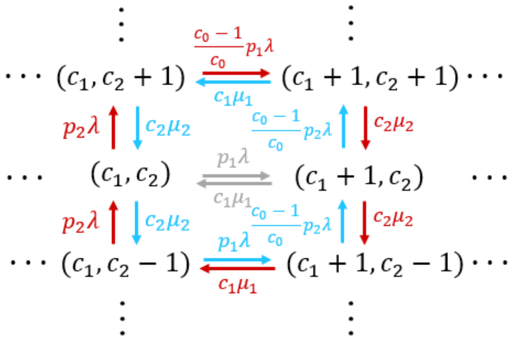


FIGURE 28: Path p and path $\bar{p}$ are visualised in the state diagram by respectively red and blue.

There are simplifications to the system that do yield a product form solution. Lam (1977) provides two state-dependent arrival processes for which a product form exists. In his first option, the total arrival rate λ_0 depends on the mean number of patients in the system, i.e. $\lambda_0(\mathbf{n}) = \sum_{r=1}^{R} n_r$, in contrast to the desired dependence on the number of patients in the queue. The second option corresponds to an arrival rate per VVT type that depends only on the number of patients of that type, i.e. $\lambda_r(n_r)$. For specific VVT types that have distinct care paths in prior care organisations, the latter simplification likely yields good results.

So, we proved that for the state-dependent arrival rate $\lambda_r(\mathbf{n}) = p_r(1 - \frac{\sum_{r=1}^{R}(n_r-c_r)^+}{c_0})\lambda$ no product form existed, while for $\lambda_r(\mathbf{n}) = \lambda_r(n_r) = p_r(1 - \frac{(n_r-c_r)^+}{c_0})\lambda$ a product form solution would exist. This begs the question for what functions $g(\mathbf{n})$ in the state dependent arrival rate $\lambda_r(\mathbf{n}) = p_r(1 - \frac{g(\mathbf{n})}{c_0})\lambda$, the system has a product form.

Boucherie and Van Dijk (1990) provide a very general expression for transition rates such that a product form exists ((3.1) in combination with (3.2) of Boucherie and Van Dijk (1990)). Applying their general expression to our situation yields the following sufficient condition on the arrival rate function,

$$\prod_{k=0}^{g(\mathbf{n}+e_r)-1}(1-\frac{k}{c_0}) = \prod_{k=0}^{g(\mathbf{n})}(1-\frac{k}{c_0}) \quad \forall r, \tag{46}$$

and thus $g(\mathbf{n}+e_r)-1 = g(\mathbf{n})$ for $g(\mathbf{n}) \geq 0$. A reader interested in the derivation of this condition is encouraged to contact the authors. First, we note that (46) is not satisfied for $g(\mathbf{n}) = \sum_{r=1}^{R}(n_r - c_r)^+$ when $n_r > c_r$. Second, condition (46) is satisfied for $g(\mathbf{n}) = \sum_{r=1}^{R} n_r$ and $g(\mathbf{n}) = n_r - c_r$, that is, the state-dependent arrival processes for which Lam (1977) showed that a product form distribution exist. Condition (46) could form a starting point for further research; either for finding arrival functions for which a product form exists or for proving that a product form distribution can not exist for a bigger class of arrival rate functions.

A.7.6 Discussion

The model corresponding to the situation in which the arrival rate to a certain VVT type is influenced by the queue length of all VVT types was proven not to have a product form solution. Systems that do not contain a product form solution are often modified to or approximated by models for which a product form solution is known to exist [Weij et al. (2012), Van Dijk and Van Der Sluis (2009)]. For the system of interest, a product form exists for a state-dependent arrival rate that only depends on the total number of patients or the number of patients of one type. Further research could be done into the accuracy of approximating the system with a queue length dependent arrival rate by these two product form systems.

Blocking in the prior care organisation by the patient waiting for VVT care, was modelled implicitly by a state-dependent arrival rate. Although several papers also use this approach, among which Dallery and Frein (1993) and Oniszczuk (2006), we could find little research that studies the accuracy of this modelling choice. One could study for example how certain state-dependent arrival functions, other than linear, influence the accuracy. Ideas for appropriate state-dependent arrival functions could be obtained from analysing the departure process of a station in which blocking occurs. In general, approximating blocking by state-dependent arrivals is reported to work well if the prior care organisation is close to saturation. This condition is reasonable for our application since hospitals and VVT organisations classically operate under high load and since the consequence of the assumption of a saturated hospital is an overestimation of the arrival rate and thus an overestimation of the necessary capacity in contrast to too little capacity.

The capacities of all VVT organisations were pooled per type of VVT care. The arrival rate to a certain department of each VVT organisation thus depended only on the type of VVT care. However, some patients might refrain from going to an organisation that is located further from Enschede. Thus, the fraction of patients that go to a specific VVT organisation might decrease with the distance of that organisation from Enschede. On the other hand, if the waiting list for this organisation further away is significantly shorter than more patients might be tempted to go there. It would be interesting to incorporate such dependencies on the location of an organisation and its waiting list into the state-dependent arrival rate.