An Overview of Brownian and Non-Brownian FCLTs for the Single-Server Queue

Ward Whitt [a]

[a] *AT&T Labs, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971*
E-mail: wow@research.att.com

Queueing Systems, to appear

We review functional central limit theorems (FCLTs) for the queue-content process in a single-server queue with finite waiting room and the first-come first-served service discipline. We emphasize alternatives to the familiar heavy-traffic FCLT with reflected Brownian motion (RBM) limit process that arise with heavy-tailed probability distributions and strong dependence. Just as for the familiar convergence to RBM, the alternative FCLTs are obtained by applying the continuous mapping theorem with the reflection map to previously established FCLTs for partial sums. We consider a discrete-time model and first assume that the cumulative net-input process has stationary and independent increments, with jumps up allowed to have infinite variance or even infinite mean. For essentially a single model, the queue must be in heavy traffic and the limit is a reflected stable process, whose steady-state distribution can be calculated by numerically inverting its Laplace transform. For a sequence of models, the queue need not be in heavy traffic, and the limit can be a general reflected Lévy process. When the Lévy process representing the net input has no negative jumps, the steady-state distribution of the reflected Lévy process again can be calculated by numerically inverting its Laplace transform. We also establish FCLTs for the queue-content process when the input process is a superposition of many independent component arrival processes, each of which may exhibit complex dependence. Then the limiting input process is a Gaussian process. When the limiting net-input process is also a Gaussian process and there is unlimited waiting room, the steady-state distribution of the limiting reflected Gaussian process can be conveniently approximated.

Keywords: queues, approximations, heavy traffic, stable laws, Lévy processes, infinitely divisible distributions, Gaussian processes, supremum of a Gaussian process, functional central limit theorems, invariance principles, Laplace transforms, numerical transform inversion

1. Introduction

In this paper we discuss extensions of the familiar heavy-traffic functional central limit theorem (FCLT) for the single-server queue in which the limit process is reflected Brownian motion (RBM) and the approximating steady-state distribution is exponential. By changing the assumptions, we obtain different limits with different scaling, different limit processes and different approximating steady-state distributions.

We assume that our queueing process of interest is a discrete-time process satisfying the finite-capacity generalization of the classical Lindley recursion, i.e.,

$$Q(k) = \max\{0, \min\{C, Q(k-1) + X(k)\}\} \ , \tag{1.1}$$

where $X(k)$ is the *net input* between periods $k-1$ and k; i.e.,

$$X(k) = A(k) - B(k), \quad k \geq 1 \ , \tag{1.2}$$

$A(k)$ is a nonnegative *input* and $B(k)$ is a nonnegative *potential* (maximum possible) *output.* The variable $Q(k)$ depicts the *queue* (or buffer) *content* in period k. In the classical Lindley recursion associated with the GI/GI/1/∞ queue, $C = \infty$, $Q(k)$ is the waiting time of the k^{th} customer before beginning service, $A(k)$ is the service time of the $(k-1)^{\text{st}}$ customer and $B(k)$ is the interarrival time between the $(k-1)^{\text{st}}$ and k^{th} customers. A popular model for communication networks is (1.1) with $B_k = c$, where c is a deterministic service capacity per period.

To obtain continuous-time stochastic processes as limits, we embed the discrete-time processes in continuous time by writing $Q(\lfloor t \rfloor)$, where $\lfloor t \rfloor$ is the greatest integer less than t, and then scale space and time, so that we work the scaled continuous-time process $\{c_n^{-1} Q(\lfloor nt \rfloor)\colon\ t \geq 0\}$ for normalization constants c_n.

First, assuming that the net inputs $X(k)$ are i.i.d., we allow the single-period inputs $A(k)$ to have infinite variances and even infinite means. It is known that a proper steady-state distribution exists for the queue-content process in this context if and only if the single-period input $A(k)$ has finite mean. If the single-period input has infinite mean, then we use the limit theorem to describe the transient behavior of the queue-content process.

The standard framework for heavy-traffic limit theorems involves a sequence of queueing processes associated with a sequence of queueing models, which we take to be indexed by n. However, it often suffices to consider essentially a single model, in the sense that the basic sequence $\{(A_n(k), B_n(k)) : k \geq 1\}$ associated with model n is a simple multiplicative scaling of the basic sequence $\{(A_1(k), B_1(k) : k \geq 1\}$ associated with model 1, i.e.,

$$A_n(k) = a_n A_1(k) \quad \text{and} \quad B_n(k) = b_n B_1(k) \text{ for } k \geq 1 \text{ and } n \geq 1 \ . \tag{1.3}$$

The familiar RBM limits can be obtained for more general model sequences, but it suffices to use the scaling (1.3).

Using the framework (1.3), and assuming that $\{X_n(k) : k \geq 1\}$ is i.i.d., we allow the input $A_1(1)$ to have infinite variances. We then show that a heavy-traffic FCLT holds for the queue-content process if and only if $A_1(1)$ has a power tail, i.e., $P(A_1(1) > x) \sim x^{-\alpha}$ for $0 < \alpha < 2$ (or, more generally, a regularly varying tail). Then the limit process is a reflected stable process.

We also consider sequences of queue-content processes associated with a more general sequence of models, not constrained to satisfy (1.3). Then we obtain FCLTs for the queue-content process that do not require that the queue be in heavy traffic. When heavy traffic does not prevail, these FCLTs can be thought of as model continuity results. Assuming that the net input sequences $\{X_n(k) : k \geq 1\}$ are i.i.d. for each n, we obtain conditions for the convergence of the sequence of normalized queue-content processes to reflected Lévy processes.

Finally, we also consider limits for the queue-content process when the inputs are superpositions of many independent component input processes, where each component input process may exhibit complex (e.g., long-range) dependence. Then we obtain reflected Gaussian processes as limits.

Interest in approximations associated with such alternative conditions has grown in recent years because of efforts to model evolving communication networks. Network traffic measurements have revealed complex stochastic behavior such as heavy-tailed probability distributions, long-range dependence and self-similarity. For background and recent related work, see Barford and Crovella [7], Boxma and Cohen [14]–[16], Cohen [25], Furrer, Michna and Weron [30], Gaver and Jacobs [31], Konstantopoulos and Lin [47], Kurtz [49], Norros [53], Resnick and Rootzén [57], Resnick and Samorodnitsky [58], Resnick and van den Berg [59], Taqqu, Willinger and Sherman [63], Tsoukatos and Makowski [64,65] and Willinger, Taqqu, Sherman and Wilson [74].

The limit theorems are of interest because they yield relatively simple approximations for intractable performance measures in complex systems. (However, even the scaling in a limit theorem by itself can be very useful; e.g., see Whitt [69].) We thus are particularly concerned about having ways to calculate the approximating distributions. Unfortunately, the approximations stemming from the nonstandard limits are more complicated than the approximations stemming from the standard heavy-traffic limit. A main contribution here is to show how the approximating distributions can be conveniently calculated in each case. For this purpose, a major technique is numerical inversion of Laplace transforms, as in Abate and Whitt [1].

To place our results in perspective, we briefly review the heavy-traffic theory. The first heavy-traffic limit theorem was obtained by Kingman [44,45]. By applying asymptotics to the previously determined transform of the steady-state queue content $Q(\infty)$ in the case $C = \infty$, Kingman showed that the relatively complicated steady-state distribution is asymptotically exponential as the traffic intensity $\rho \equiv EA(k)/EB(k)$ approaches 1, leading to the approximations

$$P(Q(\infty) > x) \approx e^{-x/EQ(\infty)} \tag{1.4}$$

and

$$EQ(\infty) \approx \frac{EA(1)\rho(c_A^2 + c_B^2)}{2(1-\rho)} \ , \tag{1.5}$$

where c_A^2 and c_B^2 are the squared coefficients of variation (SCV, variance divided by the square of the mean) of $A(1)$ and $B(1)$, respectively. A key condition, which we will be changing, is that the SCV's c_A^2 and c_B^2 in (1.5) be finite.

As reviewed in Whitt [67], it was discovered by Prohorov [55], Borovkov [13], Iglehart and Whitt [37,38] and Whitt [66] that this heavy-traffic limit and others could be approached via FCLTs (convergence in distribution for random elements of function spaces generated by partial sums and other basic processes). The essential idea is that the queue-content process can be represented, either exactly or to within an asymptotically negligible error, as a reflection map applied to an associated cumulative net-input process, i.e., the partial sums $S(k) = X(1)+\cdots+X(k)$. In the function space setting (with appropriate topology), the reflection map can be shown to be a continuous function. For background on the function space setting, see Billingsley [10], Ethier and Kurtz [27] and Jacod and Shiryaev [39]. Thus continuous mapping theorems can be applied to deduce that limits hold for a sequence of queue-content processes whenever limits hold for the associated sequence of cumulative net input processes. When $\{X(k) : k \geq 1\}$ is assumed to be i.i.d. with finite second moments, the limit holds by virtue of Donsker's FCLT; e.g., see Billingsley [10]. As a consequence, the scaled queue-content process $\{(1-\rho)Q(\lfloor t(1-\rho)^{-2}\rfloor) : t \geq 0\}$ can be approximated by RBM, which has an exponential steady-state distribution. By considering the iterated limit in which first $\rho \to 1$ and then $t \to \infty$, we obtain again the approximations (1.4) and (1.5).

The heavy-traffic FCLT is of interest, beyond Kingman's earlier result, because it is a limit for the entire queue-content process, which can generate approximations for many different functions of the queue content process, such as $\max\{Q(k) : 0 \leq k \leq n\}$. Even more important, however, is the fact that the limit can be established for more general models, for which explicit expressions for the steady-state distribution are not known. As shown by Abate, Choudhury and Whitt [3,4], the exact steady-state queue-content distribution when $\{X(k)\}$ is i.i.d. can be calculated from previously derived transforms. Thus, although the exponential approximation in (1.4) is very helpful under the i.i.d. assumption, it is not absolutely crucial. In contrast, FCLTs have been proved for net-input processes in much more general situations, allowing various forms of dependence among the interarrival times and service times. Many such theorems for net-input processes are contained in Jacod and Shiryaev [39]. In many of these situations, the steady-state queue-content distribution is unavailable. The heavy-traffic FCLTs again lead to approximations of the form (1.4) and (1.5), but now where the variability parameters depend on the asymptotic variances of the sequences $\{A(k) : k \geq 1\}$ and $\{B(k) : k \geq 1\}$; see Theorem 2 of Iglehart and Whitt [38] and Fendick, Saksena and Whitt [29].

As shown by Iglehart and Whitt [38] (for the special case of acyclic networks), Harrison and Reiman [36] and Reiman [56], the heavy-traffic limits extend from one queue to single-class networks of queues by essentially the same reasoning, using a multidimensional reflection mapping. More recently, attention has been focused on the difficult and important extension to multiclass networks, which involves different reasoning; see Bramson [17] and Williams [72,73]. However, we only discuss a single queue here.

By its nature, the FCLT approach extends directly to other conditions. The continuous-mapping argument implies that convergence of appropriately normalized net-input processes to a limit process will translate into corresponding convergence of the normalized queueing processes to a reflected limit process. This may involve new normalization constants, new limit processes and a modification of the function-space topology, but the argument is essentially the same. Thus the limiting results described here have been known for twenty years or more. Consequently, to a large extent the present paper should be regarded as a review.

Until recently, there has been little motivation for delving into this further. First, it was not evident that such generalizations can have much practical application and, second, it was not evident that the approximating process could be usefully described. However these situations have changed: There now is strong interest in these alternative limits because of the observed heavy-tailed probability distributions and strong dependence. Moreover, the numerical transform inversion makes it possible to calculate the approximating distributions.

Here is how the rest of this paper is organized: In Section 2 we review the general limit theorem for the single-server queue, which involves a sequence of queueing models. In Section 3 we consider the important special case of essentially a single model, as specified by (1.3). The essentially-single-model framework greatly restricts the class of possible limit processes. With i.i.d. summands and infinite variances, we get only convergence to a reflected stable process, which is characterized by only a few parameters. Hence, both with and without finite variances, the heavy-traffic limit theorems in the essentially-single-model framework yield parsimonious approximations.

In Section 4 we consider more general limits for more general sequences of queueing models. Particularly tractable are reflected Lévy process limits where the Lévy process has no negative jumps. That case is also most relevant in applications because it occurs when there are exceptionally large inputs. Just as in Section 3, the steady-state distributions of these limit processes can be readily calculated by numerical transform inversion.

However, the reflected Lévy processes are much more complicated than reflected stable processes, because they have essentially infinitely many parameters (the Lévy measure). They have the advantage of being more flexible, but the disadvantage of being harder to fit. Lévy processes are characterized by their infinitely divisible (ID) one-dimensional marginal distributions. Since many common distributions are ID, see Bondesson [12], the limit theorem does not greatly restrict the class of candidate approximating distributions. For example, Pareto, Weibull, lognormal and hyperexponential distributions are all ID. The general reflected Lévy process limits suggest a potentially useful class of approximations.

However work is still needed on methods for approximating net-input processes by Lévy processes. Direct analysis of models in this framework is discussed by Takács [62] and Prabhu [54].

We conclude in Section 5 by briefly considering limits for single-server queueing models in which the input is a superposition of many independent component arrival streams, allowing very general dependence in each component arrival process. This case is motivated by communication networks in which many sources are multiplexed together. Then, by the central limit theorem, the total input can be approximated by a Gaussian process. If the queue output is deterministic or if heavy-traffic conditions prevail, then the net-input process can be regarded as Gaussian and the queueing process can be approximated by a reflected Gaussian process. This produces considerable simplification because a stationary Gaussian process is characterized by its mean and its autocovariance function. As shown by Willinger et al. [74] and Taqqu et al. [63], subsequent limits after scaling time and space can lead to convergence to special Gaussian processes such as fractional Brownian motion (FBM). Conditions for convergence to FBM are also determined by Kurtz [49].

2. The Basic FCLT

We start with the queueing model defined in (1.1) and (1.2). After embedding the discrete-time processes in continuous time, the original discrete-time processes as well as the continuous-time limit process can be regarded as random elements of the same space as the limit processes. With that continuous-time representation, the process $\{Q(k) : k \geq 0\}$ and associated heavy-traffic limit processes can be defined in terms of a two-sided reflection (or regulator) map; see p. 22 of Harrison [35], Chen and Mandelbaum [19,20] and p. 16 of Berger and Whitt [8]. For this purpose, let $D \equiv D[0, \infty)$ be the space of right-continuous real-valued functions on $[0, \infty)$ with left limits at all positive times. Let D^k be the k-fold product space. The reflection map R takes elements of $x \in D$ with $0 \leq x(0) \leq C$ into (z, l, u) in D^3, where

$$z(t) \equiv R(x)(t) = x(t) + l(t) + u(t), \quad t \geq 0 , \tag{2.1}$$

$l(t)$ and $u(t)$ have nondecreasing sample paths with $l(0) = u(0) = 0$, $l(t)$ increases only when $z(t) = 0$ and $u(t)$ increases only when $z(t) = C$; i.e.,

$$\int_0^\infty z(t)dl(t) = \int_0^\infty [z(t) - C]du(t) = 0 . \tag{2.2}$$

In the infinite-capacity case, we have no upper barrier C and no increasing process u in (2.1). Then the reflection map has the equivalent from

$$R(x)(t) = x(t) - \min\{0, \inf\{x(s) : 0 \leq s \leq t\}\} . \tag{2.3}$$

The minimum in (2.3) is unnecessary when $x(0) = 0$.

Exploiting the embedding into continuous time, we obtain

$$Q(\lfloor t \rfloor) = R(S)(\lfloor t \rfloor), \quad t \geq 0 \ , \tag{2.4}$$

where $\{S(k)\}$ is the *cumulative net-input sequence*, i.e.,

$$S(k) = X(1) + \cdots + X(k), \quad k \geq 1 \ . \tag{2.5}$$

In fact, the representation (2.4) is easy to verify in discrete time by induction. At each step, the new increment $X(k)$ is either positive or negative, so that we will want to increase at most one of l and u in (2.1).

The relation (2.4) allows us to apply the continuous mapping theorem to obtain a limit theorem for the queue-content process Q whenever we have a limit theorem for the partial sums S in (2.5) e.g., see Theorem 5.1 of Billingsley [10] and Whitt [68]. To invoke a continuous mapping theorem, we need a topology on the function space D. It suffices to use the standard Skorohod [61] J_1 topology; see Ethier and Kurtz [27], Jacod and Shiryaev [39] and Whitt [68]. (We remark that the situation can be different for continuous-time models; then we may need to work with the less familiar Skorohod [61] M_1 topology; see Konstantopoulos and Lin [47] and Whitt [70,71].)

To formulate a general heavy-traffic limit theorem, we consider a sequence of queueing systems indexed by n. The n^{th} queueing system is assumed to satisfy the recursion (1.1) with capacity C_n, net-input sequence $\{X_n(k) : k \geq 1\}$ and initial content $Q_n(0)$. We get a FCLT for the scaled process $Q_n(\lfloor nt \rfloor)/c_n$ under an assumed FCLT for $S_n(\lfloor nt \rfloor)/c_n$, where $S_n(k) = X_n(1) + \cdots + X_n(k)$. Let $\Rightarrow$ denote convergence in distribution.

Theorem 1. Let c_n be deterministic scaling values and let the capacity in model n be $C_n \equiv Cc_n$ for $0 < C < \infty$. If $Q_n(0)/c_n \Rightarrow Q(0)$ in $\mathbb{R}$ as $n \to \infty$ and

$$\frac{S_n(\lfloor nt \rfloor)}{c_n} \Rightarrow S^*(t) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ , \tag{2.6}$$

then

$$\frac{Q_n(\lfloor nt \rfloor)}{c_n} \Rightarrow Q(t) \equiv R(S^*)(t) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ . \tag{2.7}$$

Proof. The assumptions allow us to write

$$c_n^{-1} Q_n(\lfloor n \cdot \rfloor) = R(c_n^{-1} S_n(\lfloor n \cdot \rfloor)) \tag{2.8}$$

for R in (2.1), so that we can apply the continuous mapping theorem with the two-sided reflection map. $\square$

The obvious consequence is the convergence of the one-dimensional distributions.

Corollary 2. Under the conditions of Theorem 1, if the limit process $\{R(S^*)(t) : t \geq 0\}$ is continuous at t with probability 1, then

$$c_n^{-1} Q_n(\lfloor nt \rfloor) \Rightarrow R(S^*)(t) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \ . \tag{2.9}$$

We apply the limit (2.9) to help justify the approximation

$$Q_n(k) \approx c_n R(S^*)(k/n) \quad \text{for each } k \ . \tag{2.10}$$

Theorem 1 does not directly imply convergence of the steady-state distributions, but assuming that $Q_n(k) \Rightarrow Q_n(\infty)$ as $k \to \infty$ for each n and $R(S^*)(t) \Rightarrow R(S^*)(\infty)$ as $t \to \infty$, we also apply Theorem 1 and Corollary 2 to help justify the approximation

$$Q_n(\infty) \approx c_n R(S^*)(\infty) \ . \tag{2.11}$$

In addition to establishing convergence for finite-dimensional distributions, the FCLT in Theorem 1 is useful to establish convergence of various functionals of the queueing process. Just as in the proof of Theorem 1, we apply the continuous mapping theorem. We illustrate with one corollary.

Corollary 3. Under the assumptions of Corollary 2,

$$c_n^{-1} \sup\{Q_n(\lfloor ns \rfloor) : 0 \leq s \leq t\} \Rightarrow \sup\{R(S^*)(s) : 0 \leq s \leq t\} \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty$$

for S^* in (2.6).

The standard way to establish the condition of Theorem 1 is to establish a joint FCLT for the partials sums of the single-period inputs and outputs. For this purpose, let $S_n^a(k)$ and $S_n^b(k)$ be the k-fold partial sums in the n^{th} model, i.e.,

$$S_n^a(k) = \sum_{j=1}^{k} A_n(j) \quad \text{and} \quad S_n^b(k) = \sum_{j=1}^{k} B_n(j) \ . \tag{2.12}$$

Theorem 4. Let c_n be the deterministic scaling values. If

$$c_n^{-1}\left(S_n^a(\lfloor nt \rfloor) - a_n nt,\ S_n^b(\lfloor nt \rfloor) - b_n nt\right) \Rightarrow \left(S^a(t), S^b(t)\right) \quad \text{in} \quad D^2 \tag{2.13}$$

as $n \to \infty$ and

$$(a_n - b_n)n/c_n \to c, \quad -\infty < c < +\infty \quad \text{as} \quad n \to \infty\,, \tag{2.14}$$

then condition (2.6) of Theorem 1 holds with

$$S^*(t) = S^a(t) - S^b(t) + ct \ , \quad t \geq 0 \ . \tag{2.15}$$

3. Essentially A Single Model with Weak Dependence

Suppose that there is only a single model and the FCLT (2.13) in Theorem 4 holds, i.e.,

$$c_n^{-1}\left(S_1^a(\lfloor nt\rfloor) - ant,\ S_1^b(\lfloor nt\rfloor) - bnt\right) \Rightarrow \left(S^a(t), S^b(t)\right) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty\ . \tag{3.1}$$

The remaining condition in Theorem 4 becomes $(a - b)n/c_n \to c$ as $n \to \infty$. Assuming that $\{(A_1(k), B_1(k)) : k \geq 1\}$ is stationary and ergodic,

$$n^{-1}(S_1^a(n), S_1^b(n)) \to (EA_1(1), EB_1(1)) \quad \text{w.p.1 as } n \to \infty\ . \tag{3.2}$$

Assuming in addition that $n/c_n \to \infty$ as $n \to \infty$, in order for the FCLT (3.1) to hold we must have

$$a = EA_1(1) = b = EB_1(1)\ . \tag{3.3}$$

In other words, in this situation the conditions of Theorem 4 can be satisfied only if $\rho \equiv EA/EB = 1$, the critical value for stability with an infinite buffer. Of course, with finite capacity, stability can hold with $\rho \geq 1$. Nevertheless, a limit is possible in this single-model framework only at the single value $\rho = 1$.

In order to develop approximations for more general traffic intensities, it is natural to consider a sequence of models such that the associated traffic intensities satisfy $\rho_n \to 1$. In the setting of Theorem 4, we choose parameters a_n and b_n such that $(a_n - b_n)n/c_n \to c$, $-\infty < c < \infty$. However, it is also natural to want to regard the framework as being *essentially a single model.* We can do so by assuming that the n^{th} model variables $(A_n(k), B_n(k))$ are obtained by simply scaling initial variables, as in (1.3).

The following theorem gives the essentially-single-model version of Theorem 4. Without loss of generality, we let the common translation constant in (3.1) be 1.

Theorem 5. If for a single model

$$c_n^{-1}(S_1^a(\lfloor nt\rfloor) - nt,\ S_1^b(\lfloor nt\rfloor) - nt) \Rightarrow (S_1^a(t), S_2^b(t)) \quad \text{in} \quad D^2 \text{ as } n \to \infty\ , \tag{3.4}$$

and a sequence of models is defined by scaling as in (1.3), where $a_n \to 1$, $b_n \to 1$, and $(a_n - b_n)n/c_n \to c$, $-\infty < c < \infty$, then the conditions of Theorem 4 are satisfied.

With the scaling (1.3), the RBM limit is obtained when $\{(A_1(k), B_1(k)\}$ is an i.i.d. sequence with finite second moments. The result follows directly from the two-dimensional version of Donsker's FCLT.

Theorem 6. If $\{(A_1(k), B_1(k)) : k \geq 1\}$ is an i.i.d. sequence with $EA_1(1) = EB_1(1) = 1$, $Var\, A_1(1) = \sigma_A^2$, $Var\, B_1(1) = \sigma_B^2$ and $Cov[A_1(1), B_1(1)] = \sigma_{AB}^2$, then the FCLT (3.4) holds with $c_n = n^{1/2}$ and $(S_1^a(t), S_1^b(t))$, being two-dimensional centered (drift 0) Brownian motion with covariance matrix

$$\Sigma \equiv (\sigma_{ij}^2) = \begin{pmatrix} \sigma_A^2 & \sigma_{AB}^2 \\ \sigma_{AB}^2 & \sigma_B^2 \end{pmatrix} . \tag{3.5}$$

If also $a_n \to 1$, $b_n \to 1$ and $(a_n - b_n)\sqrt{n} \to c$, $-\infty < c < \infty$, then the conditions of Theorem 4 hold with $c_n = \sqrt{n}$ and the limit process $S^*(t)$ in (2.15) being $\sigma B(t) + ct$, where $B(t)$ is standard (drift 0, variance 1) Brownian motion and

$$\sigma^2 = \sigma_A^2 - 2\sigma_{AB}^2 + \sigma_B^2 . \tag{3.6}$$

As in Kennedy [43] and Berger and Whitt [8]. Theorem 6 generates a relatively simple approximation for the steady-state queue-content distribution, in particular,

$$P(Q(\infty) \le x) \approx \begin{cases} \frac{1-e^{-\theta x}}{1-e^{-\theta C}}, & c \ne 0 \\ \frac{x}{C}, & c = 0 \end{cases} \tag{3.7}$$

and

$$EQ(\infty) \approx \begin{cases} \frac{1}{\theta} - \frac{C}{e^{\theta C}-1}, & c \ne 0 \\ \frac{C}{2}, & c = 0 \end{cases} \tag{3.8}$$

where $\theta = -\sigma^2/2c$. The distribution in (3.7) is the truncated exponential distribution, which approaches the exponential distribution as $C \to \infty$ if $c < 0$. For $C = \infty$, we require $c < 0$ to have a proper steady-state distribution.

A significant feature of the heavy-traffic FCLT is that it implies that the queue-content process and its steady-state distribution can be approximately characterized by a single stochastic process, canonical RBM (with drift -1, diffusion parameter 1 and upper boundary C) and the two parameters c and σ^2 appearing in (3.7) and (3.8).

The heavy-traffic FCLT in Theorem 6 has been stated under the assumption that the pairs $(A(k), B(k))$ are mutually independent for $k \ge 1$. In this i.i.d. setting the transient and steady-state queue-content distributions can actually be computed directly by numerical transform inversion, as in Abate, Choudhury and Whitt [3,4]. Thus, it is significant that essentially the same FCLT holds when the independence condition is relaxed. Indeed there now is a large literature providing the required FCLT (3.4) when independence is relaxed, i.e., when it is replaced by some form of weak dependence. Prominent among these are martingale FCLTs; see Ethier and Kurtz [27] and Jacod and Shiryaev [39]. With such weakly-dependent FCLTs, all that changes is the variance constant σ^2 in (3.6). Examples of alternative variance constants are given in Fendick, Saksena and Whitt [29]. For Markov-modulated arrival processes and more general batch

Markovian arrival processes (BMAP), the contribution to the variance parameter by the input is through the asymptotic variance; e.g., see Burman and Smith [18] and p. 284 of Neuts [52].

A critical condition in Theorem 6 is that the random variables $A_1(1)$ and $B_1(1)$ have finite second moments. We now discuss alternative FCLTs when this condition is violated. The principal case of interest occurs when the inputs occasionally may be very large, so that $A_1(1)$ has an infinite second moment, but $B_1(1)$ still has a finite second moment. Assuming that $A_1(1)$ still has a finite mean, it is natural to assume that the cdf of $A_1(1)$ has a power tail, decaying as $x^{-\alpha}$ for $1 < \alpha < 2$. Fortunately, in this case we can conclude that the FCLT (3.4) once again holds, but the limit process no longer is Brownian motion. Instead it has a component that is a stable Lévy motion, which fortunately is also characterized by only a few parameters.

A *stable Lévy motion* is a real-valued stochastic process $\{Y(t) : t \geq 0\}$ with $Y(0) = 0$, stationary and independent increments and having the stable one-dimensional marginal distribution $S_\alpha(\sigma t^{1/\alpha}, \beta, \mu)$; see p. 113 of Samorodnitsky and Taqqu [60]. The *stable law* $S_\alpha(\sigma, \beta, \mu)$ has four parameters: the *index* α, $0 < \alpha \leq 2$, the *scale parameter* $\sigma > 0$, the *skewness parameter* β, $-1 \leq \beta \leq 1$, and the location or *shift parameter* μ, $-\infty < \mu < \infty$; see Chapter 1 of Samorodnitsky and Taqqu [60]. The logarithm of the characteristic function of an $S_\alpha(\sigma, \beta, \mu)$-distributed random variable Y is

$$\log Ee^{i\theta Y} = \begin{cases} -\sigma^\alpha |\theta|^\alpha (1 - i\beta(\operatorname{sign}\theta)\tan(\pi\alpha/2)) + i\mu\theta, \alpha \neq 1 \\ \\ -\sigma|\theta|(1 + i\beta\frac{2}{\pi}(\operatorname{sign}\theta)\ln(|\theta|)) + i\mu\theta, \qquad \alpha = 1 \ , \end{cases} \tag{3.9}$$

where $\operatorname{sign}(\theta) = 1$, 0 or -1 for $\theta > 0$, $\theta = 0$ or $\theta < 0$. We will be interested in the special case $S_\alpha(\sigma, 1, 0)$ in which the distribution is totally skewed to the right ($\beta = 1$) and centered ($\mu = 0$). Brownian motion is the special case in which $\alpha = 2$. Then the skewness parameter β ceases to play a role and σ^2 is again the scale (now variance) parameter.

The stable law $S_\alpha(\sigma, 1, 0)$ has a cdf with power tail decaying as $x^{-\alpha}$; i.e., if Y is distributed as $S_\alpha(\sigma, 1, 0)$, then

$$\lim_{x\to\infty} x^\alpha P(Y > x) = K_\alpha \sigma^\alpha \ , \tag{3.10}$$

where

$$K_\alpha = \left(\int_0^\infty x^{-\alpha} \sin x \, dx\right)^{-1} = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)}, \alpha \neq 1 \\ \\ \frac{2}{\pi}, \qquad \alpha = 1 \end{cases} \tag{3.11}$$

and $\Gamma(x)$ is the gamma function. For $0 < \alpha < 1$, the stable law $S_\alpha(\sigma, 1, 0)$ is concentrated on the positive half line. The associated stable process has non-negative nondecreasing sample paths and is called a *stable subordinator*. For $1 \leq \alpha < 2$, the stable law $S_\alpha(\sigma, 1, 0)$ has support on the entire real line, but it decays sufficiently fast on the negative real line that the bilateral Laplace transform is well defined. If Y is distributed as $S_\alpha(\sigma, 1, 0)$ then the logarithm of the Laplace transform is

$$\psi_\alpha(s) \equiv \log Ee^{-sY} = \begin{cases} -\sigma^\alpha s^\alpha / \cos(\pi\alpha/2), \alpha \neq 1 \\ \\ 2\sigma s \ln(s)/\pi, \qquad \alpha = 1 \end{cases} \tag{3.12}$$

for $\mathrm{Re}(s) \geq 0$.

Closed-form representations for stable pdf's and cdf's are available in only a few cases, but numerical calculations can easily be done exploiting finite-interval integral representations in Section 2.2 of Zolotarev [76]. These integral representations have been applied to generate tables and graphs of stable pdf's, cdf's and fractiles, as indicated in Section 1.6 of Samorodnitsky and Taqqu [60].

With this background, we can state a stable-process analog of Theorem 6. We omit discussion of the somewhat pathological boundary case $\alpha = 1$. Let e denote the identity map on $[0, \infty)$, i.e., $e(t) = t$, $t \geq 0$.

Theorem 7. Let $\{(A_1(k), B_1(k)) : k \geq 1\}$ be an i.i.d. sequence with $E[B_1(1)^2] < \infty$ and $EA_1(1) = EB_1(1) = 1$. Then the FCLT (3.4) holds with $c_n = n^{1/\alpha}$ for $1 < \alpha < 2$ if and only if there exists a positive constant K for which

$$\lim_{x \to \infty} x^\alpha P(A_1(1) > x) = K \ , \tag{3.13}$$

in which case the limit process $[S^a(t), S^b(t)]$ in (3.4) has $S^b(t) = 0$ and $S^a(t)$ a stable Lévy motion with marginal distribution $S_\alpha(\sigma t^{1/\alpha}, 1, 0)$ for $\sigma = (K/K_\alpha)^{1/\alpha}$, K in (3.13) and K_α in (3.11). If also $(a_n - b_n)n^{1-\alpha^{-1}} \to c$, $-\infty < c < \infty$, then the conditions of Theorem 5, 4 and 1 are satisfied and the limit of the queue-content processes is the reflected stable process $\{R(S^a + ce)(t) : t \geq 0\}$, where $S^a(t)$ is a stable Lévy motion with $S^a(1)$ distributed as $S_\alpha(\sigma, 1, 0)$.

Proof. Since $EB_1(1)^2 < \infty$, the partial sums of $\{B_1(k)\}$ satisfy a FCLT with normalization constant $c_n = n^{1/2}$. Thus, multiplying by $n^{(\alpha-2)/2\alpha}$ yields

$$n^{-1/\alpha} \sum_{i=1}^{\lfloor nt \rfloor} (B_1(k) - 1) \Rightarrow 0 \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ . \tag{3.14}$$

By Theorems 4.1 and 4.4 of Billingsley [10], we can focus on the sequence $\{A_1(k)\}$ alone. The equivalence of (3.13) with the limits for the one-dimensional marginal distributions follows from the classical theory in Section XVII.5 of Feller [28]. Since $A_1(k)$ is nonnegative, the limiting stable law is totally skewed to the right. Given that we have centered by subtracting the means, the stable law is centered as well. The one-dimensional limits extend immediately to convergence of all finite-dimensional distributions because of the i.i.d. assumptions. Moreover, in this setting that implies convergence in the function space D with the standard

Skorohod [61] J_1 topology. For this last step, we can apply Theorems 2.52 and 3.4 on pages 368 and 373 of Jacod and Shiryaev [39]. Condition 2.48 on p. 368 holds because the normalized partial sum process with summands $(A_1(k) - 1)/n^{1/\alpha}$ is both a semimartingale and a process with independent increments (PII) (but not a process with stationary independent increments - PIIS), and the limiting stable process has no fixed discontinuities. Condition 2.53 on p. 368 holds because it is equivalent to

$$\lim_{n\to\infty} P(|A_1(1) - 1| > \epsilon n^{1/\alpha}) = 0 \tag{3.15}$$

for every $\epsilon > 0$. The property $[\beta_3 - D]$ in Theorem 2.52 is

$$\lim_{n\to\infty} nE[h((A_1(1) - 1)/n^{1/\alpha})] = b \ , \tag{3.16}$$

where h is the truncation or centering function, so that in this setting it is $[\mathrm{Sup} - \beta_3]$ in Theorem 3.4 on p. 373, i.e., the condition becomes

$$\lim_{n\to\infty} \sup_{0\le s\le t} \left| \lfloor ns \rfloor E[h((A_1(1) - 1)/n^{1/\alpha})] - bs \right| \ . \tag{3.17}$$

Remark 8. If we allow more general normalization constants c_n instead of $n^{1/\alpha}$, then Theorem 7 still holds with condition (3.13) replaced by the truncated second moment of $A_1(1)$ being regularly varying at infinity with index α, i.e.

$$\lim_{x\to\infty} x^{\alpha-2} L(x)\mu(x) = 1 \ , \tag{3.18}$$

for $1 < \alpha < 2$, where

$$\mu(x) \equiv \int_0^x y^2 dP(A_1(1) \le y) \tag{3.19}$$

and $L(x)$ is a slowly varying function; see Chapters VIII and XVII of Feller [28]. Examples of slowly varying functions are a constant and $\log x$. Given (3.18), we have

$$\lim_{x\to\infty} x^{\alpha} L(x)P(A_1(1) > x) = \frac{2-\alpha}{\alpha} \ . \tag{3.20}$$

Then the normalizing constants c_n must satisfy

$$\lim_{n\to\infty} \frac{n\mu(c_n)}{c_n^2} = K \tag{3.21}$$

and we have

$$\lim_{n\to\infty} nP(A_1(1) > c_n x) = K\frac{(2-\alpha)}{\alpha} x^{-\alpha} \tag{3.22}$$

see Section XVII.5 of Feller [28]. We did not express this more general result in Theorem 7 because we do not believe the extension has great applied value. It seems difficult to distinguish between the regularly varying tail in (3.20) and the power tail in (3.13) in applications. However, the negative result is instructive: Within the single-model framework (with i.i.d. summands), there are no normalization constants yielding a CLT with convergence to a non-normal limit unless the cdf of $A_1(1)$ has a regularly varying tail. Moreover, in that case, the limit must be a stable Lévy motion. In summary, the only possible limit process for the net input process is the stable process in Theorem 7. This limit can also be obtained, at the expense of changing the normalization constants c_n, if the power tail condition (3.13) is replaced by regular variation, but it can be obtained in no other way.

□

Remark 9. In Donsker's FCLT supporting Theorem 6, we can work with either the original scaled partial sums $c_n^{-1}[S_1^a(\lfloor nt \rfloor) - nt]$ or the associated continuous linearly-interpolated, process as in Sections 10 and 16 of Billingsley [10]. Here, however, to work with the linearly interpolated process we would have to shift from Skorohod's [61] J_1 to his M_1 topology, because the maximum jump functional is continuous in the J_1 topology; see Section VI.2 of Jacod and Shiryaev [39]. For our limit theorems, it suffices to use the J_1 topology, but in some settings it may be necessary to use the M_1 or M_2 Skorohod [61] topologies; e.g., see Kella and Whitt [40], Konstantopoulos and Lin [47] and Whitt [70,71].

For applications, it is important that the reflected stable process be relatively tractable. Fortunately, much is known about reflected stable processes and more general reflected Lévy processes, because they play a key role in their fluctuation theory; e.g., see Chapter VI of Bertoin [9], Bingham [11], Kella and Whitt [41] and Sections 24 of Takács [62]. Very nice results are available in our case in which the stable process has no negative jumps. The next result is contained in all the sources above.

Theorem 10. Let $R(S^a + ce)(t)$ be the reflected stable process arising as a limit in Theorem 7, where $c < 0$ and $S^a(1)$ is distributed as $S_\alpha(\sigma, 1, 0)$ for $1 < \alpha < 2$.

(a) If $C = \infty$, then

$$\lim_{t\to\infty} P(R(S^a + ce)(t) \le x) = H(x) , \tag{3.23}$$

where the cdf $H(x)$ has pdf $h(x)$ with Laplace transform

$$\hat{h}(s) \equiv \int_0^\infty e^{-sx} h(x)dx = \frac{1}{1 + (\nu s)^{\alpha-1}} . \tag{3.24}$$

and scaling constant ν, defined by

$$\nu^{\alpha-1} = \frac{-\sigma^\alpha}{c\cos(\pi\alpha/2)} > 0 . \tag{3.25}$$

The associated ccdf $H^c(x) \equiv 1 - H(x)$ has Laplace transform

$$\hat{H}^c(s) \equiv \int_0^\infty e^{-sx} H^c(x) dx = \frac{1 - \hat{h}(s)}{s} = \frac{\nu}{(\nu s)^{2-\alpha}(1 + (\nu s)^{\alpha-1})} . \quad (3.26)$$

(b) If $C < \infty$, then

$$\lim_{t\to\infty} P(R(S^a + ce)(t) \le x) = \frac{H(x)}{H(C)} , \quad 0 \le x \le C , \quad (3.27)$$

for H in (3.23).

Note that ν defined by (3.25) is indeed a scaling factor. To see this, let Y_ν be a random variable with cdf $H_\nu(x)$ and pdf $h_\nu(x)$, showing the dependence upon ν. Then formulas (3.24) and (3.26) are equivalent to

$$Y_\nu \stackrel{\mathrm{d}}{=} \nu Y_1, \quad H_\nu(x) = H_1(x/\nu) \quad \text{and} \quad h_\nu(x) = \frac{1}{\nu} h_1(x/\nu) . \quad (3.28)$$

For the special case $\alpha = 3/2$, the limiting pdf h and cdf H can be expressed in convenient closed form. We can apply 29.3.37 and 29.3.43 of Abramowitz and Stegun [5] to invert the Laplace transforms analytically.

Theorem 11. For $\alpha = 3/2$ and $\nu = 1$, the limiting pdf and ccdf in Theorem 10 are

$$h(x) = \frac{1}{\sqrt{\pi x}} - e^x erfc(\sqrt{x}), \quad x \ge 0 , \quad (3.29)$$

and

$$H^c(x) = e^x erfc(\sqrt{x}), \quad x \ge 0 , \quad (3.30)$$

where

$$erfc(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du = 2\Phi(x\sqrt{2}) - 1 \quad (3.31)$$

and $\Phi(x) \equiv P(N(0,1) \le x)$.

For the special case of $C = \infty$, the heavy-traffic approximation for the steady-state queue content obtained from Theorems 7 and 10 coincides with direct heavy-traffic limits for the steady-state waiting time in the M/GI/1/∞ and GI/GI/1/∞ queues obtained by Boxma and Cohen [14]–[16], Cohen [25] and Abate and Whitt [2]. Explicit results are also given there for the M/G/1 steady-state distributions for the tractable case $\alpha = 3/2$.

For any α, $1 < \alpha < 2$, we can apply Heaviside's theorem, p. 254 of Doetsch [26] to deduce the asymptotic form of the limiting pdf $h(x)$ and ccdf $H^c(x)$. We display two terms; more can be obtained in the same way.

Theorem 12. For $\nu = 1$, the limiting pdf and ccdf in Theorem 10 satisfy

$$h(x) \sim \begin{cases} \frac{-1}{\Gamma(1-\alpha)x^{\alpha}} + \frac{1}{\Gamma(2-2\alpha)x^{2\alpha-1}}, & \alpha \neq 3/2 \\ \frac{1}{2\sqrt{\pi}x^{3/2}} - \frac{3}{4\sqrt{\pi}x^{5/2}}, & \alpha = 3/2 \end{cases} \tag{3.32}$$

$$H^c(x) \sim \begin{cases} \frac{1}{\Gamma(2-\alpha)x^{\alpha-1}} - \frac{1}{\Gamma(3-2\alpha)x^{2(\alpha-1)}}, & \alpha \neq 3/2 \\ \frac{1}{\sqrt{\pi x}} - \frac{1}{2\sqrt{\pi x^3}}, & \alpha = 3/2 \end{cases} \tag{3.33}$$

as $x \to \infty$.

From (3.33), we see that the limiting cdf H fails to have a finite mean for $1 < \alpha < 2$. Theorems 7–12 imply the following iterated limit for the queue-content tail probabilities.

Corollary 13. Under the assumptions of Theorem 7 including (3.13), if $C = \infty$ and $c < 0$, then

$$\lim_{x\to\infty} \lim_{t\to\infty} \lim_{n\to\infty} x^{\alpha-1} P(n^{-1/\alpha} Q_n(\lfloor nt \rfloor) > x) = \frac{K}{c(1-\alpha)} . \tag{3.34}$$

Proof. First, Theorem 7 implies that

$$\lim_{n\to\infty} P(n^{-1/\alpha} Q_n(\lfloor nt \rfloor) > x) = P(R(S^a + ce)(t) > x) \tag{3.35}$$

where $S^a(t) \stackrel{\mathrm{d}}{=} t^{1/\alpha} S^a(1)$ and $S^a(1) \stackrel{\mathrm{d}}{=} S_\alpha(\sigma, 1, 0)$ for $\sigma = (K/K_\alpha)^{1/\alpha}$. Then Theorem 10 implies that

$$\lim_{t\to\infty} P(R(S^a + ce)(t) > x) = H^c(x/\nu) \tag{3.36}$$

for ν in (3.25). By Theorem 12,

$$\lim_{x\to\infty} x^{\alpha-1} H^c(x/\nu) = \frac{\nu^{\alpha-1}}{\Gamma(2-\alpha)} = \frac{K}{c(1-\alpha)} > 0 \tag{3.37}$$

because $\sigma = (K/K_\alpha)^{1/\alpha}$ for K_α in (3.11). □

Remark 14. Theorem 12 expresses the iterated limit in which first $\rho \to 1$ via $n \to \infty$ (in the manner of Theorem 5) and then $t \to \infty$ and $x \to \infty$. We can apply Cohen [23] to consider the iterated limit in a different order. We first let $t \to \infty$ and then $x \to \infty$. For fixed ρ (a_n and b_n), Cohen showed that condition (3.13) is equivalent to

$$\lim_{x\to\infty} x^{\alpha-1} P(Q_n(\infty) > x) = \frac{\rho_n K}{(1-\rho_n)a_n(\alpha-1)} = \frac{K}{(b_n - a_n)(\alpha-1)} . \tag{3.38}$$

We now consider $n^{-1/\alpha}Q_n(\infty)$, which is supposed to approach the limiting distribution of the reflected stable process as $n \to \infty$. By (3.38),

$$\begin{aligned}
&\lim_{n\to\infty}\lim_{x\to\infty} x^{\alpha-1}P(n^{-1/\alpha}Q_n(\infty) > x) \\
&= \lim_{n\to\infty}\lim_{x\to\infty} \frac{(xn^{1/\alpha})^{\alpha-1}}{n^{(\alpha-1)/\alpha}} P(Q_n(\infty) > xn^{1/\alpha}) \\
&= \lim_{n\to\infty} \frac{K}{n^{(\alpha-1)/\alpha}(b_n - a_n)(\alpha-1)} = \frac{K}{c(\alpha-1)} \ .
\end{aligned} \tag{3.39}$$

Hence, (3.39) agrees with (3.34). Thus, if we want to approximate the ccdf $P(Q(\infty) > x)$ when $C = \infty$, a simple approximation under condition (3.13) based on the tail asymptotics ($x \to \infty$) is $Ax^{1-\alpha}$, where the asymptotic constant A can be taken from (3.38), where we do not yet do the limit $\rho_n \to 1$. A more refined approximation exploiting the limit $\rho_n \to 1$, that essentially preserves the asymptotics, is obtained from the heavy-traffic limit Theorem 10, part (a). Except in the case $\alpha = 3/2$, that requires a numerical transform inversion to calculate the ccdf $H^c(x)$. The inversion to calculate $H^c(x)$ from $\hat{H}^c(s)$ in (3.26) is somewhat easier than directly calculating the original ccdf $P(Q > x)$ in the GI/G/1 model, but both can be done. The main advantage of the heavy-traffic limit is that, in that regime, it is not necessary to know the full distributions of the model variables $A_1(1)$ and $B_1(1)$. Only the two parameters ν and α matter in the heavy-traffic limit. The heavy-traffic limit reveals a statistical regularity (simplification) that occurs in the heavy-traffic limit. ∎

By comparing the second term to the first term of the asymptotic expansion of the ccdf $H^c(x)$ in Theorem 12, we can see that the one-term asymptote should tend to be an upper bound for $\alpha < 1.5$ and a lower bound for $\alpha > 1.5$. We also should anticipate that the one-term asymptote should be more accurate for α near $3/2$ than for other values for α. We draw this conclusion for two reasons: first, at $\alpha = 3/2$ a potential second term in the expansion does not appear; so that the relative error (ratio of appearing second term to first term) is of order $x^{-2(\alpha-1)}$ instead of $x^{-(\alpha-1)}$ for $\alpha \neq 3/2$. Second, for $\alpha \neq 3/2$ but α near $3/2$, the constant $\Gamma(3-2\alpha)$ in the denominator of the second term tends to be large, i.e., $\Gamma(x) \to \infty$ as $x \to 0$.

We now show that the anticipated structure deduced from examining Theorem 12 actually holds by making numerical comparisons with exact values computed by numerically inverting the Laplace transform in (3.26). To do the inversion, we use the Fourier series method in Abate and Whitt [1]. We display results for $\alpha = 1.5$, $\alpha = 1.9$ and $\alpha = 1.1$ in Tables 1–3.

For $\alpha = 1.5$, Table 1 shows that the one-term asymptote is a remarkably accurate approximation for x such that $H^c(x) \leq 0.20$. In Table 1 we also demonstrate a strong sensitivity to the value of α by showing the exact values for $\alpha = 1.49$ and $\alpha = 1.40$. For $x = 10^4$ when $H^c(x) = 0.056$ for $\alpha = 1.50$, the corresponding values of $H^c(x)$ for $\alpha = 1.49$ and $\alpha = 1.40$ differ by about 12% and 200%, respectively.

	$H^c(x)$ for $\alpha = 1.5$ $(\nu = 1)$			
x	exact $\alpha = 1.5$	one-term asymptote	exact $\alpha = 1.49$	exact $\alpha = 1.40$
10^{-1}	0.7236	1.78	0.7190	0.6778
10^{0}	0.4276	0.5642	0.4290	0.4421
10^{1}	0.1706	0.1784	0.1760	0.2278
10^{2}	0.5614 $e-1$	0.5642 $e-1$	0.5970 $e-1$	0.1004
10^{3}	0.1783 $e-1$	0.1784 $e-1$	0.1946 $e-1$	0.4146 $e-1$
10^{4}	0.5641 $e-2$	0.5642 $e-2$	0.6304 $e-2$	0.1673 $e-1$
10^{5}	0.1784 $e-2$	0.1784 $e-2$	0.2041 $e-2$	0.6693 $e-2$
10^{6}	0.5642 $e-3$	0.5642 $e-3$	0.6604 $e-3$	0.2670 $e-2$
10^{7}	0.1784 $e-3$	0.1784 $e-3$	0.2137 $e-3$	0.1064 $e-2$
10^{8}	0.5642 $e-4$	0.5642 $e-4$	0.6916 $e-4$	0.4236 $e-3$
10^{16}	0.5642 $e-8$	0.5642 $e-8$	0.8315 $e-8$	0.2673 $e-6$

Table 1
A comparison of the limiting ccdf $H^c(x)$ in Theorem 10 for $\alpha = 1.5$ and $\nu = 1$ with the one-term asymptote and the alternative exact values for $\alpha = 1.49$ and $\alpha = 1.40$.

Tables 2 and 3 show that the one-term asymptote is a much less accurate approximation for α away from 1.5. In the case $\alpha = 1.9$ ($\alpha = 1.1$), the one-term asymptote is a lower (upper) bound for the exact value, as anticipated. For $\alpha = 1.9$, we also compare the ccdf values $H^c(x)$ to the corresponding ccdf values for a mean-1 exponential variable (the case $\alpha = 2$). The ccdf values differ drastically in the tail, but are quite close for small x. A reasonable rough approximation for $H^c(x)$ for all x when α is near (but less than) 2 is the maximum of the one-term asymptote and the exponential ccdf e^{-x}. It is certainly far superior to either approximation alone.

Since $\alpha = 2$ is a critical boundary point for the ccdf tail behavior, one might say that a tail probability catastrophe occurs at $\alpha = 2$. Suppose that the random variable $A_1(1)$ has a power tail decaying as $x^{-\alpha}$. If $\alpha > 2$, then the limiting ccdf $H^c(x)$ is exponential, i.e., $H^c(x) = e^{-x}$, but for $\alpha < 2$ the ccdf decays as $x^{-(\alpha-1)}$.

This drastic change can be seen at the large x values in Table 2.

	$H^c(x)$ for $\alpha = 1.9$ $(\nu = 1)$		
x	exact	one-term asymptote	exponential $\alpha = 2$
$0.1 \times 2^0 = 0.1$	0.878	0.835	0.905
$0.1 \times 2^1 = 0.2$	0.786	0.447	0.819
$0.1 \times 2^2 = 0.4$	0.641	0.240	0.670
$0.1 \times 2^4 = 1.6$	0.238	0.069	0.202
$0.1 \times 2^6 = 6.4$	0.312 $e-1$	0.198 $e-1$	0.166 $e-2$
$0.1 \times 2^8 = 25.6$	0.626 $e-2$	0.568 $e-2$	0.76 $e-11$
$0.1 \times 2^{12} = 409.6$	0.472 $e-3$	0.468 $e-3$	≈ 0
$0.1 \times 2^{16} = 6553.6$	0.386 $e-4$	0.386 $e-4$	≈ 0

Table 2
A comparison of the reflected stable ccdf $H^c(x)$ in Theorem 10 for $\alpha = 1.9$ and $\nu = 1$ with the one-term asymptote and the mean-1 exponential ccdf corresponding to $\alpha = 2$.

Table 3 also illustrates how we can use the asymptotics to numerically determine its accuracy. We can conclude that the one-term asymptote is accurate at those x for which the one-term and two-term asymptotes are very close. Similarly, we can conclude that the two-term asymptote is accurate at those x for which the two-term and three-term asymptotes are close, and so on.

We now consider the case $0 < \alpha < 1$. We note that $\alpha = 1$ is another critical boundary point, because if (3.13) holds for $\alpha > 1$, then the queue-content process has a proper steady-state distribution, but if (3.13) holds for $\alpha < 1$, then the queue-content process $\{Q_1(k) : k \ge 1\}$ will fail to have a proper steady-state distribution, in particular,

$$Q_1(k) \to \infty \quad \text{as} \quad k \to \infty \quad \text{w.p.1.} \tag{3.40}$$

When (3.40) holds, we can use the heavy-traffic limit to show how the queue content should grow over finite time intervals.

Theorem 15. Let $\{(A_1(k), B_1(k)) : k \ge 1\}$ be an i.i.d. sequence with $EB_1(1) < \infty$. Then, for $0 < \alpha < 1$,

$$n^{-1/\alpha}(S_1^a(\lfloor nt \rfloor), S_1^b(\lfloor nt \rfloor)) \Rightarrow (S^a(t), 0) \quad \text{in} \quad D^2 \quad \text{as} \quad n \to \infty \tag{3.41}$$

	$H^c(x)$ for $\alpha = 1.1$ $(\nu = 1)$		
x	exact	one-term asymptote	two-term asymptote
10^{-1}	0.543	1.18	−0.19
10^{0}	0.486	0.94	0.08
10^{1}	0.428	0.74	0.20
10^{2}	0.373	0.59	0.25
10^{4}	0.272	0.373	0.237
10^{6}	0.191	0.235	0.181
10^{8}	0.129	0.148	0.126
10^{12}	0.558 $e-1$	0.590 $e-1$	0.555 $e-1$
10^{16}	0.230 $e-1$	0.235 $e-1$	0.230 $e-1$
10^{24}	0.371 $e-2$	0.373 $e-2$	0.371 $e-2$
10^{32}	0.590 $e-3$	0.590 $e-3$	0.590 $e-3$

Table 3

A comparison of the reflected stable ccdf $H^c(x)$ in Theorem 10 for $\alpha = 1.1$ and $\nu = 1$ with the one-term and two-term asymptotes from Theorem 12.

if and only if (3.13) holds, in which case $S^a(t)$ is a stable subordinator with marginal distribution $S_\alpha(\sigma t^{1/\alpha}, 1, 0)$ for $\sigma = (K/K_\alpha)^{1/\alpha}$, K in (3.13) and K_α in (3.11). Without further assumptions, condition (2.6) of Theorem 1 is satisfied, so that if $Q_n(0)/n^{1/\alpha} \Rightarrow Q(0)$, then

$$n^{-1/\alpha} Q(\lfloor nt \rfloor) \Rightarrow R(S^a)(t) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty\ , \tag{3.42}$$

where

$$R(S^a)(t) = Q(0) + S^a(t) \quad \text{if} \quad C = \infty \tag{3.43}$$

and

$$R(S^a)(t) \Rightarrow C \quad \text{as} \quad t \to \infty \tag{3.44}$$

otherwise.

Proof. Since $EB_1(1) < \infty$, $\{B_1(k) : k \geq 1\}$ obeys the strong law of large numbers, which in turn implies a functional strong law, see Theorem 4 of Glynn and Whitt [32],

$$n^{-1} S_1^b(\lfloor nt \rfloor) \to tEB_1(1) \quad \text{in} \quad D \quad \text{w.p.1.} \quad \text{as} \quad n \to \infty \; . \tag{3.45}$$

Hence, for $0 < \alpha < 1$,

$$n^{-1/\alpha} S_1^b(\lfloor nt \rfloor) \to 0 \quad \text{in} \quad D \quad \text{w.p.1.} \quad \text{as} \quad n \to \infty \; . \tag{3.46}$$

By Theorem 4.4 of Billingsley [10], it thus suffices to consider the process $\{A_1(k) : k \geq 1\}$ alone. From this point, the reasoning is just as in Theorem 7, but without any translation. Since we have no translation terms in (3.41), we can directly apply Theorem 1 to treat the queue-content process. Properties (3.43) and (3.44) hold because $\{S^a(t) : t \geq 0\}$ has nondecreasing paths. □

As a consequence, of Theorem 15, for $C = \infty$ we can approximate the transient queue-content by

$$Q_1(k) \approx n^{1/\alpha} S^a(k/n) \; , \quad k \geq 0 \tag{3.47}$$

for any k. We now consider calculating the $S_\alpha(\sigma, 1, 0)$ pdf and cdf for $0 < \alpha < 1$. Paralleling the case $\alpha = 3/2$ described in Theorem 11, the case $\alpha = 1/2$ is especially tractable. For $\alpha = 1/2$, the $S_\alpha(1, 1, 0)$ distribution has cdf

$$G_{1/2}(x) = 2\Phi^c(1/\sqrt{x}), \quad x \geq 0 \tag{3.48}$$

where $\Phi^c(x) \equiv P(N(0,1) > x)$ and pdf

$$g_{1/2}(x) = \frac{1}{\sqrt{2\pi x^3}} e^{-1/2x}, \quad x \geq 0 \tag{3.49}$$

see p. 52 of Feller [28].

More generally, we can apply numerical inversion of Laplace transforms again to calculate the pdf and ccdf of the stable subordinator $S^\alpha(t)$. We exploit the fact that the distribution $S_\alpha(\sigma, 1, 0)$ of $S^\alpha(1)$ has support on the positive halfline. We exploit self-similarity to relate the distribution at any time t to the distribution at time 1, i.e.,

$$S^a(t) \stackrel{\mathrm{d}}{=} t^{1/\alpha} S^\alpha(1) \; . \tag{3.50}$$

Hence it suffices to consider the single-parameter family of distributions $S_\alpha(1, 1, 0)$. By (3.10), we know that the ccdf of $S_\alpha(1, 1, 0)$ decays as $x^{-\alpha}$. Hence, for $0 < \alpha < 1$, $S^a(t)$ has infinite mean. By (3.50), we expect $S^\alpha(t)$ to grow like $t^{1/\alpha}$ as t increases. However, we should expect much of the growth to be in large jumps. To illustrate the form of the ccdf's, we give the ccdf values of $S_\alpha(1, 1, 0)$ for three values of α in Table 4.

We can exploit the convergence to a stable subordinator to show, asymptotically, how the queue-length process reaches new levels when the input distribution

x	G^c_α ccdf of $S_\alpha(1,1,0)$ $\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$
$(0.01)2^0 = 0.01$	0.9037	1.0000	1.0000
$(0.01)2^1 = 0.02$	0.8672	1.0000	1.0000
$(0.01)2^2 = 0.04$	0.8251	0.9996	1.0000
$(0.01)2^4 = 0.16$	0.7282	0.9229	1.0000
$(0.01)2^6 = 0.64$	0.6233	0.6232	0.7371
$(0.01)2^8 = 2.56$	0.5197	0.3415	0.1402
$(0.01)2^{10} = 10.24$	0.4242	0.1749	0.3739 $e-1$
$(0.01)2^{12} = 40.96$	0.3404	0.8798 $e-1$	0.1154 $e-1$
$(0.01)2^{16} = 655.36$	0.2112	0.2204 $e-1$	0.1220 $e-2$
$(0.01)2^{20} = 10486$	0.1269	0.5510 $e-2$	0.1324 $e-3$
$(0.01)2^{24} = 167,772$	0.7477 $e-1$	0.1377 $e-2$	0.1440 $e-4$
$(0.01)2^{28} = 2,684,000$	0.4359 $e-1$	0.3444 $e-3$	0.1567 $e-5$
$(0.01)2^{32} = 42,949,000$	0.2525 $e-1$	0.8609 $e-4$	0.1705 $e-6$

Table 4
Tail probabilities of the stable law $S_\alpha(1,1,0)$ for $\alpha = 0.2$, 0.5 and 0.8 computed by numerical transform inversion.

has such a heavy tail ($0 < \alpha < 1$). It is interesting to describe the values immediately before and after first passing some high level. The results show that very large jumps should be expected. These results are related to the generalized arc sine laws; see Sections III and VIII of Bertoin [9]. For $z > 0$, let τ_z be the *first passage time* to a level beyond z; i.e., for $x \in D$,

$$\tau_z(x) = \inf\{t \geq 0 : x(t) > z\} \tag{3.51}$$

with $\tau_z(x) = \infty$ if $x(t) \leq z$ for all t. Let γ_z be the associated *overshoot*; i.e.,

$$\gamma_z(x) = x(\tau_z(x)) - z \ . \tag{3.52}$$

Let $x(\tau_z(x)-)$ be the *value just before the first passage.*

Theorem 16. Under the conditions of Theorem 15, including (3.13) with $0 < \alpha < 1$, for $z > 0$,

$$n^{-1}\tau_{zn^{1/\alpha}}(Q) \Rightarrow \tau_z(S^a) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \ , \tag{3.53}$$

so that

$$\lim_{n\to\infty} P(\tau_{zn^{1/\alpha}}(Q) > nx) = P(S^a(x) \le z) \ ; \tag{3.54}$$

$$n^{-1/\alpha}\gamma_{zn^{1/\alpha}}(Q) \Rightarrow \gamma_z(S^a) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty \ , \tag{3.55}$$

so that, for $b > z$,

$$\lim_{n\to\infty} P(\gamma_{zn^{1/\alpha}}(Q) > (b-z)n^{1/\alpha}) = \frac{1}{\Gamma(\alpha)\Gamma(1-\alpha)} \int_0^z x^{\alpha-1}(b-x)^{-\alpha}dx \ ; \tag{3.56}$$

$$n^{-1/\alpha}Q(\tau_{zn^{1/\alpha}}(Q)-) \Rightarrow S^a(\tau_z(S^a)-) \tag{3.57}$$

and, for $0 < b < 1$,

$$\lim_{z\to\infty} P(S^a(\tau_z(S^a)-) > zb) = \int_b^1 \frac{\sin(\alpha\pi)dt}{\pi t^{1-\alpha}(1-t)^\alpha} \ . \tag{3.58}$$

Proof. Since the first-passage time, overshoot and last-value functions are continuous functions on D, we can apply the continuous mapping theorem; see Section VI of Jacod and Shiryaev [39]. Note that

$$\tau_z(n^{-1/\alpha}Q(n\cdot)) = n^{-1}\tau_{zn^{1/\alpha}}(Q) \ , \tag{3.59}$$

$$\gamma_z(n^{-1/\alpha}Q(n\cdot)) = n^{-1/\alpha}Q(\tau_{zn^{1/\alpha}}(Q)) - z = n^{-1/\alpha}\gamma_{zn^{1/\alpha}}(Q) \tag{3.60}$$

and

$$n^{-1/\alpha}Q(\tau_{zn^{1/\alpha}}(Q)-) = n^{-1/\alpha}Q(n\tau_z(n^{-1/\alpha}Q(n\cdot)-)) \ . \tag{3.61}$$

For (3.56), see Exercise 3, p. 238, and p. 241 of Bertoin [9]. For (3.58), see Theorem 6, p. 81, of Bertoin [9]. □

Note that the scale parameter σ enters in simply to the first passage time, i.e., for $y > 0$

$$\tau_z(S^a(y\cdot)) = y^{-1}\tau_z(S^a) \ ,$$

and does not appear at all in the overshoot or the last value before passage (because σ corresponds to a simple time scaling).

4. A Sequence of Models with Weak Dependence

In Section 3 we saw that there only few possible limits for the queue-content process in the framework of essentially a single model with weak dependence. In addition to the familiar RBM limit in Theorem 6, there are the reflected stable process limits in Theorems 7–15. Each of these limit processes is characterized by only a few parameters. There are many more possibilities when we consider a sequence of models.

When we consider a sequence of models we can of course obtain the same limits described in Section 3, and clearly those are important special cases, but we can also obtain different limits. *The most important generalization is that the queue need not be in heavy traffic.* As noted in the beginning of Section 3, for a single model the translation terms in the limit for the arrival and service processes imply that we must have $\rho = 1$ in order to have a nondegenerate limit for the queue-content process. When we generalize to essentially a single model, using the scaling in (1.3) and Theorem 5, we can have $\rho_n \to 1$ as $n \to \infty$, which we still regard as heavy traffic. However, when we consider a general sequence of models, the translation and normalization can be absorbed in the basic variables, so that we do *not* need the assumption $\rho_n \to 1$.

With a sequence of models, we can apply Theorem 1 directly, which has the condition of convergence of the net-input processes $c_n^{-1} S_n(\lfloor nt \rfloor)$ without translation. In order to obtain a continuous-time limit process, we want to scale time, but there is great freedom in the way the model sequence $\{(A_n(k), B_n(k) : k \geq 1\}$ can change with n. Given a limit for the sequence of net-input processes, we immediately get a limit for the associated queue-content processes.

We will focus on special case in which the sequence $\{X_n(k) : k \geq 1\}$ is i.i.d. for each n. As before, similar limits will hold when there is only weak dependence. The i.i.d. setting involves the classical limits for triangular arrays of partial sums, as in Gnedenko and Kolmogorov [33] and Chapters IX and XVII of Feller [28]. In the i.i.d. setting, the possible one-dimensional limit in $\mathbb{R}$ for the net-input process are the infinitely divisible laws. The corresponding possible process limits are Lévy processes; see Jacod and Shiryaev [39].

By a *Lévy process*, we mean a random element $\{L(t) : t \geq 0\}$ of D with $L(0) = 0$ and stationary and independent increments. The random variable $L(1)$ has an infinitely divisible (ID) distribution. Given the *truncation* (or centering) *function* $h(x) = sgn(x)(\min\{|x|, 1\})$ and the *Lévy measure* μ on $\mathbb{R}$ satisfying $\mu((-\infty, -1) \cup (1, \infty)) < \infty$, $\mu(\{0\}) = 0$ and $\int_{-1}^{1} x^2 \mu(dx) < \infty$, the characteristic function of $L(1)$ is given by the Lévy-Khinchine formula

$$Ee^{ict} = e^{t\psi(\xi)} \tag{4.1}$$

where the exponent is

$$\psi(\xi) \equiv \log Ee^{i\xi L(1)} = ib\xi - \frac{\sigma^2\xi^2}{2} + \int_{-\infty}^{\infty} (\exp(i\xi x) - 1 - i\xi h(x))\mu(dx) \ . \quad (4.2)$$

We call the triple (b, σ^2, μ) the *characteristics* of the Lévy process. If $L(1)$ has finite mean, then it is

$$EL(1) = \frac{\psi'(0)}{i} = b + \int_{-\infty}^{\infty} [x - h(x)]\mu(dx) \ , \quad (4.3)$$

where, because of the definition of h, the integrand in the second term is non-zero only in $(-\infty, -1) \cup (1, \infty)$, so that the integral is finite.

With a sequence of models, we can drop the common distribution assumptions if the summands are *infinitesimal*, i.e., if

$$\lim_{n\to\infty} \sup_{k\geq 1} P(|X_n(k)/c_n| > \epsilon) = 0 \quad (4.4)$$

e.g., see 2.33 on p. 362 of Jacod and Shiryaev [39].

Theorem 17. Consider a sequence of queueing models and suppose that the net-input sequence $\{X_n(k) : k \geq 1\}$ is i.i.d. for each n. Assume that (4.4) holds. Then

$$c_n^{-1} S_n(\lfloor nt \rfloor) \Rightarrow S^*(t) \quad \text{in} \quad D \quad \text{as} \quad n \to \infty \ , \quad (4.5)$$

i.e., condition (2.6) of Theorem 1 holds, if and only if there are nonnegative constants b and σ^2 and a Lévy measure μ such that

$$(i) \ \lim_{n\to\infty} nE[h(X_n(1)/c_n)] = b \ , \quad (4.6)$$

$$(ii) \ \lim_{n\to\infty} nVar[h(X_n(1)/c_n))] = \sigma^2 \ , \quad (4.7)$$

$$(iii) \ \lim_{n\to\infty} nE[g(X_n(1)/c_n)] = \int_{-\infty}^{\infty} g(x)\mu(dx) \quad (4.8)$$

for all continuous bounded real-valued functions g on $\mathbb{R}$ with $g(x) = 0$ in a neighborhood of $x = 0$ and $g(x) \to y$, $-\infty < y < \infty$, as $x \to \pm\infty$, in which case the limit process S^* is a Lévy process with characteristics (b, σ^2, μ).

Proof. It is well known that the possible one-dimensional limits in $\mathbb{R}$ are the ID laws; see p. 303 of Feller [28]. For the necessity and sufficiency of (i)–(iii), see 2.35 on p. 362 of Jacod and Shiryaev [39]. The extension to convergence in D follows by the same reasoning as in Theorem 7. In this case, condition 2.53 on p. 368 of Jacod and Shiryaev [39] is equivalent to (4.4). □

Remark 18. Conditions (i)–(iii) in Theorem 17 are similar to the necessary and sufficient conditions for ID cdf's F_n with characteristics (b_n, σ_n^2, μ_n) to converge to an ID cdf F with characteristics (b, σ^2, μ). The necessary and sufficient conditions are:

(i) $b_n \to b$ as $n \to \infty$
(ii) $\sigma_n^2 + \int_{-\infty}^{\infty} h(x)^2 \mu_n(dx) \to \sigma^2 + \int_{-\infty}^{\infty} h(x)^2 \mu(dx) \quad \text{as} \quad n \to \infty$
(iii) $\int_{-\infty}^{\infty} g(x)\mu_n(dx) \to \int_{-\infty}^{\infty} g(x)\mu(dx) \quad \text{as} \quad n \to \infty$

for the same class of functions g as in (4.8); see p. 355 of Jacod and Shiryaev [39]. ∎

As indicated above, similar theorems hold when the basic sequences $\{X_n(k) : k \geq 1\}$ are only weakly dependent; see Chapter VIII of Jacod and Shiryaev [39]; a martingale functional limit theorem is Theorem 2.29 on p. 426.

Similar theorems also hold for single-class networks of queues, using the multidimensional reflection map. The limit theorems in Jacod and Shiryaev [39] are stated for multidimensional processes. Properties of multidimensional (non-Brownian) reflected Lévy processes are discussed in Kella and Whitt [42], Konstantopoulos, Last and Lin [48] and references therein.

In order to apply Theorem 17, we want the reflected Lévy process to be relatively tractable. Fortunately, there are practically important cases in which a reflected Lévy process is tractable. Closely paralleling Section 3, a reflected Lévy process is tractable if the associated Lévy process has no negative jumps. For some recent results in the general case, see Konstantopoulos, Last and Lin [48]. The original result for the case with no negative jumps is due to Zolotarev [75]; also see Section 24 of Takács [62], Bingham [11] and Kella and Whitt [41], especially Section 4(a). When a Lévy process L has no negative jumps, the Lévy measure μ concentrates on $(0, \infty)$ and the bilateral Laplace-Stieltjes transform of $L(1)$ is well defined, with *Laplace exponent*

$$\begin{aligned} \phi(s) &\equiv \log E e^{-sL(1)} \\ &= -bs + \frac{\sigma^2 s^2}{2} + \int_0^{\infty} (\exp(-sx) - 1 + sh(x))\mu(dx) \ . \end{aligned} \tag{4.9}$$

An important special case is a subordinator plus a negative drift, which is just (4.9) without the second Brownian term. Storage models with such Lévy net-input processes are analyzed directly in Chapter 3 of Prabhu [54]. With (4.9), we can conveniently characterize the Laplace transform of the steady-state distribution. The following is a generalization of Theorem 10, part of Lévy process theory.

Theorem 19. Let $R(S^*)(t)$ be the reflected Lévy process arising as the limit in Theorems 1 and 17. Assume that $R(S^*)$ has no negative jumps, so that S^* has Laplace exponent ϕ in (4.9).

(a) If $C = \infty$, then

$$\lim_{t\to\infty} P(R(S^*)(t) \le x) = H(x) \ , \tag{4.10}$$

where H is a proper cdf with Laplace-Stieltjes transform

$$\hat{h}(s) \equiv \int_0^\infty e^{-sx} dH(x) = \frac{s\phi'(0)}{\phi(s)} \ , \tag{4.11}$$

with ϕ being the Laplace exponent in (4.9).

(b) If $C < \infty$, then

$$\lim_{t\to\infty} P(R(S^*)(t) \le x) = \frac{H(x)}{H(C)}, \quad 0 \le x \le C \tag{4.12}$$

for H in (4.10).

Remark 20. Unlike Section 3, the Lévy process approximation is not much more elementary than the original model. Focusing on one value of n, we start with essentially a GI/G/1 model depending on the probability law of $X_n(1)$, for which the transform of the steady-state distribution has been determined. The Lévy process S^* and the steady-state cdf H of the reflected Lévy process depend on the Laplace exponent $\phi(s)$ and thus on the triple (b, σ^2, μ). We can identify the two parameters b and σ^2 by assuming that (4.6) and (4.7) hold as equalities for a fixed n, but we must have (4.8) hold for a whole set of functions g, so that the Lévy measure μ is harder to identify from (4.8). As indicated in Theorem 2.35 on p. 362 of Jacod and Shiryaev [39], it is possible to restrict attention to a more convenient countable collection of test functions, their $C_1(\mathbb{R})$, so that approximate fitting by this route can be contemplated. We can also identify the Lévy measure μ by the limits

$$\lim_{n\to\infty} nP(X_n(1) > c_n x) = \mu((x, \infty)) \tag{4.13}$$

and

$$\lim_{n\to\infty} nP(X_n(1) < -c_n x) = \mu((-\infty, -x)) \ , \tag{4.14}$$

which should hold for all x such that $\mu(\{x\})$ and $\mu(\{-x\}) = 0$. We can approximate μ by assuming that (4.13) and (4.14) hold as equalities for x outside of the interval $(-\epsilon, \epsilon)$ for suitably large fixed n and for suitably small ϵ. Even if $\mu((-\epsilon, \epsilon)) = \infty$, the contribution to the process associated with μ on $(-\epsilon, \epsilon)$ is asymptotically negligible, as $\epsilon \to 0$; e.g., see p. 14 of Bertoin [9]. Ignoring μ on $(-\epsilon, \epsilon)$ makes the overall Lévy process the sum of a (b_1', σ^2)-Brownian motion with drift

$$b' = b + \int_{-\infty}^{\infty} [x - h(x)]\mu(dx) \ , \tag{4.15}$$

as in (4.3), and a compound Poisson process with Poisson arrival rate $\lambda = \mu((-\infty,-\epsilon)) + \mu((\epsilon,\infty)) < \infty$ and jump-size distribution

$$P(\text{Jump} < -x) = \frac{\mu((-\infty,-x))}{\lambda} \quad \text{and} \quad P(\text{Jump} > x) = \frac{\mu((x,\infty))}{\lambda} \; . \qquad (4.16)$$

The Lévy process provides more simplification when the original queueing model has some dependence, as in the case of the semi-martingale triangular array in Section VIII.2e of Jacod and Shiryaev [39]. Then the approximation steps again involve the identification of the triple (b, σ^2, μ) but from the appropriate conditional distributions. ∎

The generalization of Theorem 15 occurs when the Lévy process S^* in (4.4) is a subordinator, i.e., has nondecreasing sample paths. If $C = \infty$, then $R(S^*) = S^*$, so that $c_n^{-1}Q_n(\lfloor nt \rfloor) \Rightarrow S^*(t)$ in D. We can then use the limit to generate approximations for the transient behavior. Similarly, we have generalizations of (3.53)–(3.55) in Theorem 16.

Example 21. The workload in unfinished service time in the M/G/1 queue is a reflected Lévy process. If V is a service time and λ is the arrival rate, then the Laplace exponent of the compound-Poisson net-input process is

$$\phi(s) = s - \lambda(1 - E[\exp(-sV)]) \; . \qquad (4.17)$$

Example 22. A possible subordinator is the gamma process, which can be expressed via the Laplace exponent

$$\phi(s) = \int_0^\infty (e^{-sx} - 1)\frac{e^{-x/\eta}}{x}dx = -\log(1 + \eta s)$$

for constant η; e.g., see p. 72 of Prabhu [54]. (The centering function is not needed in this case.) If we add a constant negative drift to the gamma process then we obtain a Lévy process with negative drift but without negative jumps, having Laplace exponent $\phi(s) = bs - \log(1 + \eta s)$. if $b > \eta$, then $ES^*(1) < 0$ and we can apply Theorem 19. In this case, the steady-state ccdf H^c is easy to compute from its Laplace transform $H^c(s) = [1 - h(s)]/s$ by numerical inversion. The gamma process is a Lévy process without Brownian component; i.e., $b = \sigma^2 = 0$. The Lévy measure has density $\mu(dx) = x^{-1}e^{-x/\eta}$, $x > 0$. As in Remark 20, we can approximate the gamma process by a compound Poisson process by restricting μ to $[\epsilon, \infty)$ for some $\epsilon > 0$.

5. Gaussian Approximations to Capture Dependence

The functional limit theorems in Sections 3 and 4 depend critically on having only weak dependence. We stated results for the i.i.d. case, and mentioned that there are extensions to various forms of weak dependence, e.g., as covered by martingale FCLTs. However, in some applications there is strong dependence.

We can still exploit the basic FCLT, Theorem 1, with strong dependence, but to be of use we must find some way for the limit processes S^* and $R(S^*)$ in (2.6) and (2.7) to simplify. One way this can occur is for the input process to be the superposition of many independent, or nearly independent, component input processes. Then, assuming finite variances, we can use the central limit theorem to justify approximating the input sequence $\{A_n(k) : k \geq 1\}$ by a Gaussian process. Then there is simplification, because the marginal distributions are normal (Gaussian) with the dependence being represented via the autocovariance function. If we can also justify approximating the potential service sequence $\{B_n(k) : k \geq 1\}$ by a Gaussian process, then the net-input sequence $\{X(k) : k \geq 1\}$ can also be approximated by a Gaussian process, which is characterized by its mean and autocovariance function. One way $B_n(k)$ can be Gaussian is for the service capacity to be constant, as in a communication switch with steady output whenever there is work to be done.

It is possible to carry out this Gaussian process approximation in the original discrete-time framework of (1.1) and (1.2). Then the relevant function (sequence) space is $\mathbb{R}^\infty$. Convergence in $\mathbb{R}^\infty$ is characterized by convergence of all finite dimensional distributions; see p. 19 of Billingsley [10]. Then the reflection map is as defined in (1.1).

Theorem 23. Consider a sequence of models indexed by n of the form (1.1) and (1.2) where $\{A_n(k) : k \geq 1\}$ is the superposition of n independent component processes and $B_n(k) = b_n$ such that

$$X_n \Rightarrow Y \quad \text{in} \quad \mathbb{R}^\infty \quad \text{as} \quad n \to \infty \tag{5.1}$$

where $\{Y(k) : k \geq 1\}$ is a Gaussian process. Then

$$S_n \Rightarrow S^* \quad \text{in} \quad \mathbb{R}^\infty \quad \text{as} \quad n \to \infty \tag{5.2}$$

where $S^*(k) = Y(1) + \cdots + Y(k)$, $k \geq 1$; where $R(S^*)(0) = 0$.

We can regard the stationary Gaussian process Y in the limit (5.1) as being defined on $\{k : -\infty < k < \infty\}$. If there is unlimited waiting room, i.e., if $C = \infty$, then we can characterize the steady-state distribution of the reflected Gaussian process by

$$R(S^*)(\infty) \stackrel{\mathrm{d}}{=} \sup\{0, Y_{-1} + \cdots + Y_{-k} : k \geq 1\} \ . \tag{5.3}$$

As in Norros [53], Addie and Zukerman [6] and Choe and Shroff [21,22] we can then approximate the probability of a supremum exceeding a level by the supremum of the probabilities of exceeding that level, i.e.,

$$P(R(S^*)(\infty) > x) \approx \sup_{k \geq 1} P(S_k > x) = \sup_k P\left(\frac{S_k - km}{x - km} > 1\right) , \tag{5.4}$$

where $m = EY(1) < 0$. However, the supremum of the mean-0 Gaussian tail probabilities occurs at the k maximizing the variance of $(S_k^* - km)/(x - km)$, which can be computed given the autocovariance function of Y. That maximum Gaussian tail probability in (5.4) is the candidate approximation (which has been shown to be remarkably accurate).

Given the approximating Gaussian process $\{S^*(k) : k \geq 1\}$ and the associated reflected process, we might elect to scale space and time and then take further limits to obtain further simplification. Willinger et al. [74] and Taqqu et al. [63] do this. They actually start by considering on-off component arrival processes with power-tail (or more general regularly varying) on-time and/or off-time distributions, where the index α satisfies $1 < \alpha < 2$. They show that this produces an associated power-tail autocovariance function for the Gaussian process in the analog of Theorem 23. When they scale time and space appropriately, they obtain convergence of the Gaussian net-input to fractional Brownian motion. This second limit describes the large-time-scale behavior of the autocovariance function.

Just as in Section 4, the limit in Theorem 23 does not require that the queue be in heavy traffic. However, we might also want to consider queues with superposition arrival processes in heavy traffic. A key point then is how the traffic intensity ρ_n changes with n, where model n has a superposition of n component inputs (n i.i.d. component processes in the basic case). In fact, a FCLT for the $\sum_{i=1}^n GI/G/1$ model in the heavy-traffic case was already established by Whitt [69]; then the number n of component arrival streams is allowed to increase so that $n(1 - \rho_n)^2 \to c$ as $n \to \infty$ for $0 < c < \infty$. Then $B_n(k)$ was allowed to be non-deterministic and contributes a Brownian motion component to the limiting Gaussian net input process. Approximation (5.4) can also be applied there when $C = \infty$. (Variants of this heavy-traffic limiting regime were also subsequently considered by Knessl and Morrison [46], Kushner and Martins [50] and Kushner, Yang and Jarvis [51], the last with control considerations. By restricting attention to exponential on and off times and by keeping track of appropriate system state variables, they obtain a Markov limit process.)

These previous limits were for continuous-time processes. To extend convergence of finite-dimensional distributions to convergence in distribution in function space, we need to establish tightness in the superposition limit. As in Whitt [69], this can be done by applying central limit theorems for processes in D, as was done by Hahn [34]. For the renewal component arrival processes considered in Whitt [69], this places a constraint on the behavior of the interrenewal-time cdf at the origin, but not in the tail, so that the limit holds for heavy-tail interrenewal cdf's. That same argument applies to non-renewal component arrival processes too, such as considered by Willinger et al. [74] and Taqqu et al. [63].

6. Conclusions

We have reviewed FCLTs for the single-server queue, emphasizing non-Brownian FCLTs that hold with heavy-tailed probability distributions and strong dependence. We reviewed the continuous-mapping approach to obtain general limits in Section 2. In Section 3 we focused on the case of essentially a single model, as defined in (1.3), under weak dependence. In addition to the standard RBM limit in Theorem 6, in Theorem 7 we described the convergence to a reflected stable process when the single-period inputs have a power tail or, more generally, are regularly varying. We devoted much of Section 3 showing that the limit can yield useful practical insights; e.g., we showed that the steady-state distribution of the limit process can readily be computed using numerical transform inversion. We also used the numerical results to explore the structure of the approximating distributions.

Section 4 was devoted to the case of a general sequence of models. Theorem 17 describes the convergence to a reflected Lévy process. Theorem 19 reviews the classic result by Zolotarev [75] concluding that the steady-state distribution has a relatively simple Laplace transform (the generalized Pollaczek-Khintchine formula) when the Lévy process has no negative jumps. Takács [62] also focused on this important case, but it has not yet received the attention it deserves.

Finally, in Section 5 we discussed Gaussian approximations for the case in which the input is the superposition of inputs from a large number of independent sources, where the input from each source may have strong (e.g., long-range) dependence. For practical applications, it is significant that there is a relatively simple approximation associated with a bound. A main theme overall is the remarkable tractability of the results obtained from the limits.

References

[1] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Computing* 7 (1995) 36–43.

[2] J. Abate and W. Whitt, Explicit M/G/1 waiting-time distributions for a class of long-tail service-time distributions. *Opns. Res. Letters* 25 (1999) 25–31.

[3] J. Abate, G. L. Choudhury and W. Whitt, Calculation of the GI/G/1 waiting time distribution and its cumulants from Pollaczek's formulas. *Archiv für Elektronik und Ubertragungstechnik* 47 (1993) 311–321.

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

[5] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, D.C., 1972.

[6] R. G. Addie and M. Zukerman, An approximation for performance evaluation of stationary single server queues. *IEEE Trans. Commun.* 42 (1994) 3150–3160.
[7] P. Barford and M. Crovella, Generating representative web workloads for network and server performance evaluation. *Proc. 1998 ACM Sigmetrics* (1998) 151–160.
[8] A. W. Berger and W. Whitt, The Brownian approximation for rate-control throttles and the G/G/1/C queue. *J. Discrete Event Dyn. Systems* 2 (1992) 7–60.
[9] J. Bertoin, *Lévy Processes*, Cambridge University Press, Cambridge, U.K., 1996.
[10] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[11] N. H. Bingham, Fluctuation theory in continuous time. *Adv. Appl. Prob.* 7 (1975) 705–766.
[12] L. Bondesson, *Generalized Gamma Convolutions and Related Classes of Distributions and Densities*, Springer, New York, 1992.
[13] A. A. Borovkov, Some limit theorems in the theory of mass service, II. *Theor. Prob. Appl.* 10 (1965) 375–400.
[14] O. J. Boxma and J. W. Cohen, The M/G/1 queue with heavy-tailed service time distribution. *IEEE J. Sel. Areas Commun.* 16 (1998) 749–763.
[15] O. J. Boxma and J. W. Cohen, Heavy-traffic analysis for the GI/G/1 queue with heavy-tailed distributions. *Queueing Systems*, to appear.
[16] O. J. Boxma and J. W. Cohen, The M/G/1 queue: heavy tails and heavy traffic. In *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger (eds.), Wiley, New York, 1999, to appear.
[17] M. Bramson, State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* 30 (1998) 89–148.
[18] D. Y. Burman and D. R. Smith, An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Oper. Res.* 34 (1986) 105–119.
[19] H. Chen and A. Mandelbaum, Stochastic discrete flow networks: diffusion approximations and bottlenecks. *Ann. Probab.* 19 (1991) 1463–1519.
[20] H. Chen and A. Mandelbaum, Leontief systems, RBV's and RBM's. *Proc. Imperial College Workshop on Applied Stochastic Processes*, M. H. A. Davis and R. J. Elliott (eds.), Gordon and Breach, New York, 1991.
[21] J. Choe and N. B. Shroff, A central limit theorem based approach for analyzing queue behavior in high-speed networks. *IEEE/ACM Trans. Networking* 6 (1998) 659–671.
[22] J. Choe and N. B. Shroff, On the supremum distribution of integrated stationary Gaussian processes with negative linear drift. *Adv. Appl. Prob.* 31 (1999) 135–157.
[23] J. W. Cohen, Some results on regular variation for distributions in queueing and fluctuation theory. *J. Appl. Prob.* 10 (1973) 343–353.
[24] J. W. Cohen, *The Single Server Queue*, revised edition, North-Holland, Amsterdam, 1982.
[25] J. W. Cohen, A heavy-traffic theorem for the GI/G/1 queue with a Pareto-type service time distribution. *J. Applied Math. Stoch. Analysis*, special issue dedicated to R. Syski, 11 (1998) 247–254.
[26] G. Doetsch, *Introduction to the Theory and Application of the Laplace Transformation*, Springer-Verlag, New York, 1974.
[27] S. N. Ethier and T. G. Kurtz, *Markov Processes, Characterization and Convergence*, Wiley, New York, 1986.
[28] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. II, second edition, Wiley, New York.
[29] K. W. Fendick, V. R. Saksena and W. Whitt, Dependence in packet queues. *IEEE Trans. Commun.* 37 (1989) 1173–1183.
[30] H. Furrer, Z. Michna and A. Weron, Stable Lévy motion approximation in collective risk theory. *Insurance: Math and Econ.*, 20 (1997) 97–114.
[31] D. P. Gaver and P. A. Jacobs, Waiting times when service times are stable laws: tamed and

wild, *Advances in Applied Probability and Stochastic Processes*, liber amicorium J. Keilson, 1999, to appear.

[32] P. W. Glynn and W. Whitt, Ordinary CLT and WLLN versions of $L = \lambda W$. *Math. Oper. Res.* 13 (1988) 674–692.

[33] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, revised edition, Addison-Wesley, Reading, MA, 1968.

[34] M. G. Hahn, Central limit theorems in $D[0, 1]$. *Z. Wahrsch. verw. Geb.* 44 (1978) 89–101.

[35] J. M. Harrison, *Brownian Motion and Stochastic Flow Systems*, Wiley, New York, 1985.

[36] J. M. Harrison and M. I. Reiman, Reflected Brownian motion on an orthant. *Ann. Probab.* 9 (1981) 302–308.

[37] D. L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, I. *Adv. Appl. Prob.* 2 (1970) 150–177.

[38] D. L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, II: sequences, networks and batches. *Adv. Appl. Prob.* 2 (1970) 355–369.

[39] J. Jacod and A. N. Shiryaev, *Limit Theorems for Stochastic Processes*, Springer-Verlag, New York, 1987.

[40] O. Kella and W. Whitt, Diffusion approximations for queues with server vacations. *Adv. Appl. Prob.* 22 (1990) 706–729.

[41] O. Kella and W. Whitt, Useful martingales for stochastic storage processes with Lévy input, *J. Appl. Prob.* 29 (1992) 396–403.

[42] O. Kella and W. Whitt, Stability and structural properties of stochastic storage networks. *J. Appl. Prob.* 33 (1996) 1169–1180.

[43] D. P. Kennedy, Limit theorems for finite dams. *Stoch. Proc. Appl.* 1 (1973) 269–278.

[44] J. F. C. Kingman, The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* 57 (1961) 902–904.

[45] J. F. C. Kingman, On queues in heavy traffic. *J. Roy. Statist. Soc.* Ser B, 24 (1962) 383–392.

[46] C. Knessl and J. A. Morrison, Heavy traffic analysis of data handling system with multiple sources. *SIAM J. Appl. Math.* 51 (1991) 187–213.

[47] T. Konstantopoulos and S.-J. Lin, Macroscopic models for long-range dependent network traffic. *Queueing Systems* 28 (1998) 215–243.

[48] T. Konstantopoulos, G. Last and S.-J. Lin, On stationary reflected Lévy processes. University of Texas at Austin, 1999.

[49] T. G. Kurtz, Limit theorems for workload input models. In *Stochastic Networks: Theory and Applications*, F. P. Kelly, S. Zachary and I. Ziedins, eds., Oxford University Press, Oxford, U.K., 1996, 119–139.

[50] H. J. Kushner and L. F. Martins, Numerical methods for controlled and uncontrolled multiplexing and queueing systems. *Queueing Systems* 16 (1994) 241–285.

[51] H. J. Kushner, J. Yang and D. Jarvis, Controlled and optimally controlled multiplexing systems: a numerical exploration. *Queueing Systems* 20 (1995) 255–291.

[52] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, New York, 1989.

[53] I. Norros, A storage model with self-similar input. *Queueing Systems* 16 (1994) 387–396.

[54] N. U. Prabhu, *Stochastic Storage Processes*, Springer, New York, 1980.

[55] Yu. V. Prohorov, Transient phenomena in queueing processes. *Litovsk. Mat. Sb.* 3 (1963) 199–206 (in Russian).

[56] M. I. Reiman, Open queueing networks in heavy traffic. *Math. Oper. Res.* 9 (1984) 441–458.

[57] S. Resnick and H. Rootzén, Self-similar communication models and very heavy tails, Cornell University, 1998.

[58] S. Resnick and G. Samorodnitsky, A heavy traffic limit theorem for workload processes with heavy tailed service requirements, Cornell University, 1998.

[59] S. Resnick and E. van den Berg, Weak convergence of high-speed network traffic models, Cornell University, 1999.

[60] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes*, Chapman and Hall, New York, 1994.

[61] A. V. Skorohod, Limit theorems for stochastic processes. *Theor. Prob. Appl.* 1 (1956), 261–290.

[62] L. Takács, *Combinatorial Methods in the Theory of Stochastic Processes*, Wiley, New York, 1967.

[63] M. S. Taqqu, W. Willinger and R. Sherman, Proof of a fundamental result in self-similar traffic modeling. *Computer Communication Review* 27 (1997) 5–22.

[64] K. P. Tsoukatos and A. M. Makowski, Heavy-traffic analysis of a multiplexer driven by $M/GI/\infty$ input processes. *Teletraffic Contributions for the Information Age, Proceedings of ITC 15*, V. Ramaswami and P. E. Wirth (eds.), Elsevier, Amsterdam, 1997, 497–506.

[65] K. P. Tsoukatos and A. M. Makowski, Heavy traffic limits associated with $M/GI/\infty$ input processes. *Queueing Systems*, to appear.

[66] W. Whitt, Weak convergence theorems for priority queues: preemptive-resume discipline. *J. Appl. Prob.* 8 (1971) 74–94.

[67] W. Whitt, Heavy traffic limits for queues: a survey. *Mathematical Methods in Queueing Theory*, Proceedings of a Conference at Western Michigan University, ed. A. B. Clarke, Lecture Notes in Econ. and Math. Systems 98, Springer-Verlag, New York, 1974, 307–350.

[68] W. Whitt, Some useful functional limit theorems. *Math. Oper. Res.* 5 (1980) 67–85.

[69] W. Whitt, Queues with superposition arrival processes in heavy traffic. *Stoch. Proc. Appl.* 21 (1985) 81–91.

[70] W. Whitt, Limits for cumulative input processes to queues. *Prob Engr. Inf. Sci.* (2000) in press.

[71] W. Whitt, The reflection map is Lipschitz with appropriate Skorohod M_1 metrics, AT&T Labs, 1999.

[72] R. J. Williams, An invariance principle for semimartingale reflecting Brownian motions in an orthant. *Queueing Systems* 30 (1998) 5–25.

[73] R. J. Williams, Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems* 30 (1998) 27–88.

[74] W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson, Self-similarity through high variability: statistical analysis of Ethernent LAN traffic at the source level. *IEEE/ACM Trans. Networking* 5 (1997) 71–86.

[75] V. M. Zolotarev, the first passage time of a level and the behavior at infinity for a class of processes with independent increments. *Theor. Prob. Appl.* 9 (1964) 653–662.

[76] V. M. Zolotarev, *One-Dimensional Stable Distributions*, American Math. Society, vol. 65, Providence, RI, 1986.