Efficient Analysis of the MMAP[K]/PH[K]/1 Priority Queue[☆]

Gábor Horváth[a,b,1]

[a] *Budapest University of Technology and Economics, Department of Networked Systems and Services*
[b] *MTA-BME Information Systems Research Group*

Abstract

In this paper we consider the MMAP/PH/1 priority queue, both the case of preemptive resume and the case of non-preemptive service. The main idea of the presented analysis procedure is that the sojourn time of the low priority jobs in the preemptive case (and the waiting time distribution in the non-preemptive case) can be represented by the duration of the busy period of a special Markovian fluid model. By making use of the recent results on the busy period analysis of Markovian fluid models it is possible to calculate several queueing performance measures in an efficient way including the sojourn time distribution (both in the time domain and in the Laplace transform domain), the moments of the sojourn time, the generating function of the queue length, the queue length moments and the queue length probabilities.

Keywords: Queueing, Preemptive resume priority queue, Non-preemptive priority queue, matrix-analytic methods

1. Introduction

Priority queues belong to the most essential multi-class queueing systems that allow different job classes to receive differentiated levels of service. They play an important role in several fields like telecommunication [1], manufacturing systems [2] or, more recently, in health care [3], [4].

[☆] This work was supported by the Hungarian research project OTKA K101150, and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Email address: ghorvath@hit.bme.hu (Gábor Horváth)

Priority queues are extensively studied since the middle of the last century [5], starting with the most basic variant with Poisson arrival process and exponentially distributed service times. However, in the practice there are cases when the Poisson assumption is not reasonable. In the last two decades most research activity on priority queues has considered more general arrival proceeses like the Markovian arrival process (MAP) or the marked Markovian arrival process (MMAP).

In [6] the MAP/G/1 preemptive priority queue is analyzed based on the workload process, and the Laplace-Stieltjes transform (LST) of the sojourn time distribution of the jobs is derived. The non-preemptive case is investigated in [7] and [8], where the LST of the sojourn time, the moments of the sojourn time, the generating function (GF) of the queue length, the queue length moments and the queue length probabilities are provided. [9] studies the tail probabilities of the low priority waiting times and queue lengths in the MAP/G/1 non-preemptive priority queue.

After this overview one may think that not too much has left to be done in the field of MAP driven priority queues. However, all the aforementioned results assume a general distribution for the service time, which makes the solution complex and often difficult to implement in a proper way (in the numerical sense). To address this issue the generally distributed service times can be replaced by phase-type distributed ones in the hope of the simpler and numerically more tractable solution.

In [10] the (discrete-time) MAP/PH/1 priority queue is considered by representing the state space with a quasi birth-death process (QBD) and exploiting the special structure of the related fundamental matrices. While this approach is elegant and seems promising, there are some computational bottlenecks (as pointed out in [11]). There have been efforts to make it more efficient (see [12] and [11]), but apart from the queue length moments all performance measures can be computed only in case of a very limited number of phases.

The solution approach presented in this paper is based on the analysis of the workload process, like in [6]. The main difference is, however, that in case of PH distributed service times it is possible to analyze the workload process and the performance measures through some appropriately defined Markovian fluid models. Taking advantage of the matrix-analytic solution technique available for Markovian fluid models we managed to derive several sojourn time and queue length related quantities in an efficient and numerically stable way, both with preemptive resume and non-preemptive service. The computationally most intensive steps of the procedure are the solutions of non-symmetric algebraic Riccati equations and Sylvester equations, for which various mature implementations exist, allowing to compute the performance measures in a reasonable time even if the number of phases is relatively large.

The rest of the paper is organized as follows. Section 2 introduces the queue considered in the paper. Section 3 covers Markovian fluid flows (especially their busy period), as our solution relies on them. For two job classes, the preemptive priority queue is analyzed in Section 4, and the non-preemptive case is considered in Section 5. The extension to arbitrary many job classes is provided in the Appendix. Some numerical examples are presented in Section 6. Finally, Section 7 concludes the paper.

2. The MMAP[K]/PH[K]/1 priority queue

In the MMAP[K]/PH[K]/1 queue K types (classes) of jobs are distinguished. The arrival process of the jobs is described by a marked Markovian arrival process, and the service times are phase-type (PH) distributed. There is a single server, which always picks the job having the highest priority for service. If the ongoing service can not be interrupted when a higher priority job arrives, the service is called to be *non-preemptive*. In the *preemptive resume* case (also referred to as the preemptive case for simplicity), however, the service of jobs can be interrupted, and resumed later when all higher priority jobs leave the system.

To introduce the analysis approach, the two-class case ($K = 2$) is considered throughout the paper, and the extension to the general case ($K > 2$) is provided in the appendix.

The MMAP characterizing the arrivals [13] has a background process, that is a continuous time Markov chain (CTMC) $\{\mathcal{J}(t), t > 0\}$ with N_A states and generator matrix $\mathbf{D}$ (which is assumed to be irreducible). Some of the transitions of the background process are accompanied by the arrival of high (low) priority jobs with the corresponding transition rates given by matrix $\mathbf{D_H}$ ($\mathbf{D_L}$), respectively. The rates of the internal transitions (that do not generate arrivals) are in matrix $\mathbf{D_0}$, thus we have that $\mathbf{D} = \mathbf{D_0} + \mathbf{D_L} + \mathbf{D_H}$.

The mean arrival rate of high priority jobs is denoted by λ_H, and it is calculated by $\lambda_H = \theta \mathbf{D_H} \mathbb{1}$ with vector θ being the steady state distribution of the MMAP phase process, which is the unique solution of $\theta \mathbf{D} = 0, \theta \mathbb{1} = 1$ ($\mathbb{1}$ denotes the column vector of ones). The mean arrival rate of low priority jobs is calculated similarly, it is $\lambda_L = \theta \mathbf{D_L} \mathbb{1}$.

The random variable representing the service times of the low priority jobs S_L is PH distributed [14] with N_L phases, characterized by $\sigma_L, \mathbf{S_L}$ and s_L. Row vector σ_L is the initial vector, matrix $\mathbf{S_L}$ is the transient generator and column vector s_L holds the transition rates to the absorbing state, thus $s_L = -\mathbf{S_L}\mathbb{1}$. The probability density function (pdf) $f_{S_L}(t)$, its Laplace transform $f^*_{S_L}(s)$ and the moments $E(S_L^k)$ are

$$f_{S_L}(t) = \sigma_L e^{\mathbf{S_L}t} s_L, \quad f^*_{S_L}(s) = \sigma_L(s\mathbf{I} - \mathbf{S_L})^{-1} s_L, \quad E(S_L^k) = k!\sigma_L(-\mathbf{S_L})^{-k}\mathbb{1}, \quad (1)$$

and the mean service rate is $\mu_L = 1/E(S_L)$. The PH distribution corresponding to the high priority service times and its properties are defined similarly, by using subscript H instead of L.

The load of the queue is $\rho = \lambda_H/\mu_H + \lambda_L/\mu_L$. Throughout in this paper $\rho < 1$ is assumed.

3. Markovian fluid models

3.1. Definition and stationary solution

Markov fluid models (also known as Markovian fluid flows) are characterized by a two-dimensional Markov process $\{\mathcal{X}(t), \mathcal{Z}(t), t > 0\}$, where $\mathcal{X}(t)$ represents the fluid level and $\mathcal{Z}(t)$ is the underlying CTMC with state space $\mathcal{S}$ of size $|\mathcal{S}| = N$ and generator matrix $\mathbf{Q}$ that modulates the rate at which fluid is accumulated in the fluid buffer.

The rate at which the level of the buffer changes in state i of the background process is denoted by r_i. The diagonal matrix $\mathbf{R}$ is composed by fluid rates $r_i, i = 1, \ldots, N$. Formally, the behavior of the fluid buffer is as follows,

$$\frac{d}{dt}\mathcal{X}(t) = \begin{cases} r_{\mathcal{Z}(t)}, & \text{if } \mathcal{X}(t) > 0, \\ \max\{0, r_{\mathcal{Z}(t)}\}, & \text{if } \mathcal{X}(t) = 0. \end{cases} \quad (2)$$

Let us denote the row vector of the stationary distribution of the fluid level for $x > 0$ by $\pi(x) = \{\pi_i(x), i \in \mathcal{S}\}$ with $\pi_i(x) = \lim_{t\to\infty} \lim_{\Delta\to 0}(1/\Delta)P(\mathcal{X}(t) \in (x, x + \Delta), \mathcal{Z}(t) = i)$, and the row vector of the stationary probabilities of empty buffer by $p = \{p_i, i \in \mathcal{S}\}$ with $p_i = \lim_{t\to\infty} P(\mathcal{Z}(t) = i, \mathcal{X}(t) = 0)$.

In the recent decades it has been recognized that the matrix-analytic approach basing the efficient analysis of QBDs can be applied to fluid models as well, making it possible to solve fluid models with a large number of states (up to several thousand) in a numerically stable way (see [15],[16]). Fluid models where $|r_i| = 1, \forall i \in \mathcal{S}$ are referred to as *canonical fluid models*, and are especially simple to analyze. Here we summarize the main steps of the analysis of canonical fluid models. We assume that the state space is partitioned according to the associated fluid rates to two sets $\mathcal{S}_+ = \{i \in \mathcal{S}, r_i = 1\}$ and $\mathcal{S}_- = \{i \in \mathcal{S}, r_i = -1\}$ ($N_+ = |\mathcal{S}_+|, N_- = |\mathcal{S}_-|$) as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_{++} & \mathbf{Q}_{+-} \\ \mathbf{Q}_{-+} & \mathbf{Q}_{--} \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{I} & \\ & -\mathbf{I} \end{bmatrix}. \tag{3}$$

The analysis is based on two fundamental matrices, matrix $\mathbf{\Psi}$ and $\mathbf{K}$ (see [15]). Matrix $\mathbf{\Psi}$ has a simple probabilistic interpretation, entry $(\mathbf{\Psi})_{i,j}$, $i \in \mathcal{S}_+, j \in \mathcal{S}_-$ is the probability that the background process is in state j when the fluid level returns to 0 given that it was in state i when the busy period (a non-empty period of the fluid queue) was initiated. Matrix $\mathbf{\Psi}$ is the solution to the nonsymmetric algebraic Riccati equation (NARE)

$$\mathbf{\Psi}\mathbf{Q}_{-+}\mathbf{\Psi} + \mathbf{\Psi}\mathbf{Q}_{--} + \mathbf{Q}_{++}\mathbf{\Psi} + \mathbf{Q}_{+-} = \mathbf{0}. \tag{4}$$

Matrix $\mathbf{K}$ has an important role as well. Entry i, j of matrix $e^{\mathbf{K}x}$ is the expected number of crossings of fluid level x in phase $j \in \mathcal{S}_+$ starting from level 0 and phase $i \in \mathcal{S}_+$, before returning to level 0. If the mean fluid rate is negative, all eigenvalues of matrix $\mathbf{K}$ have negative real parts (thus it is full rank and invertible) and can be expressed from $\mathbf{\Psi}$ as

$$\mathbf{K} = \mathbf{Q}_{++} + \mathbf{\Psi}\mathbf{Q}_{-+}. \tag{5}$$

Based on these matrices the stationary fluid level density vector and the stationary probability vector of the idle buffer can be computed by the following theorem.

Theorem 1. *If the drift of the queue is negative, vector $\pi(x)$ is given by*

$$\pi(x) = \begin{bmatrix} \pi_+(x) & \pi_-(x) \end{bmatrix} = p_-\mathbf{Q}_{-+}e^{\mathbf{K}x} \begin{bmatrix} \mathbf{I} & \mathbf{\Psi} \end{bmatrix}, \quad x \geq 0, \tag{6}$$

and the probability mass vector p equals to

$$p = \begin{bmatrix} 0 & p_- \end{bmatrix}, \tag{7}$$

where p_- is the solution to the set of linear equations

$$p_-(\mathbf{Q}_{--} + \mathbf{Q}_{-+}\mathbf{\Psi}) = 0, \tag{8}$$

$$p_-\mathbf{Q}_{-+}(-\mathbf{K})^{-1} \begin{bmatrix} \mathbf{I} & \mathbf{\Psi} \end{bmatrix} \mathbb{1} + p_-\mathbb{1} = 1. \tag{9}$$

Proof. The theorem is based on [16], especially on Theorem 2.2.

3.2. Busy period analysis

In this section we briefly summarize the most essential results of [17] and [18] on the busy period analysis of fluid models.

As mentioned above, $\mathbf{\Psi}$ is the phase transition probability matrix between the beginning and the end of the busy period. If the duration of the busy period is also of interest, we can introduce matrix $\mathbf{\Psi}(t)$, the time dependent counterpart of $\mathbf{\Psi}$. Entry $(\mathbf{\Psi}(t))_{i,j},\ i \in \mathcal{S}_+, j \in \mathcal{S}_-, t > 0$ is the joint probability that the duration of the busy period is less than t and the underlying Markov chain is in state j when the fluid level returns to 0 given that it was in state i when the busy period was initiated.

According to Theorem 1 of [18], the LST of $\mathbf{\Psi}(t)$, denoted by $\mathbf{\Psi}^*(s)$ satisfies the nonsymmetric algebraic Riccati equation

$$\mathbf{\Psi}^*(s)\mathbf{Q}_{-+}\mathbf{\Psi}^*(s) + \mathbf{\Psi}^*(s)\mathbf{Q}_{--} + \mathbf{Q}_{++}\mathbf{\Psi}^*(s) + \mathbf{Q}_{+-} = 2s\mathbf{\Psi}^*(s). \tag{10}$$

Let the random variable $\mathcal{B}$ denote the length of the busy period of a canonical fluid queue characterized by matrix $\mathbf{Q}$ given that the state probability vector of the background CTMC is $\kappa = \{\kappa_i, i = 1, \ldots, N_+\}$ when the busy period starts.

Theorem 2. *The LST of the busy period $f^*_{\mathcal{B}}(s) = E(e^{-s\mathcal{B}})$ is given by*

$$f^*_{\mathcal{B}}(s) = \kappa\, \mathbf{\Psi}^*(s)\mathbb{1}. \tag{11}$$

Proof. The theorem follows from the probabilistic interpretation of $\mathbf{\Psi}(t)$.

Theorem 3. *The kth moment of the busy period is given by*

$$E(\mathcal{B}^k) = \kappa\, (-1)^k \mathbf{\Psi}^{(k)}\mathbb{1}, \tag{12}$$

where $\mathbf{\Psi}^{(0)} = \mathbf{\Psi}$ and matrices $\mathbf{\Psi}^{(k)}, k > 0$ are defined recursively as

$$\begin{aligned}(\mathbf{Q}_{++}+\mathbf{\Psi}\mathbf{Q}_{-+})\mathbf{\Psi}^{(k)} &+ \mathbf{\Psi}^{(k)}(\mathbf{Q}_{--}+\mathbf{Q}_{-+}\mathbf{\Psi}) \\ &= 2k\mathbf{\Psi}^{(k-1)} - \sum_{i=1}^{k-1} \binom{k}{i} \mathbf{\Psi}^{(i)}\mathbf{Q}_{-+}\mathbf{\Psi}^{(k-i)}.\end{aligned} \tag{13}$$

Proof. (13) follows from routine derivations with $\mathbf{\Psi}^{(k)} = \frac{d^k}{ds^k}\mathbf{\Psi}^*(s)|_{s=0}$.

Since (10) is a NARE and (13) is a Sylvester equation, the LST of the busy period and the moments can be obtained in a numerically efficient way. The distribution function in time domain, $F_{\mathcal{B}}(t) = P(\mathcal{B} < t) = \kappa\,\mathbf{\Psi}(t)\mathbb{1}$ is, however, more involved to calculate. One can rely on a generic numerical Laplace transform inversion procedure, but according to our experience they are not always reliable up to the machine precision, and need complex arithmetic. Instead, a simple and elegant procedure called *Erlangization* is available [18], according to which the order-n approximation $F_{\mathcal{B}}^{(n)}(t)$ is

$$F_{\mathcal{B}}^{(n)}(t) = \int_0^\infty f_{\mathcal{E}(n,n/t)}(u) \cdot F_{\mathcal{B}}(u)\,du, \tag{14}$$

where $f_{\mathcal{E}(n,n/t)}(u)$ is the density of an order-n Erlang distribution with rate parameter $\nu = n/t$ and we have that $F_{\mathcal{B}}^{(n)}(t) \to F_{\mathcal{B}}(t)$ as $n \to \infty$. $F_{\mathcal{B}}^{(n)}(t)$ is basically the probability that the busy period is shorter than an Erlang(n, ν) variable.

Specifically for the busy period analysis $F_{\mathcal{B}}^{(n)}(t)$ can be obtained according to the next theorem.

Theorem 4. *([18], Theorem 4) The order-n approximation of the busy period distribution is*

$$F_{\mathcal{B}}^{(n)}(t) = \kappa \sum_{k=0}^{n-1} \mathbf{\Psi}_k^\nu \mathbb{1}, \tag{15}$$

where matrices $\mathbf{\Psi}_k^\nu$ are defined recursively as

$$\begin{aligned}(\mathbf{Q}_{++}+\mathbf{\Psi}_0^\nu\mathbf{Q}_{-+}-\nu\mathbf{I})\mathbf{\Psi}_k^\nu + \mathbf{\Psi}_k^\nu(\mathbf{Q}_{--}+\mathbf{Q}_{-+}\mathbf{\Psi}_0^\nu-\nu\mathbf{I}) \\ = -2\nu\mathbf{\Psi}_{k-1}^\nu - \sum_{i=1}^{k-1}\mathbf{\Psi}_i^\nu\mathbf{Q}_{-+}\mathbf{\Psi}_{k-i}^\nu,\end{aligned} \tag{16}$$

for $k > 0$, and $\mathbf{\Psi}_0^\nu$ is the solution to the NARE

$$\mathbf{\Psi}_0^\nu\mathbf{Q}_{-+}\mathbf{\Psi}_0^\nu + \mathbf{\Psi}_0^\nu(\mathbf{Q}_{--} - \nu\mathbf{I}) + (\mathbf{Q}_{++} - \nu\mathbf{I})\mathbf{\Psi}_0^\nu + \mathbf{Q}_{+-} = \mathbf{0}. \tag{17}$$

For the detailed proof of the theorem, see [18]. The idea is to construct a special fluid model which counts the number of Exp(ν) events during the busy period. Matrix $\mathbf{\Psi}_k^\nu$ is the probability that k such events occur before the end of busy period (with the usual phase-transition probabilities being the entries of the matrix). If the number of Exp(ν) events is less than n, then the busy period is shorter than an Erlang(n, ν) variable, providing (14).

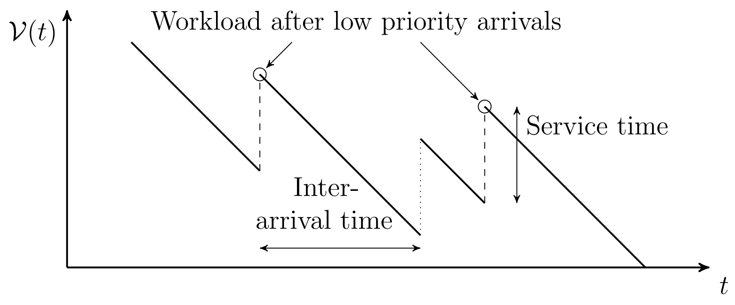

Figure 1: The workload process of the queue

4. Analysis of the preemptive resume priority queue

Our approach is based on the analysis of the workload process, just like [6] in the context of MAP/G/1 preemptive priority queues. However, by exploiting the technical simplicity of the PH distributed service times we are able to arrive to a more intuitive, simpler to implement and numerically more beneficial solution.

4.1. The workload of the system just after low priority arrival instants

For the analysis of the sojourn time we first need to derive the distribution of the workload a low priority arrival finds in the system.

The workload process $\{\mathcal{V}(t), t > 0\}$ is the amount of work in the system at time t, thus the time needed to process all the jobs in the queue if the arrival process is frozen. $\mathcal{V}(t)$ decreases by a slope of one between the arrival epochs, and jumps up at arrival epochs according to the service time requirement of the arrival; thus, $\mathcal{V}(t)$ is skip-free to the left. An example to the workload process is depicted in Figure 1. As we have two job classes, there are two kinds of jumps in the figure, the dotted one corresponds to the high, the dashed one to the low priority jobs.

To completely characterize the situation an arriving low priority job finds in the system, the stationary solution of $\{\mathcal{V}(t), \mathcal{J}(t)\}$, thus the joint distribution of the workload and the MMAP phase needs to be derived.

In our case the inter-arrival times are given by a MMAP and the size of the jumps is PH distributed, which makes it possible to apply the method of [19] to transform $\mathcal{V}(t)$, which is skip-free to the left, to $\mathcal{V}'(t)$, which is skip-free to both directions. More precisely, the continuous process with jumps $\{\mathcal{V}(t), \mathcal{J}(t)\}$, is transformed to $\{\mathcal{V}'(t), \mathcal{Z}(t)\}$ from which the stationary distribution of $\{\mathcal{V}(t), \mathcal{J}(t)\}$ at low priority arrivals is computed.

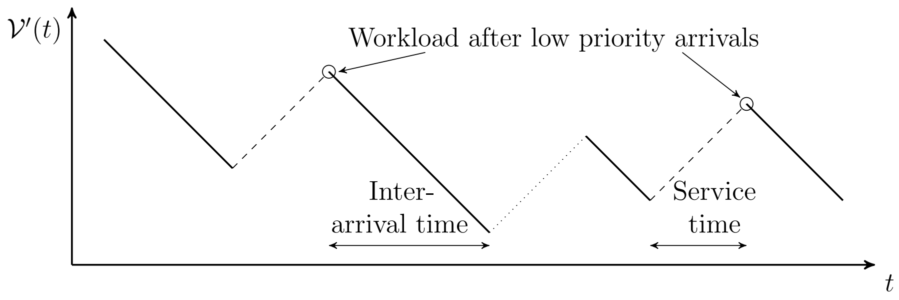

Figure 2: The modified workload process of the queue

The transformation to the skip-free process is performed as follows. Let $\{\mathcal{V}'(t), \mathcal{Z}(t)\}$ be a canonical Markovian fluid model where $\mathcal{Z}(t)$ is the underlying CTMC with generator matrix $\mathbf{Q}$ given by

$$\begin{aligned} \mathbf{Q}_{++} &= \begin{bmatrix} \mathbf{I} \otimes \mathbf{S_L} & \\ & \mathbf{I} \otimes \mathbf{S_H} \end{bmatrix}, & \mathbf{Q}_{+-} &= \begin{bmatrix} \mathbf{I} \otimes s_L \\ \mathbf{I} \otimes s_H \end{bmatrix}, \\ \mathbf{Q}_{-+} &= \begin{bmatrix} \mathbf{D_L} \otimes \sigma_L & \mathbf{D_H} \otimes \sigma_H \end{bmatrix}, & \mathbf{Q}_{--} &= \mathbf{D_0}. \end{aligned} \tag{18}$$

This fluid model behaves like $\{\mathcal{V}(t), \mathcal{J}(t)\}$ between arrivals, when it stays in the negative states $\mathcal{S}_-$. Whenever an arrival occurs, however, it switches to one of the positive state groups (depending on the class of the entering job), and accumulates the workload increment with a slope of 1. Thus, the jumps are eliminated and replaced by progressive workload accumulations. (A similar technique has been used in [20] for the analysis of a multi-type queue with impatient customers.) The transformed process obtained from Figure 1 is depicted in Figure 2.

Observe that the joint stationary density of the workload and the MMAP phase at low priority arrivals are the same in the original and in the transformed process. The stationary solution $\pi(x)$ of the transformed process (that is a canonical fluid model) is given by Theorem 1, from which, by embedding at just after low priority arrivals we get a matrix-exponential solution

$$\begin{aligned} \hat{\pi}(x) &= \frac{1}{\hat{c}}\pi(x) \begin{bmatrix} \mathbf{I} \otimes s_L \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} = \frac{1}{\hat{c}} p_- \mathbf{Q}_{-+} e^{\mathbf{K}x} \begin{bmatrix} \mathbf{I} & \mathbf{\Psi} \end{bmatrix} \begin{bmatrix} \mathbf{I} \otimes s_L \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \\ &= \underbrace{\frac{1}{\hat{c}} p_- \mathbf{Q}_{-+}}_{\hat{\beta}} e^{\mathbf{K}x} \underbrace{\begin{bmatrix} \mathbf{I} \otimes s_L \\ \mathbf{0} \end{bmatrix}}_{\hat{\mathbf{B}}} = \hat{\beta} e^{\mathbf{K}x} \hat{\mathbf{B}}, \end{aligned} \tag{19}$$

where the normalization constant is $\hat{c} = p_- \mathbf{Q}_{-+}(-\mathbf{K})^{-1}\hat{\mathbf{B}}\mathbb{1}$. Notice that from the three blocks in the last matrix term the upper two belong to $\mathcal{S}_+$ and the lower belongs to $\mathcal{S}_-$.

Due to technical reasons (which will be discussed later) the representation given by (19) will not be appropriate in the forthcoming derivations, because in general $\mathbf{K}\mathbb{1} + \hat{\mathbf{B}}\mathbb{1} \neq 0$. The following theorem provides the representation transformation that ensures the proper row-sums.

Theorem 5. *The joint density of the workload and the phase probability of the MMAP just after low priority arrivals $\hat{\pi}(x)$ can be obtained by*

$$\hat{\pi}(x) = \hat{\beta}' e^{\mathbf{K}'x}\hat{\mathbf{B}}', \tag{20}$$

with $\hat{\beta}' = \hat{\beta} \cdot diag\langle\Delta\rangle, \mathbf{K}' = diag\langle\Delta\rangle^{-1} \cdot \mathbf{K} \cdot diag\langle\Delta\rangle$ and $\hat{\mathbf{B}}' = diag\langle\Delta\rangle^{-1} \cdot \hat{\mathbf{B}}$, where $\Delta = (-\mathbf{K})^{-1}\hat{\mathbf{B}}\mathbb{1}$. Furthermore, we have that

$$\mathbf{K}'\mathbb{1} + \hat{\mathbf{B}}'\mathbb{1} = 0. \tag{21}$$

Proof. The fact that (20) equals to (19) can be proven by

$$\begin{aligned}\hat{\pi}(x) = \hat{\beta}' e^{\mathbf{K}'x}\hat{\mathbf{B}}' &= \hat{\beta} \cdot \mathrm{diag}\langle\Delta\rangle \cdot e^{\mathrm{diag}\langle\Delta\rangle^{-1}\cdot\mathbf{K}\cdot\mathrm{diag}\langle\Delta\rangle x}\mathrm{diag}\langle\Delta\rangle^{-1} \cdot \hat{\mathbf{B}} \\ &= \hat{\beta} \cdot \mathrm{diag}\langle\Delta\rangle \cdot \mathrm{diag}\langle\Delta\rangle^{-1} \cdot e^{\mathbf{K}x} \cdot \mathrm{diag}\langle\Delta\rangle\mathrm{diag}\langle\Delta\rangle^{-1} \cdot \hat{\mathbf{B}} = \hat{\beta} e^{\mathbf{K}x}\hat{\mathbf{B}}.\end{aligned} \tag{22}$$

To prove that (21) holds we have

$$\mathbf{K}'\mathbb{1} + \hat{\mathbf{B}}'\mathbb{1} = \mathrm{diag}\langle\Delta\rangle^{-1}(\mathbf{K}(-\mathbf{K})^{-1}\hat{\mathbf{B}}\mathbb{1} + \hat{\mathbf{B}}\mathbb{1}) = 0. \tag{23}$$

4.2. The sojourn time of low priority jobs

For the sojourn time analysis of low priority jobs we introduce the remaining sojourn time process $\{\mathcal{T}(t), t \geq 0\}$. At $t = 0$, $\mathcal{T}(t)$ is the workload seen by a low priority job when it arrives. For $t > 0$, $\mathcal{T}(t)$ decreases by a slope of one till a high priority arrival occurs, when $\mathcal{T}(t)$ has a jump with size given by a high priority service time. When $\mathcal{T}(t)$ reaches zero, it remains zero and the corresponding low priority job leaves the system (see Figure 3). Hence, the sojourn time of low priority jobs T_L is

$$T_L = \inf\{t > 0 : \mathcal{T}(t) = 0\}. \tag{24}$$

Just like the workload process $\mathcal{V}(t)$, the remaining sojourn time process $\mathcal{T}(t)$ is skip-free to the left and has upward jumps. As we did with the

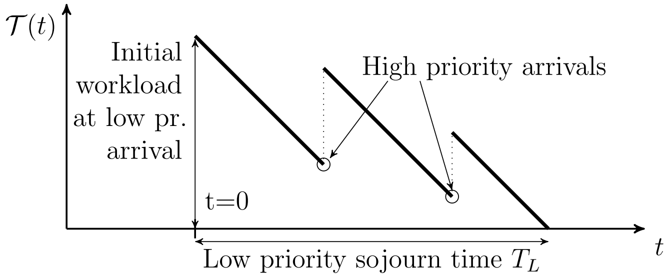
$\mathcal{T}(t)$
Initial workload at low pr. arrival
High priority arrivals
t=0
Low priority sojourn time T_L
t

Figure 3: The remaining sojourn time of a low priority job

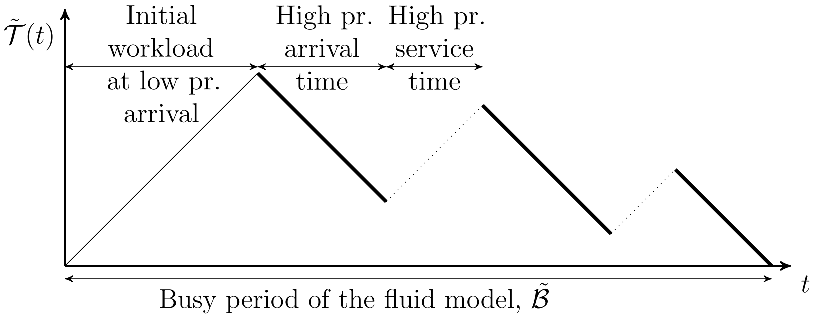
$\tilde{\mathcal{T}}(t)$
Initial workload at low pr. arrival
High pr. arrival time
High pr. service time
Busy period of the fluid model, $\tilde{\mathcal{B}}$
t

Figure 4: The fluid model for the sojourn time analysis

workload process, we transform $\mathcal{T}(t)$ to a process which is easier to handle numerically, and derive the properties of $\mathcal{T}(t)$ from the transformed process.

This transformation is based on [19] again. Let us introduce a canonical fluid model $\{\tilde{\mathcal{T}}(t), \tilde{\mathcal{Z}}(t)\}$ where the generator $\tilde{\mathbf{Q}}$ of the underlying CTMC is

$$\begin{aligned}\tilde{\mathbf{Q}}_{++} &= \begin{bmatrix}\mathbf{K}' & \\ & \mathbf{I}\otimes\mathbf{S_H}\end{bmatrix}, \quad & \tilde{\mathbf{Q}}_{+-} &= \begin{bmatrix}\hat{\mathbf{B}}' \\ \mathbf{I}\otimes s_H\end{bmatrix}, \\ \tilde{\mathbf{Q}}_{-+} &= \begin{bmatrix}\mathbf{0} & \mathbf{D_H}\otimes\sigma_H\end{bmatrix}, \quad & \tilde{\mathbf{Q}}_{--} &= \mathbf{D_0}+\mathbf{D_L},\end{aligned} \tag{25}$$

furthermore, let the distribution of $\tilde{\mathcal{Z}}(t)$ at $t = 0$ be

$$\tilde{\kappa} = \{P(\tilde{\mathcal{Z}}_+(0) = i)\} = \begin{bmatrix}\hat{\beta}' & 0\end{bmatrix}. \tag{26}$$

This fluid model has three state groups: there are two state groups in $\mathcal{S}_+$, and $\mathcal{S}_-$ is the third one.

The role of the first state group is the accumulation of the initial workload, experienced by a low priority job when it enters the system. Observe that the sojourn time density of the first state group, when started from $\tilde{\kappa}$, is exactly $\hat{\pi}(x)$, which is the density of the initial workload. The second group of states is activated when an arrival occurs, and the corresponding workload increment is accumulated. The third group of states, the negative ones represent the periods when the server is processing the low priority workload and is decreasing the remaining sojourn time of the tagged low priority job.

Note that due to Theorem 5 the usual property of Markovian generators $\tilde{\mathbf{Q}}\mathbb{1} = 0$ holds. The correctness of the solution with the non-Markovian components $\mathbf{K}'$ and $\hat{\mathbf{B}}'$ is ensured by [21].

The main idea in this section is that, by construction, the relation between the duration of the busy period $\tilde{\mathcal{B}}$ of the fluid model characterized by $(\tilde{\kappa}, \tilde{\mathbf{Q}})$ and the sojourn time of low priority jobs T_L is

$$T_L = \tilde{\mathcal{B}}/2. \tag{27}$$

This relation is clearly visible when looking at Figures 3 and 4.

Finally, the following corollary expresses the properties of the sojourn time with the properties of the busy period (detailed in Section 3.2).

Corollary 1. *The distribution of T_L in time domain, in LST domain, and its moments can be expressed by*

$$F_{T_L}(t) = P(T_L < t) = F_{\tilde{\mathcal{B}}}(2t), \tag{28}$$

$$f^*_{T_L}(s) = E(e^{-sT_L}) = f^*_{\tilde{\mathcal{B}}}(s/2), \tag{29}$$

$$E(T_L^k) = E(\tilde{\mathcal{B}}^k)/2^k. \tag{30}$$

4.3. Number of low priority jobs in the system

First we derive the distribution of the number of low priority jobs at low priority departure epochs (the corresponding random variable is denoted by X_L), then the one at a random point in time (denoted by Y_L).

When a low priority job leaves the system, the number of jobs behind it equals to the number of low priority arrivals during its sojourn in the system. To analyze this quantity, let us go back to the remaining sojourn time introduced in Section 4.2, and modify the background process of the related fluid model such that it counts the number of low priority arrivals.

Instead of $\tilde{\mathbf{Q}}$ we get $\tilde{\mathbf{Q}}'$ defined by

$$\tilde{\mathbf{Q}}' = \begin{bmatrix} \mathbf{F_0} & \mathbf{F_1} & & \\ & \mathbf{F_0} & \mathbf{F_1} & \\ & & \mathbf{F_0} & \ddots \\ & & & \ddots \end{bmatrix}, \tag{31}$$

where matrices $\mathbf{F_0}$ and $\mathbf{F_1}$ are

$$\mathbf{F_0} = \begin{bmatrix} \tilde{\mathbf{Q}}_{++} & \tilde{\mathbf{Q}}_{+-} \\ \tilde{\mathbf{Q}}_{-+} & \mathbf{D_0} \end{bmatrix}, \qquad \mathbf{F_1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D_L} \end{bmatrix}. \tag{32}$$

With this generator, matrix $\tilde{\boldsymbol{\Psi}}'$ of the corresponding canonical Markovian fluid model has an upper block-Toeplitz structure like

$$\tilde{\boldsymbol{\Psi}}' = \begin{bmatrix} \tilde{\boldsymbol{\Psi}}_\mathbf{0} & \tilde{\boldsymbol{\Psi}}_\mathbf{1} & \tilde{\boldsymbol{\Psi}}_\mathbf{2} & \cdots \\ & \tilde{\boldsymbol{\Psi}}_\mathbf{0} & \tilde{\boldsymbol{\Psi}}_\mathbf{1} & \cdots \\ & & \tilde{\boldsymbol{\Psi}}_\mathbf{0} & \cdots \\ & & & \ddots \end{bmatrix}, \tag{33}$$

where the entry $(\tilde{\boldsymbol{\Psi}}_\mathbf{i})_{k,\ell}$ is the probability that i low priority arrivals occur during the sojourn time of a low priority job and the phase of the MMAP is ℓ at the departure given that the phase was k when it entered the system.

The reason of the upper block-Toeplitz structure is that the number of low-priority arrivals during the sojourn time can only increase, and that the MMAP generating the arrivals is independent of the queue length.

Theorem 6. *Matrix $\tilde{\boldsymbol{\Psi}}_\mathbf{0}$ is the solution to the NARE*

$$\tilde{\boldsymbol{\Psi}}_\mathbf{0}\tilde{\mathbf{Q}}_{-+}\tilde{\boldsymbol{\Psi}}_\mathbf{0} + \tilde{\boldsymbol{\Psi}}_\mathbf{0}\mathbf{D_0} + \tilde{\mathbf{Q}}_{++}\tilde{\boldsymbol{\Psi}}_\mathbf{0} + \tilde{\mathbf{Q}}_{+-} = \mathbf{0}, \tag{34}$$

and for $i > 0$ matrices $\tilde{\boldsymbol{\Psi}}_\mathbf{i}$ can be obtained recursively by solving the Sylvester equation

$$(\tilde{\mathbf{Q}}_{++}+\tilde{\boldsymbol{\Psi}}_\mathbf{0}\tilde{\mathbf{Q}}_{-+})\tilde{\boldsymbol{\Psi}}_\mathbf{i} + \tilde{\boldsymbol{\Psi}}_\mathbf{i}(\mathbf{D_0}+\tilde{\mathbf{Q}}_{-+}\tilde{\boldsymbol{\Psi}}_\mathbf{0}) = -\tilde{\boldsymbol{\Psi}}_\mathbf{i-1}\mathbf{D_L} - \sum_{j=1}^{i-1}\tilde{\boldsymbol{\Psi}}_\mathbf{j}\tilde{\mathbf{Q}}_{-+}\tilde{\boldsymbol{\Psi}}_\mathbf{i-j}. \tag{35}$$

Proof. Let us partition matrix $\tilde{\mathbf{Q}}'$ according to the positive and negative states. We get

$$\tilde{\mathbf{Q}}'' = \begin{bmatrix} \tilde{\mathbf{Q}}''_{++} & \tilde{\mathbf{Q}}''_{+-} \\ \tilde{\mathbf{Q}}''_{-+} & \tilde{\mathbf{Q}}''_{--} \end{bmatrix} = \left[\begin{array}{ccc|ccc} \tilde{\mathbf{Q}}_{++} & & & \tilde{\mathbf{Q}}_{+-} & & \\ & \tilde{\mathbf{Q}}_{++} & & & \tilde{\mathbf{Q}}_{+-} & \\ & & \ddots & & & \ddots \\ \hline \tilde{\mathbf{Q}}_{-+} & & & \mathbf{D_0} & \mathbf{D_L} & \\ & \tilde{\mathbf{Q}}_{-+} & & & \mathbf{D_0} & \mathbf{D_L} \\ & & \ddots & & & \ddots \end{array}\right]. \tag{36}$$

Substituting (36) and (33) into the NARE (4) provides the theorem after some algebraic manipulation.

The probabilities for the number of low priority jobs at low priority departures $x_i^L = P(X_L = i)$ are obtained from $\tilde{\mathbf{\Psi}}_\mathbf{i}$ by taking into consideration the initial probability vector of the busy period $\tilde{\kappa}$. For later use, we also introduce row vector $\underline{x}_i^L = \{P(X_L = i, \mathcal{J} = j), j = 1, \ldots, N_A\}$, the joint probability of the number of jobs and the phase of the MMAP at departures (obviously, $x_i^L = \underline{x}_i^L \mathbb{1}$).

Corollary 2. *For the distribution of the number of low priority jobs at low priority departures we have*

$$\underline{x}_i^L = \tilde{\kappa} \tilde{\mathbf{\Psi}}_\mathbf{i}. \tag{37}$$

The significance of (37) lies in the fact that the consecutive queue length probabilities can be obtained by consecutive solutions of Sylvester equations calculating $\tilde{\mathbf{\Psi}}_\mathbf{i}$. The prior procedures of the related literature are far more expensive computationally.

Corollary 3. *The generating function (GF) of the distribution of the number of jobs at departures $X_L(z) = \sum_{i=0}^{\infty} z^i \underline{x}_i^L$ can be obtained by*

$$X_L(z) = \tilde{\kappa} \tilde{\mathbf{\Psi}}(z), \tag{38}$$

where matrix $\tilde{\mathbf{\Psi}}(z)$ satisfies the NARE

$$\tilde{\mathbf{\Psi}}(z) \tilde{\mathbf{Q}}_{-+} \tilde{\mathbf{\Psi}}(z) + \tilde{\mathbf{\Psi}}(z)(\mathbf{D_0} + z\mathbf{D_L}) + \tilde{\mathbf{Q}}_{++} \tilde{\mathbf{\Psi}}(z) + \tilde{\mathbf{Q}}_{+-} = \mathbf{0}. \tag{39}$$

Proof. Multiplying (35) by z^i, summing it from 1 to infinity, then adding (34) provides (39).

Finally, the factorial moments of X_L can be calculated by taking the derivatives of the generating function, hence

$$E(X_L^k) = \sum_{i=0}^{\infty} i^k x_i^L = \frac{d^k}{dz^k} X_L(z)|_{z=1}\mathbb{1}, \tag{40}$$

yielding a recursion introduced by the next corollary.

Corollary 4. *For the kth factorial moment of X_L we have*

$$\underline{E(X_L^k)} = \tilde{\kappa}\tilde{\mathbf{\Psi}}^{(k)}, \qquad E(X_L^k) = \underline{E(X_L^k)}\mathbb{1}, \tag{41}$$

where $\tilde{\mathbf{\Psi}}^{(k)} = \frac{d^k}{dz^k}\tilde{\mathbf{\Psi}}(z)|_{z=1}$. Matrix $\tilde{\mathbf{\Psi}}^{(0)} = \tilde{\mathbf{\Psi}}$ and for $k > 0$ matrices $\tilde{\mathbf{\Psi}}^{(k)}$ are obtained recursively by solving the following Sylvester equations

$$\begin{aligned}(\tilde{\mathbf{Q}}_{++}+\tilde{\mathbf{\Psi}}^{(0)}\tilde{\mathbf{Q}}_{-+})\tilde{\mathbf{\Psi}}^{(k)} + \tilde{\mathbf{\Psi}}^{(k)}(\tilde{\mathbf{Q}}_{--}+\tilde{\mathbf{Q}}_{-+}\tilde{\mathbf{\Psi}}^{(0)})& \\ = -k\tilde{\mathbf{\Psi}}^{(k-1)}\mathbf{D_L} - \sum_{i=1}^{k-1}\binom{k}{i}\tilde{\mathbf{\Psi}}^{(i)}\tilde{\mathbf{Q}}_{-+}\tilde{\mathbf{\Psi}}^{(k-i)}&.\end{aligned} \tag{42}$$

In the rest of the section we calculate various properties of the number of low priority jobs at random point in time denoted by Y_L. Our contribution in this subsection ends here, since the relations between X_L and Y_L are extensively studied in [7], which we adopt in this paper, and provide them for the sake of completeness.

Let us introduce row vector $\underline{y}_i^L = \{P(Y_L = i, \mathcal{J} = j), j = 1, \ldots, N_A\}$.

Theorem 7. *([7], Theorem 4.6) The generating function of $\underline{y}_i^L$, denoted by $Y_L(z) = \sum_{i=0}^{\infty} z^i \underline{y}_i^L$ is related to $X_L(z)$ as*

$$Y_L(z)(\mathbf{D_0} + \mathbf{D_H} + z\mathbf{D_L}) = \lambda_L(z-1)X_L(z). \tag{43}$$

Corollary 5. *([7], Corollary 3.11) Vectors $\underline{y}_i^L, i \geq 0$ are recursively obtained by*

$$\begin{aligned}\underline{y}_0^L &= \lambda_L \underline{x}_0^L(-\mathbf{D_0} - \mathbf{D_H})^{-1}, \\ \underline{y}_i^L &= (\underline{y}_{i-1}^L\mathbf{D_L} + \lambda_L\underline{x}_i^L - \lambda_L\underline{x}_{i-1}^L)(-\mathbf{D_0} - \mathbf{D_H})^{-1}, \qquad i > 0.\end{aligned} \tag{44}$$

Corollary 6. *([7], Corollary 3.10) The factorial moments of the number of low priority jobs at random point in time are obtained recursively as*

$$\begin{aligned}E(Y_L^k) &= E(X_L^k) + k\big(\underline{E(X_L^{k-1})} - \underline{E(Y_L^{k-1})}\mathbf{D_L}/\lambda_L\big)(\mathbb{1}\theta - \mathbf{D})^{-1}\mathbf{D_L}\mathbb{1}, \\ \underline{E(Y_L^k)} &= E(Y_L^k)\theta + k\big(\underline{E(Y_L^{k-1})}\mathbf{D_L} - \lambda_L\underline{E(X_L^{k-1})}\big)(\mathbb{1}\theta - \mathbf{D})^{-1},\end{aligned} \tag{45}$$

for $k > 0$, and $\underline{E(Y_L^0)} = \theta$.

4.4. The analysis of the high priority class

In case of the preemptive resume service policy the high priority class can be analyzed in separation, as a single-class MAP/PH/1 queue with arrival process given by matrices $(\mathbf{D_0} + \mathbf{D_L}, \mathbf{D_H})$ and service time distribution given by $(\sigma_H, \mathbf{S_H})$. The number of jobs in the system is matrix-exponentially distributed, and can be derived from the solution of a QBD (see Section 4.4.1 in [22]). The sojourn time of the jobs in a MAP/PH/1 queue is matrix exponentially distributed, as proven in [23] based on the analysis of the age process.

4.5. Extensions of the model

The two-class analysis procedure developed here can be used to solve models with more than two classes as well. When analyzing the ith class, all lower priority classes can be neglected, only the higher priority classes need to be taken into account. The details are provided in Appendix A.

The presented approach can be generalized to handle correlated service times as well, thus when the service process is a MAP. In this case the state space of the Markov chains corresponding to the fluid models have to be extended such that the phase of the service MAP is preserved in the negative states.

5. Analysis of the non-preemptive priority queue

In the non-preemptive case the service of a low priority job can not be interrupted. It turns out, that the analysis approach developed in Section 4 can still be used with a small difference. Instead of analyzing the sojourn time and the number of jobs in the system, in the non-preemptive case we will focus on the *waiting time* (which can be interrupted by a high priority arrival any time) and the *number of waiting jobs* in the system. The non-interruptible service time and the number of arrivals during it will be added afterwards to obtain the sojourn time and the number of jobs in the system.

5.1. The workload of the system just before low priority arrival instants

When a low priority job enters the system, its waiting time equals to the workload of the system just before its arrival (thus without its own service time) plus the service times of all high priority jobs arrived during waiting in the queue. To find out the workload just before the arrival in the example of Figure 1 this means that we need the distribution of $\mathcal{V}(t)$ just before the jumps, instead of just after the jumps.

This distribution can be obtained by applying the same transformation procedure which results in a canonical Markovian fluid model with stationary fluid density $\pi(x)$ and probability mass at level zero p. Embedding right before low priority arrivals we get the density

$$\begin{aligned}\breve{\pi}(x) = \frac{1}{\breve{c}}\pi(x)\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{D_L}\end{bmatrix} &= \frac{1}{\breve{c}}p_-\mathbf{Q}_{-+}e^{\mathbf{K}x}\begin{bmatrix}\mathbf{I} & \mathbf{\Psi}\end{bmatrix}\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{D_L}\end{bmatrix}\\ &= \underbrace{\frac{1}{\breve{c}}p_-\mathbf{Q}_{-+}}_{\breve{\beta}}e^{\mathbf{K}x}\underbrace{\mathbf{\Psi}\mathbf{D_L}}_{\breve{\mathbf{B}}} = \breve{\beta}e^{\mathbf{K}x}\breve{\mathbf{B}}.\end{aligned} \tag{46}$$

Notice that the workload just before the arrival can be exactly zero as well, with probability mass

$$\breve{p} = \frac{1}{\breve{c}}p_-\mathbf{D_L}. \tag{47}$$

The normalization constant is $\breve{c} = p_-\mathbf{D_L}\mathbb{1} + p_-\mathbf{Q}_{-+}(-\mathbf{K})^{-1}\breve{\mathbf{B}}\mathbb{1}$.

Similar to Theorem 5, it is again possible to similarity transform the representation $\breve{\beta}, \mathbf{K}$ and $\breve{\mathbf{B}}$ to $\breve{\beta}', \mathbf{K}'$ and $\breve{\mathbf{B}}'$ such that $\mathbf{K}'\mathbb{1} + \breve{\mathbf{B}}'\mathbb{1} = 0$ holds.

5.2. The sojourn time of low priority jobs

As mentioned before, first the waiting time (denoted by W_L) is characterized, then the service time is added afterwards to get the sojourn time.

As done in Section 4.2, it is possible to introduce the remaining waiting time process $\mathcal{W}(t)$ and construct a canonical fluid model $\{\bar{\mathcal{W}}(t), \bar{\mathcal{Z}}(t)\}$ whose busy period $\bar{\mathcal{B}}$ is closely related to the waiting time. The blocks of the generator of this fluid model are

$$\bar{\mathbf{Q}}_{++} = \begin{bmatrix}\mathbf{K}' & \\ & \mathbf{I}\otimes\mathbf{S_H}\end{bmatrix}, \quad \bar{\mathbf{Q}}_{+-} = \begin{bmatrix}\breve{\mathbf{B}}'\\ \mathbf{I}\otimes s_H\end{bmatrix}, \tag{48}$$

$$\bar{\mathbf{Q}}_{-+} = \begin{bmatrix}\mathbf{0} & \mathbf{D_H}\otimes\sigma_H\end{bmatrix}, \quad \bar{\mathbf{Q}}_{--} = \mathbf{D_0} + \mathbf{D_L}, \tag{49}$$

and the distribution of $\bar{\mathcal{Z}}(t)$ at $t = 0$ (that defines the initial distribution of the busy period) is

$$\bar{\kappa} = \{P(\bar{\mathcal{Z}}_+(0) = i)\} = \begin{bmatrix}\breve{\beta}' & 0\end{bmatrix}. \tag{50}$$

Notice that everything is the same as in Section 4.2, except the parameters of the initial workload distribution. Hence, it is not surprising that $W_L = \bar{\mathcal{B}}/2$.

Corollary 7. *The distribution of* W_L *in time domain, in LST domain, and its moments can be expressed by*

$$F_{W_L}(t) = F_{\bar{\mathcal{B}}}(2t), \qquad f^*_{W_L}(s) = f^*_{\bar{\mathcal{B}}}(s/2), \qquad E(W_L^k) = E(\bar{\mathcal{B}}^k)/2^k. \tag{51}$$

As $T_L = W_L + S_L$ holds, it is straight forward to obtain the LST of the distribution of T_L and its moments.

Corollary 8. *The LST of the distribution of* T_L *is given by*

$$f^*_{T_L}(s) = f^*_{W_L}(s) f^*_{S_L}(s). \tag{52}$$

Taking the derivatives of $f^*_{T_L}(s)$ with respect to s and tending $s \to 0$ yields the moments of the sojourn time.

Corollary 9. *The moments of the sojourn time* T_L *are given by*

$$E(T_L^k) = \sum_{i=0}^{k} \binom{k}{i} E(W_L^i) E(S_L^{k-i}). \tag{53}$$

The distribution function of the sojourn time is more involved to obtain. One could directly express it as a continuous time convolution of $F_{W_L}(t)$ and $f_{S_L}(t)$, but it would involve an integral which can be evaluated only numerically. Remind that both $F_{T_L}(t)$ in the preemptive resume case and $F_{W_L}(t)$ in the non-preemptive case are derived from the distribution of the busy period of an appropriate fluid model, which is computed in terms of Erlangization (see Section 3.2), meaning that an order-n approximation is applied where increasing n improves the accuracy. For the preemptive resume case we had that the order-n approximation is

$$F^{(n)}_{T_L,preemp.}(t) = P(\tilde{\mathcal{B}}/2 < \text{Erlang}(n, \frac{n}{t})) = P(\tilde{\mathcal{B}} < \text{Erlang}(n, \frac{n}{2t})) = \tilde{\kappa} \sum_{k=0}^{n-1} \tilde{\mathbf{\Psi}}_k^{\nu} \mathbb{1},$$

with $\nu = n/(2t)$ and $\tilde{\kappa}\tilde{\mathbf{\Psi}}_k^{\nu}\mathbb{1}$ holding the probabilities that k Exp(ν) events occur during the busy period.

In the non-preemptive case, however, busy period $\bar{\mathcal{B}}$ corresponds to the waiting time only. Thus, we have that the sojourn time distribution is

$$F^{(n)}_{T_L}(t) = P(\bar{\mathcal{B}}/2 + S_L < \text{Erlang}(n, \frac{n}{t})).$$

Theorem 8. *The order-n approximation of the distribution function of T_L is*

$$F_{T_L}^{(n)}(t) = \bar{\kappa} \sum_{k=0}^{n-1} \bar{\mathbf{\Psi}}_k^{\nu} \mathbb{1} d_{n-k} + \check{p} \mathbb{1} d_n, \tag{54}$$

where $\nu = n/(2t)$, matrices $\bar{\mathbf{\Psi}}_k^{\nu}$ are defined by Theorem 4 with using $\bar{\mathbf{Q}}$ instead of $\mathbf{Q}$, and probabilities d_n are given by

$$d_n = 1 - \sigma_L \big(\mathbf{I} - \mathbf{S_L}/(2\nu)\big)^{-n} \mathbb{1}. \tag{55}$$

Proof. We have that

$$\begin{aligned} F_{T_L}^{(n)}(t) &= P(\bar{\mathcal{B}}/2 + S_L < \text{Erlang}(n, \frac{n}{t})) = P(\bar{\mathcal{B}} + 2S_L < \text{Erlang}(n, \nu)) \\ &= \bar{\kappa} \sum_{k=0}^{n-1} \bar{\mathbf{\Psi}}_k^{\nu} \mathbb{1} \cdot \underbrace{P(2S_L < \text{Erlang}(n-k, \nu))}_{d_{n-k}} + \check{p} \mathbb{1} \cdot \underbrace{P(2S_L < \text{Erlang}(n, \nu))}_{d_n}, \end{aligned}$$

where the second term corresponds to the case when $W_L = 0$. The d_ℓ probabilities can be derived as

$$\begin{aligned} d_n &= P(2S_L < \text{Erlang}(n, \nu)) = P(S_L < \text{Erlang}(n, 2\nu)) \\ &= 1 - \int_{u=0}^{\infty} \frac{(2\nu u)^{n-1}}{(n-1)!} 2\nu e^{-2\nu u} \sigma_L e^{\mathbf{S_L} u} \mathbb{1} du \\ &= 1 - \frac{(-\nu)^{n-1}}{(n-1)!} 2\nu \sigma_L \int_{u=0}^{\infty} \frac{d^{n-1}}{d\nu^{n-1}} e^{-2\nu u} e^{\mathbf{S_L} u} \mathbb{1} du \\ &= 1 - \frac{(-\nu)^{n-1}}{(n-1)!} 2\nu \sigma_L \frac{d^{n-1}}{d\nu^{n-1}} (2\nu \mathbf{I} - \mathbf{S_L})^{-1} \mathbb{1} = 1 - (2\nu)^n \sigma_L (2\nu \mathbf{I} - \mathbf{S_L})^{-n} \mathbb{1}, \end{aligned}$$

that equals to (55).

5.3. The number of low priority jobs

As in the preemptive resume case, first the number of low priority jobs at low priority departures is analyzed, from which the results corresponding to a random point in time are derived.

To obtain the number of low priority jobs at low priority departures (X_L) a tagged low priority job is picked, and the number of low priority arrivals is counted during its stay in the system. This quantity consists of two components: the number of arrivals during the waiting time, and the number of additional arrivals during the service time.

The number of arrivals during the waiting time can be derived from the fluid model representing the remaining waiting time process introduced in Section 5.2. We follow the exactly same recipe as in Section 4.3 with the preemptive case, thus we modify the background process of the fluid model $\bar{\mathbf{Q}}$ such that it counts the number of arrivals during the busy period and get $\bar{\mathbf{Q}}'$. The blocks of the corresponding $\bar{\mathbf{\Psi}}'$ matrix, $\bar{\mathbf{\Psi}}_{\mathbf{k}}$ are holding the probabilities that k arrivals occurred during the busy period (that is, during the waiting time) given the initial phase of the MMAP. These matrices can be calculated as Theorem 6 does in the preemptive resume case, the only difference is that matrix $\bar{\mathbf{Q}}$ needs to be used instead of matrix $\tilde{\mathbf{Q}}$.

As for the second component, let us introduce matrices $\mathbf{A}_{\mathbf{i}}, i \geq 0$ whose (k, ℓ)th entry is the probability that the MMAP generates i low priority arrivals during a low priority service time starting from phase k and the MMAP phase at the end of service is ℓ. Matrices $\mathbf{A}_{\mathbf{i}}$ are matrix-geometric

$$\mathbf{A}_{\mathbf{i}} = \alpha \cdot \mathbf{A}^{i}\mathbf{a}, \quad i \geq 0, \tag{56}$$

where

$$\alpha = \mathbf{I} \otimes \sigma_L, \tag{57}$$

$$\mathbf{A} = \left(-(\mathbf{D}_{\mathbf{0}} + \mathbf{D}_{\mathbf{H}}) \oplus \mathbf{S}_{\mathbf{L}}\right)^{-1} (\mathbf{D}_{\mathbf{L}} \otimes \mathbf{I}), \tag{58}$$

$$\mathbf{a} = \left(-(\mathbf{D}_{\mathbf{0}} + \mathbf{D}_{\mathbf{H}}) \oplus \mathbf{S}_{\mathbf{L}}\right)^{-1} (\mathbf{I} \otimes s_L). \tag{59}$$

Theorem 9. *The joint probability of the number of low priority jobs in the system and the phase of the MMAP at low priority departure instants is*

$$\underline{x}_i^L = h_i \cdot \mathbf{a} + \check{p}\mathbf{A}_{\mathbf{i}}, \tag{60}$$

where matrix $h_0 = \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{0}}$ *and* $h_i, i > 0$ *is defined recursively as*

$$h_i = h_{i-1} \cdot \mathbf{A} + \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{i}}\alpha. \tag{61}$$

Proof. Let us sum the number of arrivals during the waiting time and during the service time by convolution, yielding

$$\underline{x}_i^L = \sum_{k=0}^{i} \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{k}}\mathbf{A}_{\mathbf{i-k}} + \check{p}\mathbf{A}_{\mathbf{i}} = \underbrace{\sum_{k=0}^{i} \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{k}}\alpha\mathbf{A}^{i-k}}_{h_i}\mathbf{a} + \check{p}\mathbf{A}_{\mathbf{i}}. \tag{62}$$

The recursion for h_i can be shown by

$$h_i = \sum_{k=0}^{i} \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{k}}\alpha\mathbf{A}^{i-k} = \underbrace{\sum_{k=0}^{i-1} \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{k}}\alpha\mathbf{A}^{i-1-k}}_{h_{i-1}} \cdot\mathbf{A} + \bar{\kappa}\bar{\mathbf{\Psi}}_{\mathbf{i}}\alpha. \tag{63}$$

By introducing the generating functions $\bar{\mathbf{\Psi}}(z) = \sum_{i=0}^{\infty} z^i \bar{\mathbf{\Psi}}_{\mathbf{i}}$ and $\mathbf{A}(z) = \sum_{i=0}^{\infty} z^i \mathbf{A}_{\mathbf{i}}$, the generating function $X_L(z) = \sum_{i=0}^{\infty} z^i \underline{x}_i^L$ is easy to obtain from (62) and (56).

Corollary 10. *$X_L(z)$ is expressed by*

$$X_L(z) = \bar{\kappa}\bar{\mathbf{\Psi}}(z)\mathbf{A}(z) + \check{p}\mathbf{A}(z), \tag{64}$$

where matrix $\mathbf{A}(z) = \sum_{i=0}^{\infty} z^i \mathbf{A}_{\mathbf{i}}$ has the following closed form formula

$$\mathbf{A}(z) = \alpha(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{a}. \tag{65}$$

Based on (40) the factorial moments at departures are calculated by routine derivations of (64).

Corollary 11. *For the kth factorial moment of the number of low priority jobs at low priority departures we have*

$$\underline{E(X_L^k)} = \sum_{i=0}^{k} \binom{k}{i} \bar{\kappa}\bar{\mathbf{\Psi}}^{(i)}\mathbf{A}^{(k-i)} + \check{p}\mathbf{A}^{(k)}, \tag{66}$$

where matrices $\bar{\mathbf{\Psi}}^{(i)} = \frac{d^i}{dz^i}\bar{\mathbf{\Psi}}(z)|_{z=1}$ are obtained similar to (42) and matrices $\mathbf{A}^{(i)} = \frac{d^i}{dz^i}\mathbf{A}(z)|_{z=1}$ have the following closed form:

$$\mathbf{A}^{(i)} = i!\alpha(\mathbf{I} - \mathbf{A})^{-i-1}\mathbf{A}^i\mathbf{a}. \tag{67}$$

Having characterized the number of low priority jobs at low priority departure epochs, the properties of the number of low priority jobs at a random point in time are given by Theorem 7 and Corollaries 5 and 6.

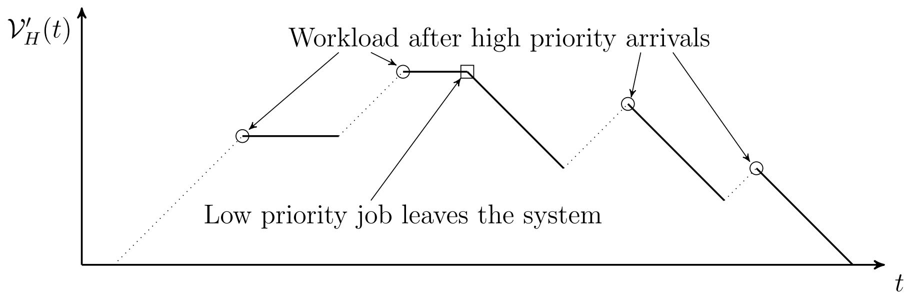

Figure 5: The modified workload process of the high priority class

5.4. The analysis of the high priority class

In the non-preemptive case the high priority class can not be analyzed in separation, since a high priority job can not be served immediately when a low priority job is in the server.

We use the workload process of the high priority class denoted by $\{\mathcal{V}_H(t), t > 0\}$ to derive the performance measures[1]. The trajectory of $\mathcal{V}_H(t)$ contains intervals where the slope is zero corresponding to the periods when the server serves low priority jobs. As before, $\mathcal{V}_H(t)$ is transformed to a fluid model (see Figure 5 for an example).

The blocks of the generator matrix of this fluid model are defined by

$$\mathbf{Q}^H_{++} = \begin{bmatrix} \mathbf{I} \otimes \mathbf{I} \otimes \mathbf{S_H} & \\ & \mathbf{I} \otimes \mathbf{S_H} \end{bmatrix}, \quad \mathbf{Q}^H_{+-} = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \otimes s_H \end{bmatrix}, \quad \mathbf{Q}^H_{+0} = \begin{bmatrix} \mathbf{I} \otimes \mathbf{I} \otimes s_H \\ \mathbf{0} \end{bmatrix},$$
$$\mathbf{Q}^H_{-+} = \begin{bmatrix} \mathbf{0} & \mathbf{D_H} \otimes \sigma_H \end{bmatrix}, \quad \mathbf{Q}^H_{--} = \mathbf{D_0} + \mathbf{D_L}, \quad \mathbf{Q}^H_{-0} = \mathbf{0},$$
$$\mathbf{Q}^H_{0+} = \begin{bmatrix} \mathbf{D_H} \otimes \mathbf{I} \otimes \sigma_H & \mathbf{0} \end{bmatrix}, \quad \mathbf{Q}^H_{0-} = \mathbf{I} \otimes s_L, \quad \mathbf{Q}^H_{00} = (\mathbf{D_0} + \mathbf{D_L}) \oplus \mathbf{S_L}.$$

Four state groups can be identified in the generator. The two state groups of $\mathcal{S}_+$ both correspond to the workload accumulation due to a new high priority arrival. The difference is that in the first state group the server works on a low priority job, thus the phase of its service needs to be maintained during the workload accumulation. In the negative states $\mathcal{S}_-$ the server is working on a high, in the zero states $\mathcal{S}_0$ the server is working on a low priority job.

The probability of the phases when the workload process leaves level zero, denoted by vector κ^H, is not easy to obtain. Regarding this vector we are

[1] Contrary to Sections 4.1 and 5.1, where the workload process of the entire system is discussed, the workload process considered here applies only to the high priority class.

relying on the results of [7], which we re-formulate and simplify at several points due to the PH distributed service times.

Let us investigate the system at the departures that leave the high priority queue empty, and introduce two probability vectors, ϕ and ϕ_0 associated to this embedded process. The ith entry of ϕ_0 is the probability that the whole system is empty at the embedded instant and the phase of the MMAP is i. Entry i of vector ϕ is the probability that the embedded process is in state i in the product space of the MMAP phase and the phase of the low priority service time.

Theorem 10. *Vector ϕ_0 is given by*

$$\phi_0 = \frac{(1-\rho)p_-(-\mathbf{D_0})}{\lambda_L p_- \mathbb{1} + (1-\rho)p_- \mathbf{D_H} \mathbb{1}}, \tag{68}$$

where p_- is the probability mass vector of the fluid queue representing the workload process of the whole system (see Sections 4.1 and 5.1).

Vector ϕ is the unique solution to the linear system

$$\begin{aligned}
\phi = {} & (\phi-\phi_0)(\mathbf{I}\otimes\sigma_L)(-(\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_L})^{-1}\begin{bmatrix}\mathbf{D_H}\otimes\mathbf{I}\otimes\sigma_H & \mathbf{0}\end{bmatrix}\mathbf{\Psi}^H \\
& + (\phi-\phi_0)(\mathbf{I}\otimes\sigma_L)(-(\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_L})^{-1}(\mathbf{I}\otimes s_L) \\
& + \phi_0(-\mathbf{D_0})^{-1}(\mathbf{D_L}\otimes\sigma_L)(-(\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_L})^{-1}\begin{bmatrix}\mathbf{D_H}\otimes\mathbf{I}\otimes\sigma_H & \mathbf{0}\end{bmatrix}\mathbf{\Psi}^H \qquad (69) \\
& + \phi_0(-\mathbf{D_0})^{-1}(\mathbf{D_L}\otimes\sigma_L)(-(\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_L})^{-1}(\mathbf{I}\otimes s_L) \\
& + \phi_0(-\mathbf{D_0})^{-1}\begin{bmatrix}\mathbf{0} & \mathbf{D_H}\otimes\sigma_H\end{bmatrix}\mathbf{\Psi}^H, \\
\phi\mathbb{1} = {} & 1, \qquad (70)
\end{aligned}$$

where $\mathbf{\Psi}^H$ is the solution of the NARE

$$\begin{aligned}
\mathbf{\Psi}^H\mathbf{Q}^H_{-+}\mathbf{\Psi}^H + \mathbf{\Psi}^H\mathbf{Q}^H_{--} + (\mathbf{Q}^H_{++} + \mathbf{Q}^H_{+0}(-\mathbf{Q}^H_{00})^{-1}\mathbf{Q}^H_{0+})\mathbf{\Psi}^H & \\
+ \mathbf{Q}^H_{+-} + \mathbf{Q}^H_{+0}(-\mathbf{Q}^H_{00})^{-1}\mathbf{Q}^H_{0-} = \mathbf{0}. &
\end{aligned} \tag{71}$$

Proof. Eq. (68) follows from [7],Theorem 3.1 and [7],Lemma 3.2.

Eq. (69) has 5 terms. The first one corresponds to the case when there are low priority jobs in the system when the last high priority job leaves. The server starts to serve a low priority job. The PH of the service process and the MMAP evolve together, and the MMAP generates a high priority arrival before the current service is completed, and initiates the workload process (see Figure 5). The next departure leaving the high priority class empty occurs when the workload of the high priority class returns to level zero, with the corresponding phase transitions given by $\mathbf{\Psi}^H$ (which satisfies the usual NARE after censoring out the zero states). According to the second term the low priority service is completed before the MMAP generates a high priority job, providing the phase of the next embedded point. In the third and fourth term the last high priority job leaves the system empty, and the next arriving job is a low priority one, while in the last term the next arriving job is a high priority one.

Let us introduce vectors q_L^H and q_0^H as the stationary phase probabilities that the server is working on a low priority job and that the system is idle when there are no high priority jobs in the system, respectively. These probability vectors can be obtained from ϕ and ϕ_0 by taking into account the mean amount of time spent in various phases in the system, yielding

$$\begin{aligned} q_L^H &= \frac{1}{c^H}(\phi - \phi_0 + \phi_0(-\mathbf{D_0})^{-1}\mathbf{D_L})(\mathbf{I}\otimes\sigma_L)(-(\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_L})^{-1}, \\ q_0^H &= \frac{1}{c^H}\phi_0(-\mathbf{D_0})^{-1}, \end{aligned} \tag{72}$$

where c^H is a normalization constant. From these vectors the initial phase distribution vector for the high priority workload process denoted by κ^H is given by

$$\kappa^H = q_L^H \begin{bmatrix}\mathbf{D_H}\otimes\mathbf{I}\otimes\sigma_H & \mathbf{0}\end{bmatrix} + q_0^H \begin{bmatrix}\mathbf{0} & \mathbf{D_H}\otimes\sigma_H\end{bmatrix} = q_L^H\mathbf{Q}_{0+}^H + q_0^H\mathbf{Q}_{-+}^H. \tag{73}$$

Finally, the next two theorems provide the performance measures for the high priority jobs.

Theorem 11. *The probability density function of the sojourn time of high priority jobs $f_{T_H}(t)$ is matrix-exponential*

$$f_{T_H}(t) = \zeta e^{\mathbf{Z}t}v, \tag{74}$$

with parameters

$$\zeta = \begin{bmatrix}\kappa^H & 0\end{bmatrix}/c, \qquad \mathbf{Z} = \begin{bmatrix}\mathbf{K}^H & \begin{bmatrix}\mathbb{1}\otimes\mathbf{I}\otimes s_H \\ \mathbf{0}\end{bmatrix} \\ \mathbf{0} & \mathbf{S_L}\end{bmatrix}, \qquad v = \begin{bmatrix}0 \\ \mathbb{1}\otimes s_H \\ s_L\end{bmatrix}, \tag{75}$$

where $\mathbf{K}^H = \mathbf{Q}_{++}^H + \mathbf{Q}_{+0}^H(-\mathbf{Q}_{00}^H)^{-1}\mathbf{Q}_{0+}^H + \mathbf{\Psi}^H\mathbf{Q}_{-+}^H$ and c is the normalization constant.

Proof. The density of the workload at high priority arrival including the service time requirement the job brought to the system is $\kappa^H e^{\mathbf{K}^H x}\mathbf{Q}^H_{+0}$ if the server works on a low priority job and it is $\kappa^H e^{\mathbf{K}^H x}\mathbf{Q}^H_{+-}$ otherwise (see the points marked by circles in Figure 5). In the latter case the sojourn time of the entering job is x. In the former case, however, the remaining service time of the low priority job has to be taken into account as well. The phase of the low priority service is also encoded in the background process, hence we have

$$f_{T_H}(t) = \left(\kappa^H \int_{x=0}^{\infty} e^{\mathbf{K}^H x}\mathbf{Q}^H_{+0}(\mathbb{1}\otimes\mathbf{I})e^{\mathbf{S_L}(t-x)}s_L\,dt + \kappa^H e^{\mathbf{K}^H x}\mathbf{Q}^H_{+-}\mathbb{1}\right)/c. \quad (76)$$

The convolution of the two matrix exponentials with parameters $\mathbf{K}^H$ and $\mathbf{S_L}$ can be represented by a single matrix exponential with parameter $\mathbf{Z}$ according to [24]. The second term can be expressed using $\zeta e^{\mathbf{Z}t}$ as well, by adding transitions from the first matrix block to the absorbing state with rates $\mathbf{Q}^H_{+-}\mathbb{1} = \begin{bmatrix} 0 \\ \mathbb{1}\otimes s_H \end{bmatrix}$. Putting together the two terms provides the theorem.

Corollary 12. *The LST of the distribution function and the moments of T_H are given by*

$$f^*_{T_H}(s) = \zeta(s\mathbf{I}-\mathbf{Z})^{-1}v, \qquad E(T_H^k) = k!\zeta(-\mathbf{Z})^{-k-1}v. \quad (77)$$

For the analysis of the number of high priority jobs in the system we introduce a quasi birth-death process (QBD, [14]), where the matrices corresponding to level backward, local and level forward transitions (denoted by $\mathbf{A}_-$, $\mathbf{A_0}$ and $\mathbf{A}_+$, respectively) are

$$\mathbf{A_0} = \begin{bmatrix} (\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_L} & \mathbf{I}\otimes s_L\sigma_H \\ & (\mathbf{D_0}+\mathbf{D_L})\oplus\mathbf{S_H} \end{bmatrix},$$

$$\mathbf{A}_- = \begin{bmatrix} & \\ & \mathbf{I}\otimes s_H\sigma_H \end{bmatrix}, \qquad \mathbf{A}_+ = \begin{bmatrix} \mathbf{D_H}\otimes\mathbf{I} & \\ & \mathbf{D_H}\otimes\mathbf{I} \end{bmatrix}.$$

In the first group of states the server is working on a low, in the second one it is working on a high priority job. It is possible to move from the first state group to second one (see matrix $\mathbf{A_0}$), but not the way around at levels > 0.

The entries of vector $\underline{y}_i^H$ are the probabilities that there are i high priority jobs in the system and the background process is in different phases. It is well known that QBDs have a matrix geometric distribution.

Theorem 12. *Vectors $\underline{y}_i^H$ have the following matrix geometric form:*

$$\underline{y}_i^H = \underline{y}_0^H \mathbf{R}^i, \tag{78}$$

where matrix $\mathbf{R}$ is the minimal non-negative solution to the matrix-quadratic equation

$$\mathbf{A}_+ + \mathbf{R}\mathbf{A_0} + \mathbf{R}^2\mathbf{A}_- = \mathbf{0}, \tag{79}$$

and the probability of level 0 is $y_0^H = \begin{bmatrix} q_L^H & q_0^H \end{bmatrix} / c'$. The normalization constant is $c' = \begin{bmatrix} q_L^H & q_0^H \end{bmatrix} (\mathbf{I} - \mathbf{R})^{-1}\mathbb{1}$.

Proof. By definition in (72), vectors q_L^H and q_0^H are the stationary phase probability vectors given that there are no high priority jobs in the system. The matrix-geometric stationary distribution is a standard property of QBDs (see [14]).

Corollary 13. *The generating function of the number of high priority jobs $Y_H(z) = \sum_{i=0}^{\infty} z^i \underline{y}_i^H \mathbb{1}$ and the factorial moments $E(Y_H^k)$ are given by*

$$Y_H(z) = y_0^H (\mathbf{I} - z\mathbf{R})^{-1}\mathbb{1}, \qquad E(Y_H^k) = k! y_0^H \mathbf{R}^k (\mathbf{I} - \mathbf{R})^{-k-1}\mathbb{1}. \tag{80}$$

5.5. Extensions of the model to arbitrary many job classes

The presented approach can be generalized to arbitrary many job classes as well. The details are given in Appendix B.

6. Numerical results

We implemented the presented analysis methods in MATLAB [2]. The implementation computes all performance measures considered in the paper by both preemptive resume and non-preemptive service, and for any number of job classes.

In our implementation the NARE problems are solved by the ADDA procedure [25]. We note that one of the two linear terms in the NAREs in this paper are block diagonal, which can be exploited by a novel technique to improve the computation speed further ([26]), but we did not use this

[2] Our implementation can be downloaded from `http://www.hit.bme.hu/~ghorvath/software`

possibility. The Sylvester equations are solved by the `lyap` function of MATLAB, which is based on the Hessenberg-Schur algorithm [27].

In this section we compare our procedure with three prior methods: the method of [10] (transformed to continuous time), its improved version published in [12], and the procedure of [11]. Note that the latter two procedures are far less general than [10] or the proposed one. They can handle only preemptive resume service, they do not analyze the sojourn time at all, and [11] is only able to provide the moments of the number of jobs.

Since all involved procedures are exact, only the scalability is investigated, that is, the analysis time as the function of the number of phases.

For this purpose let us define the MMAP matrices as

$$\mathbf{D_0}^{(K)} = \begin{bmatrix} \bullet & K\nu & & & \\ \gamma & \bullet & (K-1)\nu & & \\ & \ddots & & \ddots & \ddots \\ & & (K-1)\gamma & \bullet & \nu \\ & & & K\gamma & \bullet \end{bmatrix}, \quad \mathbf{D_L}^{(K)} = \begin{bmatrix} 0 & & & & \\ & r_L/K & & & \\ & & 2r_L/K & & \\ & & & \ddots & \\ & & & & r_L \end{bmatrix},$$

and matrix $\mathbf{D_H}^{(K)}$ is defined similarly. The diagonal entries denoted by $\bullet$ are determined uniquely such that the row sums of $\mathbf{D_0}^{(K)} + \mathbf{D_L}^{(K)} + \mathbf{D_H}^{(K)}$ are zeroes.

The service times are characterized by order-2 PH distributions with parameters

$$\begin{aligned} \sigma_H &= \begin{bmatrix} 0.16667 & 0.83333 \end{bmatrix}, & \sigma_L &= \begin{bmatrix} 0.58824 & 0.41176 \end{bmatrix}, \\ \mathbf{S_H} &= \begin{bmatrix} -0.66667 & 0.66667 \\ 0 & -4 \end{bmatrix}, & \mathbf{S_L} &= \begin{bmatrix} -3.2941 & 3.2941 \\ 0 & -5.6 \end{bmatrix}, \end{aligned}$$

having service rates $\mu_L = 2.8$ and $\mu_H = 2$. The utilization depends on K, it varies between 0.6 and 0.75.

Figure 6 depicts the analysis time required to obtain the first 10 moments of the number of low priority jobs in the system in the preemptive case as the function of K. (This is the only performance measure that is supported by all the procedures). It is clearly visible that the presented method is at least an order of magnitude faster than the prior ones, and is able to solve systems with a large number of phases. No numerical problems were encountered even with the largest model. Additionally, as opposed to [12] and [11], the presented procedure can provide sojourn time related performance measures, and is able to handle the case of non-preemptive service as well.

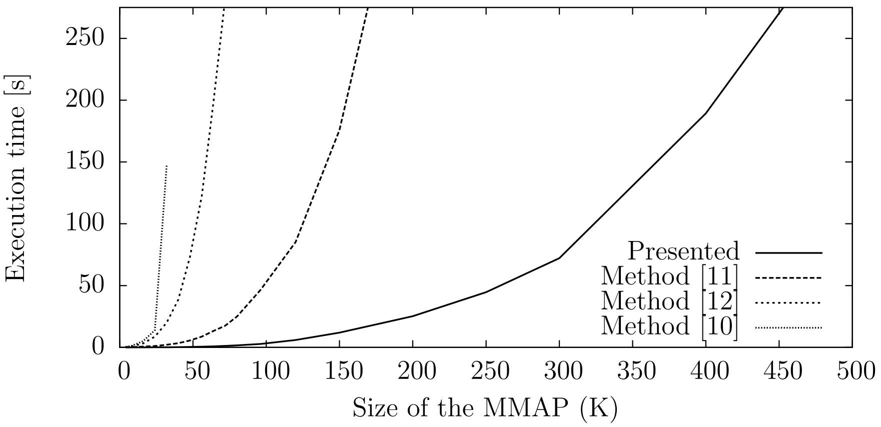

Figure 6: Comparison of the execution times of various procedures

An other simple numerical example can be found in the online supplementary material.

7. Conclusion

This paper presents a unique approach for the analysis of priority queues with MMAP input and PH distributed service times in the sense that the performance measures are derived from various properties of the busy period process of fluid queues. Several recent research results are utilized, including the workload-based queue analysis approach, the solution of fluid processes with jumps and the matrix-analytic methods for Markovian fluid models. The result is an easy to implement procedure, which, according to our numerical experiments is numerically reliable and at least an order of magnitude faster that past procedures.

References

[1] T. Demoor, J. Walraevens, D. Fiems, S. De Vuyst, H. Bruneel, Modelling Queue Sizes in an Expedited Forwarding DiffServ Router with Service Differentiation, in: Proceedings of the 4th International Conference on Queueing Theory and Network Applications, QTNA '09, 16:1–16:8, 2009.

[2] S. Shalev-Oren, A. Seidmann, P. Schweitzer, Analysis of flexible manufacturing systems with priority scheduling: PMVA, Annals of Operations Research 3 (3) (1985) 113–139.

[3] M. S. Hagen, J. K. Jopling, T. G. Buchman, E. K. Lee, Priority Queuing Models for Hospital Intensive Care Units and Impacts to Severe Case Patients, in: AMIA Annual Symposium Proceedings, vol. 2013, American Medical Informatics Association, 841–850, 2013.

[4] R. M. de Souza, R. Morabito, F. Y. Chiyoshi, A. P. Iannoni, Incorporating priorities for waiting customers in the hypercube queuing model with application to an emergency medical service system in Brazil, European Journal of Operational Research (2014) In press.

[5] R. G. Miller Jr, Priority queues, The Annals of Mathematical Statistics (1960) 86–103.

[6] T. Takine, The workload in the MAP/G/1 queue with state-dependent services: Its application to a queue with preemptive resume priority, Stochastic Models 10 (1) (1994) 183–204.

[7] T. Takine, A nonpreemptive priority MAP/G/1 queue with two classes of customers, Journal of Operations Research Society of Japan 39 (2) (1996) 266–290.

[8] T. Takine, The nonpreemptive priority MAP/G/1 queue, Operations Research 47 (6) (1999) 917–927.

[9] V. Subramanian, R. Srikant, Tail probabilities of low-priority waiting times and queue lengths in MAP/GI/1 queues, Queueing systems 34 (1-4) (2000) 215–236.

[10] A. S. Alfa, B. Liu, Q.-M. He, Discrete-time analysis of MAP/PH/1 multiclass general preemptive priority queue, Naval Research Logistics (NRL) 50 (6) (2003) 662–682.

[11] G. Horváth, Efficient analysis of the queue length moments of the MMAP/MAP/1 preemptive priority queue, Performance Evaluation 69 (12) (2012) 684–700.

[12] A. Horváth, G. Horváth, M. Telek, A traffic based decomposition of two-class queueing networks with priority service, Computer Networks 53 (8) (2009) 1235–1248.

[13] Q.-M. He, M. F. Neuts, Markov chains with marked transitions, Stochastic Processes and their Applications 74 (1) (1998) 37–52.

[14] G. Latouche, V. Ramaswami, Introduction to matrix analytic methods in stochastic modeling, 1999.

[15] V. Ramaswami, Matrix analytic methods for stochastic fluid flows, in: Teletraffic Engineering in a Competitive World - Proc. of the 16th International Teletraffic Congress (ITC 16), Elsevier Science B.V., 1019–1030, 1999.

[16] S. Soares, G. Latouche, Further results on the similarity between fluid queues and QBDs,, in: Proceedings of the 4th international conference on matrix-analytic methods, 89–106, 2002.

[17] S. Ahn, V. Ramaswami, Efficient algorithms for transient analysis of stochastic fluid flow models, Journal of Applied Probability (2005) 531–549.

[18] V. Ramaswami, D. G. Woolford, D. A. Stanford, The Erlangization method for Markovian fluid flows, Annals of Operations Research 160 (1) (2008) 215–225.

[19] T. Dzial, L. Breuer, A. da Silva Soares, G. Latouche, M.-A. Remiche, Fluid queues to solve jump processes, Performance Evaluation 62 (1) (2005) 132–146.

[20] B. Van Houdt, Analysis of the adaptive MMAP[K]/PH[K]/1 queue: a multi-type queue with adaptive arrivals and general impatience, European Journal of Operational Research 220 (3) (2012) 695–704.

[21] P. Buchholz, M. Telek, Rational processes related to communicating Markov processes, Journal of Applied Probability 49 (1) (2012) 40–59.

[22] Q.-M. He, Fundamentals of matrix-analytic methods, Springer, 2014.

[23] Q. He, Analysis of a continuous time SM[K]/PH[K]/1/FCFS queue: Age process, sojourn times, and queue lengths, Journal of Systems Science and Complexity 25 (1) (2012) 133–155.

[24] C. Van Loan, Computing integrals involving the matrix exponential, Automatic Control, IEEE Transactions on 23 (3) (1978) 395–404.

[25] W. G. Wang, W. C. Wang, R. C. Li, Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations, SIAM Journal on Matrix Analysis and Applications 33 (1) (2012) 170–194.

[26] B. Meini, On the numerical solution of a structured nonsymmetric algebraic Riccati equation, Performance Evaluation 70 (9) (2013) 682–690.

[27] G. Golub, S. Nash, C. Van Loan, A Hessenberg-Schur method for the problem AX+ XB= C, Automatic Control, IEEE Transactions on 24 (6) (1979) 909–913.

Appendix A. Preemptive resume priority queues with arbitrary many job classes

This section gives the outline of the analysis when the number of classes is greater than 2. Instead of H and L, the classes are denoted by integer numbers $1 \dots K$ such that a greater number corresponds to higher priority. The analysis is provided for class k, $1 \leq k < K$.

To characterize the amount of work in the system found by a class k job upon its arrival, classes $< k$ can be neglected. For classes $\geq k$ the workload process is similar to the one in Figure 1, but we have to distinguish various types of upward jumps corresponding to various job classes, thus, when a job arrives, it initiates the phase type distribution representing the service time of its class. Hence, the blocks of the generator of the fluid model representing the workload process are (see also (18))

$$\mathbf{Q}^{(k)}_{++} = \begin{bmatrix} \mathbf{I} \otimes \mathbf{S_k} & & \\ & \ddots & \\ & & \mathbf{I} \otimes \mathbf{S_K} \end{bmatrix}, \qquad \mathbf{Q}^{(k)}_{+-} = \begin{bmatrix} \mathbf{I} \otimes s_k \\ \vdots \\ \mathbf{I} \otimes s_K \end{bmatrix}, \tag{A.1}$$

$$\mathbf{Q}^{(k)}_{-+} = \begin{bmatrix} \mathbf{D_k} \otimes \sigma_k & \dots & \mathbf{D_K} \otimes \sigma_K \end{bmatrix}, \qquad \mathbf{Q}^{(k)}_{--} = \sum_{i=0}^{k-1} \mathbf{D_i}. \tag{A.2}$$

From the matrix-exponentially distributed stationary solution of the fluid model $\pi^{(k)}(x)$, the density of the workload at class k arrivals, is expressed by

$$\hat{\pi}^{(k)}(x) = \frac{1}{\hat{c}^{(k)}}\pi^{(k)}(x)\begin{bmatrix}\mathbf{I}\otimes s_k\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix} = \underbrace{\frac{1}{\hat{c}^{(k)}}p_-^{(k)}\mathbf{Q}_{-+}^{(k)}}_{\hat{\beta}^{(k)}} e^{\mathbf{K}^{(k)}x}\underbrace{\begin{bmatrix}\mathbf{I}\otimes s_k\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}}_{\hat{\mathbf{B}}^{(k)}}, \tag{A.3}$$

where the normalization constant is $\hat{c}^{(k)} = p_-^{(k)}\mathbf{Q}_{-+}^{(k)}(-\mathbf{K}^{(k)})^{-1}\hat{\mathbf{B}}^{(\mathbf{k})}\mathbb{1}$.

The $(\hat{\beta}^{(k)}, \mathbf{K}^{(k)}, \hat{\mathbf{B}}^{(k)})$ representation is transformed to $(\hat{\beta'}^{(k)}, \mathbf{K'}^{(k)}, \hat{\mathbf{B}'}^{(k)})$ to ensure the proper row-sums by using Theorem 5.

The blocks of the generator of the fluid model corresponding to the remaining sojourn time process are similar to the ones of the two-class case (see (25)), the difference is that now there are more than one classes that have priority over class k. Hence we get

$$\begin{aligned}\tilde{\mathbf{Q}}_{++}^{(k)} &= \begin{bmatrix}\mathbf{K'}^{(k)} & & & \\ & \mathbf{I}\otimes\mathbf{S_{k+1}} & & \\ & & \ddots & \\ & & & \mathbf{I}\otimes\mathbf{S_K}\end{bmatrix}, \qquad \tilde{\mathbf{Q}}_{+-}^{(k)} = \begin{bmatrix}\hat{\mathbf{B}'}^{(k)}\\ \mathbf{I}\otimes s_{k+1}\\ \vdots\\ \mathbf{I}\otimes s_K\end{bmatrix}, \\ \tilde{\mathbf{Q}}_{-+}^{(k)} &= \begin{bmatrix}\mathbf{0} & \mathbf{D_{k+1}}\otimes\sigma_{k+1} & \dots & \mathbf{D_K}\otimes\sigma_K\end{bmatrix}, \qquad \tilde{\mathbf{Q}}_{--}^{(k)} = \sum_{i=0}^{k}\mathbf{D_i},\end{aligned} \tag{A.4}$$

and the initial phase distribution is

$$\tilde{\kappa}^{(k)} = \begin{bmatrix}\hat{\beta'}^{(k)} & 0 & \dots & 0\end{bmatrix}. \tag{A.5}$$

The performance measures related to the sojourn time can be derived from this fluid queue exactly as in the two-class case.

For the number of jobs the results of Section 4.3 can be applied with slight modifications. Instead of $\mathbf{D_L}$ we have to use $\mathbf{D_k}$. Furthermore, in Theorem 6 and Corollaries 3, and 4 matrix $\mathbf{D_0}$ needs to be replaced by $\sum_{i=0}^{k-1}\mathbf{D_i}$, while in Theorem 7 and Corollary 5 matrix $\mathbf{D_0}+\mathbf{D_H}$ needs to be replaced by $\sum_{i=0,i\neq k}^{K}\mathbf{D_i}$.

Appendix B. Non-preemptive priority queues with many job classes

This section provides the analysis for class k, $1 < k < K$.

To characterize the amount of work in the system when a class k job arrives, we need to analyze the workload process restricted to the time periods when class $\geq k$ jobs are present in the system. This workload process is similar to the one discussed in Section 5.4, and depicted in Figure 5.

The blocks of the generator of the fluid model representing the workload process are as follows.

$$\mathbf{Q}_{++}^{(k)} = \left[\begin{array}{ccc|ccc} \mathbf{I}\otimes\mathbf{I}_1\otimes\mathbf{S_k} & & & & & \\ & \ddots & & & & \\ & & \mathbf{I}\otimes\mathbf{I}_{k-1}\otimes\mathbf{S_K} & & & \\ \hline & & & \mathbf{I}\otimes\mathbf{S_k} & & \\ & & & & \ddots & \\ & & & & & \mathbf{I}\otimes\mathbf{S_K} \end{array}\right],$$

$$\mathbf{Q}_{+-}^{(k)} = \left[\begin{array}{c} \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \hline \mathbf{I}\otimes s_k \\ \vdots \\ \mathbf{I}\otimes s_K \end{array}\right], \qquad \mathbf{Q}_{+0}^{(k)} = \left[\begin{array}{c} \mathbf{I}\otimes\mathbf{I}_1\otimes s_k \\ \vdots \\ \mathbf{I}\otimes\mathbf{I}_{k-1}\otimes s_K \\ \hline \mathbf{0} \\ \vdots \\ \mathbf{0} \end{array}\right],$$

$$\mathbf{Q}_{0+}^{(k)} = \left[\begin{array}{cccccc|c} \mathbf{D_k}\otimes\mathbf{I}_1\otimes\sigma_k & & & \mathbf{D_K}\otimes\mathbf{I}_1\otimes\sigma_K & & & \mathbf{0} \\ & \ddots & & & \ddots & & \mathbf{0} \\ & & \mathbf{D_k}\otimes\mathbf{I}_{k-1}\otimes\sigma_k & & & \mathbf{D_K}\otimes\mathbf{I}_{k-1}\otimes\sigma_K & \mathbf{0} \end{array}\right],$$

$$\mathbf{Q}_{00}^{(k)} = \left[\begin{array}{ccc} \sum_{i=0}^{k-1}\mathbf{D_i}\oplus\mathbf{S_1} & & \\ & \ddots & \\ & & \sum_{i=0}^{k-1}\mathbf{D_i}\oplus\mathbf{S_{k-1}} \end{array}\right], \quad \mathbf{Q}_{0-}^{(k)} = \left[\begin{array}{c} \mathbf{I}\otimes s_1 \\ \vdots \\ \mathbf{I}\otimes s_{k-1} \end{array}\right],$$

$$\mathbf{Q}^{(k)}_{-+} = \left[\ \mathbf{0}\ \middle|\ \mathbf{D_k}\otimes\sigma_k \quad \dots \quad \mathbf{D_K}\otimes\sigma_K\ \right], \qquad \mathbf{Q}^{(k)}_{--} = \sum_{i=0}^{k-1}\mathbf{D_i},$$

where matrix $\mathbf{I}_j$ is an identity matrix of the same size as $\mathbf{S_j}$.

The states of the background process can be grouped to four state groups. In the zero states $\mathcal{S}_0$ there are class $\geq k$ jobs in the system, but the server is working on a class $< k$ one. The class of the current job in the server and its phase are encoded into $\mathcal{S}_0$. When a class $\geq k$ job arrives, the background process moves to $\mathcal{S}_+$, where the workload is increased by the service time of the job. Two state groups can be distinguished in $\mathcal{S}_+$. In the first state group there is a low priority job in the server, thus the background process needs to keep track of 1) the phase of the arrival process, 2) the class of the job in the server, 3) the phase of the service time of the job in the server, 4) the class of the job that arrived, and 5) the phase of the service time of the job that arrived. While the workload is increasing, the arrival process and the service of the job in the server are frozen. The second state group of $\mathcal{S}_+$ takes care of the increase of the workload when there is no lower priority job in the server. Finally, $\mathcal{S}_-$ represents the periods when class $\geq k$ jobs are served and the workload decreases by a slope of one accordingly.

The fundamental matrices corresponding to this fluid model are denoted by $\mathbf{\Psi}^{(k)}$ and $\mathbf{K}^{(k)}$.

To characterize the steady state behavior of the workload process at arrivals, it remains to derive the initial phase of the background process when a class $\geq k$ job arrives when only class $< k$ jobs are present in the system. To derive the initial phase probability vector we investigate the system at those departure instants where the departing job leaves no class $\geq k$ job in the system. Similar to vectors ϕ and ϕ_0 (defined in Section 5.4), we introduce probability vectors $\phi_i^{(k)}$. Entry j of $\phi_i^{(k)}$ is the probability that there are no jobs with priority higher than i in the system at these instants and the phase of the MMAP is j. Thus,

$$\begin{aligned}\{\phi_i^{(k)}\}_j = P(&\text{phase at departure is } j \text{ and only} \leq i \text{ classes are present} \\ &\mid\ \text{no class} > k \text{ jobs in the system at departure})\end{aligned} \tag{B.1}$$

Vector $\phi_k^{(k)}$ can be obtained as a solution of a set of linear equations.

Based on Theorem 3 of [8] and (69) we have linear equations

$$
\begin{aligned}
\phi_k^{(k)} = {} & \sum_{i=1}^{k} \left(\phi_i^{(k)} - \phi_{i-1}^{(k)} + \phi_0^{(k)}(-\mathbf{D_0})^{-1}\mathbf{D_i}\right)(\mathbf{I} \otimes \sigma_i) \left(-\sum_{j=0}^{k} \mathbf{D_j} \oplus \mathbf{S_i}\right)^{-1} \\
& \cdot \left((\mathbf{I} \otimes s_i) + \sum_{j=k+1}^{K} \left[\mathbf{0} \dots \mathbf{D_j} \otimes \overset{jth}{\mathbf{I}} \otimes \sigma_j \dots \mathbf{0} \;\; | \;\; \mathbf{0} \;\; \right] \cdot \mathbf{\Psi}^{(k+1)} \right) \\
& + \phi_0^{(k)}(-\mathbf{D_0})^{-1} \sum_{j=k+1}^{K} \left[\;\; \mathbf{0} \;\; | \;\; \mathbf{0} \dots \mathbf{D_j} \overset{jth}{\otimes} \sigma_j \dots \mathbf{0} \right] \cdot \mathbf{\Psi}^{(k+1)},
\end{aligned}
\tag{B.2}
$$

where term i in the sum corresponds to the case when the class of the next job to serve is i ($i \leq k$). The service of this job is started. The next departure in the embedded process can occur when either the service of this job ends, or when a higher priority job arrives. In the latter case matrix $\mathbf{\Psi}^{(k+1)}$ determines the phase transitions between the beginning and the end of the busy period of priority $> k$ jobs. The last term is related to the case when the system is empty at departure, and the next arriving job is a high priority ($> k$) one.

To get a fully determined system of linear equations for vectors $\phi_k^{(k)}$ the following relations are also needed (see Lemma 2 and eq. (17) of [8]):

$$
\begin{aligned}
p_k &= P(\text{no class} > k \text{ jobs in the system at departure}) \\
&= \frac{1}{\lambda} \sum_{j=1}^{k} \lambda_j + \frac{1}{\lambda} \kappa \sum_{j=k+1}^{K} \mathbf{D_j} \mathbb{1}, \\
\phi_i^{(k)} \cdot p_k &= \phi_i^{(i)} \cdot p_i, \quad k, i = 1, \dots, K-1,
\end{aligned}
$$

where κ is the initial vector of the matrix-exponentially distributed workload process of the whole system (including all priority classes).

Similar to the two-class case in (72), we introduce vectors $q_i^{(k)}$, $i = 1, \dots, k-1$ representing the stationary distribution of the phase when there are no class $\geq k$ jobs in the system and the server is working on a class i job, and vector $q_0^{(k)}$ for the case when the system is idle. We have

$$q_i^{(k)} = \big(\phi_i^{(k-1)} - \phi_{i-1}^{(k-1)} + \phi_0^{(k-1)}(-\mathbf{D_0})^{-1}\mathbf{D_i}\big)(\mathbf{I} \otimes \sigma_i)\left(-\sum_{j=0}^{k-1}\mathbf{D_j} \oplus \mathbf{S_i}\right)^{-1},$$

$$q_0^{(k)} = \phi_0^{(k-1)}(-\mathbf{D_0})^{-1}.$$

Concatenating vectors $q_i^{(k)}$ to $q_L^{(k)} = \{q_i^{(k)}, i = 1, \ldots, k-1\}$ the initial vector of the workload process for class $\geq k$ jobs denoted by $\kappa^{(k)}$ is given by (see also (73) for two priorities)

$$\kappa^{(k)} = q_L^{(k)} \cdot \mathbf{Q}_{0+}^{(k)} + q_0^{(k)} \cdot \mathbf{Q}_{-+}^{(k)}. \tag{B.3}$$

Now we express the stationary density of the initial workload of class $\geq k$ jobs, the amount of workload that a class k arrival finds in the system. Four cases are distinguished:

- At the class k arrival there are no class $\geq k$ jobs in the system, but the server is working on a class $< k$ job. The initial workload for the arrival equals the remaining service time of the class $< k$ job.
- At the class k arrival there are class $\geq k$ jobs in the system, but the server is working on a class $< k$ job. The initial workload equals to the workload of class $\geq k$ jobs residing in the queue plus the remaining service time of the class $< k$ job.
- At the class k arrival there are class $\geq k$ jobs in the system, and the server is working on a class $\geq k$ job. The initial workload equals to the workload of the class $\geq k$ jobs residing in the queue.
- At the class k arrival there are no jobs in the system at all. The initial workload is zero.

Thus, the initial workload equals to either the workload of the class $\geq k$ jobs, or to the remaining service time, or to the sum of both. Similar to (75) in the two-class case, it is matrix-exponentially distributed in the general case as well,

$$\check{\pi}^{(k)}(x) = \zeta^{(k)} e^{\mathbf{Z}^{(k)} t} \mathbf{V}^{(k)}, \tag{B.4}$$

with parameters

$$\zeta^{(k)} = \left[\begin{array}{c|c} \kappa^{(k)} & q_L^{(k)} \cdot \begin{bmatrix} \mathbf{D_k} \otimes \mathbf{I}_1 & & \\ & \ddots & \\ & & \mathbf{D_k} \otimes \mathbf{I}_{k-1} \end{bmatrix} \end{array} \right] / \check{c}^{(k)}, \tag{B.5}$$

$$\mathbf{Z}^{(k)} = \left[\begin{array}{c|c} \mathbf{K}^{(k)} & (\mathbf{Q}_{+0}^{(k)} + \mathbf{\Psi}^{(k)} \mathbf{Q}_{-0}^{(k)}) \cdot \begin{bmatrix} \mathbf{D_k} \otimes \mathbf{I}_1 & & \\ & \ddots & \\ & & \mathbf{D_k} \otimes \mathbf{I}_{k-1} \end{bmatrix} \\ \hline \mathbf{0} & \begin{bmatrix} \mathbf{I} \otimes \mathbf{S_1} & & \\ & \ddots & \\ & & \mathbf{I} \otimes \mathbf{S_{k-1}} \end{bmatrix} \end{array} \right], \tag{B.6}$$

$$\mathbf{V}^{(k)} = \left[\begin{array}{c} \mathbf{\Psi}^{(k)} \mathbf{D_k} \\ \hline \mathbb{1} \otimes s_1 \\ \vdots \\ \mathbb{1} \otimes s_{k-1} \end{array} \right], \tag{B.7}$$

and the probability that the initial workload at arrival is zero is given by

$$\check{p}^{(k)} = \frac{1}{\check{c}^{(k)}} q_0^{(k)} \mathbf{D_k}. \tag{B.8}$$

The normalization constant is obtained as $\check{c}^{(k)} = 1/(\check{p}^{(k)}\mathbb{1} + \zeta^{(k)}(-\mathbf{Z}^{(k)})^{-1}\mathbf{V}^{(k)}\mathbb{1})$. The first state group of this matrix-exponential distribution generates the workload of class $\geq k$ jobs residing in the queue at the arrival, while the second state group represents the remaining service time of the low priority job residing in the server.

The fluid model representing the remaining waiting time of class k jobs can be characterized by the same matrices as in the preemptive resume case given by (A.4). The only difference is that instead of $\mathbf{K'}^{(k)}$, $\hat{\mathbf{B}}'^{(k)}$ and $\beta'^{(k)}$, the parameters of the initial workload are $\mathbf{Z}^{(k)}$, $\mathbf{V}^{(k)}$ and $\zeta^{(k)}$ in the non-preemptive case. Given the fluid model for the remaining waiting time, the performance measures are derived just like in the two-class case in Section 5.2 and Section 5.3.