Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers

TU/e technische universiteit eindhoven

/department of mathematics and computing science

Memorandum COSOR 99-20

Waiting-time asymptotics for the M/G/2 queue with heterogeneous servers

O.J. Boxma, Q. Deng and A.P. Zwart

Eindhoven, November 1999
The Netherlands

Waiting-time asymptotics for the $M/G/2$ queue with heterogeneous servers

O.J. Boxma[1], Q. Deng and A.P. Zwart

Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

November 26, 1999

Abstract

This paper considers a heterogeneous $M/G/2$ queue. The service times at server 1 are exponentially distributed, and at server 2 they have a general distribution $B(\cdot)$. We present an exact analysis of the queue length and waiting time distribution in case $B(\cdot)$ has a rational Laplace-Stieltjes transform. When $B(\cdot)$ is regularly varying at infinity of index $-\nu$, we determine the tail behaviour of the waiting time distribution. This tail is shown to be semi-exponential if the arrival rate is lower than the service rate of the exponential server, and regularly varying at infinity of index $1-\nu$ if the arrival rate is higher than that service rate.

1 Introduction

This paper considers a heterogeneous $M/G/2$ queue. Customers arrive according to a Poisson process with rate λ. The queueing discipline is first-come-first-served, where we make the additional convention that when a customer arrives and there is no other customer in the system, he receives service from server 1 immediately. The service time distribution of a customer depends on the server who serves him. The service times at server 1 are exponentially distributed with rate μ, and at server 2 they have a general distribution $B(\cdot)$ with mean β. It is known from the theory of multi-server queues, that the steady-state queue length and waiting time distributions exist if $\lambda < \mu + 1/\beta$. In the sequel, we assume this condition to hold.

We are interested in the steady-state waiting time distribution in this model, and in particular in its tail behaviour in case $B(\cdot)$ is regularly varying at infinity of index $-\nu$, i.e.,

$$1 - B(t) \sim t^{-\nu}L(t), \quad t \to \infty, \tag{1.1}$$

where $L(t)$ is a slowly varying function [4] (in the sequel, $f(t) \sim g(t)$ for $t \to \infty$ denotes $\lim_{t\to\infty} f(t)/g(t) = 1$). Our motivation to study this problem stems, generally, from the present-day importance of heavy-tailed phenomena in queueing models of modern communication systems; more specifically, it stems from recent results and conjectures [19, 20, 24] concerning finiteness of waiting-time moments and heavy-tailed phenomena in multi-server queues.

In the classical $GI/G/1$ queue, it is well-known [7] that the waiting time tail is regularly varying of index $1-\nu$ if (and only if) the service time tail is regularly varying of index $-\nu$. In fact, Pakes [17] has proven that, in the $GI/G/1$ queue with the larger class of subexponential

[1] also: CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands.

[4] residual service times, the tail of the steady-state waiting time W is related to the tail of the residual service time B^{res} in the following way:

$$\mathbf{P}(W > t) \sim \frac{\rho}{1-\rho}\mathbf{P}(B^{res} > t), \quad t \to \infty. \tag{1.2}$$

Here ρ denotes the traffic load.

The tail behaviour of the waiting time in *multi*-server queues is an almost completely open problem. Recent results suggest that the waiting time tail may not always be as heavy as the tail of the residual service time. For example, in the $GI/G/1$ queue it is known that the j-th moment of the steady-state waiting time is finite if and only if the $(j+1)$-th moment of the service time is finite, and the 'if'-part of this statement also holds for the $GI/G/s$ queue, cf. [14]; however, Scheller-Wolf and Sigman [19, 20] have shown that the 'only if'-part does not hold. In particular, if the offered traffic to the s-server queue is less than $s-1$, then the mean waiting time is finite if the mean service time is finite [20]. This suggests that the tail of $P(W > t)$ may be less heavy than that of $P(B^{res} > t)$. Bounds for the waiting time tail, that were partially proven and partially conjectured in [24], also point in the direction of different waiting time tail behaviour for different regions of the traffic load ρ. In [15], Korshunov derives asymptotic lower and upper bounds for $\mathbf{P}(W > t)$ in the GI/GI/2 queue. These bounds are (up to a constant) exact, thereby strengthening the results of [24]. Again, the crucial role of the traffic load is striking.

In Chapter 4 of his PhD thesis, Daniels [10] obtains different tail behaviour (of the buffer content distribution) for different traffic loads in a particular multi-server queue. He considers a discrete-time $DBMAP/D/c$ queue with a mix of short-range dependent and long-range dependent traffic. If the mean arrival rate of the short-range dependent background traffic is less than $c-1$, than the tail probabilities decay exponentially; if that mean arrival rate is larger than $c-1$, then they decay according to a power law. In [12], Dumas and Simonian conjecture a similar phenomenon for fluid queues. The present study ,confirms this kind of behaviour for a two-server queue.

For a particular two-server queue with one exponential and one server with regularly varying service time distribution, we are able to *prove* the following: The waiting time tail is *semi-exponential* [2] if the arrival rate λ is less than the service rate μ of the exponential server (so the exponential server would be able to handle all offered traffic on his own); and the waiting time tail is regularly varying of index $1-\nu$ if $\lambda > \mu$.

More precisely, our main asymptotic results are the following. If (1.1) holds, then for $t \to \infty$:
(i) if $\lambda > \mu$:

$$\mathbf{P}(W > t) \sim C_1 t^{1-\nu} L(t), \tag{1.3}$$

C_1 being specified in Theorem 4.1;
(ii) if $\lambda < \mu$:

$$\mathbf{P}(W > t) \sim C_2 t^{1-\nu} \mathrm{e}^{(\lambda-\mu)t}, \tag{1.4}$$

C_2 being specified in Theorem 4.2.

Apart from proving (1.3) and (1.4), we also provide heuristics in both cases that explain and interpret the occurence of each factor. These heuristics might be of independent interest as they suggest ways to generalise the above results and to generate bounds in more complicated systems.

For the moment it suffices to mention a global interpretation of the two different asymptotics. In Case (i) the exponential server is not able to handle all the traffic on its own. The most likely way for a long waiting time to occur is to arrive during a long service time at server 2. Below (4.15) we argue that

$$\mathbf{P}(W > t) \sim C_1^* \mathbf{P}\left(B^{res} > \frac{\lambda t}{\lambda - \mu}\right).$$

In Case (ii) the exponential server is able to handle all the traffic on its own. The most likely way for a long waiting time to occur is to arrive during a long service time at server 2; moreover this service time must have started sufficiently long ago for server 1 to have behaved like a deviant M/M/1 queue. Below (4.33) we argue that, with an obvious notation:

$$\mathbf{P}(W > t) \sim C_2^* \mathbf{P}\left(B^{past} > \frac{\lambda t}{\mu - \lambda}, B^{res} > t\right) \mathbf{P}(W_{M/M/1} > t).$$

The paper is organized as follows. In Section 2 we derive an expression for the steady-state distribution of the number of customers in the system. We also express the waiting time distribution in the former distribution. These expressions still involve an unknown function $Q_1(x)$, that relates to the probability of having one customer in the system, at server 2. In Section 3 we show how $Q_1(x)$ can be determined in case the service time distribution at server 2 has a rational LST (Laplace-Stieltjes Transform). Unfortunately, we are not able to determine $Q_1(x)$ in the general case, and regularly varying distributions do not have a rational LST. However, in Section 4 we show that, in the latter case, the expression for the waiting time LST, that was obtained in Section 2, still is sufficient for determining the tail behaviour of the waiting time distribution. We provide explicit asymptotics for this tail behaviour, distinguishing between $\lambda < \mu$ and $\lambda > \mu$.

2 The steady-state distributions

In this section we derive expressions for the steady-state distribution of the number of customers in the system and for the equilibrium waiting time distribution. We compute the former and then derive an expression for the latter utilizing a distributional form of Little's law.

Before we proceed, we introduce some notation. Let $\beta(s)$ be the LST of the general service time distribution $B(t)$. We also need the LST of the residual service time B^{res}, which is given by

$$\beta_e(s) := \int_0^\infty \mathrm{e}^{-st} \mathrm{d}\mathbf{P}(B^{res} \leq t) = \int_0^\infty e^{-st} \frac{1 - B(t)}{\beta} \mathrm{d}t = \frac{1 - \beta(s)}{\beta s}, \quad \mathrm{Re}\ s \geq 0.$$

We denote by $w(s)$ the LST of the waiting time distribution.

2.1 The number of customers in the system

The goal of this subsection is to compute the generating function of the steady-state number of customers in the system. To accomplish this goal, we use the supplementary variable technique. We refer to Section II.6.2 in [8] for an application of this technique to the M/G/1 queue. We shall consider the process $(X_t, \zeta_t)_{t \geq 0}$, with X_t the number of customers at time t and ζ_t the past service time of the customer in service at the second server. The second server is idle at time t iff $\zeta_t = 0$. It is easy to see that $(X_t, \zeta_t)_{t \geq 0}$ is a Markov process. It will be assumed that the service time distribution is absolutely continuous. Also, we assume that $X_0 = 0$. Define for $t \geq 0$:

$$\begin{aligned} R_{0,t} &= \mathbf{P}(X_t = 0), \\ R_{1,t} &= \mathbf{P}(X_t = 1, \text{server 2 idle}), \\ R_{j,t}(\eta)\mathrm{d}\eta &= \mathbf{P}(X_t = j, \eta \leq \zeta_t < \eta + \mathrm{d}\eta), \qquad \eta > 0, j = 1, 2, \ldots . \end{aligned}$$

As stated in Section 1, it is assumed that the stability condition $\lambda < \mu + \frac{1}{\beta}$ is satisfied.

Denote by X the steady-state number of customers in the system and by ζ the steady-state past service time of the customer in service at server-2. If the system is stable, $R_{0,t}, R_{1,t}$ and $R_{j,t}(\eta)$ converge for $t \to \infty$ to R_0, R_1 and $R_j(\eta)$, which correspond to the distribution of (X, ζ). It can easily be verified that R_0, R_1 and $R_j(\eta)$ satisfy the following differential equations: For $\eta > 0$,

$$\begin{aligned}
\lambda R_0 &= \mu R_1 + \int_0^\infty \frac{R_1(x)}{1-B(x)} \mathrm{d}B(x), \\
(\mu+\lambda) R_1 &= \lambda R_0 + \int_0^\infty \frac{R_2(x)}{1-B(x)} \mathrm{d}B(x), \\
R_1'(\eta) &= -\left(\lambda + \frac{B'(\eta)}{1-B(\eta)}\right) R_1(\eta) + \mu R_2(\eta), \\
R_j'(\eta) &= -\left(\lambda + \frac{B'(\eta)}{1-B(\eta)}\right) R_j(\eta) + \lambda R_{j-1}(\eta) + \mu R_{j+1}(\eta), \qquad j = 2, 3, ..., \\
R_1(0+) &= 0, \\
R_2(0+) &= \int_0^\infty \frac{R_3(x)}{1-B(x)} \mathrm{d}B(x) + \lambda R_1, \\
R_j(0+) &= \int_0^\infty \frac{R_{j+1}(x)}{1-B(x)} \mathrm{d}B(x), \qquad j \geq 3.
\end{aligned}$$

These equations can be derived in the same way as in the ordinary M/G/1 queue, see Section II.6.2 in [8]. We also transform these differential equations in a similar way as in [8]: Define $Q_0 = R_0, Q_1 = R_1$, and for $j \geq 1$, $\eta > 0$,

$$Q_j(\eta) = \frac{R_j(\eta)}{1-B(\eta)}.$$

Q_0, Q_1 and $Q_j(\eta)$ satisfy

$$\begin{aligned}
\lambda Q_0 &= \mu Q_1 + \int_0^\infty Q_1(x) \mathrm{d}B(x), \\
(\mu+\lambda) Q_1 &= \lambda Q_0 + \int_0^\infty Q_2(x) \mathrm{d}B(x), \\
Q_1'(\eta) &= -\lambda Q_1(\eta) + \mu Q_2(\eta), \qquad \eta > 0, \\
Q_j'(\eta) &= -(\lambda+\mu) Q_j(\eta) + \lambda Q_{j-1}(\eta) + \mu Q_{j+1}(\eta), \qquad j \geq 2,\ \eta > 0, \\
Q_1(0+) &= 0, \\
Q_2(0+) &= \int_0^\infty Q_3(x) \mathrm{d}B(x) + \lambda Q_1, \\
Q_j(0+) &= \int_0^\infty Q_{j+1}(x) \mathrm{d}B(x), \qquad j \geq 3.
\end{aligned}$$

Define for $0 \leq p \leq \max(1, \mu/\lambda)$, $\eta \geq 0$,

$$\begin{aligned} G(p,\eta) &:= \sum_{j=1}^{\infty} Q_j(\eta) p^j, && (2.1) \\ f(p) &:= \lambda(1-p) + \mu\left(1 - \frac{1}{p}\right). && (2.2) \end{aligned}$$

If $\mu > \lambda$, it is not difficult to see that $G(p,\eta)$ is well-defined for $1 \leq p \leq \mu/\lambda$ by using similar arguments as in the proof of Theorem 4.2 below. From the last set of differential equations we get

$$\frac{\partial}{\partial \eta} G(p,\eta) = \mu(p-1)Q_1(\eta) - f(p)G(p,\eta), \tag{2.3}$$

which satisfies the following boundary condition,

$$G(p,0+) = \frac{1}{p}\int_0^{\infty} G(p,\eta)\mathrm{d}B(\eta) + \lambda Q_1 p^2 - [(\mu+\lambda)Q_1 - \lambda Q_0]p - (\lambda Q_0 - \mu Q_1). \tag{2.4}$$

The general solution to (2.3) is given by

$$G(p,\eta) = e^{-f(p)\eta}\left[c_1(p) - \mu(1-p)\int_0^{\eta} e^{f(p)x} Q_1(x)\mathrm{d}x\right], \tag{2.5}$$

where $c_1(p)$ is independent of η. It is easy to see that

$$c_1(p) = G(p,0+) = \sum_{j=2}^{\infty} Q_j(0+)p^j. \tag{2.6}$$

We now derive two different expressions for $c_1(p)$ which will be used for the two respective cases in which $f(p) < 0$ and $f(p) \geq 0$. Subsequently we can get expressions for $G(p,\eta)$ which do not contain $c_1(p)$.

If $0 < p < \min(1, \mu/\lambda)$, then $f(p) < 0$ and $G(p,\eta) \leq G(1,\eta)$. From (2.3) we know that $G(1,\eta)$ is a constant which is not related to η (In fact, its interpretation is: $G(1,\eta) = \frac{1}{\beta}\mathbf{P}(X \geq 1, \text{server 2 busy})$; notice that the density of ζ is $\frac{1-B(\eta)}{\beta}$). Multiplying by $e^{f(p)\eta}$ on both sides of (2.5) and taking the limit for $\eta \to \infty$, we obtain

$$\lim_{\eta\to\infty}\left[c_1(p) - \mu(1-p)\int_0^{\eta} e^{f(p)x}Q_1(x)\mathrm{d}x\right] = \lim_{\eta\to\infty}[G(p,\eta)e^{f(p)\eta}] = 0,$$

which implies that, for $0 < p < \min(1, \mu/\lambda)$,

$$c_1(p) = \mu(1-p)\int_0^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x. \tag{2.7}$$

Substitute (2.7) into (2.5) to get

$$G(p,\eta) = \mu(1-p)e^{-f(p)\eta}\int_{x=\eta}^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x. \tag{2.8}$$

If $\min(1, \mu/\lambda) \leq p \leq \max(1, \mu/\lambda)$, then $f(p) \geq 0$. Substituting (2.5) into (2.4), we obtain

$$\begin{aligned} c_1(p) &= \lambda Q_1 p^2 - [(\mu+\lambda)Q_1 - \lambda Q_0]p - (\lambda Q_0 - \mu Q_1) + \frac{c_1(p)}{p}\int_0^{\infty} e^{-f(p)\eta}\mathrm{d}B(\eta) \\ &\quad -\frac{\mu(1-p)}{p}\int_{x=0}^{\infty} e^{f(p)x}Q_1(x)\int_{\eta=x}^{\infty} e^{-f(p)\eta}\mathrm{d}B(\eta)\mathrm{d}x, \end{aligned}$$

which implies that,

$$c_1(p) = \frac{p(1-p)[(\mu - \lambda p)Q_1 - \lambda Q_0] - \mu(1-p)\int_{x=0}^{\infty} e^{f(p)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-f(p)\eta} \mathrm{d}B(\eta)\mathrm{d}x}{p - \beta(f(p))}. \tag{2.9}$$

We may rewrite (2.9) as: For $\min(1, \mu/\lambda) \leq p \leq \max(1, \mu/\lambda)$,

$$c_1(p) = \frac{p[(\lambda p - \mu)Q_1 + \lambda Q_0] + \mu \int_{x=0}^{\infty} e^{f(p)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-f(p)\eta} \mathrm{d}B(\eta)\mathrm{d}x}{1 - \frac{(\lambda p - \mu)\beta}{p}\beta_e(f(p))}. \tag{2.10}$$

We are now ready to calculate the generating function $X(p) := \mathrm{E}[p^X]$ of the steady-state number of customers in the system. We have

$$\begin{aligned} X(p) &= \sum_{j=0}^{\infty} p^j \mathbf{P}(X = j) = R_0 + R_1 p + \sum_{j=1}^{\infty} p^j \int_0^{\infty} R_j(\eta)\mathrm{d}\eta \\ &= Q_0 + Q_1 p + \int_0^{\infty} G(p,\eta)(1 - B(\eta))\mathrm{d}\eta. \end{aligned} \tag{2.11}$$

From (2.11) we can derive that $G(1,\eta) = (1 - Q_0 - Q_1)/\beta$. For the sake of simplicity, put

$$\tilde{Q}_1 := \mathbf{P}(X = 1, \text{server 1 idle}) = \int_0^{\infty} Q_1(\eta)(1 - B(\eta))\mathrm{d}\eta. \tag{2.12}$$

Taking $p = 1$ in (2.10) and noting that $c_1(1) = G(1,\eta) = (1 - Q_0 - Q_1)/\beta$, we get the following equation:

$$\frac{1}{\beta} + \mu - \lambda = \frac{1}{\beta}Q_0 + \mu Q_0 + \frac{1}{\beta}Q_1 + \mu\tilde{Q}_1. \tag{2.13}$$

If $0 < p < \min(1, \mu/\lambda)$, substitute (2.8) into (2.11) to get

$$\begin{aligned} X(p) &= Q_0 + Q_1 p + \mu(1-p)\int_{\eta=0}^{\infty}(1 - B(\eta))e^{-f(p)\eta}\int_{x=\eta}^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x\mathrm{d}\eta \\ &= Q_0 + Q_1 p + \mu(1-p)\int_{\eta=0}^{\infty}\int_{t=\eta}^{\infty}\int_{x=\eta}^{\infty} e^{f(p)(x-\eta)}Q_1(x)\mathrm{d}x\mathrm{d}B(t)\mathrm{d}\eta \\ &= Q_0 + Q_1 p + \mu(1-p)\left[\int_{t=0}^{\infty}\int_{x=0}^{t}\int_{\eta=0}^{x} e^{f(p)(x-\eta)}Q_1(x)\mathrm{d}\eta\mathrm{d}x\mathrm{d}B(t)\right. \\ &\quad \left. + \int_{t=0}^{\infty}\int_{x=t}^{\infty}\int_{\eta=0}^{t} e^{f(p)(x-\eta)}Q_1(x)\mathrm{d}\eta\mathrm{d}x\mathrm{d}B(t)\right] \\ &= Q_0 + Q_1 p - \frac{\mu(1-p)}{f(p)}\left[\int_0^{\infty} Q_1(x)(1 - B(x))\mathrm{d}x\right. \\ &\quad \left. - \int_0^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x + \int_{x=0}^{\infty} e^{f(p)x}Q_1(x)\int_{\eta=0}^{x} e^{-f(p)\eta}\mathrm{d}B(\eta)\mathrm{d}x\right], \end{aligned}$$

which in combination with (2.12) and (2.16) below leads to

$$X(p) = Q_0 + Q_1 p - \frac{p}{\lambda p - \mu}\left[\mu\tilde{Q}_1 - \lambda Q_0 p + (\mu - \lambda p)pQ_1 - \mu(1-p)\int_{x=0}^{\infty} e^{f(p)x}Q_1(x)\mathrm{d}x\right]. \tag{2.14}$$

If $\min(1,\mu/\lambda) \leq p \leq \max(1,\mu/\lambda)$, substitute (2.5) and (2.10) into (2.11) to get

$$\begin{aligned} X(p) &= Q_0 + Q_1 p + c_1(p)\frac{1-\beta(f(p))}{f(p)} - \mu(1-p)\int_0^\infty e^{-f(p)\eta}(1-B(\eta))\int_{x=0}^{\eta} e^{f(p)x}Q_1(x)\mathrm{d}x\mathrm{d}\eta \\ &= Q_0 + Q_1 p + c_1(p)\frac{1-\beta(f(p))}{f(p)} - \frac{\mu(1-p)}{f(p)}\left[\int_0^\infty Q_1(x)(1-B(x))\mathrm{d}x \right. \\ &\quad \left. - \int_0^\infty e^{f(p)x}Q_1(x)\int_{\eta=x}^\infty e^{-f(p)\eta}\mathrm{d}B(\eta)\mathrm{d}x\right] \\ &= Q_0 + Q_1 p + \frac{p^2\beta[(\lambda p-\mu)Q_1+\lambda Q_0]\beta_e(f(p))}{p-(\lambda p-\mu)\beta\beta_e(f(p))} \\ &\quad + \frac{\mu p}{\lambda p-\mu}\left[\frac{p}{p-(\lambda p-\mu)\beta\beta_e(f(p))}\int_{x=0}^\infty e^{f(p)x}Q_1(x)\int_{\eta=x}^\infty e^{-f(p)\eta}\mathrm{d}B(\eta)\mathrm{d}x - \tilde{Q}_1\right]. \end{aligned} \tag{2.15}$$

As we can see, the expressions (2.14) and (2.15) for $X(p)$ contain an unknown function $Q_1(x)$. Replacing $G(p,\eta)$ in (2.4) by (2.8), we derive an equation for $Q_1(x)$ which is given by: For $0 < p < \min(1,\mu/\lambda)$,

$$\mu\int_0^\infty e^{f(p)x}Q_1(x)\mathrm{d}x = \frac{\mu}{p}\int_{x=0}^\infty e^{f(p)x}Q_1(x)\int_{\eta=0}^{x} e^{-f(p)\eta}\mathrm{d}B(\eta)\mathrm{d}x + (\mu-\lambda p)Q_1 - \lambda Q_0. \tag{2.16}$$

Unfortunately, we are not able to obtain $Q_1(x)$, and hence $X(p)$, for general service time distribution. In case $B(\cdot)$ has a rational LST, we can determine $X(p)$ completely; this is done in Section 3. In Section 4, in the case of regularly varying $B(\cdot)$, we perform an asymptotic analysis of the waiting time distribution. In the next subsection we establish a link between the waiting time distribution and the queue length generating function $X(p)$.

2.2 The waiting time distribution

In order to get an explicit formula for the LST $w(s)$ of the waiting time distribution, we introduce the queue length N which is the number of customers who are waiting in the system, and its probability generating function $N(p) := \mathrm{E}[p^N]$. By the distributional form of Little's law (cf. [13]), $w(s)$ is related to $N(p)$ as follows:

$$w(s) = N(1-s/\lambda), \qquad 0 \leq s \leq \lambda. \tag{2.17}$$

Since $N = \max(X-2,0)$, it follows that

$$N(p) = Q_0 + Q_1 + \tilde{Q}_1 + \frac{1}{p^2}\left[X(p) - Q_0 - Q_1 p - \tilde{Q}_1 p\right],$$

which in combination with (2.17) implies

$$w(s) = Q_0 + Q_1 + \tilde{Q}_1 + \frac{1}{(1-s/\lambda)^2}[X(1-s/\lambda) - Q_0 - (1-s/\lambda)Q_1 - (1-s/\lambda)\tilde{Q}_1]. \tag{2.18}$$

For the sake of simplicity, put

$$\hat{f}(s) := f(1-s/\lambda) = \frac{s(\lambda-\mu-s)}{\lambda-s}. \tag{2.19}$$

If $\max(0, \lambda - \mu) < s < \lambda$, then $0 < 1 - s/\lambda < \min(1, \mu/\lambda)$. So we can substitute (2.14) into (2.18) to get

$$w(s) = Q_0 + 2Q_1 + \tilde{Q}_1 + \frac{\lambda}{\mu - \lambda - s}(\tilde{Q}_1 - Q_0) - \frac{\mu s}{(\lambda - s)(\mu - \lambda - s)} \int_0^\infty e^{\hat{f}(s)x} Q_1(x) \mathrm{d}x. \quad (2.20)$$

If $\min(0, \lambda - \mu) \leq s \leq \max(0, \lambda - \mu)$, then $\min(1, \mu/\lambda) \leq 1 - s/\lambda \leq \max(1, \mu/\lambda)$. Substitute (2.15) into (2.18) to get

$$\begin{aligned} w(s) \;=\; & Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s}\tilde{Q}_1 + \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))} \\ & + \frac{\mu}{(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]} \int_{x=0}^\infty e^{\hat{f}(s)x} Q_1(x) \int_{\eta = x}^\infty e^{-\hat{f}(s)\eta} \mathrm{d}B(\eta) \mathrm{d}x. \end{aligned} \quad (2.21)$$

The above formulas are quite useful in deriving the asymptotic behaviour of the waiting time distribution in the case that $B(t)$ has a regularly varying tail, although they contain an unknown function $Q_1(x)$.

3 Rational service time distribution

In this section we assume that the service time distribution has a rational LST, i.e., $\beta(s) = \frac{\beta_1(s)}{\beta_2(s)}$ where $\beta_1(s)$ and $\beta_2(s)$ are relatively prime polynomials, the degree of $\beta_2(s)$ being higher than that of $\beta_1(s)$. Without loss of generality we can write

$$\beta_2(s) \;=\; \prod_{j=1}^{n} (s - s_j)^{m_j},$$

where $s_1, \ldots, s_n$ are different from each other, $m_j \in \{1, 2, \ldots\}$ and Re $s_j < 0$, $j = 1, \ldots, n$ (because $\beta(s)$ is analytic for Re $s \geq 0$). We outline how, in this case of rational service time LST, one can obtain the Laplace transform, $q_1(s) := \int_0^\infty e^{-sx} Q_1(x) \mathrm{d}x$, of $Q_1(x)$.

Step 1: Obtaining an expression for $q_1(-f(p))$.
Replace $Q_1(x)$ in (2.16) by the following representation of the inverse of $q_1(s)$:

$$Q_1(x) \;=\; \frac{1}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{e}^{sx} q_1(s) \mathrm{d}s.$$

Formula (2.16) then becomes, after some interchange of integrals and division by μ: For $0 < p < \min(1, \mu/\lambda)$,

$$q_1(-f(p)) \;=\; -\frac{1}{p}\frac{1}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{q_1(s)}{f(p) + s} \beta(-s) \mathrm{d}s \;+\; (1 - \frac{\lambda}{\mu}p)Q_1 - \frac{\lambda}{\mu}Q_0. \quad (3.1)$$

The integral in the righthand side can be handled by observing that all the poles of its integrand are located in the righthalf plane: $-s_1, \ldots, -s_n$, and also $-f(p) > 0$ since $0 < p < \min(1, \mu/\lambda)$. Consider the semi-circle with center in the origin and radius R in the righthalf plane. Choose R so large that all $n + 1$ above-mentioned poles are inside the semi-circle. Then the integral along the line segment from $-\mathrm{i}R$ to $+\mathrm{i}R$ and then along the semi-circle back to $-\mathrm{i}R$ equals minus the sum of the residues of those poles. Since the integral along the semi-circle disappears when $R \to \infty$ (remember that the degree of $\beta_2(s)$ is larger than that of $\beta_1(s)$), we have:

$$\frac{1}{2\pi \mathrm{i}} \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \frac{q_1(s)}{f(p)+s} \beta(-s) \mathrm{d}s \;=\; -q_1(-f(p))\beta(f(p)) \;+\; \sum_{j=1}^{n} \frac{(-1)^{m_j}}{(m_j-1)!} \frac{\mathrm{d}^{m_j-1}}{\mathrm{d}a^{m_j-1}} \{ \frac{q_1(a)}{f(p)+a} \frac{\beta_1(-a)}{\prod_{i \neq j}(-a-s_i)^{m_i}} \}|_{a=-s_j}. \tag{3.2}$$

Formula (3.1) thus reduces to: For $0 < p < \min(1, \mu/\lambda)$,

$$q_1(-f(p)) \;=\; \frac{\frac{1}{p}\sum_{j=1}^{n} \frac{(-1)^{m_j-1}}{(m_j-1)!} \frac{\mathrm{d}^{m_j-1}}{\mathrm{d}a^{m_j-1}} \{ \frac{q_1(a)}{f(p)+a} \frac{\beta_1(-a)}{\prod_{i \neq j}(-a-s_i)^{m_i}} \}|_{a=-s_j} + (1-\frac{\lambda}{\mu}p)Q_1 - \frac{\lambda}{\mu}Q_0}{1 - \frac{\beta(f(p))}{p}}. \tag{3.3}$$

The numerator of (3.3) contains $\sum_{j=1}^{n} m_j + 2$ unknown constants: Q_0, Q_1, and the $\sum_{j=1}^{n} m_j$ terms relating to the ith derivative of $q_1(s)$ at $s = -s_j$, $i = 0, \ldots, m_j - 1$, $j = 1, \ldots, n$.

Step 2: Determining the unknown constants.
Noting that $q_1(-f(p)) = \frac{c_1(p)}{\mu(1-p)} = \frac{G(p,0+)}{\mu(1-p)}$ (cf. (2.6) and (2.7)), it follows by analytic continuation that the righthand side of (3.3) is analytic inside the unit circle. Let us consider the poles of the denominator of (3.3). Multiply numerator and denominator of the righthand side of (3.3) by $p^{\sum_{i=1}^{n} m_i} \beta_2(f(p))$. We shall prove that

$$p^{\sum_{i=1}^{n} m_i} \beta_2(f(p))(1 - \frac{\beta(f(p))}{p}) \tag{3.4}$$

has $\sum_{i=1}^{n} m_i - 1$ different roots inside the unit circle. Remember that $f(p) = \lambda(1-p) + \mu(1-1/p)$. One can easily check that $\mathrm{Re}\, f(p) \geq f(|p|)$ for $1 \leq |p| \leq 1+\epsilon$, and hence $|\beta(f(p))| \leq |\beta(\mathrm{Re}\, f(p))| \leq |\beta(f(|p|))|$. For $|p| = 1 + \epsilon$ we then have:

$$|p| > |\beta(f(|p|))| \geq |\beta(f(p))|,$$

the inequality sign holding because $\lambda < \mu + 1/\beta$ (the ergodicity condition).
The resulting inequality $|p| > |\beta(f(p))|$ is equivalent with: For $|p| = 1 + \epsilon$,

$$|p^{\sum_{i=1}^{n} m_i} \beta_2(f(p))| > |p^{\sum_{i=1}^{n} m_i - 1} \beta_1(f(p))|. \tag{3.5}$$

Notice that the multiplication by suitable powers of p has led to functions that are analytic inside $|p| = 1 + \epsilon$. Application of Rouché's theorem now implies that $p^{\sum_{i=1}^{n} m_i} \beta_2(f(p))$ $(1 - \frac{\beta(f(p))}{p})$ has the same number of zeros as $p^{\sum_{i=1}^{n} m_i} \beta_2(f(p))$ inside the circle $|p| = 1 + \epsilon$.
We shall prove that the latter number of zeros is $\sum_{i=1}^{n} m_i$; observe that there is one zero $p = 1$, but we need the $\sum_{i=1}^{n} m_i - 1$ zeros *inside* the unit circle. Consider

$$p^{\sum_{i=1}^{n} m_i} \beta_2(f(p)) \;=\; \prod_{j=1}^{n} (\lambda p(1-p) + \mu(p-1) - ps_j)^{m_j}. \tag{3.6}$$

Each factor $\lambda p(1-p) + \mu(p-1) - ps_j$ has exactly one zero inside the unit circle, and one zero outside it. This can be seen, e.g., by yet another application of Rouché's theorem. In fact, comparison with the expression for the LST of the busy period P of an $M/M/1$ queue with

arrival rate λ and service rate μ reveals that the zero inside the unit circle is $\mathrm{E}[\mathrm{e}^{s_j P}]$.
The analyticity of $q_1(-f(p))$ inside the unit circle implies that the $\sum_{j=1}^{n} m_j - 1$ zeros of the denominator of (3.3) inside the unit circle should also be zeros of the numerator. This yields $\sum m_i - 1$ linear equations. As remarked at the end of Step 1, we have $\sum m_i + 2$ unknowns. Two additional equations result from the fact that $p = 0$ is a double root of the righthand side of (3.3) which follows from the observation that (see (2.6)) $c_1(0) = 0$ and $c_1'(0) = 0$. Noticing that $\tilde{Q}_1$ can be represented by a linear combination of the $\sum m_i$ terms which relate to the ith derivative of $q_1(s)$ at $s = -s_j$, $i = 0, \ldots, m_j - 1$, $j = 1, \ldots, n$, the final equation is provided by (2.13). Solution of the resulting $\sum m_i + 2$ linear equations yields the $\sum m_i + 2$ unknowns, and finally $q_1(-f(p))$. Once $q_1(-f(p))$ and hence $c_1(p) = \mu(1-p)q_1(-f(p))$ have been obtained, the generating function $X(p)$ of the number of customers follows from (2.14) and (2.15) and the waiting time LST follows from Subsection 2.2.

The Erlang-n and hyperexponential distributions are examples of distributions with rational LST. In the special case that $B(t) = 1 - \mathrm{e}^{-t/\beta}$, (3.4) reduces to $(p-1)(\mu + 1/\beta - \lambda p)$, which does not have a zero in $|p| < 1$. The numerator of the righthand side of (3.3) reduces to $-\frac{1/\beta}{pf(p)+p/\beta} + (1 - \frac{\lambda}{\mu}p)Q_1 - \frac{\lambda}{\mu}Q_0$. Noting that $p = 0$ is a double root of this function, it follows that

$$\frac{\tilde{Q}_1}{\mu\beta} + Q_1 - \frac{\lambda}{\mu}Q_0 = 0, \tag{3.7}$$

$$\frac{(\lambda + \mu + 1/\beta)\tilde{Q}_1}{\mu\beta} - \lambda Q_1 = 0. \tag{3.8}$$

Combining the above two equations and (2.13) leads to

$$\begin{aligned}
Q_1 &= \frac{\lambda Q_0(\lambda + \mu + 1/\beta)}{\mu(2\lambda + \mu + 1/\beta)}, \\
\tilde{Q}_1 &= \frac{\lambda^2 \beta Q_0}{2\lambda + \mu + 1/\beta}, \\
Q_0 &= \frac{\mu(2\lambda + \mu + 1/\beta)(1 - \frac{\lambda}{\mu + 1/\beta})}{\lambda\beta[\lambda(\mu + 1/\beta) + 1/\beta^2 + \mu/\beta] + \mu(1 - \frac{\lambda}{\mu+1/\beta})(2\lambda + \mu + 1/\beta)}.
\end{aligned} \tag{3.9}$$

Remark 3.1. In [16] Knessl et al. also study the $M/G/2$ queue with heterogeneous servers. Using a supplementary variable approach, they derive integral equations for the joint steady-state distribution of numbers of customers and elapsed service times. They construct the solution of these equations for several mixtures of exponential, Erlang and hyperexponential service time distributions.

4 Main asymptotic results

In this section we shall present our main results: the asymptotic behaviour of the waiting time distribution under the assumption that $B(t)$ has a regularly varying tail. First of all, it is useful to recall that Abate, Choudhury and Whitt [1] and Abate and Whitt [2] divide probability distributions on the positive halfline into three classes according to the rightmost singularity of the LST and the value of the LST at this singularity. Let $G(t)$ be a probability distribution function with LST $\zeta(s)$ and let $-s^*$ be the rightmost singularity of $\zeta(s)$, with $-s^* = -\infty$ if $\zeta(s)$

is analytic everywhere. In this setting $G(t)$ and its LST $\zeta(s)$ are classified as follows:

$$\begin{array}{ll} \text{class I}: & s^* > 0 \text{ and } \zeta(-s^*) = \infty, \\ \text{class II}: & s^* > 0 \text{ and } 1 < \zeta(-s^*) < \infty, \\ \text{class III}: & s^* = 0 \text{ and } \zeta(-s^*) = 1. \end{array} \tag{4.1}$$

As indicated in [1, 2], class-I distributions are the 'well behaved' distributions and class-III distributions are the long-tailed distributions. Class-II distributions are called *semi-exponential distributions*, because they are dominated by an exponential, i.e., $\lim_{t\to\infty} e^{\gamma t}(1 - G(t)) = 0$ for all real $\gamma < s^*$. Since $\zeta(-s^*) < \infty$, the rightmost singularity $-s^*$ is necessarily a branch point singularity, not a pole (note that $-s^*$ can still be a branch point when G is in class I). For more discussions, see [1, 2].

The results in this section show that, when $1 - B(t)$ is regularly varying, the waiting time distribution $W(t)$ can be in all classes considered above. In particular, we show that

- $W(t)$ belongs to class I if $\lambda < \mu$ and $\beta_2 = \infty$,
- $W(t)$ belongs to class II if $\lambda < \mu$ and $\beta_2 < \infty$,
- $W(t)$ belongs to class III if $\lambda > \mu$.

In the next two subsections, we consider the cases $\lambda > \mu$ and $\lambda < \mu$. In both cases, we analyse the tail $1 - W(t)$ of the waiting time distribution, using the expression for $w(s)$ developed in Section 2 and an appropriate Tauberian theorem. In the case $\lambda > \mu$ we hase $s^* = 0$ and apply Theorem 8.1.6 of [4]. The case $\lambda < \mu$ is more intricate and here we apply a theorem of Sutton (cf. [21]).

4.1 The regime $\lambda > \mu$

When $\lambda > \mu$, the exponential server alone cannot cope with all the traffic: The second, ill-behaved, server is necessary for stability of the system. This makes it plausible that the heavy-tailed service times at the second server give rise to a heavy-tailed waiting time. In fact, we have

Theorem 4.1 *Suppose that $\lambda > \mu$ and*

$$1 - B(t) \sim t^{-\nu} L(t), \qquad t \to \infty. \tag{4.2}$$

$\nu \in (m, m+1)$ ($m \in \mathbf{N}$) and $L(t)$ is a slowly varying function. Then

$$1 - W(t) \sim \frac{1 - Q_0 - Q_1}{(\nu - 1)(1 - \lambda\beta + \mu\beta)\beta} \left(\frac{\lambda - \mu}{\lambda}\right)^{\nu-1} t^{1-\nu} L(t), \qquad t \to \infty. \tag{4.3}$$

Proof. By using Theorem 8.1.6 in [4], we have

$$\beta(s) = 1 + \sum_{i=1}^{m} (-1)^i b_i s^i + (-1)^{m+1}\Gamma(1 - \nu) s^\nu L(1/s) + \mathrm{o}(s^\nu L(1/s)), \qquad s \downarrow 0, \tag{4.4}$$

where $b_i > 0$ for $i = 1, ..., m$ and $\Gamma(\cdot)$ is the Gamma function. Again, by using Theorem 8.1.6 in [4], it is sufficient to prove that $w(s)$ can be written as

$$\begin{aligned} w(s) \;\; = \;\; & 1 + \sum_{i=1}^{m-1} (-1)^i d_i s^i + (-1)^m \frac{\Gamma(1 - \nu)(1 - Q_0 - Q_1)}{(1 - \lambda\beta + \mu\beta)\beta} \left(\frac{\lambda - \mu}{\lambda}\right)^{\nu-1} s^{\nu-1} L(1/s) \\ & + \mathrm{o}(s^{\nu-1} L(1/s)), \qquad s \downarrow 0, \end{aligned} \tag{4.5}$$

where $d_i > 0$ for $i = 1, ..., m-1$. For $0 \leq s \leq \lambda - \mu$, $w(s)$ is given by (2.21). The expression (2.21) includes the function $\hat{q}_1(s) := \int_{x=0}^{\infty} e^{\hat{f}(s)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-\hat{f}(s)\eta} \mathrm{d}B(\eta)\mathrm{d}x$ of which the asymptotic expansion in the neighbourhood of the origin is not known. From (2.10), we observe that $\hat{q}_1(s)$ for $0 < s < \lambda - \mu$ can be expressed in terms of $c_1(\frac{\mu - \hat{f}(s) + s}{\lambda})$ and $\beta(\hat{f}(s))$. We will analyse the behaviour of the latter functions in the origin. Taking $p = 1 - s/\lambda$ in (2.10), we obtain

$$c_1(1 - s/\lambda) = \frac{(1 - s/\lambda)[(\lambda - \mu - s)Q_1 + \lambda Q_0] + \mu \hat{q}_1(s)}{1 - \frac{(\lambda - \mu - s)\beta}{1 - s/\lambda} \beta_e(\hat{f}(s))}. \tag{4.6}$$

For $s \downarrow 0$, there exists $s_1 = s_1(s) \uparrow \lambda - \mu$ such that $\hat{f}(s) = \hat{f}(s_1)$ where $\hat{f}(s)$ is given by (2.19). It is not difficult to see that $s_1 = \lambda - \mu + \hat{f}(s) - s$. Using (4.6), we may write, for $0 \leq s \leq \lambda - \mu$,

$$\begin{aligned} \hat{q}_1(s) &= \frac{1}{\mu} c_1(1 - s_1/\lambda) \left[1 - \frac{(\lambda - \mu - s_1)\beta}{1 - s_1/\lambda} \beta_e(\hat{f}(s))\right] - \frac{1}{\mu}(1 - s_1/\lambda)[(\lambda - \mu - s_1)Q_1 + \lambda Q_0] \\ &= \frac{1}{\mu} c_1(\frac{\mu - \hat{f}(s) + s}{\lambda}) \left[1 - \frac{\beta\lambda(s - \hat{f}(s))\beta_e(\hat{f}(s))}{\mu + s - \hat{f}(s)}\right] - \frac{\mu - \hat{f}(s) + s}{\lambda\mu}[(s - \hat{f}(s))Q_1 + \lambda Q_0]. \end{aligned} \tag{4.7}$$

Then, replacing $\hat{q}_1(s)$ in the last term of (2.21) by the expression in (4.7) gives, for $0 \leq s \leq \lambda - \mu$,

$$\begin{aligned} w(s) &= Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s}\tilde{Q}_1 + \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))} \\ &\quad + \frac{c_1(\frac{\mu - \hat{f}(s) + s}{\lambda})\left(1 - \frac{\beta\lambda(s - \hat{f}(s))\beta_e(\hat{f}(s))}{\mu + s - \hat{f}(s)}\right) - \frac{1}{\lambda}(\mu - \hat{f}(s) + s)[(s - \hat{f}(s))Q_1 + \lambda Q_0]}{(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]} \\ &= Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s}\tilde{Q}_1 + \frac{1}{(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]} \\ &\quad \left[c_1(\frac{\mu - \hat{f}(s) + s}{\lambda}) - \frac{1}{\lambda}(\mu - \hat{f}(s) + s)[(s - \hat{f}(s))Q_1 + \lambda Q_0]\right. \\ &\quad \left. + \left((\lambda - \mu - s)[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta - \frac{\beta\lambda(s - \hat{f}(s))}{\mu - \hat{f}(s) + s} c_1(\frac{\mu - \hat{f}(s) + s}{\lambda})\right)\beta_e(\hat{f}(s))\right]. \end{aligned} \tag{4.8}$$

Letting $s = 0$ in (4.7), the lefthand side is equal to $\tilde{Q}_1$. Therefore, we get

$$c_1(\frac{\mu}{\lambda}) = \mu(\tilde{Q}_1 + Q_0). \tag{4.9}$$

Since $c_1(p)$ is well-defined for $|p| \leq 1$, the Taylor expansion of $c_1(p)$ in the neighbourhood of μ/λ exists which is given as follows:

$$c_1(\frac{\mu + s}{\lambda}) = \mu(\tilde{Q}_1 + Q_0) + \sum_{i=1}^{\infty} c_{\mu/\lambda,i} s^i, \qquad |s| < \lambda - \mu, \tag{4.10}$$

where $c_{\mu/\lambda,i}$ are constants for $i = 1, 2, ...$ Because $\hat{f}(s)$ also has a Taylor expansion in the neighbourhood of the origin and $\hat{f}(0) = 0$, we may write

$$c_1(\frac{\mu + s - \hat{f}(s)}{\lambda}) = \mu(\tilde{Q}_1 + Q_0) + \sum_{i=1}^{\infty} \tilde{c}_{\mu/\lambda,i} s^i, \qquad \text{for } |s| < \delta, \tag{4.11}$$

where δ is some positive constant and $\tilde{c}_{\mu/\lambda,i}$ are all constants for $i = 1, 2, \ldots$. By (4.4), we have for $s \downarrow 0$:

$$\beta_e(s) = 1 + \sum_{i=1}^{m-1} (-1)^i \frac{b_{i+1}}{\beta} s^i + (-1)^m \frac{\Gamma(1-\nu)}{\beta} s^{\nu-1} L(1/s) + o(s^{\nu-1} L(1/s)). \tag{4.12}$$

From (4.12) we get

$$\begin{aligned} \frac{1}{(\lambda-\mu-s)[1-s/\lambda-(\lambda-\mu-s)\beta\beta_e(\hat{f}(s))]} &= \frac{1}{(\lambda-\mu)[1-(\lambda-\mu)\beta]} + g_1(s)s \\ &+(-1)^m \frac{\Gamma(1-\nu)}{(1+\mu\beta-\lambda\beta)^2} \left(\frac{\lambda-\mu}{\lambda}\right)^{\nu-1} s^{\nu-1} L(1/s) + o(s^{\nu-1}L(1/s)), \; s \downarrow 0, \end{aligned} \tag{4.13}$$

and

$$\begin{aligned} & \Bigg[c_1\left(\frac{\mu-\hat{f}(s)+s}{\lambda}\right) - \frac{1}{\lambda}(\mu-\hat{f}(s)+s)[(s-\hat{f}(s))Q_1+\lambda Q_0] \\ & + \left((\lambda-\mu-s)[(\lambda-\mu-s)Q_1+\lambda Q_0]\beta - \frac{\beta\lambda(s-\hat{f}(s))}{\mu-\hat{f}(s)+s} c_1\left(\frac{\mu-\hat{f}(s)+s}{\lambda}\right) \right) \beta_e(\hat{f}(s)) \Bigg] \\ = \; & \mu\tilde{Q}_1 + (\lambda-\mu)(\lambda Q_1 - \mu Q_1 + \lambda Q_0)\beta + (-1)^m \Gamma(1-\nu)(\lambda-\mu) \\ & (\lambda Q_1 - \mu Q_1 + \lambda Q_0) \left(\frac{\lambda-\mu}{\lambda}\right)^{\nu-1} s^{\nu-1} L(1/s) + g_2(s)s + o(s^{\nu-1}L(1/s)), \quad s \downarrow 0, \end{aligned} \tag{4.14}$$

where $g_i(s)$ $(i = 1, 2)$ are polynomials of degree $m-1$. In combination with (4.8), (4.13), (4.14) and (4.11), this leads to (4.5). □

An alternative characterisation for the tail of $W(t)$ is

$$\mathbf{P}(W > t) \sim \frac{1-Q_0-Q_1}{1-\lambda\beta+\mu\beta} \mathbf{P}\left(B^{res} > \frac{\lambda t}{\lambda-\mu}\right), \tag{4.15}$$

when $t \to \infty$.

It is possible to give a heuristic explanation of (4.15) by identifying a possible way (which we claim is the most probable way) for W to become large. The heuristics given in the remainder of this subsection are very similar to those in [18] for a fluid queue with $M/G/\infty$ input.

First, we make two preliminary observations:

1. The long-term fraction of customers served by server 2 equals $\frac{1-Q_0-Q_1}{\lambda\beta}$ (note that the mean number of customers handled by server 2 per time unit equals $\frac{1-Q_0-Q_1}{\beta}$).

2. If both servers are busy (i.e. if the waiting time is larger than zero), the fraction of customers that goes to server 1 equals $\frac{\mu}{\mu+\beta^{-1}} = \frac{\beta\mu}{1+\beta\mu}$. Hence, the workload then decreases at rate
$$\frac{\lambda}{\mu}\frac{\beta\mu}{1+\beta\mu} + \lambda\beta\frac{1}{1+\beta\mu} - 2.$$

We now turn to the heuristic explanation of (4.15). Suppose a customer enters the system in steady state at time τ (say) and is served by server 2. This happens with probability $\frac{1-Q_0-Q_1}{\lambda\beta}$ (due to PASTA and observation 1). Let the service time of this customer be equal to B. Assume that the total workload in the system is very small compared to B. Then, the workload at the second server is roughly equal to B and the workload at server 1 is approximately zero. This means that all incoming traffic will be allocated to server 1, implying that the workload at server 1 will increase linearly with drift $\rho - 1 = \lambda/\mu - 1$. As no work is allocated to the second server, the workload of server 2 decreases with drift -1. This carries on until both workloads are the same, which happens at time $\tau + B/\rho$, see Figure 1.

After time $\tau + B/\rho$, the waiting time decreases at rate $1 - \frac{\lambda}{\mu+\beta^{-1}}$, by observation 2. Hence, at time $\tau + \frac{B}{(\mu-\lambda)\beta+1}$ the effect of the large customer entering the system at time 0 has expired, see again Figure 1.

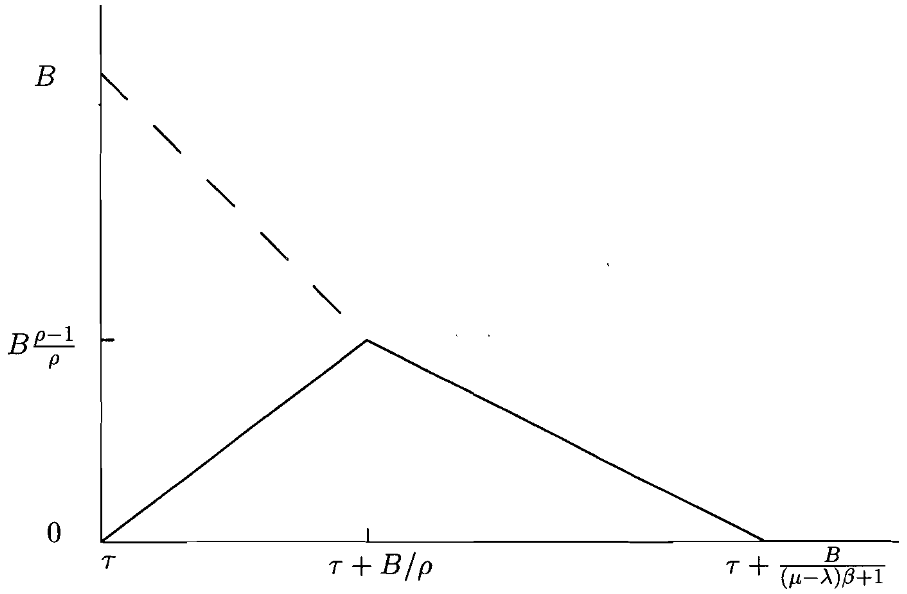

Figure 1: Evolution of the waiting time.

Suppose that we observe the system at time 0 and that $W > t$. Our claim is that the waiting time is large because at time $\tau = -y$, a customer entered the system which was being served by server 2. This customer had a large service time B. Keeping Figure 1 in mind, there are two possibilities:

1. $y < B/\rho$. In this case, we are still in the first part of the excursion described in Figure 1 (where all incoming customers are served by the first server). To get $W > t$, we need $y > t/(\rho - 1)$.

2. $y > B/\rho$. Using observation 2 (to get the drift after time B/ρ in Figure 1) we obtain the condition

$$t \;<\; B\frac{\rho-1}{\rho} - \frac{1+(\mu-\lambda)\beta}{1+\mu\beta}\left(y - \frac{B}{\rho}\right)$$

$$= \frac{B}{1+\mu\beta} - \frac{1+(\mu-\lambda)\beta}{1+\mu\beta}y.$$

Together with the condition $y > B/\rho$, this can be rewritten into

$$B > (1+\mu\beta)t + (1+(\mu-\lambda)\beta)y, \qquad y > \frac{t}{1-\rho}.$$

To summarise, the event $W > t$ occurs if at time $y > t/(\rho - 1)$ a customer enters the system which is served by server 2 and which has a service time $B > (1+\mu\beta)t + (1+(\mu-\lambda)\beta)y$. By observation 1, the probability that the customer is served by server 2 equals $\frac{1-Q_0-Q_1}{\lambda\beta}$. We conclude that, after a straightforward computation,

$$\begin{aligned} \mathbf{P}(W > t) &\approx \int_{\frac{t}{\rho-1}}^{\infty} \frac{1-Q_0-Q_1}{\lambda\beta}\mathbf{P}(B > (1+\mu\beta)t + (1+(\mu-\lambda)\beta)y)\lambda \mathrm{d}y \\ &= \frac{1-Q_0-Q_1}{1+(\mu-\lambda)\beta}\frac{1}{\beta}\int_{\frac{\rho t}{\rho-1}}^{\infty} \mathbf{P}(B > z)\mathrm{d}z, \end{aligned}$$

which is equal to (4.15).

The heuristics given above can easily be formalized to prove an *asymptotic lower bound* for the waiting time distribution. Furthermore, the heuristic arguments do not depend on the service time distribution of server 1 and can thus be extended to more general multi-server queues

In addition to the heuristics given above one may try to find a different way of explaining (4.15), namely by relating the M/G/2 queue to an M/G/1 queue with arrival rate $\mu - \lambda$ and service time B. This might lead to some type of *reduced load* equivalence, as is often the case in fluid queues under the FIFO [3] or GPS [5] discipline: Those studies show that under certain conditions, in the asymptotic analysis one can ignore exponentially tailed sources, apart from reducing the outflow rate by the load offered by those exponential sources.

4.2 The regime $\lambda < \mu$

Now we turn to the case $\lambda < \mu$. This case is more intricate. If one wants to apply a similar technique as in the proof of Theorem 4.1, one needs to consider the function $\mathrm{e}^{s^* t}(1-W(t))$ (which has LST $w(s - s^*)$). However, this function need not be monotone, so a standard Tauberian theorem does not work.

Instead, we shall use a Theorem of Sutton [21]. This theorem does not need a monotonicity assumption, but requires other regularity conditions. In order to meet these conditions, we make the following assumptions on the general service time distribution $B(t)$; these assumptions are similar to the ones in Section 2 of [6], see also [9]. It is assumed that $\beta(s)$ can be represented as: for Re $s \geq 0$,

$$1 - \frac{1-\beta(s)}{\beta s} = h(s) + s^{\nu-1}l(s), \tag{4.16}$$

where

(i) $m < \nu < m+1$ $(m \in \mathbf{N})$;

(ii) $h(s)$ is analytic in s for Re $s > -\epsilon_0$ $(\epsilon_0 > 0)$, $h(0) = 0$;

(iii) $l(s)$ is analytic in $s \in \{s :$ Re s > 0, or $|s| < \epsilon_0\}$ $(\epsilon_0 > 0)$ and continuous for Re $s \geq 0$, $l(0) \neq 0$.

Note that our assumptions are slightly stronger than those in [6]. The conditions above are satisfied by various distributions, like the distributions considered in Examples (i) and (ii) in Section 3 of [6]. From Theorem 8.1.6 in [4], it is easily shown that the assumptions above imply that $1 - B(t)$ varies regularly with index $-\nu$.

Here is our main result for the case $\lambda < \mu$:

Theorem 4.2 *Suppose $\lambda < \mu$ and (4.16) holds. Then*

$$1 - W(t) \sim \frac{\lambda l(0)(1 - Q_0 - Q_1)}{\mu\Gamma(2-\nu)} \left(\frac{\mu - \lambda}{\mu}\right)^{\nu-1} t^{1-\nu} e^{(\lambda-\mu)t}, \qquad t \to \infty. \tag{4.17}$$

Proof. Since $\lambda < \mu$, the ordinary $M/M/1$ queue with input rate λ and service rate μ is stable. Denote by $W_{M/M/1}$ the waiting time in this ordinary $M/M/1$ queue. It is easy to see that W is stochastically smaller than $W_{M/M/1}$, i.e.,

$$1 - W(t) \leq 1 - W_{M/M/1}(t) = \frac{\lambda}{\mu} e^{(\lambda-\mu)t}, \qquad t > 0, \tag{4.18}$$

which implies that the rightmost singularity $-s^* \leq \lambda - \mu$.

Next, we shall show that $w(s)$ is analytic in the region $\{s : \text{Re } s \geq \lambda - \mu - \delta\} \backslash \{\lambda - \mu\}$ for some $\delta > 0$. By (4.18), we know that $w(s)$ is an analytic function in the region $\{s : \text{Re } s \geq \lambda - \mu + \epsilon\}$ for any $\epsilon > 0$. So it is sufficient to show that $w(s)$ is an analytic function for $s \in \{s : \lambda - \mu - \delta \leq \text{Re } s < 0\} \backslash \{\lambda - \mu\}$. Noting that $\beta_e(s)$ is analytic in the region $\{s : \text{Re } s > 0 \text{ or } |s| < \epsilon_0\} \backslash \{0\}$, we may continue $w(s)$ as given by (2.21) analytically into $\{s : \text{Re } \hat{f}(s) \geq 0 \text{ or } |\hat{f}(s)| < \epsilon_0\} \cap \{s : \text{Re } s > \lambda - \mu - \delta\}$:

$$\begin{aligned} w(s) &= Q_0 + Q_1 - \frac{\mu + s}{\lambda - \mu - s} \tilde{Q}_1 + \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))} \\ &\quad + \frac{\mu \int_{x=0}^{\infty} e^{\hat{f}(s)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-\hat{f}(s)\eta} \mathrm{d}B(\eta) \mathrm{d}x}{(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]}. \end{aligned} \tag{4.19}$$

For $s \in \{s : \text{Re } \hat{f}(s) > 0 \text{ and Re } s < 0\}$, we have

$$1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s)) = \frac{\lambda - s}{s}(\beta(\hat{f}(s)) - 1 + s/\lambda) \neq 0. \tag{4.20}$$

Since

$$\Big[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))\Big]_{s=\lambda-\mu} = \mu/\lambda > 0,$$

it follows that there exists $\epsilon_1 > 0$ such that, for $|s - \lambda + \mu| < \epsilon_1$, we have

$$\text{Re } [1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))] > 0.$$

Hence, from the analytic continuation of $w(s)$ as given in (4.19), we conclude that $w(s)$ is analytic in the region $\{s : \text{Re } \hat{f}(s) > 0 \text{ and Re } s < 0\} \cup \{s : |s - \lambda + \mu| < \epsilon_1\} \backslash \{\lambda - \mu\}$. Taking $s = x + yi$, we have

$$\hat{f}(s) = \frac{x(x - \lambda + \mu)(x - \lambda) + (x + \mu)y^2}{(x - \lambda)^2 + y^2} + i\frac{y^3 + y((\lambda - x)^2 - \lambda\mu)}{(\lambda - x)^2 + y^2}. \tag{4.21}$$

It is easy to check that there exist $\delta > 0$ such that $\{s : \lambda - \mu - \delta < \text{Re } s < 0\} \subseteq \{s : \text{Re } \hat{f}(s) \geq 0\} \cup \{s : |s - \lambda + \mu| < \epsilon_1 \text{ and } |\hat{f}(s)| < \epsilon_0\}$.

In order to apply the results in [21], we define

$$\tilde{w}(s) := \frac{1 - w(s)}{s} - \frac{1 - Q_0 - Q_1 - \tilde{Q}_1}{s - \lambda + \mu}. \tag{4.22}$$

We may write, for $t > 0$ and some real a,

$$\begin{aligned} 1 - W(t) &= \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} e^{ts} \frac{1 - w(s)}{s} \mathrm{d}s \\ &= \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} e^{ts} \tilde{w}(s) \mathrm{d}s + (1 - Q_0 - Q_1 - \tilde{Q}_1) e^{(\lambda-\mu)t}. \end{aligned} \tag{4.23}$$

By (2.21) and (4.22), we have

$$\begin{aligned} \tilde{w}(s) &= \frac{(\lambda - \mu)(1 - Q_0 - Q_1) + \mu \tilde{Q}_1}{s(\lambda - \mu - s)} - \frac{\beta[(\lambda - \mu - s)Q_1 + \lambda Q_0]\beta_e(\hat{f}(s))}{s[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]} \\ &\quad - \frac{\mu \int_{x=0}^{\infty} e^{\hat{f}(s)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-\hat{f}(s)\eta} \mathrm{d}B(\eta) \mathrm{d}x}{s(\lambda - \mu - s)[1 - s/\lambda - (\lambda - \mu - s)\beta\beta_e(\hat{f}(s))]}. \end{aligned} \tag{4.24}$$

It is not difficult to check from (4.24) that $\tilde{w}(s) \to 0$ as $y \to \pm\infty$, uniformly in x for $\lambda - \mu - \delta \leq x \leq \frac{\lambda-\mu}{2}$, and in such a manner that $\int_0^\infty |\tilde{w}(s)| \mathrm{d}y < \infty$.

In the following we shall concentrate on the asymptotic behaviour of $w(s)$ in the neighbourhood of $\lambda - \mu$. We apply similar arguments as in the proof of Theorem 4.1. In order to simplify the notation, we introduce $z := s - \lambda + \mu$. There exists $z_1(z) = \mu - \lambda - z + \hat{f}(z + \lambda - \mu)$ such that $\hat{f}(z + \lambda - \mu) = \hat{f}(z_1 + \lambda - \mu)$. Taking $p = 1 - s/\lambda = (\mu - z)/\lambda$ in (2.10), we obtain, for $|z - \mu| < \mu$,

$$c_1(\mu/\lambda - z/\lambda) = \frac{(\mu - z)(\lambda Q_0 - Q_1 z)/\lambda + \mu \int_{x=0}^{\infty} e^{\hat{f}(z+\lambda-\mu)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-\hat{f}(z+\lambda-\mu)\eta} \mathrm{d}B(\eta) \mathrm{d}x}{1 + \frac{\lambda\beta z}{\mu - z} \beta_e(\hat{f}(z + \lambda - \mu))}. \tag{4.25}$$

Using the above relation we may write, for $z \in \{z : |z| \leq \mu - \lambda,\ |\lambda + z - \hat{f}(z + \lambda - \mu)| < \mu\}$,

$$\begin{aligned} &\int_{x=0}^{\infty} e^{\hat{f}(z+\lambda-\mu)x} Q_1(x) \int_{\eta=x}^{\infty} e^{-\hat{f}(z+\lambda-\mu)\eta} \mathrm{d}B(\eta) \mathrm{d}x \\ &= \frac{1}{\mu} c_1(\mu/\lambda - z_1/\lambda) \left[1 + \frac{\lambda\beta z_1}{\mu - z_1} \beta_e(\hat{f}(z_1 + \lambda - \mu))\right] - \frac{1}{\lambda\mu}(\mu - z_1)(\lambda Q_0 - Q_1 z_1) \\ &= \frac{1}{\mu} c_1(1 + z/\lambda - \hat{f}(z + \lambda - \mu)/\lambda) \left[1 + \frac{(\mu - \lambda - z + \hat{f}(z + \lambda - \mu))\lambda\beta}{\lambda + z - \hat{f}(z + \lambda - \mu)} \beta_e(\hat{f}(z + \lambda - \mu))\right] \\ &\quad - \frac{1}{\lambda\mu}(\lambda + z - \hat{f}(z + \lambda - \mu))[\lambda Q_0 - (\mu - \lambda - z + \hat{f}(z + \lambda - \mu))Q_1]. \end{aligned} \tag{4.26}$$

Noting that $c_1(p)$ is analytic in $|p| < \mu/\lambda$ and $c_1(1) = (1 - Q_0 - Q_1)/\beta_1$, we may write: for $|p - 1| < (\mu - \lambda)/\lambda$,

$$c_1(p) = (1 - Q_0 - Q_1)/\beta + \sum_{i=1}^{\infty} c_{1,i}(p - 1)^i,$$

where $c_{1,i}$ $(i = 1, 2, ...)$ are real constants. Again, since $\hat{f}(z + \lambda - \mu)$ is analytic in the region $|z| < \mu - \lambda$ and $\hat{f}(\lambda - \mu) = 0$, it follows that $c_1(\frac{\lambda+z-\hat{f}(z+\lambda-\mu)}{\lambda})$ is also analytic in the region $\{z : |z + (\mu - \lambda - z)z/(\mu - z)| < \min(\lambda, \mu - \lambda), |z| < \mu - \lambda\}$. Thus, $c_1(\frac{\lambda+z-\hat{f}(z+\lambda-\mu)}{\lambda})$ can be represented as:

$$c_1(\frac{\lambda + z - \hat{f}(z + \lambda - \mu)}{\lambda}) = (1 - Q_0 - Q_1)/\beta + z\tilde{c}_1(z), \tag{4.27}$$

where $\tilde{c}_1(z)$ is analytic in the region $|z| < \delta$ for some $\delta > 0$. Then substituting (4.26) into (4.19), yields, for $|z| < \delta$,

$$\begin{aligned} w(s) &= w(z + \lambda - \mu) \\ = \quad & Q_0 + Q_1 + \tilde{Q}_1 + \frac{(\lambda Q_0 - Q_1 z)\beta\beta_e(\hat{f}(z + \lambda - \mu))}{\mu/\lambda - z/\lambda + z\beta\beta_e(\hat{f}(z + \lambda - \mu))} \\ & + \frac{\lambda}{z}\tilde{Q}_1 - \frac{c_1(1 + z/\lambda - \hat{f}(z + \lambda - \mu)/\lambda)\left[1 + \frac{(\mu-\lambda-z+\hat{f}(z+\lambda-\mu))\lambda\beta}{\lambda+z-\hat{f}(z+\lambda-\mu)}\beta_e(\hat{f}(z + \lambda - \mu))\right]}{z[\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z + \lambda - \mu))]} \\ & + \frac{(1 + z/\lambda - \hat{f}(z + \lambda - \mu)/\lambda)[\lambda Q_0 - (\mu - \lambda - z + \hat{f}(z + \lambda - \mu))Q_1]}{z[\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z + \lambda - \mu))]}. \end{aligned} \tag{4.28}$$

By using (2.13) and (4.27), one can easily check that

$$\begin{aligned} A(z) \quad := \quad & \frac{1}{z}\Bigg[\mu\tilde{Q}_1 - \tilde{Q}_1 z + \lambda\beta z\beta_e(\hat{f}(z + \lambda - \mu)) + (1 + z/\lambda - \hat{f}(z + \lambda - \mu)/\lambda) \\ & [\lambda Q_0 - (\mu - \lambda - z + \hat{f}(z + \lambda - \mu))Q_1] + c_1(1 + z/\lambda - \hat{f}(z + \lambda - \mu)/\lambda) \\ & \left(1 + \frac{(\mu - \lambda - z + \hat{f}(z + \lambda - \mu))\lambda\beta}{\lambda + z - \hat{f}(z + \lambda - \mu)}\beta_e(\hat{f}(z + \lambda - \mu))\right)\Bigg] \\ = \quad & h_1(z) + h_2(z)\beta_e(\hat{f}(z + \lambda - \mu)) \\ & + \frac{\lambda(\mu - \lambda)(1 - Q_0 - Q_1)}{\lambda + z - \hat{f}(z + \lambda - \mu)}\frac{1 - \beta_e(\hat{f}(z + \lambda - \mu))}{z}, \end{aligned} \tag{4.29}$$

where $h_j(z)$ $(j = 1, 2)$ are both analytic functions for $|z| < \delta_1$ with δ_1 some positive constant. Combining (4.28) and (4.29), we obtain, for $|z| < \delta_1$,

$$\begin{aligned} w(z + \lambda - \mu) = Q_0 + Q_1 + \tilde{Q}_1 + \frac{h_1(z) + (h_2(z) + \lambda\beta Q_0 - \beta Q_1 z)\beta_e(\hat{f}(z + \lambda - \mu))}{\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z + \lambda - \mu))} \\ + \frac{\lambda(\mu - \lambda)(1 - Q_0 - Q_1)}{(\lambda + z - \hat{f}(z + \lambda - \mu))[\mu/\lambda - z/\lambda + \beta z\beta_e(\hat{f}(z + \lambda - \mu))]}\frac{1 - \beta_e(\hat{f}(z + \lambda - \mu))}{z}. \end{aligned} \tag{4.30}$$

From (4.16), (4.22) and (4.30), we conclude that

$$\tilde{w}(s) = \sum_{j=-1}^{\infty} d_j(s - \lambda + \mu)^{r_j}, \quad (-1 = r_{-1} < r_0 < r_1 < ...), \quad |s - \lambda + \mu| < \delta_2, \tag{4.31}$$

where δ_2 is some positive constant. If $1 < \nu < 2$, we have

$$\begin{aligned} r_{-1} &= -1, \qquad & d_{-1} &= -(1 - Q_0 - Q_1 - \tilde{Q}_1), \\ r_0 &= \nu - 2, \qquad & d_0 &= l(0)(1 - Q_0 - Q_1)\frac{\lambda}{\mu}\left(\frac{\mu - \lambda}{\mu}\right)^{\nu - 1}, \end{aligned} \tag{4.32}$$

where $l(\cdot)$ is given in (4.16); if $m < \nu < m+1$ $(m \geq 2)$, we have

$$\begin{aligned} r_{-1} &= -1, \qquad & d_{-1} &= -(1 - Q_0 - Q_1 - \tilde{Q}_1), \\ r_j &= j; \qquad & j &= 0, ..., m-2, \\ r_{m-1} &= \nu - 2, \qquad & d_{m-1} &= l(0)(1 - Q_0 - Q_1)\frac{\lambda}{\mu}\left(\frac{\mu - \lambda}{\mu}\right)^{\nu - 1}. \end{aligned}$$

Therefore, applying the result in [21], it follows from (4.23) that

$$\begin{aligned} & 1 - W(t) - (1 - Q_0 - Q_1 - \tilde{Q}_1)e^{(\lambda - \mu)t} \\ \approx \quad & -(1 - Q_0 - Q_1 - \tilde{Q}_1)e^{(\lambda - \mu)t} + \sum_{j=0}^{\infty} \frac{d_j}{\Gamma(-r_j)} t^{-r_j - 1}, \quad \left(\frac{1}{\Gamma(-r_j)} = 0 \text{ for } r_j = 0, 1, 2, ...\right), \end{aligned}$$

which implies that (4.17) holds. □

Like in the previous subsection, we rewrite (4.17). A straightforward computation (using Theorem 8.1.6 in [4] to obtain the tail behaviour of the distribution of B^{res} from (4.16)) shows that, for $t \to \infty$,

$$1 - W(t) \sim (1 - Q_0 - Q_1)\mathbf{P}\left(B^{res} > \frac{\mu t}{\mu - \lambda}\right)\mathbf{P}(W_{M/M/1} > t). \tag{4.33}$$

This result has the following intuitive interpretation: A large waiting time W occurs as a consequence of a large service time at server 2, which lets the system function as an M/M/1 queue. It is well known from standard large deviations theory that the most probable way of the workload in an M/M/1 queue ($W_{M/M/1}$) getting large is in a linear fashion, with a positive drift of $\mu/\lambda - 1$ (see e.g. p. 276 of [22]). Hence, the time it takes until $W_{M/M/1} > t$ (given that this event occurs) is equal to $\lambda t/(\mu - \lambda)$.

In order to let the deviant behaviour of the M/M/1 take place, server 2 needs to be occupied (which has probability $1 - Q_0 - Q_1$) and the past service time B^{past} of the customer must be bigger than $\lambda t/(\mu - \lambda)$. Finally, the residual service time B^{res} of the customer at server 2 must be bigger than t. Standard renewal theory (see e.g. [8], p. 113) gives

$$\mathbf{P}\left(B^{past} > \frac{\lambda t}{\mu - \lambda}, B^{res} > t\right) = \mathbf{P}\left(B^{res} > \frac{\mu t}{\mu - \lambda}\right).$$

Combining all these observations yields (4.33). The above interpretation shows an interesting feature of this model: A waiting time can become very large by the simultaneous occurence of two events: A very long waiting time at an exponential server (M/M/1 large deviations) and one large service time of a heavy-tailed server.

5 Conclusion

The main results of our study of the heterogeneous $M/G/2$ queue with one exponential and one general server are: (i) an exact analysis of the queue length and waiting time distribution if the general service time distribution has a rational LST, and (ii) an asymptotic analysis of the waiting time tail if the general service time distribution is regularly varying at infinity. The analysis of (i) may be extended to the case of an $M/G/s$ queue with $s-1 \geq 1$ exponential servers and one general server with rational service time LST. The exact and heuristic analysis in (ii) should form just the beginning of an investigation of waiting time asymptotics in multi-server queues. In the two-server case, we have not yet been able to handle the intricate case $\lambda = \mu$; another line of research would be to generalize the class of service time distributions for server 1.

Acknowledgment The authors gratefully acknowledge interesting discussions with S.C. Borst, D. Korshunov and W. Whitt.

References

[1] J. Abate, G.L. Choudhury, and W. Whitt (1994). Waiting time tail probabilities in queues with long-tail service-time distributions, *Queueing Systems*, **16**, 311-338.

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

[3] R. Agrawal, A. Makowski, and P. Nain (1999). On a reduced load equivalence for a fluid model under subexponential assumptions. Report INRIA Sophia-Antipolis, June 1998. *Queueing Systems*, to appear.

[4] N.H. Bingham, C.M. Goldie, and J.L. Teugels (1989). *Regular Variation*, Cambridge Univ. Press, Cambridge, England.

[5] S.C. Borst, O.J. Boxma, and P.R. Jelenković (1999). *Generalized processor sharing with long-tailed traffic sources.* In: *Teletraffic Engineering in a Competitive World, Proc. ITC-16*, Edinburgh, UK, eds. P. Key, D. Smith (North-Holland, Amsterdam), 345-354.

[6] O.J. Boxma, and J.W. Cohen (1999). Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. CWI Report PNA-R9710, 1997. *Queueing Systems*, to appear.

[7] J.W. Cohen (1973). Some results on regular variation for distributions in queueing and fluctuation theory, *Journal of Applied Probability*, **10**, 343-353.

[8] J.W. Cohen (1982). *The Single Server Queue,* Second edition, North Holland, Amsterdam.

[9] J.W. Cohen (1998). A heavy-traffic theorem for the $GI/G/1$ queue with a Pareto-type service. time distribution, *Journal of Applied Mathematics and Stochastic Analysis*, **11**, 339-355.

[10] T. Daniëls (1999). *Asymptotic Behaviour of Queueing Systems.* PhD Thesis, Universiteit Antwerpen.

[11] G. Doetsch (1974). *Introduction to the Theory and Applications of the Laplace Transformation,* Springer, New York.

[12] V. Dumas, and A. Simonian (1998). Asymptotic bounds for the fluid queue fed by subexponential on/off sources. Technical report, Univ. Bordeaux.

[13] R. Haji, and G.F. Newell (1971). A relation between stationary queue length and waiting time distributions, *Journal of Applied Probability*, **8**, 617-620.

[14] J. Kiefer and J. Wolfowitz (1956). On the characteristics of the general queueing process with applications to random walk, *Annals of Mathematical Statistics*, **27**, 147-161.

[15] D.A. Korshunov (1999). On waiting time distribution in $GI/GI/2$ queue system with heavy tailed service times. Unpublished manuscript.

[16] C. Knessl, B.J. Matkowsky, Z. Schuss and C. Tier (1990). An integral equation approach to the M/G/2 queue, *Operations Research*, **38**, 506-518.

[17] A.G. Pakes (1975). On the tails of waiting-time distributions, *Journal of Applied Probability*, **12**, 555-564.

[18] S. Resnick and G. Samorodnitsky (1999). Steady state distribution of the buffer content for $M/G/\infty$ input fluid queues. Report no. 1242, Department of ORIE, Cornell University.

[19] A. Scheller-Wolf, and K. Sigman (1997). Delay moments for FIFO GI/GI/s queues, *Queueing Systems*, **25**, 77-95.

[20] A. Scheller-Wolf (1999). Further delay moment results for FIFO multiserver queues. Report GSIA, Carnegie Mellon University, Pittsburgh, March 1999.

[21] W.G.L. Sutton (1934). The asymptotic expansion of a function whose operational equivalent is known, *Journal of the London Mathematical Society*, **9**, 131-137.

[22] A. Shwartz and A. Weiss (1995). *Large Deviations for Performance Analysis.* Chapman and Hall, London.

[23] E.C. Titchmarsh (1952). *The Theory of Functions*, Oxford University Press, London.

[24] W. Whitt (1998). The impact of a heavy-tailed service-time distribution upon the M/GI/s waiting-time distribution. Report AT&T Labs, December 1998.