Asymptotic behaviour of the tandem queueing system with identical service times at both queues

Citation for published version (APA):
Boxma, O. J., & Deng, Q. (1999). *Asymptotic behaviour of the tandem queueing system with identical service times at both queues*. (Memorandum COSOR; Vol. 9902). Technische Universiteit Eindhoven.

Memorandum COSOR 99-02, 1999, Eindhoven University of Technology

Asymptotic behaviour of the tandem queueing system with identical service times at both queues

O.J. Boxma[1] and Q. Deng

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

Abstract

Consider a tandem queue consisting of two single-server queues in series, with a Poisson arrival process at the first queue and arbitrarily distributed service times, which for any customer are *identical* in both queues. For this tandem queue, we relate the tail behavior of the sojourn time distribution and the workload distribution at the second queue to that of the (residual) service time distribution. As a by-result, we prove that both the sojourn time distribution and the workload distribution at the second queue are regularly varying at infinity of index $1-\nu$ if the service time distribution is regularly varying at infinity of index $-\nu$ ($\nu > 1$). Furthermore, in the latter case we derive a heavy-traffic limit theorem for the sojourn time $S^{(2)}$ at the second queue when the traffic load $\rho \uparrow 1$. It states that, for a particular contraction factor $\Delta(\rho)$, the contracted sojourn time $\Delta(\rho)S^{(2)}$ converges in distribution to the limit distribution $H(\cdot)$ as $\rho \uparrow 1$ where $H(w) = \dfrac{\exp\{-w^{1-\nu}\}}{1+\nu w^{1-\nu}}$.

AMS subject classification: 60K25, 90B22.
Keywords and phrases: tandem queue, identical service times, sojourn time distribution, workload distribution, regular variation, heavy-traffic limit theorem.

1 Introduction

In this paper we consider a queueing system consisting of two single-server queues Q_1 and Q_2 in series with infinite waiting space at each queue. Customers arrive at Q_1 according to a Poisson process; Q_1 is an ordinary $M/G/1$ queue. The special feature of the model is that the service time experienced by any customer in Q_2 is *exactly the same* as the one he experienced in Q_1. We are in particular interested in the asymptotic behavior of the steady-state sojourn time and workload distributions in Q_2, paying special attention to the case of a heavy-tailed service time distribution.
The tandem system with identical service times at both nodes is interesting for some practical communication nets, as it reflects the situation in which a message retains the same length while being transmitted through various communication channels. The two-node case has been studied in detail in [6]. A nice feature of this model is, that it allows explicit expressions for the sojourn time and workload distributions at the second node (without taking recourse to Laplace-Stieltjes transforms). These explicit expressions enable us in the present paper to obtain precise relations between the tail behaviour of the sojourn time and workload distributions, and that of the (residual) service time distribution. Such tail behaviour relations presently receive much attention, because of recent traffic measurements in, a.o., Ethernet Local Area Networks [23],

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

Wide Area Networks [21] and VBR video [3]. Those measurements have revealed that this traffic often exhibits features like *self-similarity* and *long-range dependence.* These phenomena can be modeled by considering fluid or ordinary queues with some heavy-tailed input distribution (see the survey [11] and the forthcoming book [20]). The class of regularly varying distributions with index $-\nu$ where $1 < \nu < 2$ is an important and useful class of heavy-tailed distributions. While the present paper studies tail behaviour of key performance measures in tandem queues for *general* service time distributions, indeed some special attention is paid to the case of service time distributions with a regularly varying tail.

Main contributions of the paper

In the present study we establish direct relations between the tail behaviour of the (residual) service time distribution and the sojourn time and workload distributions at the second queue, see Theorems 4.1 and 5.1. By using those relations we obtain asymptotic results for the sojourn time and workload distributions in the case of a service time distribution with regularly varying tail, see Theorems 4.2 and 5.2. In particular, both the sojourn time distribution and the workload distribution at Q_2 are shown to be regularly varying at infinity of index $1-\nu$, if the service time distribution is regularly varying at infinity of index $-\nu$. Finally a heavy-traffic limit theorem, see Theorem 6.1, is provided. It states that if the service time distribution is regularly varying of index $-\nu$ $(1 < \nu < 2)$, and the traffic load $\rho \uparrow 1$, then the contracted sojourn time $\Delta(\rho)S^{(2)}$ converges in distribution for an appropriately chosen coefficient of contraction $\Delta(\rho)$, and the limit distribution function is given by $H(w) = \dfrac{\exp\{-w^{1-\nu}\}}{1+\nu w^{1-\nu}}$.

Related work

Vinogradov [22] considers a tandem system consisting of an arbitrary number of queues, with identical service times at all queues. He studies the joint steady-state distribution of the sojourn time at the first queue and the total sojourn time at the remaining queues, in the case of heavy traffic. He assumes that the service time distribution has a finite third moment. In [18], a tandem queueing system with identical service times at both nodes is considered for various service disciplines (e.g., FCFS at the first queue and LCFS preemptive resume at the second queue) in the case of heavy traffic. It is assumed that the service time distribution has a finite second moment.

While the tail behaviour of the waiting time, sojourn time and workload distributions with heavy-tailed service time distributions is presently a hot topic in performance analysis, hardly any *network* results have been obtained. Anantharam [1] and Boxma and Dumas [12] obtain results regarding the propagation of long-range dependence in networks of (fluid) queues. Baccelli, Schlegel and Schmidt [2] consider tandem queues with a (Palm) stationary arrival process at the first node and *independent* service times at the various nodes, that have a subexponential distribution in at least one node. They derive lower and upper bounds for the tails of the sojourn time distributions; in some cases, these bounds coincide and hence the precise tail behaviour is established. Huang and Sigman [15] consider tandem queues with renewal input process at the first node and *independent* service times at the various nodes. They obtain several tail results, partly building upon [2]. Particularly in the two-node case, it is shown that if the service time distribution at the second node is subexponential and the service time distribution at the first node has a lighter tail, then the tail behaviour of the waiting time at the second node has the same asymptotics as if it were an ordinary $GI/G/1$ queue in isolation.

The occurrence of service time distributions which are regularly varying at infinity of index $-\nu$ with $1 < \nu < 2$, hence with infinite variance, has recently triggered the study of heavy-traffic behaviour of such queueing systems (such heavy-traffic behaviour had always been studied under the assumption of finite second moments). Heavy-traffic limit theorems for the $GI/G/1$ queue with regularly varying interarrival and/or service time distributions and infinite variance have been obtained in [9, 13], and for the $M/G/1$ queue with priority classes in [10]. In the present paper, such a heavy-traffic limit theorem is obtained for the sojourn time distribution in Q_2.

Organization of the paper
Section 2 summarizes the notation and the main results from [5, 6, 7] that will be used in the sequel. In Section 3 we obtain tail asymptotics for some performance measures for Q_1. These results are used in Sections 4 and 5 to obtain the tail behaviour of, respectively, the sojourn time distribution and workload distribution at Q_2. In Section 6 we derive a heavy-traffic limit theorem for the sojourn time distribution in Q_2, in the case of a regularly varying service time distribution with *infinite* or *finite* variance.

2 The basic equations

First we introduce some notations. λ denotes the arrival intensity, $B(\cdot)$ the service time distribution and $\beta(\cdot)$ the Laplace-Stieltjes Transform (LST) of $B(\cdot)$. Note that when an arbitrary customer arrives at Q_1, his service time is a random variable with distribution $B(\cdot)$; when he enters Q_2, his service time is *identical* to his previous service time in the first queue. We assume that $B(\cdot)$ has a finite first moment β and that the traffic load $\rho = \lambda\beta < 1$. This ensures [6] that steady-state distributions of the sojourn time and workload distributions at both queues exist.

Let $S^{(j)}$ be a random variable with distribution the steady-state distribution of the sojourn time at Q_j, $j = 1, 2$; the sojourn time distributions are denoted by $S^{(j)}(\cdot)$ and their LST by $s^{(j)}(\cdot)$, for $j = 1, 2$. To introduce an explicit expression for $S^{(2)}(\cdot)$, we need the following distributions. $m(\cdot)$ denotes the steady-state distribution function of the supremum, m, of the service times of customers during a busy period of Q_1; $G(\cdot)$ denotes the steady-state distribution function of the supremum, G, of the service times of an arbitrary customer C and of those customers who have arrived before C and belong to the same busy period of Q_1 as C. As shown in [5, 7], $m(w)$ is the unique zero inside the unit circle of the following equation,

$$m(w) = \int_0^w \exp\{-\lambda(1 - m(w))t\}\mathrm{d}B(t), \quad w > 0, \tag{2.1}$$

and $G(w)$ is given by

$$G(w) = (1 - \rho)\frac{1 - m(w)}{1 - B(w)}B(w), \quad w > 0. \tag{2.2}$$

Let X be the supremum of the service times of those customers who arrived before an arbitrary customer C and belong to the same busy period of Q_1 as C; $X = 0$ if C is the first customer during a busy period. Let $X(\cdot)$ be the distribution function of X. Apparently we have

$$G(w) = X(w)B(w), \quad w > 0, \tag{2.3}$$

which in combination with (2.2) implies that

$$X(w) = (1 - \rho)\frac{1 - m(w)}{1 - B(w)}, \quad w > 0. \tag{2.4}$$

$Y(\cdot)$ denotes the steady-state distribution function of the amount of work, Y, in Q_2 at the epoch that a busy cycle of Q_1 starts. From Theorem 6.1 in [6] we know that

$$Y(w) = \exp\left\{-\lambda\int_w^\infty (1 - m(t))\mathrm{d}t\right\}, \quad w > 0. \tag{2.5}$$

Now we turn to the sojourn time distributions. The LST of the sojourn time distribution in the $M/G/1$ queue Q_1 follows immediately from the Pollaczek-Khinchine formula:

$$s^{(1)}(s) = \frac{(1-\rho)\beta(s)}{1-\rho\frac{1-\beta(s)}{\beta s}}. \tag{2.6}$$

A probabilistic reasoning shows that the steady-state sojourn time $S^{(2)}$ at Q_2 is the maximum of two independent random variables with distribution $G(\cdot)$ and $Y(\cdot)$. I.e., cf. Theorem 6.4 in [6]: for $w > 0$,

$$S^{(2)}(w) = G(w)Y(w) = (1-\rho)\frac{1-m(w)}{1-B(w)}B(w)\exp\left\{-\lambda\int_w^\infty(1-m(t))\mathrm{d}t\right\}. \tag{2.7}$$

3 Preliminaries

In this section we investigate the asymptotic behavior of $1-X(w)$, $1-G(w)$ and $1-Y(w)$ for $w\to\infty$. These asymptotics will turn out to play a key role in the asymptotic behaviour of the sojourn time and workload distribution in Q_2, cf. Sections 4 and 5. In this paper we assume that service time is unbounded, i.e., $1-B(w) > 0$ for $w > 0$. In the sequel, $f(w)\sim g(w)$ for $w\to\infty$ denotes $\lim_{w\to\infty}f(w)/g(w) = 1$.

Lemma 3.1

$$1-X(w) = \Pr\{X > w\} \sim \frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t) \quad \textit{for } w\to\infty. \tag{3.1}$$

Proof. Rewrite (2.1) as

$$1-m(w) = 1-B(w) + \int_0^w (1-\exp\{-\lambda(1-m(w))t\})\mathrm{d}B(t). \tag{3.2}$$

It follows from the fact that $X(w)$ is a proper probability distribution, cf. (2.4), that

$$\lim_{w\to\infty}\frac{1-m(w)}{1-B(w)} = \frac{1}{1-\rho}. \tag{3.3}$$

This implies that

$$\lim_{w\to\infty} w(1-m(w)) = 0, \tag{3.4}$$

since $\lim_{w\to\infty} w(1-B(w)) = 0$ which follows from the fact that $B(\cdot)$ has finite first moment. Hence for any $0<\epsilon<1$, if w is sufficiently large, then for $0<t<w$,

$$\begin{aligned}\lambda(1-m(w))t-(1+\epsilon)\frac{\lambda^2}{2}(1-m(w))^2t^2 &< 1-\exp\{-\lambda(1-m(w))t\}\\ &< \lambda(1-m(w))t-(1-\epsilon)\frac{\lambda^2}{2}(1-m(w))^2t^2.\end{aligned} \tag{3.5}$$

For the sake of simplicity, we define

$$F(w) = \frac{\lambda^2(1-m(w))^2}{2(1-B(w))}\int_0^w t^2\mathrm{d}B(t). \tag{3.6}$$

Dividing both sides of (3.2) by $1 - B(w)$ and applying (3.5) gives

$$\begin{aligned} 1 + \lambda\frac{1-m(w)}{1-B(w)}\int_0^w t\mathrm{d}B(t) - (1+\epsilon)F(w) &< \frac{1-m(w)}{1-B(w)} \\ &< 1 + \lambda\frac{1-m(w)}{1-B(w)}\int_0^w t\mathrm{d}B(t) - (1-\epsilon)F(w). \end{aligned}$$

Subtract $\lambda\frac{1-m(w)}{1-B(w)}\int_0^w t\mathrm{d}B(t)$ and multiply by $(1-\rho)/(1-\lambda\int_0^w t\mathrm{d}B(t))$ on both sides of the above inequality to obtain

$$\frac{1-(1+\epsilon)F(w)}{1+\frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t)} < (1-\rho)\frac{1-m(w)}{1-B(w)} < \frac{1-(1-\epsilon)F(w)}{1+\frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t)}. \tag{3.7}$$

Replacing $(1-\rho)(1-m(w))/(1-B(w))$ by $X(w)$, it follows from the above equality that

$$\frac{\frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t) + (1-\epsilon)F(w)}{1+\frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t)} < 1 - X(w) < \frac{\frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t) + (1+\epsilon)F(w)}{1+\frac{\lambda}{1-\rho}\int_w^\infty t\mathrm{d}B(t)}. \tag{3.8}$$

We now investigate the asymptotic behavior of $F(w)$ for $w \to \infty$. We show that the first term in the numerator of the left- and righthand sides of (3.8) dominates the second term. By (3.3) we have for large enough w,

$$\begin{aligned} \frac{F(w)}{\int_w^\infty t\mathrm{d}B(t)} &= \frac{\lambda^2(1-m(w))^2\int_0^w t^2\mathrm{d}B(t)}{2(1-B(w))\int_w^\infty t\mathrm{d}B(t)} \\ &\leq \frac{\lambda^2(1-m(w))^2\int_0^w t^2\mathrm{d}B(t)}{2(1-B(w))^2 w} \\ &\leq \frac{\lambda^2\int_0^w t^2\mathrm{d}B(t)}{2(1-\rho)^2 w} \\ &= \lambda^2\frac{-w^2(1-B(w)) + 2\int_0^w(1-B(t))t\mathrm{d}t}{2(1-\rho)^2 w}. \end{aligned} \tag{3.9}$$

Since

$$\int_0^\infty (1-B(t))\mathrm{d}t < \infty,$$

by the Dominated Convergence Theorem it follows that

$$\lim_{w\to\infty}\frac{1}{w}\int_0^w (1-B(t))t\mathrm{d}t = 0.$$

Combining the above equation and (3.9) yields

$$\lim_{w\to\infty}\frac{F(w)}{\int_w^\infty t\mathrm{d}B(t)} = 0. \tag{3.10}$$

The result follows from the above relation and (3.8). □

From (2.3) we can immediately derive the asymptotic behavior of $1 - G(w)$ for $w \to \infty$.

Lemma 3.2

$$1 - G(w) = \Pr\{G > w\} \sim \frac{\lambda}{1-\rho} \int_w^\infty t \mathrm{d}B(t) \quad \textit{for } w \to \infty. \tag{3.11}$$

Proof. It follows from (2.3) that

$$1 - G(w) = 1 - X(w) + 1 - B(w) - (1 - X(w))(1 - B(w)).$$

By Lemma 3.1, we have

$$\lim_{w\to\infty} \frac{1 - B(w)}{1 - X(w)} = 0.$$

Combining the above two relations yields that

$$1 - G(w) \sim 1 - X(w), \quad w \to \infty,$$

which in combination with Lemma 3.1 implies (3.11). It should be noted that the latter relation can also be written as (with $(B > w)$ the indicator function of the event $B > w$):

$$\Pr\{G > w\} \sim \frac{\lambda}{1-\rho} \mathrm{E}[B(B > w)] \ \text{ for } w \to \infty.$$

□

Lemma 3.3

$$1 - Y(w) = \Pr\{Y > w\} \sim \frac{\rho}{1-\rho} \Pr\{B^* > w\} \quad \textit{for } w \to \infty, \tag{3.12}$$

where $Y(w)$ is given by (2.5) and B^* is the residual service time which has density function $(1 - B(w))/\beta$.

Proof. As seen in (2.4), $(1-\rho)(1-m(w))/(1-B(w))$ is the probability distribution of a proper random variable X, the supremum of the service times of the customers who arrived before an arbitrary customer C in the same busy period of Q_1 as C. Hence, cf. also (3.3), with an arbitrary $\epsilon > 0$ and for w large enough,

$$\left(\frac{1}{1-\rho} - \epsilon\right)(1 - B(w)) \le 1 - m(w) \le \frac{1}{1-\rho}(1 - B(w)).$$

Thus

$$\left(\frac{1}{1-\rho} - \epsilon\right) \int_w^\infty (1 - B(t))\mathrm{d}t \le \int_w^\infty (1 - m(t))\mathrm{d}t \le \frac{1}{1-\rho} \int_w^\infty (1 - B(t))\mathrm{d}t,$$

which implies that

$$\lim_{w\to\infty} \frac{\int_w^\infty (1 - m(t))\mathrm{d}t}{\int_w^\infty (1 - B(t))\mathrm{d}t} = \frac{1}{1-\rho}. \tag{3.13}$$

Hence it follows that

$$\begin{aligned} \lim_{w\to\infty} \frac{1 - Y(w)}{\int_w^\infty (1 - B(t))\mathrm{d}t} &= \lim_{w\to\infty} \left(\frac{1 - Y(w)}{\int_w^\infty (1 - m(t))\mathrm{d}t} \frac{\int_w^\infty (1 - m(t))\mathrm{d}t}{\int_w^\infty (1 - B(t))\mathrm{d}t} \right) \\ &= \frac{\lambda}{1-\rho}. \end{aligned}$$

□

4 Asymptotic behaviour of the sojourn time distribution

In this section we apply the lemmas that were obtained in the previous section to derive the asymptotic behavior of $1 - S^{(2)}(w)$ for $w \to \infty$. Moreover, we show how $1 - S^{(2)}(w)$ behaves for $w \to \infty$ if the service time distribution is regularly varying. In fact, if the service distribution is regularly varying of index $-\nu$ ($\nu > 1$), the sojourn time in the second queue is shown to be regularly varying of index $1 - \nu$, which is one degree higher than that of the service time distribution.

Theorem 4.1

$$1 - S^{(2)}(w) = \Pr\{S^{(2)} > w\} \sim \frac{\lambda}{1-\rho} w \Pr\{B > w\} + \frac{2\rho}{1-\rho} \Pr\{B^* > w\} \quad \textit{for } w \to \infty. \tag{4.1}$$

Proof. Since $\lim_{w\to\infty} w(1 - B(w)) = 0$, it follows that

$$\int_w^\infty t \mathrm{d}B(t) = w(1 - B(w)) + \int_w^\infty (1 - B(t))\mathrm{d}t. \tag{4.2}$$

By applying Lemmas 3.2 and 3.3, it follows from (2.7) that

$$\begin{aligned} 1 - S^{(2)}(w) &\sim 1 - G(w) + 1 - Y(w) \\ &\sim \frac{\lambda}{1-\rho} \int_w^\infty t\mathrm{d}B(t) + \frac{\rho}{1-\rho} \int_w^\infty \frac{1 - B(t)}{\beta} \mathrm{d}t \\ &= \frac{\lambda}{1-\rho} w \Pr\{B > w\} + \frac{2\rho}{1-\rho} \Pr\{B^* > w\}, \quad w \to \infty, \end{aligned} \tag{4.3}$$

where the last equation follows from (4.2). □

Theorem 4.1 gives a precise expression for the tail behaviour of the sojourn time distribution at Q_2 into the service time distribution, for arbitrary service time distributions. Below we specify this sojourn time tail behaviour (and that of m, G, X and Y) for the case of a regularly varying service time distribution, a case that presently receives much attention in the literature of communication network performance. We refer to [4] for an excellent discussion of regular variation in probability theory. A probability distribution $F(\cdot)$ of a non-negative random variable is said to be regularly varying at infinity of index $-\nu$ if, for all $x > 0$, $(1 - F(xt))/(1 - F(t)) \to x^{-\nu}$, $t \to \infty$. When $\nu = 0$, one speaks of a slowly varying function; in the sequel, $L(\cdot)$ denotes a slowly varying function.

Theorem 4.2 *Let $\nu > 1$. If $\Pr\{B > w\}$ is regularly varying at infinity of index $-\nu$, then $\Pr\{m > w\}$ is regularly varying at infinity of index $-\nu$, while $\Pr\{G > w\}$, $\Pr\{X > w\}$, $\Pr\{Y > w\}$ and $\Pr\{S^{(2)} > w\}$ are regularly varying at infinity of index $1 - \nu$. More precisely, if*

$$\Pr\{B > w\} \sim w^{-\nu} L(w), \quad w \to \infty, \tag{4.4}$$

then

$$\Pr\{m > w\} \sim \frac{1}{1-\rho} w^{-\nu} L(w), \quad w \to \infty, \tag{4.5}$$

$$\Pr\{X > w\} \sim \Pr\{G > w\} \sim \frac{\lambda}{1-\rho} \frac{\nu}{\nu - 1} w^{1-\nu} L(w), \quad w \to \infty, \tag{4.6}$$

$$\Pr\{Y > w\} \sim \frac{\lambda}{1-\rho}\frac{1}{\nu-1}w^{1-\nu}L(w), \quad w \to \infty, \tag{4.7}$$

$$\Pr\{S^{(2)} > w\} \sim \frac{\lambda}{1-\rho}\frac{\nu+1}{\nu-1}w^{1-\nu}L(w), \quad w \to \infty. \tag{4.8}$$

Proof. It follows from the Karamata theorem and the Monotone Density theorem, see [4] Section 1.5.6 and 1.7.3, that $\Pr\{B^* > w\} \sim \frac{w^{1-\nu}}{(\nu-1)\beta}L(w)$ as $w \to \infty$. The results now follow immediately from (3.3), Lemma's 3.1, 3.2, 3.3 and Theorem 4.1. □

Remark 4.1. If the residual service time distribution is a Weibull distribution, viz., $\Pr\{B^* > w\} = \mathrm{e}^{-w^\delta}$, $0 < \delta < 1$, then Theorem 4.1 implies that

$$\Pr\{S^{(2)} > w\} \sim \frac{\rho}{1-\rho}\delta w^\delta \mathrm{e}^{-w^\delta}, \quad w \to \infty. \tag{4.9}$$

In this case, the second term in the righthand side of (4.1) becomes negligible compared to the first one.

Remark 4.2. The Weibull distribution of the previous remark is a subexponential distribution, cf. [4]. Pakes [19] has proven for the $GI/G/1$ queue that, if the residual service time distribution is subexponential, then the tail of the sojourn time distribution is asymptotically equivalent to the tail of the residual service time distribution, up to a multiplicative factor $\rho/(1-\rho)$. So in this subexponential case we have, for the $M/G/1$ queue Q_1:

$$\Pr\{S^{(1)} > w\} \sim \frac{\rho}{1-\rho}\Pr\{B^* > w\}, \quad w \to \infty. \tag{4.10}$$

Hence, in the case of the above-mentioned Weibull distribution, the following holds for the $M/G/1$ queue Q_1:

$$\Pr\{S^{(1)} > w\} \sim \frac{\rho}{1-\rho}\mathrm{e}^{-w^\delta}, \quad w \to \infty,$$

which is less heavy than the tail $\Pr\{S^{(2)} > w\}$ as given in (4.9).
In the case of a regularly varying service time distribution, the above-mentioned asymptotics imply that

$$\Pr\{S^{(2)} > w\} \sim (\nu+1)\Pr\{S^{(1)} > w\}, \quad w \to \infty.$$

5 Asymptotic behaviour of the workload distribution

Let $V^{(2)}$ denote the steady-state workload at Q_2. It is shown in [6] that

$$\Pr\{V^{(2)} < w\} = (1-\rho)Y(w) + \lambda\int_{\eta=0}^{\infty}(1-B(\eta))\frac{1-m(w+\eta)}{1-B(w+\eta)}(1-\rho)Y(w+\eta)\mathrm{d}\eta, \quad w > 0,$$

which can be rewritten as (note that $(1-B(\eta))/\beta$ is the density of the residual service time distribution; in the sequel, B^* and B_1^* will denote independent random variables with this density):

$$\Pr\{V^{(2)} < w\} = (1-\rho)\Pr\{Y < w\} + \rho\Pr\{X < B_1^* + w\}\Pr\{Y < B_1^* + w\}, \quad w > 0. \tag{5.1}$$

Hence

$$\begin{aligned}\Pr\{V^{(2)} > w\} &= (1-\rho)\Pr\{Y > w\} + \rho\Pr\{X > B_1^* + w\} + \rho\Pr\{Y > B_1^* + w\} \\ &- \rho\Pr\{X > B_1^* + w\}\Pr\{Y > B_1^* + w\}, \quad w > 0. \end{aligned} \tag{5.2}$$

Using Lemma 3.3, as $w \to \infty$,

$$\Pr\{V^{(2)} > w\} \sim \rho[\Pr\{B^* > w\} + \Pr\{X > B_1^* + w\} + \Pr\{Y > B_1^* + w\}]. \tag{5.3}$$

Using Lemma 3.3 again, now in the last term of (5.3), we obtain for $w \to \infty$,

$$\Pr\{V^{(2)} > w\} \sim \rho[\Pr\{B^* > w\} + \Pr\{X > B_1^* + w\} + \frac{\rho}{1-\rho}\Pr\{B^* > B_1^* + w\}]. \tag{5.4}$$

Using Lemma 3.1 and (4.2), for $w \to \infty$,

$$\begin{aligned}\Pr\{V^{(2)} > w\} &\sim \rho[\Pr\{B^* > w\} + \frac{2\rho}{1-\rho}\Pr\{B^* > B_1^* + w\} \\ &+ \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta}(w+z)(1-B(w+z))\mathrm{d}z]. \end{aligned} \tag{5.5}$$

Slightly rewriting this result, we have proven the following:

Theorem 5.1

$$\begin{aligned}\Pr\{V^{(2)} > w\} &\sim \rho[\Pr\{B^* > w\} + \frac{2\rho}{1-\rho}\Pr\{B^* > B_1^* + w\} + \frac{\lambda}{1-\rho}w\Pr\{B > B_1^* + w\} \\ &+ \frac{\lambda}{1-\rho}\int_{z=0}^{\infty} \frac{1-B(z)}{\beta}z(1-B(w+z))\mathrm{d}z], \quad \textit{for } w \to \infty. \end{aligned} \tag{5.6}$$

For general service time distributions, we have now expressed the tail behaviour of the distribution of the workload at Q_2 into the (residual) service time distribution. As in the case of Section 4 (the sojourn time distribution), it would be easy to specify the workload tail behaviour for particular service time distributions with an exponential tail. Instead, we now restrict our attention to the case that the service time distribution is *long-tailed*, i.e., $\frac{\Pr\{B > t + u\}}{\Pr\{B > t\}} \to 1$ as $t \to \infty$ for all real u. The class of long-tailed distributions contains the class of subexponential distributions, which in turn contains the class of regularly varying distributions, cf. [4]. It is easy to prove that, if $\Pr\{B > t\}$ is long-tailed and D is any non-negative random variable that is independent of B, then $\frac{\Pr\{B - D > t\}}{\Pr\{B > t\}} \to 1$ as $t \to \infty$. We can apply this rule to replace $\Pr\{B > B_1^* + w\}$ by $\Pr\{B > w\}$ in (5.6). Actually, $\Pr\{B > t\}$ is long-tailed implies that $\Pr\{B^* > w\}$ is long-tailed by using l'Hospital's rule (but the converse is not true in general, cf. [17]). Therefore the second term in the righthand side of (5.6) can be replaced by $\frac{2\rho}{1-\rho}\Pr\{B^* > w\}$. If furthermore $\mathrm{E}B^2 < \infty$, then the last term in the righthand side of (5.6) can be replaced by $\frac{\lambda}{1-\rho}\frac{\mathrm{E}B^2}{2\beta}\Pr\{B > w\}$.

Below we restrict ourselves to the subclass of regularly varying service time distributions. It then follows from Theorem 4.2 that $\Pr\{X > w\}$ is regularly varying, hence long-tailed; hence

$$\Pr\{X > B_1^* + w\} \sim \Pr\{X > w\}, \quad w \to \infty.$$

We can now conclude from (5.4) that the following result holds.

Theorem 5.2 *Let $\nu > 1$. If $\Pr\{B > w\}$ is regularly varying at infinity of index $-\nu$, then $\Pr\{V^{(2)} > w\}$ is regularly varying at infinity of index $1-\nu$. More precisely, if*

$$\Pr\{B > w\} \sim w^{-\nu}L(w), \quad w \to \infty, \tag{5.7}$$

then

$$\Pr\{V^{(2)} > w\} \sim \frac{1}{\nu-1}\frac{\lambda}{1-\rho}(1+\rho\nu)w^{1-\nu}L(w), \quad w \to \infty. \tag{5.8}$$

Remark 5.1. Under the conditions of Theorem 5.2, the tail of the waiting time distribution in Q_1 is regularly varying of index $1-\nu$; this tail behaviour in fact coincides with that of the distribution of $S^{(1)}$. Moreover, in the M/G/1 queue Q_1, the steady-state workload $V^{(1)}$ has the same distribution as the steady-state waiting time. Hence:

$$\Pr\{V^{(1)} > w\} \sim \frac{1}{\nu-1}\frac{\lambda}{1-\rho}w^{1-\nu}L(w), \quad w \to \infty, \tag{5.9}$$

which should be compared with (5.8).

6 A heavy-traffic limit theorem for the sojourn time distribution

In [9], Boxma and Cohen have obtained heavy-traffic limit theorems for the waiting time distribution in the $GI/G/1$ queue, when the variance of the interarrival time distribution and/or the service time distribution is *infinite*. One of their main cases concerns the $M/G/1$ queue with regularly varying service time distribution of index $-\nu$, with $1 < \nu < 2$, as given by (5.7); in this case, the service time variance is indeed infinite. Their theorem for this case states the following. *The 'contracted' waiting time $\delta(\rho)W$ converges in distribution for $\rho \uparrow 1$ to a limiting distribution $R_{\nu-1}(t)$. This distribution is specified by having $\frac{1}{1+s^{\nu-1}}$ as its Laplace-Stieltjes transform. The coefficient of contraction $\delta(\rho)$ is the only solution to the 'contraction equation'*

$$-\Gamma(1-\nu)x^{\nu-1}L(1/x) = \frac{1-\rho}{\lambda}, \quad x > 0, \tag{6.1}$$

where $\Gamma(\cdot)$ is the Gamma function, with the property that $\delta(\rho) \downarrow 0$ for $\rho \uparrow 1$.
Exactly the same limit theorem holds in Q_1 for the sojourn time $S^{(1)}$ which is the sum of the waiting time and the (independent) service time. In the present section, we derive a heavy-traffic limit theorem for the sojourn time $S^{(2)}$ in the case of a regularly varying service time distribution of index $-\nu$, $\nu > 1$.

Theorem 6.1 *For the stable tandem queue with Poisson input process and identical service times at both queues, and with the service time distribution satisfying the condition of Theorem 4.2, i.e.,*

$$\Pr\{B > w\} \sim w^{-\nu}L(w), \quad w \to \infty, \tag{6.2}$$

where $\nu > 1$, the 'contracted' sojourn time $\Delta(\rho)S^{(2)}$ converges in distribution for $\rho \uparrow 1$. The limit distribution function $H(w)$ is given by:

$$H(w) = \frac{\exp\{-w^{1-\nu}\}}{1+\nu w^{1-\nu}}, \quad w > 0, \tag{6.3}$$

and the coefficient of contraction $\Delta(\rho)$ is the unique root of the following equation

$$x^{\nu-1}L(1/x)=\frac{(\nu-1)(1-\rho)}{\lambda}, \tag{6.4}$$

with the property that $\Delta(\rho)\downarrow 0$ for $\rho\uparrow 1$.

Proof. Let $\Delta(\rho)$ be the solution to Equation (6.4) with the property $\Delta(\rho)\downarrow 0$ for $\rho\uparrow 1$. As proved in Lemma 10 in [10], the solution $\Delta(\rho)$ with such a property is unique. Using Theorem 1.6.1 in [4], it follows from (4.5) that for $w>0$,

$$\lambda\int_{w/\delta}^{\infty}(1-m(t))\mathrm{d}t\sim\frac{\lambda}{1-\rho}\frac{1}{\nu-1}\frac{w^{1-\nu}}{\delta^{1-\nu}}L(w/\delta),\quad \delta\downarrow 0, \tag{6.5}$$

which in combination with the definition of $\Delta(\rho)$ yields

$$\lim_{\rho\uparrow 1}\lambda\int_{w/\Delta(\rho)}^{\infty}(1-m(t))\mathrm{d}t=w^{1-\nu}. \tag{6.6}$$

Thus, for $w>0$,

$$\lim_{\rho\uparrow 1}Y(w/\Delta(\rho))=\lim_{\rho\uparrow 1}\exp\left\{-\lambda\int_{w/\Delta(\rho)}^{\infty}(1-m(t))\mathrm{d}t\right\}=\exp\{-w^{1-\nu}\}. \tag{6.7}$$

By (6.2), it is easy to get

$$\lim_{\rho\uparrow 1}B(w/\Delta(\rho))=1. \tag{6.8}$$

Applying Theorem 1.6.5 in [4], (4.4) implies that, for $w>0$,

$$\int_{w/\delta}^{\infty}t\mathrm{d}B(t)\sim\frac{\nu}{\nu-1}\frac{w^{1-\nu}}{\delta^{1-\nu}}L(w/\delta),\quad \delta\downarrow 0,$$

which further implies that

$$\lim_{\rho\uparrow 1}\frac{\lambda}{1-\rho}\int_{w/\Delta(\rho)}^{\infty}t\mathrm{d}B(t)=\nu w^{1-\nu}. \tag{6.9}$$

Since (3.10) and (6.9) implies that $\lim_{\rho\uparrow 1}F(w/\Delta(\rho))=0$ where $F(\cdot)$ is given by (3.6), it follows from (2.4) and (3.7) that, for $w>0$,

$$\lim_{\rho\uparrow 1}X(w/\Delta(\rho))=\frac{1}{1+\nu w^{1-\nu}}. \tag{6.10}$$

By (2.2) and (2.7), we can rewrite $S^{(2)}(w)$ as

$$S^{(2)}(w)=X(w)Y(w)B(w). \tag{6.11}$$

Combining (6.7), (6.8), (6.10) and (6.11) leads to

$$\lim_{\rho\uparrow 1}S^{(2)}(w/\Delta(\rho))=\frac{\exp\{-w^{1-\nu}\}}{1+\nu w^{1-\nu}},$$

which finally implies that, for $w>0$,

$$\lim_{\rho\uparrow 1}\Pr\{\Delta(\rho)S^{(2)}\le w\}=\lim_{\rho\uparrow 1}S^{(2)}(w/\Delta(\rho))=\frac{\exp\{-w^{1-\nu}\}}{1+\nu w^{1-\nu}}.$$ □

Remark 6.1. The limiting distribution function $H(w)$ is easily seen to have a regularly varying tail of the same index as that of the tail of $S^{(2)}(w)$. It is interesting to observe that $\exp\{-w^{1-\nu}\}$, $w > 0$, is a Weibull distribution, cf. Feller [14] p. 52.

Remark 6.2. The above heavy-traffic limit theorem may be used to provide an approximation for $S^{(2)}(w)$; for such an approach to respectively the ordinary $M/G/1$ queue and the $M/G/1$ queue with priority classes, see [8] and [10].

Remark 6.3. In case both service time and interarrival time distributions have a finite second moment, Kingman [16] derives a standard heavy-traffic limit theorem for the stationary waiting time W in the $GI/G/1$ queue. In our tandem model, if $\nu > 2$, a similar limit theorem holds for the sojourn time $S^{(1)}$ at Q_1, i.e.,

$$\lim_{\rho\uparrow 1} \Pr\{\zeta(\rho) S^{(1)} \leq w\} = 1 - \mathrm{e}^{-w}, \quad w \geq 0,$$

with $\zeta(\rho) := 2\lambda(1-\rho)/[1+\lambda^2(\mathrm{E}B^2 - \beta^2)]$ (cf. [16]).

Remark 6.4. In fact, it is not surprising that when $\nu > 2$, the contraction coefficient $\Delta(\rho)$ of $S^{(2)}$ is much larger than the above contraction coefficient $\zeta(\rho)$ of $S^{(1)}$ for $\rho \uparrow 1$. As has been shown in [6], if the third moment β_3 of the service time is finite, then

$$\lim_{\rho\uparrow 1} \frac{\mathrm{E}S^{(2)}}{\mathrm{E}S^{(1)}} = 0, \qquad \lim_{\rho\uparrow 1} \frac{\mathrm{Var}(S^{(2)})}{\mathrm{Var}(S^{(1)})} = 0.$$

In fact, using the technique used in [6] to derive the above limit results, one can show that if the $(n+1)$-th moment β_{n+1} is finite $(n \geq 2)$, then

$$\mathrm{E}[(S^{(2)})^n] \leq \frac{C}{(1-\rho)^{\frac{2n+1}{n+1}}},$$

for some positive constant C.

Acknowledgment. The authors are indebted to V. Dumas and K. Sigman for interesting discussions.

References

[1] V. Anantharam (1996). *Networks of queues with long-range dependent traffic streams.* In: P. Glasserman, K. Sigman and D.D. Yao (eds.), Stochastic Networks - Stability and Rare Events (Springer Verlag, Berlin) 237-256.

[2] F. Baccelli, S. Schlegel, and V. Schmidt (1998). *Asymptotics of stochastic networks with subexponential service times.* Queueing Systems, to appear.

[3] J. Beran, R. Sherman, M.S. Taqqu, and W. Willinger (1995). *Long-range dependence in variable-bit-rate video.* IEEE Transactions on Communications **43**, 1566-1579.

[4] N.H. Bingham, C.M. Goldie, and J.L. Teugels (1987). *Regular Variation.* Cambridge University Press, Cambridge.

[5] O.J. Boxma (1978). *On the longest service time in a busy period of the M/G/1 queue.* Stochastic Processes and their Applications **8**, 93-100.

[6] O.J. Boxma (1979). *On a tandem queueing model with identical service times at both counters, I,II.* Adv. Appl. Probab. **11**, 616-643; 644-659.

[7] O.J. Boxma (1980). *The longest service time in a busy period.* ZOR **24**, 235-242.

[8] O.J. Boxma and J.W. Cohen (1998). *The $M/G/1$ queue with heavy-tailed service time distribution.* IEEE J. Sel. Areas Commun. **16**, 749-763.

[9] O.J. Boxma and J.W. Cohen (1999). *Heavy-traffic analysis for the $GI/G/1$ queue with heavy-tailed distribution.* Queueing Systems, to appear.

[10] O.J. Boxma, J.W. Cohen and Q. Deng (1998). *Heavy-traffic analysis of the $M/G/1$ queue with priority classes.* Center discussion paper 98102, Center for Economic Research, Tilburg University.

[11] O.J. Boxma and V. Dumas (1998). *Fluid queues with heavy-tailed activity period distributions.* CWI Report PNA-R9705. To appear in *Computer Communications.*

[12] O.J. Boxma and V. Dumas (1998). *The busy period in the fluid queue.* Performance Evaluation Review **26**, 100-110.

[13] J.W. Cohen (1997). *Heavy-traffic limit theorems for the heavy-tailed $GI/G/1$ queue.* CWI Report PNA-R9719.

[14] W. Feller (1970). *An Introduction to Probability Theory and its Applications, Vol. II.* Wiley, New York.

[15] T. Huang and K. Sigman (1998). *Delay asymptotics for tandem, split & match, and other feedforward queues with heavy tailed service.* Queueing Systems, to appear.

[16] J.F.C. Kingman (1965). *The heavy traffic approximation in the theory of queues.* In: W.L. Smith, W.E. Wilkinson (eds.), Proceedings of the Symposium on Congestion Theory (The University of North Carolina Press, Chapel Hill) 137-159.

[17] C. Klüppelberg (1988). *Subexponential distributions and integrated tails.* J. Appl. Probab. **25**, 132-141.

[18] A.V. Makarichev *Analysis of a tandem queueing system with identical service times at both counters for various service disciplines.* Proc. 3rd ITC Seminar (Moscow, June 1984) 298-301.

[19] A.G. Pakes (1975). *On the tails of waiting-time distributions.* J. Appl. Probab. **12**, 555-564.

[20] K. Park and W. Willinger, eds. (1999). *Self-similar Network Traffic and Performance Evaluation.* Wiley, New York.

[21] V.Paxson, S. Floyd (1995). *Wide area traffic: the failure of Poisson modeling.* IEEE/ACM Transactions on Networking **3**, 226-244.

[22] O.P. Vinogradov (1984) *On the distribution of sojourn time in the tandem system with identical service times.* Proc. 3rd ITC Seminar (Moscow, June 1984) 449-450.

[23] W. Willinger, M.S. Taqqu, W.E. Leland, and D.V. Wilson (1995). *Self-similarity in high-speed packet traffic: analysis and modeling of Ethernet traffic measurements.* Statistical Science **10**, 67-85.