The Annals of Applied Probability
1992, Vol. 2, No. 3, 555–574

LIGHT TRAFFIC EQUIVALENCE IN SINGLE-SERVER QUEUES

BY SØREN ASMUSSEN

Chalmers University of Technology

A light traffic limit theorem is proved for random walks in a triangular array setting similar to the heavy traffic situation, the basic assumption being on the moments in the right tail of the increment distribution. When specialized to GI/G/1 queues, this result is shown to contain the known types of light traffic behaviour in this setting (Daley and Rolski) as well as some additional ones. Intuitively, the results state that typically delay in light traffic occurs with just one customer in the system, and then as a result of long service times and/or short interarrival times in a balance which depends on the particular parameters of the model. Particular attention is given to queues with phase-type service times, for example of Coxian type.

1. Introduction. The study of light traffic limit theorems for queues goes back at least to [5], but compared to heavy traffic approximations, which were extensively studied and applied in the 1970s and 1980s, the area has remained relatively unexplored for many years. However, the interest in light traffic phenomena has been renewed in recent years, starting with the work of Burman and Smith [6, 7] and Daley and Rolski [9]. Besides the intrinsic interest of the topic, one main motivation has been to study queues in moderate traffic by interpolating between the light traffic and heavy traffic approximations; see [8], [17], [19] and [21]. Further important references in the area include Wolff [22], Pinedo and Wolff [14] and Reiman and Simon [16, 18].

To indicate the flavour of the topic, consider the GI/G/1 queue with generic service time U, generic interarrival time T and generic steady-state waiting time W. Intuitively, light traffic means that T is much larger than U, implying that typically, customers do not have to wait at all so that $p_+ = \mathbb{P}(W > 0)$ is close to 0. The problem is to study W in the unlikely event that $W > 0$, and the crux for this may be argued to be formalizing the relation

$$W \approx (U - T)^+, \tag{1.1}$$

which comes out by a number of intuitive or semi-mathematical lines of thought [e.g., one may note that $W = (U - T)^+$ exactly in the typical case of the preceding customer arriving to an idle system].

In Daley and Rolski [9], the mathematical setting for (1.1) is to provide conditions for

$$\frac{\mathbb{E}W}{\mathbb{E}(U - T)^+} \to 1 \tag{1.2}$$

Received March 1990; revised June 1991.

AMS 1980 *subject classification*. 60K25.

Key words and phrases. Light traffic, random walk, GI/G/1 queue, phase-type distribution, Coxian distribution.

in a triangular array setting where $U = U^{(\gamma)}$, $T = T^{(\gamma)}$ (and hence $W = W^{(\gamma)}$) depend on a parameter γ tending to ∞ such that $T^{(\gamma)}$ ultimately dominates $U^{(\gamma)}$. More precisely, Daley and Rolski [9] consider the dilation case

$$(1.3) \qquad U^{(\gamma)} = U_*, \qquad T^{(\gamma)} = \gamma T_*,$$

with U_*, T_* independent of γ (by rescaling, this situation is equivalent to $U = \varepsilon U_*$, $T = T_*$, $\varepsilon \to 0$). It is noted in [9] that (1.2) holds for M/G/1 or GI/M/1 queues, but also counterexamples (of D/G/1 type) are given, showing that (1.2) may fail even in the simple setting of (1.3). These investigations are continued by Daley and Rolski in [10] and [11], but the picture which emerges is more diverse. For example in [10], asymptotic formulas for $\mathbb{E}W$ are given subject to thinning of the arrival process and in the special case

$$(1.4) \qquad \mathbb{P}(T_* \le x) \approx c_{A_*} x^\alpha, \qquad x \to 0,$$

of (1.3) [here c_{A_*} is a constant dependent on the distribution A_* of T_*; as in [10], the class of distributions satisfying (1.4) is denoted $\mathscr{S}_\alpha$ in the following]. From these and related results, Daley and Rolski [10] draw some general conclusions such as:

> "The dominant feature of light traffic characteristics is their dependence on the clustering tendency... of the arrival process";
> "... there do not exist perfectly general conditions for... [(1.2)]."

Discussions of these matters with Daryl Daley in the spring of 1988 and with Tomasz Rolski in the fall of 1989 provided the stimulus for the present research.

The points that we hope to make with the present paper are the following:

1. The relation (1.1) between W and $(U - T)^+$ is at the core of light traffic limit theory. Furthermore, it is a *distributional* property and does not refer to expected values as in (1.2) alone. Some indications of this are in fact already present in [10] and, even more, in [11], and the spirit is certainly also rather much the same as in [18].
2. One can comprise *all* known instances of light traffic behaviour in single-server queues into a single random walk triangular array setting (just as in the heavy traffic case; cf. [1], Chapter VIII.6).
3. Light traffic behaviour is the result of a delicate interaction between short interarrival times (clustering) *and* long service times—in some situations one of these features may be the predominant one (say short interarrival times in GI/D/1 or long service times in D/G/1), in others both may be present (e.g., Example 3.2 below).

The paper is organized as follows. In Section 2 we introduce the notion of *light traffic equivalence* which is our distributional setting, and a light traffic limit theorem for random walks is proved under a certain condition (somewhat reminiscent of uniform integrability) on the moments in the right tail in the family of increment distributions. Section 3 deals with GI/G/1 queues: We investigate the form of the basic condition in a number of cases (in particular, we show how a number of results of Daley and Rolski [9, 10] come out as special cases), and we give examples showing some of the possible types of behaviour exhibited by interarrival times and service times when delay occurs in light traffic. In Section 4 we study the specific form of the light traffic approximations for the GI/PH/1 queue. It is shown that the light traffic limit is phase-type with the same phase generator as the service time distribution, and we derive computationally tractable expressions for the entrance distribution and $p_+ = \mathbb{P}(W > 0)$. In particular, Coxian distributions are exploited, and we introduce a class $\mathscr{S}_\infty$ of distributions which fade away at the origin faster than for any of the classes $\mathscr{S}_\alpha$ (e.g., T is in $\mathscr{S}_\infty$ if $T \geq \varepsilon$ for some $\varepsilon > 0$). The main result for GI/PH/1 queues in this setting states that when the interarrival distribution is in $\mathscr{S}_\infty$, then the light traffic approximation is simply exponential. Finally, some concluding discussion is given in Section 5.

The following notation is used throughout the paper: $\mathbb{E}[X; A]$ means $\mathbb{E}XI(A)$; for a given distribution function F, $\bar{F}$ denotes the tail, $\bar{F}(x) = 1 - F(x)$, and $F(\cdot|x)$ denotes the overshoot distribution, $F(y|x) = F(y + x)/F(x)$.

2. Light traffic equivalence in random walks.

To formalize our results, we need the following notion.

Definition 2.1. Two families $\{R^{(\gamma)}\}_{\gamma > 0}, \{S^{(\gamma)}\}_{\gamma > 0}$ of random variables with values in $[0, \infty)$ are *distributional light traffic equivalent* if

$$\mathbb{P}(R^{(\gamma)} > 0) \to 0, \qquad \mathbb{P}(S^{(\gamma)} > 0) \to 0, \qquad \frac{\mathbb{P}(R^{(\gamma)} > 0)}{\mathbb{P}(S^{(\gamma)} > 0)} \to 1 \tag{2.1}$$

as $\gamma \to \infty$ and the total variation distance

$$\left\| \mathbb{P}(R^{(\gamma)} \in \cdot | R^{(\gamma)} > 0) - \mathbb{P}(S^{(\gamma)} \in \cdot | S^{(\gamma)} > 0) \right\| \tag{2.2}$$

converges to 0 as $\gamma \to \infty$. The families are *light traffic equivalent of order p* if the moments of order $q \leq p$ behave asymptotically equivalent when $\gamma \to \infty$ as well,

$$\frac{\mathbb{E}R^{(\gamma)q}}{\mathbb{E}S^{(\gamma)q}} \to 1, \qquad 0 \leq q \leq p. \tag{2.3}$$

For basic facts about total variation convergence, see the appendix of [1]. Note that it is not essential for the definition that $R^{(\gamma)}$, $S^{(\gamma)}$ are defined w.p.1; for defective random variables then, for example, the event $\{R^{(\gamma)} > 0\}$ means that $R^{(\gamma)}$ is defined and is greater than 0. Examples of this occur in connection with ladder variables below.

The concept may, of course, also be relevant outside the area of light traffic limit theorems for queues, and for this reason Daley and Rolski used the term *asymptotic conditional equivalence* in a revision of [11] following the first version of this paper.

Now consider a triangular array $\{S_n^{(\gamma)}\}$ of random walks with increments $X_1^{(\gamma)}, X_2^{(\gamma)}, \ldots,$ increment distributions $F^{(\gamma)}(x) = \mathbb{P}(X^{(\gamma)} \leq x)$, and define

$M^{(\gamma)} = \max_{n=0,1,\dots} S_n^{(\gamma)}$. Our goal for light traffic analysis is then to show that $M^{(\gamma)}$ and $X^{(\gamma)^+}$ are light traffic equivalent, and this will be shown to hold true under conditions of the following type.

CONDITION $\mathscr{L}\mathscr{T}(p)$. For all q with $0 \le q \le p$,

$$\lim_{K \uparrow \infty} \limsup_{\gamma \to \infty} \frac{\mathbb{E}\left[X^{(\gamma)^{q+1}}; X^{(\gamma)} > K\right]}{\mathbb{E}\left[X^{(\gamma)^q}; X^{(\gamma)} > 0\right]} = \lim_{K \uparrow \infty} \limsup_{\gamma \to \infty} \frac{\int_K^\infty x^{q+1} F^{(\gamma)}(dx)}{\int_0^\infty x^q F^{(\gamma)}(dx)} = 0.$$

One may note that Condition $\mathscr{L}\mathscr{T}(p)$ with the denominator removed and the integration carried out over $\{x\colon |x| > K\}$ instead of $\{x\colon x > K\}$ means uniform integrability of the $(p+1)$th moment which the setting of heavy traffic limit theorems is exactly the required regularity condition needed for the study of the pth moment of M; cf. [4] and [1], Chapter VIII.6. If, as will typically be the case, the denominator (considered as a function of γ) is of the same order of magnitude as $\tilde{p}_+^{(\gamma)} = \mathbb{P}(X^{(\gamma)} > 0)$, we are back to uniform integrability, only now in the conditional distribution $F^{(\gamma)}(\cdot\,|0)$ of $X^{(\gamma)}$ given $X^{(\gamma)} > 0$ [in particular, $\mathscr{L}\mathscr{T}(0)$ always means that the family $\{F^{(\gamma)}(\cdot\,|0)\}_{\gamma > 0}$ is uniformly integrable]. An example of this type of behaviour which, it would appear, would cover a wide range of examples is given in Corollary 2.1 (some cases not included are studied in Corollary 3.5 and Example 3.2). Condition $\mathscr{L}\mathscr{T}$ would fail, typically, if $F^{(\gamma)}$ exhibits wild fluctuations in the overshoot distributions. See, for example, Example 3.1 below.

Here is our main result for the random walk setting.

THEOREM 2.1. *Assume that* $X^{(\gamma)} \to_{\mathscr{D}} -\infty$ *when* $\gamma \to \infty$, *and that Condition* $\mathscr{L}\mathscr{T}(0)$ *holds. Then* $M^{(\gamma)}$ *and* $X^{(\gamma)^+}$ *are distributional light traffic equivalent. If Condition* $\mathscr{L}\mathscr{T}(p)$ *also holds, then* $M^{(\gamma)}$ *and* $X^{(\gamma)^+}$ *are light traffic equivalent of order* p.

For the proof, we shall employ a Wiener–Hopf inspired argument, the essence of which is to first express $M^{(\gamma)}$ as a geometric sum of ascending ladder heights which in terms of distributions means

$$(2.4) \qquad \left(1 - \|G_+^{(\gamma)}\|\right) \sum_{n=0}^{\infty} G_+^{(\gamma)^{*n}}$$

(this idea is advocated at an early historical stage in [12]) and next to express the ascending ladder height distribution $G_+^{(\gamma)}$ in terms of the renewal measure $U_-^{(\gamma)} = \sum_0^\infty G_-^{(\gamma)^{*n}}$ associated with the descending ladder height distribution $G_-^{(\gamma)}$ (for further closely related applications of this technique, see [3], [1], pages 184 and 185, and [23]). Here as usual $G_+^{(\gamma)}(x) = \mathbb{P}(S_{\tau_+}^{(\gamma)} \le x)$, where $\tau_+ = \inf\{n \ge 1\colon S_n^{(\gamma)} > 0\}$.

Define $p_+^{(\gamma)} = \mathbb{P}(M^{(\gamma)} > 0)$, $\tilde{p}_+^{(\gamma)} = \mathbb{P}(X^{(\gamma)^+} > 0)$. Then also $p_+^{(\gamma)} = \|G_+^{(\gamma)}\|$ in view of (2.4).

Lemma 2.1. *Assume that* $X^{(\gamma)} \to_{\mathscr{D}} -\infty$ *when* $\gamma \to \infty$, *and that Condition* $\mathscr{L}\mathscr{T}(0)$ *holds. Then the ascending ladder height* $S^{(\gamma)}_{\tau_+}$ *and* $X^{(\gamma)^+}$ *are distributional light traffic equivalent. If also Condition* $\mathscr{L}\mathscr{T}(p)$ *holds, then* $S^{(\gamma)}_{\tau_+}$ *and* $X^{(\gamma)^+}$ *are light traffic equivalent of order* p.

Proof. It is standard ([1], page 173) that $G_+^{(\gamma)}$ is the restriction of $F^{(\gamma)} * U_-^{(\gamma)}$ to $(0,\infty)$ so that we can write $G_+^{(\gamma)} = H^{(\gamma)} + K^{(\gamma)}$, where $H^{(\gamma)}$ is the restriction to $(0,\infty)$ of $F^{(\gamma)}$ (i.e., $H^{(\gamma)}$ is the contribution from the atom of $U_-^{(\gamma)}$ at 0) and $K^{(\gamma)} = F^{(\gamma)} * \sum_1^\infty G_-^{(\gamma)*n}$. For simplicity of notation, let $R^{(\gamma)}$ be the measure $R^{(\gamma)}(dx) = \sum_{n=1}^\infty G_-^{(\gamma)*n}(d(-x))$ on $(0,\infty)$. Then for $z \ge 0$,

$$
\begin{aligned}
\overline{K}^{(\gamma)}(z) &= \sum_{n=1}^{\infty} \int_{-\infty}^{0} \overline{F}^{(\gamma)}(z-x) G_-^{(\gamma)*n}(dx) \\
&= \int_0^\infty \overline{F}^{(\gamma)}(z+x) R^{(\gamma)}(dx) \\
&= \int_z^\infty R^{(\gamma)}(y-z) F^{(\gamma)}(dy) \\
&\le \int_z^\infty R^{(\gamma)}(y) F^{(\gamma)}(dy).
\end{aligned}
\tag{2.5}
$$

To proceed from (2.5), we need to show that

$$
R^{(\gamma)}(t) \le \varphi^{(\gamma)}(t)(1+t), \tag{2.6}
$$

where $\varphi^{(\gamma)}(t)$ is bounded, nondecreasing and tends to 0 for any fixed t as $\gamma \to \infty$. First $X^{(\gamma)} \to_{\mathscr{D}} -\infty$ implies $S_{\tau_-}^{(\gamma)} \to_{\mathscr{D}} -\infty$ (with high probability $S_{\tau_-}^{(\gamma)}$ coincides with $X_1^{(\gamma)}$). In particular, $\overline{G}_-^{(\gamma)}(-t) \to 0$ for all $t > 0$. Since $R^{(\gamma)}(1) \le \overline{G}_-^{(\gamma)}(-1)(1 + R^{(\gamma)}(1))$, this implies that $R^{(\gamma)}(1)$ is bounded. Similarly,

$$
R^{(\gamma)}(n) - R^{(\gamma)}(n-1) \le \{G_-^{(\gamma)}(1-n) - G_-^{(\gamma)}(-n)\}(1 + R^{(\gamma)}(1))
$$

so that

$$
R^{(\gamma)}(n) \le \overline{G}_-^{(\gamma)}(-n) n (1 + R^{(\gamma)}(1)),
$$

and from these estimates (2.6) follows.

Letting $z = 0$ in (2.5), we get for $K \ge 1$,

$$
\begin{aligned}
\limsup_{\gamma\to\infty} \frac{\|K^{(\gamma)}\|}{\tilde{p}_+^{(\gamma)}} &\le \frac{\int_0^\infty \varphi^{(\gamma)}(y)(1+y)F^{(\gamma)}(dy)}{\int_0^\infty F^{(\gamma)}(dy)} \\
&\le \limsup_{\gamma\to\infty} \left\{ (1+K)\varphi^{(\gamma)}(K) + 2\varphi^{(\gamma)}(\infty) \frac{\int_K^\infty y F^{(\gamma)}(dy)}{\int_0^\infty F^{(\gamma)}(dy)} \right\} \\
&= \limsup_{\gamma\to\infty} 2\varphi^{(\gamma)}(\infty) \frac{\int_K^\infty y F^{(\gamma)}(dy)}{\int_0^\infty F^{(\gamma)}(dy)}.
\end{aligned}
$$

Letting $K \to \infty$, this converges to 0 according to $\mathscr{L}\mathscr{T}(0)$. That the requirements corresponding to (2.1) and (2.2) are satisfied now follows easily from $\tilde{p}_+^{(\gamma)} = \|H^{(\gamma)}\| \to 0$.

For (2.3), fix $q \le p$ and write $m(H^{(\gamma)}) = \int_0^\infty x^q H^{(\gamma)}(dx)$. Then we must show $m(G_+^{(\gamma)})/m(H^{(\gamma)}) \to 1$. Multiplying (2.5) by qz^{q-1}, inserting (2.6) and integrating, we get similarly as above that

$$\begin{aligned}\frac{m(K^{(\gamma)})}{m(H^{(\gamma)})} &\le \frac{\int_0^\infty qz^{q-1}\,dz\int_z^\infty \varphi^{(\gamma)}(y)(1+y)F^{(\gamma)}(dy)}{m(H^{(\gamma)})}\\ &= \frac{\int_0^\infty \varphi^{(\gamma)}(y)(1+y)y^qF^{(\gamma)}(dy)}{\int_0^\infty y^qF^{(\gamma)}(dy)}\\ &\le (1+K)K^q\varphi^{(\gamma)}(K) + 2\varphi^{(\gamma)}(\infty)\frac{\int_K^\infty y^{q+1}F^{(\gamma)}(dy)}{\int_0^\infty y^qF^{(\gamma)}(dy)},\end{aligned}$$

$$\limsup_{\gamma\to\infty}\frac{m(K^{(\gamma)})}{m(H^{(\gamma)})} \le \limsup_{\gamma\to\infty} 2\varphi^{(\gamma)}(\infty)\frac{\int_K^\infty y^{q+1}F^{(\gamma)}(dy)}{\int_0^\infty y^qF^{(\gamma)}(dy)}.$$

Letting $K \uparrow \infty$ and using $\mathscr{L}\mathscr{T}(p)$ completes the proof. □

Proof of Theorem 2.1. Since $p_+^{(\gamma)} = \|G_+^{(\gamma)}\|$, the requirements for distributional light traffic equivalence [corresponding to (2.1) and (2.2)] are immediate consequences of (2.4) and Lemma 2.1. For (2.3), we must show $\mathbb{E}W^{(\gamma)^q}/m(H^{(\gamma)}) \to 1$. By (2.4),

$$\text{(2.7)}\qquad \mathbb{E}W^{(\gamma)^q} = \left(1-p_+^{(\gamma)}\right)m\left(G_+^{(\gamma)}\right) + \left(1-p_+^{(\gamma)}\right)\sum_{n=2}^\infty \int_0^\infty x^qG_+^{(\gamma)^{*n}}(dx).$$

As in [1], page 184, we can bound the second term by

$$\sum_{n=2}^\infty n^p\|G_+^{(\gamma)}\|^n\frac{m\left(G_+^{(\gamma)}\right)}{\|G_+^{(\gamma)}\|} = o\left(m\left(G_+^{(\gamma)}\right)\right).$$

Thus (2.7) becomes

$$\mathbb{E}W^{(\gamma)^q} = m\left(G_+^{(\gamma)}\right) + o\left(m\left(G_+^{(\gamma)}\right)\right) = m(H^{(\gamma)})(1+o(1)). \qquad \square$$

Corollary 2.1. *Suppose that there exists a distribution G which has finite $(p+1)$th moment such that for any γ the conditional distribution $F^{(\gamma)}(\cdot|0)$ of $X^{(\gamma)}$ given $X^{(\gamma)} > 0$ is stochastically dominated by G, and suppose further that $\liminf_{\gamma\to\infty}\overline{F}^{(\gamma)}(\varepsilon|0) > 0$ for some $\varepsilon > 0$. Then Condition $\mathscr{L}\mathscr{T}(p)$ holds.*

Proof. The corollary is an obvious consequence of the inequality

$$\frac{\int_K^\infty x^{p+1}F^{(\gamma)}(x)}{\int_0^\infty x^pF^{(\gamma)}(x)} \le \frac{\tilde{p}_+^{(\gamma)}\int_K^\infty x^{p+1}G(dx)}{\tilde{p}_+^{(\gamma)}\varepsilon^p\overline{F}^{(\gamma)}(\varepsilon|0)}. \qquad \square$$

3. Light traffic behaviour of GI/G/1 queues. Now consider a system (triangular array) of GI/G/1 queues in the notation of the Introduction.

Denote the interarrival distribution by $A^{(\gamma)}$ and assume throughout that the service time U or equivalently the service-time distribution B is fixed, that is, does not depend on the parameter γ. Then, in view of the well-known random walk representation of the waiting time, Theorem 2.1 states that $W^{(\gamma)}$ and $(U - T^{(\gamma)})^+$ are light traffic equivalent provided $X^{(\gamma)} = U - T^{(\gamma)}$ satisfies Condition $\mathscr{L}\mathscr{T}$. We shall here carry out the relevant translation to conditions in terms of $A^{(\gamma)}$, B, show that the main cases considered by Daley and Rolski [9, 10] are included in Theorem 2.1 (Corollaries 3.1 and 3.2) and give some examples of a different spirit (Examples 3.1 and 3.2). Also we look into the problem of describing the conditional distribution of $U, T^{(\gamma)}$ given $X^{(\gamma)} = U - T^{(\gamma)} > 0$ for the purpose of providing a more intuitive description of how delay occurs in light traffic.

The first example is thinning of the arrival where the results of [10] are given in terms of the renewal function $H = \sum_1^\infty A^{*n}$.

Corollary 3.1. *Given a* GI/G/1 *queueing system specified in terms of* U, T, *define another* GI/G/1 *system by thinning of the arrival process with retention probability* $1/\gamma$. *That is,* $T^{(\gamma)} = T_1 + \cdots + T_{N^{(\gamma)}}$, *where* $N^{(\gamma)}$ *is independent of* $T_1, T_2, \dots$ *with* $\mathbb{P}(N^{(\gamma)} = k) = (1 - 1/\gamma)^{k-1}/\gamma$. *Then* $W^{(\gamma)}$ *and* $(U - T^{(\gamma)})^+$ *are distributional light traffic equivalent provided that* $\mathbb{E}U^2 < \infty$, *in which case*

$$p_+^{(\gamma)} = \mathbb{P}(W^{(\gamma)} > 0) \approx \mathbb{P}\Big(\big(U - T^{(\gamma)}\big)^+ > 0\Big) \approx \frac{1}{\gamma}\mathbb{E}H(U), \tag{3.1}$$

where $H(u) = \sum_1^\infty A^{*k}(u)$ *is the renewal function. If furthermore* $\mathbb{E}U^{p+2} < \infty$, *then also light traffic equivalence of order p holds, in which case*

$$\mathbb{E}W^{(\gamma)p} \approx \mathbb{E}\Big(\big(U - T^{(\gamma)}\big)^+\Big)^p \approx \frac{1}{\gamma}\mathbb{E}\int_0^U p y^{p-1} H(U - y)\,dy. \tag{3.2}$$

Proof. Obviously,

$$\mathbb{P}(U - T^{(\gamma)} > y) = \int_y^\infty \sum_{k=1}^\infty \frac{1}{\gamma}(1 - 1/\gamma)^{k-1} A^{*k}(u - y) B(du)$$

for $y > 0$ so that

$$\gamma \mathbb{P}(U - T^{(\gamma)} > y) \uparrow \int_y^\infty H(u - y) B(du), \qquad \gamma \to \infty. \tag{3.3}$$

In particular, monotone convergence yields

$$\begin{aligned}\limsup_{\gamma\to\infty} \frac{\int_K^\infty x^{p+1} F^{(\gamma)}(dx)}{\int_0^\infty x^p F^{(\gamma)}(dx)} &= \limsup_{\gamma\to\infty} \frac{\int_K^\infty (p+1) y^p \mathbb{P}(U - T^{(\gamma)} > y)\,dy}{\int_0^\infty p y^{p-1} \mathbb{P}(U - T^{(\gamma)} > y)\,dy} \\ &= \frac{\int_K^\infty (p+1) y^p \int_y^\infty H(u - y) B(du)\,dy}{\int_0^\infty p y^{p-1} \int_y^\infty H(u - y) B(du)\,dy}.\end{aligned} \tag{3.4}$$

Now according to the renewal theorem, we have $H(u) \le c_1 + c_2 u$, and thus the numerator in (3.4) can be bounded by $c_3 \int_K^\infty u^{p+2} B(du)$ which tends to 0 as $K \uparrow \infty$ when $\mathbb{E}U^{p+2} < \infty$, and thus $\mathscr{LT}(p)$ holds. It only remains to prove that $p_+^{(\gamma)}$ and $\mathbb{E}((U - T^{(\gamma)})^+)^p$ behave as asserted in (3.1) and (3.2). Interchanging the order of integrations, this comes out as a special case of (3.3) for $p_+^{(\gamma)}$ and by inspecting the denominator of (3.4) for $\mathbb{E}((U - T^{(\gamma)})^+)^p$. □

In the rest of this section, we consider the dilation case $T^{(\gamma)} = \gamma T_*$ with $A_*(t) = \mathbb{P}(T_* \le t)$ independent of γ. Recall the definition (1.4) of the class $\mathscr{S}_\alpha$.

Corollary 3.2. *Assume $T^{(\gamma)} = \gamma T_*$ with T_* in the class $\mathscr{S}_\alpha$. Then $W^{(\gamma)}$ and $(U - T^{(\gamma)})^+$ are distributional light traffic equivalent provided that $\mathbb{E}U^{\alpha+1} < \infty$, in which case*

$$p_+^{(\gamma)} = \mathbb{P}(W^{(\gamma)} > 0) \approx \mathbb{P}\Big((U - T^{(\gamma)})^+ > 0\Big) \approx \frac{c_{A_*}}{\gamma^\alpha}\mathbb{E}U^\alpha. \tag{3.5}$$

If furthermore $\mathbb{E}U^{\alpha+p+1} < \infty$, then also light traffic equivalence of order p holds, in which case

$$\mathbb{E}W^{(\gamma)^p} \approx \mathbb{E}\Big((U - T^{(\gamma)})^+\Big)^p \approx \frac{p c_{A_*} \mathscr{B}(\alpha + 1, p)}{\gamma^\alpha}\mathbb{E}U^{\alpha+p}, \tag{3.6}$$

where $\mathscr{B}(\alpha + 1, p) = \int_0^1 x^\alpha (1 - x)^{p-1}\,dx$ is the beta function.

Proof. For $y > 0$,

$$\mathbb{P}(U - \gamma T_* > y) = \int_y^\infty A_*\left(\frac{u - y}{\gamma}\right) B(du).$$

Letting $y = 0$, we get

$$\gamma^\alpha p_+^{(\gamma)} = \int_0^\infty \gamma^\alpha A_*\left(\frac{u}{\gamma}\right) B(du) \to c_{A_*}\int_0^\infty u^\alpha B(du)$$

[using dominated convergence and $\sup_t A_*(t)/t^\alpha < \infty$]. Similarly,

$$\gamma^\alpha \mathbb{E}\Big((U - T^{(\gamma)})^+\Big)^p \to \int_0^\infty p y^{p-1}\,dy\, c_{A_*}\int_y^\infty (u - y)^\alpha B(du),$$

which after some calculations reduces to (3.6). It only remains to verify condition $\mathscr{LT}(p)$, which follows easily from

$$\gamma^\alpha \int_K^\infty y^{p+1} F^{(\gamma)}(dy) = \gamma^\alpha \int_K^\infty (p + 1) y^p\,dy \int_y^\infty A_*\left(\frac{u - y}{\gamma}\right) B(du)$$

$$\to \int_K^\infty (p + 1) y^p\,dy\, c_{A_*}\int_y^\infty (u - y)^\alpha B(du). \qquad \square$$

COROLLARY 3.3. *Assume* $T^{(\gamma)} = \gamma T_*$ *with* T_* *in the class* $\mathscr{S}_\alpha$. *If* $\mathbb{E}U^{\alpha+1} < \infty$, *then for* $0 < t < u$,

$$\mathbb{P}\left(U \le u, T_* \le \frac{t}{\gamma}\,\middle|\, U - \gamma T_* > 0\right) \to \frac{\int_0^u y^\alpha (t/y)^\alpha B(dy)}{\int_0^\infty y^\alpha B(dy)}.$$

PROOF. This follows by combining the previous estimates with

$$\mathbb{P}\left(U \le u, T_* \le \frac{t}{\gamma}\right) = \int_0^u A_*\left(\frac{t}{\gamma}\right) B(dy) \approx \gamma^{-\alpha} c_{A_*} \int_0^u t^\alpha B(dy). \qquad \square$$

REMARK 3.1. The intuitive content of Corollary 3.3 is the following: If delay occurs at all in light traffic, then typically the preceding customer entered an empty system, his service time $U = u$ was chosen from the distribution B_α with density $u^\alpha/\mathbb{E}U^\alpha$ w.r.t. $B(du)$ and the time T ago when he arrived has density $\alpha(t/u)^{\alpha-1}$ on $(0, u)$. In terms of T_*, this is the same as saying that T_* exhibits atypical behaviour by being concentrated on $(0, u/\gamma)$ (thus T_* is one order of magnitude smaller than typical), whereas U pertains to the typical order of magnitude (but is moderately larger than typical by being chosen from B_α which is stochastically larger than B).

REMARK 3.2. As a trivial example of behaviour opposite to Example 3.1, consider the D/M/1 queue, say $T_* = 1$, where the distribution of $U - \gamma T$ given $U - \gamma T > 0$ is exponential. Thus if delay occurs in light traffic, then U is greater than γ and thus atypically large (on the contrary the behaviour of T is deterministic and thus trivial). Slightly more complicated examples of this type are in Section 4.

Here is a fairly general result allowing for a completely arbitrary arrival mechanism.

COROLLARY 3.4. *Assume that there exists a distribution* G *with finite* $(p+1)$*th moment such that the overshoot distribution* $B(\cdot|b)$ *is stochastically dominated by* G *for all* b, *and also that* $\overline{B}(\varepsilon|b) \ge \delta$ *for some* $\varepsilon, \delta > 0$ *and all* b. *Then* $W^{(\gamma)}$ *and* $(U - T^{(\gamma)})^+$ *are distributional light traffic equivalent when* $p = 0$ *whereas if* $p > 0$, *then also light traffic equivalence of order* p *holds.*

PROOF. For $y > 0$,

$$\mathbb{P}(U - T^{(\gamma)} > y) = p_+^{(\gamma)} \mathbb{P}(U > T^{(\gamma)} + y | U > T^{(\gamma)}) \le p_+^{(\gamma)} \overline{G}(y),$$

$$\mathbb{P}(U - T^{(\gamma)} > \varepsilon) = p_+^{(\gamma)} \mathbb{P}(U > T^{(\gamma)} + \varepsilon | U > T^{(\gamma)}) \ge p_+^{(\gamma)} \delta.$$

Hence

$$\limsup_{\gamma\to\infty} \frac{\int_K^\infty x^{p+1} F^{(\gamma)}(dx)}{\int_0^\infty x^p F^{(\gamma)}(dx)} = \limsup_{\gamma\to\infty} \frac{\int_K^\infty (p+1) y^p \mathbb{P}(U - T^{(\gamma)} > y)\, dy}{\int_0^\infty x^p F^{(\gamma)}(dx)}$$

$$\leq \frac{\int_K^\infty (p+1) y^p \overline{G}(y)\, dy}{\delta \varepsilon^p},$$

and letting $K \to \infty$ shows that Condition $\mathscr{L}\mathscr{T}(p)$ is satisfied. □

Example 3.1. It seems reasonable to ask whether it is really necessary (as in Corollary 3.4) to impose conditions on the family of overshoot distributions $\{B(\cdot\,|t)\}_{t>0}$, or whether instead a sufficiently strong moment condition on B alone would suffice (this question is further motivated by the fact that the counterexamples in [10] have infinite third moment). The following example strongly supports, however, the relevance of conditions on the overshoots. In fact, we shall exhibit a service-time distribution B such that $\int_0^\infty \mathrm{e}^x\, B(dx)$ is finite but (1.2) fails in the D/G/1 queue. To this end, let $T^{(\gamma)} = \gamma T_*$ with $T_* = 1$ and let B be concentrated on $n_0 = 3$, $n_1 = 3 \cdot 4, \dots, n_k = 3 \cdot 4^k, \dots$ with weights $p_k = c\,\mathrm{e}^{-4^{k+1}}$, where $c^{-1} = \sum_0^\infty \mathrm{e}^{-4^{k+1}}$. Then

$$\int_0^\infty \mathrm{e}^x\, B(dx) = \sum_{k=0}^\infty p_k\, \mathrm{e}^{n_k} = c \sum_{k=0}^\infty \mathrm{e}^{-4^k} < \infty,$$

$$\mathbb{E}\bigl(U - T^{(4k)}\bigr)^+ = \sum_{l=k}^\infty p_l (3 \cdot 4^l - 4^k) \approx 2 p_k 4^k = 2c 4^k\, \mathrm{e}^{-4^{k+1}},$$

$$\mathbb{E}\Bigl(U_1 + U_2 - T_1^{(4^k)} - T_2^{(4^k)}\Bigr)^+ \geq \mathbb{E}(U - 2 \cdot 4^k)$$

$$= \sum_{l=k}^\infty p_l (3 \cdot 4^l - 2 \cdot 4^k) \approx p_k 4^k = c 4^k\, \mathrm{e}^{-4^{k+1}}.$$

Thus $\liminf \mathbb{E} S_2^{(\gamma)^+} / \mathbb{E} S_1^{(\gamma)^+} \geq 2 > 0$, and since $\mathbb{E} W = \sum_1^\infty \mathbb{E} S_n^{(\gamma)^+}/n$ (Spitzer's identity [1], page 177), it follows that (1.2) cannot hold.

A second question raised by Corollary 3.4 is the role of the condition $\overline{B}(\varepsilon|b) \geq \delta$. The following result allows us to dispense with this in a number of cases (see also Example 3.2 below for a concrete case).

Corollary 3.5. *Assume that U has a nondecreasing failure rate. Then Condition $\mathscr{L}\mathscr{T}(p)$ is satisfied for all p, and hence $W^{(\gamma)}$ and $(U - T^{(\gamma)})^+$ are light traffic equivalent (distributional and of any order p).*

Proof. The case where the failure rate of U has a finite limit as $u \to \infty$ is covered by Corollary 3.4, and hence we may assume that the limit is ∞. It is then easy to see that the failure rate $r^{(\gamma)}(x)$ of $X^{(\gamma)} = U - T^{(\gamma)}$ is defined for all $x > 0$, nondecreasing in x for fixed γ and satisfies $r^{(\gamma)}(x) \to \infty$, $x \to \infty$. Let $\lambda^{(\gamma)} = r^{(\gamma)}(1)$ and consider $K > 1$. By monotonicity, we have $\overline{F}^{(\gamma)}(x|0) \geq \mathrm{e}^{-\lambda^{(\gamma)} x}$

when $x \le 1$, and hence

$$\begin{aligned}\int_0^\infty x^p F^{(\gamma)}(x) &\ge \tilde{p}_+^{(\gamma)} \mathbb{E}\big[X^{(\gamma)^p} \wedge 1 | X^{(\gamma)} > 0\big]\\ &\ge \tilde{p}_+^{(\gamma)} \int_0^1 x^p \, \mathrm{e}^{-\lambda^{(\gamma)} x}\, dx\\ &\approx \frac{\tilde{p}_+^{(\gamma)}}{\lambda^{(\gamma)p+1}} \int_0^\infty y^p \, \mathrm{e}^{-y}\, dy\end{aligned}$$

(using a change of variables and $\lambda^{(\gamma)} \to \infty$ in the last step). Similarly, $\mathbb{P}(X^{(\gamma)} > K) \le \tilde{p}_+^{(\gamma)} \mathrm{e}^{-\lambda^{(\gamma)}(K-1)}$ and

$$\begin{aligned}\int_K^\infty x^{p+1} F^{(\gamma)}(x) &\le \tilde{p}_+^{(\gamma)} \mathrm{e}^{-\lambda^{(\gamma)}(K-1)} \int_K^\infty x^{p+1} \lambda^{(\gamma)} \mathrm{e}^{-\lambda^{(\gamma)}(x-K)}\, dx\\ &= \frac{\tilde{p}_+^{(\gamma)} \mathrm{e}^{\lambda^{(\gamma)}}}{\lambda^{(\gamma)^{p+2}}} \int_{K\lambda^{(\gamma)}}^\infty y^{p+1} \mathrm{e}^{-y}\, dy.\end{aligned}$$

That $\mathscr{L}\mathscr{T}(p)$ holds now follows from

$$\limsup_{\gamma\to\infty} \frac{\mathrm{e}^{\lambda^{(\gamma)}}}{\lambda^{(\gamma)}} \int_{K\lambda^{(\gamma)}}^\infty y^{p+1} \mathrm{e}^{-y}\, dy = \limsup_{\gamma\to\infty} \frac{\mathrm{e}^{\lambda^{(\gamma)}}}{\lambda^{(\gamma)}} O\Big(\big(K\lambda^{(\gamma)}\big)^{p+1} \mathrm{e}^{-K\lambda^{(\gamma)}}\Big) = 0. \quad \square$$

Example 3.2. We take $\mathbb{P}(U > u) = \mathrm{e}^{-u^2}$, $T^{(\gamma)} = \gamma T_*$, $\mathbb{P}(T_* \le t) = \mathrm{e}^{-1/\sqrt{t}}$ (since the failure rate at $U = u$ is $2u$, Corollary 3.5 immediately shows that we have light traffic equivalence of all orders). We shall show that conditionally upon $U - \gamma T_* > 0$, we have

$$\text{(3.7)} \qquad \frac{U}{\gamma^{1/5}} \to_{\mathbb{P}} K, \qquad \gamma^{4/5} T_* \to_{\mathbb{P}} K, \qquad U - \gamma T_* \to_{\mathbb{P}} 0$$

[here $K = 4^{-2/5}$ is the unique point where $\varphi(k) = k^{-1/2} + k^2$ attains its minimum]. We thereby provide an example where it is necessary to have *both* long service times and short interarrival times if delay is to occur in light traffic, and we observe the somewhat peculiar phenomenon that $\mathbb{E}W^{(\gamma)}$ is of smaller order of magnitude than $p_+^{(\gamma)} = \mathbb{P}(W^{(\gamma)} > 0)$.

Obviously,

$$\mathbb{P}(U - \gamma T_* > 0) \ge \mathbb{P}\big(T \le K\gamma^{-4/5}, U > K\gamma^{1/5}\big) = \mathrm{e}^{-\gamma^{2/5}\varphi(K)}.$$

Furthermore, let $L > K$. Then

$$\mathbb{P}\big(U - \gamma T_* > 0, T_* > L\gamma^{-4/5}\big) = \int_{L\gamma^{-4/5}}^1 \frac{1}{2t^{3/2}} \mathrm{e}^{-1/\sqrt{t}} \, \mathrm{e}^{-\gamma^2 t^2}\, dt.$$

Substituting $s = t\gamma^{4/5}$, we have

$$\frac{1}{\sqrt{t}} + \gamma^2 t^2 = \gamma^{2/5}\left(\frac{1}{\sqrt{s}} + s^2\right) = \gamma^{2/5}\varphi(s) \le \gamma^{2/5}\varphi(L)$$

when $t \geq L\gamma^{-4/5}$. Thus

$$\frac{\mathbb{P}(U - \gamma T_* > 0, T_* > L\gamma^{-4/5})}{\mathbb{P}(U - \gamma T_* > 0)} \leq \frac{\gamma^{6/5}}{2L^{3/2}} e^{-\gamma^{2/5}(\varphi(L) - \varphi(K))},$$

which tends to 0 when $\gamma \to \infty$ because $\varphi(L) > \varphi(K)$. Similarly,

$$\frac{\mathbb{P}(U - \gamma T_* > 0, U > L\gamma^{1/5})}{\mathbb{P}(U - \gamma T_* > 0)} \to 0$$

so that

$$\mathbb{P}(T_* \leq K\gamma^{-4/5}, U \leq K\gamma^{1/5} | U - \gamma T_* > 0) \to 1.$$

That (3.7) holds now follows from

$$\lim_{t \downarrow 0} \mathbb{P}(T_* < (1 - \varepsilon)t | T_* \leq t) = 0, \qquad \lim_{u \uparrow \infty} \mathbb{P}(U_* > (1 + \varepsilon)u | U_* \geq u) = 0.$$

4. Light traffic approximations for GI/PH/1 queues. We now assume that $U = U^{(\gamma)} = U_*$ has a phase-type distribution, say with representation $(\pi, \mathbf{Q}, E)$ not dependent on the parameter γ. In the standard setup ([13] or [1], Chapter VIII.6), this means that we may think of U as the time until absorption in δ in a Markov process which moves on a finite state space E according to the intensity matrix $\mathbf{Q}$, has initial distribution π (written as a row vector) concentrated on $E(\pi e = 1$ where e is the column vector with all 1's) and is eventually absorbed in $\delta \notin E$. Note that $\mathbf{Q}$ is a subintensity (meaning $\mathbf{Q}e \leq 0$) since we may identify the entries of $q = -\mathbf{Q}e$ with the exit rates $E \to \delta$.

Standard analytical identities for the distribution function, the density, the moment generating function, resp. the moments, are

$$B(t) = \mathbb{P}(U \leq t) = 1 - \pi e^{\mathbf{Q}t} e, \tag{4.1}$$

$$b(t) = B'(t) = \pi e^{\mathbf{Q}t} q, \tag{4.2}$$

$$\hat{B}[s] = \int_0^\infty e^{st} B(dt) = \pi(-s\mathbf{I} - \mathbf{Q})^{-1} e, \tag{4.3}$$

$$\mathbb{E}U^n = \hat{B}^{(n)}[0] = n! \pi(-\mathbf{Q})^{-n} e. \tag{4.4}$$

Some further crucial properties of phase-type distributions are given by the following lemma.

LEMMA 4.1. *Let B be phase-type with representation $(\pi, \mathbf{Q}, E)$. Then*:

(a) *There exist $\eta > 0$, $k = 0, 1, \dots$ and c_1 such that*

$$\overline{B}(t) \approx c_1 t^k e^{-\eta t}, \qquad t \to \infty.$$

(b) *There exist c_2, c_3, c_4 and c_5 such that*

$$(c_2 + c_3 t^k) e^{-\eta t} \leq \overline{B}(t) \leq (c_4 + c_5 t^k) e^{-\eta t} \quad \textit{for all } t.$$

(c) *There exist c_6 and c_7 such that*

$$\overline{B}(s|t) \le \left(c_6 + c_7 s^k\right)\mathrm{e}^{-\eta s} \quad \textit{for all } s \textit{ and } t.$$

(d) $\overline{B}(s|t) \approx \mathrm{e}^{-\eta s}$ *when $t \to \infty$ with s fixed.*

PROOF. The lemma is essentially well known. For example, part (a) follows from the Perron–Frobenius theory (with $k = 0$) if $\mathbf{Q}$ is irreducible and can be obtained for the general case by writing $\mathbf{Q}$ in the Jordan canonical form. For part (b), one can first apply part (a) to get the desired result for all sufficiently large t and next adjust the constants to apply to all t. Part (c) follows from part (b) for $t \ge t_0$, whereas for $t \le t_0$ we can bound $\overline{B}(s+t)$ from above using part (b) and $\overline{B}(t)$ from below by $\overline{B}(t_0)$. Finally, part (d) is an easy consequence of part (a). □

In the definition of a phase-type distribution, it is occasionally convenient to allow π to be defective, $\pi e < 1$. This may either be interpreted as the phase-type distribution having the defect $1 - \pi e$ (as for ladder variables) or an atom of size $1 - \pi e$ at 0 (this occurs for the waiting time W in Theorem 4.1 below). We recall that $A^{(\gamma)}$ denotes the interarrival distribution and $\hat{A}^{(\gamma)}$ its moment-generating function [having possibly a matrix-valued argument as in (4.5) below]. The following representation of waiting-time distribution was recently obtained in [2] and is substantially simpler than those of the matrix-geometric literature (e.g., [15] and references therein; see, however, also [20]).

THEOREM 4.1. *$W^{(\gamma)}$ is phase-type with representation $(\nu^{(\gamma)}, \mathbf{Q} + q\nu^{(\gamma)}, E)$, where $\nu^{(\gamma)}$ is the unique subprobability vector satisfying*

$$\nu^{(\gamma)} = \pi\hat{A}^{(\gamma)}\left[\mathbf{Q} + q\nu^{(\gamma)}\right] = \int_0^\infty \pi\, \mathrm{e}^{(\mathbf{Q}+q\nu^{(\gamma)})t}\, A^{(\gamma)}(dt). \tag{4.5}$$

Note that in the GI/M/1 case, Theorem 4.1 reduces to the classical solution of the waiting-time problem which states that

$$\mathbb{P}(W^{(\gamma)} > x) = p_+^{(\gamma)}\mathrm{e}^{-\eta^{(\gamma)}x}, \quad \text{where } \eta^{(\gamma)} = \mu\left(1 - p_+^{(\gamma)}\right),$$

μ is the service intensity and $p_+^{(\gamma)}$ is the solution of

$$p_+^{(\gamma)} = \hat{A}^{(\gamma)}\left[\mu\left(1 - p_+^{(\gamma)}\right)\right]. \tag{4.6}$$

Indeed, here E reduces to one point and we have $\nu^{(\gamma)} = p_+^{(\gamma)}$, $\mu = -\mathbf{Q} = q$, $\pi = 1$.

The intuitive content of Corollary 4.1 below is that in light traffic one can neglect $\nu^{(\gamma)}$ on the r.h.s of (4.5) so that

$$\tilde{\nu}^{(\gamma)} = \pi\hat{A}^{(\gamma)}[\mathbf{Q}] = \int_0^\infty \pi\, \mathrm{e}^{\mathbf{Q}t}\, A^{(\gamma)}(dt) \tag{4.7}$$

is useful as an approximation to $\nu^{(\gamma)}$ and hence

$$\tilde{p}_{+}^{(\gamma)} = \int_0^\infty \pi \, e^{\mathbf{Q}t} eA^{(\gamma)}(dt) = \tilde{\nu}^{(\gamma)}e \tag{4.8}$$

is useful as an approximation to $p_{+}^{(\gamma)} = \nu^{(\gamma)}e$. Note the interpretation of $\tilde{\nu}^{(\gamma)}$ given by part (b) of the following easily proved lemma.

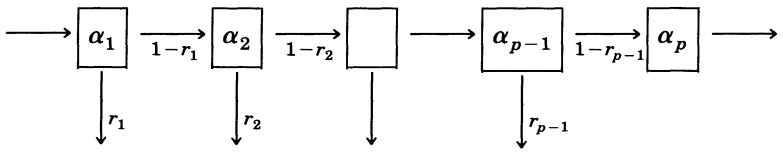
LEMMA 4.2. (a) *The distribution of* $(U - x)^+$ *is (defective) phase-type with representation* $(\pi \, e^{\mathbf{Q}x}, \mathbf{Q}, E)$.

(b) *The distribution of* $(U - T^{(\gamma)})^+$ *is (defective) phase-type with representation* $(\tilde{\nu}^{(\gamma)}, \mathbf{Q}, E)$.

Combining Lemmas 4.1(b) and 4.2(b) with Corollary 3.4, we now get the following corollary.

COROLLARY 4.1. *Consider the* GI/PH/1 *queue in light traffic and assume that* $T^{(\gamma)} \to_{\mathscr{D}} \infty$ *when* $\gamma \to \infty$, *with* $U^{(\gamma)} = U_* = U$ *or equivalently the phase-type representation fixed. Then* $W^{(\gamma)}$ *is light traffic equivalent (distributional and of any order p) to a phase-type random variable with representation* $(\tilde{\nu}^{(\gamma)}, \mathbf{Q}, E)$.

EXAMPLE 4.1. Assume that the service time U has a Coxian distribution corresponding to the phase diagram



Equivalently, $E = \{1, \dots, p\}$, π is degenerate at state 1 and

$$\mathbf{Q} = \begin{pmatrix} -\alpha_1 & \beta_1 & 0 & \cdots & 0 & 0 \\ 0 & -\alpha_2 & \beta_2 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & -\alpha_{p-1} & \beta_{p-1} \\ 0 & 0 & 0 & \cdots & 0 & \alpha_p \end{pmatrix},$$

where $\beta_i = \alpha_i(1 - r_i)$. We shall assume that no two α_i's are equal so that $\mathbf{Q}$ has diagonal form

$$\mathbf{Q} = \sum_{i=1}^{p} -\alpha_i h^{(i)} \otimes \mu^{(i)}, \tag{4.9}$$

where $\mu^{(i)}$ and $h^{(i)}$ are the left (row), resp. right (column), eigenvectors of $\mathbf{Q}$ corresponding to the eigenvalue $-\alpha_i$, normalized by $\mu^{(i)}h^{(i)} = 1$. By elementary calculations, it is readily checked that

$$\mu_j^{(i)} = \begin{cases} 0, & j < i, \\ 1, & j = i, \\ \dfrac{\beta_i \cdots \beta_{j-1}}{(\alpha_{i+1} - \alpha_i) \cdots (\alpha_j - \alpha_i)}, & j > i, \end{cases}$$

$$h_j^{(i)} = \begin{cases} \dfrac{\beta_j \cdots \beta_{i-1}}{(\alpha_j - \alpha_i) \cdots (\alpha_{i-1} - \alpha_i)}, & j < i, \\ 1, & j = i, \\ 0, & j > i. \end{cases}$$

Since (4.9) implies that

$$\text{(4.10)} \qquad \mathrm{e}^{\mathbf{Q}t} = \sum_{i=1}^{p} \mathrm{e}^{-\alpha_i t}\, h^{(i)} \otimes \mu^{(i)},$$

$$\text{(4.11)} \qquad \tilde{\nu} = \pi \hat{A}[\mathbf{Q}] = \sum_{i=1}^{p} \hat{A}[-\alpha_i] h_1^{(i)} \mu^{(i)},$$

it follows that the light traffic approximation for W is phase-type with representation $(\tilde{\nu}, \mathbf{Q}, E)$ with $\tilde{\nu}$ given by (4.11) or, equivalently, by the coordinate representation

$$\text{(4.12)} \quad \tilde{\nu}_j = \sum_{i=1}^{p} \hat{A}[-\alpha_i] \frac{\beta_1 \cdots \beta_{i-1}}{(\alpha_1 - \alpha_i) \cdots (\alpha_{i-1} - \alpha_i)(\alpha_{i+1} - \alpha_i) \cdots (\alpha_j - \alpha_i)}.$$

This distribution is not defective Coxian since $\tilde{\nu}$ is not concentrated at state 1, but rather a mixture $\Sigma_1^p \tilde{\nu}_j B^{(j)}$, where $B^{(j)}$ is the Coxian distribution obtained from the given one $B = B^{(1)}$ by deleting phases $1, \ldots, j-1$.

Note also that the diagonal form allows moments to be calculated readily. Thus, combining with (4.4), we get

$$\text{(4.13)} \qquad \mathbb{E}W^n \approx n! \sum_{i=1}^{p} \hat{A}[-\alpha_i] h_1^{(i)} \mu^{(i)} (-\mathbf{Q})^{-n} e = n! \sum_{i=1}^{p} \frac{\hat{A}[-\alpha_i]}{\alpha_i^n} \gamma_i,$$

where

$$\begin{aligned} \gamma_i &= h_1^{(i)} \mu^{(i)} e \\ &= \sum_{j=1}^{p} \frac{\beta_1 \cdots \beta_{i-1}}{(\alpha_1 - \alpha_i) \cdots (\alpha_{i-1} - \alpha_i)(\alpha_{i+1} - \alpha_i) \cdots (\alpha_j - \alpha_i)}. \end{aligned}$$

Even though the model of Example 4.1 is fairly general and the solution reasonably simple, there may be cases where alternative expressions are more convenient and illuminating. Consider first the class $\mathscr{S}_\alpha$ of [10].

COROLLARY 4.2. *Assume $T^{(\gamma)} = \gamma T_*$ with T_* in the class $\mathscr{S}_\alpha$. Then $W^{(\gamma)}$ is light traffic equivalent to the phase-type distribution with representation $(\gamma^{-\alpha} c_{A_*} \pi^{(\alpha)}, \mathbf{Q}, E)$, where $\pi^{(\alpha)} = \Gamma(\alpha + 1)\pi(-\mathbf{Q})^{-\alpha}$.*

[Note that when α is not an integer, $(-\mathbf{Q})^{-\alpha}$ may still be given a meaning in the standard functional analytic sense.]

PROOF. Using integration by parts, we get

$$\begin{aligned}\tilde{\nu}^{(\gamma)} &= \int_0^\infty \pi\, e^{\mathbf{Q}\gamma t} A(dt)\\ &= -\int_0^\infty \gamma\pi\mathbf{Q}\, e^{\mathbf{Q}\gamma t} A(t)\, dt\\ &= -\int_0^\infty \pi\mathbf{Q}\, e^{\mathbf{Q}u} A\left(\frac{u}{\gamma}\right) du.\end{aligned}$$

Here $\gamma^\alpha A(u/\gamma)$ is bounded with limit $c_{A_*} u^\alpha$ and the components of $e^{\mathbf{Q}u}$ decay exponentially fast. Thus, by dominated convergence,

$$\begin{aligned}\tilde{\nu}^{(\gamma)} &\approx \gamma^{-\alpha} \int_0^\infty \pi\mathbf{Q}\, e^{\mathbf{Q}u} c_{A_*} u^\alpha\, du\\ &= \gamma^{-\alpha}\Gamma(\alpha + 1) c_{A_*} \mathbf{Q}(-\mathbf{Q})^{-\alpha-1} = \gamma^{-\alpha} c_{A_*} \pi^{(\alpha)}.\end{aligned}$$ □

As a motivating example for the following, consider the D/PH/1 queue where $\overline{B}(x|\gamma T_*) \approx e^{-\eta x}$ in view of Lemma 4.2(d) so that

$$\begin{aligned}\mathbb{P}(W^{(\gamma)} > x) &\approx \mathbb{P}(U > \gamma T_* + x) = \mathbb{E}\overline{B}(x|\gamma T_*)\overline{B}(\gamma T_*)\\ &\approx e^{-\eta x}\, \mathbb{E}\overline{B}(\gamma T_*) = e^{-\eta x} p_+^{(\gamma)}.\end{aligned}$$

That is, in the limit $W^{(\gamma)}$ is conditionally exponential. To generalize this, we shall invoke the following concept.

DEFINITION 4.1. A random variable T_* with m.g.f. $\hat{A}_*[s]$ belongs to the class $\mathscr{S}_\infty$ if

$$\frac{\hat{A}_*[-\gamma u]}{\hat{A}_*[-\gamma v]} \to 0 \quad \text{as } \gamma \to \infty \text{ when } u > v, \tag{4.14}$$

$$\frac{\gamma^l \hat{A}_*^{(l)}[-\gamma u]}{\gamma^k \hat{A}_*^{(k)}[-\gamma u]} \to 0 \quad \text{as } \gamma \to \infty \text{ when } 0 \le l < k. \tag{4.15}$$

Then, with η, k and c_1 as in Lemma 4.1(a), we obtain the following corollary.

COROLLARY 4.3. *Assume $T^{(\gamma)} = \gamma T_*$ with T_* in the class $\mathscr{S}_\infty$. Then $W^{(\gamma)}$ is light traffic equivalent to the exponential distribution with density $c_1 \gamma^k \hat{A}_*^{(k)}[-\gamma\eta]\eta\, e^{-\eta x}$.*

PROOF. Using the whole Jordan canonical form of $\mathbf{Q}$, we may write

$$\overline{B}(t) = \sum_{l=1}^{k} c_{k-l+1} t^l \,\mathrm{e}^{-\eta t} + O(\mathrm{e}^{-\eta_1} t),$$

$$b(t) = \sum_{l=1}^{k} c_{k-l+1}^{(1)} t^l \,\mathrm{e}^{-\eta t} + O(\mathrm{e}^{-\eta_1} t),$$

where $0 \le \eta_1 < \eta$, $c_i^{(1)} = \eta c_i$. Then

$$\begin{aligned}\bar{p}_+^{(\gamma)} &= \int_0^\infty \overline{B}(\gamma t) A_*(dt) \\ &= \sum_{l=1}^{k} c_{k-l+1} \gamma^l \hat{A}_*^{(l)}[-\gamma\eta] + O\big(\hat{A}_*^{(l)}[-\gamma\eta_1]\big),\end{aligned}$$

which by assumption behaves like $c_1 \gamma^k \hat{A}_*^{(k)}[-\gamma\eta]$. Furthermore, we may write

$$b(\gamma t + x) = \mathrm{e}^{-\eta x} \sum_{l=1}^{k} c_{k-l+1}^{(2)} t^l \,\mathrm{e}^{-\eta t} + O(\mathrm{e}^{-\eta_1} t),$$

where $c_1^{(2)} = c_1^{(1)} = \eta c_1$ (the remaining $c_i^{(2)}$ depend on x). Thus similarly, we get the density of $W^{(\gamma)}$ at x as

$$\begin{aligned}&\int_0^\infty b(\gamma t + x) A_*(dt) \\ &\quad = \mathrm{e}^{-\eta x} \sum_{l=1}^{k} c_{k-l+1}^{(2)} \gamma^l \hat{A}_*^{(l)}[-\gamma\eta] + O\big(\hat{A}_*^{(l)}[-\gamma\eta_1]\big) \\ &\quad \approx \mathrm{e}^{-\eta x} c_1^{(2)} \gamma^k \hat{A}_*^{(k)}[-\gamma\eta] \eta \,\mathrm{e}^{-\eta x} = c_1 \gamma^k \hat{A}_*^{(k)}[-\gamma\eta] \eta \,\mathrm{e}^{-\eta x}. \qquad \square\end{aligned}$$

Easy estimates show that any distribution A_* with support contained in $[\varepsilon, \infty)$ is in $\mathscr{S}_\infty$. More generally, we have the following proposition.

PROPOSITION 4.1. *A sufficient condition for a distribution* A_* *to be in* $\mathscr{S}_\infty$ *is*

$$A_*\left(\frac{r}{\gamma}\right) = o(\mathrm{e}^{-\gamma s}) \quad \textit{for all } r < \infty,\ s > 0. \tag{4.16}$$

PROOF. Choose $s, \varepsilon > 0$ with $\delta = A_*(s) - A_*(s - \varepsilon) > 0$. Then when γ is so large that $r/\gamma < s - \varepsilon$,

$$\frac{\int_0^{r/\gamma} (\gamma t)^l \,\mathrm{e}^{-\eta t} A_*(dt)}{\int_{r/\gamma}^{\infty} (\gamma t)^l \,\mathrm{e}^{-\eta t} A_*(dt)} \le \frac{\int_0^{r/\gamma} \mathrm{e}^{-\eta t} A_*(dt)}{\int_{r/\gamma}^{\infty} \mathrm{e}^{-\eta t} A_*(dt)} \le \frac{A_*(r/\gamma)}{\delta \,\mathrm{e}^{-\gamma s}} \to 0$$

for $l = 0, 1, \ldots$. In particular,

$$\frac{\hat{A}_*[-\gamma u]}{\hat{A}_*[-\gamma v]} \approx \frac{\int_{r/\gamma}^{\infty} \mathrm{e}^{-\gamma u t} A_*(dt)}{\int_{r/\gamma}^{\infty} \mathrm{e}^{-\gamma v t} A_*(dt)} \leq \mathrm{e}^{-r(u-v)},$$

$$\frac{\gamma^l \hat{A}_*^{(l)}[-\gamma u]}{\gamma^k \hat{A}_*^{(k)}[-\gamma u]} \approx \frac{\int_{r/\gamma}^{\infty} (\gamma t)^l \mathrm{e}^{-\gamma u t} A_*(dt)}{\int_{r/\gamma}^{\infty} (\gamma t)^k \mathrm{e}^{-\gamma u t} A_*(dt)} \leq \frac{1}{r^{k-l}},$$

and the desired conclusion follows by letting $r \to \infty$. □

A sufficient condition for (4.16) and hence the $\mathscr{S}_\infty$ property is $a_*(t) = O(\mathrm{e}^{-t^{-\alpha}})$, $t \to 0$, for some $\alpha > 1$ (here $a_* = A'_*$ is the density). An example not covered by this is the inverse Gaussian distribution where

$$\hat{A}_*[-u] = \exp\left\{\xi c - \sqrt{2\xi + u}\right\}$$

([1], page 263) and the $\mathscr{S}_\infty$ property follows by easy explicit calculus. Note that here $a_*(t)$ is essentially of the order of magnitude $\mathrm{e}^{-c^2/2t}$ when $t \to 0$, so that we are on the border of (4.16).

5. Concluding remarks. Since the results of Section 4 and those of Reiman and Simon [18] for the dual problem of phase-type arrivals are more explicit and easier to compute than for general GI/G/1 queues, it seems reasonable to ask whether one could not simply approximate either U or T by a phase-type random variable and use the resulting simplification in analysis. Using the explicit solution (say Theorem 4.1) for a fixed queueing system, this procedure is basically correct (the steady-state characteristics are continuous under weak conditions; see, e.g., [3] and the references therein), but for the purpose of light traffic approximations it seems that one can easily obtain quite misleading results. Suppose, for example, that U is exponential with rate η and $T^{(\gamma)} = \gamma T_*$ with $T_* = 1$ so that $\mathbb{P}(W^{(\gamma)} > 0) \approx \mathrm{e}^{-\gamma\eta}$, say by Corollary 4.3. If we approximate the distribution of T_* by an Erlang distribution with k phases (with k large) and mean 1, (3.5) leads to $\mathbb{P}(W^{(\gamma)} > 0) = O(\gamma^{-k})$ which is of a different order of magnitude. For a given moderately small traffic intensity ρ, the difference may or may not be really crucial, but the effect is certainly worrying if one goes all the way to the light traffic limit as is done, for example, for computing the constants underlying the interpolation approximations in [8], [17] and [21]. Some empirical investigations seem to be needed here.

Similar examples are easily constructed to show that the fact that a phase-type approximation may not mimic the tail behaviour of U may be fatal. In short, light traffic behaviour is sensitive to certain small changes in the distribution. Thus our point in Section 4 is certainly not to say that phase-type approximations are always the proper tool in light traffic, but rather to provide some more detailed examples of light traffic behaviour. The lesson for the practitioner to be learned is that phase-type (or other types of) approximations should be done with care in light traffic, and it is crucial that one gets a good fit of the tail of the service-time distribution and the behaviour near 0 of the interarrival-time distribution (note, however, that the tail of the service-time distribution matters somewhat less when the interarrival distribution is in, say, $\mathscr{S}_\alpha$ rather than in $\mathscr{S}_\infty$; compare Corollaries 3.2 and 4.3).

For practical purposes, the GI/G/1 queue is sometimes argued to be a toy model, and it would seem that at present the work of Reiman and Simon [18] has the greatest potential in the direction of incorporating very complex models in light traffic analysis. However, a Markovian structure (often paramount to phase-type assumptions) is needed in [18]. For non-Markovian models, one can readily guess what the extensions of the results of the present paper to the many-server queue GI/G/k should be. Proofs of parts of this are in [11], but the general case is still open. For even more general queueing systems having general stationary arrival streams, even the intuition behind the present paper (the effect of the single customer) breaks down.

Acknowledgments. The seminal debt to Daryl Daley and Tomasz Rolski has already been acknowledged. I am also grateful to them and the referees for a very careful reading of the manuscript which helped eliminate numerous errors and clarify the exposition.

REFERENCES

[1] Asmussen, S. (1987). *Applied Probability and Queues*. Wiley, Chichester.

[2] Asmussen, S. (1992). Phase-type representation in random walk and queueing problems. *Ann. Probab.* **20** 772–789.

[3] Asmussen, S. and Johansen, H. (1986). Über eine Stetigkeitsfrage betreffend das Bedienungssystem GI/GI/s. *Elektron. Informationsverarb. Kybernet.* **22** 565–570.

[4] Blomqvist, N. (1969). Estimation of waiting-time parameters in the GI/G/1 queueing system. II. Heavy-traffic approximations. *Skand. Aktuar. Tidsskr.* **1969** 125–136.

[5] Bloomfield, P. and Cox, D. R. (1972). A low traffic approximation for queues. *J. Appl. Probab.* **9** 832–840.

[6] Burman, D. Y. and Smith, D. R. (1983a). A light traffic approximation for multi-server queues. *Math. Oper. Res.* **8** 15–25.

[7] Burman, D. Y. and Smith, D. R. (1983b). Asymptotic analysis of a queueing model with bursty traffic. *Bell System Tech. J.* **62** 1433–1452.

[8] Burman, D. Y. and Smith, D. R. (1986). An analysis of a queueing system with Markov-modulated arrivals. *Oper. Res.* **34** 105–119.

[9] Daley, D. and Rolski, T. (1984). A light traffic approximation for a single-server queue. *Math. Oper. Res.* **9** 624–628.

[10] Daley, D. and Rolski, T. (1991). Light traffic approximations in queues. *Math. Oper. Res.* **16** 57–71.

[11] Daley, D. and Rolski, T. (1992). Light traffic approximations in many server queues. *Adv. in Probab.* To appear.

[12] Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, 2nd ed. Wiley, New York.

[13] Neuts, M. F. (1981). *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins Univ. Press.

[14] Pinedo, M. and Wolff, R. W. (1982). A comparison between tandem queues with dependent and independent service times. *Oper. Res.* **30** 464–479.

[15] Ramaswami, V. and Lucantoni, D. (1988). Moments of the stationary waiting time distribution in the GI/PH/1 queue. *J. Appl. Probab.* **25** 636–641.

[16] Reiman, M. I. and Simon, B. (1988a). Light traffic limits of sojourn time distributions in Markovian queueing networks. *Stochastic Models* **4** 191–223.

[17] Reiman, M. I. and Simon, B. (1988b). An interpolation approximation for queueing systems with Poisson input. *Oper. Res.* **36** 454–469.

[18] Reiman, M. I. and Simon, B. (1989). Open queueing systems in light traffic. *Math. Oper. Res.* **14** 26–59.

[19] Reiman, M. I. and Weiss, A. (1989). Light traffic derivatives via likelihood ratios. *IEEE Trans. Inform. Theory* **35** 648–654.

[20] Sengupta, B. (1989). Markov processes whose steady-state distribution is exponential with an application to the GI/PH/1 queue. *Adv. in Appl. Probab.* **21** 159–180.

[21] Whitt, W. (1989). An interpolation approximation for the mean workload in a GI/G/1 queue. *Oper. Res.* **37** 936–952.

[22] Wolff, R. W. (1982). Tandem queues with dependent service times in light traffic. *Oper. Res.* **30** 619–635.

[23] Wolff, R. W. (1984). Conditions for finite ladder height and delay moments. *Oper. Res.* **32** 909–916.

Institute of Electronic Systems
Aalborg University
DK-9220 Aalborg
Denmark