Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

On a generic class of two-node queueing systems

I.J.B.F. Adan, M.R.H. Mandjes, W.R.W. Scheinhardt,
E. Tzenova

REPORT PNA-R0802 FEBRUARY 2008

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

On a generic class of two-node queueing systems

ABSTRACT
This paper analyzes a generic class of two-node queueing systems. A first queue is fed by an on-off Markov fluid source; the input of a second queue is a function of the state of the Markov fluid source as well, but now also of the first queue being empty or not. This model covers the classical two-node tandem queue and the two-class priority queue as special cases. Relying predominantly on probabilistic argumentation, the steady-state buffer content of both queues is determined (in terms of its Laplace transform). Interpreting the buffer content of the second queue in terms of busy periods of the first queue, the (exact) tail asymptotics of the distribution of the second queue are found. Two regimes can be distinguished: a first in which the state of the first queue (that is, being empty or not) hardly plays a role, and a second in which it explicitly does. This dichotomy can be understood by using large-deviations heuristics.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: queueing, fluid models, networks, asymptotics
Note: This research has been funded by the Dutch Bsik/BRICKS (Basic Research in Informatics for Creating the Knowledge Society) project.

On a generic class of two-node queueing systems*

Ivo Adan† Michel Mandjes‡ Werner Scheinhardt§ Elena Tzenova¶

February 18, 2008

Abstract

This paper analyzes a generic class of two-node queueing systems. A first queue is fed by an on-off Markov fluid source; the input of a second queue is a function of the state of the Markov fluid source as well, but now also of the first queue being empty or not. This model covers the classical two-node tandem queue and the two-class priority queue as special cases. Relying predominantly on probabilistic argumentation, the steady-state buffer content of both queues is determined (in terms of its Laplace transform). Interpreting the buffer content of the second queue in terms of busy periods of the first queue, the (exact) tail asymptotics of the distribution of the second queue are found. Two regimes can be distinguished: a first in which the state of the first queue (that is, being empty or not) hardly plays a role, and a second in which it explicitly does. This dichotomy can be understood by using large-deviations heuristics.

*This work has been carried out partly in the Dutch BSIK/BRICKS project.

†IA (email: i.j.b.f.adan@tue.nl is with Eindhoven University of Technology, Department of Mathematics and Computing Science, P.O. Box 513, 5600 MB Eindhoven, the Netherlands. He is also affiliated to EURANDOM, Eindhoven, the Netherlands.

‡MM (email: michel@cwi.nl) is with Korteweg-de Vries Institute, University of Amsterdam, Plantage Muidergracht 24, 1018 TV Amsterdam, the Netherlands. He is also affiliated to CWI, Amsterdam, the Netherlands, and EURANDOM, Eindhoven, the Netherlands.

§WS (email: W.R.W.Scheinhardt@utwente.nl) is with Faculty of Electrical Engineering, Mathematics, and Computer Science, P.O. Box 217, 7500 AE Enschede, the Netherlands. He is also affiliated to CWI, Amsterdam, the Netherlands.

¶Written while ET was affiliated to EURANDOM, P.O. Box 513, 5600 MB Eindhoven, the Netherlands.

1 Introduction

In a variety of operational applications, one needs to analyze the performance experienced by traffic streams flowing through a network — one could think of production systems, logistic systems, communication networks, etc. Queueing theory offers a natural framework for this. More specifically, in queueing theory, the network nodes are modeled as queues at which traffic arrives, these queues are served according to some discipline, and after being served the output of one node can serve as input for a next node or leave the system. Also, nodes could operate under scheduling disciplines that are more sophisticated than simply first-in-first-out; one could for instance prioritize certain traffic streams.
Queueing theory aims at analyzing the performance (in terms of loss, delay, throughput, etc.) of these nodes. However, most studies address performance issues just for single nodes, and do not consider end-to-end metrics. In some cases, it is well understood how the probabilistic properties of the traffic stream are affected by traversing a node (for instance in M/M/1-type of networks where the output streams have the same statistical properties as the input stream), but in many situations just partial results are available. The same applies to queues operating under non-standard scheduling disciplines.

In the present paper we consider a network of two queues, that, interestingly, covers the two-node tandem queue and the priority queue as special cases (and, in fact, a variety of combinations of these two). The first queue is fed by an on-off Markovian fluid source, and can be analyzed by standard techniques. The input of the second queue, however, is strongly affected by the buffer content of the first queue: it is again a function of the state of the Markov fluid source, but now also of the first queue being empty or not. The fact that the second queue cannot be solved in isolation from the first queue, makes this queue considerably harder to analyze.
The main contribution of our work is that we explicitly characterize the distribution of the buffer content of this second queue (in terms of its Laplace transform). We do so exclusively relying on elementary probabilistic techniques; for instance no martingale methods are needed. Remarkably, we can express the buffer content of the second queue in terms of the busy period of the first queue, which yields appealing probabilistic interpretations. As a second contribution we also derive the tail asymptotics of the second queue, and this we do without resorting to techniques from complex function analysis. In addition, we provide the intuition behind these asymptotic results; a number of regimes can be distinguished, and large-deviations argumentation can be used to develop understanding for these.

Our results touch on those derived in several other papers. Rough (that is, logarithmic) asymptotics for tandem networks (but the results partially generalize to the framework of the present paper) were derived in Chang *et al.* [3] — albeit in a discrete-time setting — and Mandjes [14]. With Q_2 being the steady-state buffer content of the downstream queue, they identify the limit $-\theta$ of $x^{-1}\log\mathbb{P}(Q_2 > x)$, implying that $\mathbb{P}(Q_2 > x) = f(x)\exp(-\theta x)$ for some (unknown) subexponential function $f(\cdot)$ (i.e., $\log f(x)/x \to 0$ as $x \to \infty$). A main conclusion from these papers is that essentially two regimes exist: one in which the first queue is 'transparent', in that its behavior hardly affects the overflow asymptotics of the second queue, and one in which the impact of the buffer content of the first queue is more explicitly visible.
Abate and Whitt [1] consider asymptotics, for compound Poisson input, of a priority system, and they

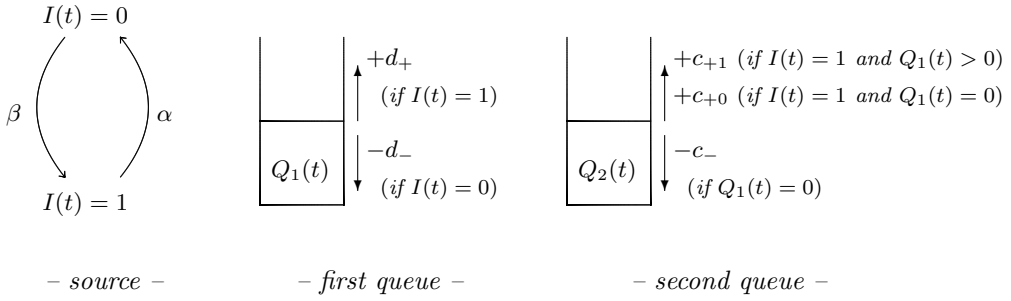

Figure 1.1: The fluid model.

also identify the two regimes. Importantly, the asymptotics in [1] are 'exact', in that an (explicitly given) function $g(\cdot)$ is found such that $\mathbb{P}(Q_2 > x)/g(x) \to 1$ as $x \to \infty$, with Q_2 being the steady-state buffer content of the low-priority queue. More precisely, in the transparent regime mentioned above, the exact asymptotics are of the type $\alpha \exp(-sx)$ for positive constants α, s, whereas in the other regime they look like $\alpha'/(x\sqrt{x})\exp(-s'x)$ for positive constants α', s'. Our results indicate that this dichotomy carries over to the more general two-node network that we briefly introduced above.
Exact analyses of the buffer content distribution of the second queue, in a tandem setting, are given by Scheinhardt and Zwart [19] and Kella [11], predominantly relying on martingale techniques; see also [13] and [18] for related results. Dieker and Mandjes [8] consider networks in which the input is a Markov additive process (that is, a Markov-modulated Lévy process), and in this sense more general than just an on-off Markov fluid source; their results are, however, considerably less explicit, and they do not consider tail asymptotics either.

The paper is organized as follows. Section 2 introduces our model. It also shows that a number of important queueing systems are covered as special cases. In Section 3 we concentrate on the Laplace transform of the buffer content of the second queue, and we probabilistically interpret the result. The remainder of the paper is devoted to the analysis of the tail asymptotics of the buffer content of the second queue. First we present (Section 4) heuristics for the logarithmic asymptotics: relying on a large-deviations motivation, we show why one would expect two regimes to appear. These regimes are indeed identified in Section 5: using the probabilistic interpretation mentioned above, we characterize the exact tail asymptotics of the buffer content of the second queue. Section 6 concludes.

2 Model and preliminaries

In this section we will first introduce the model and some interesting special cases. Then we present preliminary results concerning stability of the system and the distribution of the first queue.

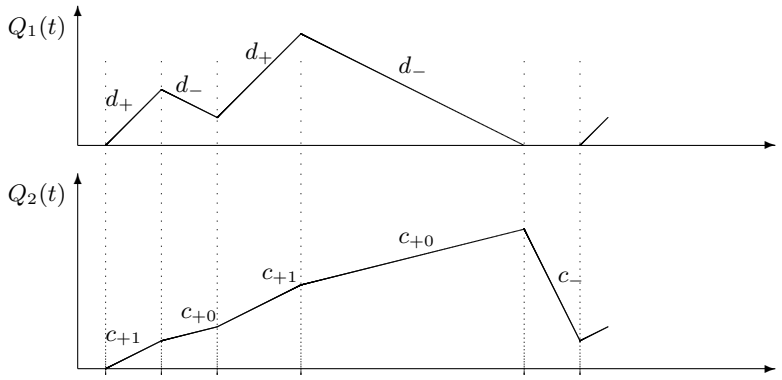

Figure 2.2: Sample-path of the two-node queueing system.

Thus, consider a stochastic fluid model with two infinite capacity buffers, which have at time t respective contents $Q_i(t),\ i = 1, 2$, see Fig. 1.1. The first buffer is fed by a Markovian on-off source $\{I(t) \in \{0, 1\},\ t \geq 0\}$ with mean off-time β^{-1} and mean on-time α^{-1}. Writing the steady state of this process as I without time index t, (as we will henceforth do for all stochastic processes), we clearly have $\mathbb{P}(I = 0) = 1 - \mathbb{P}(I = 1) = \alpha/(\alpha + \beta)$.
When $I(t) = 1$ the first buffer increases at rate d_+; otherwise it decreases at rate d_-, as long as it is not empty. The second buffer is driven by the first one and the input source in the following way: its content increases at rate c_{+1} (c_{+0}, respectively) when the first buffer is not empty and $I(t) = 1$ ($I(t) = 0$); otherwise it decreases at rate c_- (of course, provided that it is not empty).

Some special cases. We now show that a number of standard models are specific cases of our generic model.

- *Model 1: Priority system.* While $I(t) = 0$ there is no input to the first buffer and the input rate to the second buffer is $p_{20} > 0$. While $I(t) = 1$ the input rates to the first and the second buffers are $p_1 > 0$ and $p_{21} > 0$, respectively. The first buffer receives strict priority and is served at rate $c > 0$. The second buffer is served at rate $c > 0$ only when the first one is empty. To avoid trivialities we assume that $p_1 > c$, so that the first buffer is not always empty.

- *Model 2: Priority/tandem system.* This is a modification of the first model where type one fluid is served at rate $c_1 > 0$ and the output of the first buffer is input to the second. The second buffer is again served at rate $c_2 > 0$ only when the first one is empty. Again we assume that $p_1 > c_1$.

- *Model 3: Tandem/priority system.* This is a tandem fluid model with priorities. The two fluid buffers, with constant output rates $c_1 > 0$ and $c_2 > 0$, are placed in series. The first one is fed by the on-off source: while $I(t) = 1$ the input rate is $p_1 > c_1$. The output of the first buffer is the (only) input to the second. The second buffer is served only when the first is empty.

- *Model 4: Tandem system.* This is a classical tandem fluid model, as was also studied in [13, 18]. It is the same as model 3 with the modification that the second buffer is always served at rate $c_2 > 0$, provided that it is not empty. Here we assume that $p_1 > c_1 > c_2 > 0$.

The correspondence between these four models and the general model can be summarized as follows:

	Model 1	Model 2	Model 3	Model 4
d_+	$p_1 - c$	$p_1 - c_1$	$p_1 - c_1$	$p_1 - c_1$
d_-	c	c_1	c_1	c_1
c_{+1}	p_{21}	$c_1 + p_{21}$	c_1	$c_1 - c_2$
c_{+0}	p_{20}	$c_1 + p_{20}$	c_1	$c_1 - c_2$
c_-	$c - p_{20}$	$c_2 - p_{20}$	c_2	c_2

Stability conditions. The stability condition of the first queue is $d_+\mathbb{P}(I = 1) < d_-\mathbb{P}(I = 0)$, which is equivalent to

$$\alpha d_- - \beta d_+ > 0. \tag{2.1}$$

Under (2.1) the stationary distribution of $(I(t), Q_1(t))$ is known to exist and is given by (see e.g. [18])

$$\begin{aligned}
\mathbb{P}(I = 0, Q_1 \leq x) &= \frac{\alpha}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}\frac{d_+}{d_-}e^{-(\alpha/d_+ - \beta/d_-)x}, \\
\mathbb{P}(I = 1, Q_1 \leq x) &= \frac{\beta}{\alpha+\beta} - \frac{\beta}{\alpha+\beta}e^{-(\alpha/d_+ - \beta/d_-)x},
\end{aligned}$$

where $\alpha/d_+ - \beta/d_-$ is positive due to (2.1). The utilization of the first buffer is defined as $\rho_1 := \mathbb{P}(Q_1 > 0)$ and is given by

$$\rho_1 = \frac{\beta}{\alpha+\beta}\frac{d_- + d_+}{d_-}. \tag{2.2}$$

Similarly, stability of the second queue is ensured if and only if the input rate is smaller than the output rate; this means that the condition $c_{+1}\mathbb{P}(I = 1) + c_{+0}\mathbb{P}(I = 0, Q_1 > 0) < c_-\mathbb{P}(Q_1 = 0)$ should be satisfied, or equivalently

$$c_{+1}\frac{\beta}{\alpha+\beta} + c_{+0}\frac{\beta}{\alpha+\beta}\frac{d_+}{d_-} < c_-\left(1 - \frac{\beta}{\alpha+\beta}\frac{d_+ + d_-}{d_-}\right),$$

which can also be written as

$$\frac{\alpha d_-}{c_{+1}d_- + c_{+0}d_+ + c_-d_+} - \frac{\beta}{c_-} > 0. \tag{2.3}$$

Notice that (2.1) is implied by (2.3), as can be seen by multiplying the latter with c_-d_+. Hence, under (2.3), the stationary distribution of $(I(t), Q_1(t), Q_2(t))$ exists. The distribution of Q_1 being known, this paper focuses on the distribution of Q_2 and its tail asymptotics.

3 Distribution of queue 2

In this section we express the distribution of Q_2 in terms of other, known distributions. In particular, we present an explicit expression for the Laplace Transform (LT) of Q_2 in Theorem 3.6. The approach is based on Kella and Whitt [10], where we condition on the state of the first buffer.

3.1 Distribution of queue 2 when queue 1 is idle

We consider the buffer content process $Q_2(t)$ and delete the busy periods of queue 1 from the time axis. The resulting process, which has positive jumps at the beginning of each idle period of queue 1, will be called $W(t)$. In fact it is identical to the workload process in an M/G/1 queue, drained at rate c_-, with arrival rate β, in which the service times are distributed as the typical increase of the second buffer content during a busy period of the first buffer.

To analyze this increase, we relate it to the length of a busy period of buffer 1, denoted by B (realize that these busy periods are independent and identically distributed random variables). Consider then a typical sample path during a busy period of buffer 1 with length B, and let N denote the number of times the source turns on during this busy period (including the one that initiates the busy period) and let $X_i, Y_i,\ i = 1, \ldots, N$, denote the lengths of the source's respective on-times and off-times during this busy period, see Fig. 2.2. (Notice that Y_N only includes the part of the off-time that overlaps with the busy period of buffer 1.) Then we have the following two equations:

$$d_+ \sum_{i=1}^{N} X_i = d_- \sum_{i=1}^{N} Y_i, \quad \text{and} \quad \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} Y_i = B.$$

We thus find that $\sum_{i=1}^{N} X_i = d_-/(d_- + d_+) \cdot B$ and $\sum_{i=1}^{N} Y_i = d_+/(d_- + d_+) \cdot B$, so that we have for the total increase during B,

$$c_{+1} \sum_{i=1}^{N} X_i + c_{+0} \sum_{i=1}^{N} Y_i \stackrel{\mathrm{d}}{=} \frac{c_{+1} d_- + c_{+0} d_+}{d_- + d_+} \cdot B, \tag{3.4}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. Notice that the factor in front of B may be viewed as the (weighted) average increase rate of the second buffer content during a busy period of queue 1. In special cases where $c_{+0} = c_{+1} = c_+$, as in models 3 and 4, it is immediately clear that the increase should indeed be $c_+ B$. In the remainder we shall also use the shorthand notation c_+ to denote the weighted average of c_{+0} and c_{+1} when they are not equal.

Turning back to the process $W(t)$, when scaling time to arrive at a standard M/G/1 queue drained at rate 1, we have the following result for the distribution of the steady state random variable $W \stackrel{\mathrm{d}}{=} (Q_2 \mid Q_1 = 0)$.

Lemma 3.1 *W is distributed as the steady-state workload of an* M/G/1 *queue drained at unit rate, with arrival rate β/c_- and service times distributed as $c_+ B$, where B is the typical busy period of the first buffer, and*

$$c_+ := \frac{c_{+1} d_- + c_{+0} d_+}{d_- + d_+}. \tag{3.5}$$

The LT of W is given by

$$\mathbb{E}e^{-sW} = \frac{\left(1 - \frac{\beta}{c_-} c_+ \mathbb{E}B\right) s}{\frac{\beta}{c_-} \mathbb{E}e^{-sc_+ B} - \frac{\beta}{c_-} + s}. \tag{3.6}$$

Proof: The form of $\mathbb{E}e^{-sW}$ is immediate from the Pollaczek-Khinchine formula. ♢

To obtain the distribution of B, we consider the buffer content process $Q_1(t)$ and delete the on-periods X_i from the time axis, in a similar way as how we constructed the process $W(t)$ from the process

$Q_2(t)$. The resulting process is now identical to the workload process in an M/M/1 queue drained at rate d_- with arrival rate β and mean service time d_+/α. In this case we prefer to scale the buffer space to arrive at a standard M/M/1 queue drained at rate 1; this queue then also has arrival rate β, but mean service time $d_+/(\alpha d_-)$. The total busy period of the first queue, including the on-times, is then $(d_+ + d_-)/d_+$ times the busy period of this M/M/1 queue, which we denote as P. This leads to the following.

Lemma 3.2 *The busy period B of queue 1 is distributed as m times P, the busy period of an* M/M/1 *queue with arrival rate β and service rate $\alpha d_-/d_+$, i.e., $B \stackrel{\mathrm{d}}{=} mP$, where*

$$m := \frac{d_+ + d_-}{d_+}. \tag{3.7}$$

The LT and mean of B are given by

$$\mathbb{E}e^{-sB} = \frac{\beta + \frac{\alpha d_-}{d_+} + ms - \sqrt{\left(\beta + \frac{\alpha d_-}{d_+} + ms\right)^2 - 4\beta\frac{\alpha d_-}{d_+}}}{2\beta}, \qquad \textit{and} \tag{3.8}$$

$$\mathbb{E}B = \frac{d_- + d_+}{\alpha d_- - \beta d_+}. \tag{3.9}$$

Proof: To show (3.8), note that the LT $\mathbb{E}e^{-sP}$ of the busy period of an M/M/1 queue with arrival rate λ and service rate μ is found by solving (under the condition that it should have value 1 for $s = 0$)

$$\lambda(\mathbb{E}e^{-sP})^2 - (\lambda + \mu + s)\mathbb{E}e^{-sP} + \mu = 0; \tag{3.10}$$

it is therefore given by

$$\mathbb{E}e^{-sP} = \frac{\lambda + \mu + s - \sqrt{(\lambda + \mu + s)^2 - 4\lambda\mu}}{2\lambda}. \tag{3.11}$$

It suffices to choose $\lambda = \beta$ and $\mu = d_-\alpha/d_+$ in this expression, and then evaluate it at ms to find $\mathbb{E}e^{-sB}$. Equation (3.9) follows from the fact that $\mathbb{E}P = 1/(\mu - \lambda)$. ♢

Corollary 3.3 *The LT of W can be rewritten as*

$$\mathbb{E}e^{-sW} = \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right)\left(1 + \frac{\beta/c_-}{s_\mathrm{p}}\frac{s_\mathrm{p}}{s + s_\mathrm{p}} - \frac{\beta/c_-}{(1 + mc_+/c_-)s_\mathrm{p}}\frac{s_\mathrm{p}}{s + s_\mathrm{p}}\mathbb{E}e^{-sc_+B}\right) \tag{3.12}$$

where

$$s_\mathrm{p} := \frac{\alpha d_-}{c_+(d_- + d_+) + c_- d_+} - \frac{\beta}{c_-}. \tag{3.13}$$

Proof: After substitution of (3.8) and (3.9) into (3.6), we find a square root in the denominator. By multiplying numerator and denominator with a factor

$$\frac{\beta}{c_-}\frac{\beta + \frac{\alpha d_-}{d_+} + mc_+s + \sqrt{\left(\beta + \frac{\alpha d_-}{d_+} + mc_+s\right)^2 - 4\beta\frac{\alpha d_-}{d_+}}}{2\beta} - \frac{\beta}{c_-} + s, \tag{3.14}$$

this square root vanishes, while the square root that arises in the numerator can be written in terms of $\mathbb{E}e^{-sB}$. After some rewriting the result follows. ♢

3.2 Distribution of queue 2 when queue 1 is busy

Our next concern is to find the distribution of Q_2 during busy periods of the first buffer. To do so, let us consider an arbitrary point in time t during a busy period of buffer 1 (i.e., buffer 1 is non-empty), and define A as the amount of fluid that flowed into buffer 2 since the start of the current busy period. Since the amount of fluid in buffer 2 at the beginning of a busy period of queue 1 is the waiting time in the corresponding M/G/1 queue, we have by PASTA that it has the same distribution as W. Hence we have $(Q_2 \mid Q_1 > 0) \stackrel{\mathrm{d}}{=} W + A$, with W and A independent, and all we need to do is find the distribution of A.

Before proceeding, recall that, if we consider the buffer content process $Q_1(\cdot)$ and delete the on-periods X_i from the time axis, then it is identical to the workload process in the M/M/1 queue drained at rate d_- with arrival rate β and mean service time d_+/α. This relation with the M/M/1 is crucial for what follows.

The fraction of time the source is off (on) during a busy period is equal to $d_+/(d_- + d_+)$ (resp. $d_-/(d_- + d_+)$) due to the discussion above (3.4). Hence, with probability $d_+/(d_- + d_+)$, the source is off at time t. In that case let Y_N denote the length of the (whole) off-period at time t, the random variable $N \geq 1$ being the number of on-periods before the current off-period. Then we have, sample-path-wise,

$$A = c_{+1} \sum_{i=1}^{N} X_i + c_{+0} \left(\sum_{i=1}^{N-1} Y_i + A_Y \right), \tag{3.15}$$

where A_Y denotes the age of Y_N at time t; an empty sum is interpreted as 0. Note that, with V denoting the content of buffer 1 at time t, then

$$d_+ \sum_{i=1}^{N} X_i = d_- \left(\sum_{i=1}^{N-1} Y_i + A_Y \right) + V,$$

since the left-hand side is the total increase in buffer 1 during on-times from the start of the busy period up to time t, and the right-hand side is the total decrease in buffer 1 during off-times up to time t plus what is left in the buffer at time t. Substitution into (3.15) and then using (3.7) and (3.5) yields

$$\begin{aligned} A &= c_{+1} \frac{V}{d_+} + \left(c_{+0} + c_{+1} \frac{d_-}{d_+} \right) \left(\sum_{i=1}^{N-1} Y_i + A_Y \right) \\ &= c_{+1} \frac{V}{d_+} + m c_+ \left(\sum_{i=1}^{N-1} Y_i + A_Y \right). \end{aligned}$$

Note that the random variables in the right-hand side are dependent, but as we will see below, *conditionally* they are independent.

We proceed by conditioning on the number of jobs in the corresponding M/M/1 queue at the arbitrary point in time t during the busy period. The probability p_n that there are n jobs in the system at time t is

$$p_n = (1 - \rho)\rho^{n-1}, \qquad n = 1, 2, \ldots,$$

where

$$\rho := \frac{\beta d_+}{\alpha d_-}. \tag{3.16}$$

Given that there are n jobs in the system at time t, it follows from the memoryless property that

$$V \stackrel{\mathrm{d}}{=} d_+ \sum_{i=1}^{n} X_i,$$

and V and $\sum_{i=1}^{N-1} Y_i + A_Y$ are (conditionally) independent. Further, the age of the busy period $\sum_{i=1}^{N-1} Y_i + A_Y$ is the same as the remaining busy period of the time-reversed M/M/1 queue, and since the M/M/1 is reversible, the remaining busy period is the sum of n busy periods of an M/M/1 (with the same parameters as the original M/M/1). Hence, given that there are n jobs in the system at time t,

$$\sum_{i=1}^{N-1} Y_i + A_Y \stackrel{\mathrm{d}}{=} \sum_{i=1}^{n} P_i,$$

where P_i is a busy period in an M/M/1 with arrival rate β and mean service time $d_+/(d_-\alpha)$. Putting all ingredients together, we find that with probability $p_n \cdot d_+/(d_- + d_+)$,

$$A \stackrel{\mathrm{d}}{=} c_{+1} \sum_{i=1}^{n} X_i + mc_+ \sum_{i=1}^{n} P_i.$$

Now assume that at the arbitrary point in time t the source is on. Then we have

$$A = c_{+1} \left(\sum_{i=1}^{N} X_i + A_X \right) + c_{+0} \sum_{i=1}^{N} Y_i, \tag{3.17}$$

where $N \geq 0$ is the number of on-periods *before* the current on-period (possibly taking the value 0) and A_X denotes the age of X_{N+1} at time t. Further,

$$d_+ \sum_{i=1}^{N} X_i = d_- \sum_{i=1}^{N} Y_i + V, \tag{3.18}$$

where V now denotes the amount of fluid in the first buffer at the beginning of X_{N+1}, or in the M/M/1 queue, it denotes the amount of work in the system just prior to the $(N+1)$-st arrival in the busy period. Substitution of (3.18) into (3.17) and again using (3.7) and (3.5) yields

$$A = c_{+1} \frac{V}{d_+} + c_{+1} A_X + mc_+ \sum_{i=1}^{N} Y_i.$$

Note that the $(N+1)$-st arrival in the busy period of the M/M/1 queue is (statistically) the same as an arbitrarily chosen one, so the probability that there are n jobs in the system just prior to the arrival is equal to

$$q_n = (1-\rho)\rho^n, \qquad n = 0, 1, \ldots.$$

Conditioning on the number of jobs in the M/M/1 system being n, we have

$$V \stackrel{\mathrm{d}}{=} d_+ \sum_{i=1}^{n} X_i,$$

while A_X has the same distribution as any of the X_i, and the random variables V, A_X and $\sum_{i=1}^{N} Y_i$ are (conditionally) independent. Also, we have again (by using time-reversibility)

$$\sum_{i=1}^{N} Y_i \stackrel{\mathrm{d}}{=} \sum_{i=1}^{n} P_i.$$

So, summarizing, with probability $q_n \cdot d_-/(d_- + d_+)$,

$$A \stackrel{\mathrm{d}}{=} c_{+1} \sum_{i=1}^{n+1} X_i + mc_+ \sum_{i=1}^{n} P_i.$$

Finally, putting the results during off and on periods together, and using the fact that $mP \stackrel{\mathrm{d}}{=} B$, we obtain the following result.

Lemma 3.4 *The distribution of A is given by*

$$A \stackrel{\mathrm{d}}{=} \begin{cases} c_{+1} \sum_{i=1}^{n+1} X_i + c_+ \sum_{i=1}^{n+1} B_i & w.p. \quad (1-\rho)\rho^n d_+/(d_- + d_+), \quad n = 0, 1, \ldots, \\ c_{+1} \sum_{i=1}^{n+1} X_i + c_+ \sum_{i=1}^{n} B_i & w.p. \quad (1-\rho)\rho^n d_-/(d_- + d_+), \quad n = 0, 1, \ldots, \end{cases}$$

where the X_i are distributed as the on-times and the B_i are distributed as the busy periods of queue 1. The X_i and B_i are independent.

The LT of A is found as

$$\mathbb{E}e^{-sA} = \frac{d_+ \mathbb{E}e^{-sc_+ B} + d_-}{d_- + d_+} \sum_{n=0}^{\infty} (1-\rho)\rho^n \left(\frac{\alpha}{\alpha + c_{+1}s}\right)^{n+1} \left(\mathbb{E}e^{-sc_+ B}\right)^n \tag{3.19}$$

$$= \frac{(1-\rho)\alpha m^{-1} \left(d_-/d_+ + \mathbb{E}e^{-sc_+ B}\right)}{\alpha + c_{+1}s - \rho\alpha \mathbb{E}e^{-sc_+ B}}. \tag{3.20}$$

A more insightful form is presented next.

Corollary 3.5 *The LT of A is given by*

$$\mathbb{E}(e^{-sA}) = R(sc_+) \frac{c_+}{c_{+1}} \left(1 + \frac{c_{+1} - c_{+0}}{mc_+} \left(\frac{\alpha/c_{+1}}{\gamma} \frac{\gamma}{\gamma + s} + \frac{\beta/c_{+0}}{\gamma} \frac{\gamma}{\gamma + s} \mathbb{E}e^{-sc_+ B}\right)\right), \tag{3.21}$$

where $\gamma := \alpha/c_{+1} + \beta/c_{+0}$ and

$$R(s) := \left(\frac{\alpha d_-}{d_+} - \beta\right) \frac{1 - \mathbb{E}e^{-sB}}{ms}.$$

Proof: We first replace s by s/mc_+ in (3.20) since we prefer to work with $\mathbb{E}e^{-sP} = \mathbb{E}e^{-(s/m)B}$. Multiplying numerator and denominator by $1 - \mathbb{E}e^{-sP}$, and using that $\mathbb{E}e^{-sP}$ satisfies (3.10) with $\lambda = \beta$ and $\mu = d_-\alpha/d_+$ (see the proof of Lemma 3.2), we can rewrite the above expression as

$$\mathbb{E} \exp\left(-\frac{s}{mc_+} A\right) = \hat{R}(s) \frac{c_+}{c_{+1}} \frac{1 + (d_+/d_-) \cdot \mathbb{E}e^{-sP}}{1 + (d_+/d_-)(c_{+0}/c_{+1}) \cdot \mathbb{E}e^{-sP}},$$

where $\hat{R}(s)$ is the LT of the residual busy period in an M/M/1 queue with arrival rate β and mean service time $d_+/(d_-\alpha)$, that is,

$$\hat{R}(s) = \left(\frac{\alpha d_-}{d_+} - \beta\right) \frac{1 - \mathbb{E}e^{-sP}}{s}.$$

The last term in the above expression can be rewritten as

$$\begin{aligned} &\frac{1 + (d_+/d_-) \cdot \mathbb{E}e^{-sP}}{1 + (d_+/d_-)(c_{+0}/c_{+1}) \cdot \mathbb{E}e^{-sP}} \\ &= 1 + \left(1 - \frac{c_{+0}}{c_{+1}}\right) \frac{d_+}{d_-} \frac{\mathbb{E}e^{-sP}}{1 + (d_+/d_-)(c_{+0}/c_{+1}) \cdot \mathbb{E}e^{-sP}} \\ &= 1 + \left(1 - \frac{c_{+0}}{c_{+1}}\right) \frac{d_+}{d_-} \left(\frac{\mu}{\hat{\gamma} + s} + \frac{d_-}{d_+} \frac{c_{+1}}{c_{+0}} \frac{\lambda}{\hat{\gamma} + s} \mathbb{E}e^{-sP}\right), \end{aligned}$$

where in the last step we removed the square root from the denominator by exploiting the explicit form for $\mathbb{E}e^{-sP}$, see (3.11). The constant $\hat{\gamma}$ is given by

$$\hat{\gamma} := \lambda\left(1 + \frac{d_-}{d_+}\frac{c_{+1}}{c_{+0}}\right) + \mu\left(1 + \frac{d_+}{d_-}\frac{c_{+0}}{c_{+1}}\right).$$

Summarizing, and substituting λ and μ, we find

$$\mathbb{E}\exp\left(-\frac{s}{mc_+}A\right) = \hat{R}(s)\frac{c_+}{c_{+1}}\left(1 + \left(1 - \frac{c_{+0}}{c_{+1}}\right)\left(\frac{\alpha}{\hat{\gamma}+s} + \frac{c_{+1}}{c_{+0}}\frac{\beta}{\hat{\gamma}+s}B(s)\right)\right).$$

Finally, replacing s by smc_+ and, in addition, letting $\gamma = \hat{\gamma}/(mc_+)$ and $R(s) = \hat{R}(sm)$ yields the desired result. ♢

Remark: From the proof and the fact that $B \stackrel{\mathrm{d}}{=} mP$, it can be understood that $R(s)$ is the LT of B^*, the *residual* busy period of buffer 1. In fact, when $c_{+0} = c_{+1} = c_+$, we find that $\mathbb{E}e^{-sA} = R(c_+s)$ and hence $A \stackrel{\mathrm{d}}{=} c_+B^*$, as should be the case.

3.3 Result

We are now ready to present the main result of this section.

Theorem 3.6 *The stationary content of buffer 2 can be decomposed as, with ρ_1 given through* (2.2)*,*

$$Q_2 \stackrel{\mathrm{d}}{=} \begin{cases} W & w.p. \quad 1-\rho_1, \\ W + A & w.p. \quad \rho_1. \end{cases}$$

Here, W is distributed as the workload of an M/G/1 *queue with arrival rate β/c_- and service times distributed as c_+B, and A is distributed as the geometric sum involving on-times and busy periods as in Lemma 3.4. Finally all random variables involved are independent.*
Hence, the LT of the stationary content of buffer 2 is given, with ρ_1 given through (2.2)*, by*

$$\mathbb{E}e^{-sQ_2} = \left(1 - \rho_1 + \rho_1\mathbb{E}e^{-sA}\right)\mathbb{E}e^{-sW}, \tag{3.22}$$

where $\mathbb{E}e^{-sW}$ and $\mathbb{E}e^{-sA}$ are given in Corollaries 3.3 and 3.5 respectively.

Proof: Immediate from the preceding. ♢

3.4 Properties of the Laplace transform of Q_2

As an introduction to the next sections, this subsection concentrates on the singularities of the LT of Q_2. We do so, as it is expected that the largest negative singularity (that is, closest to 0) is the exponential rate at which the probability $\mathbb{P}(Q_2 > x)$ decays as $x \to \infty$, i.e.,

$$\lim_{x\to\infty} \frac{1}{x}\log\mathbb{P}(Q_2 > x).$$

We do not give a formal proof of this at this stage, as it will follow from the exact asymptotics in Section 5.
There may be two types of singularities for the LT in (3.22), as presented in the following lemmas, viz. poles and branching points.

Lemma 3.7 *$\mathbb{E}e^{-sQ_2}$ has a branching point at $s = -s_{\rm b}$, where*

$$s_{\rm b} = \frac{(\sqrt{\beta d_+} - \sqrt{\alpha d_-})^2}{c_{+1}d_- + c_{+0}d_+} > 0, \tag{3.23}$$

for all parameter values that satisfy the stability condition (2.3).

Proof: The LT of the busy period of an M/M/1 has branching points at $s = -(\sqrt{\lambda} \pm \sqrt{\mu})^2$, see the proof of Lemma 3.2. Therefore the branching points of $\mathbb{E}e^{-sc_+B}$ (and by (3.22) also those of $\mathbb{E}e^{-sQ_2}$) are given by the solutions to

$$mc_+s = -\left(\sqrt{\beta} \pm \sqrt{\frac{\alpha d_-}{d_+}}\right)^2,$$

so that the largest of these is $-s_{\rm b}$ as given in (3.23). ♢

Lemma 3.8 *$\mathbb{E}e^{-sQ_2}$ has a pole at $s = -s_{\rm p}$, where*

$$s_{\rm p} = \frac{\alpha d_-}{c_{+1}d_- + c_{+0}d_+ + c_-d_+} - \frac{\beta}{c_-} > 0, \tag{3.24}$$

for all parameter values that satisfy (2.3) *and the following criterion,*

$$\alpha c_-^2 d_- d_+ \leq \beta(c_{+1}d_- + c_{+0}d_+ + c_-d_+)^2. \tag{3.25}$$

If (3.25) *is not fulfilled, $\mathbb{E}e^{-sQ_2}$ has no negative pole.*

Proof: Since $\mathbb{E}e^{-sc_+B}$ has no poles, $\mathbb{E}e^{-sQ_2}$ as given in (3.22) only may have poles at the value(s) of s for which either $\beta\mathbb{E}e^{-c_+sB} - \beta + c_-s = 0$ or $\alpha + c_{+1}s - \rho\alpha\mathbb{E}e^{-c_+sB} = 0$, see (3.6) and (3.20) respectively. The latter equation leads to

$$\sqrt{(\beta + \frac{\alpha d_-}{d_+} + mc_+s)^2 - 4\beta\frac{\alpha d_-}{d_+}} = \beta - \frac{\alpha d_-}{d_+} + (mc_+ + 2c_{+1}\beta/\alpha)s,$$

which cannot hold for negative s due to (2.1). The other equation leads to

$$\sqrt{(\beta + \frac{\alpha d_-}{d_+} + mc_+s)^2 - 4\beta\frac{\alpha d_-}{d_+}} = -\beta + \frac{\alpha d_-}{d_+} + (mc_+ + 2c_-)s. \tag{3.26}$$

After squaring both sides, and dividing by $4s$ we obtain $s = -s_{\rm p}$ as in (3.24), but only if the right-hand side of (3.26) is positive, which is equivalent to (3.25). The fact that $s_{\rm p} > 0$ follows from the stability condition (2.3). ♢

Remark: In the proof of Lemma 3.8 we used the LT of W in (3.6). The alternative form in (3.12) seems to suggest that $\mathbb{E}e^{-sW}$ always has a pole at $s = -s_{\rm p}$, but this is not the case. The reason is that in the proof of Corollary 3.3 we multiplied with the factor (3.14), which equals zero for $s = -s_{\rm p}$ if (3.25) does not hold.

Lemma 3.9 *For the quantities in* (3.23) *and* (3.24) *we have*

$$s_{\rm p} \leq s_{\rm b},$$

where equality holds if and only if (3.25) *holds with equality. Therefore, if the pole $-s_{\rm p}$ exists, it is larger than or equal to the branching point $-s_{\rm b}$.*

Proof: Using the expressions for $s_{\rm p}$ and $s_{\rm b}$, and using (3.5) to alleviate the notational burden somewhat, we have

$$\begin{aligned} -s_{\rm p}+s_{\rm b} &= \frac{\beta}{c_-}-\frac{\alpha d_-}{c_+(d_-+d_+)+c_-d_+}+\frac{(\sqrt{\beta d_+}-\sqrt{\alpha d_-})^2}{c_+(d_-+d_+)} \\ &= \frac{\beta d_+}{c_-d_+}-\frac{\alpha d_-}{c_+(d_-+d_+)+c_-d_+}+\frac{\beta d_++\alpha d_-}{c_+(d_-+d_+)}-\frac{2\sqrt{\alpha\beta d_+d_-}}{c_+(d_-+d_+)} \\ &= \frac{\left(\sqrt{\alpha d_-}(c_-d_+)-\sqrt{\beta d_+}\,(c_+(d_-+d_+)+c_-d_+)\right)^2}{c_-d_+(c_+(d_-+d_+))\,(c_+(d_-+d_+)+c_-d_+)} \geq 0. \end{aligned}$$

Obviously, the numerator of this expression is always positive, being zero only when (3.25) holds with equality. ♢

Thus we can distinguish between the following cases:

1. (3.25) holds with strict inequality; hence the pole $-s_{\rm p}$ exists and since it is larger than $-s_{\rm b}$, we conjecture it determines the logarithmic asymptotics (dominating the branching point);

2. (3.25) does not hold; hence a pole does not exist, so the branching point $-s_{\rm b}$ supposedly determines the logarithmic asymptotics.

In Section 5 we will prove these claims. In fact, we even provide *exact* asymptotics (that is, we identify a function $f(\cdot)$ such that $\mathbb{P}(Q_2 > x)/f(x) \to 1$ as $x \to \infty$); the form of this function will obviously depend on the case involved. It turns out there is a third case, namely the situation in which (3.25) holds with equality; then pole and branching point coincide, and determine the logarithmic asymptotics. For the latter (boundary) case similar techniques can be used, leading to yet another form for the function $f(\cdot)$.

4 Intuition behind overflow behavior

In this section we use the theory of *large deviations* to further substantiate our educated guess about the type of asymptotic behavior for the second queue content. Indeed we find that the singularities found in the previous section determine the decay, again depending on whether or not the criterion in (3.25) holds. Moreover, the current approach yields insight in the interpretation of the two different outcomes.
Let $y \in [0,1]$ denote the fraction of time the source is on. Then one could define some sort of 'cost' (per unit of time) of generating traffic at rate y by [12]

$$I(y) := \left(\sqrt{\alpha y}-\sqrt{\beta(1-y)}\right)^2.$$

Indeed, when inserting $y := \beta/(\alpha+\beta)$ — which corresponds with the source's 'average mode' — one obtains cost 0. As we will see below, this cost heuristic is rather helpful when generating guesses for decay rates.

First queue. To demonstrate how the approach works, let us first consider the decay rate of the first queue. Supposing that the source is on a fraction y of the time ($y \in [0,1]$), the first queue grows

roughly at a rate $d_+ y - d_-(1-y) =: r(y)$. In order to let the buffer build up, y needs to be larger than $\delta_1 := d_-/(d_+ + d_-)$. As argued in, among several other references, [12], it holds that

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(Q_1 > x) = - \inf_{y \geq \delta_1} \frac{I(y)}{r(y)}.$$

The interpretation is the following: if the source is on a fraction y of time, then it takes $x/r(y)$ time to exceed level x. The y that minimizes $I(y)/r(y)$ is the *most likely* fraction of time the source is on during the trajectory to overflow. This approach extends to a large class of inputs; notably, these need to be short-range dependent [9].

Tandem. A similar approach can be followed in case of a tandem queue, i.e., Model 4 in Section 2. If the source is on (that is, generating traffic at rate p_1) a fraction y of the time, the first queue grows at rate $p_1 y - c_1$ if $y > c_1/p_1$, and otherwise it remains empty. This implies that the rate of the growth of the second queue is $c_1 - c_2$ if $y > c_1/p_1$ (as traffic leaves the first queue at a rate c_1), and $p_1 y - c_2$ if $c_2/p_1 < y < c_1/p_1$ (as traffic leaves the first queue at rate $p_1 y$). We thus (heuristically) obtain

$$\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}(Q_2 > x) = - \inf_{\delta_2^{(\ell)} \leq y \leq \delta_2^{(\mathrm{u})}} \frac{I(y)}{r(y)},$$

where $\delta_2^{(\ell)} := c_2/p_1$ and $\delta_2^{(\mathrm{u})} := c_1/p_1$, and $r(y) := \min\{p_1 y, c_1\} - c_2$.
Put differently: the most likely fraction of time the source is on, is, during the path to overflow, not larger than c_1/p_1. A fraction larger than c_1/p_1 leads to queue 1 building up, but does not help building up queue 2 (compared to a fraction of exactly c_1/p_1). This heuristic was made rigorous in [14]; see also [3].
Performing the minimization, one obtains the decay rate as the minimal cost value, which equals

$$s_{\mathrm{p}} = \frac{\alpha}{p_1 - c_2} - \frac{\beta}{c_2}, \quad \text{if } c_1 \geq c_1^* := \frac{\alpha c_2^2 p_1}{\alpha c_2^2 + \beta(p_1 - c_2)^2},$$

and

$$s_{\mathrm{b}} = \frac{(\sqrt{\beta(p_1 - c_1)} - \sqrt{\alpha c_1})^2}{(c_1 - c_2)p_1}$$

else. These results can be understood as follows. If c_1 is relatively large, then the first queue is essentially 'transparent', in that it does not 'shape' the traffic that flows into the second queue — the decay rate is the same as if the traffic streams feeds immediately in the second queue (and does not depend on the particular value of c_1). If c_1 is relatively small, the buildup of the second queue is hampered by the fact that traffic can leave the first queue at a rate of at most c_1; as a result, traffic is most likely generated at a rate of exactly c_1, leading to overflow (over level x) in the second queue around time $x/(c_1 - c_2)$ — here c_1 plays a crucial role. This dichotomy has been observed for Markov fluid sources in [14], but also for other input processes; see [5, 15].

Our two-node model. We can follow the same recipe for our two-node queueing system. It is readily verified that

$$r(y) = c_{+1} y + c_{+0} \cdot y \cdot \frac{d_+}{d_-} - c_- \left(1 - \frac{y(d_+ + d_-)}{d_-}\right),$$

as $c_{+1} y + c_{+0} y d_+/d_-$ is the input rate of the second queue (when the fraction of time the source is on is y) while $c_- \mathbb{P}(Q_1 = 0) = c_- \, (1 - y(d_+ + d_-)/d_-)$ is its service rate, see (2.2). With

$$\delta_3^{(\ell)} := \frac{c_-}{c_{+1} + (c_{+0} + c_-) \cdot d_+/d_- + c_-}, \quad \delta_3^{(\mathrm{u})} := \frac{d_-}{d_- + d_+},$$

the above line of reasoning gives

$$\lim_{x\to\infty}\frac{1}{x}\log\mathbb{P}(Q_2>x) = -\inf_{\delta_3^{(\ell)}\le y\le\delta_3^{(\mathrm{u})}}\frac{I(y)}{r(y)}. \tag{4.27}$$

With y^* the optimizer in the right-hand side of the above variational problem, we distinguish two cases: $y^*=\delta_3^{(\mathrm{u})}$ and $y^*\in[\delta_3^{(\ell)},\delta_3^{(\mathrm{u})})$. We first solve the 'unconstrained' problem

$$\inf_{y\ge\delta_3^{(\ell)}}\frac{I(y)}{r(y)}.$$

Tedious computations yield that the minimum is attained at

$$y=\frac{\alpha c_-^2 d_-^2}{\alpha c_-^2 d_-^2+\beta(c_{+1}d_-+c_{+0}d_++c_-d_+)^2}. \tag{4.28}$$

If this value is smaller than $\delta_3^{(\mathrm{u})}$, i.e.,

$$\frac{\alpha c_-^2 d_-^2}{\alpha c_-^2 d_-^2+\beta(c_{+1}d_-+c_{+0}d_++c_-d_+)^2}<\frac{d_-}{d_-+d_+}, \tag{4.29}$$

we obviously have that y^* equals (4.28). But now, remarkably, observe that criterion (4.29) is equivalent to (3.25)! Then it is readily verified that in this situation the decay rate in (4.27) equals the pole s_{p}, as given in (3.24). In the other case, i.e., $y^*=d_-/(d_-+d_+)$, the decay rate in (4.27) equals the branching point s_{b}, as given in (3.23). Thus, in both cases we find the same decay rate as through the explicit derivation above, and also the criterion that determines which of the two dominates is the same. Heuristics regarding the path to overflow are similar to those presented for the tandem.

5 Exact asymptotics

In this section we will prove the exact asymptotics of the density $f_{Q_2}(x)$ of the second queue content as $x\to\infty$. The proof will be based on Theorem 3.6, for which we will derive the exact asymptotics of $f_W(x)$ and $f_A(x)$. We then combine this knowledge to find the asymptotics of Q_2. We first focus on the cases in which pole and branching point do not coincide, leaving the boundary case for the last subsection.

We start off by by stating a number of useful results. The first, dealing with the M/M/1 busy-period distribution, can be found in, e.g., [4].

Lemma 5.1 *For the density of the busy period P of an* M/M/1 *queue with arrival rate λ and service rate μ, we have*

$$f_P(t) \;=\; \frac{1}{t\sqrt{\lambda/\mu}}e^{-(\lambda+\mu)t}I_1(2t\sqrt{\lambda\mu}) \quad\sim\quad K_P t^{-3/2}e^{-(\sqrt{\mu}-\sqrt{\lambda})^2 t},\quad t\to\infty,$$

where

$$K_P := \frac{1}{2\sqrt{\pi\lambda}}\frac{1}{(\lambda/\mu)^{1/4}}.$$

For the density of the residual busy period R we have

$$f_R(t) \;=\; \frac{\mathbb{P}(P>t)}{\mathbb{E}P} \quad\sim\quad \frac{1}{(\sqrt{\mu}-\sqrt{\lambda})^2\mathbb{E}P}f_P(t) \;=\; \frac{\sqrt{\mu}+\sqrt{\lambda}}{\sqrt{\mu}-\sqrt{\lambda}}f_P(t)$$

and thus

$$f_R(t) \sim K_R t^{-3/2} e^{-(\sqrt{\mu}-\sqrt{\lambda})^2 t}, \quad t \to \infty,$$

where

$$K_R := \frac{\sqrt{\mu}+\sqrt{\lambda}}{\sqrt{\mu}-\sqrt{\lambda}} K_P.$$

The other useful lemma follows below. Although most, if not all, of this lemma is known, see e.g. [1, 16], we include it, since it plays an important role in what follows. We also provide a proof in the appendix, which elegantly shows how large values of $X+Y$ are typically attained; e.g., in case (ii) this typically happens due to a large value of X or Y, but *not* by both taking large values (even though X and Y are not heavy-tailed).

Lemma 5.2 *Let X and Y be independent random variables with densities satisfying*

$$f_X(x) \sim K_X x^{-p} e^{-\sigma x}, \quad f_Y(x) \sim K_Y x^{-q} e^{-\tau x},$$

as $x \to \infty$, for some constants $p, q \geq 0$ and $\sigma, \tau, K_X, K_Y > 0$.

(i) If either $\sigma < \tau$ holds, or it holds that $\sigma = \tau$ and $p < q$ and $q > 1$, then we have as $x \to \infty$,

$$f_{X+Y}(x) \;\sim\; \mathbb{E}e^{\sigma Y} f_X(x) \;\sim\; \mathbb{E}e^{\sigma Y} K_X x^{-p} e^{-\sigma x}.$$

(ii) If both $\sigma = \tau$ and $p = q > 1$ hold, then we have as $x \to \infty$,

$$f_{X+Y}(x) \;\sim\; \mathbb{E}e^{\sigma Y} f_X(x) + \mathbb{E}e^{\tau X} f_Y(x) \;\sim\; \left(K_X \mathbb{E}e^{\sigma Y} + K_Y \mathbb{E}e^{\sigma X}\right) x^{-p} e^{-\sigma x}.$$

5.1 Exact asymptotics of the density of W

Our starting point is the LT of W in (3.12), from which we immediately obtain the density as

$$\begin{aligned} f_W(x) &= \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right)\frac{\beta}{c_-}\left(e^{-s_{\rm p} x} - \frac{1}{1+mc_+/c_-}\int_0^x f_{c_+B}(u) e^{-s_{\rm p}(x-u)} du\right) \\ &= \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right)\frac{\beta}{c_-}\left(1 - \frac{1}{1+mc_+/c_-}\int_0^x f_{c_+B}(u) e^{s_{\rm p} u} du\right) e^{-s_{\rm p} x} \\ &= \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right)\frac{\beta}{c_-}\left(1 - \frac{\mathbb{E}e^{s_{\rm p} c_+ B}}{1+mc_+/c_-} + \frac{1}{1+mc_+/c_-}\int_x^\infty f_{c_+B}(u) e^{s_{\rm p} u} du\right) e^{-s_{\rm p} x}, \end{aligned}$$

where $f_{c_+B}(u)$ denotes the density of c_+B. Since $c_+B \stackrel{\rm d}{=} c_+ m P$, where P is the busy period of an M/M/1 with arrival rate β and service rate $\alpha d_-/d_+$, it holds by Lemma 5.1 that

$$f_{c_+B}(u) = \frac{1}{c_+ m} f_P\left(\frac{u}{c_+ m}\right) = \frac{1}{u}\sqrt{\frac{\beta d_+}{\alpha d_-}} e^{-\frac{\alpha d_- + \beta d_+}{d_+ + d_-}\frac{u}{c_+}} I_1\left(2\frac{u}{c_+}\frac{\sqrt{\alpha d_- \beta d_+}}{d_+ + d_-}\right)$$

and as $u \to \infty$,

$$\begin{aligned} f_{c_+B}(u) \;&\sim\; \frac{K_P}{mc_+}\left(\frac{u}{mc_+}\right)^{-3/2} \exp\left(-\left(\sqrt{\frac{\alpha d_-}{d_+}} - \sqrt{\beta}\right)^2 \frac{u}{mc_+}\right) \;=\; K_{c_+B}\, u^{-3/2} e^{-s_{\rm b} u}, \\ &\text{with } K_{c_+B} := \frac{1}{2}\sqrt{\frac{mc_+}{\pi\beta}}\left(\frac{\alpha d_-}{\beta d_+}\right)^{1/4}. \end{aligned}$$

The above implies that as $x \to \infty$ (note that $s_{\rm p} - s_{\rm b} < 0$ by Lemma 3.9 and the assumption that $s_{\rm p} \neq s_{\rm b}$),

$$\int_x^\infty f_{c_+B}(u) e^{s_{\rm p} u} du \sim \frac{K_{c_+B}}{s_{\rm p} - s_{\rm b}} x^{-3/2} e^{(s_{\rm p} - s_{\rm b})x}. \tag{5.30}$$

Furthermore, using (3.8) we can write

$$\begin{aligned} 1 - \frac{\mathbb{E}e^{s_{\rm p} c_+ B}}{1 + mc_+/c_-} &= 1 - \frac{1}{\frac{c_- + mc_+}{c_-}} \frac{\beta + \frac{\alpha d_-}{d_+} - mc_+ s_{\rm p} - \sqrt{(\beta + \frac{\alpha d_-}{d_+} - mc_+ s_{\rm p})^2 - 4\beta \frac{\alpha d_-}{d_+}}}{2\beta} \\ &= \frac{2\beta \frac{c_- + mc_+}{c_-} - \beta - \frac{\alpha d_-}{d_+} + mc_+ s_{\rm p} + \sqrt{(\beta + \frac{\alpha d_-}{d_+} - mc_+ s_{\rm p})^2 - 4\beta \frac{\alpha d_-}{d_+}}}{2\beta \frac{c_- + mc_+}{c_-}}. \end{aligned}$$

Since

$$\beta + \frac{\alpha d_-}{d_+} - mc_+ s_{\rm p} = \beta \frac{c_- + mc_+}{c_-} + \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+},$$

the above simplifies to

$$\begin{aligned} 1 - \frac{\mathbb{E}e^{s_{\rm p} c_+ B}}{1 + mc_+/c_-} &= \frac{\beta \frac{c_- + mc_+}{c_-} - \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+} + \sqrt{\left(\beta \frac{c_- + mc_+}{c_-} - \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+}\right)^2}}{2\beta \frac{c_- + mc_+}{c_-}} \\ &= \frac{\beta \frac{c_- + mc_+}{c_-} - \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+} + \left|\beta \frac{c_- + mc_+}{c_-} - \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+}\right|}{2\beta \frac{c_- + mc_+}{c_-}}. \end{aligned}$$

Hence,

$$1 - \frac{\mathbb{E}e^{s_{\rm p} c_+ B}}{1 + mc_+/c_-} = \begin{cases} 1 - \frac{\alpha d_-}{\beta d_+} \frac{c_-^2}{(c_- + mc_+)^2} & \text{if } \beta \frac{c_- + mc_+}{c_-} \geq \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+}, \\ 0 & \text{otherwise.} \end{cases}$$

Since the condition

$$\beta \frac{c_- + mc_+}{c_-} \geq \frac{\alpha d_-}{d_+} \frac{c_-}{c_- + mc_+}$$

is equivalent with condition (3.25), the density of W can be written as

$$\begin{aligned} f_W(x) &= \left(1 - \frac{\beta}{c_-} \frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right) \frac{\beta}{c_-} \\ &\quad \times \left(1 - \frac{\alpha d_-}{\beta d_+} \frac{c_-^2}{(c_- + mc_+)^2} + \frac{c_-}{c_- + mc_+} \int_x^\infty f_{c_+B}(u) e^{s_{\rm p} u} du\right) e^{-s_{\rm p} x}, \end{aligned}$$

if (3.25) holds, or otherwise as

$$f_W(x) = \left(1 - \frac{\beta}{c_-} \frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right) \left(\frac{\beta}{c_- + mc_+} \int_x^\infty f_{c_+B}(u) e^{s_{\rm p} u} du\right) e^{-s_{\rm p} x}.$$

We can now state the following.

Lemma 5.3 *The asymptotic behavior of $f_W(x)$ as $x \to \infty$ is given by either of the following. When condition (3.25) holds with strict inequality, $f_W(x) \sim K_{W,{\rm p}}\, e^{-s_{\rm p} x}$, with*

$$K_{W,{\rm p}} := \left(1 - \frac{\beta}{c_-} \frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right) \frac{\beta}{c_-} \left(1 - \frac{\alpha d_-}{\beta d_+} \frac{c_-^2}{(c_- + mc_+)^2}\right). \tag{5.31}$$

When condition (3.25) *does not hold,* $f_W(x) \sim K_{W,\mathrm{b}}\, x^{-3/2} e^{-s_\mathrm{b} x}$, *with*

$$K_{W,\mathrm{b}} := \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right) \frac{\beta/2}{c_- + m c_+} \sqrt{\frac{c_+ m}{\pi\beta}} \left(\frac{\alpha d_-}{\beta d_+}\right)^{1/4} \frac{1}{s_\mathrm{p} - s_\mathrm{b}}. \tag{5.32}$$

Proof: Immediate from the above. ♢

Remark: By explicitly inverting the LT, we above found the density of W, as well as its asymptotics. We could also have used the fact that W is the waiting time in an M/G/1 queue (see Lemma 3.1). If the so-called *Lundberg equation* $\mathbb{E}e^{sc_+B} = 1 + s/(\beta/c_-)$ has a positive solution s_p, which is equivalent to (3.25) — see Lemma 3.8 and its proof — the Cramér-Lundberg approximation leads to the purely exponential form displayed in Lemma 5.3. When the Lundberg equation fails to have a positive solution, the asymptotics could be found by applying the random walk results of, e.g., Section 5.2 of Dieker [7], specialized to the M/G/1 case (with i.i.d. increments distributed as $c_+B - c_-Y$); then we obtain the mixed polynomial-exponential form mentioned in Lemma 5.3.

5.2 Exact asymptotics of the densities of A and Q_2

In the following lemma we use the expresssion for the LT of A in Corollary 3.5 to derive the asymptotic behavior of the density $f_A(x)$ as $x \to \infty$.

Lemma 5.4 *The asymptotic behavior of the density of* A *as* $x \to \infty$ *is given by*

$$f_A(x) \sim K_A x^{-3/2} e^{-s_\mathrm{b} x},$$

where s_b *is the same as in Lemma* 5.3 *and*

$$\begin{aligned} K_A \;:=\; & \frac{1}{2}\sqrt{\frac{mc_+}{\pi\beta}} \left(\frac{\alpha d_-}{\beta d_+}\right)^{1/4} \frac{\sqrt{\alpha d_-} + \sqrt{\beta d_+}}{\sqrt{\alpha d_-} - \sqrt{\beta d_+}} \\ & \times \left[\frac{c_+}{c_{+1}} + \left(\frac{c_+}{c_{+0}} - \frac{c_+}{c_{+1}}\right) \frac{\alpha d_+ c_{+0}/c_{+1} + 2\sqrt{\alpha\beta d_+ d_-} - \beta d_+}{\alpha d_+ c_{+0}/c_{+1} + 2\sqrt{\alpha\beta d_+ d_-} + \beta d_- c_{+1}/c_{+0}}\right]. \end{aligned}$$

Proof: We first collect the exact asympotics for the densities of c_+B and c_+B^*. Those of c_+B were already found in the previous subsection (using that $c_+B \stackrel{d}{=} mc_+P$ and using Lemma 5.1); they satisfy

$$f_{c_+B}(x) \sim K_{c_+B}\, x^{-3/2} e^{-s_\mathrm{b} x}.$$

Similarly, since $c_+B^* \stackrel{\mathrm{d}}{=} mc_+R$, where R is the residual busy period of the related M/M/1 queue, we can again use Lemma 5.1 to find that

$$f_{c_+B^*}(x) \;\sim\; K_{c_+B^*} x^{-3/2} e^{-s_\mathrm{b} x} \quad \text{with} \quad K_{c_+B^*} = \frac{\sqrt{\alpha d_-} + \sqrt{\beta d_+}}{\sqrt{\alpha d_-} - \sqrt{\beta d_+}} K_{c_+B}.$$

Hence, both c_+B and c_+B^* have asymptotic behavior as X in Lemma 5.2, with $\sigma = s_\mathrm{b}$ and $p = 3/2$. Looking at the expression in (3.21), we can immediately apply this lemma (notice that $R(sc_+)$ is the LT of c_+B^*, and that $s_\mathrm{b} < \gamma$) to find:

$$\begin{aligned} f_A(x) \;\sim\; & \frac{c_+}{c_{+1}} f_{c_+B*}(x) \;+\; \frac{c_{+1} - c_{+0}}{mc_{+1}} \frac{\alpha/c_{+1}}{\gamma} \frac{\gamma}{\gamma - s_\mathrm{b}} f_{c_+B^*}(x) \\ & + \frac{c_{+1} - c_{+0}}{mc_{+1}} \frac{\beta/c_{+0}}{\gamma} \frac{\gamma}{\gamma - s_\mathrm{b}} \left(R(-s_\mathrm{b}c_+) f_{c_+B}(x) + \mathbb{E}e^{s_\mathrm{b}c_+B} f_{c_+B^*}(x)\right). \end{aligned}$$

Hence, $f_A(x) \sim K_A x^{-3/2} e^{-s_b x}$, for some constant K_A. To find this constant we note that

$$R(-s_b c_+) = 1 + \sqrt{\frac{\alpha d_-}{\beta d_+}} \quad \text{and} \quad \mathbb{E}e^{s_b c_+ B} = \sqrt{\frac{\alpha d_-}{\beta d_+}},$$

so that we can write

$$K_A = \frac{c_+}{c_{+1}} K_{c_+ B*} \left[1 + \frac{c_{+1} - c_{+0}}{m c_+ (\gamma - s_b)} \left(\frac{\alpha}{c_{+1}} + \frac{\beta}{c_{+0}} \left(2\sqrt{\frac{\alpha d_-}{\beta d_+}} - 1 \right) \right) \right].$$

Using the fact that

$$m c_+ d_+ (\gamma - s_b) = \alpha d_- \frac{c_{+0} d_+}{c_{+1} d_-} + \beta d_+ \frac{c_{+1} d_-}{c_{+0} d_+} + 2\sqrt{\alpha \beta d_+ d_-},$$

this can be rewritten to the form of K_A as stated in the lemma. ♢

Finally, now that we have the asymptotic behaviors of $f_W(x)$ and $f_A(x)$ at our disposal, we come back to Theorem 3.6, from which we have

$$f_{Q_2}(x) = (1 - \rho_1) f_W(x) + \rho_1 \int_0^x f_W(u) f_A(x - u) du. \tag{5.33}$$

We apply Lemma 5.2 again to find the following:

Theorem 5.5 *The asymptotic behavior of $f_{Q_2}(x)$ as $x \to \infty$ is given by either of the following. When condition* (3.25) *holds with strict inequality, $f_{Q_2}(x) \sim K_{Q_2,\mathrm{p}}\, e^{-s_\mathrm{p} x}$, with*

$$K_{Q_2,\mathrm{p}} := (1 - \rho_1) K_{W,\mathrm{p}} + \rho_1 (K_{W,\mathrm{p}}\, \mathbb{E}e^{s_\mathrm{p} A}).$$

When condition (3.25) *does not hold, $f_{Q_2}(x) \sim K_{Q_2,\mathrm{b}}\, x^{-3/2} e^{-s_\mathrm{b} x}$, with*

$$K_{Q_2,\mathrm{b}} := (1 - \rho_1) K_{W,\mathrm{b}} + \rho_1 (K_A\, \mathbb{E}e^{s_\mathrm{b} W} + K_{W,\mathrm{b}}\, \mathbb{E}e^{s_\mathrm{b} A}).$$

Proof: Immediate from (5.33) and both parts of Lemma 5.2 (noting that $s_\mathrm{p} < s_\mathrm{b}$ in the first case). ♢

Corollary 5.6 *The asymptotic behavior of the tail probability $\mathbb{P}(Q_2 > x)$ as $x \to \infty$ is given by either of the following. When condition (3.25) holds with strict inequality,*

$$\mathbb{P}(Q_2 > x) \sim \frac{K_{Q_2,\mathrm{p}}}{s_\mathrm{p}}\, e^{-s_\mathrm{p} x}. \tag{5.34}$$

When condition (3.25) does not hold,

$$\mathbb{P}(Q_2 > x) \sim \frac{K_{Q_2,\mathrm{b}}}{s_\mathrm{b}}\, x^{-3/2} e^{-s_\mathrm{b} x}. \tag{5.35}$$

Proof: Immediate. ♢

5.3 Exact asymptotics of the density of Q_2 when $s_{\rm b} = s_{\rm p}$

When (3.25) holds with equality, we know from Lemma 3.9 that the pole $s_{\rm p}$ and the branching point $s_{\rm b}$ coincide, both being equal to

$$s_{\rm pb} := \frac{\beta m c_+}{c_-^2}.$$

As a consequence, we find a different asymptotic behavior for $f_W(x)$, in some sense lying in between the two outcomes in Lemma 5.3. In the derivation of the analogue of this lemma, we (again) find that

$$f_W(x) = \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right)\left(\frac{\beta}{c_- + m c_+}\int_x^\infty f_{c_+B}(u)e^{s_{\rm pb}u}du\right)e^{-s_{\rm pb}x},$$

but instead of (5.30) we now find

$$\int_x^\infty f_{c_+B}(u)e^{s_{\rm pb}u}du \sim 2K_{c_+B}\,x^{-1/2}.$$

The asymptotic behavior of $f_W(x)$ as $x \to \infty$ is therefore given by $f_W(x) \sim K_{W,\rm pb}\,x^{-1/2}e^{-s_{\rm pb}x}$, with

$$K_{W,\rm pb} := \left(1 - \frac{\beta}{c_-}\frac{c_+(d_+ + d_-)}{\alpha d_- - \beta d_+}\right)\frac{\beta}{c_- + m c_+}\sqrt{\frac{c_+ m}{\pi\beta}}\left(\frac{\alpha d_-}{\beta d_+}\right)^{1/4}. \tag{5.36}$$

Since the behavior of $f_A(x)$ is the same as before, we only need to apply the first part of Lemma 5.2 with $\sigma = \tau = s_{\rm pb}$, $p = 1/2$ and $q = 3/2$ to find the following.

Theorem 5.7 *When condition* (3.25) *holds with equality, the asymptotic behavior of* $f_{Q_2}(x)$ *as* $x \to \infty$ *is given by* $f_{Q_2}(x) \sim K_{Q_2,\rm pb}\,x^{-1/2}e^{-s_{\rm pb}x}$, *with*

$$K_{Q_2,\rm pb} := (1-\rho_1)K_{W,\rm pb} + \rho_1(K_{W,\rm pb}\,\mathbb{E}e^{s_{\rm pb}A}).$$

The asymptotic behavior of the tail probability $\mathbb{P}(Q_2 > x)$ *as* $x \to \infty$ *is given by*

$$\mathbb{P}(Q_2 > x) \sim \frac{K_{Q_2,\rm pb}}{s_{\rm pb}}\,x^{-1/2}e^{-s_{\rm pb}x}. \tag{5.37}$$

6 Concluding remarks

In this paper we considered a rather general class of two-node fluid queues, that includes the classical tandem and priority systems. In these systems the first queue can be analyzed in isolation by applying standard techniques; the evolution of the other queue, however, is affected by the first queue being empty or not, which makes this queue substantially harder to analyze. We explicitly derived the buffer-content distribution of this second queue (in terms of its Laplace transform), as well as its tail asymptotics, relying exclusively on probabilistic argumentation. Interestingly, there is a sharp dichotomy, in that two asymptotic regimes can be distinguished; large deviations theory provides an appealing interpretation of these regimes.

A direction for further research is to broaden the class of input models. In this paper we restricted ourselves to fairly elementary Markov fluid input, but, suggested by e.g. [3, 14], one would expect that the dichotomy of the tail asymptotics carries over to a considerably larger class of inputs. The recent results in [8] may give a handle on resolving this issue.

Appendix: Proof of Lemma 5.2

In this appendix we provide a proof of Lemma 5.2. The proof explains how large values of $X+Y$ are typically attained. As can be expected this happens due to X taking a large value when the tail of X is heavier than that of Y. However, when both tails are equally heavy, it typically happens due to a large value of either X or Y, but *not* by both taking large values, even though X and Y are not heavy-tailed.
We repeat the statement of the lemma for convenience.

Lemma 5.2 *Let X and Y be independent random variables with densities satisfying*

$$f_X(x) \sim K_X x^{-p} e^{-\sigma x}, \quad f_Y(x) \sim K_Y x^{-q} e^{-\tau x},$$

as $x \to \infty$, for some constants $p, q \geq 0$ and $\sigma, \tau, K_X, K_Y > 0$.

(i) If either $\sigma < \tau$ holds, or it holds that $\sigma = \tau$ and $p < q$ and $q > 1$, then we have as $x \to \infty$,

$$f_{X+Y}(x) \ \sim \ \mathbb{E}e^{\sigma Y} f_X(x) \ \sim \ \mathbb{E}e^{\sigma Y} K_X x^{-p} e^{-\sigma x}.$$

(ii) If both $\sigma = \tau$ and $p = q > 1$ hold, then we have as $x \to \infty$,

$$f_{X+Y}(x) \ \sim \ \mathbb{E}e^{\sigma Y} f_X(x) + \mathbb{E}e^{\tau X} f_Y(x) \ \sim \ \left(K_X \mathbb{E}e^{\sigma Y} + K_Y \mathbb{E}e^{\sigma X}\right) x^{-p} e^{-\sigma x}.$$

Proof: To prove the first part of the lemma we first fix some small $\epsilon > 0$ and write

$$\begin{aligned} f_{X+Y}(x) &= \int_0^x f_Y(u) f_X(x-u) du \\ &= \int_0^{\epsilon x} f_Y(u) f_X(x-u) du + \int_{\epsilon x}^{(1-\epsilon)x} f_Y(u) f_X(x-u) du + \int_{(1-\epsilon)x}^{x} f_Y(u) f_X(x-u) du. \end{aligned}$$

For the first integral in this sum we have

$$x^p e^{\sigma x} \int_0^{\epsilon x} f_Y(u) f_X(x-u) du = \int_0^{\epsilon x} e^{\sigma u} f_Y(u) (x-u)^p e^{\sigma(x-u)} f_X(x-u) \left(\frac{x}{x-u}\right)^p du$$

The given asymptotics of f_X imply that for any $\delta > 0$ we have, for x sufficiently large (and any $u \in [0, \epsilon x]$),

$$K_X - \delta \leq (x-u)^p e^{\sigma(x-u)} f_X(x-u) \leq K_X + \delta,$$

so that we find

$$x^p e^{\sigma x} \int_0^{\epsilon x} f_Y(u) f_X(x-u) du \geq \int_0^{\epsilon x} e^{\sigma u} f_Y(u) (K_X - \delta) du,$$

and hence

$$\liminf_{x \to \infty} x^p e^{\sigma x} \int_0^{\epsilon x} f_Y(u) f_X(x-u) du \geq \mathbb{E}e^{\sigma Y} K_X.$$

Keeping in mind the asymptotic behavior of f_Y it may be good to note that $\mathbb{E}e^{\sigma Y}$ is indeed finite when $\sigma < \tau$, while it is also finite when $\sigma = \tau$, due to $q > 1$.

To find an upper bound for the first integral, we write it in a slightly different form; with $\delta > 0$ and sufficiently large x we have

$$\begin{aligned} x^p e^{\sigma x} \int_0^{\epsilon x} f_Y(u) f_X(x-u) du &= (1-\epsilon)^{-p} \int_0^{\epsilon x} e^{\sigma u} f_Y(u) (1-\epsilon)^p x^p e^{\sigma(x-u)} f_X(x-u) du \\ &\leq (1-\epsilon)^{-p} \int_0^{\epsilon x} e^{\sigma u} f_Y(u) (x-u)^p e^{\sigma(x-u)} f_X(x-u) du \\ &\leq (1-\epsilon)^{-p} \int_0^{\epsilon x} e^{\sigma u} f_Y(u) (K_X + \delta) du, \end{aligned}$$

and hence

$$\limsup_{x\to\infty} x^p e^{\sigma x} \int_0^{\epsilon x} f_Y(u) f_X(x-u) du \leq (1-\epsilon)^{-p} \mathbb{E} e^{\sigma Y} K_X.$$

For the second integral we can write

$$\begin{aligned} &\limsup_{x\to\infty} x^p e^{\sigma x} \int_{\epsilon x}^{(1-\epsilon)x} f_Y(u) f_X(x-u) du \\ &= \limsup_{x\to\infty} \int_{\epsilon x}^{(1-\epsilon)x} \frac{x^p}{u^q (x-u)^p} e^{(\sigma-\tau)u} u^q e^{\tau u} f_Y(u) (x-u)^p e^{\sigma(x-u)} f_X(x-u) du \\ &\leq \limsup_{x\to\infty} \int_{\epsilon x}^{(1-\epsilon)x} \frac{x^p}{(\epsilon x)^q (\epsilon x)^p} e^{(\sigma-\tau)u} K_Y K_X du \\ &= \limsup_{x\to\infty} \frac{K_X K_Y}{\epsilon^{p+q} x^q} \int_{\epsilon x}^{(1-\epsilon)x} e^{(\sigma-\tau)u} du \;=\; 0 \end{aligned}$$

when $\sigma < \tau$, but also when $\sigma = \tau$ and $q > 1$.
Finally, for the third integral we have, assuming that $\sigma < \tau$, that

$$\begin{aligned} &\limsup_{x\to\infty} x^p e^{\sigma x} \int_{(1-\epsilon)x}^{x} f_Y(u) f_X(x-u) du \\ &\leq \limsup_{x\to\infty} \int_{(1-\epsilon)x}^{x} \frac{x^{p-q}}{(1-\epsilon)^q} u^q e^{\tau u} f_Y(u) e^{(\sigma-\tau+\tau\epsilon)x} f_X(x-u) du \\ &\leq \limsup_{x\to\infty} (1-\epsilon)^{-q} K_Y x^{p-q} e^{(\sigma-\tau+\tau\epsilon)x} \;=\; 0, \end{aligned}$$

assuming that ϵ is chosen such that $\sigma - \tau + \tau\epsilon < 0$. On the other hand, when $\sigma = \tau$ we have

$$\begin{aligned} &\limsup_{x\to\infty} x^p e^{\sigma x} \int_{(1-\epsilon)x}^{x} f_Y(u) f_X(x-u) du \\ &\leq \limsup_{x\to\infty} x^p \int_{(1-\epsilon)x}^{x} u^{-q} K_Y e^{\sigma(x-u)} f_X(x-u) du \qquad (6.38) \\ &= \limsup_{x\to\infty} x^p \left(\int_0^{\epsilon} (x-u)^{-q} K_Y e^{\sigma u} f_X(u) du + \int_{\epsilon}^{\epsilon x} (x-u)^{-q} K_Y e^{\sigma u} f_X(u) du \right) \\ &= 0 + \limsup_{x\to\infty} x^p \int_{\epsilon}^{\epsilon x} u^{-p} (x-u)^{-q} K_X K_Y du \\ &= \limsup_{x\to\infty} \frac{K_X K_Y}{1-q} x^p \left(\left[-u^{-p} (x-u)^{1-q} \right]_{\epsilon}^{\epsilon x} - \int_{\epsilon}^{\epsilon x} p u^{-p-1} (x-u)^{1-q} du \right) \;=\; 0, \end{aligned}$$

where the last step is due to integration by parts; note that we used $p < q$ and $q > 1$.

Taking the three terms together, we have

$$\begin{aligned}\liminf_{x\to\infty} x^p e^{\sigma x}\int_0^x f_Y(u)f_X(x-u)du &\geq \mathbb{E}e^{\sigma Y}K_X,\\ \limsup_{x\to\infty} x^p e^{\sigma x}\int_0^x f_Y(u)f_X(x-u)du &\leq (1-\epsilon)^{-p}\mathbb{E}e^{\sigma Y}K_X,\end{aligned}$$

and hence, letting $\epsilon \to 0$,

$$\lim_{x\to\infty} x^p e^{\sigma x}\int_0^x f_Y(u)f_X(x-u)du = \mathbb{E}e^{\sigma Y}K_X,$$

which proves the part *(i)* of the lemma. The proof of part *(ii)*, in which $\sigma = \tau$ and $p = q > 1$, is completely similar, except for the third integral. The limsup in (6.38) now becomes $\leq (1-\epsilon)^{-q}K_Y\mathbb{E}e^{\sigma X}$, and since the liminf can be shown to be $K_Y\mathbb{E}e^{\sigma X}$ (or by using the full symmetry with the first integral term), the result is easily shown. $\diamondsuit$

References

[1] J. Abate and W. Whitt (1997). Asymptotics for M/G/1 low-priority waiting-time tail probabilities. *Queueing Systems* 25, pp. 173–233.

[2] S. Asmussen (2003), *Applied Probability and Queues*, second edition, Springer.

[3] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin (1994). Effective bandwidth and fast simulation of ATM intree networks. *Perf. Eval.* 20, pp. 45-65.

[4] D. Cox and W. Smith (1961). *Queues,* Methuen.

[5] K. Dębicki, M. Mandjes, and M. van Uitert (2007). A tandem queue with Lévy input: a new representation of the downstream queue length. *Probab. Engg. Inf. Sci.* 21, pp. 83–107.

[6] T. Dieker (2006). *Extremes and fluid queues*, Ph.D. Thesis, University of Amsterdam.

[7] T. Dieker (2006). Applications of factorization embeddings for Lévy processes. *Adv. Appl. Probab.* 38, pp. 768–791.

[8] T. Dieker and M. Mandjes (2007). Extremes of Markov-additive processes with one-sided jumps, with queueing applications. Submitted.

[9] P. Glynn and W. Whitt (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Studies in Applied Probability, Papers in Honour of Lajos Takács,* J. Galambos and J. Gani (eds.), Applied Probability Trust, Sheffield, England, pp. 131–156.

[10] O. Kella and W. Whitt (1992.) A storage model with a two state random environment. *Oper. Res.* 40, pp. S257–S262.

[11] O. Kella (2001). Markov-modulated feedforward fluid networks. *Queueing Systems* 37, pp. 141–161.

[12] G. Kesidis, J. Walrand and C.-S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. *IEEE/ACM Trans. Netw.* 1, pp. 424–428.

[13] D. Kroese and W. Scheinhardt (2001). Joint distributions for interacting fluid queues. *Queueing Systems* 37, pp. 99–139.

[14] M. Mandjes (1998). Asymptotically optimal importance sampling for tandem queues with Markov fluid input. *AEÜ Int. J. Electr. Comm.* 52, pp. 152–161.

[15] M. Mandjes and M. van Uitert (2005). Sample-path large deviations for tandem and priority queues with Gaussian inputs. *Ann. Appl. Probab.* 15, pp. 1193–1226.

[16] A. Pakes (2004). Convolution equivalence and infinite divisibility, *J. Appl. Probab.* 41, pp. 407–424.

[17] Z. Palmowski, T. Rolski (2004). On the busy period asymptotics of $GI/G/1$ queues. *Adv. Appl. Probab.* 38, 792–803.

[18] W. Scheinhardt (1998). *Markov-modulated and feedback fluid queues*, Ph.D. Thesis, University of Twente.

[19] W. Scheinhardt and B. Zwart (2002). A tandem fluid queue with gradual input. *Probab. Engg. Inf. Sci.* 16, pp. 29–45.