UNIVERSITY OF TWENTE.

UNIVERSITY of TASMANIA

Priority queues and the stationary distribution calculated using tandem fluid queues

Hiska M. Boelema

M.Sc. Thesis

June 2017

Supervisors:
dr. M. M. O'Reilly
dr. ir. W. R. W. Scheinhardt

Preface

Twelve months ago I started looking for a place to finish my Master degree. I wanted to go abroad: The further away, the better! The University of Tasmania offered me a desk and a research project, supervised by Małgorzata O'Reilly. Without thinking twice, I packed my bags and flew to the other side of the world. This thesis concludes the seven-month research I did on the maximum priority process. We constructed a mapping to a tandem fluid queue to derive a stationary distribution for this process.

First of all, I would like to thank Małgorzata O'Reilly for her dedication and helpfulness. Every meeting she took the time to give notes and comments on my work which resulted in a better understanding of the project. I want to give a special thanks to her and ACEMS for making it possible for me to attend the workshop at Uluru and giving me the possibility to give a short talk about my project.

Next, I would like to thank Werner Scheinhardt. Not only for bringing me in contact with Małgorzata O'Reilly, but also for the ideas he gave during the Skype-meetings. His angle on the project and knowledge about fluid queues were invaluable to my research.

Also, I would like to thank Peter Taylor. His idea of mapping the maximum priority process to a tandem fluid queue initiated this research.

Subsequently, I would like to thank my fellow students at my office for giving me an amazing time at the university. The cups of coffee and the bowls of chips were never consumed alone.

Last but not least, I want to give special note of thanks to my parents for supporting my stay in Australia and Rogier Heeg for coming over so I wouldn't be alone for the whole time I've spent in Australia.

I hope you enjoy reading.

Hiska Boelema
Hobart, June 2017

Contents

Chapter 1

Introduction

Traditionally in queueing theory, the way to analyse a priority queue is to assume that every customer upon arrival has a fixed priority, which is class-dependent [8, 10, 14]. Also, no customer from a given class commences service if a customer of a higher class is still present in the queue. However, in situations where a performance target has to be met, such as maximum waiting time of a customer, this model of priority queueing is not satisfactory. There are situations where high priority classes easily meet their performance target in terms of the maximum waiting time of a customer, while lower classes do not meet their targets in terms of maximum waiting time.

For example, for a two-class case, it is possible that the high-class customer almost always starts service within the hour, while the low-class customer only has a 70% chance to start service within an hour. However, the performance target of this queue is such that 95% of the high class customers start service within the hour and the same holds for 80% of the low class customers. Hence, the traditional models of priority queueing do not meet our requirements.

Kleinrock [11] introduced a time-dependent priority queue in 1964. In this paper, results are derived for a delay dependent priority system in which a customer's priority is increasing, from zero, linearly with time in proportion to a rate assigned to the customer's priority class. The advantage of this new priority structure is that it provides a number of degrees of freedom with which to manipulate the relative waiting times for each customer class. Upon a departure, the customer with highest priority in queue (if any) commences service.

In 2013, Stanford, Taylor and Ziedins [18] pointed out that the performance of many queues, particularly in the healthcare and human services sectors, is specified in terms of tails of waiting time distributions for customers of different classes. They use this time-dependent priority queue, which is referred to as the accumulating priority queue in [18], to construct a corresponding stochastic process to derive its waiting time distributions, rather than just the mean waiting times. This stochastic process is referred to as the *maximum priority process*.

We are interested in the stationary distribution at the times of commencement of service of this maximum priority process. Until now, there is no explicit expression for this distribution. We construct a mapping of the maximum priority process to a *tandem fluid queue* [3–7, 15–17] which enables us to find expressions for this stationary distribution using techniques derived in [15, 16] by O'Reilly and Scheinhardt.

Previous work on this topic was done by Dams [9]. However, he was not able to obtain the stationary distribution of the maximum priority process at the times of the commencement of service. This stationary distribution is interesting because it gives information on the maximum waiting time of the customer.

The remainder of this report is organized as follows. In Chapter 2 we describe the accumulating priority queue and the corresponding maximum priority process. We also provide some explanations and diagrams in order to explain the model in detail, for the benefit of the reader. In Chapter 3, we give the definition of the tandem fluid queue and construct the mapping of the maximum priority process to a tandem fluid queue. Chapter 4 describes the numerical treatment of the tandem fluid queue. Moreover, we give an example of how the maximum priority process can be mapped as a tandem fluid queue. We conclude this report with Chapter 5, here we provide comments on our method and suggestions for further research are given.

Chapter 2

The priority process

In this chapter we consider the accumulating priority queue introduced in [18], in which two classes of customers accumulate priority over time at linear and class-dependent rates. We give the details of the construction of this process and describe a related *maximum priority process.* The latter will form the key focus for this report.

2.1 Accumulating priority queue

Here, we give details of the multiclass accumulating priority queue as defined in [18], for convenience of the reader.

Consider a single-server queue with Poisson arrivals such that customers of class i arrive to the queue at some rate $\lambda_i > 0$. Upon arrival to the queue, a customer of class i starts accumulating priority at rate $b_i > 0$. After completion of the service, the server starts serving the customer with the highest accumulated priority, regardless of their class. Note that this means that the queue discipline is not first in, first out (FIFO). That is, when the server selects the next customer to be served, they do not choose the customer who arrived first, but the customer with the highest accumulated priority.

Assume that the single server has a general distribution $B^{(i)}$ of service time with mean $1/\mu_i$ for customer class i. Let $X^{(i)}$ be the random variable recording the service time of customer class i. The distribution function $B^{(i)}$ for a class i customer has a Laplace-Stieltjes transform $\tilde{B}^{(i)}(s) = E(e^{-sX^{(i)}})$, which is defined in the right complex half-plane for at least some s with $Re(s) < 0$. Assume that the inter-arrival times and service processes are mutually independent and that the system is stable.

Let $\mathbf{\Gamma} = \{\Gamma_n; n = 1, 2, \ldots\}$ be the corresponding stochastic process of the inter-arrival times with $\gamma_n = \sum_{k=1}^{n} \Gamma_k$ denoting the time of the n^{th} arrival. Further, let

$\boldsymbol{\chi} = \{\chi(n); n = 1, 2, \ldots\}$ be the process recording the customer classes of the arrivals such that $\chi(n)$ is the customer class of the n^{th} arrival. Also, let $\mathbf{X} = \{X_n; n = 1, 2, \ldots\}$ be the service time of the n^{th} arriving customer.

Now, we define the accumulated priority function $V_n(t)$ by

$$V_n(t) = b_{\chi(n)}[t - \gamma_n]^+, \tag{2.1}$$

where $V_n(t)$ denotes the accumulated priority of the n^{th} customer. Note that $b_{\chi(n)}$ is the rate of the n^{th} arriving customer. Also note that if the n^{th} customer arrived after time t, that is when $\gamma_n > t$, then the accumulating priority at time t is set to 0.

Let $n(m)$ be the function recording the position in the arrival sequence of the m^{th} customer to be served. For example, if the third customer to be served was the fourth arrival then $n(3) = 4$.

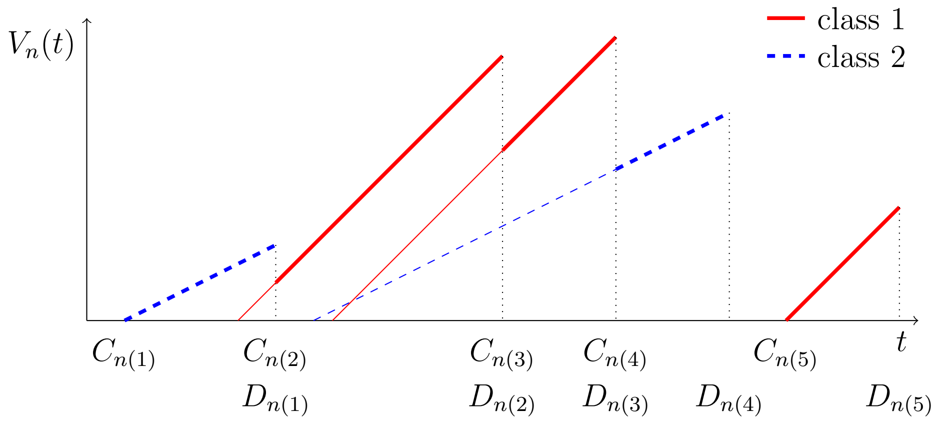

Figure 2.1: An example of the evolution of the accumulating priority function $V_n(t)$.

Let C_n be the time that the n^{th} arrival starts service and D_n be the departure time of this customer, with clearly $D_n = C_n + X_n$. The time that the m^{th} customer commences service is therefore $C_{n(m)}$ and the departure time of this customer is $D_{n(m)}$. After departure of a customer there are two possibilities, the queue is empty, or the queue is non-empty and the customer with the highest priority commences service. In mathematical form this can be written as

$$n(m+1) = \min\{\arg\max_{n \notin \{n(i)\ 1 \leq i \leq m\}} V_n(D_{n(m)})\}, \tag{2.2}$$

where we use the minimum function since it is possible that the set in (2.2) contains more than one element, though the probability of this occurring is 0. An example of the evolution of the accumulating priority function $V_n(t)$ is presented in Figure 2.1. Here, we observe 5 arrivals, with the corresponding start-of-service times $C_{n(m)}$ and departure times $D_{n(m)}$, for $m = 1, 2, \ldots, 5$.

2.2 Maximum priority process

In this section we describe the maximum priority process $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$, as defined in [18], that corresponds to the accumulating priority queue of Section 2.1 with two classes of customers such that $b_1 > b_2$, so that a class 1 customer accumulates priority at a higher rate than class 2 customer. Intuitively, this process records the least upper bounds $M_1(t)$ and $M_2(t)$ of the accumulated priority $V_n(t)$ for customer classes 1 and 2, respectively. The values of $M_1(t)$ and $M_2(t)$ grow at class-dependent rates during service, with $M_1(t)$ always and $M_2(t)$ possibly observing a jump down at the end of the service.

Definition 2.1. *For the two-class accumulating priority queue, the maximum priority process* $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$ *is defined as follows.*

1. *For an empty queue at time* t*, we let* $M_1(t) = M_2(t) = 0$*.*
2. *For a non-empty queue, at the departure times* $\{D_{n(m)}, m = 1, 2, \ldots\}$*, we let*

$$M_1(D_{n(m)}) = \max_{n \notin \{n(k); 1 \leq k \leq m\}} V_n(D_{n(m)}), \tag{2.3}$$

$$M_2(D_{n(m)}) = \min\{M_1(D_{n(m)}), M_2(C_{n(m)} + b_2 X_{n(m)})\}. \tag{2.4}$$

3. *For a non-empty queue during the* m^{th} *service at time* t*, that is for* $t \in [C_{n(m)}, D_{n(m)})$*, for* $i = 1, 2$*, we let*

$$M_i(t) = M_i(C_{n(m)}) + b_i(t - C_{n(m)}). \tag{2.5}$$

By the above definition, at departure times the accumulated priority of the customer with the highest accumulated priority determines $M_1(t)$. This customer is also the person who commences service at that time point. Clearly, the accumulated priority of the customer commencing service must be lower than the accumulated priority of the customer leaving service. Consequently, $M_1(t)$ will always experience jumps down at the departure times.

On the other hand, since $b_1 > b_2$, $M_2(t)$ will experience jump down at a departure time only when the accumulated priority of the customer with the highest accumulated priority is strictly less than the accumulated priority of the customer that is leaving the service.

An example of a sample path of the maximum priority process $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$ is shown in Figure 2.2. We observe the jumps in $M_1(t)$ at all departure times, while this is not always true for $M_2(t)$.

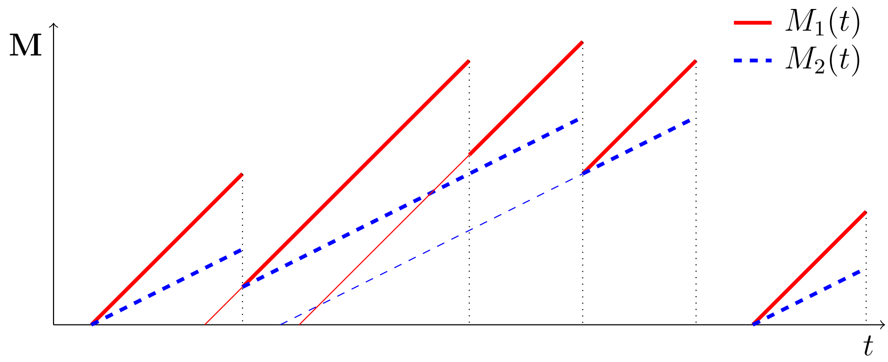

Figure 2.2: The maximum priority process corresponding to Figure 2.1.

2.3 Jumps in $V_n(t)$

Consider the behaviour of the accumulated priority function $V_n(t)$ at the departure times. Denote by E_k the value of the jump in $V_n(t)$ at the k^{th} departure time, defined as

$$E_k = b_{i(k)}X_{n(k)} - b_{i(k+1)}[D_{n(k)} - \gamma_{n(k+1)}]^+ + M_1(C_{n(k)}), \tag{2.6}$$

where $i(k)$ is the customer class corresponding to the k^{th} position in the start-of-service sequence, and $D_{n(k)}$ is the departure time of that customer. Note that $\gamma_{n(k+1)}$ is the arrival time of the customer that commences the $(k+1)^{th}$ service. $M_1(C_{n(k)})$ is the accumulated priority of the k^{th} customer at the beginning of the service and $X_{n(k)}$ is the service time of this customer. See Figure 2.3.

The second term on the right-hand side of equation (2.6) becomes 0 if the queue is empty at the k^{th} departure time. This is due to the fact that in such case the arrival of the $(k+1)^{th}$ customer did not occur yet, and so $\gamma_{n(k+1)} > D_{n(k)}$. In this case we have $V_n(D_{n(k)}) = 0$.

2.4 Jumps in the maximum priority process M

The behaviour of the jumps in the maximum priority process **M** is somewhat similar to that of the jumps in $V_n(t)$. The difference is that the size of the jump affects the behaviour of the process **M**. This is because the maximum priority process **M** records the least upper bounds of both classes of customers in the queue. Also, the jumps in the process **M** will always be jums down.

At the non-departure times t, both variables $M_1(t)$ and $M_2(t)$ are increasing at constant rates, b_1 and b_2, respectively. The jumps down may only occur at the departure times. We consider *three different types of behaviour* at the departure times.

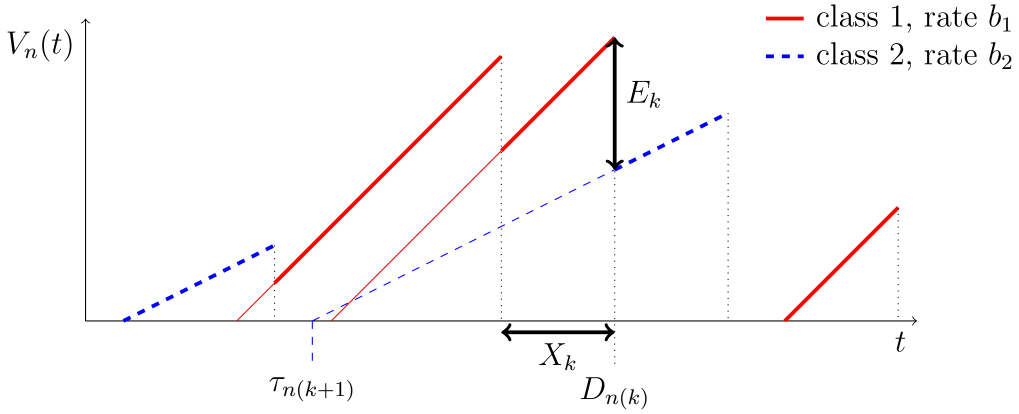

Figure 2.3: Example of the behaviour of the jump down in the accumulating priority process.

Type 1: Priority of customer that commences service is higher then the upper bound of the class 2 customers

We illustrate this in Figure 2.4. In this case, at the departure times t, the least upper bound of class 2 customers, $M_2(t)$, remains unchanged, while the least upper bound of the class 1 customer, $M_1(t)$, jumps down to the maximum accumulated priority of the customers still in queue. Therefore, the difference $Z(t) = M_1(t) - M_2(t)$ is equal precisely to the difference of the accumulated priority of the customer that commences service and $M_2(t)$, and so $Z(t) > 0$.

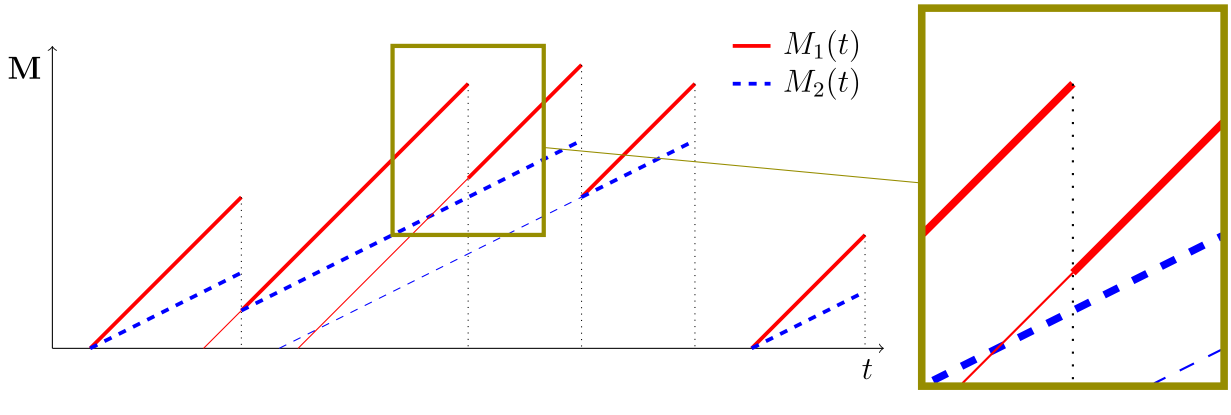

Figure 2.4: Type 1 jump in the maximum priority process **M**.

Type 2: Priority of customer that commences service is smaller than the upper bound of the class 2 customers

We illustrate this in Figure 2.5. In this case, at the departure times t, there are some customers still present in the queue after the service of the $n(k)^{th}$ customer was completed, with a priority smaller than $M_2(t)$. Consequently, both upper bounds, $M_1(t)$ and $M_2(t)$, are set to the value of the priority of the customer that commences service, and so $M_1 = M_2 > 0$ and $Z(t) = 0$.

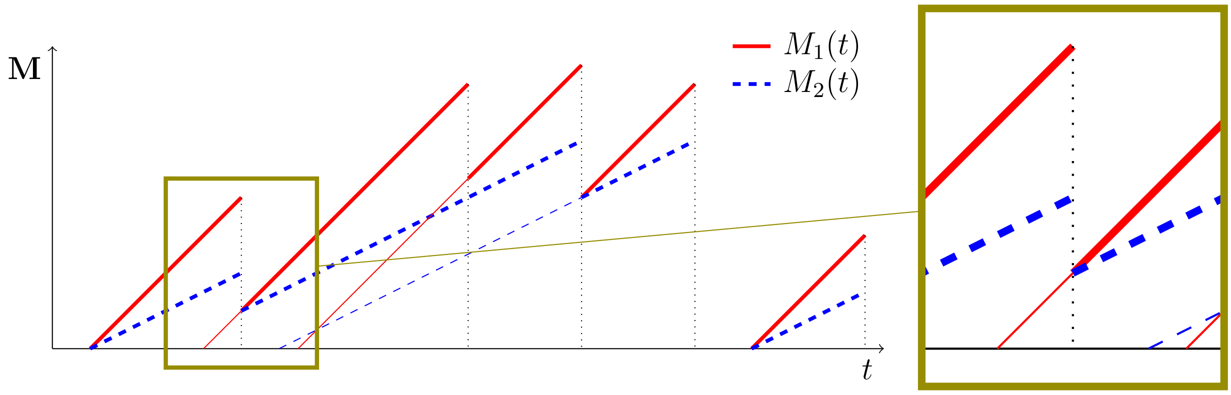

Figure 2.5: Type 2 jump in the maximum priority process **M**.

Type 3: There are no customers in the queue after the completion of the service

We illustrate this in Figure 2.6. At the departure times t, the queue is empty and the least upper bound for both classes is $M_1(t) = M_2(t) = 0$ and $Z(t) = 0$, and these quantities will remain zero until the next arrival.

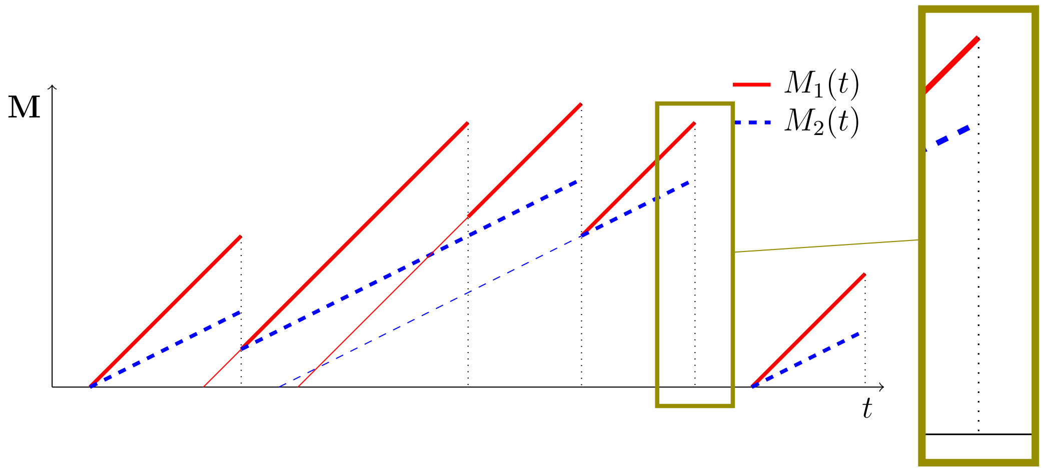

Figure 2.6: Type 3 jump in the maximum priority process **M**.

Chapter 3

The tandem fluid model

Consider the two-class maximum priority process $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$ of Section 2.2 as defined in [18]. We will construct a mapping of this process to a tandem fluid queue $\{(\varphi(t), X(t), Y(t)); t \geq 0\}$, which was analysed in [15, 16]. The goal is to derive a stationary distribution for the two-class maximum priority process $\mathbf{M}$ using the results in [15, 16].

3.1 Tandem fluid queue

Consider two fluid queues, collecting fluid in buffers X and Y. The level variables recording the content of the buffers at time t are given by $X(t)$ and $Y(t)$, respectively. These level variables are driven by the same background continuous-time Markov chain, denoted by $\{\varphi(t); t \geq 0\}$ with some finite state space $\mathcal{S}$ and irreducible generator $\mathbf{T}$.

We partition the state space $\mathcal{S}$ as $\mathcal{S} = \mathcal{S}_+ \cup \mathcal{S}_-$, where $\mathcal{S}_+ = \{i : r_i > 0\}$, $\mathcal{S}_- = \{i : r_i < 0\}$, and refer to $i \in \mathcal{S}_+$ as the up-phases and $i \in \mathcal{S}_-$ as the down-phases.

The first level variable $X(t)$ has a lower boundary at level 0, and depends on $\varphi(t)$ and real-valued fluid rates r_i, for all $i \in \mathcal{S}$, as follows. When the buffer is non-empty, the level in the buffer changes at rates r_i. However, when the buffer is empty and $i \in \mathcal{S}_-$, the level of the fluid stays 0. That is,

$$\frac{d}{dt}X(t) = r_{\varphi(t)} \qquad \text{when } X(t) > 0, \tag{3.1}$$

$$\frac{d}{dt}X(t) = \max(0, r_{\varphi(t)}) \qquad \text{when } X(t) = 0. \tag{3.2}$$

The second fluid queue $Y(t)$ depends on $X(t)$, $\varphi(t)$ and rates c_i, for all $i \in \mathcal{S}$, as follows. When the first buffer is non-empty, the level in the second buffer changes at non-negative fluid rates $\hat{c}_i$. However, when the first buffer is empty, the level in the second buffer changes at negative fluid rates $\check{c}_i$. That is,

$$\frac{d}{dt}Y(t) = \hat{c}_{\varphi(t)} \geq 0 \qquad \text{when } X(t) > 0, \tag{3.3}$$

$$\frac{d}{dt}Y(t) = \check{c}_{\varphi(t)} < 0 \qquad \text{when } X(t) = 0, Y(t) > 0, \tag{3.4}$$

$$\frac{d}{dt}Y(t) = \hat{c}_{\varphi(t)} \cdot 1\{\varphi(t) \in S_+\} \qquad \text{when } X(t) = 0, Y(t) = 0. \tag{3.5}$$

Remark 1. *The tandem fluid queue analysed in [15, 16] only considers $\hat{c}_{\varphi(t)} > 0$, however the result still holds if $\hat{c}_{\varphi(t)} \geq 0$.*

We denote such defined process as $\{(\varphi(t), X(t), Y(t)); t \geq 0\}$. The stationary distribution of this tandem fluid queue was derived in [15, 16]. Next, we map the two-class maximum priority process $\mathbf{M}$ into $\{(\varphi(t), X(t), Y(t)); t \geq 0\}$.

3.2 Mapping of the maximum priority process with exponential service times to a tandem fluid queue

We map the maximum priority process $\mathbf{M}$ described in Section 2.2 into a tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ as follows.

First, let $\{\varphi(t); t \geq 0\}$ be some background continuous-time Markov chain with state space $\mathcal{S} = \{+, -\}$, where $+$ is referred to as the up-phase, and $-$ as the down-phase. In our mapping, these phases correspond to the service time and the jump down in the process $\mathbf{M}$ of Section 2.2, respectively. The generator of this chain is assumed to be

$$\mathbf{T} = \begin{bmatrix} -\mu & \mu \\ 1 & -1 \end{bmatrix}. \tag{3.6}$$

Note that the distribution of the time spent in phase $+$ is equal to the distribution of the service time in the process $\mathbf{M}$.

Next, with the variables $M_1(t)$ and $M_2(t)$ as described in Section 2.2, let the process $\{(Z(t), M_2(t)); t \geq 0\}$ be an adjusted representation of the maximum priority process $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$, where

$$Z(t) = M_1(t) - M_2(t), \tag{3.7}$$

and define the variables $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ with the following desired properties.

1. To map the behaviour of the process $\mathbf{M}$ *during service times*, we assume that when $\varphi(t) = +$, then $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ increase at rates $(b_1 - b_2)$ and b_2, respectively. This means the distribution of shift in $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ is equivalent to that of $Z(t)$ and $M_2(t)$, respectively, during service times.

2. To map the behaviour of the process $\mathbf{M}$ *at the end of service times*, that is when we observe the jumps, we assume that when $\varphi(t) = -$, the variables $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ will change at some appropriate rates, such that the distributions of jumps in $Z(t)$ and $M_2(t)$ at the end of the service times are the same as the distributions of shift in $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ at the end of the down phase $-$, respectively. We choose these rates according to part 2 of Theorem 3.2 in [18].

For completeness, we state part 2 of Theorem 3.2 in [18] below.

Theorem 3.1. *The accumulated priorities $\{V_k(t); k = 1, 2, \ldots\}$ of all customers still present in the queue are distributed as a Poisson process with piecewise constant rates, with rate zero on the interval $[M_1(t), \infty)$, rate $\frac{\lambda_1}{b_1}$ on the interval $[M_2(t), M_1(t))$, and rate $\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)$ on the interval $[0, M_2(t))$.*

To this end, we assume

$$\frac{d}{dt}\widetilde{Z}(t) = b_1 - b_2 \qquad \text{when } \widetilde{Z}(t) > 0, \varphi(t) \in S_+, \tag{3.8}$$

$$\frac{d}{dt}\widetilde{Z}(t) = -\left(\frac{\lambda_1}{b_1}\right) \qquad \text{when } \widetilde{Z}(t) > 0, \varphi(t) \in S_-, \tag{3.9}$$

$$\frac{d}{dt}\widetilde{Z}(t) = \max(0, b_1 - b_2) \qquad \text{when } \widetilde{Z}(t) = 0, \tag{3.10}$$

and

$$\frac{d}{dt}\widetilde{M}_2(t) = b_2 \qquad \text{when } \widetilde{Z}(t) > 0, \varphi(t) \in S_+, \tag{3.11}$$

$$\frac{d}{dt}\widetilde{M}_2(t) = 0 \qquad \text{when } \widetilde{Z}(t) > 0, \varphi(t) \in S_-, \tag{3.12}$$

$$\frac{d}{dt}\widetilde{M}_2(t) = -\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right) \qquad \text{when } \widetilde{Z}(t) = 0, \widetilde{M}_2(t) > 0, \tag{3.13}$$

$$\frac{d}{dt}\widetilde{M}_2(t) = b_2 \cdot 1\{\varphi(t) \in S_+\} \qquad \text{when } \widetilde{Z}(t) = 0, \widetilde{M}_2(t) = 0. \tag{3.14}$$

Below we show that the desired properties 1-2 are met by assuming the rates (3.8)–(3.14).

Lemma 3.2. *The following properties stated below hold.*

1. *The distributions of shift in $Z(t)$ and $M_2(t)$ during service times are equivalent to that of shift in $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ during up-phase $+$, respectively.*
2. *The distributions of jumps in $Z(t)$ and $M_2(t)$ at the end of the service times are equivalent to that of shift in $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ at the end of the down phase $-$, respectively.*

Proof. First, we prove property 1.

By assumption, the service time $X_{n(m)}$ of a customer in the maximum priority process $\mathbf{M}$ is exponentially distributed with parameter μ, for all $m = 1, 2, \ldots$. The rate at which $Z(t)$ increases during service time is $(b_1 - b_2)$, while $M_2(t)$ increases at rate b_2. That is, during the time interval $[C_{n(m)}, D_{n(m)})$, for $m = 1, 2, \ldots$, we have

$$Z(D_{n(m)}) - Z(C_{n(m)}) \sim Exp((b_1 - b_2) \cdot \mu), \tag{3.15}$$

$$M_2(D_{n(m)}) - M_2(C_{n(m)}) \sim Exp(b_2 \cdot \mu). \tag{3.16}$$

That is, the distribution of shift in $Z(t)$ during service time is exponentially distributed with parameter $((b_1 - b_2) \cdot \mu)$, while the distribution of shift in $M_2(t)$ during service time is exponentially distributed with parameter $(b_2 \cdot \mu)$.

Let τ_i be the time spent in phase i before leaving phase i. Then by the choice of generator $\mathbf{T}$ in the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ as described in 3.6, the time spent τ_+ in up-phase $+$ is exponentially distributed with parameter μ. The rate at which $\widetilde{Z}(t)$ increases during time τ_+ is $(b_1 - b_2)$, while $\widetilde{M}_2(t)$ increases at rate b_2. Therefore, at the end of time τ_+, we have

$$\widetilde{Z}(\tau_+) \sim Exp((b_1 - b_2) \cdot \mu), \tag{3.17}$$

$$\widetilde{M}_2(\tau_+) \sim Exp(b_2 \cdot \mu). \tag{3.18}$$

That is, the distribution of shift in $\widetilde{Z}(t)$ during up-phase is exponentially distributed with parameter $((b_1 - b_2) \cdot \mu)$, while the distribution of shift in $\widetilde{M}_2(t)$ during up-phase is exponentially distributed with parameter $(b_2 \cdot \mu)$.

By (3.15)–(3.18), the distribution of shift in $\widetilde{Z}(t)$ and $\widetilde{M}_2(t)$ in up-phase is equivalent to that of $Z(t)$ and $M_2(t)$ during service times, which proves property 1.

It is left to prove property 2.

Let $D^-_{n(m)}$ be the moment at the end of service time just before the jump. By Theorem 3.1, the accumulated priorities $\{V_k(t); k = 1, 2, \ldots\}$ of all customers still present in the queue at the end of service times $D^-_{n(m)}$ are distributed as a Poisson process with piecewise constant rates in the maximum priority process $\mathbf{M}$. The constant rate in the interval $[M_2(D^-_{n(m)}), M_1(D^-_{n(m)}))$ is $\frac{\lambda_1}{b_1}$. The constant rate in the interval $[0, M_2(D^-_{n(m)}))$ is $\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)$ if $Z(t) = 0$ for $t \in [D^-_{n(m)}, D_{n(m)}]$. Therefore, for $t \in [D^-_{n(m)}, D_{n(m)})$ and

$V_k(t) \in [M_2(t), M_1(t))$, we have

$$Z(D_{n(m)}) - Z(D^-_{n(m)}) \sim Exp\left(\frac{\lambda_1}{b_1}\right), \tag{3.19}$$

$$M_2(D_{n(m)}) - M_2(D^-_{n(m)}) = 0, \tag{3.20}$$

and for $V_k(t) \in [0, M_2(t))$, we have

$$Z(D_{n(m)}) - Z(D^-_{n(m)}) = 0, \tag{3.21}$$

$$M_2(D_{n(m)}) - M_2(D^-_{n(m)}) \sim Exp\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right). \tag{3.22}$$

That is, the distribution of shift in $Z(t)$ during jump time $D_{n(m)}$ when $V_k(t) \in [M_2(t), M_1(t))$, is exponentially distributed with parameter $\left(\frac{\lambda_1}{b_1}\right)$, while the distribution of shift in $M_2(t)$ during jump time when $V_k(t) \in [M_2(t), M_1(t))$, is zero. The distribution of shift in $Z(t)$ during jump time $D_{n(m)}$ when $V_k(t) \in [0, M_2(t))$, is zero, while the distribution of shift in $M_2(t)$ during jump time when $V_k(t) \in [M_2(t), M_1(t))$, is exponentially distributed with parameter $\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)$.

The time τ_- spent in down-phase $-$ in the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ is exponential distributed with parameter 1, by (3.6). The rate at which $\widetilde{Z}(t)$ changes during down-phase, that is for $t \in [0, \tau_-)$, when $V_k \in [M_2(t), M_1(t))$ is $\left(\frac{\lambda_1}{b_1}\right)$. The rate at which $\widetilde{Z}(t)$ changes when $V_k(t) \in [0, M_2(t))$ is $\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)$ if $\widetilde{Z}(t) = 0$. Therefore, at the end of time τ_+, when $V_k(t) \in [M_2(t), M_1(t))$, we have

$$\widetilde{Z}(\tau_-) \sim Exp\left(\frac{\lambda_1}{b_1}\right), \tag{3.23}$$

$$\widetilde{M_2}(\tau_-) = 0, \tag{3.24}$$

and for $V_k \in [0, M_2(t))$, we have

$$\widetilde{Z}(\tau_-) = 0, \tag{3.25}$$

$$\widetilde{M_2}(\tau_-) \sim Exp\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right). \tag{3.26}$$

That is, the distribution of shift in $\widetilde{Z}(t)$ at the end of service time τ_- when $V_k(t) \in [M_2(t), M_1(t))$, is exponentially distributed with parameter $\left(\frac{\lambda_1}{b_1}\right)$, while the distribution of shift in $M_2(t)$ at the end of service time when $V_k(t) \in [M_2(t), M_1(t))$, is zero. The distribution of shift in $Z(t)$ at the end of service time when $V_k(t) \in [0, M_2(t))$, is zero, while the distribution of shift in $M_2(t)$ at the end of service time when $V_k(t) \in [M_2(t), M_1(t))$, is exponentially distributed with parameter $\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)$.

By (3.23)–(3.26), the distribution of shift in $\widetilde{Z}(t)$ and $\widetilde{M_2}(t)$ at the end of down-phase is equivalent to that of $Z(t)$ and $M_2(t)$ at jump times, which proves property 2.

Now, consider the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ observed only when $\varphi(t) = +$. Then by (3.15)–(3.26), the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ is a mapping of the process $\{(Z(t), M_2(t)); t \geq 0\}$. Indeed, the time spent in up-phase and down-phase in the $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ observed only when $\varphi(t) = +$ is now equal to that of the maximum priority process $\{(Z(t), M_2(t)); t \geq 0\}$. □

Recall that the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ is a tandem fluid queue as described in Section 3.1. Here, $\widetilde{Z}(t)$ corresponds to $X(t)$ and $\widetilde{M_2}(t)$ to $Y(t)$. Further, the rates $(b_1 - b_2)$ and $\left(-\left(\frac{\lambda_1}{b_1}\right)\right)$ correspond to the rates $r_{\varphi(t)}$, the rates b_2 and 0 correspond to $\hat{c}_{\varphi(t)}$ and the rate $\left(-\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)\right)$ corresponds to $\check{c}_{\varphi(t)}$.

The stationary distribution of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ can be readily derived using the results in [15, 16]. Below, we explain how to derive the stationary distribution of the the maximum priority process $\mathbf{M}$ using the stationary distribution of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$.

3.3 The stationary distribution of M

We consider the stationary distribution of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ defined in Section 3.2, and the stationary distribution of its embedded chain $\{J_k; k = 1, 2, \ldots\}$ observed at the moments of transition to up-phase $+$. We show a relationship between these processes and the stationary distribution of $\{(Z(t), M_2(t)); t \geq 0\}$.

First, we give the intuitive behaviour of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$. The process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ alternates between two types of behaviour.

- (i) *The first buffer $\widetilde{Z}$ is empty*, while $\widetilde{M_2}(t)$ is decreasing, possibly until the second buffer $\widetilde{M_2}$ becomes empty, and $\varphi(t) \in \mathcal{S}_-$. This period ends when $\varphi(t)$ makes a transition from $\mathcal{S}_-$ to $\mathcal{S}_+$, at which a type (ii) behaviour begins.
- (ii) *The first buffer $\widetilde{Z}$ is non-empty*, with $\widetilde{M_2}(t)$ non-decreasing, while $\widetilde{Z}(t)$ can either increase or decrease. This period ends when the first buffer $\widetilde{Z}$ becomes empty with $\varphi(t) \in \mathcal{S}_-$.

Denote by z and m_2 as the values of level variables $\widetilde{Z}(t)$ and $\widetilde{M_2}(t)$, respectively. Observe that in stationarity, the process can not be at $m_2 = 0$, $z > 0$. Indeed, if a type (ii) behaviour starts with $z = 0$, $m_2 = 0$, that is when both buffers are empty, then levels $\widetilde{Z}(t)$ and $\widetilde{M}(t)$ increase at rates $(b_1 - b_2)$ and b_2, respectively. The slope of leaving the origin $(z, m_2) = (0, 0)$ is $b_2/(b_1 - b_2) > 0$, and any future slope of (z, m_2) cannot be lower than this value. Consequently, after leaving the origin $(z, m_2) = (0, 0)$, the values $\{(z, m_2); m_2 < z \cdot b_2/(b_1 - b_2)\}$ can never occur.

As a result of this, the stationary distribution of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ described in Section 3.2 will have the following components.

- A *probability point mass* at $(0, 0)$, denoted $\widetilde{\mathbf{p}}(0, 0) = [\widetilde{p}_+(0, 0) \;\; \widetilde{p}_-(0, 0)]$.
- A *one-dimensional density*, denoted $\widetilde{\boldsymbol{\pi}}(0, m_2) = [\widetilde{\pi}_+(0, m_2) \;\; \widetilde{\pi}_-(0, m_2)]$ for $m_2 > 0$.
- A *two-dimensional density*, denoted $\widetilde{\boldsymbol{\pi}}(z, m_2) = [\widetilde{\pi}_+(z, m_2) \;\; \widetilde{\pi}_-(z, m_2)]$ for the set $\{(z, m_2); z > 0, m_2 > z \cdot b_2/(b_1 - b_2)\}$.

In rigorous terms, we define the components of the stationary distribution of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$, when $\varphi(t) = -$, as

$$\widetilde{\pi}_-(z, m_2) = \lim_{t \to \infty} \frac{\partial^2 P(\varphi(t) = -, \widetilde{Z}(t) \leq z, \widetilde{M}_2(t) \leq m_2)}{\partial z \partial m_2}, \tag{3.27}$$

$$\widetilde{\pi}_-(0, m_2) = \lim_{t \to \infty} \frac{\partial P(\varphi(t) = -, \widetilde{Z}(t) = 0, \widetilde{M}_2(t) \leq m_2)}{\partial m_2}, \tag{3.28}$$

$$\widetilde{p}_-(0, 0) = \lim_{t \to \infty} P(\varphi(t) = -, \widetilde{Z}(t) = 0, \widetilde{M}_2(t) = 0). \tag{3.29}$$

Similar expressions follow when $\varphi(t) = +$. Note that $\widetilde{p}_+(0, 0) = 0$ and $\widetilde{\pi}_+(0, m_2) = 0$, because at the moment of a transition from down-phase $-$ to up-phase $+$, the fluid level of buffer Z starts to increase, which implies that $z > 0$.

Next, we introduce the following embedded Markov chain of the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$, observed at the moments of transition to up-phase $+$. Note that the end of a type (i) behaviour in the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ coincides with the moment right after jump of the process $\{(Z(t), M_2(t)); t \geq 0\}$. We will show in Theorem 3.4 below that the stationary distributions of these processes at these times are equivalent.

Let ν_k for $k = 1, 2, \ldots$ be the k^{th} time that $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ transitions to an up-phase $+$. Also, let η_k for $k = 1, 2, \ldots$ be the k^{th} time that $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ transitions to the down-phase $-$.

Let $J_k = (\varphi(\nu_k), \widetilde{Z}(\nu_k), \widetilde{M}_2(\nu_k))$ and note that $\{J_k; k = 1, 2, \ldots\}$ is a discrete-time Markov chain with discrete/continuous state space, such that state J_k at time k records the position of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ at times ν_k. That is, $\widetilde{Z}(\nu_k)$ and $\widetilde{M}_2(\nu_k)$ are the contents of the first buffer and second buffer, respectively, at the beginning of the k^{th} up-phase. This is equal to the content at the end of the $(k-1)^{th}$ down-phase. A corresponding sample path is drawn in Figure 3.1.

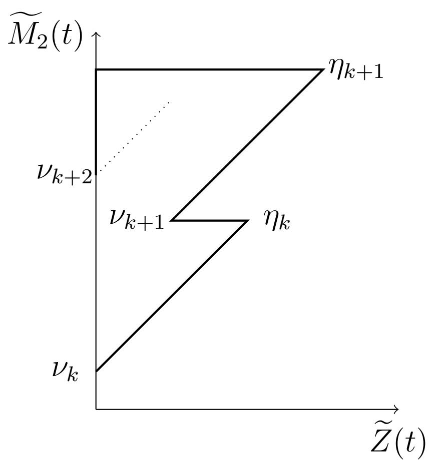

Figure 3.1: Sample path of the the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ and the discrete time points ν_k and η_k.

We define the components of the stationary distribution of the process $\{J_k; k = 1, 2, \ldots\}$ as a probability point mass, a one-dimensional density and a two-dimensional density in a manner similar to that of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$. That is,

$$\widehat{\pi}_-(z, m_2) = \lim_{k \to \infty} \frac{\partial^2 P(\varphi(\nu_k) = -, \widetilde{Z}(\nu_k) \leq z, \widetilde{M_2}(\nu_k) \leq m_2)}{\partial z \partial m_2}, \tag{3.30}$$

$$\widehat{\pi}_-(0, m_2) = \lim_{k \to \infty} \frac{\partial P(\varphi(\nu_k) = -, \widetilde{Z}(\nu_k) = 0, \widetilde{M_2}(\nu_k) \leq m_2)}{\partial m_2}, \tag{3.31}$$

$$\widehat{p}_-(0, 0) = \lim_{k \to \infty} P(\varphi(\nu_k) = -, \widetilde{Z}(\nu_k) = 0, \widetilde{M_2}(\nu_k) = 0). \tag{3.32}$$

Further, consider the stationary distribution of the process $\{(Z(t), M_2(t)); t \geq 0\}$ embedded right before commencement of service times $C^-_{n(k)}$, referred to as $H_k = (Z(C^-_{n(k)}), M_2(C^-_{n(k)}))$. Note that $\{H_k; k = 1, 2, \ldots\}$ is a discrete-time Markov chain such that state H_k at time k records the position of $\{(Z(t), M_2(t)); t \geq 0\}$ at times $C^-_{n(k)}$. The components of the stationary distribution of this process H_k are,

$$\pi(z, m_2) = \lim_{k \to \infty} \frac{\partial^2 P(Z(C^-_{n(k)}) \leq z, M_2(C^-_{n(k)}) \leq m_2)}{\partial z \partial m_2}, \tag{3.33}$$

$$\pi(0, m_2) = \lim_{k \to \infty} \frac{\partial P(Z(C^-_{n(k)}) = 0, M_2(C^-_{n(k)}) \leq m_2)}{\partial m_2}, \tag{3.34}$$

$$p(0, 0) = \lim_{k \to \infty} P\left(Z(C^-_{n(k)}) = 0, M_2(C^-_{n(k)}) = 0\right). \tag{3.35}$$

Now, we prove that the stationary distribution of the process $\{H_k; k = 1, 2, \ldots\}$, defined as the stationary distribution of the process $\{(Z(t), M_2(t)); t \geq 0\}$ embedded right before commencement of service times, is the same as the stationary distribution of the process $\{J_k; k = 1, 2, \ldots\}$.

Theorem 3.3. *For the stationary distributions of the embedded Markov chains H_k and J_k, we have for all $z > 0$ and $m_2 > 0$,*

$$\begin{aligned} \widehat{\pi}_-(z,m_2) &= \pi(z,m_2), & (3.36)\\ \widehat{\pi}_-(0,m_2) &= \pi(0,m_2), & (3.37)\\ \widehat{p}_-(0,0) &= p(0,0). & (3.38) \end{aligned}$$

Proof. From Lemma 3.2, in particular (3.15)–(3.26), it follows that the stationary distribution of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ at the end of down-phase $-$ is equal to that of $\{(Z(t), M_2(t)); t \geq 0\}$ after jump times for both buffers. After a jump there are two possibilities.

- The queue is non-empty. This implies that a customer commences service and the down-phase ends.
- The queue is empty. This implies that both buffers are empty, i.e. there is no shift in both the buffers until arrival of a new customer. Upon arrival of a new customer the down-phase ends.

In other words, the stationary distribution of H_k is equal to that of J_k. Hence, (3.36)–(3.38) are true. □

Also, we prove that the stationary distribution of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ at the end of the down-phase, also referred to as the process $\{J_k; k = 1, 2, \ldots\}$, is the same as the stationary distribution of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ during the down-phase.

Theorem 3.4. *For the stationary distributions of the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ and its embedded Markov chain J_k at times ν_k, we have for all $z > 0$ and $m_2 > 0$,*

$$\begin{aligned} \widehat{\pi}_-(z,m_2) &= \widetilde{\pi}_-(z,m_2)/\beta, & (3.39)\\ \widehat{\pi}_-(0,m_2) &= \widetilde{\pi}_-(0,m_2)/\beta, & (3.40)\\ \widehat{p}_-(0,0) &= \widetilde{p}_-(0,0)/\beta, & (3.41) \end{aligned}$$

where β, the probability being in down-phase, is a normalizing constant given by

$$\beta = \widetilde{p}_-(0,0) + \int_{m_2=0}^{\infty} \widetilde{\pi}_-(0,m_2)dm_2 + \int_{z=0}^{\infty}\int_{m_2=0}^{\infty} \widetilde{\pi}_-(z,m_2)dzdm_2, \qquad (3.42)$$

which alternatively can be calculated using the generator of the continuous-time Markov chain $\varphi(t)$.

Proof. *Intuition:* the time spent in down-phase of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ is exponentially distributed. We use the memoryless property of this distribution to show that the moment of transition to the up-phase does not depend on how much time has elapsed already in the down-phase.

First, we establish expressions for the stationary distribution at the end of the down-phase of the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$, i.e. the process $\{J_k; k = 1, 2, \ldots\}$. The content of the buffers $\widetilde{Z}(t)$ and $\widetilde{M}_2$ at the end of the down-phase is

$$\widetilde{Z}(\nu_{k+1}) = [\widetilde{Z}(\eta_k) - |r_-| \cdot \tau_{-(k)}]^+, \tag{3.43}$$

$$\widetilde{M}_2(\nu_{k+1}) = \widetilde{M}_2(\eta_k) - D_k, \tag{3.44}$$

respectively, where $\tau_{-(k)}$ denotes the time spent in the k^{th} down-phase, $r_- = -\left(\frac{\lambda_1}{b_1}\right) < 0$, and D_k is given by

$$D_k = \begin{cases} 0 & \text{for } \tau_{-(k)} < \frac{\widetilde{Z}(\eta_k)}{|r_-|}, & (3.45) \\ |\check{c}_-| \left(\tau_{-(k)} - \dfrac{\widetilde{Z}(\eta_k)}{|r_-|}\right) & \text{for } \frac{\widetilde{Z}(\eta_k)}{|r_-|} \leq \tau_{-(k)} < \frac{\widetilde{Z}(\eta_k)}{|r_-|} + \frac{\widetilde{M}_2(\eta_k)}{|\check{c}_-|}, & (3.46) \\ \widetilde{M}_2(\eta_k) & \text{for } \tau_{-(k)} \geq \frac{\widetilde{Z}(\eta_k)}{|r_-|} + \frac{\widetilde{M}_2(\eta_k)}{|\check{c}_-|}. & (3.47) \end{cases}$$

Note that $\nu_k = \eta_k + \tau_{-(k)}$, and that we express the fluid level of the buffer at the end of the down-phase in terms of the fluid level at the beginning of the down-phase. The three cases in D_k refer to the the three types of jumps in the maximum priority process $\mathbf{M}$ described in Section 2.4.

For case (3.45), if the time spent in down-phase is less than $\frac{\widetilde{Z}(\eta_k)}{|r_-|}$, that is the time that is required to empty the first buffer, nothing will happen to the second buffer $\widetilde{M}_2$. This refers to a type 1 jump described in Section 2.4. For case (3.46), if the time spent in down-phase is more than $\frac{\widetilde{Z}(\eta_k)}{|r_-|}$, but less than $\frac{\widetilde{Z}(\eta_k)}{|r_-|} + \frac{\widetilde{M}_2(\eta_k)}{|\check{c}_-|}$, which is the time that is required to empty both the buffers, than the second buffer will decrease in level but will not reach zero. Recall that the first buffer has to be empty before the level of the second starts decreasing. Therefore, the term $\left(\tau_{-(k)} - \frac{\widetilde{Z}(\eta_k)}{|r_-|}\right)$ is the excess time after the first buffer is empty. The rate of decrease in the second buffer in this excess time $\check{c}_- = -\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right)$. This refers to a type 2 jump described in Section 2.4. For case (3.47), if the time spent in down-phase is more than $\frac{\widetilde{Z}(\eta_k)}{|r_-|} + \frac{\widetilde{M}_2(\eta_k)}{|\check{c}_-|}$, than both buffers will be empty at the end of down-phase and the corresponding fluid levels are zero for both buffers. This refers to a type 3 jump described in Section 2.4.

Further, for the distribution during the down-phase we can create a similar expression. Define $\eta_k + \tau^*_{-(k)}$ as some time during the k^{th} down-phase, where $\tau^*_{-(k)}$ represents some residual time. The content of the buffers at that time point are

$$\widetilde{Z}(\nu_{k+1}) = [\widetilde{Z}(\eta_k) - |r_-| \cdot \tau^*_{-(k)}]^+ \tag{3.48}$$

$$\widetilde{M}_2(\nu_{k+1}) = \widetilde{M}_2(\eta_k) - D^*_k, \tag{3.49}$$

where D_k^* is given by

$$D_k^* = \begin{cases} 0 & \text{for } \tau^*_{-(k)} < \frac{\widetilde{Z}(\eta_k)}{|r_-|}, & (3.50) \\ |\check{c}_-| \left(\tau^*_{-(k)} - \dfrac{\widetilde{Z}(\eta_k)}{|r_-|} \right) & \text{for } \frac{\widetilde{Z}(\eta_k)}{|r_-|} \leq \tau^*_{-(k)} < \frac{\widetilde{Z}(\eta_k)}{|r_-|} + \frac{\widetilde{M_2}(\eta_k)}{|\check{c}_-|}, & (3.51) \\ \widetilde{M_2}(\eta_k) & \text{for } \tau^*_{-(k)} \geq \frac{\widetilde{Z}(\eta_k)}{|r_-|} + \frac{\widetilde{M_2}(\eta_k)}{|\check{c}_-|}. & (3.52) \end{cases}$$

The expressions for the distribution of J_k are similar to that of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ during down-phase, except for the time spent in down-phase $\tau_{-(k)}$ and $\tau^*_{-(k)}$, respectively. For these random variables, we have

$$\tau_{-(k)} \sim Exp(\mu), \tag{3.53}$$
$$\tau^*_{-(k)} \sim Exp(\mu). \tag{3.54}$$

Hence, the stationary distribution of J_k is follows the same distribution as to that of $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ during down-phase. □

Theorem 3.3 and Theorem 3.4 hold, therefore we have the following result.

Corollary 1. *The stationary distribution of the process* $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ *in down-phase − is equal up to a factor* β*, given by (3.42), to the stationary distribution of the process* $\{(Z(t), M_2(t)); t \geq 0\}$ *embedded at commencement of service times, that is*

$$\pi_-(z, m_2) = \widetilde{\pi}_-(z, m_2)/\beta, \tag{3.55}$$
$$\pi_-(0, m_2) = \widetilde{\pi}_-(0, m_2)/\beta, \tag{3.56}$$
$$p_-(0, 0) = \widetilde{p}_-(0, 0)/\beta. \tag{3.57}$$

3.4 Mapping of the maximum priority process with phase-type distributed service times to a tandem fluid queue

We generalize the results of Section 3.2 to the case that the service time of the process $\mathbf{M}$ has a phase-type distribution. We map the maximum priority process $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$ of Section 2.2 into a tandem fluid queue $\{(\varphi(t), X(t), Y(t)); t \geq 0\}$ of Section 3.1. The class of phase-type distributions is dense in the field of all positive-valued distributions, that is, it can be used to approximate any positive-valued distribution.

Definition 3.5. *Consider a continuous-time Markov chain* $\{\varphi(t); t \leq 0\}$ *with state space* $\mathcal{S} = \{0, 1, \ldots, m\}$ *where* m *is an absorbing state, and generator*

$$\mathbf{T} = \begin{bmatrix} \mathbf{Q} & \mathbf{q} \\ \mathbf{0} & 0 \end{bmatrix}, \tag{3.58}$$

where $\mathbf{Q}$ *is the matrix of transition rates between non-absorbing states,* $\mathbf{q}$ *is the column vector of transition rates from non-absorbing states to the absorbing state* m*. Note that* $\mathbf{q} = -\mathbf{Q}\mathbf{1}$*. Further, let the process have an initial probability distribution vector* $\boldsymbol{\alpha} = [\alpha_i]$ *of starting in any of the non-absorbing states such that* $\alpha_i = P(\varphi(0) = i)$*. Let* X *be a random variable which records the time until absorption to state* m*. We say that* X *follows phase-type distribution with parameters* $(\boldsymbol{\alpha}, \mathbf{Q}, \mathbf{q})$ *and write*

$$X \sim PH(\boldsymbol{\alpha}, \mathbf{Q}, \mathbf{q}). \tag{3.59}$$

We now construct a tandem fluid queue $\{(\varphi(t), X(t), Y(t)); t \geq 0\}$ with the property that the time spent in the set of up-phases $\mathcal{S}_+$ until a transition to the down-phase $-$, which models service times in the maximum priority process $\mathbf{M} = \{(M_1(t), M_2(t)); t \geq 0\}$, follows phase-type distribution.

Consider the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ described in Section 3.2 by (3.8)–(3.14). Let $\{\varphi(t); t \geq 0\}$ be some background continuous-time Markov chain with state space $\mathcal{S} = \mathcal{S}_+ \cup \{-\}$, where $\mathcal{S}_+$ is referred to as the set of up-phases, and $-$ as the single down-phase. For every state in $\mathcal{S}_+$, $\frac{d}{dt}\widetilde{Z}(t) = b_1 - b_2$ and $\frac{d}{dt}\widetilde{M_2}(t) = b_2$. In our mapping of Section 3.2, the phases in $\mathcal{S}_+$ correspond to the service time and the down-phase $-$ to the jump down in the process $\mathbf{M}$ of Section 2.2, respectively. The generator of this chain is

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{++} & \mathbf{T}_{+-} \\ \mathbf{T}_{-+} & \mathbf{T}_{--} \end{bmatrix}, \tag{3.60}$$

partitioned according to $\mathcal{S} = \mathcal{S}_+ \cup \{-\}$. Note that $\mathbf{T}_{++}$ is the transition matrix between the phases in $\mathcal{S}_+$, $\mathbf{T}_{+-}$ is the transition vector from phases in $\mathcal{S}_+$ to the down-phase $-$. The vector $\mathbf{T}_{-+}$ is the transition vector from down-phase $-$ to the up-phases in $\mathcal{S}_+$. Since the down-phase $-$ is exponentially distributed with parameter 1, $\mathbf{T}_{--} = -1$. Also, since $\mathbf{T}_{-+}\mathbf{1} = 1$, we may interpret the vector $\mathbf{T}_{-+}$ as the initial distribution of starting in any of the phases in $\mathcal{S}_+$ after a down-phase.

Further, define τ_+ as the time spent in the set of up-phases $\mathcal{S}_+$ until transition to state $-$. Then the random variable τ_+ follows a phase-type distribution, that is

$$\tau_+ \sim PH(\mathbf{T}_{-+}, \mathbf{T}_{++}, \mathbf{T}_{+-}). \tag{3.61}$$

As described in Section 3.2, the service time in the process $\mathbf{M}$ corresponds to the random variable τ_+. Therefore, the above construction of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq$

$0\}$ with phase-type distributed up-phases instead of exponential up-phases is a tandem fluid queue as described in Section 3.2.

We can now follow a similar analysis as in Section 3.3 to show that the stationary distribution of a tandem fluid with phase-type distributed set of phases is in distribution equal up to a factor to that of the process $\{(Z(t), M_2(t)); t \geq 0\}$ embedded at commencement of service times, where the service times are phase-type distributed.

Theorem 3.6. *In the current setting where the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ has phase-type distributed set of up-phases instead of exponential up-phases, the statements in Theorem 3.3 and Theorem 3.4 and hence Corollary 1 still hold.*

Chapter 4

Numerical treatment

To evaluate the stationary distribution discussed in Chapter 3, we use a numerical approach given in [15, 16]. First, we give a summary of this approach in Section 4.1. After that, the expressions for the Laplace-Stieltjes transform (LST) of the densities $\widetilde{\boldsymbol{\pi}}(x, y)$ are given in Section 4.2. At last, in Section 4.3, we give two simple examples of this numerical treatment with respect to the maximum priority process $\mathbf{M}$.

4.1 Numerical approach

In this numerical approach we discretize and bound the state space of the fluid level in the second buffer. We choose some large integer L, with $\ell = 1, 2, \ldots, L$, to be the number of uniform intervals of size Δu such that maximum level of the second buffer is $L\Delta u$.

First, we discretize the process J_k for moments when $x = 0$, discussed in Section 3.3, to the discrete-time Markov chain $\{\bar{J}_k; k = 0, 1, 2, \ldots\}$ with state space $\{(i, \ell); i \in \mathcal{S}_-, \ell = 1, 2, \ldots, L\}$. This discretization is constructed such that when $J_k = (i, 0, z)$ for some $(\ell - 1)\Delta u \leq z < \ell\Delta u$, $\ell = 1, 2, \ldots, L - 1$, we have $\bar{J}_k = (j, \ell)$. For $J_k = (i, 0, z)$ with $z \geq (L - 1)\Delta u$, we have $\bar{J}_k = (j, L)$. Note that for the models we discussed in 3, we only have one down-phase i.

Now, we discretize the corresponding transition probabilities $\mathbf{P}_{z,y}$ [16, equation (30) of Section 3.2] to $\bar{P}_{i,\ell,j,m}$, where

$$\bar{P}_{i,\ell,j,m} = P(\bar{J}_{k+1} = (j, m) \mid \bar{J}_k = (i, \ell)). \tag{4.1}$$

This matrix records the probabilities for hitting 0 in buffer X in phase j with level m in buffer Y, starting the last time hitting 0 in buffer X in phase i with level ℓ in buffer Y.

Because we only consider one down-phase, we can collect this values $\bar{P}_{i,\ell,j,m}$ in a matrix

$\bar{\mathbf{P}} = [\bar{P}_{\ell m}]_{\ell,m=1,2,...L}$, where

$$\bar{\mathbf{P}}_{\ell m} \approx \Delta u \mathbf{P}_{\ell\Delta u, m\Delta u}, \tag{4.2}$$

derived in [16, equations (74)–(75) of Section 4].

With the use of this discretized probability matrix $\bar{\mathbf{P}}$ we derive the stationary distribution vector of the process $\{\bar{J}_k; k = 0, 1, 2, \ldots\}$ denoted by $\bar{\boldsymbol{\xi}} = [\bar{\boldsymbol{\xi}}_\ell]_{\ell=1,2,\ldots,L}, \bar{\boldsymbol{\xi}}_\ell = [\bar{\xi}_{j,\ell}]_{j\in\mathcal{S}_-}$, where $lim_{k\to\infty} P(\bar{J}_k = (j, \ell)) = \bar{\xi}_{j,\ell}$. This vector $\bar{\boldsymbol{\xi}}$ exist if the process $\{\bar{J}_k; k = 0, 1, 2, \ldots\}$ is stable.

We approximate the the stationary distribution of the process $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M_2}(t)); t \geq 0\}$ using the limiting distribution vector $\bar{\boldsymbol{\xi}}_\ell$ such that

$$\boldsymbol{\xi}_z \approx \frac{\bar{\boldsymbol{\xi}}_\ell}{\Delta u}, \tag{4.3}$$

for any z with $(\ell - 1)\Delta u \leq z < \ell\Delta u$, $\ell = 1, 2, \ldots, L$ [16, equation (77) of Section 4]. Modifying an using [16, equations (42),(58),(78)], we have that

$$\widetilde{\mathbf{p}}(0, 0)_- \approx \delta \sum_{\ell=1}^{L} \bar{\boldsymbol{\xi}}_\ell e^{(|\check{c}_-|)^{-1}T_{--}\ell\Delta u}(-T_{--})^{-1}, \tag{4.4}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{\pi}}(0, \cdot)(s)_- \approx \delta \sum_{\ell=1}^{L} \bar{\boldsymbol{\xi}}_\ell \cdot e^{(|\check{c}_-|)^{-1}T_{--}\ell\Delta u}((|\check{c}_-|)^{-1}T_{--} + sI)^{-1} \\ \times \left(I - e^{((|\check{c}_-|)^{-1}T_{--}+sI)\ell\Delta u}\right)(|\check{c}_-|)^{-1},\end{aligned} \tag{4.5}$$

where δ is a normalizing constant and I is the identity matrix of appropriate size. See Appendix A for the explicit expression of δ.

With the use of equation 4.5 and [16, equation (61) of Section 3] we can now approximate $\widetilde{\boldsymbol{\pi}}(x, y), x > 0, y > 0$ with the use of the Euler-Euler inversion method in [1]. Recall that we only considered models with one down-phase.

4.2 Inverting the Laplace-Stieltjes transform

We evaluate the densities $\widetilde{\boldsymbol{\pi}}(x, y) = [\widetilde{\boldsymbol{\pi}}(x, y)_+ \;\; \widetilde{\boldsymbol{\pi}}(x, y)_-]$ by inverting the LST $\boldsymbol{\pi}(\cdot, \cdot)(v, s) = [\boldsymbol{\pi}(\cdot, \cdot)(v, s)_+ \;\; \boldsymbol{\pi}(\cdot, \cdot)(v, s)_-]$. For every phase j, $[\pi(\cdot, \cdot)(v, s)]_j$ is defined as

$$[\pi(\cdot, \cdot)(v, s)]_j = E(e^{-vX}e^{-sY}I\{\varphi = j\}), \tag{4.6}$$

where $X = lim_{t\to\infty}X(t)$, $Y = lim_{t\to\infty}Y(t)$ and $\varphi = lim_{t\to\infty}\varphi(t)$. We calculate the LST with the Euler-Euler 2D-inversion method of Abate and Whitt in [1]. With the use of [16, Corollary 3], we are now able to derive the LST.

Note that, as explained in Section 3.3, the level variable Y satisfies $Y \geq X\cdot\min_{i\in\mathcal{S}_+}\{\hat{c}_i/r_i\}$. Therefore, in order to apply the algorithm in [1], we perform the following shift of the LST. Define $\tilde{Y} = Y - X \cdot \min_{i\in\mathcal{S}_+}\{\hat{c}_i/r_i\}$ and the LST of $\widetilde{\pi}(x, \tilde{y})$ for phase j such that

$$
\begin{aligned}
[\pi(\cdot,\cdot)(v - \min_{i\in\mathcal{S}_+}\{\hat{c}_i/r_i\}\cdot s, s)]_j &= E(e^{-(v-\min_{i\in\mathcal{S}_+}\{\hat{c}_i/r_i\}\cdot s)X}\cdot e^{-sY}I\{\varphi = j\}) &(4.7)\\
&= E(e^{-vX}\cdot e^{-s(Y-X\cdot\min_{i\in\mathcal{S}_+}\{\hat{c}_i/r_i\})}I\{\varphi = j\}) &(4.8)\\
&= E(e^{-vX}\cdot e^{-s\tilde{Y}}I\{\varphi = j\}). &(4.9)
\end{aligned}
$$

4.3 Examples

In this section we construct two examples using the numerical approach of Section 4.1. In the first example we consider a server with exponentially distributed service times. The second example considers a server with Coxian-2 distributed service times.

Example 1. We consider a single server queue with two classes of customers with Poisson arrival rates λ_1 and λ_2. These customers gain priority over time with class-dependent rates b_1 and b_2 where b_1 is larger than b_2. The single server has an exponential distribution B for service times with mean $1/\mu$. We choose the following parameters: $b_1 = 1, b_2 = 0.5, \lambda_1 = 1, \lambda_2 = 2, \mu = 4$.

In order to derive the stationary distribution embedded at commencement of service of the maximum priority process $\mathbf{M}$, we first derive the stationary distribution of the tandem fluid queue $\{(\varphi(t), \widetilde{Z}(t), \widetilde{M}_2(t)); t \geq 0\}$ discussed in Section 3.2. We translate the above parameters of the maximum priority process to parameters of the tandem fluid queue, where (3.1)–(3.5) are compared to (3.8)–(3.14): $\mathcal{S} = \{+, -\}$, $r_+ = b_1 - b_2 = 0.5$, $r_- = -\left(\frac{\lambda_1}{b_1}\right) = -1$, $\hat{c}_+ = b_2 = 0.5$, $\hat{c}_- = 0$, $\check{c}_+$ does not exist, $\check{c}_- = -\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right) = -5$ and

$$
\mathbf{T} = \begin{bmatrix} -4 & 4 \\ 1 & -1 \end{bmatrix}.
$$

We note that the stability conditions are met, since

$$
\sum_{i\in\mathcal{S}} r_i P(\varphi = i) = -0.7 < 0,
$$

$$
\sum_{i\in\mathcal{S}} \hat{c}_i P(\varphi = i, X > 0) - \sum_{i\in\mathcal{S}_-} |\check{c}_i| P(\varphi = i, X = 0) = -3.4 < 0.
$$

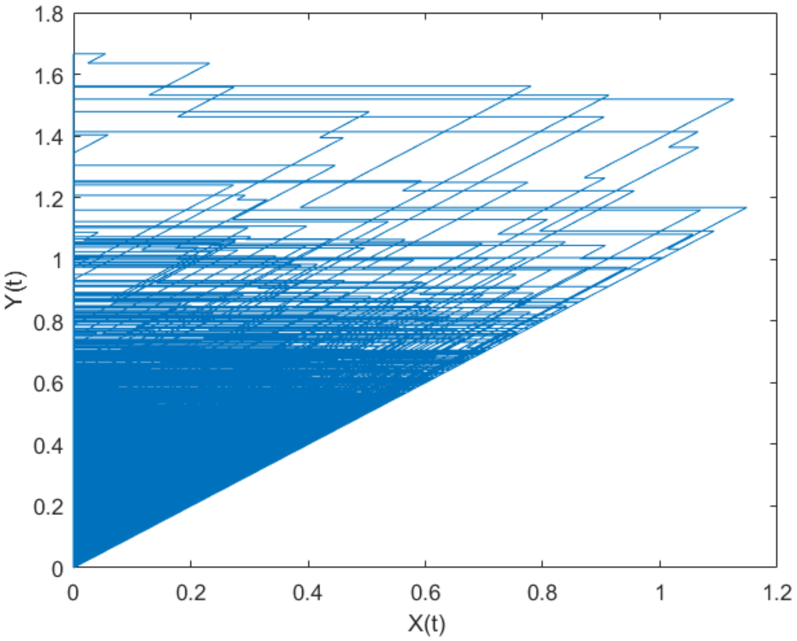

Figure 4.1: Simulated values for $\widetilde{Z}(t)$ and $\widetilde{M_2}(t)$, with $0 \leq t \leq 10^5$, in Example 1.

First, we simulate the above process to find appropriate parameters L and Δu. The result of one of these simulations is displayed in Figure 4.1. As seen in Figure 4.1, the level of the second buffer does not exceed 2. Figure 4.2 shows the probability of exceeding 2 is nearly 0. As a result, we approximate this process with parameters $L = 200$ and $\Delta u = 0.01$, which truncates the values of y to the interval $[0, 2]$.

The values $\boldsymbol{\xi}_z$ based on (77) in [16] are plotted in Figure 4.3. We find $\widetilde{\mathbf{p}}(0,0)_- = 0.6796$ and $\widetilde{\mathbf{p}}_- = 0.7$, which indicates that the stationary probability mass at $(0, y), y > 0$ is $\widetilde{\mathbf{p}}_- - \widetilde{\mathbf{p}}(0,0)_- = 0.0204$. With the use of the Euler inverse method of Abate and Whitt in [2], we find the values of $\widetilde{\boldsymbol{\pi}}(0, y)$, which are displayed in Figure 4.4.

The stationary probability of both buffers being non-empty is $1 - \widetilde{\mathbf{p}}_- = 0.3$. In Figure 4.5 are selected values of $\widetilde{\boldsymbol{\pi}}(x, y)$ plotted, using the Euler-Euler inversion method of Abate and Whitt [1].

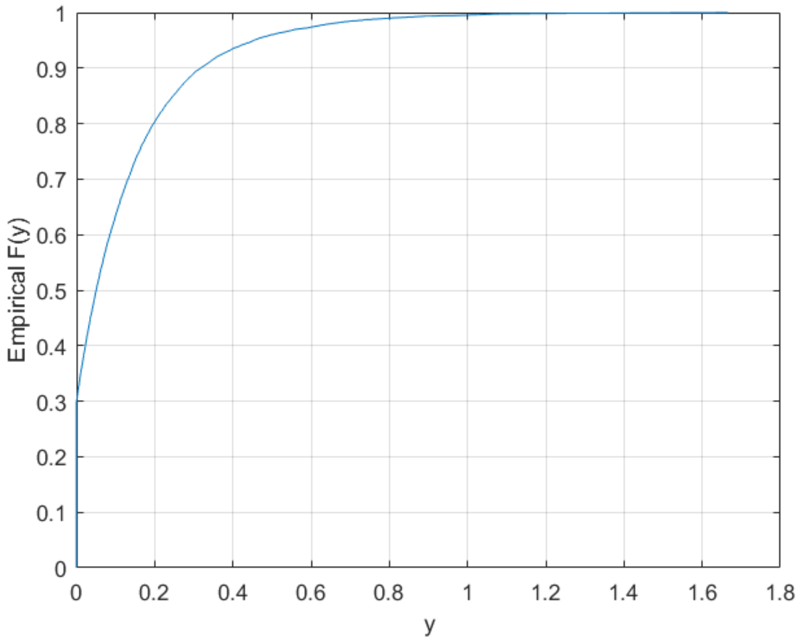

Figure 4.2: Empirical values of $F(y) = P(Y \leq y)$ in the simulation of Example 1.

Recall from Corollary 1 that we are interested in the stationary distribution of the fluid model in the down-phase $-$ to calculate the stationary distribution at times of commencement of service of the maximum priority process $\mathbf{M}$. In this example we use the generator to calculate the factor $\beta = 4/5$.

The interpretation of the solution is that in stationarity 87.5% of the times that a service is finished, the first person in line will commence service. This is regardless of type due to the fact that the priority levels for customer class 1 and customer class 2 are equal, i.e. $M_1 = M_2$. The probability that the queue is empty when a customer arrives is $\mathbf{p}(0,0)_- = \widetilde{\mathbf{p}}(0,0)_-/\beta \approx 85\%$.

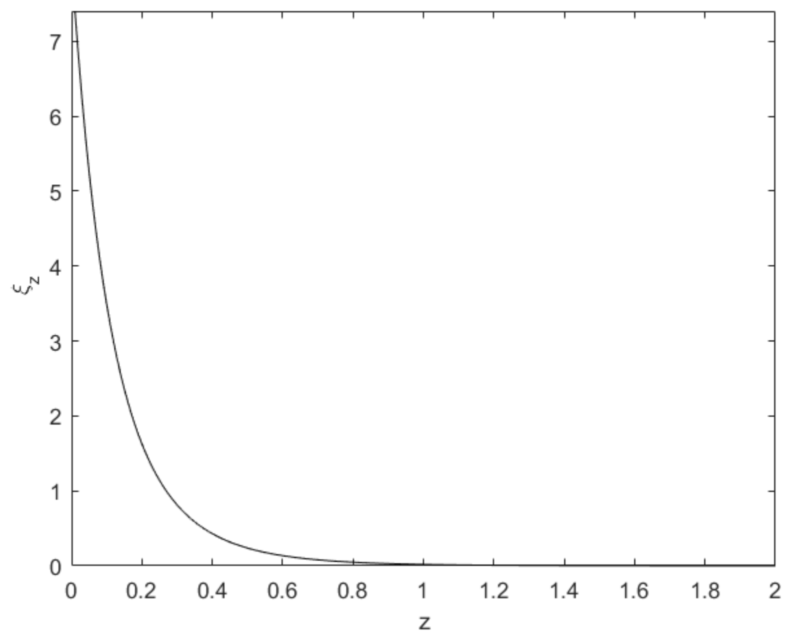

Figure 4.3: The values $[\bar{\boldsymbol{\xi}}_z]_i$ for $i = 2$ in Example 1.

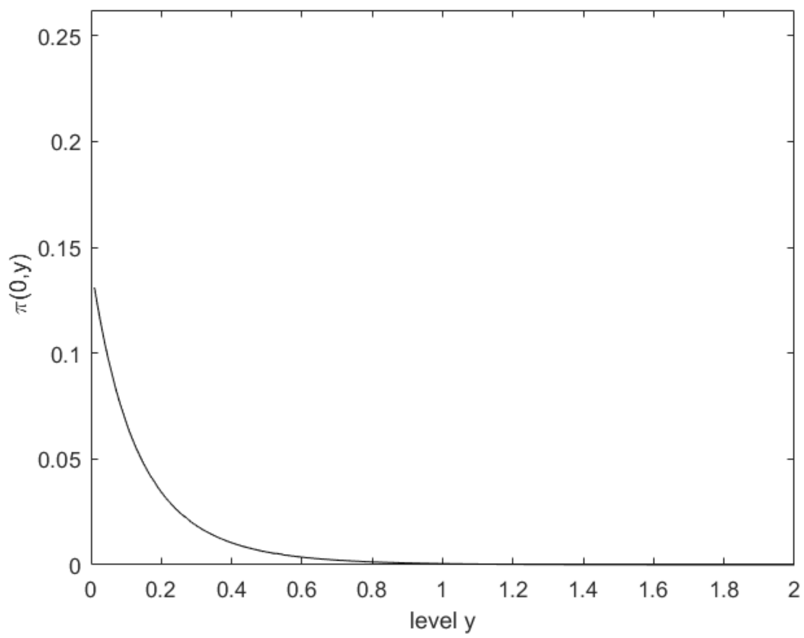

Figure 4.4: The values $[\widetilde{\boldsymbol{\pi}}(0, y)]_i$ for $i = 2$ in Example 1.

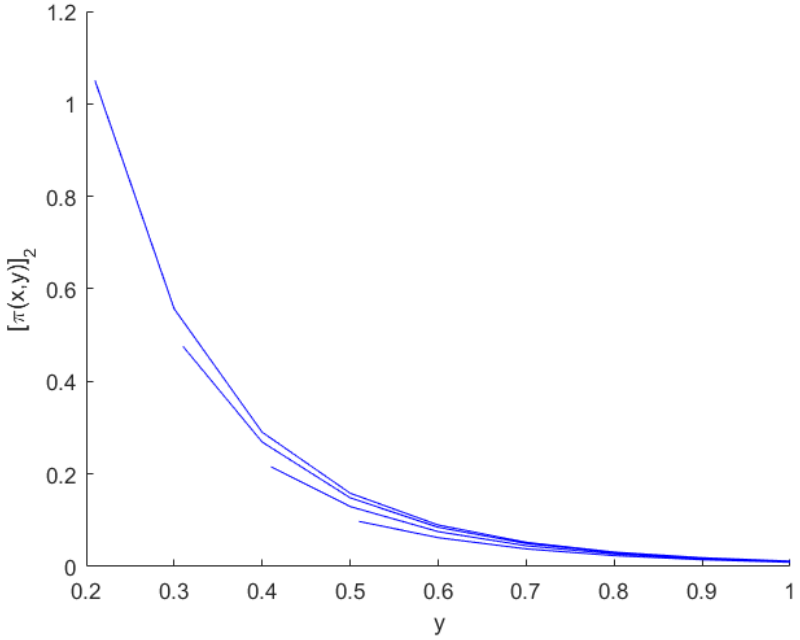

Figure 4.5: The values $[\widetilde{\boldsymbol{\pi}}(x, y)]_i$ for $i = 2$ and selected values of x, y in Example 1. The lines (left to right) correspond to $x = 0.2, \ldots, 0.5$, with $y = x + 0.01, x + 0.1, x + 0.2, \ldots, 1$. We note that the range of x and y is chosen such that it meets the condition $y > x\widehat{c}_1/r_1$ (here equivalent to $y > x$ since $\widehat{c}_1/r_1 = 1$), as detailed in Section 3.2.

Example 2. We consider a single server queue with two classes of customers with Poisson arrival rates λ_1 and λ_2. Again, these customers gain priority over time with class-dependent rates b_1 and b_2 where b_1 is larger than b_2. The single server has a Coxian-2 distribution B for service times. For customer arriving at the server, he has a service time which is exponentially distributed with parameter $1/\mu$. With probability p_1 he finishes service and with probability $1 - p_1$ he has another service, which is also exponentially distributed with parameter $1/\mu$. The situation is displayed in Figure 4.6. We choose the following parameters: $b_1 = 1, b_2 = 0.5, \lambda_1 = 1, \lambda_2 = 2, \mu = 4, p_1 = 0.5$.

We translate the above parameters of the maximum priority process with phase-type distributed service times to parameters of the tandem fluid queue, where (3.1)–(3.5) are compared to (3.8)–(3.14): $\mathcal{S} = \{1, 2, 3\}$, $r_{1,2} = b_1 - b_2 = 0.5$, $r_3 = -\left(\frac{\lambda_1}{b_1}\right) = -1$, $\hat{c}_{1,2} = b_2 = 0.5$, $\hat{c}_3 = 0$, $\check{c}_{1,2}$ does not exist, $\check{c}_3 = -\left(\frac{\lambda_1}{b_1} + \frac{\lambda_2}{b_2}\right) = -5$ and

$$\mathbf{T} = \begin{bmatrix} -4 & 2 & 2 \\ 0 & -4 & 4 \\ 1 & 0 & -1 \end{bmatrix}.$$

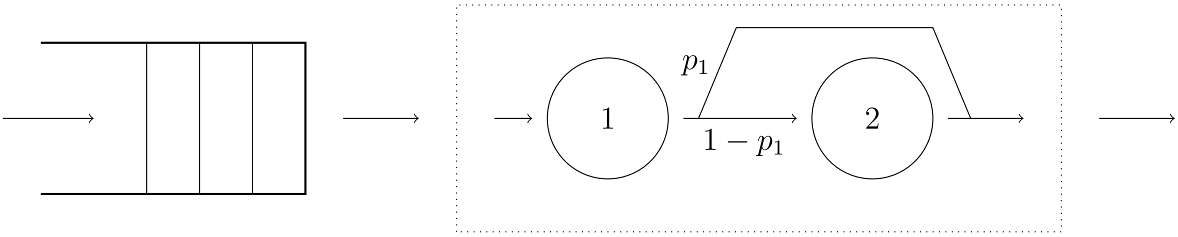

Figure 4.6: The Coxian-2 distribution. In Example 2 the parameter $p_1 = 0.5$.

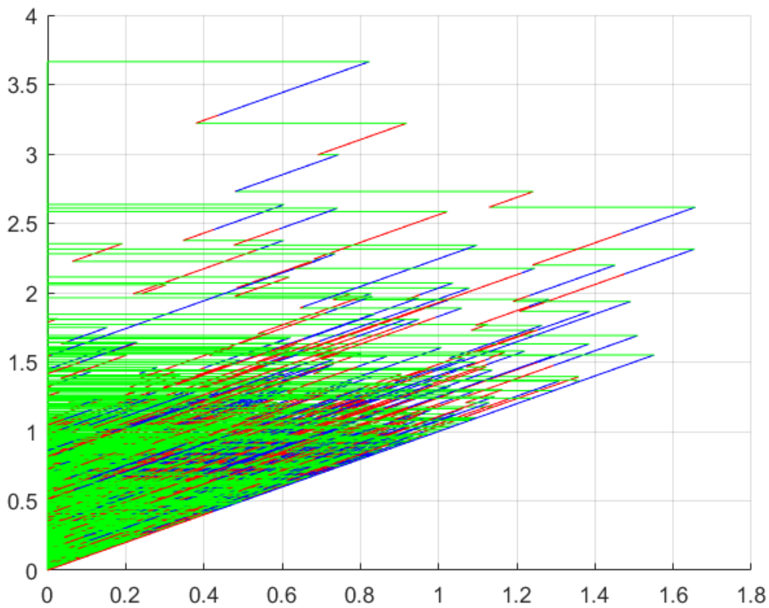

Figure 4.7: Simulated values for $\widetilde{Z}(t)$ and $\widetilde{M_2}(t)$, with $0 \leq t \leq 10^5$. The colors represent the time spent in different phases where red is phase 1, blue is phase 2 and green is phase 3.

We note that the stability conditions are met, since

$$\sum_{i \in \mathcal{S}} r_i P(\varphi = i) = -0.5909 < 0,$$

$$\sum_{i \in \mathcal{S}} \hat{c}_i P(\varphi = i, X > 0) - \sum_{i \in \mathcal{S}_-} |\check{c}_i| P(\varphi = i, X = 0) = -2.8182 < 0.$$

We simulate the above process to find appropriate parameters L and Δu. The result of one of these simulations is displayed in Figure 4.7. As seen in Figure 4.7, the level of the second buffer does not exceed 4. Figure 4.8 shows the probability of exceeding 4 is nearly 0. As a result, we approximate this process with parameters $L = 400$ and $\Delta u = 0.01$, which truncates the values of y to the interval $[0, 4]$.

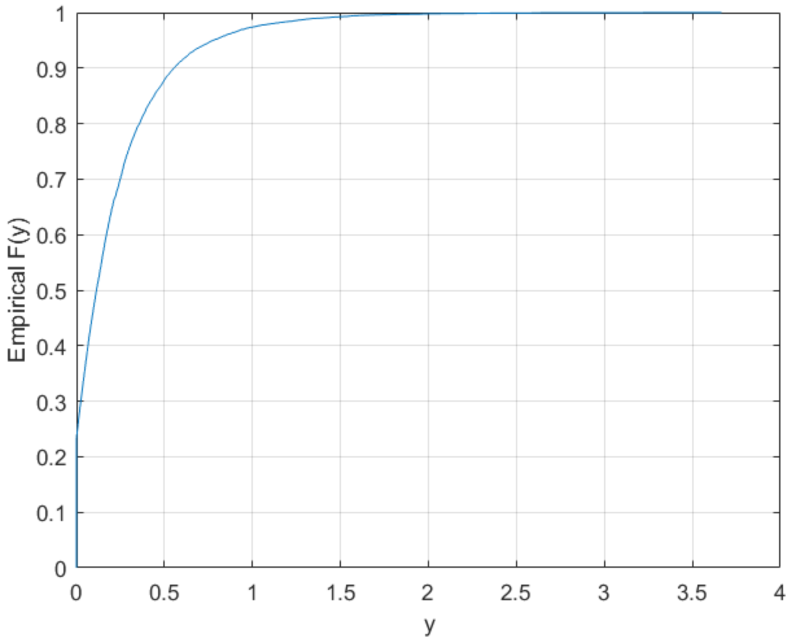

Figure 4.8: Empirical values of $F(y) = P(Y \leq y)$ in the simulation of Example 2.

We find $\widetilde{\mathbf{p}}(0,0)_- = 0.5631$ and $\widetilde{\mathbf{p}}_- = 0.5909$, which indicates that the stationary probability mass at $(0, y), y > 0$ is $\widetilde{\mathbf{p}}_- - \widetilde{\mathbf{p}}(0,0)_- = 0.0278$. With the use of the Euler inverse method of Abate and Whitt in [2], we find the values of $\boldsymbol{\pi}(0, y)$. These values are displayed in Figure 4.9.

The stationary probability of both buffers being non-empty is $1 - \widetilde{\mathbf{p}}_- = 0.4091$. In Figure 4.10 are selected values of $\widetilde{\boldsymbol{\pi}}(x, y)$ plotted, using the Euler-Euler inversion method of Abate and Whitt [1].

Recall from Corollary 1 that we are interested in the stationary distribution of the fluid model in the down-phase $-$ to calculate the stationary distribution at times of commencement of service of the maximum priority process $\mathbf{M}$. In this example we use the generator to calculate the factor $\beta = 8/11$.

The interpretation of the solution is that in stationarity 81.25% of the times that a service is finished, the first person in line will commence service. This is regardless of type due to the fact that the priority levels for customer class 1 and customer class 2 are equal, i.e. $M_1 = M_2$. The probability that the queue is empty when a customer arrives is $\mathbf{p}(0,0)_- = \widetilde{\mathbf{p}}(0,0)_-/\beta \approx 77.43\%$.

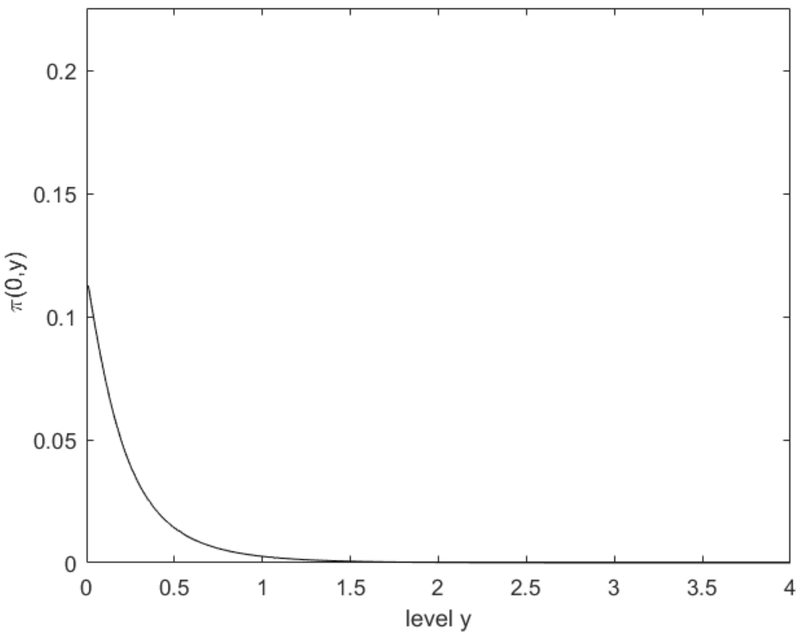

Figure 4.9: The values $[\boldsymbol{\pi}(0, y)]_i$ for $i = 3$ in Example 2.

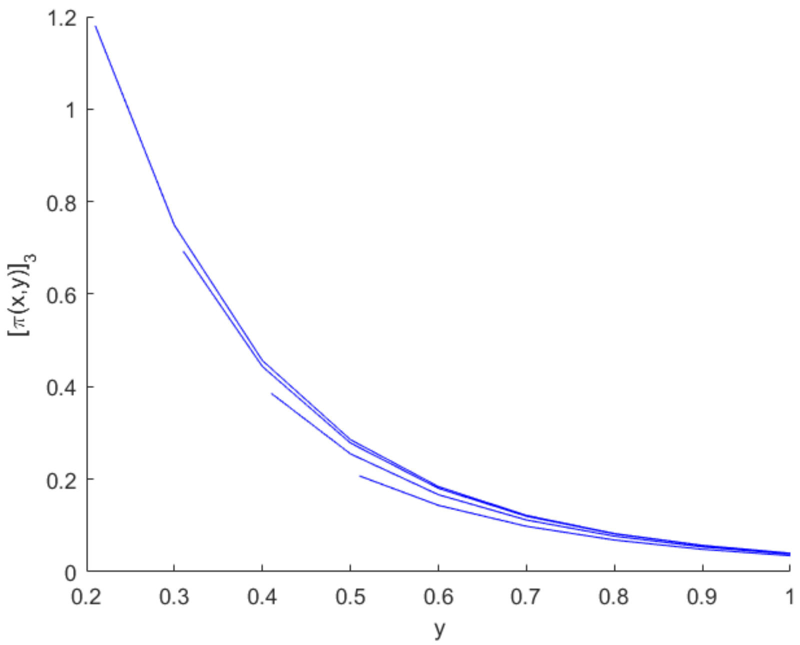

Figure 4.10: The values $[\boldsymbol{\pi}(x, y)]_i$ for $i = 3$ and selected values of x, y in Example 2. The lines (left to right) correspond to $x = 0.2, \ldots, 0.5$, with $y = x + 0.01, x + 0.1, x + 0.2, \ldots, 1$. We note that the range of x and y is chosen such that it meets the condition $y > x\widehat{c}_1/r_1$ (here equivalent to $y > x$ since $\widehat{c}_j/r_j = 1$ for every $j \in \mathcal{S}_+$).

Chapter 5

Conclusions and recommendations

In this chapter the conclusions and recommendation of the research are provided. In the Section 5.1 the reader will find the conclusions, while some recommendations for further research are given in Section 5.2.

5.1 Conclusions

In this project we considered the maximum priority process $\mathbf{M}$ studied in [18], and derived the results for the stationary distribution of this process embedded at times of commencement of service. The stationary distribution of the maximum priority process at the times of the commencement of service gives information on the maximum waiting time of a customer. At the time of commencement of a customer, his priority is known. With this information, since the priority rates accumulate at a constant rate, we know the maximum time this customer waited.

First, we considered the two class accumulating priority queue and defined the corresponding maximum priority process [9, 18]. We explained the model in detail with pictures for convenience of the reader.

Next, we mapped the maximum priority process to a tandem fluid queue analysed in [15, 16]. The assumptions in these papers were to strict to model the maximum priority process at first, however we were still able to use the techniques derived in [15, 16] with a slightly loosened assumption. With this adjustment we were able to map the maximum priority process to a tandem fluid queue and we derived the stationary distribution at the times of the commencement of service, where the service times were exponentially distributed.

Further, we extended these results to the maximum priority process in which service times are phase-type distributed. We assumed that the service times for every customer class followed the same distribution and that priority rates were equal for every phase. Again, with the use of the results of the tandem fluid queue in [15, 16] we derived the stationary distribution at the times of the commencement of service.

Finally, we constructed two simple numerical examples to illustrate the theory.

5.2 Recommendations

As future work we are interested in the analysis of the models with more than only two classes of customers, so that we can consider more realistic scenarios. For example, the emergency department of a hospital does not only collect urgent and non-urgent patients. There are also patients that are more urgent then non-urgent patients, but not life-threatening. These different types of patients can be modelled as the classes of customers in priority queues.

Also, we are interested in class-dependent service times. The assumption of a service time that is the same for every customer class is generally not true. To address this, different requests can be modelled as different classes with class-dependent service times.

Further, it would be interesting to construct time-varying models in order to analyse problems where the arrival/service rates vary in time. Such models could be useful in the analysis of peak-hours in supermarkets, as an example. This could be achieved by applying the methodology derived in this research to a time-varying tandem model in [13], built on the results in [12, 15].

Bibliography

[1] J. Abate. and W. Whitt, "A unified framework for numerically inverting laplace transforms," *INFORMS Journal on Computing*, vol. 18, no. 4, pp. 408–421, 2006.

[2] J. Abate. and W. Whitt, "Numerical inversion of laplace transforms of probability distributions," *ORSA Journal of Computing*, vol. 7, no. 1, pp. 36–43, 1995.

[3] N. G. Bean, G. T. Nguyen, M. M. O'Reilly and V. Sunkara, "A numerical framework for computing the limiting distribution of a stochastic fluid-fluid process," *Proceedings of the Ninth International Conference on Matrix-Analytic Methods in Stochastic Models*, pp. 21–22, 2016.

[4] N. G. Bean and M. M. O'Reilly, "A stochastic two-dimensional fluid model," *Stochastic Models*, vol. 29, no. 1, pp. 31–63, 2013.

[5] N. G. Bean and M. M. O'Reilly, "The stochastic fluid-fluid model: A stochastic fluid model driven by an uncountable-state process, which is a stochastic fluid model itself," *Stochastic Processes and their Applications*, vol. 124, no. 5, pp. 1741–1772, 2014.

[6] N. G. Bean, M. O'Reilly, and P. G. Taylor, "Algorithms for the laplace-stieltjes transforms of first return times for stochastic fluid flows," *Methodology and Computing in Applied Probability*, vol. 10, no. 3, pp. 381–408, 2008.

[7] N. G. Bean, M. M. O'Reilly, and P. G. Taylor, "Hitting probabilities and hitting times for stochastic fluid flows," *Stochastic Processes and their Applications*, vol. 115, no. 9, pp. 1530–1556, 2005.

[8] R. W. Conway, W. L. Maxwell, and L. W. Miller, "Theory of scheduling," 1967.

[9] D. J. J. Dams, "Stationary waiting time distribution of the maximum priority process in the accumulating priority $M/M/1$ queue," internship dissertation, University of Melbourne, 2013.

[10] M. S. Hagen, J. K. Jopling, T. G. Buchman, and E. K. Lee, "Priority queuing models for hospital intensive care units and impacts to severe case patients," *AMIA Annual Symposium proceedings*, vol. 2013, pp. 841–850, 2013.

[11] L. Kleinrock, "A delay dependent queue discipline," *Naval Research Logistics Quarterly*, vol. 11, no. 3-4, pp. 329–341, 1964.

[12] B. Margolius and M. M. O'Reilly, "The analysis of cyclic stochastic fluid flows with time-varying transition rates," *Queueing Systems*, vol. 82, no. 1-2, pp. 43–73, 2016.

[13] B. Margolius, M. M. O'Reilly and W. R. W. Scheinhardt, "The analysis of Markov-modulated tandem fluid queues with time-varying transition rates," *In preparation.*

[14] G. R. Miller, "Priority queues," *The Annals of Mathematical Statistics*, vol. 31, no. 1, pp. 86–10, 03 1960.

[15] M. M. O'Reilly and W. R. W. Scheinhardt, "Analysis of tandem fluid queues," *Proceedings of the Ninth International Conference on Matrix-Analytic Methods in Stochastic Models*, pp. 85–92, 2016.

[16] M. M. O'Reilly and W. R. W. Scheinhardt, "Stationary distributions for a class of Markov-modulated tandem fluid queues," Accepted for publication in *Stochastic Models*, 2017.

[17] A. Samuelson, M. M. O'Reilly, and N. G. Bean, "On the generalised reward generator for stochastic fluid models: a new equation for Ψ," Accepted for publication in *Stochastic Models*, 2017.

[18] D. A. Stanford, P. Taylor, and I. Ziedins, "Waiting time distributions in the accumulating priority queue," *Queueing Systems*, vol. 77, no. 3, pp. 297–330, 2014.

Appendix A

The normalizing constant δ

The normalizing constant δ of equation (4.5) is given in [16, equation (43) of Section 3]. For convenience we state the expression below.

$$\delta = \left\{ \boldsymbol{\xi}(-T_{--})^{-1} \left(\mathbf{1} + \mathbf{T}_{-+}\mathbf{K}^{-1} \left[\ (\mathbf{R}_{+})^{-1} \quad \boldsymbol{\Psi}(|\mathbf{R}_{-}|)^{-1} \ \right] \right) \right\}^{-1}, \tag{A.1}$$

where $\boldsymbol{\xi} = \int_{z=0}^{\infty} \boldsymbol{\xi}_z dz$, $\boldsymbol{\Psi} = \widehat{\boldsymbol{\Psi}}(s)|_{s=0}$ and $\mathbf{K} = \widehat{\mathbf{K}}(s)|_{s=0}$, with $\widehat{\boldsymbol{\Psi}}(s)$ the minimum non-negative solution of the Ricatti equation and where $\widehat{\mathbf{K}}(s)$ solves

$$\widehat{\mathbf{K}}(s) = \widehat{\mathbf{Q}}(s)_{++} + \widehat{\boldsymbol{\Psi}}(s)\widehat{\mathbf{Q}}(s)_{-+}. \tag{A.2}$$

The definition of the key generator matrix $\widehat{\mathbf{Q}}(s)$ can be found in [equation (18)] [4].