Chapter 8

Fluid Queues with On-Off Sources

8.1. Introduction

In this chapter we consider a queueing model introduced to help understand the performance of evolving communication networks. As indicated in Section 2.4.1, traffic measurements have shown that the traffic carried on these networks is remarkably bursty and complex, exhibiting features such as heavy-tailed probability distributions, strong positive dependence and self-similarity. These traffic studies have generated strong interest in the impact of heavy-tailed probability distributions and other forms of traffic burstiness upon queueing performance.

A useful model for studying such phenomena is a fluid queue having input from multiple on-off sources; e.g., see Anick et al. (1982), Roberts (1992), Willinger et al. (1997), Taqqu et al. (1997), Choudhury and Whitt (1997), Boxma and Dumas (1998) and Zwart (2001). The queue represents a switch or router in the communication network, where data must be temporarily stored and then forwarded to its destination. The queue may have constant or random release rate, representing the available bandwidth. The actual flow of data in many small packets is modelled as fluid. Each of the many sources alternates between periods when it is busy (active or on) and periods when it is idle (inactive or off). During busy periods, the source transmits data at a constant or random rate; during idle periods, the source is idle, not transmitting anything. The total input to the queue is the superposition (sum) of the inputs from the separate sources, which usually are assumed to be stochastically independent. Given such a model, with stochastic elements specified in more detail, the object is to describe the distributions of quantities such as the buffer content, data loss and end-to-end delay experienced by users.

In this chapter we establish heavy-traffic stochastic-process limits for fluid-queue models with multiple on-off sources. Much of the literature on the fluid queue with multiple on-off sources focuses on the relatively tractable special case of homogeneous (IID) sources in which the busy periods and idle periods come from independent sequences of IID exponentially distributed random variables; e.g., see Anick et al. (1982). We consider more general models, aiming to capture the performance impact of features such as heavy-tailed distributions and strong dependence. We show how the heavy-traffic limits can identify key features determining performance in more complicated queueing models.

The on-off source traffic model represents stochastic fluctuations at two different time scales. The pattern of busy periods and idle periods produces stochastic fluctuations in a longer time scale. The stochastic process depicting the fluid flow during busy periods represents stochastic fluctuations in a shorter time scale. If we let the flow during a busy period be at a deterministic constant rate, then we are deciding in advance that the stochastic fluctuations in the shorter time scale are neglibible compared to the stochastic fluctuations in the longer time scale. More generally, the model gives us the opportunity to compare the impact of stochastic fluctuations in the two different time scales.

We could consider fluctuations in an even longer time scale by letting the number of sources itself evolve as a stochastic process, but here we consider a fixed number of sources. The heavy-traffic stochastic-process limits can be extended to the more general setting by treating the number of sources as a random environment, as in Example 9.6.2.

It will be obvious that in the principal cases the stochastic processes of interest have continuous sample paths. Thus, if there is convergence of a sequence of fluid-queue stochastic processes to a limiting stochastic process with discontinuous sample paths, as we establish in Section 8.5, then the Skorohod M_1 topology must be used. Such limits with discontinuous sample paths will arise under heavy-traffic conditions and heavy-traffic scaling when the busy-period or idle-period distributions of some sources have heavy tails (infinite variance).

Here is how this chapter is organized: In Section 8.2 we introduce the more-detailed multi-source on-off model for the input to a fluid queue, Then in Section 8.3 we apply the continuous-mapping approach again to establish heavy-traffic stochastic-process limits for the more-detailed fluid-queue model.

We consider the special cases of Brownian-motion and stable-Lévy-motion heavy-traffic stochastic-process limits for fluid-queue models in Sections 8.4 and 8.5. In these sections we discuss properties of the reflected limit processes to demonstrate that the stochastic-process limits lead to tractable approximations. In some cases, probability distributions of random quantities associated with the limit process can be given explicitly in closed form. In other cases, the probability distributions can be conveniently characterized via transforms. Then numerical transform inversion can be exploited to calculate the probability distributions. There is a great potential for combining asymptotic and numerical methods.

In some cases, such as with convergence to reflected stable Lévy motion, the limit process is relatively complicated. Then, for applications, there may be interest in developing approximations for the limit process. In Section 8.6 we show how a second stochastic-process limit can be used for that purpose.

We consider strongly-dependent net-input processes in Section 8.7. When the input comes from many independent sources, the central limit theorem for processes in Section 7.2 implies that the net-input process can be approximated by a Gaussian process. With strong dependence, the scaled net-input processes converge to fractional Brownian motion (FBM). Then the associated sequence of scaled workload processes converges to a reflected FBM (RFBM). We develop approximations for the steady-state distribution of RFBM and more general reflected stationary Gaussian processes in Section 8.8.

As in Chapter 5, we give proofs for the theorems in this chapter, but we are primarily interested in the result statements and their applied significance. The proofs draw on material in later chapters.

Remark 8.1.1. *Literature on non-Brownian heavy-traffic limits.* Our discussion of heavy-traffic stochastic-process limits for fluid queues, emphasizing non-Brownian limit processes, follows Whitt (2000a, b). Non-Brownian heavy-traffic limits for queues and related models have been established by Brichet et al. (1996, 2000), Boxma and Cohen (1998, 1999, 2000), Cohen (1998), Furrer, Michna and Weron (1997), Konstantopoulos and Lin (1996, 1998), Kurtz (1996), Resnick and Rootzén (2000), Resnick and Samorodnitsky (2000), Resnick and van den Berg (2000) and Tsoukatos and Makowski (1997, 2000).

8.2. A Fluid Queue Fed by On-Off Sources

In this section we add extra detail to the fluid-queue model introduced in Section 5.2. In particular, we consider multiple on-off sources.

In the fluid queue model, there can be infinite (unlimited) or finite storage space, as discussed in Section 5.2. We assume that there is a single shared buffer receiving the input from all the sources. (We discussed the case of separate source queues in Section 2.4.2.) Fluid can be processed according to the general stochastic process S, as described in Section 5.2. As indicated there, an important special case is $S(t) = \mu t$ for all $t \geq 0$ w.p.1; i.e., the fluid can be processed continuously at constant rate μ whenever there is work to process. However, we consider the general case. For example, it allows us to model service interruptions; for further discussion about service interruptions, see Remark 8.3.2.

8.2.1. The On-Off Source Model

Input arrives from each of m sources. Each source is alternatively busy (active or on) and idle (inactive or off) for random *busy periods* B_i and *idle periods* I_i, $i \geq 1$. Without loss of generality, let the first busy period begin at time 0. (That is without loss of generality, because we can redefine B_1 and I_1 to represent alternative initial conditions; e.g., to start idle, let $B_1 = 0$. To have the busy and idle periods well defined, we assume that $I_i > 0$ for all i and $B_i > 0$ for all $i \geq 2$.) For mathematical tractability, it is natural to assume that the successive pairs (B_i, I_i) after the first are IID, but we do not make that assumption. Our key assumption will be a FCLT for the associated partial sums, which from Chapter 4 we know can hold without that IID assumption.

A *busy cycle* is a busy period plus the following idle period. Thus the *termination time* of the j^{th} busy cycle is

$$T_j \equiv \sum_{i=1}^{j} (B_i + I_i) \ . \tag{2.1}$$

(As before, we use $\equiv$ instead of $=$ to designate equality by a definition.) As a regularity condition, we assume that $T_j \to \infty$ with probability one (w.p.1) as $j \to \infty$. Let $N \equiv \{N(t) : t \geq 0\}$ be the *busy-cycle counting process*, defined by

$$N(t) = \min\{j \geq 0 : T_j \leq t\}, \quad t \geq 0\,, \tag{2.2}$$

where $T_0 \equiv 0$. Let A_j be the set of times when the j^{th} busy period occurs, i.e.,

$$A_j \equiv \{t : T_{j-1} \le t < T_{j-1} + B_j) \ , \tag{2.3}$$

where $T_0 \equiv 0$. Let A be the source *activity period* — the set of times when the source is busy; i.e.,

$$A \equiv \bigcup_{n=1}^{\infty} A_n \ . \tag{2.4}$$

For any set S, let I_S be the *indicator function* of the set S; i.e., $I_S(x) = 1$ if $x \in S$ and $I_S(x) = 0$ otherwise. Thus, for A in (2.4), $\{I_A(t) : t \ge 0\}$ is the *activity process*; $I_A(t) = 1$ if the source is active at time t and $I_A(t) = 0$ otherwise. Let $B(t)$ represent the *cumulative busy time* in $[0, t]$; i.e., let

$$B(t) \equiv \int_0^t 1_A(s)ds, \quad t \ge 0 \ . \tag{2.5}$$

Possible realizations for $\{I_A(t) : t \ge 0\}$ and $\{B(t) : t \ge 0\}$ are shown in Figure 8.1.

Let input come from the source when it is active according to the stochastic process $\{\Lambda(t) : t \ge 0\}$; i.e., let the *cumulative input* during the interval $[0, t]$ be

$$C(t) \equiv (\Lambda \circ B)(t) \equiv \Lambda(B(t)), \quad t \ge 0 \tag{2.6}$$

where $\circ$ is the composition map. (This definition ignores complicated end effects at the beginning and end of busy periods and idle periods.)

We assume that the sample paths of Λ are nondecreasing. A principal case is

$$P(\Lambda(t) = \hat{\lambda}t, \ t \ge 0) = 1$$

for a positive deterministic scalar $\hat{\lambda}$, in which case $C(t) = \hat{\lambda}B(t)$, but we allow other possibilities. (We use the notation $\hat{\lambda}$ here because we have already used λ as the overall input rate, i.e., the rate of $C(t)$.) To be consistent with the busy-period concept, one might require that the sample paths of Λ be strictly increasing, but we do not require it. To be consistent with the fluid concept, the sample paths of Λ should be continuous, in which case the sample paths of C will be continuous. That is the intended case, but we do not require it either.

Notice that the random quantities $\{T_j : j \ge 1\}$ in (2.1), $\{A_j : j \ge 1\}$ in (2.3), A in (2.4), $\{I_A(t) : t \ge 0\}$ and $\{B(t) : t \ge 0\}$ in (2.5) and $\{C(t) : t \ge 0\}$ in (2.6) are all defined in terms of the *basic model elements* — the stochastic processes $\{(B_j, I_j) : j \ge 1\}$ and $\{\Lambda(t) : t \ge 0\}$. Many measures of system

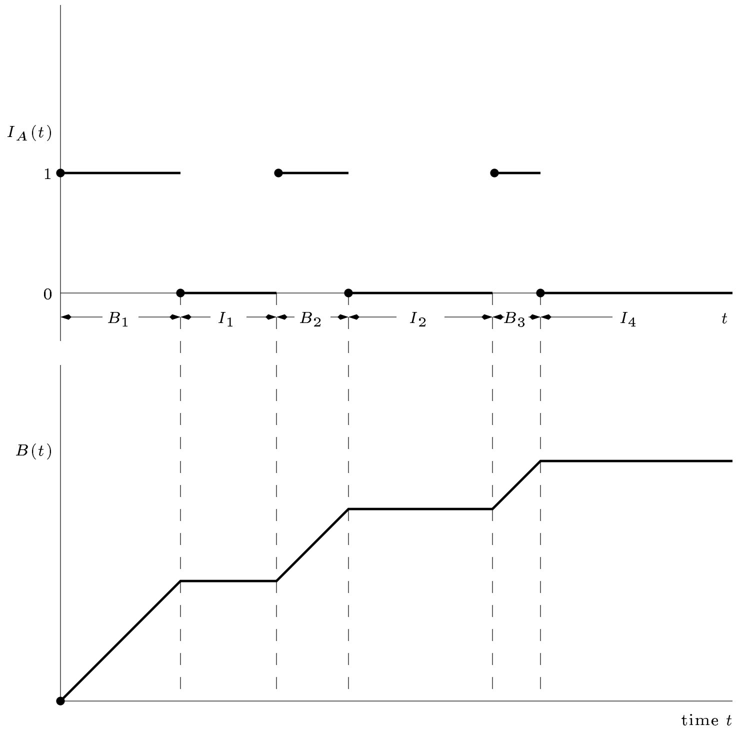

Figure 8.1: Possible realizations for the initial segments of a source-activity process $\{I_A(t) : t \geq 0\}$ and cumulative-busy-time process $\{B(t) : t \geq 0\}$.

performance will depend only on these source characteristics only via the cumulative-input process $\{C(t) : t \geq 0\}$. Also notice that we have imposed no stochastic assumptions yet. By using the continuous-mapping approach, we are able to show that desired heavy-traffic stochastic-process limits hold whenever corresponding stochastic-process limits hold for stochastic processes associated with the basic model data.

Now suppose that we have m sources of the kind defined above. We add an extra subscript l, $1 \leq l \leq m$, to index the source. Thus the basic model elements are $\{\{(B_{l,j}, I_{l,j}) : j \geq 1\}, 1 \leq l \leq m\}$ and $\{\{\Lambda_l(t) : t \geq 0\}, 1 \leq l \leq m\}$. The cumulative input from source l over $[0, t]$ is $C_l(t)$. The cumulative input from all m sources over $[0, t]$ is

$$C(t) \equiv \sum_{l=1}^{m} C_l(t), \quad t \geq 0 \ . \tag{2.7}$$

As indicated above, we suppose that the input from these m sources is fed to a single queue where work is processed according to the available-processing stochastic process S. At this point we can apply Section 5.2 to map the cumulative-input process C and the available-processing stochastic process S into a net-input process, a potential-workload process and the workload process, exploiting a reflection map. We use the one-sided reflection map if there is unlimited storage capacity and the two-sided reflection map if there is limited storage capacity.

Motivated by the fluid notion, it is natural to assume that the sample paths of the single-source cumulative-input processes are continuous. Then the aggregate (for all sources) cumulative-input stochastic process and the buffer-content stochastic process in the fluid queue model also have continuous sample paths. However, as a consequence of the heavy-tailed busy-period and idle-period distributions, the limiting stochastic processes for appropriately scaled versions of these stochastic processes have discontinuous sample paths. Thus, stochastic-process limits with unmatched jumps in the limit process arise naturally in this setting.

Just as for the renewal processes in Section 6.3, it is obvious here that any jumps in the limit process must be unmatched. What is not so obvious, again, is that there can indeed be jumps in the limit process. Moreover, the setting here is substantially more complicated, so that it is more difficult to explain where the jumps come from. To show that there can indeed be jumps in the limit process, we once again resort to simulations and plots. Specifically, we plot the buffer-content process for several specific cases.

8.2.2. Simulation Examples

To have a concrete example to focus on, suppose that we consider a fluid queue with two IID sources. Let the mean busy period and mean idle period both be 1, so that each source is busy half of the time. Let the sources transmit at constant rate during their busy periods. Let the input rate for each source during its busy period be 1. Thus the long-run input rate for each source is $1/2$ and the overall input rate is 1. The instantaneous input rate then must be one of $0, 1$ or 2.

In the queueing example in Section 2.3 we saw that it was necessary to do some careful analysis to obtain the appropriate scaling. In particular, in that discrete-time model, we had to let the finite capacity K and the output rate μ depend on the sample size n in an appropriate way. We avoid that complication here by assuming that the capacity is infinite (by letting $K = \infty$) and by letting the output rate exactly equal the input rate (by letting $\mu = 1$, where μ here is the output rate).

With those parameter choices, the instantaneous net-input rate (input rate minus output rate) at any time must be one of $-1, 0$ or $+1$. To understand the plots, it is good to think about the consequence of having exceptionally long busy periods and idle periods. If both sources are simultaneouly in long busy periods, then there will be a long interval over which the net-input rate is $+1$. Similarly, if both sources are simultaneously in long idle periods, then there will be a long interval over which the net-input rate is -1. If only one source is in a long busy period, then the other source will oscillate between busy and idle periods, so that the net-input rate will oscillate between 0 and $+1$, yielding an average rate of about $+1/2$. Similarly, if only one source is in a long idle period, then again the other source will oscillate between busy and idle periods, so that the net-input rate will oscillate between 0 and -1, yielding an average rate of about $-1/2$.

The likelihood of exceptionally long busy periods (idle periods) depends on the busy-period (idle-period) probability distribution, and these probability distributions have yet to be specified. To illustrate light-tailed and heavy-tailed alternatives, we consider the exponential and Pareto(1.5) distributions. We consider four cases: We consider every combination of the two possible distributions assigned to the busy-period distribution and the idle-period distribution. We call the model Pareto/exponential if the busy-period distribution is Pareto and the idle-period distribution is exponential, and so on.

We plot four possible realizations of the workload sample path for each of the four combinations of busy-period and idle-period distributions in Figures 8.2 – 8.5. We generate $1,000$ busy cycles (idle period plus following busy period) for each source, starting at the beginning of an idle period. We plot the workload process in the interval $[0, T]$, where $T = \min(T_1, T_2)$ with T_i being the time that the $1,000^{th}$ busy cycle ends for source i. Note that $E[T_i] = 2000$.

The differences among the plots are less obvious than before. Places in the plots corresponding to jumps in the limit process have steep slopes. Jumps up only are clearly discernable in the plot of the Pareto/exponential workload in Figure 8.3, while jumps down only are clearly discernable in the plot of the exponential/Pareto workload in Figure 8.4. In contrast, both jumps up and down are discernable in the plots of the Pareto/Pareto workload in Figure 8.5. However, these jumps are not always apparent in every realization. The plots of the exponential/exponential workloads in Figure 8.2 are approaching plots of reflected Brownian motion. Just as for plots of random walks, sometimes plots that are approaching a relected

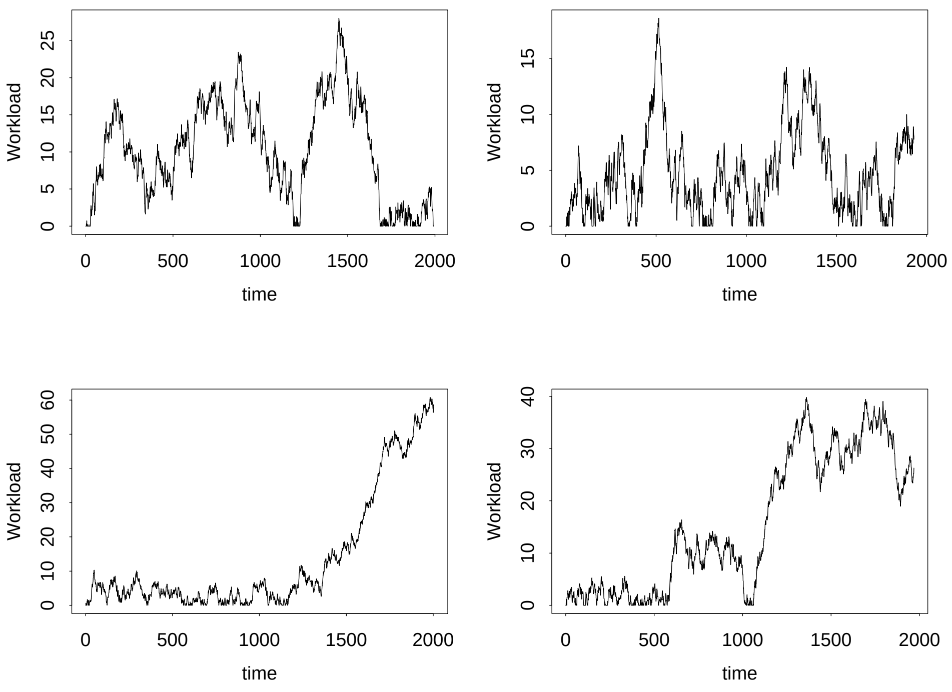

Figure 8.2: Plots of four possible realizations of the workload process in the two-source infinite-capacity exponential/exponential fluid queue having exponential busy-period and idle-period distributions with mean 1, where the input rate equals the output rate. Each source has up to 10^3 busy cycles.

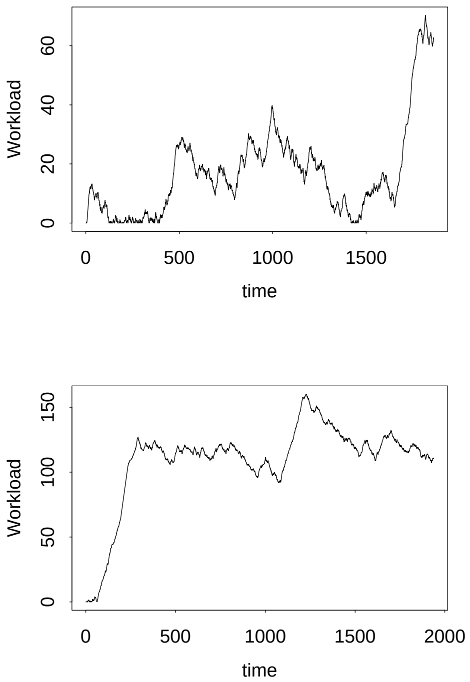

Figure 8.3: Plots of four possible realizations of the workload process in the two-source infinite-capacity Pareto/exponential fluid queue having Pareto(1.5) busy-period distributions and exponential idle-period distributions with mean 1, where the input rate equals the output rate. Each source has up to 10^3 busy cycles.

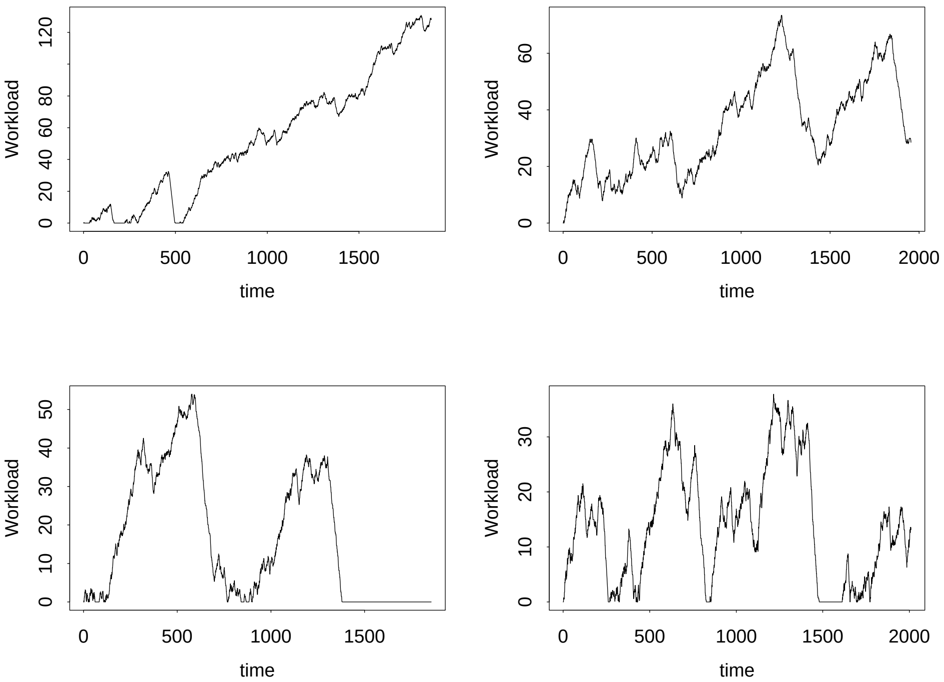

Figure 8.4: Plots of four possible realizations of the workload process in the two-source infinite-capacity exponential/Pareto fluid queue having exponential busy-period distributions and Pareto(1.5) idle-period distributions with mean 1, where the input rate equals the output rate. Each source has up to 10^3 busy cycles.

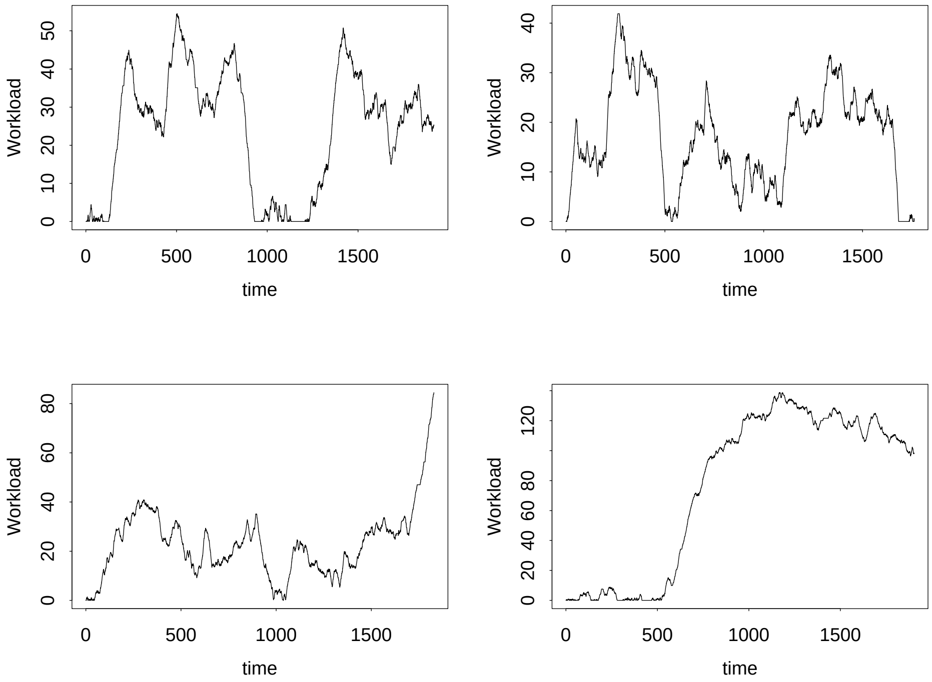

Figure 8.5: Plots of four possible realizations of the workload process in the two-source infinite-capacity Pareto/Pareto fluid queue having Pareto(1.5) busy-period and idle-period distributions with mean 1, where the input rate equals the output rate. Each source has up to 10^3 busy cycles.

stable Lévy motion appear quite similar to other plots that are approaching a reflected Brownian motion. However, additional replications and statistical tests can confirm the differences.

8.3. Heavy-Traffic Limits for the On-Off Sources

In this section, following Whitt (2000b), we establish stochastic-process limits for the more-detailed multi-source on-off fluid-queue model in Section 8.2. From Theorems 5.4.1 and 5.9.1, we see that it suffices to establish a stochastic-process limit for the cumulative-input processes, so that is our goal.

Let model n have m_n on-off sources. Thus the basic model random elements become $\{\{\{(B_{n,l,i}, I_{n,l,i}) : i \geq 1\},\ 1 \leq l \leq m_n\},\ n \geq 1\}$ and $\{\{\{\Lambda_{n,l}(t) : t \geq 0\},\ 1 \leq l \leq m_n\},\ n \geq 1\}$. Let the source cumulative-input processes $C_{n,l}$ be defined as in Section 8.2. Let the aggregate cumulative-input process be the sum as before, i.e.,

$$C_n(t) \equiv C_{n,1}(t) + \cdots + C_{n,m_n}(t), \quad t \geq 0 \ . \tag{3.1}$$

We establish general limit theorems for the stochastic processes $\{N_{n,l}(t) : t \geq 0\}$, $\{B_{n,l}(t) : t \geq 0\}$ and $\{C_{n,l}(t) : t \geq 0\}$. Recall that $N_{n,l}(t)$ represents the number of completed busy cycles in the interval $[0, t]$, while $B_{n,l}(t)$ represents the cumulative busy time in the interval $[0, t]$, both for source l in model n. We first consider the stochastic processes $\{N_{n,l}(t) : t \geq 0\}$ and $\{B_{n,l}(t) : t \geq 0\}$.

8.3.1. A Single Source

We first focus on a single source, so we omit the subscript l here. We now define the scaled stochastic processes in (D, M_1). As before, we use bold capitals to represent the scaled stochastic processes and associated limiting stochastic processes in D. We use the same scaling as in Section 5.4; i.e., we scale time by n and space by c_n, where $c_n/n \to 0$ as $n \to \infty$. Let

$$\begin{aligned}
\mathbf{B}_n(t) &\equiv c_n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (B_{n,i} - m_{B,n}) \\
\mathbf{I}_n(t) &\equiv c_n^{-1} \sum_{i=1}^{\lfloor nt \rfloor} (I_{n,i} - m_{I,n}) \\
\mathbf{N}_n(t) &\equiv c_n^{-1}[N_n(nt) - \gamma_n nt] \\
\mathbf{B}'_n(t) &\equiv c_n^{-1}[B_n(nt) - \xi_n nt], \quad t \geq 0 \ ,
\end{aligned} \tag{3.2}$$

where again $\lfloor nt \rfloor$ is the integer part of nt,

$$\xi_n \equiv \frac{m_{B,n}}{m_{B,n} + m_{I,n}} \quad \text{and} \quad \gamma_n \equiv \frac{1}{m_{B,n} + m_{I,n}} \ . \tag{3.3}$$

We think of $m_{B,n}$ in (3.2) as the mean busy period, $EB_{n,i}$, and $m_{I,n}$ as the mean idle period, $EI_{n,i}$, in the case $\{(B_{n,i}, I_{n,i}) : i \geq 1\}$ is a stationary sequence for each n, but in general that is not required. Similarly, we think of ξ_n in (3.3) as the *source on rate* and γ_n^{-1} in (3.3) as the *mean source cycle time*.

When we consider a sequence of models, as we have done, scaling constants can be incorporated in the random variables. However, in (3.2) and later, we include space and time scaling consistent with what occurs with a single model in heavy traffic.

We first show that a limit for $(\mathbf{B}_n, \mathbf{I}_n)$ implies a limit for $(\mathbf{N}_n, \mathbf{B}'_n)$ jointly with $(\mathbf{B}_n, \mathbf{I}_n)$. That follows by applying the continuous-mapping approach with the addition, inverse and composition functions. Here is a quick sketch of the argument: Since N_n is the counting process associated with the partial sums of $B_{n,i} + I_{n,i}$, we first apply the addition function to treat $\mathbf{B}_n + \mathbf{I}_n$ and then the inverse function to treat $\mathbf{N}_n$. The reasoning for the counting process is just as in Section 7.3. Then the cumulative busy-time $B_n(t)$ is approximately a random sum, i.e.,

$$B_n(t) \approx \sum_{i=1}^{N_n(t)} B_{n,i} \ ,$$

so that we can apply composition plus addition; see Chapter 13. For the case of limit processes with discontinuous sample paths, the last step of the argument is somewhat complicated. Thus, the proof has been put in the Internet Supplement; see Section 5.3 there.

As before, let $Disc(X)$ be the random set of discontinuities in $[0, \infty)$ of X. Let ϕ be the empty set. (We will refer to the reflection map by ϕ_K to avoid any possible confusion.)

Theorem 8.3.1. (FCLT for the cumulative busy time) *If*

$$(\mathbf{B}_n, \mathbf{I}_n) \Rightarrow (\mathbf{B}, \mathbf{I}) \quad in \quad (D, M_1)^2 \tag{3.4}$$

for $\mathbf{B}_n$ and $\mathbf{I}_n$ in (3.2), $c_n \to \infty$, $c_n/n \to 0$, $m_{B,n} \to m_B$, $m_{I,n} \to m_I$, with $0 < m_B + m_I < \infty$, so that $\xi_n \to \xi$ with $0 \leq \xi \leq 1$ and $\gamma_n \to \gamma > 0$ for ξ_n and γ_n in (3.3), and

$$P(Disc(\mathbf{B}) \cap Disc(\mathbf{I}) = \phi) = 1 \ , \tag{3.5}$$

then

$$(\mathbf{B}_n, \mathbf{I}_n, \mathbf{N}_n, \mathbf{B}'_n) \Rightarrow (\mathbf{B}, \mathbf{I}, \mathbf{N}, \mathbf{B}') \quad in \quad (D, M_1)^4 \ , \tag{3.6}$$

for $\mathbf{N}_n$, $\mathbf{B}'_n$ *in* (3.2) *and*

$$\begin{aligned} \mathbf{N}(t) &\equiv -\gamma[\mathbf{B}(\gamma t) + \mathbf{I}(\gamma t)] \\ \mathbf{B}'(t) &\equiv (1-\xi)\mathbf{B}(\gamma t) - \xi \mathbf{I}(\gamma t) \ . \end{aligned} \tag{3.7}$$

Next, given that the single-source cumulative-input process is defined as a composition in (2.6), the continuous-mapping approach can be applied with the convergence preservation of the composition map established in Chapter 13 to show that a joint limit for $B_n(t)$ and $\Lambda_n(t)$ implies a limit for $C_n(t)$, all for one source. Note that this limit does not depend on the way that the cumulative-busy-time processes $B_n(t)$ are defined. Also note that the process $\{\Lambda_n(t) : t \geq 0\}$ representing the input when the source is active can have general sample paths in D. The fluid idea suggests continuous sample paths, but that is not required.

We again omit the source subscript l. Let $\mathbf{e}$ be the identity map on $\mathbb{R}_+$, i.e., $\mathbf{e}(t) = t$, $t \geq 0$. Let

$$\begin{aligned} \mathbf{\Lambda}_n(t) &\equiv c_n^{-1}[\Lambda_n(nt) - \hat{\lambda}_n nt] \\ \mathbf{C}_n(t) &\equiv c_n^{-1}[C_n(nt) - \lambda_n nt], \quad t \geq 0 \ . \end{aligned} \tag{3.8}$$

Theorem 8.3.2. (FCLT for the cumulative input) *If*

$$(\mathbf{\Lambda}_n, \mathbf{B}'_n) \Rightarrow (\mathbf{\Lambda}, \mathbf{B}') \quad in \quad (D, M_1)^2 \tag{3.9}$$

for $\mathbf{\Lambda}_n$ *in* (3.8) *and* $\mathbf{B}'_n$ *in* (3.2), *where* $c_n \to \infty$, $c_n/n \to 0$, $\xi_n \to \xi$, $\hat{\lambda}_n \to \hat{\lambda}$ *for* $0 < \hat{\lambda} < \infty$ *and* $\mathbf{\Lambda} \circ \xi \mathbf{e}$ *and* $\mathbf{B}'$ *have no common discontinuities of opposite sign, then*

$$(\mathbf{\Lambda}_n, \mathbf{B}'_n, \mathbf{C}_n) \Rightarrow (\mathbf{\Lambda}, \mathbf{B}', \mathbf{C}) \quad in \quad (D, M_1)^3 \ . \tag{3.10}$$

for $\mathbf{C}_n$ *in* (3.8) *with* $\lambda_n = \xi_n \hat{\lambda}_n$ *and*

$$\mathbf{C}(t) \equiv \mathbf{\Lambda}(\xi t) + \hat{\lambda} \mathbf{B}'(t) \ . \tag{3.11}$$

Proof. Given that $\mathbf{B}'_n \Rightarrow \mathbf{B}'$, we have $\mathbf{B}''_n \Rightarrow \xi e$ for $\mathbf{B}''_n(t) = n^{-1}B(nt)$. Then we can apply the continuous-mapping approach with composition and addition to treat $\mathbf{C}_n$, i.e., by Corollary 13.3.2,

$$\mathbf{C}_n = \mathbf{\Lambda}_n \circ \mathbf{B}''_n + \hat{\lambda}_n \mathbf{B}'_n \Rightarrow \mathbf{\Lambda} \circ \xi e + \hat{\lambda} \mathbf{B}' \ . \quad \blacksquare$$

From (3.11), we see that the limit processes $\mathbf{\Lambda}$ and $\mathbf{B}'$ appear in $\mathbf{C}$ as a simple sum with deterministic scalar modification. The stochastic fluctuations in the cumulative-input process over a longer (shorter) time scale are captured by the component $\hat{\lambda}\mathbf{B}'(t)$ $(\mathbf{\Lambda}(\xi t)$). Thus, the contribution of each component to the limit process $\mathbf{C}$ can easily be identified and quantified.

As shown in Section 5.4, we can apply the continuous mapping theorem with the reflection map ϕ_K in (2.9) to convert a limit for the cumulative-input processes C_n into a limit for the workload processes W_n and the associated processes L_n, U_n and D_n.

8.3.2. Multiple Sources

Now we are ready to combine Theorems 5.4.1, 8.3.1 and 8.3.2 to obtain simultaneous joint limits for all processes with m sources. To treat the fluid queues, we introduce the sequence of available-processing processes $\{\{S_n(t) : t \geq 0\} : n \geq 1\}$ and the sequence of storage capacities $\{K_n : n \geq 1\}$.

We define the following random elements of (D, M_1) associated with source l, $1 \leq l \leq m$:

$$\begin{aligned}
\mathbf{B}_{n,l}(t) &\equiv c_n^{-1}\sum_{i=1}^{\lfloor nt\rfloor}(B_{n,l,i} - m_{B,n,l}) \\
\mathbf{I}_{n,l}(t) &\equiv c_n^{-1}\sum_{i=1}^{\lfloor nt\rfloor}(I_{n,l,i} - m_{I,n,l}) \\
\mathbf{\Lambda}_{n,l}(t) &\equiv c_n^{-1}[\Lambda_{n,l}(nt) - \hat{\lambda}_{n,l}nt] \\
\mathbf{N}_{n,l}(t) &\equiv c_n^{-1}[N_{n,l}(nt) - \gamma_{n,l}nt] \\
\mathbf{B}'_{n,l}(t) &\equiv c_n^{-1}[B_{n,l}(nt) - \xi_{n,l}nt] \\
\mathbf{C}_{n,l}(t) &\equiv c_n^{-1}[C_{n,l}(nt) - \hat{\lambda}_{n,l}\xi_{n,l}nt] \qquad (3.12)
\end{aligned}$$

Theorem 8.3.3. (heavy-traffic limit for the fluid-queue model with m sources) *Consider a sequence of fluid queues indexed by n with $m_n = m$ sources, capacities K_n, $0 < K_n \leq \infty$, and cumulative-available-processing processes $\{S_n(t) : t \geq 0\}$. Suppose that $K_n = c_nK$, $0 < K \leq \infty$, $W_{n,l}(0) \geq 0$ and $\sum_{l=1}^{m} W_{n,l}(0) \leq K_n$,*

$$\begin{aligned}
&(\mathbf{B}_{n,l}, \mathbf{I}_{n,l}, \mathbf{\Lambda}_{n,l}, c_n^{-1}W_{n,l}(0), 1 \leq l \leq m, \mathbf{S}_n) \\
&\qquad \Rightarrow (\mathbf{B}_l, \mathbf{I}_l, \mathbf{\Lambda}_l, W'_l(0), 1 \leq l \leq m, \mathbf{S}) \qquad (3.13)
\end{aligned}$$

in $(D, M_1)^{3m+1} \times \mathbb{R}^m$ for $\mathbf{B}_{n,l}$, $\mathbf{I}_{n,l}$, $\mathbf{\Lambda}_{n,l}$ in (3.12) and $\mathbf{S}_n$ in (4.4) of Chapter 5, where $c_n \to \infty$, $c_n/n \to 0$, $\hat{\lambda}_{n,l} \to \hat{\lambda}_l$ for $0 < \hat{\lambda}_l < \infty$, $m_{B,n,l} \to m_{B,l}$ and

$m_{I,n,l} \to m_{I,l}$ *with* $0 < m_{B,l} + m_{I,l} < \infty$, *so that*

$$\xi_{n,l} \equiv \frac{m_{B,n,l}}{m_{B,n,l} + m_{I,n,l}} \to \xi_l \tag{3.14}$$

and

$$\gamma_{n,l} \equiv \frac{1}{m_{B,n,l} + m_{I,n,l}} \to \gamma_l > 0 \ . \tag{3.15}$$

If, in addition, $Disc(\mathbf{B}_l \circ \gamma_l \mathbf{e})$, $Disc(\mathbf{I}_l \circ \gamma_l \mathbf{e})$ *and* $Disc(\mathbf{\Lambda}_l \circ \xi_l \mathbf{e})$, $1 \le l \le m$, *are pairwise disjoint w.p.1, and* $\lambda_n - \mu_n \to 0$ *so that*

$$\eta_n \equiv n\left(\lambda_n - \mu_n\right) \Big/ c_n \to \eta \quad as \quad n \to \infty \tag{3.16}$$

for $-\infty < \eta < \infty$, *where*

$$\lambda_n \equiv \sum_{l=1}^{m} \lambda_{n,l} \quad and \quad \lambda_{n,l} \equiv \xi_{n,l} \hat{\lambda}_{n,l} \tag{3.17}$$

for each l, $1 \le l \le m$, *then*

$$\begin{aligned}&(\mathbf{B}_{n,l}, \mathbf{I}_{n,l}, \mathbf{\Lambda}_{n,l}, \mathbf{N}_{n,l}, \mathbf{B}'_{n,l}, \mathbf{C}_{n,l}, 1 \le l \le m, \mathbf{C}_n, \mathbf{S}_n, \mathbf{X}_n, \mathbf{W}_n, \mathbf{U}_n, \mathbf{L}_n) \\ &\quad \Rightarrow (\mathbf{B}_l, \mathbf{I}_l, \mathbf{\Lambda}_l, \mathbf{N}_l, \mathbf{B}'_l, \mathbf{C}_l, 1 \le l \le m, \mathbf{C}, \mathbf{S}, \mathbf{X}, \mathbf{W}, \mathbf{U}, \mathbf{L})\end{aligned}$$

in $(D, M_1)^{6m+6}$ *for* $\mathbf{N}_{n,l}$, $\mathbf{B}'_{n,l}$, $\mathbf{C}_{n,l}$ *in* (3.12), $\mathbf{C}_n$ *in* (3.8) *with* $(\mathbf{X}_n, \mathbf{W}_n, \mathbf{U}_n, \mathbf{L}_n)$ *in* (4.4) *of Chapter* 5 *and*

$$\begin{aligned}
\mathbf{N}_l(t) &\equiv -\gamma_l[\mathbf{B}_l(\gamma_l t) + \mathbf{I}_l(\gamma_l t)] \\
\mathbf{B}'_l(t) &\equiv (1-\xi_l)\mathbf{B}_l(\gamma_l t) - \xi_l \mathbf{I}_l(\gamma_l t) \\
\mathbf{C}_l(t) &\equiv \mathbf{\Lambda}_l(\xi_l t) + \hat{\lambda}_l[(1-\xi_l)\mathbf{B}_l(\gamma_l t) - \xi_l \mathbf{I}_l(\gamma_l t)] \\
\mathbf{C}(t) &\equiv \mathbf{C}_1(t) + \cdots + \mathbf{C}_m(t) \\
\mathbf{X}(t) &\equiv \sum_{l=1}^{m} W'_l(0) + \mathbf{C}(t) - \mathbf{S}(t) + \eta t \\
\mathbf{W}(t) &\equiv \phi_K(\mathbf{X})(t) \\
(\mathbf{U}(t), \mathbf{L}(t)) &\equiv (\psi_U(\mathbf{X})(t), \psi_L(\mathbf{X})(t)), \quad t \ge 0\,,
\end{aligned} \tag{3.18}$$

for ϕ_K *and* (ψ_U, ψ_L) *in* (2.9) *or* (2.10) *and* η *in* (3.16).

In many heavy-tailed applications, the fluctuations of the idle periods $I_{n,l,i}$ and the process $\Lambda_{n,l}(t)$ will be asymptotically negligible compared to the fluctuations of the busy periods $B_{n,l,i}$. In that case, condition (3.13) will hold with $\mathbf{I}_l(t) = \mathbf{\Lambda}_l(t) = 0$, $1 \le l \le m$. Then $\mathbf{C}_l$ in (3.18) simplifies to a simple scaling of the limit process $\mathbf{B}_l$, i.e., then

$$\mathbf{C}_l(t) = \hat{\lambda}_l(1-\xi_l)\mathbf{B}_l(\gamma_l t), \quad t \ge 0 \ . \tag{3.19}$$

Moreover, if some sources have more bursty busy periods than others, then only the ones with highest burstiness will impact the limit. In the extreme case, $\mathbf{B}_l(t)$ will be the zero function for all but one l, say l^*, and

$$\mathbf{C}(t) = \mathbf{C}_{l^*}(t) = \hat{\lambda}_{l^*}(1-\xi_{l^*})\mathbf{B}_{l^*}(\gamma_{l^*} t), \quad t \ge 0 \ , \tag{3.20}$$

and the stochastic nature of the limit for the workload process will be determined, asymptotically, by the single limit process $\{\mathbf{B}_{l^*}(t) : t \ge 0\}$ in (3.13).

In summary, we have shown that a heavy-traffic stochastic-process limit holds for the scaled workload process, with appropriate scaling (including (3.16)), whenever associated stochastic-process limits hold for the basic model elements. In the common case in which we have no initial workloads, deterministic processing according to the rates μ_n and the source input during busy periods is deterministic (i.e., when $\Lambda_{n,l}(t) = \hat{\lambda}_{n,l}t$ for deterministic constants $\hat{\lambda}_{n,l}$), there is a heavy-traffic stochastic-process limit for the workload process whenever the partial sums of the busy periods and idle periods satisfy a joint FCLT. Donsker's theorem and its extensions in Chapter 4 thus yield the required initial FCLTs.

Applying Chapter 4 and Section 2.4 of the Internet Supplement, we obtain convergence of the appropriate scaled cumulative-input processes to Brownian motion, stable Lévy motion and more general Lévy processes, all of which have stationary and independent increments. We describe consequences of those stochastic-process limits in Sections 8.4 and 8.5 below and in Section 5.2 in the Internet Supplement.

Remark 8.3.1. *The effect of dependence.* At first glance, the independent-increments property of the limit process may seem inconsistent with the dependence observed in network traffic measurements in the communications network context, but recall that the limits are obtained under time scaling. Even if successive busy and idle periods for each source are nearly independent, the cumulative inputs

$$C(t_1+h) - C(t_1) \quad \text{and} \quad C(t_2+h) - C(t_2)$$

in disjoint intervals $(t_1, t_1+h]$ and $(t_2, t+h]$ with $t_1 + h < t_2$ are likely to be dependent because a single busy cycle for one source can fall within both intervals.

However, when we introduce time scaling by n, as in (4.4), the associated two scaled intervals $(nt_1, n(t_1+h)]$ and $(nt_2, n(t_2+h)]$ become far apart, so that it is natural that the dependence should disappear in the limit. It is also intuitively clear that the dependence in the original cumulative-input process (before scaling space and time) should have an impact upon performance, causing greater congestion. And that is confirmed by measurements. It is thus important that the limit can capture that performance impact.

Even though the dependence present in the cumulative-input process disappears in the limit, that dependence has a significant impact upon the limit process: *After scaling time, the large busy periods which cause strong dependence over time tend to cause large fluctuations in space.* This effect of time scaling is shown pictorially in Figure 8.6. Thus, even though the limit processes have independent increments, the burstiness in the original cumulative-input process (e.g., caused by a heavy-tailed busy-period distribution) leads to approximating workload distributions that reflect the burstiness. For example, with heavy-tailed busy-period distributions, the limiting workload process will have marginal and steady-state distributions with heavy tails (with related decay rates), which is consistent with what is seen when a single-server queue is simulated with the trace of measured cumulative-input processes. (However, as noted in Section 2.4.1, the pres-

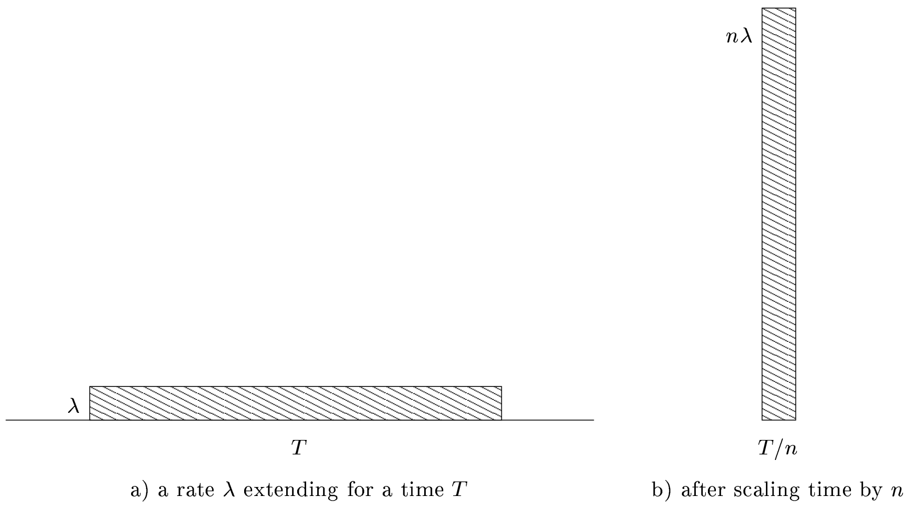

Figure 8.6: The effect of time scaling: transforming dependence over time into jumps in space.

ence of flow controls such as that used by the Transmission Control Protocol

(TCP) significantly complicates interpretations of simulations based on actual network traces.)

Remark 8.3.2. *Structure for the available-processing processes.* Theorem 8.3.3 goes beyond Theorem 5.4.1 by giving the cumulative-input process additional structure. However, in both cases, the available-processing processes $\{S_n(t) : t \geq 0\}$ are kept abstract; we simply assumed that the available-processing processes satisfy a FCLT. However, structure also can be given to the available-processing processes, just as we gave structure to the cumulative-input processes. Indeed, the *same structure* is meaningful. In particular, we can have multiple output channels or servers, each of which is subject to service interruption. We can stipulate that the total available processing in the interval $[0, t]$ is the superposition of the component available-processing processes, i.e., with q output channels,

$$S(t) = S_1(t) + \cdots + S_q(t), \quad t \geq 0 .$$

If server i is subject to service interruptions, then it is natural to model the available-processing process S_i by an on-off model. The busy periods then represent intervals when processing can be done, and the idle periods represent intervals when processing cannot be done. Thus, Theorems 8.3.1 and 8.3.2 can be applied directly to obtain stochastic-process limits for structured available-processing processes. And we immediately obtain generalizations of Theorem 8.3.3 and the Brownian limit in Theorem 8.4.1. For further discussion about queues with service interruptions, see Section 14.7. ■

8.3.3. $M/G/\infty$ Sources

So far we have assumed that the number of sources is fixed, i.e., that $m_n = m$ as $n \to \infty$. An alternative limiting regime is to have $m_n \to \infty$ as $n \to \infty$. We can let $m_n \to \infty$ and still keep the total input rate unchanged by letting the rate from each source decrease. One way to do this is by letting the off periods in each source grow appropriately; e.g., we can let $I_{n,l,i} = m_n I_{l,i}$ and $B_{n,l,i} = B_{l,i}$. Let $N_n(t)$ be the number of new busy cycles started in $[0, t]$ by all sources.

Under regularity conditions, with this scaling of the off periods, $N_n \Rightarrow N$ as $n \to \infty$, where N is a Poisson process. Indeed that stochastic-process limit is the classical stochastic-process limit in which a superposition of independent renewal processes with fixed total intensity (in which each component process is asymptotically negligible) converges to a Poisson process; see

Theorem 9.8.1. Moreover, by a simple continuity argument, the associated cumulative busy-time for all sources, say $\{B_n(t) : t \geq 0\}$ converges to the integral of the number of busy servers in an $M/G/\infty$ queue (a system with infinitely many servers, IID service times and a Poisson arrival process). We state the result as another theorem. Note that there is no additional scaling of space and time here.

Theorem 8.3.4. (an increasing number of sources) *Consider a sequence of fluid models indexed by n with m_n IID fluid sources in model n, where $m_n \to \infty$. Let $I_{n,l,i} = m_n I_{l,i}$ and $B_{n,l,i} = B_{l,i}$ where $\{(B_{l,i}, I_{l,i}) : i \geq 1\}$ is an IID sequence. Then*

$$(N_n, B_n) \Rightarrow (N, B) \quad in \quad (D, M_1)^2 \ , \tag{3.21}$$

where N is a Poisson process with arrival rate $1/EI_{1,1}$,

$$B(t) = \int_0^t Q(s)ds, \quad t \geq 0 \ , \tag{3.22}$$

and $Q(t)$ is the number of busy servers at time t in an $M/G/\infty$ queueing model with arrival process N and IID service times distributed as $B_{1,1}$.

Heavy-traffic stochastic-process limits can be obtained by considering a sequence of models with changing $M/G/\infty$ inputs, as has been done by Tsoukatos and Makowski (1997, 2000) and Resnick and van der Berg (2000).

Let $\{Q_n(t) : t \geq 0\}$ be a sequence processes counting the number of busy servers in $M/G/\infty$ systems. Let $B_n(t)$ be defined in terms of Q_n by (3.22). Let

$$\mathbf{Q}_n(t) \equiv c_n^{-1}[Q_n(t) - \eta_n], \quad t \geq 0 \ , \tag{3.23}$$

and

$$\mathbf{B}'_n(t) \equiv c_n^{-1}[B_n(t) - \eta_n t], \quad t \geq 0 \ . \tag{3.24}$$

(We use the prime to be consistent with (3.2).) We think of $\{Q_n(t) : t \geq 0\}$ as being a stationary process, or at least an asymptotically stationary process, for each n. Hence its translation term in (3.23) is η_n independent of t. In contrast, $B_n(t)$ grows with t, so the translation term of $\mathbf{B}'_n(t)$ in (3.24) is $\eta_n t$.

We can apply the continuous mapping theorem with the integral function in (3.22), using Theorems 3.4.1 and 11.5.1, to show that limits for $\mathbf{Q}_n$ imply corresponding limits for $\mathbf{B}'_n$.

Theorem 8.3.5. (changing M/G/∞ inputs) *Consider a sequence of fluid queue models indexed by n with $M/G/\infty$ inputs characterized by the processes $\{Q_n(t) : t \geq 0\}$. If $\mathbf{Q}_n \Rightarrow \mathbf{Q}$ in (D, M_1) for $\mathbf{Q}_n$ in (3.23), then*

$$\mathbf{B}'_n \Rightarrow \mathbf{B}' \quad in \quad (D, M_1) \ ,$$

for $\mathbf{B}'_n$ in (3.24) and

$$\mathbf{B}'(t) = \int_0^t \mathbf{Q}(s)ds, \quad t \geq 0 \ . \tag{3.25}$$

From heavy-traffic limit theorems for the $M/G/\infty$ queue due to Borovkov (1967) [also see Section 10.3], we know that the condition of Theorem 8.3.5 holds for proper initial conditions when the Poisson arrival rate is allowed to approach infinity. Then $\mathbf{Q}$ is a Gaussian process, which can be fully characterized. In the case of exponential service times, the limit process $\mathbf{Q}$ is the relative tractable Ornstein-Uhlenbeck diffusion process.

We discuss infinite-server models further in Section 10.3. As noted in Remark 10.3.1, the $M/G/\infty$ queue is remarkably tractable, even with a Poisson arrival process having a time-dependent rate. We discuss limits with an increasing number of sources further in Section 8.7.

8.4. Brownian Approximations

We now supplement the discussion of Brownian approximations in Section 5.7 by establishing a Brownian limit for the more detailed multi-source on-off fluid-queue model discussed in Sections 8.2 and 8.3. We discuss Brownian approximations further in Section 9.6.

8.4.1. The Brownian Limit

For simplicity, we let heavy traffic be achieved by only suitably changing the processing rate μ_n, so that the output rate μ_n approaches a fixed input rate λ. Moreover, we can construct the available-processing processes indexed by n by scaling time in a fixed available-processing process S. Let the cumulative available processing in the interval $[0, t]$ in model n be $S_n(t)$. We assume that the available-processing processes S_n satisfy a FCLT, in particular,

$$\mathbf{S}_n \Rightarrow \mathbf{S} \quad \text{in} \quad (D, M_1) \ , \tag{4.1}$$

where $\mathbf{S}$ is a zero-drift Brownian motion with variance parameter σ_S^2 and

$$\mathbf{S}_n(t) \equiv n^{-1/2}(S_n(nt - \mu_n nt), \quad t \geq 0 . \tag{4.2}$$

We use the M_1 topology in (4.1) and later, but it does not play a crucial role in this section because the limit processes here have continuous sample paths.

Since we only change the processing rate μ_n, we can have a single model for the cumulative input process. In the setting of Theorem 8.3.3, we thus assume that the random variables $B_{n,l,i}$ and $I_{n,l,i}$ are independent of n. Moreover, we assume that the processes $\{\Lambda_{n,l}(t) : t \geq 0\}$ are independent of n. Hence we drop the subscript n from these random quantities. We call this a *single fluid model*.

In order to be able to invoke Donsker's FCLT (specifically, the multidimensional version in Theorem 4.3.5), we make several independence assumptions. First, we assume that the m sources are mutually independent. By that we mean that the stochastic processes $\{(B_{l,i}, I_{l,i}) : i \geq 1\}$ and $\{\Lambda_l(t) : t \geq 0\}$ for different l, $1 \leq l \leq m$ are mutually independent. Second, for each source l, we assume that the rate process $\{\Lambda_l(t) : t \geq 0\}$ is independent of the sequence $\{(B_{l,i}, I_{l,i}) : i \geq 1\}$. We also assume that the sources are independent of the available-processing process.

To invoke Donsker's FCLT for $\{(B_{l,i}, I_{l,i}) : i \geq 1\}$, we assume that $\{(B_{l,i}, I_{l,i}) : i \geq 1\}$ is a sequence of IID random vectors in $\mathbb{R}^2$ with finite second moments. In particular, let

$$\begin{aligned} &m_{B,l} \equiv EB_{l,1}, \quad m_{I,l} \equiv EI_{l,1}, \\ &\sigma_{B,l}^2 \equiv Var\, B_{l,1}, \quad \sigma_{I,l}^2 \equiv Var\, I_{l,1}, \\ &\sigma_{B,I,l}^2 \equiv Cov(B_{l,1}, I_{l,1}) . \end{aligned} \tag{4.3}$$

The second-moment assumption implies that all the quantities in (4.3) are finite. The positivity assumption on $B_{l,i}$ and $I_{l,i}$ implies that $m_{B,l} > 0$ and $m_{I,l} > 0$.

Instead of describing the rate processes $\{\Lambda_l(t) : t \geq 0\}$ in detail, we simply assume that they satisfy FCLT's. In particular, we assume that

$$\mathbf{\Lambda}_{n,l} \Rightarrow \mathbf{\Lambda}_l \quad \text{in} \quad (D, M_1) \tag{4.4}$$

for each l, where $\mathbf{\Lambda}_l$ is a zero-drift Brownian motion with variance coefficient $\sigma_{\Lambda,l}^2$ (possibly zero) and

$$\mathbf{\Lambda}_{n,l}(t) \equiv n^{-1/2}[\Lambda_l(nt) - \hat{\lambda}_l nt], \quad t \geq 0 . \tag{4.5}$$

Given $\hat{\lambda}_l$ in (4.5), we can define the overall "input rate" λ by

$$\lambda \equiv \sum_{l=1}^{m} \lambda_l \ , \tag{4.6}$$

where

$$\lambda_l \equiv \xi_l \hat{\lambda}_l \quad \text{and} \quad \xi_l \equiv \frac{m_{B,l}}{m_{B,l} + m_{I,l}} \ . \tag{4.7}$$

The Brownian heavy-traffic stochastic-process limit follows directly from Theorems 8.3.3, 4.3.5 and 11.4.4. We obtain convergence to the same Brownian limit (with different variance parameter) if we replace the assumed independence by associated weak dependence, as discussed in Section 4.4.

Theorem 8.4.1. (Brownian limit for the fluid queue with m on-off sources) *Consider a single fluid-queue model with m on-off sources satisfying the independence assumptions above, the moment assumptions in* (4.3)*, the scaling assumption in* (4.6) *and the convergence assumptions in* (4.4) *and* (4.5)*. Let a sequence of systems indexed by n be formed by letting the capacity in system n be $K_n = \sqrt{n}K$ for some K, $0 < K \leq \infty$, and introducing available-processing processes satisfying the limits in* (4.1) *and* (4.2)*. Let the initial random workloads from the m sources in system n satisfy*

$$W_{n,l}(0) \geq 0, \quad 1 \leq l \leq m, \quad \sum_{l=1}^{m} W_{n,l}(0) \leq K_n \tag{4.8}$$

and

$$n^{-1/2} W_{n,l}(0) \Rightarrow y_l \quad \text{in} \quad \mathbb{R}, \tag{4.9}$$

where y_l is a deterministic scalar. Assume that

$$\eta_n \equiv \sqrt{n}\,(\lambda - \mu_n) \to \eta \quad \text{as} \quad n \to \infty \tag{4.10}$$

for $-\infty < \eta < \infty$, λ in (4.6) *and μ_n in* (4.2)*. Then the conditions and conclusions of Theorems* 5.4.1, 5.9.1, 8.3.1, 8.3.2 *and* 8.3.3 *hold with $c_n \equiv \sqrt{n}$, $W_l'(0) \equiv y_l$, $\gamma_l \equiv (m_{B,l}+m_{I,l})^{-1}$, ξ_l in* (4.7) *and $(\mathbf{B}_l, \mathbf{I}_l, \mathbf{\Lambda}_l)$, $1 \leq l \leq m$, being mutually independent three-dimensional zero-drift Brownian motions, independent of the standard Bownian motion $\mathbf{S}$ in* (4.1)*. For each l, the limit processes $(\mathbf{B}_l, \mathbf{I}_l)$ and $\mathbf{\Lambda}_l$, are mutually independent zero-drift Brownian motions, with $(\mathbf{B}_l, \mathbf{I}_l)$ having covariance matrix*

$$\Sigma_l = \begin{pmatrix} \sigma^2_{B,l} & \sigma^2_{B,I,l} \\ \sigma^2_{B,I,l} & \sigma^2_{I,l} \end{pmatrix} \ . \tag{4.11}$$

The limit processes $\mathbf{N}_l$, $\mathbf{B}'_l$, $\mathbf{C}_l$ and $\mathbf{C}$ are all one-dimensional zero-drift Brownian motions. In particular,

$$\begin{aligned} \mathbf{N}_l &\stackrel{\mathrm{d}}{=} \sigma_{N,l}\mathbf{B}, \qquad & \mathbf{B}'_l &\stackrel{\mathrm{d}}{=} \sigma_{B,l}\mathbf{B} \\ \mathbf{C}_l &\stackrel{\mathrm{d}}{=} \sigma_{\mathbf{C},l}\mathbf{B}, \qquad & \mathbf{C} &\stackrel{\mathrm{d}}{=} \sigma_C\mathbf{B} \ , \end{aligned} \tag{4.12}$$

where $\mathbf{B}$ is a standard Brownian motion and

$$\begin{aligned} \sigma^2_{N,l} &\equiv \gamma_l^3(\sigma^2_{B,l} + 2\sigma^2_{B,I,l} + \sigma^2_{I,l}) \ , \\ \sigma^2_{B',l} &\equiv \gamma_l((1-\xi_l)^2\sigma^2_{B,l} - 2(1-\xi_l)\xi_l\sigma^2_{B,I,l} + \xi_l^2\sigma^2_{I,l}) \ , \\ \sigma^2_{C,l} &\equiv \xi_l\sigma^2_{\Lambda,l} + \hat{\lambda}_l^2\gamma_l((1-\xi_l)^2\sigma^2_{B,l} - 2(1-\xi_l)\xi_l\sigma^2_{B,I,l} + \xi_l^2\sigma^2_{I,l}) \ , \\ \sigma^2_C &\equiv \sum_{l=1}^{m}\sigma^2_{C,l} \ , \\ \sigma^2_X &\equiv \sigma^2_C + \sigma^2_S \ . \end{aligned} \tag{4.13}$$

The limit process $\mathbf{X}$ is distributed as

$$\mathbf{X} \stackrel{\mathrm{d}}{=} \{\mathbf{B}(t;\eta,\sigma^2_X,y) : t \ge 0\} \stackrel{\mathrm{d}}{=} \{y + \eta t + \sigma_X\mathbf{B}(t) : t \ge 0\} \ , \tag{4.14}$$

for η in (4.10), σ^2_X in (4.13) and $y = y_1 + \cdots + y_m$ for y_l in (4.9).

8.4.2. Model Simplification

We now observe that the heavy-traffic stochastic-process limit produces a significant model simplification. As a consequence of Theorem 8.4.1, the scaled workload processes converge to a one-dimension reflected Brownian motion (RBM). In particular, the limit $\mathbf{W}$ is $\phi_K(\mathbf{X})$, where $\mathbf{X}$ is the one-dimensional Brownian motion in (4.14) and ϕ_K is the two-sided reflection map in (2.9) and Section 14.8. The process $\mathbf{W}$ depends on only four parameters: the initial position y specified in (4.9), the drift η specified in (4.10), the diffusion coefficient σ^2_X specified in (4.13) and the upper barrier K for the two-sided reflection. If we are only interested in the steady-state distribution, then we can ignore the initial position y, which leaves only three parameters. From the expression for σ^2_X in (4.13), we can determine the impact of various component sources of variability.

In contrast, the original fluid model has much more structure. The structure was reduced substantially by our independence assumptions, but still there are many model data. First, we need to know the m probability distributions in $\mathbb{R}^2$ of the busy periods $B_{l,1}$ and idle periods $I_{l,1}$, $1 \le l \le m$. These m distributions only affect the limit in Theorem 8.4.1 through their first two moments and covariances. Second, we have the m rate processes $\{\Lambda_l(t) : t \ge 0\}$ and the available-processing process $\{S(t) : t \ge 0\}$. The m rate processes affect the limit only through the scaling parameters $\hat{\lambda}_l$ and $\sigma^2_{\Lambda,l}$ in the assumed FCLT in (4.4) and (4.5). The available-processing process affects the limit only through the scaling parameter σ^2_S in (4.1). We also have the m initial random workloads $W_{n,l}(0)$. The random vector $(W_{n,1}(0), \ldots, W_{n,m}(0))$ can have a very general m-dimensional distribution. That distribution affects the limit only via the deterministic limits y_l in (4.8). Since we have convergence in distribution to a deterministic limit, we automatically get convergence in the product space; i.e.,

$$n^{-1}(W_{n,1}(0), \ldots, W_{n,m}(0)) \Rightarrow (y_1, \ldots, y_m) \quad \text{in} \quad \mathbb{R}^m$$

as $n \to \infty$ by Theorem 11.4.5. Finally, we have the capacity K_n and the processing rate μ_n. Thus, for the purpose of determining the limit, there are $8m + 3$ relevant parameters.

It is significant that the stochastic-process limit clearly shows how the $8m+3$ parameters in the original fluid model should be combined to produce the final four parameters characterizing the limiting RBM. First, the final drift η depends only upon the parameters μ_n, ξ_l and λ_l for $1 \le l \le m$ and n as indicated in (4.10). Second, the final variance parameter σ^2_X depends on the basic model parameters as indicated in (4.13). We can thus quickly evaluate the consequence of altering the rate or variability of the stochastic processes characterizing the model.

We now consider the Brownian approximation for the distribution of the steady-state workload in the multi-source on-off model. The steady-state distribution of RBM is given in Theorem 5.7.2. By Theorem 8.4.1, the key parameter is

$$\theta \equiv \frac{2\eta}{\sigma^2} \approx \frac{2\sqrt{n}\,((\sum_{l=1}^m \lambda_l) - \mu_n)}{\sigma^2_X} \tag{4.15}$$

where $\lambda_l \equiv \xi_l \hat{\lambda}_l$ and $\mathbf{W}(\infty)$ has pdf in (7.10). As $K \to \infty$, the mean of $\mathbf{W}(\infty)$ approaches $-\theta^{-1}$ when $\eta < 0$. Then $W_n(\infty)$ has approximately an exponential distribution with mean

$$\begin{aligned} E\mathbf{W}_n(\infty) &\approx \sqrt{n}E\mathbf{W}(\infty) \\ &\approx \frac{\sqrt{n}\sigma^2_X}{2\sqrt{n}\,(\mu_n - \sum_{l=1}^m \lambda_l)} = \frac{\sigma^2_X}{2\,(\mu_n - \sum_{l=1}^m \lambda_l)} \ . \end{aligned} \tag{4.16}$$

Notice that the $\sqrt{n}$ factors in (7.14) and (4.15) cancel out in (4.16).

It is instructive to consider the contribution of each source to the overall mean steady-state workload. Expanding upon formula (4.16) in the case $K = \infty$, we obtain the approximation

$$EW(\infty) \approx \frac{(\sigma_S^2 + \sum_{l=1}^m \sigma_{C,l}^2)}{2(\mu - \sum_{l=1}^m \lambda_l)} \ . \tag{4.17}$$

Source l contributes to the approximate mean steady-state workload via both its rate λ_l and its variance parameter $\sigma_{C,l}^2$.

8.5. Stable-Lévy Approximations

Paralleling Theorem 8.4.1, we now combine the FCLT obtaining convergence of normalized partial-sum processes to stable Lévy motion (SLM) in Section 4.5 and the general limits for multi-source on-off fluid queues in Section 8.3 in order to obtain a reflected-stable-Lévy-motion (RSLM) stochastic-process limit for the multi-source on-off fluid queue. In particular, we assume that the busy-period distribution has a heavy tail, which makes the limit for the normalized cumulative-input process be a centered totally-skewed stable Lévy motion (having $\mu = 0$ and $\beta = 1$). The limit for the normalized workload process will be the reflection of a constant drift plus this centered totally-skewed Lévy stable motion. See Section 6.4.3 for RSLM limits for a more conventional queueing model.

It is possible to obtain more general reflected Lévy process limits for general sequences of models, which also have remarkably tractable steady-state distributions when the underlying Lévy process has no negative jumps. We discuss these more general limits in Sections 2.4 and 5.2 of the Internet Supplement.

8.5.1. The RSLM Heavy-Traffic Limit

As in Section 5.7, the limit will be achieved under heavy-traffic conditions. As before, we can let heavy-traffic be achieved by suitably changing the deterministic processing rate μ_n by scaling a fixed available-processing stochastic process S, so that the output rate approaches the constant input rate. As in Section 5.7, we thus have a single model for the cumulative-input process in the fluid queue, so that we are in the setting of the single-sequence limit in Section 4.5. In the setting of Theorem 8.3.3, the variables $B_{n,l,i}$ and $I_{n,l,i}$ are independent of n. As in Section 5.7, we assume that the processes $\{\Lambda_{n,l}(t) : t \geq 0\}$ also are independent of n. Hence we drop the subscript n from these quantities. We again call this a single fluid model.

When the random variables have infinite second moments and appropriately scaled versions of the random walk converge to a stable process, the scaling depends critically on the tail probability decay rate or, equivalently, the stable index α. Hence it is natural for one component in the model to dominate in the sense that it has a heavier tail than the other components. We will assume that the busy-period distributions have the heaviest tail, so that the stochastic fluctuations in the idle periods I_l, the rate processes $\{\Lambda_t(t) : t \geq 0\}$ and the available-processing process $\{S(t) : t \geq 0\}$ become asymptotically negligible. (That is conveyed by assumption (5.5) in Theorem 8.5.1 below.) It is straightforward to obtain the corresponding limits in the other cases, but we regard this case as the common case. It has the advantage of not requiring conditions involving joint convergence.

We need fewer independence assumptions than we needed in Section 5.7. In particular, now we assume that the m busy-period sequences $\{B_{l,i} : i \geq 1\}$ are mutually independent sequences of IID random variables. We also assume that the idle-periods come from sequences $\{I_{l,i} : i \geq 1\}$ of IID random variables for each l, but that could easily be weakened. As in Section 5.7, we make a stochastic-process-limit assumption on the rate processes $\{\Lambda_l(t) : t \geq 0\}$ and the available-processing process $\{S(t) : t \geq 0\}$ instead of specifying the structure in detail.

Theorem 8.5.1. (RSLM limit for the multi-source fluid queue) *Consider a single fluid model with m sources satisfying the independence assumptions above and the scaling assumption in* (4.6)*. Let a sequence of systems indexed by n be formed by having the capacity in system n be $K_n = n^{1/\alpha}K$ for some K, $0 < K \leq \infty$. . Suppose that $1 < \alpha < 2$. Let the initial random workloads from the m sources in system n satisfy*

$$W_{n,l}(0) \geq 0, \quad 1 \leq l \leq m, \quad \sum_{l=1}^{m} W_n(0) \leq K_n \ , \tag{5.1}$$

and

$$n^{-1/\alpha} W_{n,l}(0) \Rightarrow y_l \quad in \quad \mathbb{R}, \quad 1 \leq l \leq m \ . \tag{5.2}$$

Assume that

$$x^{\alpha} P(B_{l,1} > x) \to A_l \quad as \quad x \to \infty \ ; \tag{5.3}$$

where $0 \leq A_l < \infty$, and

$$x^{\alpha} P(I_{l,1} > x) \to 0 \quad as \quad x \to \infty \tag{5.4}$$

for $1 \le l \le m$. *Assume that*

$$\mathbf{\Lambda}_{n,l} \Rightarrow \mathbf{0} \quad \text{and} \quad \mathbf{S}_n \Rightarrow \mathbf{0} \quad \text{in} \quad (D, M_1)\ , \tag{5.5}$$

where

$$\mathbf{\Lambda}_{n,l}(t) \equiv n^{-1/\alpha}(\Lambda_l(nt) - \hat{\lambda}_l nt), \quad t \ge 0 \tag{5.6}$$

for each l, $1 \le l \le m$ *and*

$$\mathbf{S}_{n,l}(t) \equiv n^{-1/\alpha}(S(nt) - \mu_n nt), \quad t \ge 0\ . \tag{5.7}$$

Assume that

$$\eta_n \equiv n^{(1-\alpha^{-1})}(\lambda - \mu_n) \to \eta \quad \text{as} \quad n \to \infty \tag{5.8}$$

for λ *in* (4.6) *and* μ_n *in* (5.7). *Then the conditions and conclusions of Theorems* 8.3.3, 5.9.1 *and* 5.9.2 (*when* $K = \infty$ *in the last case*) *hold with* $c_n = n^{1/\alpha}$, $m_{B,n,l} = m_{B,l} = EB_{l,1}$, $m_{I,n,l} = m_{I,l} = EI_{l,1}$, $\xi_{n,l} = \xi_l$, $\gamma_{n,l} = \gamma_l > 0$, $\mathbf{I}_l = \mathbf{\Lambda}_l = \mathbf{S} = \mathbf{0}$, $W_l'(0) = y_l$, $1 \le l \le m$, *and* $\mathbf{B}_1, \ldots, \mathbf{B}_m$ *being* m *mutually independent stable Lévy motions with*

$$\mathbf{B}_l(t) \stackrel{\mathrm{d}}{=} S_\alpha(\sigma_l t^{1/\alpha}, 1, 0), \quad t \ge 0\ , \tag{5.9}$$

where

$$\sigma_l \equiv (A_l/C_\alpha)^{1/\alpha} \tag{5.10}$$

for A_l *in* (5.3) *and* C_α *in* (5.14) *in Section* 4.5. *Moreover,*

$$\begin{aligned}
&(\mathbf{N}_l, \mathbf{B}_l', \mathbf{C}_l, \mathbf{C})(t) \\
&\quad = (-\gamma_l \mathbf{B}_l(\gamma_l t), (1-\xi_l)\mathbf{B}_l(\gamma_l t), \\
&\qquad\quad \hat{\lambda}_l(1-\xi_l)\mathbf{B}_l(\gamma_l t), \sum_{l=1}^m \hat{\lambda}_l(1-\xi_l)\mathbf{B}_l(\gamma_l t)) \\
&\quad \stackrel{\mathrm{d}}{=} \Big(-\gamma_l^{1+\alpha^{-1}}\mathbf{B}_l(t), (1-\xi_l)\gamma_p^{\alpha^{-1}}\mathbf{B}_l(t), \\
&\qquad\quad \hat{\lambda}_l(1-\xi_l)\gamma_l^{\alpha^{-1}}\mathbf{B}_l(t), \sum_{l=1}^m \hat{\lambda}_l(1-\xi_l)\gamma_l^{\alpha^{-1}}\mathbf{B}_l(t)\Big),
\end{aligned} \tag{5.11}$$

so that $\mathbf{C}$ *is a centered stable Lévy motion with*

$$\mathbf{C}(t) \stackrel{\mathrm{d}}{=} \left(\sum_{l=1}^m \hat{\lambda}_l^\alpha (1-\xi_l)^\alpha \gamma_l A_l/C_l\right)^{1/\alpha} S_\alpha(t^{\alpha^{-1}}, 1, 0) \tag{5.12}$$

and

$$\mathbf{X}(t) = y + \eta t + \mathbf{C}(t), \quad t \geq 0 \ , \tag{5.13}$$

for $y = y_1 + \cdots + y_m$. *Hence the limit* $\mathbf{W} = \phi_K(\mathbf{X})$ *is the reflection of the linear drift* $\eta\mathbf{e}$ *plus standard stable Lévy motion* $\mathbf{C}$ *having parameter four-tuple* $(\alpha, \sigma, \beta, \mu) = (\alpha, \sigma_C, 1, 0)$ *with*

$$\sigma_C = \left(\sum_{l=1}^{m} \hat{\lambda}_l^{\alpha} (1 - \xi_l)^{\alpha} \gamma_l \sigma_l^{\alpha} \right)^{1/\alpha} , \tag{5.14}$$

starting at initial position y, $0 \leq y \leq K$, *with reflection having upper barrier at* K, $0 < K \leq \infty$.

Proof. We apply Theorem 8.3.3. First note that (5.3) and (5.4) imply that

$$0 < EB_{l,1} < \infty \quad \text{and} \quad 0 < EI_{l,1} < \infty \ . \tag{5.15}$$

To apply Theorem 8.3.3, we need to verify condition (3.13). The limits for $\mathbf{B}_{n,l}$ individually follow from Theorem 4.5.3, using the normal-domain of attraction result in Theorem 4.5.2. The joint FCLT for $(\mathbf{B}_{n,1}, \ldots, \mathbf{B}_{n,m})$ then follows from Theorem 11.4.4. Since the limits for $\mathbf{I}_{n,l}$, $\mathbf{\Lambda}_{n,l}$ and $\mathbf{S}_n$ are deterministic, we obtain an overall joint FCLT from the individual FCLTs, using Theorem 11.4.5. That yields the required stochastic-process limit in (3.13). See (5.7) – (5.11) in Section 4.5 for the scaling yielding (5.12). For Theorem 5.9.2 with $K = \infty$, we use the fact that $\mathbf{S} = 0\mathbf{e}$ and $\mathbf{C}$ has nonnegative jumps to conclude that $P(\mathbf{L} \in C) = 1$ and that all the conditions on the discontinuities of the limit processes in Theorem 5.9.2 are satisfied in this setting. ∎

Theorem 8.5.1 tells us what are the key parameters governing system performance. First, the nonstandard space scaling by $c_n = n^{\alpha}$ shows that the scaling exponent α in (5.3) is critical. But clearly the values of the means, which necessarily are finite with $1 < \alpha < 2$, are also critical. The values of the means appear via the asymptotic drift η in (5.8).

Since the stable laws with $1 < \alpha < 2$ have infinite variance, there are no variance parameters describing the impact of variability, as there were in Section 5.7. However, essentially the same variability parameters appear; they just cannot be interpreted as variances. In both cases, the variability parameters appear as *scale factors* multiplying canonical limit processes. In Theorem 8.4.1, the variability parameter is the parameter σ_X appearing as a multiplicative factor before the standard Brownian motion $\mathbf{B}$ in (4.14). Correspondingly, in Theorem 8.5.1, the variability parameter is the stable scale parameter σ_C in (5.14) and (5.12). That scale parameter depends critically on the parameters A_l, which are the second-order parameters appearing in the tail-probability asymptotics for the busy-period distributions in (5.3).

The stable Lévy motions and reflected stable Lévy motions appearing in Theorem 8.5.1 are less familiar than the corresponding Brownian motions and reflected Brownian motions appearing in Section 5.7, but they have been studied quite extensively too; e.g., see Samorodnitsky and Taqqu (1994) and Bertoin (1996).

8.5.2. The Steady-State Distribution

It is significant that the limiting stable Lévy motion for the cumulative-input process is *totally skewed*, i.e., has skewness parameter $\beta = 1$, so that the stable Lévy motion has *sample paths without negative jumps*. That important property implies that the reflected stable Lévy motion with negative drift has a relatively simple steady-state distribution, both with and without an upper barrier. That is also true for more general reflected Lévy processes; see Section 5.2 in the Internet Supplement, Theorem 4.2 of Kella and Whitt (1991) and Section 4 (a) of Kella and Whitt (1992c). With a finite upper barrier, the steady-state distribution is the steady-state distribution with infinite capacity, truncated and renormalized. Equivalently, the finite-capacity distribution is the conditional distribution of the infinite-capacity content, given that the infinite-capacity content is less than the upper barrier.

Theorem 8.5.2. (steady-state distribution of RSLM) *Let* $\mathbf{R} \equiv \phi_K(\eta\mathbf{e} + \mathbf{S})$ *be the reflected stable Lévy motion with negative drift* ($\eta < 0$) *and SLM* $\mathbf{S}$ *having stable index* α, $1 < \alpha < 2$, *and scaling parameter* σ *in* (5.12) (*with* $\beta = 1$ *and* $\mu = 0$) *arising as the limit in Theorem* 8.5.1 (*where* $\mathbf{S}(1) \stackrel{\mathrm{d}}{=} S_\alpha(\sigma, 1, 0) \stackrel{\mathrm{d}}{=} \sigma S_\alpha(1, 1, 0)$).

(a) If $K = \infty$, *then*

$$\lim_{t\to\infty} P(\mathbf{R}(t) \le x) = H(x) \quad \textit{for all} \quad x \; , \tag{5.16}$$

where H *is a proper cdf with pdf* h *on* $(0, \infty)$ *with Laplace transform*

$$\hat{h}(s) \equiv \int_0^\infty e^{-sx} h(x)dx = \frac{s\hat{\phi}'(0)}{\hat{\phi}(s)} = \frac{1}{1 + (\nu s)^{\alpha-1}} \; , \tag{5.17}$$

with

$$\hat{\phi}(s) \equiv \log Ee^{-s[\eta + S_\alpha(\sigma,1,0)]} = -\eta s(\sigma^\alpha s^\alpha / \cos(\pi\alpha/2)) \tag{5.18}$$

and the scale factor ν $(H_\nu(x) = H_1(x/\nu))$ *being*

$$\nu^{\alpha-1} \equiv \frac{\sigma^\alpha}{\eta \cos(\pi\alpha/2)} > 0 \ . \tag{5.19}$$

The associated ccdf $H^c \equiv 1 - H$ *has Laplace transform*

$$\begin{aligned} \hat{H}^c(s) \equiv \int_0^\infty e^{-sx} H^c(x)dx &= \frac{1-\hat{h}(s)}{s} \\ &= \frac{\nu}{(\nu s)^{2-\alpha}(1+(\nu s)^{\alpha-1})} \ . \end{aligned} \tag{5.20}$$

(b) If $c < 0$ *and* $K < \infty$, *then*

$$\lim_{t\to\infty} P(\mathbf{R}(t) \le x) = \frac{H(x)}{H(K)}, \quad 0 \le x \le K \ , \tag{5.21}$$

where H *is the cdf in* (5.16).

We have observed that the heavy-traffic limit for the workload processes depends on the parameter five-tuple $(\alpha, \eta, \sigma, K, y)$. When we consider the steady-state distribution, the initial value y obviously plays no role, but we can reduce the number of relevant parameters even further: From Theorem 8.5.2, we see that the steady-state distribution depends on the parameter four-tuple $(\alpha, \eta, \sigma, K)$ only via the parameter triple (α, ν, K) for ν in (5.19).

As should be anticipated, the steady-state distribution in Theorem 8.5.2 has a heavy tail. Indeed it has infinite mean. We can apply Heaviside's theorem, p. 254 of Doetsch (1974) to deduce the asymptotic form of the ccdf H^c and pdf h in Theorem 8.5.2. We display the first two terms of the asymptotic expansions below. (The second term for $\alpha = 3/2$ seems inconsistent with the second term for $\alpha \neq 3/2$; that occurs because the second term for $\alpha = 3/2$ actually corresponds to the third term for $\alpha \neq 3/2$, while the second term is zero.)

Theorem 8.5.3. (tail asymptotics for steady-state distribution) *For* $\nu = 1$, *the steady-state ccdf and pdf in Theorem* 8.5.2 *satisfy*

$$H^c(x) \sim \begin{cases} \frac{1}{\Gamma(2-\alpha)x^{\alpha-1}} - \frac{1}{\Gamma(3-2\alpha)x^{2(\alpha-1)}}, & \alpha \neq 3/2 \\ \frac{1}{\sqrt{\pi}x^{1/2}} - \frac{1}{2\sqrt{\pi}x^{3/2}}, & \alpha = 3/2 \end{cases} \tag{5.22}$$

and

$$h(x) \sim \begin{cases} \frac{-1}{\Gamma(1-\alpha)x^{\alpha}} + \frac{1}{\Gamma(2-2\alpha)x^{2\alpha-1}}, & \alpha \neq 3/2 \\ \\ \frac{1}{2\sqrt{\pi}x^{3/2}} - \frac{3}{4\sqrt{\pi}x^{5/2}}, & \alpha = 3/2 \end{cases} \tag{5.23}$$

as $x \to \infty$, where $\Gamma(x)$ is the gamma function.

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 (1972) to invert the Laplace transforms analytically in terms of the error function, which is closely related to the standard normal cdf. A similar explicit expression is possible for a class of $M/G/1$ steady-state workload distributions; see Abate and Whitt (1998). That can be used to make numerical comparisons.

Theorem 8.5.4. (explicit expressions for $\alpha = 3/2$) *For $\alpha = 3/2$ and $\nu = 1$, the limiting pdf and ccdf in Theorem* 8.5.2 *are*

$$h(x) = \frac{1}{\sqrt{\pi x}} - 2e^x \Phi^c(\sqrt{2x}), \quad x \geq 0 \ , \tag{5.24}$$

and

$$H^c(x) = 2e^x \Phi^c(\sqrt{2x}), \quad x \geq 0 \ , \tag{5.25}$$

where Φ^c is again the standard normal ccdf.

Theorems 8.5.2 – 8.5.4 provide important practical engineering insight. Note the the ccdf $H^c(x)$ of the steady-state distribution of RSLM with no upper barrier decays slowly. Specifically, from (5.22), we see that it decays as the power $x^{-(\alpha-1)}$. In fact, the exponent $-(\alpha-1)$, implies a slower decay than the decay rate $-\alpha$ of the heavy-tailed busy-period ccdf $B^c_{l,1}(x)$ in (5.3). Indeed, the density $h(x)$ of the steady-state RSLM decays at exactly the same rate as the busy-period ccdf.

In contrast, for the RBM limit (with negative drift) in Section 5.7, the steady-state distribution decays exponentially. In both cases, the steady-state distribution with a finite buffer is the steady-state distribution with an infinite buffer truncated and renormalized. Equivalently, the steady state distribution with a finite buffer is the steady-state distribution with an infinite buffer condioned to be less than the buffer capacity K. That property supports corresponding approximations for finite-capacity queueing systems; see Whitt (1984a). The fact the ccdf $H^c(x)$ of the steady-state distribution of RSLM decays as a power implies that a buffer will be less effective in reducing losses than it would be for a workload process that can be approximated by RBM.

Also note that the asymptote in (5.22) provides a useful "back-of-the envelope" approximation for the ccdf of the steady-state distribution:

$$P(W_\rho(\infty) > x) \approx P((\zeta/(1-\rho))^{1/(\alpha-1)} Z > x) \ , \tag{5.26}$$

where

$$P(Z > x) \approx C x^{-(\alpha-1)} \tag{5.27}$$

for constants ζ and C determined by (5.11) and (5.22).

The asymptotic tail in (5.26) and (5.27) is consistent with known asymptotics for the exact steady-state workload ccdf in special cases; e.g., see the power-tail references cited in Remark 5.4.1.

By comparing the second term to the first term of the asymptotic expansion of the ccdf $H^c(x)$ in Theorem 8.5.3, 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 $x \to 0$.

8.5.3. Numerical Comparisons

We now show that the anticipated structure deduced from examining Theorem 8.5.3 actually holds by making numerical comparisons with exact values computed by numerically inverting the Laplace transform in (5.20). To do the inversion, we use the Fourier series method in Abate and Whitt (1995a); see Abate, Choudhury and Whitt (1999) for an overview.

We display results for $\alpha = 1.5$, $\alpha = 1.9$ and $\alpha = 1.1$ in Tables 8.1–8.3. For $\alpha = 1.5$, Table 8.1 shows that the one-term asymptote is a remarkably accurate approximation for x such that $H^c(x) \leq 0.20$. In Table 8.1 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.

Tables 8.2 and 8.3 show that the one-term asymptote is a much less accurate approximation for α away from 1.5. In the case $\alpha = 1.9$ ($\alpha = 1.1$),

	$H^c(x)$ 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 8.1: A comparison of the limiting ccdf $H^c(x)$ in Theorem 8.5.2 for $\alpha = 1.5$ and $\nu = 1$ with the one-term asymptote and the alternative exact values for $\alpha = 1.49$ and $\alpha = 1.40$.

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.

The exponent $\alpha = 2$ is a critical boundary point for the ccdf tail behavior: 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 8.2.

Table 8.3 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.

	$H^c(x)$ for $\alpha = 1.9$ $(\nu = 1)$		
		one-term	exponential
x	exact	asymptote	$\alpha = 2$
$0.1 \times 2^0 = 0.1$	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 8.2: A comparison of the reflected stable ccdf $H^c(x)$ in Theorem 8.5.2 for $\alpha = 1.9$ and $\nu = 1$ with the one-term asymptote and the mean-1 exponential ccdf corresponding to $\alpha = 2$.

Remark 8.5.1. *First passage times.* We have applied numerical transform inversion to calculate steady-state tail probabilities of RSLM. To describe the transient behavior, we might want to compute first-passage-time probabilities. Rogers (2000) has developed a numerical transform inversion algorithm for calculating first-passage-time probabilities in Lévy processes with jumps in at most one direction.

8.6. Second Stochastic-Process Limits

The usual approximation for a stochastic process $\mathbf{X}_n$ based on a stochastic-process limit $\mathbf{X}_n \Rightarrow \mathbf{X}$ is $\mathbf{X}_n \approx \mathbf{X}$. However, if the limit process $\mathbf{X}$ is not convenient, then we may want to consider developing approximations for the limit process $\mathbf{X}$. We can obtain such additional approximations by considering yet another stochastic-process limit. We describe two approaches in this section.

The first approach is based on having another stochastic-process limit with the same limit process $\mathbf{X}$: We may be able to establish both $\mathbf{X}_n \Rightarrow \mathbf{X}$ and $\mathbf{Y}_n \Rightarrow \mathbf{X}$, where the process $\mathbf{X}_n$ is the scaled version of the relatively complicated process of interest and $\mathbf{Y}_n$ is the scaled version of another more tractable process. Then we can use the double approximation

$$\mathbf{X}_n \approx \mathbf{X} \approx \mathbf{Y}_m ,$$

where n and m are suitably large. We discuss two possible approximations

	$H^c(x)$ for $\alpha = 1.1$ $(\nu = 1)$		
		one-term	two-term
x	exact	asymptote	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 8.3: A comparison of the reflected stable ccdf $H^c(x)$ in Theorem 8.5.2 for $\alpha = 1.1$ and $\nu = 1$ with the one-term and two-term asymptotes from Theorem 8.5.3.

for processes associated with a reflected-stable-Lévy-motion (RSLM) in the first subsection below.

The second approach is to approximate the limit process $\mathbf{X}$ by establishing a further stochastic-process limit for scaled versions of the limit process $\mathbf{X}$ itself. That produces an approximation that should be relevant in an even longer time scale, since time is now scaled twice. We discuss this second approach in the second subsection below.

8.6.1. M/G/1/K Approximations

We now show how the two-limit approach can be used to generate $M/G/1/K$ approximations for complicated queueing models such as the multi-source on-off fluid-queue model with heavy tailed busy-period distributions. We first apply Theorem 8.5.1 to obtain a reflected-stable-Lévy-motion (RSLM) approximation for the workload process in the fluid queue. We then construct a sequence of $M/G/1/K$ models with scaled workload processes converging to the RSLM.

Consider a stable Lévy motion $\sigma\mathbf{S} + \eta\mathbf{e}$, where $\mathbf{S}(1) \stackrel{d}{=} S_\alpha(1,1,0)$ with $1 < \alpha < 2$ and $\sigma > 0$, modified by two-sided reflection at 0 and K. We now show that, for any choice of the four paramerters α, σ, η and K, we can construct a sequence of M/G/1/K fluid queues such that the scaled M/G/1/K net-input processes, workload processes, overflow processes and departure processes converge to those associated with the given RSLM. Since we can apply the continuous-mapping approach with the two-sided reflection map, it suffices to show that the scaled net-input processes converge to $\sigma\mathbf{S} + \eta\mathbf{e}$. We can apply Theorems 5.4.1 and 5.9.1.

The specific M/G/1/K fluid-queue model corresponds to the standard M/G/1/K queue with bounded virtual waiting time, which is often called the *finite dam*; see Section III.5 of Cohen (1982). For the workload process, there is a constant output rate of 1. As in Example 5.7.1, The cumulative input over the interval $[0,t]$ is the sum of the service times of all arrivals in the interval $[0,t]$, i.e., the cumulative input is

$$C(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0 \ ,$$

where $\{A(t) : t \geq 0\}$ is a rate-λ Poisson arrival process independent of a sequence $\{V_k : k \geq 1\}$ of IID service times. The workload process is defined in terms of the net-input process $X(t) \equiv C(t) - t$ as described in Section 5.2. Any input that would take the workload above the storage capacity K overflows and is lost.

The $M/G/1/K$ model is appealing because $C \equiv \{C(t) : t \geq 0\}$ is a compound Poisson process, which is a special case of a renewal-reward process; see Section 7.4. Like the SLM, the processes C and X are Lévy processes, but unlike the SLM, the processes C and X almost surely have only finitely many jumps in a bounded interval.

The key to achieving the asymptotic parameters α and $\beta = 1$ is to use a heavy-tailed service-time distribution with power tail, where the ccdf decays as $x^{-\alpha}$ as $x \to \infty$. Thus, we let the service times be mV_k, where $\{V_k : k \geq 1\}$ is a sequence of IID nonnegative random variables with mean $EV_1 = 1$ and an appropriate power tail:

$$P(V_1 > x) \sim \gamma_1 x^{-\alpha} \quad \text{as} \quad x \to \infty \ .$$

Let $A \equiv \{A(t) : t \geq 0\}$ be a rate-1 Poisson process. Let the net-input process in the n^{th} $M/G/1/K$ model be

$$X_n(t) \equiv \sum_{k=1}^{A(\lambda_n t)} mV_k - t, \quad t \geq 0 \ ;$$

i.e., we let the service times be distributed as mV_1 and we let the arrival process be a Poisson process with rate λ_n. The associated *scaled net-input processes* are

$$\mathbf{X}_n(t) \equiv n^{-1/\alpha} X_n(nt), \quad t \geq 0 \ .$$

We will choose the parameters m and λ_n in order to obtain the desired convergence

$$\mathbf{X}_n \Rightarrow \sigma \mathbf{S} + \eta \mathbf{e} \ .$$

We must also make appropriate definitions for the queue. Since we scale space by dividing by $n^{1/\alpha}$ in the stochastic-process limit, we let the upper barrier in the n^{th} $M/G/1/K$ queue be $K_n = n^{1/\alpha} K$. We also must match the initial conditions appropriately. If the RSLM starts at the origin, then the $M/G/1/K$ queue starts empty.

To determine the parameters m and λ_n, observe that

$$\mathbf{X}_n = \mathbf{S}_n + \eta_n \mathbf{e} \ ,$$

where

$$\mathbf{S}_n(t) \equiv n^{-1/\alpha} (\sum_{k=1}^{A(\lambda_n nt)} mV_k - \lambda_n mnt), \quad t \geq 0 \ ,$$

and

$$\eta_n \equiv n^{1-\alpha^{-1}} (\lambda_n m - 1), \quad n \geq 1 \ .$$

Hence, we can obtain $\eta_n = \eta$ by letting

$$\lambda_n = m^{-1}(1 + \eta n^{-(1-\alpha^{-1})}), \quad n \geq 1 \ . \tag{6.1}$$

Since $\lambda_n \to m^{-1}$ as $n \to \infty$, $\mathbf{S}_n$ has the same limit as

$$\tilde{\mathbf{S}}_n(t) \equiv n^{-1/\alpha} (\sum_{k=1}^{A(m^{-1}nt)} mV_k - nt), \quad t \geq 0 \ .$$

Note that

$$P(mV_k > x) \sim \gamma_1 m^{\alpha} x^{-\alpha} \quad \text{as} \quad x \to \infty \ .$$

Hence, we can apply Theorem 7.4.2 for renewal-reward processes to deduce that

$$\tilde{\mathbf{S}}_n \Rightarrow \tilde{\sigma} \mathbf{S} \quad \text{in} \quad (D, J_1) \ ,$$

where

$$\tilde{\sigma}^{\alpha} = \gamma_1 m^{\alpha} m^{-1} / C_{\alpha}$$

for C_α in (5.14) of Section 4.5.1. To achieve our goal of $\tilde{\sigma} = \sigma$, we must have

$$m = (\sigma C_\alpha / \gamma_1)^{1/(\alpha-1)} . \tag{6.2}$$

With that choice of m, we have $\tilde{\mathbf{S}}_n \Rightarrow \sigma\mathbf{S}$. For λ_n in (6.1) and m in (6.2), we obtain $\mathbf{S}_n \Rightarrow \sigma\mathbf{S}$ and $\mathbf{X}_n \Rightarrow \sigma\mathbf{S} + \eta\mathbf{e}$ as desired.

For example, as the service-time distribution in the M/G/1/K fluid queue, we can use the Pareto distribution used previously in Chapter 1. Recall that a Pareto(p) random variable Z_p with $p > 1$ has ccdf

$$F_p^c(x) \equiv P(Z_p > x) \equiv x^{-p}, \quad x \ge 1 , \tag{6.3}$$

and mean

$$m_p = 1 + (p-1)^{-1} . \tag{6.4}$$

To achieve the specified power tail, we must let $p = \alpha$ for $1 < \alpha < 2$.

To put the Pareto distribution in the framework above, we need to rescale to obtain mean 1. Clearly, $m_\alpha^{-1} Z_\alpha$ has mean 1 and

$$P(m_\alpha^{-1} Z_\alpha > x) = P(Z_\alpha > m_\alpha x) = (m_\alpha x)^{-\alpha}, \quad x \ge 1 ,$$

so that we let $V_1 \stackrel{\mathrm{d}}{=} m_\alpha^{-1} Z_\alpha$ and have

$$\gamma_1 = m_\alpha^{-\alpha} = (1 + (\alpha-1)^{-1})^{-\alpha} .$$

Given those model specifications, we approximate the limiting RSLM, say $\mathbf{W}$, by the n^{th} scaled M/G/1/K workload process, i.e.,

$$\{\mathbf{W}(t) : t \ge 0\} \approx \{n^{-1/\alpha}(W_n(nt) : t \ge 0\} \tag{6.5}$$

for suitably large n, where $\{W_n(t) : t \ge 0\}$ is the workload (or virtual waiting time) process in the n^{th} M/G/1/K queueing system with upper barrier at $K_n = n^{1/\alpha} K$ specified above.

This approximation can be very useful because the M/G/1/K fluid-queue model or finite dam has been quite thoroughly studied and is known to be tractable; e.g., see Takács (1967), Cohen (1982) and Chapter 1 of Neuts (1989).

For the remaining discussion, consider an $M/G/1/K$ model with sevice times V_k and arrival rate λ. For $K = \infty$ and $\rho = \lambda E V_1 < 1$, the M/G/1 workload process is especially tractable; e.g., see Abate and Whitt (1994a). Then the steady-state distribution is characterized by the Pollaczek-Khintchine transform. For $K < \infty$, the steady-state distribution is the infinite-capacity steady-state truncated and renormalized, just as in Theorems 8.5.2 here and Theorem 5.2.1 in the Internet Supplement.

Given the steady-state workload $W(\infty)$ with $K < \infty$, the overflow rate is

$$\beta = \lambda E[W(\infty) + V_1 - K]^+ ,$$

where V_1 is a service time independent of $W(\infty)$. The rate of overflows of at least size x is

$$\beta_x = \lambda P(W(\infty) + V_1 - K > x) .$$

(For asymptotics, see Zwart (2000).) The ccdf $P(W(\infty) + V_1 > x + K)$ is easily computed by numerical transform inversion. For that purpose, we first remove the known atom (positive probability mass) at zero in the distribution of $W(\infty)$, as discussed in Abate and Whitt (1992a). Since $P(W(\infty) = 0) = 1 - \rho$, we can write

$$\begin{aligned} P(W(\infty) + V_1 > x + K) &= (1-\rho)P(V_1 > x + K) \\ &\quad + \rho P(W(\infty) + V_1 > x + K | W(\infty) > 0) . \end{aligned}$$

We then calculate the conditional ccdf $P(W(\infty)+V_1 > x+K|W(\infty) > 0)$ by numerically inverting its Laplace transform $(1-Ee^{-s(W\infty)|W(\infty)>0)}Ee^{-sV_1})/s$.

For any K with $0 < K \leq \infty$, the M/G/1/K fluid-queue departure process is an on-off cumulative-input process with mutually independent sequences of IID busy periods and IID idle periods, with the idle periods being exponentially distributed with mean λ^{-1}. The departure rate during busy periods is 1. For $K = \infty$, the busy-period distribution is characterized by its Laplace transform, which satisfies the Kendall functional equation; e.g., as in equation (28) of Abate and Whitt (1994a). The numerical values of the Laplace transform for complex arguments needed for numerical transform inversion can be computed by iterating the Kendall functional equation; see Abate and Whitt (1992b).

In the heavy-tailed case under consideration, $P(V_1 > x) \sim Ax^{-\alpha}$, where $A = C_\alpha \sigma^\alpha$ by virtue of (**??**) and (**??**). By de Meyer and Teugels (1980), the busy-period distribution in this M/G/1/∞ model inherits the power tail; i.e.,

$$P(B_1 > x) \sim A(1-\rho)^{-(\alpha+1)}x^{-\alpha} \quad \text{as} \quad x \to \infty .$$

Hence, we can apply Theorems 8.3.1 and 8.3.2 to establish a stochastic-process limit for the M/G/1 departure process with ρ assumed fixed. However, as noted at the end of Section 5.3.2, for fixed ρ with $\rho < 1$, the departure process obeys the same FCLT as the input process, given in Theorem 7.4.2. The two approaches to this FCLT can be seen to be consistent.

We have indicated that many M/G/1/K fluid-queue random quantities can be conveniently expressed in terms of the Laplace transform of the service-time distribution. Unfortunately, the Laplace transform of the Pareto distribution does not have a convenient simple explicit form, but because of its connection to the gamma integral, the transform values can easily be computed by exploiting efficient algorithms based on continued fractions, as shown by Abate and Whitt (1999a). That algorithm to compute Pareto distributions by numerical transform inversion is applied in Ward and Whitt (2000).

For some applications it might be desirable to have even more tractable "Markovian" service-time distributions. One approach is to approximate the Pareto distribution by a mixture of exponentials. A specific procedure is described in Feldmann and Whitt (1998). Since the Pareto distribution is completely monotone, it is directly a continuous mixture of exponentials. Thus, finite mixtures can provide excellent approximations for the Pareto distribution, but since it is impossible to match the entire tail, it is often necessary to use ten or more component exponentials to obtain a good fit for applications. Having only two component exponentials usually produces a poor match.

Alternatively, the M/G/1/K approximations can be produced with different heavy-tailed service-time distributions. Other heavy-tailed distributions that can be used instead of the scaled Pareto distribution or the approximating hyperexponential distribution are described in Abate, Choudhury and Whitt (1994) and Abate and Whitt (1996, 1999b, c).

We can also consider other approximating processes converging to the RSLM. Attractive alternatives to the M/G/1/K models just considered are discrete-time random walks on a discrete lattice. Such random-walk approximations are appealing because we can then employ well-known numerical methods for finite-state Markov chains, as in Kemeny and Snell (1960) and Stewart (1994). For the RSLM based on the totally skewed ($\beta = 1$) α-stable Lévy motion that arises as the stochastic-process limit in Theorem 8.5.1, it is natural to construct the random walk by appropriately modifying an initial random walk with IID steps distributed as the random variable Z with the zeta distribution, which has probability mass function

$$p(k) \equiv P(Z = k) \equiv 1/\zeta(\alpha + 1)k^{\alpha+1}. \quad k \geq 1 \ , \tag{6.6}$$

where $\zeta(s)$ is the Riemann zeta function; see p. 240 of Johnson and Kotz (1969) and Chapter 23 of Abramowitz and Stegun (1972). The zeta distribution has mean $\zeta(\alpha)/\zeta(\alpha+1)$ and the appropriate tail asymptotics, i.e.,

$$F^c(x) \equiv P(Z > x) \sim 1/\alpha\zeta(\alpha+1)x^{\alpha} \quad \text{as} \quad x \to \infty \; . \tag{6.7}$$

We can apply Section 4.5 to construct scaled random walks converging to the totally-skewed α-stable Lévy motion.

8.6.2. Limits for Limit Processes

We now consider the second approach for approximating the limit process $\mathbf{X}$ associated with a stochastic process-limit $\mathbf{X}_n \Rightarrow \mathbf{X}$. Now we assume that $\mathbf{X}_n$ is the scaled process

$$\mathbf{X}_n(t) \equiv n^{-H}(X_n(nt) - \nu_n nt), \quad t \geq 0 \; , \tag{6.8}$$

where $0 < H < 1$. If $\mathbf{X}$ is not sufficiently tractable, we may try to establish to establish a further stochastic-process limit for scaled versions of the limit process $\mathbf{X}$.

Suppose that we can construct the new scaled process

$$\tilde{\mathbf{X}}_n(t) \equiv n^{-\tilde{H}}(\mathbf{X}(nt) - \eta nt), \quad t \geq 0 \tag{6.9}$$

and show that $\tilde{\mathbf{X}}_n \Rightarrow \tilde{\mathbf{X}}$ as $n \to \infty$. Then we can approximate the first limit process $\mathbf{X}$ by

$$\{\mathbf{X}(t) : t \geq 0\} \approx \{\eta t + n^{\tilde{H}}\tilde{\mathbf{X}}(t/n) : t \geq 0\} \; . \tag{6.10}$$

Given the two stochastic-process limits, we can combine the two approximations with the scaling in (6.8) and (6.9) to obtain the overall approximation

$$\begin{aligned} \{X_n(t) : t \geq 0\} &\approx \{\nu_n t + n^H \mathbf{X}(t/n) : t \geq 0\} \\ &\approx \{\nu_n t + n^H \eta t/n + n^H m^{\tilde{H}} \tilde{\mathbf{X}}(t/nm) : t \geq 0\} \end{aligned} \tag{6.11}$$

for appropriate n and m. In the queueing context, we can fix n by letting n be such that the traffic intensity ρ_n coincides with the traffic intensity of the queue being approximated.

Since the limit processes for scaled overflow and departure processes in the fluid-queue model are somewhat complicated, even in the light-tailed weakly-dependent case considered in Section 5.7, it is natural to apply this two-limit approach to overflow and departure processes. We can apply Theorem 5.7.4 to obtain FCLT's for the boundary regulator processes $\mathbf{U}$ and $\mathbf{L}$ in the Brownian case. For that purpose, introduce the normalized processes

$$\begin{aligned} \tilde{\mathbf{U}}_n(t) &\equiv n^{-1/2}(\mathbf{U}(nt) - \beta nt), \\ \tilde{\mathbf{L}}_n(t) &\equiv n^{-1/2}(\mathbf{L}(nt) - \alpha nt), \quad t \geq 0 \,, \end{aligned} \tag{6.12}$$

where α and β are the rates determined in Theorem 5.7.3.

By using the regenerative structure associated with Theorem 5.7.4 (see Theorem 2.3.8 in the Internet Supplement), we obtain the following stochastic-process limit.

Theorem 8.6.1. (FCLT for RBM boundary regulator processes) *The normalized boundary regulation processes for RBM in* (6.12) *satisfy*

$$\begin{aligned} \tilde{\mathbf{U}}_n &\Rightarrow \sigma_U \mathbf{B}, \\ \tilde{\mathbf{L}}_n &\Rightarrow \sigma_L \mathbf{B} \quad in \quad D \,, \end{aligned} \tag{6.13}$$

where $\mathbf{B}$ *is standard BM and* σ_U^2 *and* σ_L^2 *are given in* (7.26), (7.27) *and* (7.30)–(7.32).

In general, a corresponding limit for the departure process is more complicated, but one follows directly from Theorems 8.6.1 and 5.9.1 when the processing is deterministic, so that the limit process $\mathbf{S}$ in (9.2) is the zero function.

Corollary 8.6.1. (second limit for the departure process) *Let* $\mathbf{D} = \mathbf{S} - \mathbf{L}$ *be the heavy-traffic limit for the departure process in Theorem* 5.9.1 *under the conditions of the general Brownian limit in Theorem* 5.7.1. *If* $\mathbf{S} = 0\mathbf{e}$, *then*

$$\tilde{\mathbf{D}}_n = -\tilde{\mathbf{L}}_n \Rightarrow \sigma_L \mathbf{B} \quad in \quad (D, M_1)$$

for $\tilde{\mathbf{L}}_n$ *in* (6.12), $\mathbf{B}$ *standard Brownian motion and* σ_L *as in* (7.26), (7.30) *and* (7.32).

The condition $\mathbf{S} = 0\mathbf{e}$ in Corollary 8.6.1 is satisfied under the common assumption of deterministic processing. We can apply Corollary 8.6.1 to obtain the approximation

$$\begin{aligned} \{D_n(t) : t \geq 0\} &\approx \mu_n t - n^{1/2}\mathbf{L}(t/n) \\ &\approx \mu_n t - n^{1/2}\alpha t/n - n^{1/2}m^{1/2}\sigma_L \mathbf{B}(t/nm) \,. \end{aligned} \tag{6.14}$$

where

$$\{m^{1/2}\mathbf{B}(t/m) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{\mathbf{B}(t) : t \geq 0\} \ .$$

As above, we may fix n by choing ρ_n to match the traffic intensity in the given queueing system.

8.7. Reflected Fractional Brownian Motion

In this section we discuss heavy-traffic stochastic-process limits in which the limit process for the scaled net-input process does *not* have independent increments. Specifically, we consider heavy-traffic limits in which the scaled workload processes converge to reflected fractional Brownian motion (RFBM). With time scaling, such limits arise because of strong dependence. As in the previous section, the RFBM limit occurs in a second stochastic-process limit.

8.7.1. An Increasing Number of Sources

We consider the same on-off model introduced in Section 8.2, but now we let the number of sources become large before we do the heavy-traffic scaling. (We also discuss limits in which the number of sources grows in Sections 8.3.3 and 9.8.) When we let the number of sources become large, we can apply the central limit theorem for processes in Section 7.2 to obtain convergence to a Gaussian process. Then, again following Taqqu, Willinger and Sherman (1997), we can let the busy-period and idle-period distributions have heavy tails, and scale time, to obtain a stochastic-process limit for the Gaussian process in which the limit process is FBM. The continuous-mapping approach with the reflection map then yields convergence of the scaled workload processes to RFBM. This heavy-traffic limit supplements and supports direct modeling using FBM, as in Norros (1994, 2000).

With the on-off model, heavy-tailed busy-period and idle-period distributions cause the scaling exponent H to exceed $1/2$. Specifically, as indicated in (2.20) in Section 7.2,

$$H = (3 - \alpha_{min})/2 \ , \tag{7.1}$$

where α_{min} is the minimum of the idle-period and busy-period ccdf decay rates if both have power tails with decay rates $\alpha_i < 2$. Otherwise, it is the decay rate of the one ccdf with a power tail if only one has a power tail. If both the busy-period and the idle-period have finite variance, then $\alpha_{min} = 2$ and $H = 1/2$. The following result is an immediate consequence of Theorem 7.2.5 and the continuous-mapping approach.

Theorem 8.7.1. (RFBM limit for the on-off model with many sources) *Consider a family of fluid-queue models indexed by the parameter pair (n, τ). Let all the models be based on a single on-off source model as specified before Theorem 7.2.5. In model (n, τ) there are n IID on-off sources and time scaling by τ. Let the processing in model (n, τ) be at a constant deterministic rate*

$$\mu_{n,\tau} \equiv \lambda n \tau + c n^{1/2} \tau^H \ , \tag{7.2}$$

where $\lambda \equiv m_1/(m_1 + m_2)$ is the single-source input rate. Let the capacity in model (n, τ) be $K_{n,\tau} \equiv n^{1/2}\tau^H K$. Let $W_{n,\tau}(0) = 0$. Then

$$(\mathbf{X}_{n,\tau}, \mathbf{W}_{n,\tau}) \Rightarrow (\sigma_{lim}\mathbf{Z}_H - c\mathbf{e}, \phi_K(\sigma_{lim}\mathbf{Z}_H - c\mathbf{e})) \tag{7.3}$$

in $(C, U)^2$ as first $n \to \infty$ and then $\tau \to \infty$, where H is in (7.1), σ^2_{lim} is in (2.17) or (2.18) in Chapter 4, $\mathbf{Z}_H$ is standard FBM, $\mathbf{X}_{n,\tau}$ is the scaled net-input process

$$\mathbf{X}_{n,\tau} \equiv \tau^{-H} n^{-1/2} (\sum_{i=1}^{n} C_i(\tau t) - \mu_{n,\tau} t), \quad t \geq 0 \ , \tag{7.4}$$

and $\mathbf{W}_{n,\tau}$ is the scaled workload process, i.e.,

$$\mathbf{W}_{n,\tau}(t) \equiv \phi_K(\mathbf{X}_{n,\tau}) \ . \tag{7.5}$$

Remark 8.7.1. *Double limits.* Theorems 8.5.1 and 8.7.1 establish heavy-traffic limits with *different* limit processes for the *same* fluid queue model with multiple on-off sources having heavy-tailed busy and idle periods. Theorem 8.5.1 establishes convergence to RSLM with a fixed number of sources, whereas Theorem 8.7.1 establishes convergence to RFBM when the number of sources is sent to infinity before the time (and space) scaling is performed. As indicated in Section 7.2.2, Mikosch et al. (2001) establish more general double limits that provide additional insight. Then RFBM (RSLM) is obtained if the number of sources increases sufficiently quickly (slowly) in the double limit. ■

8.7.2. Gaussian Input

When looking at traffic data from communication networks, we do not directly see the source busy and idle periods of the on-off model. Instead, we see an irregular stream of packets. Given such a packet stream, we must estimate source busy and idle periods if we are to use the on-off source traffic model. When we do fit busy and idle periods to traffic data, we may find that the busy and idle periods do not nearly have the independence assumed in Sections 5.7 and 8.5. Thus we might elect not to use the on-off model.

Instead, we might directly analyze the aggregate cumulative-input process. Measurements of the aggregate cumulative-input process $\{C(t) : t \geq 0\}$ may reveal strong positive dependence. From the data, we may fairly conclude that the variance of $C(t)$ is finite, but that it grows rapidly with t. In particular, we may see asymptotic growth of the variance as a function of t according to a power as

$$Var(C(t)) \sim t^{2H} \quad \text{as} \quad t \to \infty \tag{7.6}$$

for

$$1/2 < H < 1 \ , \tag{7.7}$$

which indicates strong positive dependence, as we saw in Section 4.6.

Given the asymptotic growth rate of the variance in (7.6), it is natural to look for a FCLT capturing strong positive dependence with light tails. From Section 4.6, it is natural to anticipate that properly scaled cumulative-input processes should converge to fractional Brownian motion (FBM). In particular, letting the input rate be 1 as before, it is natural to anticipate that

$$\mathbf{C}_n \Rightarrow \sigma \mathbf{Z}_H \quad \text{in} \quad (C, U) \quad \text{as} \quad n \to \infty \ , \tag{7.8}$$

where

$$\mathbf{C}_n(t) \equiv n^{-H}(C(nt) - nt), \quad t \geq 0 \ , \tag{7.9}$$

the process $\mathbf{Z}_H$ is standard FBM, as characterized by (6.13) or (6.14) in Section 4.6, and σ is a positive scaling constant. Since both $\mathbf{C}_n$ and $\mathbf{Z}_H$ have continuous sample paths, we can work in the function space C.

However, the principal theorems in Section 4.6, Theorems 4.6.1 and 4.6.2, do *not* directly imply the stochastic-process limit in (7.8), because there are extra structural conditions beyond the variance asymptotics in (7.6). In order to apply the theorems in Section 4.6, we need additional structure – either linear structure or Gaussian structure.

Fortunately, it is often reasonable to assume additional Gaussian structure in order to obtain convergence to FBM. With communication networks, it is often natural to regard the aggregate cumulative-input stochastic process as a Gaussian process, because the aggregate cumulative-input process is usually the superposition of a large number of component cumulative-input processes associated with different sources, which may be regarded as approximately independent. The intended activities of different users can usually be regarded as approximately independent. However, network controls in response to lost packets such as contained in TCP can induce dependence among sources. Hence approximate independence needs to be checked.

Not only are sources approximately independent, but the individual source input over bounded intervals is usually bounded because of a limited access rate. Thus we may apply Theorem 7.2.1 to justify approximating the aggregate cumulative-input process as a Gaussian process, just as we did for Theorem 8.7.1. Then there is a strong theoretical basis for approximating the workload process by reflected FBM without assuming that we have on-off sources.

We summarize by stating the theorem. Define additional random elements of C by

$$\begin{aligned} \mathbf{X}_n(t) &\equiv n^{-H}(C(nt) - \mu_n nt) = \mathbf{C}_n(t) + n^{1-H}(1-\mu_n)t \;, \\ \mathbf{W}_n(t) &\equiv \phi_K(\mathbf{X}_n)(t), \quad t \geq 0 \;. \end{aligned} \tag{7.10}$$

We apply Theorems 5.4.1 and 4.6.2 to obtain the following result. We apply Theorem 7.2.1 to justify the Gaussian assumption.

Theorem 8.7.2. (RFBM limit with strongly-dependent Gaussian input) *Consider a sequence of fluid queues indexed by n with capacities K_n, $0 < K_n \leq \infty$, and output rates μ_n, $n \geq 1$. Suppose that $\{C(t) - t : t \geq 0\}$ is a zero-mean Gaussian process with*

$$Var(C(t)) \sim \sigma t^{2H} \quad as \quad t \to \infty \tag{7.11}$$

and

$$Var(C(t)) \leq M t^{2H} \quad for\ all \quad t > 0 \tag{7.12}$$

for some positive constants H, σ and M with $1/2 \leq H < 1$. If, in addition, $K_n = n^H K$, $0 < K \leq \infty$, $0 \leq W_n(0) \leq K_n$ for all n,

$$n^{-H} W_n(0) \Rightarrow y \quad in \quad \mathbb{R} \quad as \quad n \to \infty \tag{7.13}$$

and

$$n^{1-H}(1-\mu_n) \to \eta \quad as \quad n \to \infty \tag{7.14}$$

for $-\infty < \eta < \infty$, then

$$(\mathbf{C}_n, \mathbf{X}_n, \mathbf{W}_n) \Rightarrow (\sigma \mathbf{Z}_H, y + \sigma \mathbf{Z}_H + \eta \mathbf{e}, \phi_K(y + \sigma \mathbf{Z}_H + \eta \mathbf{e})) \tag{7.15}$$

in $(C, U)^3$, where $\mathbf{Z}_H$ is standard FBM with parameter H and $(\mathbf{C}_n, \mathbf{X}_n, \mathbf{W}_n)$ is in (7.9) and (7.10).

Unfortunately, the limit process RFBM is relatively intractable; see Norros (2000) for a discussion. The asymptotic behavior of first passage times to high levels has been characterized by Zeevi and Glynn (2000). We discuss approximations for the steady-state distribution in the next section.

Of course, it may happen that traffic measurements indicate, not only that the aggregate cumulative-input process fails to have independent increments, but also that the aggregate cumulative-input process is not nearly Gaussian. Nevertheless, the FBM approximation might be reasonable after scaling. The FBM approximation would be supported by the theory in Section 4.6 if the cumulative-input process had the linear structure described in Section 4.6. With different structure, there might be a relevant stochastic-process limit to a different limit process. An invariance principle like that associated with Donsker's theorem evidently does not hold in the strongly-dependent case. Thus, careful analysis in specific settings may lead to different approximating processes.

Remark 8.7.2. *Bad news and good news.* From an applied perspective, the FBM heavy-traffic stochastic-process limit provides both bad news and good news. The main bad news is the greater congestion associated with greater space scaling as H increases. Part of the bad news also is the fact that the limit process is relatively difficult to analyze. Moreover, as discussed in Section 5.5, the greater time scaling as H increases means that significant relative changes in the process take longer to occur. That is part of the bad news if we are concerned about recovery from a large congestion event; see Duffield and Whitt (1997).

The good news is the greater possibility of *real-time prediction* of future behavior based on observations of the system up to the present time. As discussed in Section 4.6, the dependence of the increments in FBM provides a basis for exploiting the history, beyond the present state, to predict the future. For further discussion about predicting congestion in queues, see Duffield and Whitt (1997, 1998, 2000), Srikant and Whitt (2001), Ward and Whitt (2000) and Whitt (1999a,b). ∎

8.8. Reflected Gaussian Processes

In the last section we established convergence to reflected fractional Brownian motion (RFBM) for workload processes in fluid-queue models. In this section we consider approximations that can be obtained for the steady-state distributions of RFBM and more general stationary Gaussian processes when we have one-sided reflection. To do so, we exploit a lower bound due to Norros (1994), which has been found to be often an excellent approximation; e.g. see Addie and Zuckerman (1994) and Choe and Shroff (1998, 1999).

First let X be a general stochastic process with stationary increments defined on the entire real line $(-\infty, \infty)$ with $X(0) = 0$. Then the (one-sided) reflection of X is

$$\begin{aligned} \phi(X)(t) &\equiv X(t) - \inf_{0 \le s \le t} X(s) \\ &= \sup_{0 \le s \le t} \{X(t) - X(s)\} \stackrel{\mathrm{d}}{=} \sup_{0 \le s \le t} \{-X(-s)\} \, . \end{aligned} \tag{8.1}$$

Assuming that

$$\sup_{0 \le s \le t} \{-X(s)\} \to \sup_{s \ge 0} \{-X(-s)\} < \infty \quad \text{w.p.1} \quad \text{as} \quad t \to \infty \, , \tag{8.2}$$

$$\phi(X)(t) \Rightarrow \phi(X)(\infty) \stackrel{\mathrm{d}}{=} \sup_{s \ge 0} \{-X(-s)\} \quad \text{as} \quad t \to \infty \, . \tag{8.3}$$

A finite limit in (8.2) will hold when the process X has negative drift, i.e., when $EX(t) = -mt$ for some $m > 0$ and X is ergodic.

We propose approximating the steady-state tail probability $P(\phi(X)(\infty) > x)$ by a lower bound obtained by interchanging the probability and the supremum, i.e.,

$$\begin{aligned} P(\phi(X)(\infty) > x) &= P\left(\sup_{t \ge 0} \{-X(-t)\} > x\right) \\ &\ge \sup_{t \ge 0} P(-X(-t) > x) \, . \end{aligned} \tag{8.4}$$

Assuming that $X(t) \to -\infty$ as $t \to \infty$, we have $-X(-t) \to -\infty$ as $t \to \infty$. Hence $P(-X(-t) > x)$ will be small for both small t and large t, so it is natural to anticipate that there is an intermediate value yielding the maximum in (8.4). Moreover, large deviation arguments can be developed to show that the lower bound is asymptotically correct, in a logarithmic sense, as $x \to \infty$ under regularity conditions; see Duffield and O'Connell (1995), Botvich and Duffield (1995) Choe and Shroff (1998, 1999), Norros (2000) and Wischik (2001a). The moderate-deviations limit by Wischik (2001a) is especially insightful because it applies, not just to the Gaussian process, but also to superposition processes converging to Gaussing processes (in a CLT for processes).

In general, the lower bound may not get us very far, because it tends to be intractable. However, if we assume that X is also a Gaussian process, then we can conveniently evaluate the lower bound. For a Gaussian process X, we can calculate the lower bound in (8.4) Once we find the optimum t, t^*, the probability $P(-X(-t^*) > x)$ is Gaussian.

We can find t^* by transforming the variables to variables with zero means. In particular, let

$$Z(t) = \frac{-X(t) - E[-X(-t)]}{x - E[-X(-t)]}, \quad t \geq 0 , \tag{8.5}$$

and note that $Z(t)$ has mean 0 (assuming that $E[-X(-t)] \leq 0$) and

$$Z(t) \geq 1 \quad \text{if and only if} \quad -X(t) > x . \tag{8.6}$$

Hence, the optimum t^* for the lower bound in (8.4) is the t^* maximizing the variance of $Z(t)$, where

$$Var\, Z(t) = \frac{Var\, X(-t)}{(x - E[-X(-t)])^2} . \tag{8.7}$$

Given the mean and covariance function of the Gaussian process X, the variance $Var\, Z(t)$ in (8.7) is computable. The final approximation is thus

$$\begin{aligned} P(\phi(X)(\infty) > x) &\approx P(-X(-t^*) > x) \\ &= \Phi^c([x - E[-X(-t^*)]]/\sqrt{Var\, X(-t^*)}) \\ &= \Phi^c(1/\sqrt{Var\, Z(t^*)}) \end{aligned} \tag{8.8}$$

where $\Phi^c(x) \equiv 1 - \Phi(x)$ is the standard normal ccdf.

The approximation can be applied to any stationary Gaussian process with negative drift. From the last section, we are especially interested in the case in which X is FBM. So suppose that

$$X(t) \equiv \sigma \mathbf{Z}_H(t) + \eta t, \quad t \geq 0 ,$$

where $\eta < 0$ and $\mathbf{Z}_H$ is standard FBM. Since $Var(X(t)) = \sigma^2 t^{2H}$, the variance of $Z(t)$ in (8.5) is mazimized for

$$t^* = xH/|\eta|(1 - H).$$

Consequently, the desired lower bound is

$$P(\phi(X)(\infty) > x) \geq \Phi^c(\sigma^{-1}(|\eta|/H)^H (x/(1 - H))^{1-H}) , \tag{8.9}$$

where Φ^c is the standard normal ccdf. Using the approximation

$$\Phi^c(x) \approx e^{-x^2/2} ,$$

we obtain the approximation

$$P(\phi(X)(\infty) > x) \approx e^{-\gamma x^{2(1-H)}} , \tag{8.10}$$

where

$$\gamma \equiv \frac{1}{2\sigma^2}(\frac{|\eta|}{H})^{2H}(\frac{1}{1-H})^{2(1-H)} . \tag{8.11}$$

For $H > 1/2$, approximation (8.10) is a Weibull distribution with relatively heavy tail. Thus the lower-bound distribution has a Weibull tail. Note that for $H = 1/2$ approximation (8.10) agrees with the exact value for RBM in Theorem 5.7.2. Additional theoretical support for approximation (8.10) and asymptotic refinements are contained in Narayan (1998), Hüsler and Piterbarg (1999) and Massoulie and Simonian (1999); see Section 4.5 of Norros (2000). They show that there is an additional prefactor $Kx^{-\gamma}$ in (8.10) as $x \to \infty$ for $\gamma = (1-H)(2H-1)/H$. The exponent in approximation (8.10) was shown to be asymptotically correct as $x \to \infty$ by Duffield and O'Connell in their large-deviations limit.