Queueing Syst (2016) 84:355–379
DOI 10.1007/s11134-016-9500-3

A tandem fluid network with Lévy input in heavy traffic

D. T. Koops[1] · **O. J. Boxma**[2] · **M. R. H. Mandjes**[1]

Received: 15 December 2015 / Revised: 5 August 2016 / Published online: 17 September 2016

Abstract In this paper we study the stationary workload distribution of a fluid tandem queue in heavy traffic. We consider different types of Lévy input, covering compound Poisson, α-stable Lévy motion (with $1 < \alpha < 2$), and Brownian motion. In our analysis, we separately deal with Lévy input processes with increments that have finite and infinite variance. A distinguishing feature of this paper is that we do not only consider the usual heavy traffic regime, in which the load at one of the nodes goes to unity, but also a regime in which we *simultaneously* let the load of both servers tend to one, which, as it turns out, leads to entirely different heavy traffic asymptotics. Numerical experiments indicate that under specific conditions the resulting simultaneous heavy traffic approximation significantly outperforms the usual heavy traffic approximation.

Keywords Queueing theory · Tandem queue · Lévy processes · Fluid queue · Heavy traffic · Steady-state distribution · Heavy tails

Mathematics Subject Classification 60K25

1 Introduction

In this paper we study a fluid tandem queue that consists of two servers in series. A spectrally positive Lévy process serves as the input process of the first queue (also: upstream queue). The first server empties the upstream queue at a deterministic rate r_1,

✉ D. T. Koops
d.t.koops@uva.nl

[1] Korteweg-de Vries Institute, University of Amsterdam, Amsterdam, The Netherlands

[2] Eurandom and Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

immediately feeding the second (also: downstream) queue. The downstream server leaks at some deterministic rate r_2; to make the system non-trivial we throughout assume $r_2 < r_1$. After the fluid has passed the second server, it leaves the system. We are interested in the stationary workloads in both queues in heavy traffic regimes that we specify below.

The heavy traffic regime was first considered in [1]: one lets the load of the system tend to one, while simultaneously scaling the workload in such a way that a non-degenerate limiting distribution is obtained. Kingman's approach was mainly based on manipulating Laplace–Stieltjes transforms; this approach we also follow in our paper. Another approach relies on the functional central limit theorem in combination with the continuous mapping theorem; see, for example, [2]. In [3], both approaches are compared, and the traditional heavy traffic results, which assume the increments of the input process have a finite variance, are generalized to the infinite variance case. For excellent surveys, we refer to [4] and the book [5]. Tandem queueing systems in which both queues are experiencing heavy traffic conditions have been studied before. Harrison [6] has focused on the classical setting of a GI/G/1-type tandem in which discrete entities ('customers') receive service in each server and move to the next queue (or leave the system) only after its full service has been completed. In such queueing systems, the correlation between both queues is typically *negative*, as the first queue being relatively large could be a consequence of long service times in that queue, which in turn result in long inter-arrival times in the second queue, and hence a relative small number of customers in the second queue. Harrison manages to quantify the resulting (negative) covariance between the populations in both queues in heavy traffic. Importantly, in the fluid setting considered in our work, this reasoning does *not* hold. More specifically, for the types of models we study, the correlation between both workloads is *positive*: large workloads in the upstream queue likely correspond to large workloads in the downstream queue.

Fluid tandem queues with spectrally positive Lévy input were initially scrutinized in a series of papers starting with [7] and the follow-up paper [8]. The results concerning the joint distribution of the steady state of the workloads were studied in a more general network setting in, for example, [9]. These results play an important role for our analysis and are therefore summarized in Sect. 2.2. An extensive account of Lévy-driven networks can be found in Chaps. 12 and 13 of [10].

The load of a server is defined as the average input rate into the server divided by its service rate. The load can thus be increased by increasing the average input rate, or lowering the service rate. In case of a single-node system, both methods are equivalent in the sense that they lead to the same heavy traffic results. However, for multi-node systems (such as tandem queues), increasing the average input to the first server only leads to heavy traffic in the downstream server (recall that $r_1 > r_2$). To be more general, we therefore adapt the *service rates* appropriately, while keeping the input process fixed. Taking this approach opens up the possibility that the servers experience heavy traffic simultaneously. In this paper, we study both types of heavy traffic and refer to them as follows:

- *Regime I*, If only the downstream server has a load that tends to unity (whereas the first queue does not operate under heavy traffic);

- *Regime II*, If the up- and downstream server have loads that *simultaneously* tend to unity.

Even though this particular approach to heavy traffic in a tandem setting is new, several related approaches have been developed earlier. More specifically, we would like to refer to Example 9.9.1 on p. 335 in [5]. There it is mentioned that the 'standard' heavy traffic approach may lead to poor approximations in a network setting. In particular, the expected waiting time of the last server in a tandem is estimated by using different approximations of the squared coefficient of variation of the arrival process at the corresponding server. The resulting approximations of the expected waiting time are then compared to simulated values, and it is noted that substantial improvements can be made by making use of a suitable choice for the squared coefficient of variation. In our paper, however, we use a different scaling that leads to new results on the *distribution* of the steady-state workload at the downstream node, rather than merely its mean.

In general terms, the results we find for Regime I are much in line with those for heavy traffic in single queues, whereas for Regime II we obtain limiting distributions which, to the best of our knowledge, have not appeared before. More specifically, our contributions are as follows:

- For Regime I, we find that the steady-state distribution of the workload in the second queue is similar to the one of the first queue. Moreover, the up- and downstream workloads are asymptotically independent in the heavy traffic limit.
- In Regime II, we establish the interesting feature that the workloads do not decouple in heavy traffic, i.e. some dependence between the up- and downstream workloads remains. Moreover, the marginal steady-state distribution of the downstream queue is crucially different from the one obtained in Regime I. This has practical implications: as verified through a set of experiments, Regime II approximations tend to outperform those based on Regime I, particularly when the load of both servers is large.

We find that, as in the single-server case, there is a dichotomy between input processes that have increments with finite and infinite variance; as a consequence, they have to be dealt with separately. We have derived Regime I results in both cases, and for the case of finite variance, we have also succeeded in addressing the technically more demanding Regime II.

In Regime I, we prove that the stationary workload of the downstream queue has an exponential distribution (for the case of finite variance) or Mittag-Leffler distribution (for infinite variance). Remarkably, the same distributions (up to some factor) were found for single fluid queues; apparently, the fact that there is an additional fluid server that modifies the process hardly affects the limiting distribution. In addition, similar results were also found for waiting times in non-fluid single GI/G/1 queues; see [11] for the case of infinite variance.

The paper is organized as follows. In Sect. 2 we introduce our framework of queueing models with Lévy input; we subsequently explain the fluid Lévy tandem queueing model that we consider and recall results that play a key role throughout the paper. As mentioned above, there is a dichotomy between the case of finite (Sect. 3) and infinite variance (Sect. 4). In Sect. 3 we first consider Brownian input, for which all computations can be done explicitly, and then turn to general spectrally positive Lévy input. This

section also includes numerical experiments that indicate that the Regime II approximation typically outperforms the Regime I approximation. Section 4, which focuses on infinite variance input, covers results for compound Poisson input and α-stable input. Finally, in the Appendix, we state Tauberian theorems that are used in Sect. 4.

2 Lévy driven queues

In this section we briefly introduce the fluid tandem queueing model, and we state some results that are important for the remainder of the paper.

2.1 A fluid tandem queueing model

We consider a Lévy driven fluid tandem queue consisting of two servers. The Lévy input process $J = \{J_t, t \geq 0\}$ feeds the first server (upstream server). The workload from the first server then flows continuously, at a fixed rate r_1, to the second server (downstream server). The downstream server empties itself at a fixed rate r_2 and the exiting fluid leaves the system. We denote the workload in queue 1, 2 as $Q^{(1)}$, $Q^{(2)}$, respectively, and define $X_t^{(i)} := J_t - r_i t$, for $i = 1, 2$. Then we can precisely define the workload process of the first node in the Lévy-driven queue as

$$Q_t^{(1)} := Q_0^{(1)} + X_t^{(1)} + \sup_{0\leq s\leq t} \left(Q_0^{(1)} + X_s^{(1)}\right)^-,$$

in which $(x)^-$ denotes $-\min\{x, 0\}$. The output process of the first server can be represented as

$$D_t := r_1 t - \sup_{0\leq s\leq t} \left(Q_0^{(1)} + X_s^{(1)}\right)^-.$$

This output process is then the input process of the downstream server, which results in the following workload representation:

$$Q_t^{(2)} := Q_0^{(2)} + D_t - r_2 t + \sup_{0\leq s\leq t} \left(Q_0^{(2)} + D_s - r_2 s\right)^-.$$

Consider Fig. 1 for a diagram of this model and consider Fig. 2 for a typical sample path when the arrival process is a renewal process. Assume $r_2 < r_1$, as otherwise the second queue would remain empty. We use two different parametrizations in this paper. In Regime I, we parametrize

$$r_1 = \mathbb{E}\, J_1 + r, \quad \text{for some fixed } r > 0, \text{ and } r_2 = \mathbb{E}\, J_1 + \epsilon.$$

For Regime II, we take

$$r_1 = \mathbb{E}\, J_1 + \gamma\epsilon \quad \text{and} \quad r_2 = \mathbb{E}\, J_1 + \epsilon, \quad \text{in which } \gamma > 1 \text{ to guarantee that } r_1 > r_2.$$

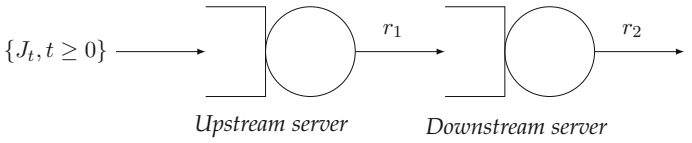

Fig. 1 A diagram of the fluid tandem queueing system that we consider in this paper

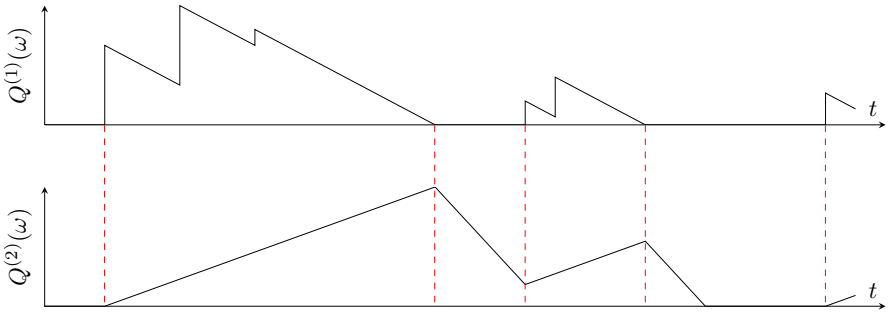

Fig. 2 An arbitrary sample path in the fluid tandem queue. During busy periods of the *upstream* queue, the downstream server fills up with a net rate of $r_1 - r_2$. During idle periods, the workload of the *downstream* server decreases with rate r_2

In Regime I, the upstream server will have a fixed load of $\rho_1 = \mathbb{E}\, J_1/r_1 < 1$ as $\epsilon \downarrow 0$, whereas the load of the downstream server will tend to one: $\rho_2 = \mathbb{E}\, J_1/r_2 \uparrow 1$ as $\epsilon \downarrow 0$. In Regime II, on the contrary, *both* the up- and downstream server will have loads that tend to one: $\rho_1, \rho_2 \uparrow 1$ as $\epsilon \downarrow 0$. To avoid the workload from increasing indefinitely, we scale the workloads so as to obtain a non-degenerate limit. Only the queues for which the load is increasing, an appropriate scaling is required. The specific way in which the workloads should be scaled depends on the type of input (more specifically, it matters whether the increments have finite variance or not); this will be pointed out in detail later in the paper. In addition, note that r_i contains a term $\mathbb{E}\, J_i$, which negates the drift of the input process. Therefore, the drift of the input process is not important and can be assumed to be zero in the remainder of the paper.

We now introduce some additional notation. We denote by ϕ the *Laplace exponent*

$$\phi(\alpha) = \log \mathbb{E}\left[e^{-\alpha X_1^{(1)}}\right],$$

and the inverse function of ϕ by $\phi^{-1} \equiv \psi$.

2.2 Useful results on transforms

To ensure stability, it is required that the average input rate is less than the speed of the slowest server, i.e. $\mathbb{E}\, J_1 < r_2$. Therefore, it is possible to define a random variable $(Q_0^{(1)}, Q_0^{(2)})$, so that the resulting bivariate process $\{(Q_t^{(1)}, Q_t^{(2)}), t \geq 0\}$ is stationary. We write $Q^{(i)}$ for a random variable with distribution equal to $Q_t^{(i)}$, for a fixed t, when the process is initiated as mentioned above.

The theorems stated below, which uniquely characterize the distributions of the $Q^{(i)}$, play a crucial role throughout the paper. The following assertions are Theorems 3.2, 12.11, and 12.3, respectively, copied from the book [10] (mostly using their notation). Closely related results were originally developed in [7], cf. Eq. (4.12) in their paper. Theorem 2.1 gives the Laplace–Stieltjes tranform (LST) for the stationary workload if there is only one server and can be considered to be a generalization of the well-known Pollaczek–Khinchine formula. The LST for the joint stationary workload in the fluid tandem system is presented in Theorem 2.2, which also provides us with the LST for the downstream queue only (Corollary 2.3).

Theorem 2.1 [Generalized Pollaczek–Khinchine (PK)] *Let $J \in \mathcal{S}^+$. For $s \geq 0$,*

$$\mathbb{E}\, e^{-sQ^{(1)}} = \frac{s\phi'(0)}{\phi(s)}.$$

Theorem 2.2 (Two-dimensional PK for fluid tandem) *Let $J \in \mathcal{S}^+$. For $s_1, s_2 \geq 0$,*

$$\mathbb{E}\, e^{-s_1 Q^{(1)} - s_2 Q^{(2)}} = \frac{-\mathbb{E}\, X_1^{(2)} s_2}{s_2 - \psi\big(s_2(r_1 - r_2)\big)} \frac{\psi\big(s_2(r_1 - r_2)\big) - s_1}{(r_1 - r_2)s_2 - \phi(s_1)}.$$

Corollary 2.3 (One-dimensional PK for fluid tandem) *Let $J \in \mathcal{S}^+$. For $s \geq 0$,*

$$\mathbb{E}\, e^{-sQ^{(2)}} = \frac{-\mathbb{E}\, X_1^{(2)}}{r_1 - r_2} \frac{\psi\big(s(r_1 - r_2)\big)}{s - \psi\big(s(r_1 - r_2)\big)}.$$

Remark 2.4 Throughout the remainder of the paper, we assume $J \in \mathcal{S}^+$ and $\mathbb{E}\,|J_1| < \infty$. It is straightforward to extend our results to spectrally negative input processes $J \in \mathcal{S}^-$, by making use of Laplace–Stieltjes transforms for $\mathcal{S}^-$-processes, which can be found in, for example, Theorem 12.12 of [10].

3 Input processes with finite variance

In this section we consider the fluid tandem queue for various types of input processes that have increments with finite variance. Since, for Brownian input, an explicit analysis can be performed, we consider this case first (Sect. 3.1). Using appropriate expansions, we show in Sect. 3.2 how these results extend to spectrally positive Lévy processes. In both cases, we establish Regime I and Regime II results. Finally, in Sect. 3.3, we provide a numerical comparison between the Regime I and Regime II approximations.

3.1 Brownian input

Assume that the input is Brownian, that is, $J_t = \sigma W_t$, where W denotes a standard Brownian motion. Recall that we can assume, without loss of generality, that the input process has zero drift. Then we have

$$\phi(s) = \log \mathbb{E}\left[e^{-sJ_1}\right] = \log \mathbb{E}\, e^{-s(\sigma W_1 - r)} = \frac{1}{2}\sigma^2 s^2 + rs,$$

and after some elementary algebra we find that the inverse is given by

$$\psi(s) = -\frac{r}{\sigma^2} + \frac{1}{\sigma}\sqrt{\frac{r^2}{\sigma^2} + 2s}. \tag{1}$$

Regime I

In this case, the upstream server has a fixed load $\rho_1 < 1$, and the downstream server has a load that tends to one. Therefore, we have to scale the workload of the downstream server: to obtain a non-degenerate limit, we scale by ϵ. Relying on Theorem 2.2,

$$\begin{aligned}\mathbb{E}\, e^{-s_1 Q^{(1)} - s_2 \epsilon Q^{(2)}} &= \frac{\epsilon^2 s_2}{\epsilon s_2 - \psi\left(\epsilon s_2 (r-\epsilon)\right)} \frac{\psi\left(\epsilon s_2 (r-\epsilon)\right) - s_1}{(r-\epsilon)s_2\epsilon - \phi(s_1)} \\ &= \frac{1}{(r-\epsilon)s_2\epsilon - s_1 r - \frac{1}{2}s_1^2\sigma^2} \\ &\quad \times \frac{(\epsilon s_2 - s_1)\frac{1}{\sigma}\sqrt{\frac{r^2}{\sigma^2} + 2s_2\epsilon(r-\epsilon)} - s_1 s_2 \epsilon - \frac{s_1 r}{\sigma^2} + \frac{s_2\epsilon(r-\epsilon)}{\sigma^2}}{s_2 + \frac{2}{\sigma^2}},\end{aligned} \tag{2}$$

yielding the following proposition.

Proposition 3.1 *Suppose that the input process is a Brownian motion. Then, in Regime I, the joint stationary workload in heavy traffic is given by*

$$\mathbb{E}\, e^{-s_1 Q^{(1)} - s_2 \epsilon Q^{(2)}} \xrightarrow{\epsilon \downarrow 0} \frac{2r/\sigma^2}{2r/\sigma^2 + s_1} \frac{2/\sigma^2}{2/\sigma^2 + s_2}. \tag{3}$$

In particular, this implies that the distribution of $\epsilon Q^{(2)}$ converges to an exponential distribution with rate $2/\sigma^2$, which is equal to the distribution of the total workload. Moreover, it turns out that $Q^{(1)}$ and $\epsilon Q^{(2)}$ are asymptotically independent in the limit $\epsilon \downarrow 0$. Asymptotic independence should not be very surprising. In a pre-limit setting, there is a positive correlation between both buffer contents (cf. [7], Corollary 4.2). It follows from, for example, Eq. (4.11) in the same paper that the correlation tends to zero as the load in (only) the second node increases to one. Furthermore, one should realize that $Q^{(2)}$ is scaled by a factor ϵ, whereas $Q^{(1)}$ is not. It turns out that, due to the asymmetry in the spatial scaling, we obtain asymptotic independence.

Although this asymptotic independence is an interesting finding from a theoretical point of view, it has the intrinsic drawback that the original dependency structure is lost. Another drawback of this approximation is that it leads to significant errors if ρ_1 is large as well, as will be illustrated in Sect. 3.3. This prompts us to consider Regime II.

Regime II

In this regime, we scale both workloads, and we choose the service rates as explained in Sect. 2.1. Thus, we take $r = \gamma\epsilon$ in Eq. (1), so as to obtain

$$\psi(s) = -\frac{\gamma\epsilon}{\sigma^2} + \frac{1}{\sigma}\sqrt{\frac{\gamma^2\epsilon^2}{\sigma^2} + 2s}. \tag{4}$$

By Theorem 2.2,

$$\mathbb{E}\, e^{-s_1\epsilon Q^{(1)} - s_2\epsilon Q^{(2)}} = \frac{-\mathbb{E}\, X_1^{(2)} s_2\epsilon}{s_2\epsilon - \psi\big(s_2\epsilon(r_1 - r_2)\big)} \frac{\psi\big(s_2\epsilon(r_1 - r_2)\big) - s_1\epsilon}{(r_1 - r_2)s_2\epsilon - \phi(s_1\epsilon)}.$$

Using Eq. (4) yields

$$\mathbb{E}\, e^{-s_1\epsilon Q^{(1)} - s_2\epsilon Q^{(2)}} = \frac{s_2}{s_2 + \frac{\gamma}{\sigma^2} - \frac{1}{\sigma}\sqrt{\frac{\gamma^2}{\sigma^2} + 2(\gamma-1)s_2}} \frac{-\frac{\gamma}{\sigma^2} + \frac{1}{\sigma}\sqrt{\frac{\gamma^2}{\sigma^2} + 2(\gamma-1)s_2} - s_1}{(\gamma-1)s_2 - \gamma s_1 - \frac{1}{2}s_1^2\sigma^2},$$

where it should be noted that the expression on the right-hand side does not contain any ϵ anymore. This indicates that, for Brownian input, the joint distribution in the heavy traffic limit is of the same type as the distribution for 'non-heavy traffic loads' loads ρ_1 and ρ_2. After further simplification, we obtain

$$\mathbb{E}\, e^{-s_1\epsilon Q^{(1)} - s_2\epsilon Q^{(2)}} = \frac{s_2\big(\gamma - 2 - s_1\sigma^2\big) - s_1\gamma + (s_2 - s_1)\sigma\sqrt{\frac{\gamma^2}{\sigma^2} + 2(\gamma-1)s_2}}{(s_2\sigma^2 + 2)\Big((\gamma-1)s_2 - \gamma s_1 - \frac{1}{2}s_1^2\sigma^2\Big)}. \tag{5}$$

We can find the marginal distributions of the stationary workload of the first and second queue by plugging in $s_2 = 0$, respectively, $s_1 = 0$. This yields

$$\mathbb{E}\, e^{-s\epsilon Q^{(1)}} = \frac{2\gamma/\sigma^2}{2\gamma/\sigma^2 + s}, \tag{6}$$

$$\mathbb{E}\, e^{-s\epsilon Q^{(2)}} = \frac{1}{\gamma - 1}\frac{-2 + \gamma + \sqrt{\gamma^2 + 2s\sigma^2(\gamma-1)}}{2 + s\sigma^2}. \tag{7}$$

After lengthy but elementary calculations, we obtain

$$\mathbb{E}\Big[Q^{(1)}Q^{(2)}\Big] = \lim_{s_1\downarrow 0}\lim_{s_2\downarrow 0}\frac{\partial^2}{\partial s_1\partial s_2}\mathbb{E}\, e^{-s_1Q^{(1)} - s_2Q^{(2)}} = \frac{\gamma^2 - 1}{4\gamma^3}\sigma^4.$$

Using that $\mathbb{E}[Q^{(1)}] = \frac{1}{2}\sigma^2/\gamma$ and $\mathbb{E}[Q^{(2)}] = \frac{1}{2}\sigma^2(\gamma-1)/\gamma$,

$$\mathrm{Cov}\Big(Q^{(1)}, Q^{(2)}\Big) = \sigma^4\left(\frac{\gamma^2 - 1}{4\gamma^3} - \frac{\gamma - 1}{4\gamma^2}\right) = \frac{\gamma - 1}{4\gamma^3}\sigma^4. \tag{8}$$

To calculate the correlation coefficient, we also compute the variances. Since $Q^{(1)}$ has an $\mathrm{Exp}(2\gamma/\sigma^2)$ distribution, its variance is given by $\mathrm{Var}(Q^{(1)}) = \frac{1}{4}\sigma^4/\gamma^2$. By making use of the LST of $Q^{(2)}$, we also find

$$\mathrm{Var}(Q^{(2)}) = \frac{(\gamma-1)^2(\gamma+2)\sigma^4}{4\gamma^3}.$$

It now follows that the correlation coefficient is given by

$$\mathrm{Corr}\left(Q^{(1)}, Q^{(2)}\right) = c(\gamma) = \frac{1}{\sqrt{\gamma(\gamma+2)}}. \tag{9}$$

Observe that, when decreasing γ from ∞ to 1, $c(\gamma)$ increases from 0 to $1/\sqrt{3}$. This result is in line with Corollary 4.1 in [8]: there $c(\gamma)$ is studied without heavy traffic, and it is concluded that $c(\gamma) \in (0, 1/\sqrt{3})$. In the introduction, we already argued why $c(\gamma)$ is anticipated to be positive, but it can also be seen that $c(\gamma)$ decreases in γ. Indeed, as γ grows, the service rate in the upstream server increases. This implies that it becomes more likely that the downstream server has a large workload, while the workload in the first server may be relatively small due to its fast service.

3.2 General input

We now extend the results for the Brownian case in the previous section to spectrally positive Lévy input. Again we consider both regimes, starting with Regime I.

Regime I

In this section we prove the following main result.

Proposition 3.2 *Let the input process $J \in \mathcal{S}^+$ be such that $\mathrm{Var}\, J_1 = \sigma^2 < \infty$. Then, in Regime I, the stationary workloads of the up- and downstream queue are asymptotically independent, with $Q^{(1)}$ given by Theorem 2.1, and $Q^{(2)} \overset{\mathrm{d}}{=} \mathrm{Exp}(\frac{2}{\sigma^2})$.*

To prove this proposition, we require the following lemma.

Lemma 3.3 *Let*

$$\phi(s) = sr + \frac{1}{2}\sigma^2 s^2 + K_1 s^{\eta_1} + o(s^{\eta_1}), \tag{10}$$

with $\eta_1 > 2$. Then the inverse function ψ with argument $s\epsilon(r-\epsilon)$ satisfies, for $\epsilon \downarrow 0$,

$$\psi\left(s\epsilon(r-\epsilon)\right) = s\epsilon - \frac{1}{r}s\epsilon^2 - \frac{\sigma^2}{2r}s^2\epsilon^2 + o(\epsilon^2).$$

Proof of Lemma 3.3 Suppose that

$$\psi\left(s\epsilon(r-\epsilon)\right) = C_1 s\epsilon + C_2 s\epsilon^2 + C_3 s^2\epsilon^2 + o(\epsilon^2).$$

Consider

$$\begin{aligned}\phi\Big(\psi\big(s\epsilon(r-\epsilon)\big)\Big) - s\epsilon(r-\epsilon) &= \psi\big(s\epsilon(r-\epsilon)\big)r + \frac{1}{2}\sigma^2\psi\big(s\epsilon(r-\epsilon)\big)^2 \\ &\quad + K_1\psi\big(s\epsilon(r-\epsilon)\big)^{\eta_1} - s\epsilon(r-\epsilon) + o(\epsilon^2) \\ &= (C_1 r - r)s\epsilon + (rC_2 + 1)s\epsilon^2 \\ &\quad + \epsilon^2\Big(\sigma^2 C_1 + 2rC_3\Big)\frac{1}{2}s^2 + o(\epsilon^2).\end{aligned}$$

For ψ to be the inverse of ϕ for $\epsilon \downarrow 0$, we equate the above to zero. This is achieved by taking the constants $C_1 = 1$, $C_2 = -\frac{1}{r}$ and $C_3 = -\frac{\sigma^2}{2r}$. This proves the lemma. □

At first glance, it may be unclear why ψ in Lemma 3.3 has this specific form. However, in case of, for example, compound Poisson input, this shape arises naturally, as is demonstrated in Example 3.4 below. We first prove the main result.

Proof of Proposition 3.2 Assume that $\operatorname{Var} J_1 = \sigma^2 < \infty$. We first develop a general expansion for ϕ. From the definition of ϕ, we have $\phi(s) = sr_1 + \log \mathbb{E}\, e^{-sJ_1}$. Note that $\phi(s)$ is linear in r at $s = 0$:

$$\phi'(0) = r_1 - \mathbb{E}\, J_1 = \mathbb{E}\, J_1 + r - \mathbb{E}\, J_1 = r.$$

Now note that $\phi''(0) = \operatorname{Var} J_1 = \sigma^2$. This means that the coefficient of s^2 must be $\frac{1}{2}\sigma^2$. Upon combining all of the above, we see that necessarily

$$\phi(s) = sr + \frac{1}{2}\sigma^2 s^2 + o(s^2).$$

We can write

$$\phi(s) = sr + \frac{1}{2}\sigma^2 s^2 + K_1 s^{\eta_1} + o(s^{\eta_1}),$$

for some $K_1 \in \mathbb{R}$, where $\eta_1 = 3$ corresponds to the existence of a finite third moment, and $2 < \eta_1 < 3$ corresponds to an infinite third moment. It thus follows that ϕ as in (10) covers all input processes with finite second moment. Therefore, we can use the functions ϕ and ψ in Lemma 3.3 and apply them to Theorem 2.2. By scaling only the workload of the second queue by a factor ϵ and taking the heavy traffic limit, we find

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\, e^{-s_1 Q^{(1)} - s_2 \epsilon Q^{(2)}} = \frac{s_1 r}{\phi(s_1)} \frac{1}{1 + \frac{1}{2}\sigma^2 s_2} = \frac{s_1 \phi'(0)}{\phi(s_1)} \frac{1}{1 + \frac{1}{2}\sigma^2 s_2}.$$

The result follows. □

Example 3.4 Suppose that the input process is a compound Poisson process in which the first two moments of the job sizes are finite: $\mathbb{E}\, B, \mathbb{E}\, B^2 < \infty$. The goal is to find an asymptotic expression for $\psi(s\epsilon(r-\epsilon))$ as $\epsilon \downarrow 0$, while $s \geq 0$ is fixed. The proof

of Lemma 3.3 is by validation. How such an expression for $\psi(s\epsilon(r-\epsilon))$ can be constructed becomes clear in this example. We approach the problem in the following steps:

- Derive the Takács equation (describing the LST π of the busy period in an M/G/1 queue) with service rate equal to r_1;
- Use this Takács equation to express ψ in terms of π;
- Expand π, which yields an expansion for ψ.

Since we have a compound Poisson input process, the Laplace exponent is given by

$$\phi(s) = sr_1 - \lambda + \lambda b(s), \tag{11}$$

with $b(s) = \mathbb{E}\, e^{-sB}$. Let τ^0 denote the busy period started by a job arriving at an empty system. Using the standard argumentation, it turns out that

$$\pi(s) = b\Big(\frac{1}{r_1}\big(\lambda - \lambda\pi(s) + s\big)\Big); \tag{12}$$

this functional equation is well-known for $r_1 = 1$, cf. Sect. 1.3 in [12], but it can be readily extended to general r_1. Eqs. (11) and (12) imply

$$\frac{1}{\lambda}\phi\Big(\frac{1}{r_1}\big(\lambda - \lambda\pi(s) + s\big)\Big) - \frac{s}{\lambda} = b\Big(\frac{1}{r_1}\big(\lambda - \lambda\pi(s) + s\big)\Big) - \pi(s) = 0.$$

Applying the inverse function ψ, we obtain

$$\psi(s) = \frac{\lambda - \lambda\pi(s) + s}{1 + r}. \tag{13}$$

Now we will find an expansion for π, which in turn yields an expansion for ψ. Using Eq. (12) and some elementary calculus, we find

$$\pi'(0) = \frac{-\mathbb{E}\, B}{r_1 - \lambda\, \mathbb{E}\, B} \quad \text{and} \quad \pi''(0) = \frac{r_1^2\, \mathbb{E}\, B^2}{\big(r_1 - \lambda\, \mathbb{E}\, B\big)^3}.$$

Recall that the above quantities are finite by the conditions we imposed on the moments of B, and since the loads are assumed to be less than one. Therefore,

$$\pi(s) = 1 - \frac{\mathbb{E}\, B}{r_1 - \lambda\, \mathbb{E}\, B}s + \frac{1}{2}\frac{r_1^2\, \mathbb{E}\, B^2}{\big(r_1 - \lambda\, \mathbb{E}\, B\big)^3}s^2 + o(s^2).$$

Substituting this into Eq. (13) yields

$$\psi(s) = \frac{1}{r}s - \frac{1}{2}\frac{\lambda\, \mathbb{E}\, B^2}{r^2}s^2 + o(s^2).$$

It follows that

$$\psi\big(s\epsilon(r_1-r_2)\big)=\psi\big(s\epsilon(r-\epsilon)\big)=s\epsilon-\frac{1}{r}\Big(s\epsilon^2+\frac{1}{2}\lambda\,\mathbb{E}\,B^2s^2\epsilon^2\Big)+o(\epsilon^2).$$

Noting that $\lambda\,\mathbb{E}\,B^2=\sigma^2$, we find the structure of $\psi(s\epsilon(r_1-r_2))$ as in Lemma 3.3.

Regime II

In the following we consider the corresponding Regime II result. It should be noted that the methodology is similar to that for Regime I. However, since the ϵ now plays a different role, we cannot use Lemma 3.3, but we develop Lemma 3.7 instead.

Proposition 3.5 *Let the input process $J\in\mathcal{S}^+$ be such that* $\mathrm{Var}\,J_1=\sigma^2<\infty$. *Then, in Regime II, the joint scaled workload is given by*

$$\lim_{\epsilon\downarrow0}\mathbb{E}\,e^{-s_1\epsilon Q^{(1)}-s_2\epsilon Q^{(2)}}=\frac{s_2\big(\gamma-2-s_1\sigma^2\big)-s_1\gamma+(s_2-s_1)\sigma\sqrt{\frac{\gamma^2}{\sigma^2}+2(\gamma-1)s_2}}{(s_2\sigma^2+2)\Big((\gamma-1)s_2-\gamma s_1-\frac{1}{2}s_1^2\sigma^2\Big)}.$$

Remark 3.6 Note that the result in Proposition 3.5 corresponds to Eq. (5), i.e. the LST we found in case of Brownian input, except now we do take a proper heavy traffic limit, whereas Eq. (5) holds for all $\epsilon>0$.

Lemma 3.7 *Let*

$$\phi(s)=s\epsilon+\frac{1}{2}s^2+K_1s^{\eta_1}+o(s^{\eta_1}),$$

for some constant $K_1\in\mathbb{R}$ and $2<\eta_1\le3$. Then, asymptotically for $\epsilon\downarrow0$, we have

$$\psi\big(s\epsilon^2(\gamma-1)\big)=-\epsilon+\epsilon\sqrt{1+2s(\gamma-1)}+o(\epsilon).$$

Proof of Lemma 3.7 Suppose that

$$\psi\big(s\epsilon^2(\gamma-1)\big)=-\epsilon+\epsilon\sqrt{1+2s(\gamma-1)}+K_2(s)\epsilon^{2\eta_2},$$

for some function K_2 of s (and independent of ϵ) and for some constant η_2. If we show that

$$\lim_{\epsilon\downarrow0}\frac{\phi\Big(\psi\big(s\epsilon^2(\gamma-1)\big)\Big)}{s\epsilon^2(\gamma-1)}=1$$

and $\eta_2\ge\frac{1}{2}$, then we have proved the lemma.

Indeed, for all $s \geq 0$,

$$\begin{aligned}\phi\Big(\psi\big(s\epsilon^2(\gamma-1)\big)\Big)-s\epsilon^2(\gamma-1) &= -\epsilon^2+\epsilon^2\sqrt{1+2s(\gamma-1)}+K_2(s)\epsilon^{2\eta_2+1}+o(\epsilon^{2\eta_1})\\ &\quad+\frac{1}{2}\Big(-\epsilon+\epsilon\sqrt{1+2s(\gamma-1)}+K_2(s)\epsilon^{2\eta_2}\Big)^2\\ &\quad+K_1\Big(-\epsilon+\epsilon\sqrt{1+2s(\gamma-1)}+K_2(s)\epsilon^{2\eta_2}\Big)^{\eta_1}\\ &\quad-s\epsilon^2(\gamma-1).\end{aligned}$$

Hence, after simplification, we see that the following should hold for all $s \geq 0$:

$$\begin{aligned}&\frac{1}{2}K_2(s)^2\epsilon^{4\eta_2}+K_2(s)\epsilon^{2\eta_2+1}\sqrt{1+2s(\gamma-1)}\\ &\quad+K_1\Big(-\epsilon+\epsilon\sqrt{1+2s(\gamma-1)}+K_2(s)\epsilon^{2\eta_2}\Big)^{\eta_1}+o(\epsilon^{2\eta_1})=0.\end{aligned}\tag{14}$$

Case 1 If $\eta_2 < \frac{1}{2}$, then using (14) we obtain for all $s \geq 0$,

$$\frac{1}{2}K_2(s)^2\epsilon^{4\eta_2}+K_1K_2(s)^{\eta_1}\epsilon^{2\eta_1\eta_2}+o(\epsilon^{2\eta_1})=0,$$

which holds for $\epsilon \downarrow 0$ if and only if $4\eta_2 = 2\eta_1\eta_2$. This implies $\eta_1 = 2$, but this contradicts $\eta_1 > 2$. We conclude that $\eta_2 \geq \frac{1}{2}$.

Case 2 If $\eta_2 = \frac{1}{2}$, then we can write

$$\frac{1}{2}K_2(s)^2\epsilon^2+K_2(s)\epsilon^2\sqrt{1+2s(\gamma-1)}+o(\epsilon^2)=0,$$

which is solved by $K_2 = 0$. Note that the conclusion of the lemma holds in this case.

Case 3 If $\eta_2 > \frac{1}{2}$, then we can write for all $s \geq 0$,

$$K_2(s)\epsilon^{2\eta_2+1}\sqrt{1+2s(\gamma-1)}+K_1\epsilon^{\eta_1}\Big(-1+\sqrt{1+2s(\gamma-1)}\Big)^{\eta_1}+o\Big(\epsilon^{\min\{4\eta_2,\eta_1\}}\Big)=0.$$

We see that we have to make sure that $2\eta_2+1=\eta_1$, since the equation has to hold for all $s \geq 0$. So define $\eta_2 = \frac{1}{2}(\eta_1-1)$. Then, for all $s \geq 0$,

$$K_2(s)\sqrt{1+2s(\gamma-1)}+K_1\Big(-1+\sqrt{1+2s(\gamma-1)}\Big)^{\eta_1}+o(1)=0.$$

The conclusion of the lemma holds in this case, noting that $\eta_2 > \frac{1}{2}$, where

$$K_2(s) := -\frac{K_1\Big(-1+\sqrt{1+2s(\gamma-1)}\Big)^{\eta_1}}{\sqrt{1+2s(\gamma-1)}}.$$

This proves the claim. □

Table 1 The values in this table correspond to the left and right plot in Figure 3

	Figure 3, left plot			Figure 3, right plot			
x	Simul	R1	R2	x	Simul	R1	R2
1	0.975	0.990	0.984	1	0.884	0.990	0.888
20	0.800	0.817	0.810	5	0.750	0.951	0.753
40	0.653	0.668	0.662	10	0.654	0.904	0.656
80	0.436	0.446	0.442	15	0.584	0.859	0.585
100	0.356	0.364	0.362	20	0.527	0.817	0.528
150	0.215	0.220	0.219	25	0.480	0.777	0.480
200	0.129	0.133	0.132	30	0.439	0.739	0.439
250	0.077	0.080	0.080	35	0.403	0.702	0.403
300	0.047	0.048	0.049	40	0.372	0.668	0.372
400	0.017	0.018	0.019	45	0.344	0.635	0.343
500	0.006	0.006	0.008	50	0.318	0.603	0.318

Simul stands for the simulated values, and R1, R2 stand for the approximated values using Regime I, II, respectively

Proof of Proposition 3.5 This result follows from Lemma 3.7 and Theorem 2.2, and taking the limit $\epsilon \downarrow 0$. The calculations are similar to those in the Brownian case, except there are some additional terms of small order ϵ that cancel in the heavy traffic limit. □

3.3 Numerical approximations for exponential jobs

Example 3.8 (Comparison of Regime I and Regime II) Suppose that we have a system with compound Poisson input with exponential jobs, with $\lambda = 1$, $\mu = 1$. By Eq. (3), one obtains the Regime I approximation $Q^{(2)} \stackrel{\mathrm{d}}{=} \mathrm{Exp}(\epsilon)$. Due to Eq. (7), the Regime II approximation entails numerically inverting

$$\mathbb{E}\, e^{-sQ^{(2)}} = \frac{1}{\gamma - 1} \frac{-2 + \gamma + \sqrt{\gamma^2 + 2\frac{s}{\epsilon}\sigma^2(\gamma - 1)}}{2 + \frac{s}{\epsilon}\sigma^2}.$$

In addition, we estimated the probabilities by simulation. The results are gathered in Tables 1 and 2, and are plotted in Figs. 3 and 4. Observe from Fig. 3 that the Regime II approximation is substantially more accurate than the Regime I approximation when ρ_2 is high (in this case $\rho_2 = 0.99$). By comparing the two plots in Fig. 3, we see that increasing ρ_1 negatively affects the performance of the Regime I approximation. Figure 4 shows that the Regime II approximation works remarkably well even when relatively low loads are imposed on both servers. Our experiments reveal that it is only reasonable to use Regime I approximations in a tandem queue when the load of the first server ρ_1 is low; in all other cases, it is outperformed by the Regime II approximation. If ρ_1 is high, then there is a stronger dependence between the up- and downstream workloads (cf. Eq. (9), noting that ρ_1 increases as γ decreases). Apparently, the dependence between both workloads, which is ignored in Regime I, has a crucial impact.

Table 2 The values in this table correspond to the left and right plot in Figure 4

	Figure 4, left plot			Figure 4, right plot			
x	Simul	R1	R2	x	Simul	R1	R2
1	0.498	0.777	0.551	0.5	0.719	0.945	0.761
2	0.362	0.605	0.387	1.0	0.630	0.894	0.665
3	0.273	0.471	0.283	1.5	0.565	0.846	0.594
4	0.210	0.367	0.210	2.0	0.512	0.800	0.537
5	0.163	0.286	0.157	2.5	0.468	0.757	0.489
6	0.128	0.223	0.119	3.0	0.429	0.716	0.447
7	0.101	0.173	0.090	3.5	0.395	0.677	0.410
8	0.080	0.135	0.069	4.0	0.365	0.640	0.378
9	0.064	0.105	0.053	4.5	0.338	0.606	0.349
10	0.051	0.082	0.040	5.0	0.313	0.573	0.323

The abbreviations are as in Table 1

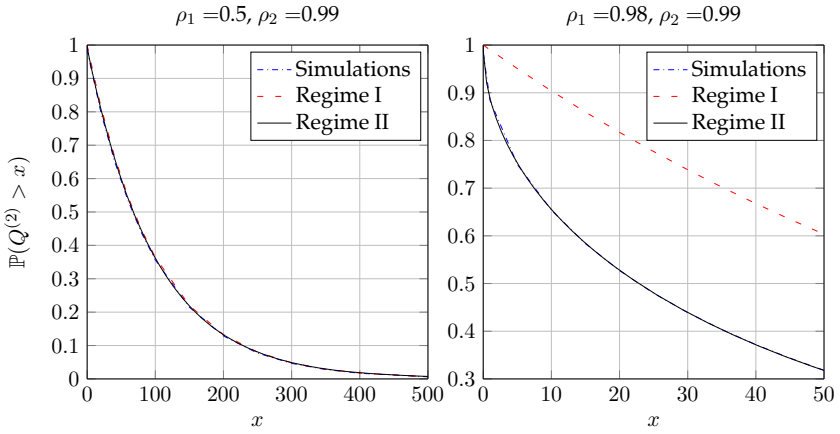

Fig. 3 Varying ρ_1 while keeping $\rho_2 = 0.99$. It appears that the Regime II approximation is almost perfect, and the Regime I approximation becomes worse the higher ρ_1 becomes

4 Heavy-tailed input

In this section we consider spectrally positive Lévy input processes with increments that have infinite variance. Unlike in the finite variance case, the precise form of the heavy traffic limit depends on the specific features of the Lévy input process. In Sect. 4.1, we consider compound Poisson input with heavy-tailed jumps, and in Sect. 4.2, we consider α-stable Lévy input (where $1 < \alpha < 2$). Note that α-stable Lévy motion can be regarded as a generalization of Brownian motion. Indeed, for $\alpha = 2$, an α-stable Lévy motion reduces to a Brownian motion.

Remark 4.1 We only consider Regime I results, because we have not managed to compute Regime II results here. In the finite variance case we relied on the existence of the inverse function of ϕ in the Brownian case to construct $\psi(s\epsilon(r_1 - r_2))$

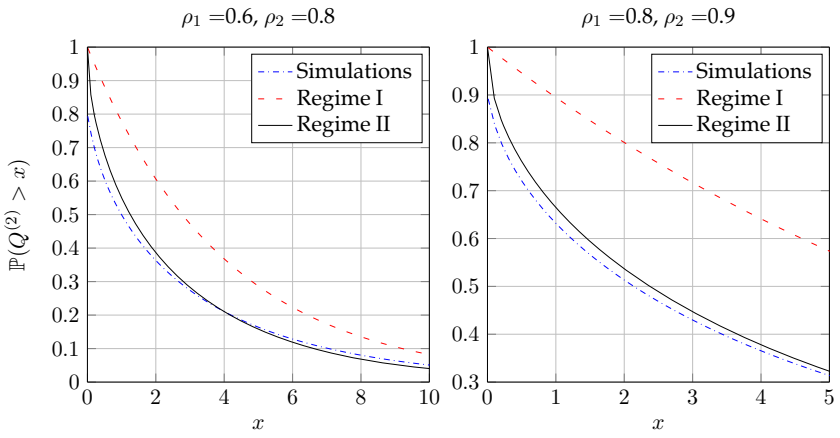

Fig. 4 In the above two plots, we see that again the Regime II approximation works significantly better than the Regime I approximation. Even for $\rho_1 = 0.6$ and $\rho_2 = 0.8$, the Regime II approximation shows remarkably good fit

as in Lemma 3.7. However, for heavy-tailed input, there is in general no inverse function of ϕ available, except for some special cases, such as $\frac{3}{2}$-stable Lévy motion.

Before we state the result, we introduce some notation. We write

$$E_\alpha(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(\alpha n + 1)}$$

for the Mittag-Leffler function with parameter α. Random variables that have a distribution function $1 - E_\alpha(x)$ are called *Mittag-Leffler distributed* with parameter α. Suppose that M is Mittag-Leffler distributed with parameter α, then the LST of M is given by

$$\mathbb{E}\, e^{-sM} = \frac{1}{1 + s^\alpha}.$$

Furthermore, suppose a measurable function L defined on some neighbourhood of ∞, $[x, \infty)$, $x \in \mathbb{R}$, satisfies

$$\lim_{x\to\infty} \frac{L(ax)}{L(x)} = 1, \quad \forall a > 0,$$

then it is called a slowly varying function [13]. For notational brevity, we sometimes write $f(x) \sim g(x)$ (as $x \to \infty$) to denote $\lim_{x\to\infty} f(x)/g(x) = 1$, for generic functions f, g.

4.1 Compound Poisson

In this section we consider spectrally positive compound Poisson input processes with heavy-tailed jumps.

Remark 4.2 In [11] a heavy traffic problem for heavy-tailed input was studied in a GI/G/1 setting. In their paper the correct scaling function $\Delta(\epsilon)$ was also found by letting it be the zero of an appropriate equation. We follow a similar approach.

Proposition 4.3 *Let the input process $J \in \mathcal{S}^+$ to the first queue be a compound Poisson process with heavy-tailed service requirements, that is, the distribution of the service requirement B satisfies*

$$\mathbb{P}(B > x) \sim x^{-\nu} L(x), \quad as \quad x \to \infty, \tag{15}$$

where L is some slowly varying function. Suppose that the load of the first queue is fixed and the load of the second queue is increasing to one as $\epsilon \downarrow 0$. For $\epsilon > 0$ small enough, there is a unique solution $s = \Delta(\epsilon)$ to

$$-\lambda\Gamma(1-\nu)\frac{(r-\epsilon)^{\nu}}{r^{\nu+1}} s_2^{\nu-1} L(1/s_2) = \epsilon,$$

such that $\Delta(\epsilon) \downarrow 0$. It holds that

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\, e^{-s_1 Q^{(1)} - s_2 \Delta(\epsilon) Q^{(2)}} = \frac{r s_1}{\phi(s_1)} \cdot \frac{1}{1 + s_2^{\nu-1}}.$$

Proof Suppose the input process J is of the compound Poisson type. More precisely, we have a Poisson process N with rate λ and we assume

$$J_t = \sum_{k=1}^{N(t)} B_k, \text{ where the } B_k \text{ are i.i.d., independent of } N(t), \text{ and such that } \mathbb{E}\, J_1 = 1.$$

Then the cumulative net input processes for the first server and the whole system ($i = 1, 2$, respectively) are defined by

$$X_t^{(i)} = \sum_{k=1}^{N(t)} B_k - r_i t.$$

Suppose we have a compound Poisson input process, then $\phi(s) = sr_1 - \lambda + \lambda b(s)$, where $b(s) = \mathbb{E}\, e^{-sB}$ [cf. Eq. (11)]. Suppose the service time B is regularly varying, with index $1 < \nu < 2$. Then it takes the form of Eq. (15). By applying Theorem 5.1,

$$b(s) - 1 - c_1 s \sim -\Gamma(1-\nu) s^{\nu} L(1/s) \quad \text{as} \quad s \downarrow 0,$$

with $1 < \nu < 2$. Substitution yields $\phi(s) \sim (\lambda c_1 + r_1)s - \lambda\Gamma(1-\nu)s^\nu L(1/s)$. We assumed $\lambda\,\mathbb{E}\,B = 1$, so $b'(0) = -\frac{1}{\lambda} = c_1$. Recall that $r_1 - 1 = r$, and therefore

$$\phi(s) - rs \sim -\lambda\Gamma(1-\nu)s^\nu L(1/s).$$

By Lemma 9.2 from [10] (see also Lemma 5.2 in the Appendix), we find

$$\psi(s) - \frac{1}{r}s \sim \lambda\Gamma(1-\nu)\frac{1}{r^{\nu+1}}s^\nu L(1/s) \quad \text{as} \quad s \downarrow 0. \tag{16}$$

We now identify a scaling function $\Delta(\epsilon)$ such that we have convergence to a non-degenerate distribution. By making use of Eq. (16) and by scaling the workload of the downstream queue by a function $\Delta(\epsilon)$, for which $\Delta(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$, we obtain

$$\begin{aligned}\mathbb{E}\,e^{-s_1 Q^{(1)} - s_2\Delta(\epsilon)Q^{(2)}} \sim{}& \frac{1}{1 + \frac{1}{\epsilon}C(r-\epsilon)^\nu s_2^{\nu-1}\Delta(\epsilon)^{\nu-1}L(\frac{1}{s_2\Delta(\epsilon)})} \\ &\times \frac{(r-\epsilon)s_2\Delta(\epsilon) - Cs_2^\nu(r-\epsilon)^\nu\Delta(\epsilon)^\nu L\big(\frac{1}{s_2\Delta(\epsilon)(r-\epsilon)}\big) - rs_1}{(r-\epsilon)s_2\Delta(\epsilon) - \phi(s_1)},\end{aligned} \tag{17}$$

where $C := -\lambda\Gamma(1-\nu)r^{-\nu-1}$, for $s_1, s_2 \geq 0$ fixed and $\epsilon \downarrow 0$. Consider the equation

$$C(r-\epsilon)^\nu s_2^{\nu-1}L(1/s_2) = \epsilon. \tag{18}$$

We will show that this equation has a unique zero for ϵ close enough to zero, and we call the zero $\Delta(\epsilon)$. Indeed, by Theorem 1.5.4 in [13], we have that

$$C(r-\epsilon)^\nu s^{\nu-1}L(1/s) \sim \xi(1/s), \quad s \downarrow 0,$$

where $s \mapsto \xi(s)$ is a non-decreasing function (hence $s \mapsto \xi(1/s)$ non-increasing). So if ϵ is chosen small enough, the s solving Eq. (18) also becomes small and

$$\frac{C(r-\epsilon)^\nu s^{\nu-1}L(1/s)}{\xi(1/s)}\xi(1/s) \approx \xi(1/s),$$

so the left-hand side of Eq. (18) is asymptotically monotone. This ensures that there is exactly one root $\Delta(\epsilon)$ for all $\epsilon > 0$ small enough. Moreover, note that $\Delta(\epsilon)$ indeed satisfies $\Delta(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$.

Therefore, we have

$$C(r-\epsilon)^\nu\Delta(\epsilon)^{\nu-1}L\big(1/(s_2\Delta(\epsilon))\big) = \epsilon. \tag{19}$$

Now consider the first factor on the right-hand side in Eq. (17). Substituting Eq. (19) into this factor,

$$\frac{1}{1+\frac{1}{\epsilon}C(r-\epsilon)^{\nu}s_2^{\nu-1}\Delta(\epsilon)^{\nu-1}L\left(\frac{1}{s_2\Delta(\epsilon)}\right)}=\frac{1}{1+s_2^{\nu-1}\frac{L\left(1/\left(s_2\Delta(\epsilon)\right)\right)}{L(1/s_2)}}\xrightarrow{\epsilon\downarrow 0}\frac{1}{1+s_2^{\nu-1}},$$

where we make use of the fact that L is slowly varying at ∞. Now consider the following part of the second factor in Eq. (17):

$$\begin{aligned}Cs_2^{\nu}(r-\epsilon)^{\nu}\Delta(\epsilon)^{\nu}L\left(\frac{1}{s_2\Delta(\epsilon)(r-\epsilon)}\right)&=\left[C(r-\epsilon)^{\nu}\Delta(\epsilon)^{\nu-1}L\left(\frac{1}{s_2\Delta(\epsilon)}\right)\right]\\&\quad\Delta(\epsilon)\frac{L\left(\frac{1}{s_2\Delta(\epsilon)(r-\epsilon)}\right)}{L\left(\frac{1}{s_2\Delta(\epsilon)}\right)}\\&=\epsilon\Delta(\epsilon)\frac{L\left(\frac{1}{s_2\Delta(\epsilon)(r-\epsilon)}\right)}{L\left(\frac{1}{s_2\Delta(\epsilon)}\right)}\sim\epsilon\Delta(\epsilon),\end{aligned}$$

where we substituted the part between square brackets by making use of Eq. (19) and used that L is slowly varying. By again exploiting the fact that $\Delta(\epsilon)\downarrow 0$ as $\epsilon\downarrow 0$, the result now follows from Eq. (17). □

Example 4.4 Suppose that we are in the setting of Proposition 4.3, but we are in the special case that $\lim_{x\to\infty}L(x)=L\in\mathbb{R}$. Then a correct scaling function is

$$\Delta(\epsilon)=\left(\frac{\epsilon}{\frac{\lambda}{r}\Gamma(1-\nu)L}\right)^{\frac{1}{\nu-1}}.$$

Proposition 4.3 can be used to find a heavy traffic approximation as follows. We have

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\,e^{-s\Delta(\epsilon)Q^{(2)}}=\frac{1}{1+s^{\nu-1}},$$

so that, for $x\geq 0$, and $\epsilon>0$ small,

$$\mathbb{P}\left(\Delta(\epsilon)Q^{(2)}>x\right)\approx E_{\nu-1}(-x^{\nu-1}).$$

By substitution we thus obtain the heavy traffic approximation for $x\geq 0$, and $\epsilon>0$ small,

$$\mathbb{P}(Q^{(2)}>x)\approx E_{\nu-1}\Big(-\Delta(\epsilon)^{\nu-1}x^{\nu-1}\Big).$$

4.2 α-Stable Lévy motion

In this subsection we prove the following result. It entails that the workloads are asymptotically independent in the heavy traffic limit and that the marginals correspond to scaled Mittag-Leffler distributed random variables.

Proposition 4.5 *Let the input process $J \in \mathcal{S}^+$ to the first queue be a spectrally positive α-stable Lévy motion, with $1 < \alpha < 2$. Suppose that the load of the first queue is fixed and the load of the second queue is increasing as $\epsilon \downarrow 0$ and scaled by ϵ^β, with $\beta := (\alpha - 1)^{-1}$. It holds that*

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\, e^{-s_1 (C/r)^\beta Q^{(1)} - s_2 (\epsilon C)^\beta Q^{(2)}} = \frac{1}{1 + s_1^{\alpha-1}} \cdot \frac{1}{1 + s_2^{\alpha-1}},$$

with $C := (\cos(\pi(\frac{\alpha}{2} - 1)))^{-1}$.

Proof of Proposition 4.5 The Laplace exponent is given by $\phi(s) = (r_1 - 1)s + Cs^\alpha$. It follows by Lemma 9.2 from [10], that

$$\psi(s) = c_1 + c_2 s - \frac{C}{r_1 - 1}\left(\frac{s}{r_1 - 1}\right)^\alpha + o(s^\alpha).$$

We know that $\psi(0) = 0$, hence $c_1 = 0$, and $\psi'(0) = \frac{1}{\phi'(0)} = \frac{1}{r_1 - 1}$, hence $c_2 = \frac{1}{r_1 - 1}$. This leads to

$$\psi(s) = \frac{s}{r_1 - 1} - \frac{C}{r_1 - 1}\left(\frac{s}{r_1 - 1}\right)^\alpha + o(s^\alpha).$$

It follows from Theorem 2.2 that

$$\begin{aligned}\mathbb{E}\, e^{-s_1 Q^{(1)} - \epsilon^{\frac{1}{\alpha-1}} s_2 Q^{(2)}} = &\frac{\epsilon s_2 \epsilon^{\frac{1}{\alpha-1}} \left(1 + o\left(s^\alpha \epsilon^{\frac{\alpha}{\alpha-1}}\right)\right)}{\epsilon^{\frac{1}{\alpha-1}} s_2 - \epsilon^{\frac{1}{\alpha-1}} \frac{s_2(r-\epsilon)}{r} + \frac{C}{r}\left(\frac{s_2(r-\epsilon)}{r}\right)^\alpha \epsilon^{\frac{\alpha}{\alpha-1}}} \\ &\times \frac{\frac{1}{r}\epsilon^{\frac{1}{\alpha-1}} s_2 (r - \epsilon) - \frac{C}{r}\left(\frac{s_2(r-\epsilon)}{r}\right)^\alpha \epsilon^{\frac{\alpha}{\alpha-1}} - s_1}{\epsilon^{\frac{1}{\alpha-1}}(r - \epsilon)s_2 - r s_1 - C s_1^\alpha}.\end{aligned}$$

Consequently,

$$\lim_{\epsilon \downarrow 0} \mathbb{E}\, e^{-s_1 Q^{(1)} - \epsilon^{\frac{1}{\alpha-1}} s_2 Q^{(2)}} = \frac{r}{r + C s_1^{\alpha-1}} \frac{1}{1 + C s_2^{\alpha-1}}, \tag{20}$$

which implies the claim. □

In the case $\alpha = \frac{3}{2}$, ψ can be calculated explicitly and the result can be obtained without the use of Tauberian theorems. We include this in the paper, as the calculations potentially contain clues as to how Regime II results can be eventually obtained.

Example 4.6 (Explicit calculations for $\alpha = \frac{3}{2}$) We assume a $\frac{3}{2}$-stable input process, so that the Laplace exponent is given by

$$\phi(s) = rs + \frac{1}{\cos\left(\pi(\frac{\alpha}{2} - 1)\right)} s^{\frac{3}{2}} = rs + \sqrt{2} s\sqrt{s}.$$

Define

$$R(s) := -\frac{r^3}{54\sqrt{2}} + \sqrt{\frac{1}{8}s^2 - \frac{sr^3}{108}} + \frac{s}{2\sqrt{2}}. \tag{21}$$

By making a substitution $s^2 \leftarrow s$, ϕ turns into a third-order polynomial, which can be inverted by using Cardano's formula. It follows that the inverse function of ϕ is given by

$$\psi(s) = \left(R(s)^{\frac{1}{3}} + \frac{r^2}{18R(s)^{\frac{1}{3}}} - \frac{r}{3\sqrt{2}} \right)^2. \tag{22}$$

Note that $s = \phi(\psi(s)) = r\psi(s) + \sqrt{2}\psi(s)^{\frac{3}{2}}$. Define the function ζ such that $\zeta(s)^2 = \psi(s)$. Then $\psi(s) = \zeta(s)^2 = r^{-1}(s - \sqrt{2}\zeta(s)^3)$. So

$$\psi\left(s\epsilon^2(r-\epsilon)\right) = \frac{1}{r} s\epsilon^2(r-\epsilon) - \frac{\sqrt{2}}{r} \zeta\left(s\epsilon^2(r-\epsilon)\right)^3. \tag{23}$$

Now we can focus on

$$\zeta\left(s\epsilon^2(r-\epsilon)\right) = R\left(s\epsilon^2(r-\epsilon)\right)^{\frac{1}{3}} + \frac{r^2}{18R\left(s\epsilon^2(r-\epsilon)\right)^{\frac{1}{3}}} - \frac{r}{3\sqrt{2}}, \tag{24}$$

where we can simplify by constructing the Taylor series of $R^{\frac{1}{3}}$. First note that

$$R\left(s\epsilon^2(r-\epsilon)\right) = -\frac{r^3}{54\sqrt{2}} + \sqrt{\frac{1}{8}s^2\epsilon^4(r-\epsilon)^2 - \frac{s\epsilon^2(r-\epsilon)r^3}{108}} + o(\epsilon).$$

By rewriting this and using Taylor expansions for the square roots, neglecting all terms of smaller order than ϵ, we obtain

$$R\left(s\epsilon^2(r-\epsilon)\right) = -\frac{r^3}{54\sqrt{2}} + \sqrt{\frac{r^4 s}{108}}\epsilon i\sqrt{1 + o(\epsilon)} + o(\epsilon) = -\frac{r^3}{54\sqrt{2}}(1 - g\epsilon) + o(\epsilon),$$

where i denotes $\sqrt{-1}$ and we defined $g := \frac{3\sqrt{6}}{r}\sqrt{s}i$. Again using a Taylor expansion, we find

$$R\big(s\epsilon^2(r-\epsilon)\big)^{\frac{1}{3}} = (-1)^{\frac{1}{3}}\frac{r}{3\sqrt{2}}\Big(1-\frac{1}{3}g\epsilon\Big)+o(\epsilon),$$
$$R\big(s\epsilon^2(r-\epsilon)\big)^{-\frac{1}{3}} = (-1)^{-\frac{1}{3}}\frac{3\sqrt{2}}{r}\Big(1+\frac{1}{3}g\epsilon\Big)+o(\epsilon).$$

Substituting this into Eq. (24) yields

$$\zeta\big(s\epsilon^2(r-\epsilon)\big) = \frac{rg\epsilon}{9\sqrt{2}}\Big(-(-1)^{\frac{1}{3}}+(-1)^{-\frac{1}{3}}\Big) = \frac{rg\epsilon}{9\sqrt{2}}\big(-\sqrt{3}i\big)+o(\epsilon),$$

by making use of $(-1)^{\frac{1}{3}} = e^{i\pi/3} = \frac{1}{2}+\frac{1}{2}\sqrt{3}i$ and $(-1)^{-\frac{1}{3}} = e^{-i\pi/3} = \frac{1}{2}-\frac{1}{2}\sqrt{3}i$. Recalling the definition of g, we find $\zeta(s\epsilon^2(r-\epsilon)) = \epsilon\sqrt{s}+o(\epsilon)$. Substituting this into Eq. (23) yields $\psi(s\epsilon^2(r-\epsilon)) = s\epsilon^2-(1+\sqrt{2s})\frac{s\epsilon^3}{r}+o(\epsilon^3)$. It can be verified that terms of smaller magnitudes do not contribute to the heavy traffic version of Corollary 2.3. Using this corollary yields

$$\lim_{\epsilon\downarrow 0}\mathbb{E}\,e^{-s\epsilon^2Q^{(2)}} = \lim_{\epsilon\downarrow 0}\frac{1-\big(1+\sqrt{2s}\big)\frac{\epsilon}{r}+o(\epsilon)}{1+\sqrt{2s}+o(1)} = \frac{1}{1+\sqrt{2s}},$$

which corresponds to Eq. (20) with $\alpha = \frac{3}{2}$ (and here we considered $s_1 = 0$).

4.3 Numerical heavy traffic approximations

Suppose the tandem system is fed by a compound Poisson input process with jobs that are Pareto distributed. In this case the slowly varying function from Proposition 4.3 is actually a constant. In Example 4.4, we obtained the corresponding heavy traffic approximation. Figure 5 facilitates a comparison between estimates obtained from simulations and the Mittag-Leffler (Regime I) heavy traffic approximation. As expected, we see that as ρ_2 increases the heavy traffic approximation becomes more accurate, by comparing the left plot (where $\rho_2 = 0.95$) to the right plot (where $\rho_2 = 0.99$). We show the plotted values in Table 3, along with the relative difference between the two values.

5 Discussion and concluding remarks

In this paper we considered two types of heavy traffic regimes for a two-node fluid tandem queue with spectrally positive Lévy input. In Regime I, only the second server experiences heavy traffic. In this case, the load of the first server has no influence on the steady-state distribution of the workload in the second server. In Regime II, where both servers experience heavy traffic, the dependence structure between both workloads is preserved. In the case where the increments of the Lévy input process

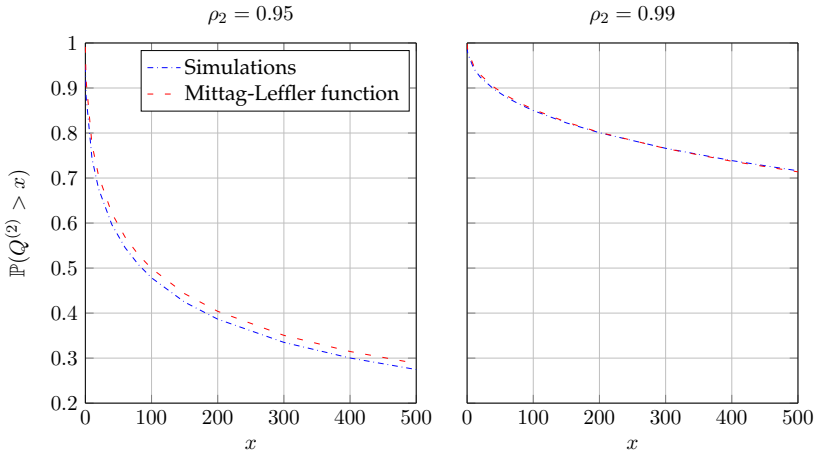

Fig. 5 Using the Mittag-Leffler function as an approximation. We simulated 288 sample paths each consisting of $50 \cdot 10^6$ arrivals of Pareto distributed jobs. In both cases, we used $\lambda = 1$, $\nu = 1.5$, $\rho_1 = \frac{1}{2}$, and we only varied ρ_2 as indicated above the plots

Table 3 This table corresponds to the left and right plot in Fig. 5

	$\rho_2 = 0.95$			$\rho_2 = 0.99$		
x	Simul	M-L	diff (%)	Simul	M-L	diff (%)
10	0.744	0.775	4.2	0.943	0.949	0.64
20	0.676	0.705	4.3	0.924	0.929	0.54
40	0.597	0.622	4.2	0.900	0.903	0.33
60	0.546	0.569	4.2	0.879	0.883	0.46
80	0.508	0.530	4.3	0.863	0.867	0.46
100	0.478	0.499	4.4	0.850	0.853	0.35
150	0.424	0.443	4.5	0.823	0.824	0.12
200	0.387	0.404	4.4	0.801	0.802	0.12
300	0.335	0.351	4.8	0.766	0.766	0.00
400	0.300	0.315	5.0	0.739	0.737	−0.27
500	0.274	0.288	5.1	0.716	0.714	−0.28

The columns Simul and M-L show the probabilities $\mathbb{P}(Q^{(2)} > x)$, for the simulated sample paths and the heavy traffic approximation from Example 4.4, respectively. The last column shows the relative difference between the two values, that is, diff equals (M-L − Simul)/ Simul ·100 %

have finite variance, we have obtained Regime I and II results, whereas for the infinite variance case we established Regime I results.

The numerical experiments led to the interesting insight that (for finite variance input processes) the Regime II approximation performs typically better than the Regime I approximation, particularly when the load of the first server is high as well. This leads us to wonder if results of this kind carry over to a more general setting.

An open problem concerns Regime II results in the case where the increments of the input process have infinite variance. It is not clear how such results can be established. In the finite variance case we could define an inverse Laplace exponent that was in line with the exact inverse for Brownian motion. However, in the case of

heavy-tailed input, for example for α-stable Lévy motion, there is no explicit inverse Laplace exponent for all $1 < \alpha < 2$, and hence a fundamentally different approach needs to be developed.

Another direction for further research concerns stochastic-process limits. In the single-node case there is convergence to reflected Brownian motion (in the finite variance case) and to a reflected stable process (in the infinite variance case), and the question is whether we can establish the counterpart of such results for the downstream node in a tandem system, or even for the joint distribution of both workloads.

Acknowledgments The research for this paper is partly funded by the NWO Gravitation Project NETWORKS, Grant Number 024.002.003. The research of Onno Boxma was also partly funded by the Belgian Government, via the IAP Bestcom Project.

Appendix: Useful Tauberian results

We turn to a law F defined on $[0, \infty)$. We study the LST $\hat{F}$. Write

$$\mu_n := \mathbb{E}\, X^n = \int_{[0,\infty)} x^n dF(x) \quad (n = 0, 1, \ldots)$$

for the n-th moment. When $\mu_n < \infty$, $\hat{F}(s)$ may be expanded in a Taylor series at zero as far as the s^n term:

$$\hat{F}(s) = \sum_{r=0}^{n} \mu_r (-s)^r / r!.$$

To relate the tail behaviour of F to the behaviour of $\hat{F}$ at zero, one needs to eliminate the polynomial $\sum_{r=0}^{n} \mu_r (-s)^r / r!$, which can be done by subtraction or differentiation. This leads to the following definitions:

$$f_n(s) := (-1)^{n+1} \left(\hat{F}(s) - \sum_{r=0}^{n} \mu_r (-s)^r / r! \right),$$

$$g_n(s) := \frac{d^n f_n(s)}{ds^n} = \mu_n - (-1)^n \hat{F}^{(n)}(s),$$

thus $f_0(s) = g_0(s) = 1 - \hat{F}(s)$. Now we are ready to state the following important theorem.

Theorem 5.1 (Theorem 8.1.6 in [13]) *Let L be a slowly varying function, $\mu_n < \infty$, where $n \in \mathbb{Z}^+$, and $\nu = n + \beta$ with $0 \le \beta \le 1$. Then the following are equivalent:*

- $f_n(s) \sim s^\nu L(1/s)$ *as* $s \downarrow 0$;

- $1 - F(x) \sim \frac{(-1)^n}{\Gamma(1-\nu)} x^{-\nu} L(x)$ *as* $x \to \infty$ *when* $0 < \beta < 1$.

Lemma 5.2 (Lemma 9.2 in [10]) *Let ϕ be a Laplace exponent, such that for its first derivative ϕ', for $s \downarrow 0$,*

$$\phi'(s) \sim \sum_{i=0}^{n-1} c_i s^i + \eta s^{\nu-1} L(1/s),$$

for some constants $c_0, \ldots, c_{n-1}$ with $\nu \in (n, n+1)$, and L a slowly varying function. Then, for ψ, the inverse function of ϕ, it holds that as $s \downarrow 0$:

$$\psi(s) \sim \sum_{i=0}^{n} \hat{c}_i s^i - \frac{\eta}{\nu} \frac{1}{\left(\phi'(0)\right)^{\nu+1}} s^{\nu-1} L(1/s),$$

for some constants $\hat{c}_0, \ldots, \hat{c}_n$.

References

1. Kingman, J.F.C.: On queues in heavy traffic. J. R. Stat. Soc. Ser. B (Methodol.) **24**(2), 383–392 (1962)
2. Prohorov, V.: Transition phenomena in queueing processes. I. Litov. Mat. Sb. **3**, 199–205 (1963)
3. Shneer, S., Wachtel, W.: Heavy-traffic analysis of the maximum of an asymptotically stable random walk. Theory Probab. Appl. **55**, 332–341 (2011)
4. Glynn, P.: Diffusion approximations. In: Heyman, D., Sobel, M. (eds.) Handbooks on Operations Research & Management Science, vol. 2, pp. 145–198. Elsevier, New York (1990)
5. Whitt, W.: Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and Their Application to Queues. Springer, New York (2002)
6. Harrison, J.M.: The diffusion approximation for tandem queues in heavy traffic. Adv. Appl. Probab. **10**(4), 886–905 (1978)
7. Kella, O., Whitt, W.: A tandem fluid network with Lévy input. In: Basawa, I., Bhat, U.N. (eds.) Queueing and Related Models, pp. 112–128. Oxford University Pres, Oxford (1992)
8. Kella, O.: Parallel and tandem fluid networks with dependent Lévy inputs. Ann. Appl. Probab. **3**(3), 682–695 (1993)
9. Dębicki, K., Dieker, A.B., Rolski, T.: Quasi-product forms for Lévy-driven fluid networks. Math. Oper. Res. **32**(3), 629–647 (2007)
10. Dębicki, K., Mandjes, M.: Queues and Lévy Fluctuation Theory. Springer, New York (2015)
11. Boxma, O.J., Cohen, J.W.: Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distributions. Queueing Syst. **33**(1–3), 177–204 (1999)
12. Takács, L.: Introduction to the Theory of Queues. Oxford University Press, New York (1962)
13. Bingham, N.H., Goldie, C.M., Teugels, J.L.: Regular Variation. Cambridge University Press, Cambridge (1987)