Periodic Load Balancing

Gisli Hjálmtýsson [a] Ward Whitt [a]

[a] *AT&T Labs, 180 Park Avenue, Building 103, Florham Park, NJ 07932-0971*
E-mail: {gisli,wow}@research.att.com

Queueing Systems 30 (1998) 203–250

Multiprocessor load balancing aims to improve performance by moving jobs from highly loaded processors to more lightly loaded processors. Some schemes allow only migration of new jobs upon arrival, while other schemes allow migration of jobs in progress. A difficulty with all these schemes, however, is that they require continuously maintaining detailed state information. In this paper we consider the alternative of periodic load balancing, in which the loads are balanced only at each T time units for some appropriate T. With periodic load balancing, state information is only needed at the balancing times. Moreover, it is often possible to use slightly stale information collected during the interval between balancing times.

In this paper we study the performance of periodic load balancing. We consider multiple queues in parallel with unlimited waiting space to which jobs come either in separate independent streams or by assignment (either random or cyclic) from a single stream. Resource sharing is achieved by periodically redistributing the jobs or the work in the system among the queues. The performance of these systems of queues coupled by periodic load balancing depends on the transient behavior of a single queue. We focus on useful approximations obtained by considering a large number of homogeneous queues and a heavy load. When the number of queues is sufficiently large, the number of jobs or quantity of work at each queue immediately after redistribution tends to evolve deterministically, by the law of large numbers. The steady-state (limiting) value of this deterministic sequence is obtained as the solution of a fixed point equation, where the initial value is equal to the expected transient value over the interval between successive redistributions conditional on the initial value. A refined approximation based on the central limit theorem is a normal distribution, where the mean and variance are obtained by solving a pair of fixed-point equations. With higher loads, which is natural to consider when load balancing is performed, a heavy-traffic limit theorem shows that one-dimensional reflected Brownian motion can be used to approximately describe system performance, even with general arrival and service processes. With these approximations, we show how performance depends on the assumed arrival pattern of jobs and the model parameters. We do numerical calculations and conduct simulation experiments to show the accuracy of the approximations.

Keywords: load balancing, resource sharing, periodic load balancing, heavy traffic diffusion approximations, reflected Brownian motion, transient behavior

1. Introduction

There is now a substantial literature on dynamic multiprocessor load balancing; e.g., see Eager, Lazowska and Zahorjan [14], Hajek [20], Harchol-Balter and Downey [21], Leland and Ott [28], Willebeck-LeMair and Reeves [49], Zhou [53] and references therein. The basic scheme is to move jobs from a highly loaded originating processor to another more lightly loaded processor. There can be significant overhead associated with this load balancing, but it is nevertheless often worthwhile. There is a tradition in multiprocessor load balancing of only moving entire jobs at the time they originate, but migration of jobs in process is now beginning to be used as well, e.g., see Barak, Shai and Wheeler [8]. There is typically substantially more overhead with migration of jobs in process, but it has been shown to yield significant performance improvement by Harchol-Balter and Downey [21].

A difficulty with any form of dynamic load balancing, however, is that it involves real-time control, requiring continuous maintenance of state information. It is thus natural to consider whether it is possible to achieve much of the load balancing benefit with less work. Hence, in this paper we study the alternative of periodic load balancing. With periodic load balancing, no elaborate control is done for each arriving job or at each time. Instead, the loads are balanced only periodically, at each T units of time for some appropriate T.

Another motivation for the present paper is to lend support for a notion of lightweight call setup, supporting connection and connectionless services in communication networks; see Hjálmtýsson [22] and Hjálmtýsson and Ramakrishnan [23]. The main idea is to quickly provide service to new connections at a low or moderate quality and, over time, gradually meet higher quality-of-service requirements as requested. In that context, the periodic load balancing considered here is an abstraction of slower-time-scale reconfiguring that might be done in the network instead of quality-of-service routing immediately upon arrival.

In this paper we study the performance of periodic load balancing. Specifically, we consider m queues in parallel with unlimited waiting space. Every T time units, we redistribute the jobs or the remaining work in the system among the queues to balance the loads. Like other forms of load balancing, periodic load balancing corrects for systematic differences in the loads; e.g., when the arrival rates or service requirements at some queues are greater than at other queues. Load balancing also can significantly improve performance in a system with homogeneous queues. Then the load balancing compensates for stochastic fluctuations which make the loads at some queues temporarily greater than the loads at other queues. Here we primarily consider the benefits of periodic load balancing with homogeneous queues, but we also consider the case in which a proportion of the queues are temporarily down (arrivals come but no service is provided); see Section 10. Consistent with intuition, we show that load balancing is even more important in unbalanced scenarios.

We consider two different redistribution schemes. In the first scheme, every T time units the jobs in the system are redistributed among the queues, so that after each redistribution the numbers of jobs in any two queues differ by at most one. We do not focus on alternative ways to assign the jobs to the queues. In our simulations we assign jobs to the queues in a round robin fashion in order of arrival times, with the older job getting assigned first. When we redistribute jobs, we assume that the service discipline for each separate queue is first-come first-served (FCFS), but our results for the FCFS discipline may also serve as useful approximations for other disciplines such as round robin (RR) or processor sharing (PS).

In the second redistribution scheme, every T time units we redistribute the remaining work (in service time) evenly among the queues. When we redistribute the work, we assume that we know the remaining service requirements of all jobs in the system and that the remaining work of each job can be divided up and assigned to different queues. When we redistribute work, we assume that there is a general work-conserving discipline. (There is never an idle server at a queue when there is work to be done there.) There are many work-conserving disciplines; examples are FCFS, RR and PS.

Dividing jobs into pieces is currently not possible in a non-parallel environment, but may become so. As is, our analysis of redistribution of work describes a lower bound which available alternatives can try to achieve. We believe that alternative policies can be developed without job splitting and without knowledge of remaining service requirements that nearly achieve this lower bound. The main idea is to focus on work, since the work associated with different jobs may be very different. The alternative policies can be based on estimates of the remaining service requirements given available information, including elapsed service times. (We intend to discuss such alternative periodic load balancing schemes in a subsequent paper.)

Our main contributions are analytical models and formulas describing the performance of periodic load balancing. Our goal is to describe the distribution of the workload at each queue as a function of time, especially just before and just after each reconfiguration (balancing). During the interval between reconfigurations, the degree of inbalance and the likelihood of a larger workload at any one queue tends to increase with time. The workloads after reconfiguring equal, at least approximately, the average of the workloads before reconfiguring. From the workload distributions just before and after balancing, we can determine the distribution of the number of jobs and the amount of work that must be moved and, thus, the overhead associated with periodic load balancing.

We also describe how the performance of periodic load balancing depends upon the balancing interval T, the number of queues m and the other model parameters. We show how the performance depends on the arrival pattern. We consider three possible arrival patterns: Each queue may have its own arrival process or all arrivals may come in a single arrival process, after which they are assigned to the queues either at random or deterministically (in a cyclic or round robin order).

We obtain relatively tractable explicit formulas by considering the limiting case in which the number of queues, m, and the traffic intensity (or server utilization), ρ, are both large, i.e., as $m \to \infty$ and $\rho \to 1$, where $\rho = 1$ is the critical value for stability. The case of large m is currently of great interest, e.g., for understanding large computers constructed from many smaller computers. Moreover, the limit as $m \to \infty$ may serve as a useful approximation when m is not too large, e.g., when $m = 10$. When there are many servers, higher utilizations tend to be more feasible. We consider the limit as $\rho \to 1$ to generate approximations for typical (not small) utilizations.

In addition to the literature on dynamic multiprocessor load balancing, our work is also related to the literature on resource sharing within general queueing theory. In many situations multiple jobs must be processed on multiple resources. It is known that greater efficiency usually (but not always) can be achieved if the resources can be shared or pooled; e.g., see Smith and Whitt [36], Rothkopf and Rech [33], Laws [27], Whitt [47] and Mandelbaum and Reiman [29]. For example, consider two separate finite-server queues with infinite-waiting room, the FCFS discipline, all service times i.i.d. and general stationary arrival processes that are independent of the service times. Then the number of customers in the system at any time is stochastically smaller if the two systems are combined into one, having the aggregate superposition arrival process, the combined number of servers and the FCFS discipline; see Theorem 6 of Smith and Whitt [36], which draws on Wolff [51]. (As noted in [36], this result depends critically on the service-time distributions being identical, or at least not too different.)

Quantitatively, the (great) advantage of multi-server systems over a collection of separate single-server systems with common total load is well described by approximation formulas for basic performance measures. For example, the simple heavy-traffic approximation (limit after normalization) for the steady-state distribution of the waiting time before beginning service in a GI/GI/s queue (in which interarrival times and service times each come from i.i.d. sequences) is an exponential distribution with mean

$$EW \approx \frac{\rho}{s(1-\rho)} \frac{(c_1^2 + c_s^2)}{2} \ , \tag{1.1}$$

where the mean service time is taken to be 1, the traffic intensity (utilization of each server) is ρ and the squared coefficient of variations (SCV, variance divided by the square of the mean) of the interarrival and service times are c_a^2 and c_s^2, respectively; e.g., see (2.13) of Whitt [48]. (For supporting theory, see Iglehart and Whitt [24] and Köllerström [26].) Formula (1.1) shows that the mean EW is inversely proportional to s for fixed ρ.

A more refined approximation for the mean characterized by the parameter quadruple (s, ρ, c_a^2, c_s^2) is

$$EW(s, \rho, c_a^2, c_s^2) = \frac{(c_a^2 + c_s^2)}{2} EW(M/M/s) \tag{1.2}$$

where $EW(M/M/s)$ is the mean in the associated M/M/s model (exponential interarrival and service times with the same means), which can easily be calculated numerically, and can be further approximated by the Sakasegawa [34] approximation

$$EW(M/M/s) = \frac{\rho^{(\sqrt{2(s+1)}-1)}}{s(1-\rho)} \; ; \tag{1.3}$$

see (2.12) and (2.14) of [48]. From (1.2) and (1.3), we see that the heavy-traffic formula (1.1) actually underestimates the advantage of sharing. Numerical examples in [48] show that these formulas accurately describe the way the mean waiting time depends on s and the other parameters.

The expected number of jobs in the system, say EN, is the expected number of jobs in service, $s\rho$, plus the expected number of jobs in queue, $\lambda EW = s\rho EW$ (both by Little's law), so that the expected number of jobs in the system per server is $\rho(1 + EW)$. The EW component exhibits the strong dependence on s shown above.

The advantage of multi-server systems over separate single-server systems is also seen in other performance measures. For example, the probability of experiencing delay before beginning service remains approximately constant as the number of servers increases if the traffic intensity increases as well with $(1 - \rho)\sqrt{s}$ held fixed; see Whitt [47]. In other words, the utilization as a function of s is approximately

$$\rho \approx 1 - \gamma/\sqrt{s} \tag{1.4}$$

for some constant γ, if we also hold the probability of delay fixed. Formula (1.4) illustrates that the greater efficiency with multiple servers can be realized by higher utilization for a given level of congestion instead of less congestion. Alternatively, resource sharing can yield a combination of higher utilization and reduced delays.

Unfortunately, however, it is not always possible to fully share resources. In this paper we consider partial-sharing schemes that yield performance in between the single-server and multi-server cases in formulas (1.1)–(1.4). One way to partially share resources when the queues are separate is to assign new jobs upon arrival to the more lightly loaded queues. When the service-time distribution is exponential or has increasing failure rate, if jobs must be assigned to queues upon arrival without further intervention, then it is optimal to assign the job to the shortest queue; see Winston [50] and Weber [41]. However, somewhat surprisingly, for other service-time distributions, the shortest queue (SQ) rule need not be optimal; see Whitt [45]. More generally, it is natural to assign each job to the queue that will minimize its expected delay, although this rule is not always optimal either [45]. The advantage of the SQ rule is illustrated by the heavy-traffic limit, which shows that SQ behaves as well as the combined system as $\rho \to 1$; see Foschini and Salz [19], Reiman [32] and Zhang, Hsu and Wang [52].

Instead of assigning jobs upon arrival, here we consider the alternative of periodically redistributing the jobs to balance the queue lengths (number in system); i.e., so that they differ by at most 1 after redistribution. Periodic redistribution has two potential advantages over dynamic assignment of arrivals. First, the periodic redistribution gives an alternative way to balance the loads, which may be more robust. Even with the SQ rule, after a rare period of high congestion (with very large queue lengths), a few queues may remain very long after most queues have emptied (because of especially long service times, e.g., when the servers at one queue are temporarily unavailable). Then load balancing only through routing of new arrivals may be less effective than periodically redistributing jobs. Second, with periodic redistribution, we need not perform any control upon arrival. Dynamic assignment of arrivals may be very costly, because we need to constantly maintain system state. In contrast, with periodic load balancing, system state information is only needed at redistribution times. Moreover, the most current state is often not actually needed. Under relatively heavy loads, it is possible to determine the appropriate redistribution during a short interval before the actual redistribution time. Under heavy loads, when processing system state becomes difficult, the queue lengths tend to change relatively slowly (the snapshot principle; see Reiman [32]), so that little is lost if the system state is somewhat stale. As shown by Foschini [18], even under heavy loads, the system state can be communicated without significantly further increasing the loads.

Even less state information is required if redistribution is done with a large number of queues. Then the required number at each queue can be closely estimated without actually looking at the queue lengths, provided one knows the queueing model reasonably accurately. Even if the queueing model is not known, the average number *after the last redistribution* usually will be a good estimate for the number that should be present after the next redistribution, because these averages tend to evolve deterministically when there are many queues. Given that the target level is known in advance, local adjustments can be made among the queues in a distributed manner.

Here is how the rest of this paper is organized. In Section 2 we define the basic stochastic processes and characterize the steady-state number of jobs after redistributing jobs when each queue is an M/M/s queue (has s servers, Poisson arrivals and exponential service times). We also make stochastic comparisons showing that the performance with periodic load balancing falls in between the single s-server queue and the combined ms-server queue (showing that formulas (1.1)–(1.4) provide upper and lower bounds).

In Section 3 we consider the limiting behavior as $m \to \infty$ in the M/M/s setting, and show that, asymptotically, the total number of jobs evolves deterministically, having a limit characterized by a fixed-point equation. We also show that the fixed-point equation has a unique solution for each redistribution interval, and that the fixed-point level is a continuous strictly increasing function of the interval length. We show how to calculate the fixed-point level and the queue performance between redistributions by exploiting previous transient results for M/M/s and M/M/1 queues. We also show that for suitably large m the number of jobs per queue immediately after load balancing tends to be normally distributed. The approximate steady-state normal distribution can be obtained by solving a pair of equations for the mean and variance.

In Section 4 we establish a heavy-traffic diffusion approximation for the case of general arrival and service processes (satisfying central limit theorems, e.g., i.i.d. sequences with finite second moments), which yields reflected Brownian motion (RBM) as the model for the single-queue evolution between redistributions. The heavy-traffic limit shows how the interval between redistributions and the level after redistributions should scale with increasing load. Indeed by appropriate scaling, all cases are reduced to the single case of canonical RBM (with drift -1 and diffusion coefficient 1). (See Abate and Whitt [1] and Whitt [46] for further discussion.) We also show how approximate system performance can be described explicitly. Heavy-traffic limits seem very appropriate in this setting, because when we couple m s-server queues, they can usually operate at a higher server utilization; e.g., recall formula (1.4).

In Section 5 we apply the new asymptotic results and previous ones to compare the performance of load balancing to the performance of the two basic alternatives: (1) m separate single-server queues and (2) one combined m-server queue. In Section 6 we make comparisons between the RBM approximation and simulations of M/G/1 queues coupled by periodic load balancing. We consider exponential and Pareto service-time distributions (with finite variance).

In Section 7 we consider the redistribution of remaining work in single-server queues using a work-conserving discipline. The resulting RBM heavy-traffic approximation is the same as in Section 4. In Section 8 we show that the periodic load balancing significantly reduces the likelihood of severe congestion by showing that the tail probabilities with periodic load balancing decay much more rapidly than they do without load balancing.

In Section 9 we discuss the case of long-tail service-time distributions, which Leland and Ott [28] and Harchol and Downey [21] have shown to be present in computer systems. We show that a heavy-traffic limit involving an extra jump process can be used to approximate system behavior when the service-time distributions fail to have finite second moments or even first moments. We provide both transient and steady-state descriptions. With such high variability, transient descriptions tend to be more useful.

In Section 10 we consider the situation in which a proportion of the queues are down, so that at these queues in the interval between redistributions jobs arrive but no service is performed. The primary purpose of this section is to show how periodic load balancing performs in unbalanced scenarios. It should be clear that load balancing is even more important when the queues are not homogeneous. In an unbalanced environment, queues will often be unstable, i.e., the processes will grow without bound, when no form of load balancing is performed. The analysis in Section 10 can also be used to describe the effect of long-tail service-time distributions. The down times can represent exceptionally long service times. When we focus on the jobs in each queue, the few jobs with exceptionally long service time are themselves asymptotically negligible as $\rho \to 1$, but their impact on the processing of other jobs can be great.

2. Redistributing Jobs in the Markov Case

We start by considering the Markov special case, in which the service times are exponential and the arrival processes to the queues are i.i.d. Poisson processes. Let the mean service time be 1 and let the arrival rate at each queue be λ. We consider s-server queues, but we are primarily interested in the case of relatively small s, e.g., $s = 1$. With m s-server queues, the overall traffic intensity (or server utilization) is

$$\rho = \frac{\lambda m}{sm} = \frac{\lambda}{s} \ . \tag{2.1}$$

In this context we redistribute the jobs in the system every T time units, so that the numbers of jobs at any two queues differ by at most 1 after each redistribution. In between redistributions, the queues evolve independently (conditional on the initial values after the last redistribution). Let N_{in} be the number of jobs at the i^{th} queue (waiting and in service) after the n^{th} redistribution at time nT. Let N_n be the total number of jobs at all m queues after the n^{th} redistribution. Without loss of generality, let the elements of the vector $(N_{in}, 1 \leq i \leq m)$ be ordered so that they are nondecreasing in i. Let ϕ_m be the function that maps N_n into $(N_{1n}, \ldots, N_{mn})$ with this ordering; e.g., $\phi_5(7) = (1,1,1,2,2)$. With the ordering imposed upon the vectors $(N_{1n}, \ldots, N_{mn})$, there is a one-to-one correspondence between the processes $\{N_n : n \geq 1\}$ and $\{(N_{1n}, \ldots, N_{mn}) : n \geq 1\}$; i.e.,

$$N_n = N_{1n} + \ldots + N_{mn} \tag{2.2}$$

and

$$(N_{1n}, \ldots, N_{mn}) = \phi_m(N_n) \ . \tag{2.3}$$

We now characterize the stochastic process $\{N_n : n \geq 1\}$. Let $\stackrel{d}{=}$ denote equality in distribution. Recall that a Markov chain is *stochastically monotone* if the conditional distribution of N_{n+1} given $N_n = i$ is stochastically increasing as i increases, i.e., if

$$E[g(N_{n+1})|N_n = i] \le E[g(N_{n+1})|N_n = j] \quad \text{whenever } i \le j \tag{2.4}$$

for all nondecreasing real-valued functions g for which the expectations are well defined; e.g., see Stoyan [37] or Baccelli and Brémaud [7].

Theorem 1. For the Markov special case, the stochastic process $\{N_n : n \ge 1\}$ is a stochastically monotone, irreducible, aperiodic Markov chain with transition probability

$$\begin{aligned} P_{jk} &\equiv P(N_{n+1} = k|N_n = j) \\ &= P(\sum_{i=1}^{m} Q_i(T) = k|Q_1(0) = \ell_1, \dots, Q_m(0) = \ell_m) \end{aligned} \tag{2.5}$$

where $(\ell_1, \dots, \ell_m) = \phi_m(j)$ and $\{(Q_i(t)|Q_i(0) = \ell_i) : t \ge 0\}$, $1 \le i \le m$, are independent M/M/s queue-length (number in system) stochastic processes. If $\rho < 1$, then this Markov chain is positive recurrent with stationary random element N_∞ characterized by the equation

$$(N_{1\infty}, \dots, N_{m\infty}) \stackrel{d}{=} \phi_m\left(\sum_{i=1}^{m} (Q_i(T)|Q_i(0) = N_{i\infty})\right) , \tag{2.6}$$

where $(N_{1\infty}, \dots, N_{m\infty}) = \phi_m(N_\infty)$ and $\{(Q_1(t), \dots, Q_m(t)) : t > 0\}$ is independent of $(Q_1(0), \dots, Q_m(0) = (N_{1\infty}, \dots, N_{m\infty})$ on the right in (2.6).

Proof. The Markov property for $\{N_n : n \ge 1\}$ follows immediately from the lack of memory property associated with the exponential interarrival and service times. Since it is possible to go from 0 to 0 in one step, the chain is aperiodic. Since it is possible to get from any state to any other, the chain is irreducible. The stochastic monotonicity follows from comparison results for the M/M/s queue; e.g., see Whitt [44] or Baccelli and Bremaud [7]. If $Q_i(0)$ increases, then the distribution of $Q_i(T)$ (and the entire sample path $Q_i(t), t \ge 0$) increases stochastically. The positive recurrence follows from the mean drift criterion: As x increases,

$$E[N_{n+1} - N_n|N_n = x] \ \to \ smT(\rho - 1) < 0 \tag{2.7}$$

while

$$E[N_{n+1} - N_n|N_n = x] \le \lambda mT \quad \text{for all } x \ ; \tag{2.8}$$

see p. 262 of Meyn and Tweedie [31]. Finally, the steady-state equation (2.6) corresponds to the usual steady-state equation $\pi = \pi P$. $\square$

The overhead associated with the load balancing can also be described. The number of jobs that must be moved from the i^{th} queue in steady state, say $J_{i\infty}$, is

$$J_{i\infty} = \left[(Q_i(T)|Q_i(0) = N_{i\infty}) - \frac{1}{m} \sum_{j=1}^{m} (Q_j(T)|Q_j(0) = N_{j\infty}) \right]^+ , \tag{2.9}$$

where $[x]^+ = \max\{x_1 0\}$. The number of jobs moved in is described similarly. The limiting case as $m \to \infty$ considered in the next section provides a convenient simple approximation.

Because of the stochastic monotonicity property, we can deduce that the random sequence $\{N_n : n \geq 1\}$ increases stochastically as n increases when $N_0 = 0$. Recall that one random variable X_1 is *stochastically less than or equal to* another X_2, denoted by $X_1 \leq_{st} X_2$, if $Eg(X_1) \leq Eg(X_2)$ for all nondecreasing real-valued functions g for which the expectations are well defined.

Corollary 2. If $N_0 = 0$ in the setting of Theorem 1, then for all $n \geq 0$

$$N_n \leq_{st} N_{n+1} \leq_{st} N_\infty \ .$$

It is intuitively clear that periodic load balancing helps, i.e., that the steady-state distribution is in some sense smaller with load balancing than for separate M/M/s queues without load balancing. On the other hand, periodic load balancing should not be as good as one combined queue with sm servers. We now establish supporting stochastic comparisons. Let $N(t)$ be the number of jobs at all queues in the m-queue load balancing model at time t.

Theorem 3. With m M/M/s queues,

$$(N(t)|N(0) = n) \geq_{st} (Q(t)|Q(0) = n) \quad for\ all\ t\ , \tag{2.10}$$

where $\{Q(t) : t \geq 0\}$ is the queue length process in a single combined M/M/ms system with arrival rate λm.

Proof. As in Whitt [44], we can artificially construct the two processes on the same probability space so that the sample paths are ordered with probability 1. Let the two processes have the same arrival process. Since $N(0) = Q(0)$, the departure rate in Q is initially no smaller. At each transition, we maintain $N(t) \geq Q(t)$ by having departures in Q whenever $N(t) = Q(t)$ and there is a departure in N. Whenever $N(t) = Q(t)$, the departure rates are ordered. There may be strict inequality because some servers are idle with N but not Q. This special construction implies stochastic order as expressed in (2.10). □

The stochastic process $\{N(t) : t \geq 0\}$ has a periodic structure. The variables $N(kT + t)$ for $0 \leq t < T$ converge in distribution as $k \to \infty$ to limits $N_t(\infty)$ by virtue of Theorem 1. We can apply Theorem 3 to obtain a stochastic comparison for these periodic steady-state limits.

Corollary 4. In the M/M/s setting, the periodic steady-state variables are ordered by

$$N_t(\infty) \geq_{st} Q(\infty) \ , \quad 0 \leq t < T \ ,$$

where $Q(\infty)$ has the steady-state distribution of the combined M/M/sm system with arrival rate λm.

Since the infinite-server M/M/∞ system is a lower bound to the M/M/s system, we can obtain a further lower bound, which should be more useful when s and λ are large.

Corollary 5. In the M/M/s setting,

$$P(N_t(\infty) \geq k) \geq \sum_{j=k}^{\infty} \frac{e^{-\lambda m}(\lambda m)^j}{j!} \ , \quad k \geq 1 \ , \tag{2.11}$$

so that

$$EN_t(\infty) \geq \lambda m \ . \tag{2.12}$$

Proof. Recall that the steady-state distribution of $Q(t)$ in the M/M/∞ model is Poisson with mean equal to the offered load, which here is λm. □

We now show that load balancing helps by making stochastic comparisons with m separate queues. In this case we establish results only for the case $s = 1$. Recall that one random variable X_1 is less than or equal to another X_2 in the *(increasing) convex stochastic order*, denoted by $X_1 \leq_c X_2$ ($X_1 \leq_{ic} X_2$), if $Eg(X_1) \leq Eg(X_2)$ for all (increasing) convex real-valued functions g for which the expectations are well defined; e.g., see Stoyan [37] and Baccelli and Bremaud [7].

We use the following result for the transient M/M/1 queue, which is analogous to part of Theorem 5.2.1 of Stoyan [37].

Theorem 6. Consider two M/M/1 queue length processes $\{Q_i(t) : t \geq 0\}$ differing only in their initial values. If $Q_1(0) \leq_{ic} Q_2(0)$, then $Q_1(t) \leq_{ic} Q_2(t)$ for all $t \geq 0$.

Proof. Note that

$$Q(t) = \max\{Q(0) + X(t) \ , \quad X(t) - \inf_{0 \leq s \leq t} X(s)\} \ ,$$

where $X(t) = A(t) - S(t)$, $t \geq 0$, with $\{A(t) : t \geq 0\}$ and $\{S(t) : t \geq 0\}$ being independent Poisson processes, so that $Q(t)$ is an increasing convex function of $Q(0)$. Hence $f(Q(t))$ is an increasing convex function of $Q(0)$ for each increasing convex f. □

Theorem 7. Let N(t) be the total number of jobs at time t in the m M/M/1 queues coupled by periodic load balancing. Then

$$(N(t)|N(0) = \sum_{i=1}^{n} k_i) \leq_{ic} \sum_{i=1}^{m} (Q_i(t)|Q_i(0) = k_i) \quad \textit{for all } t$$

and initial vectors $(k_1, \ldots, k_m)$, where $\{(Q_i(t)|Q_i(0) = k_i) : t \geq 0\}$, $1 \leq i \leq m$, are independent M/M/1 queue length processes.

Proof. Each load balancing makes the vector of m queue lengths, after permuting the queues randomly, smaller (no larger) in the convex stochastic order ($\leq_c$). This initial convex order implies that increasing stochastic order ($\leq_{ic}$) is maintained throughout the interval between redistributions by Theorem 6. Thus the result follows by induction on the redistribution times. □

Corollary 8. Let $N_t(\infty)$ have the periodic steady-state distribution of $N(t)$ in the setting of Theorem 7. Then

$$N_t(\infty) \leq_{ic} \sum_{i=1}^{m} Q_i(\infty) \ ,$$

where $Q_i(\infty)$, $1 \leq i \leq m$, are i.i.d. with $P(Q_i(\infty) = k) = (1 - \rho)\rho^k$, $k \geq 0$.

Proof. Increasing convex order is inherited by the limits with convergence in distribution. □

We close this section by pointing out that an interesting open problem is to describe customer waiting times. Within one cycle, the waiting time of a new arrival is just the random sum of exponential service times, where the random number is the number of customers in the queue. Thus the waiting-time distribution will depend on the arrival time within a cycle. When we consider times extending beyond one cycle, we must properly take account of the way jobs are assigned to queues at redistribution points, which introduces considerable potential complexity. In the heavy-traffic limit in Seciton 5, we will observe that jobs tend to get served in the same cycle in which they arrive, so that this extra complexity does not arise.

3. Many Queues and Exponential Service Times

The system behavior simplifies when there are many queues. First, suppose that there is random assignment to the queues from a single general stationary arrival process, where by random assignment we mean that each queue is selected with equal probability and that successive assignments are mutually independent and independent of the service times. Then, as $m \to \infty$, the arrival processes to the queues approach independent Poisson processes; e.g., see Çinlar [12], Serfozo [35] and references therein. Hence, we have additional justification for considering independent Poisson arrival processes.

Second, with exponential service times, as the number m of queues gets large, the redistribution tends to put the constant, expected value at each queue. In the limit $m \to \infty$, the only deviation from this expected value is due to the requirement that the initial number of jobs at each queue must be an integer. Hence, a proportion of the queues will have n jobs, while the remainder of the queues will have $n+1$ for some deterministic n. We now state this property as a theorem. Let $N_n^{(m)}$ denote the random total number of jobs just after the n^{th} redistribution in the model with m queues. For $x \geq 0$, let $\lfloor x \rfloor$ be the integer part of x.

Theorem 9. Consider the m-queue model with periodic load balancing at times nT for $n \geq 1$. Let the queues be M/M/s queues. If $N_0^{(m)}/m \to x_0$ w.p.1 as $m \to \infty$, then

$$\frac{N_n^{(m)}}{m} \quad \to \quad x_n \quad w.p.1 \quad as \quad m \to \infty \quad \text{for each } n \geq 0 \; , \tag{3.1}$$

where $\{x_n : n \geq 1\}$ is a deterministic sequence evolving as

$$x_{n+1} = f_T(x_n) \; , \quad n \geq 0 \; , \tag{3.2}$$

for a function f_T independent of n with $f_T(x_n) = M(T, x_n)$ and

$$\begin{aligned} M(t,x) = {} & (x - \lfloor x \rfloor) E[Q(t) | Q(0) = \lfloor x \rfloor + 1] \\ & + (1 - x + \lfloor x \rfloor) E[Q(t) | Q(0) = \lfloor x \rfloor] \; , \end{aligned} \tag{3.3}$$

where $\{Q(t) : t \geq 0\}$ is an M/M/s queue-length process.

Proof. Note that

$$\begin{aligned} & (m^{-1} N_{n+1}^{(m)} | N_n^{(m)} = mj + k) \;\to \\ & \quad pE[Q(T)|Q(0) = j+1] + (1-p) E[Q(T)|Q(0) = j] \quad \text{w.p.1} \end{aligned}$$

as $m \to \infty$ with $k/m \to p$, by virtue of the strong law of larger numbers. Apply mathematical induction on n. □

We propose the limiting case as $m \to \infty$ as an approximation, i.e., the deterministic sequence $\{x_n : n \geq 0\}$ specified by (3.2). Clearly, for large finite m, the state variable x_n means that a proportion $(x_n - \lfloor x_n \rfloor)$ of the m queues will be assigned $\lfloor x_n \rfloor + 1$ jobs, while the remainder of the queues are assigned $\lfloor x_n \rfloor$ jobs. The central limit theorem can be used to describe deviations from the limiting behavior for finite m. Let $N(a, b)$ denote a normally distributed random variable with mean a and variance b. Let $\Rightarrow$ denote convergence in distribution; e.g., see Billingsley [9].

Theorem 10. In the setting of Theorem 9,

$$\frac{(N_{n+1}^{(m)} - mx_{n+1} | N_n^{(m)} = mx_n)}{\sqrt{m}} \Rightarrow N(0, v_{n+1}) \quad as \quad m \to \infty \ , \tag{3.4}$$

where $v_{n+1} = V(T, x_n)$ with

$$\begin{aligned} V(t,x) = & (x - \lfloor x \rfloor) Var[Q(t) | Q(0) = \lfloor x \rfloor + 1] \\ & + (1 - x + \lfloor x \rfloor) Var[Q(t) | Q(0) = \lfloor x \rfloor] \ . \end{aligned} \tag{3.5}$$

Proof. We can apply the central limit theorem after noting that the quantity on the left in (3.4) is the sum of m independent random variables, where $m(x_n - \lfloor x_n \rfloor)$ have one distribution, while $m(1 - x_n + \lfloor x_n \rfloor)$ have another distribution. The second moments are finite, being bounded above by the second moments of a constant (the initial value) plus the Poisson number of arrivals. □

We now draw implications for the distribution at n steps. For this purpose, we use the following elementary lemma.

Lemma 11. Suppose that $X \stackrel{d}{=} N(Y, \sigma_1^2)$, where $Y \stackrel{d}{=} N(m, \sigma_2^2)$. Then

$$X \stackrel{d}{=} N(m, \sigma_1^2 + \sigma_2^2) \ .$$

Proof. Note that

$$\begin{aligned} Ee^{itX} &= E[Ee^{itX} | Y] = E \exp(itY + t^2 \sigma_1^2 / 2) \\ &= \exp(itm + t^2(\sigma_1^2 + \sigma_2^2)/2) \ . \end{aligned}$$

□

We now apply Theorem 10 and Lemma 11 with mathematical induction to obtain the asymptotic distribution of $N_n^{(m)}$.

Corollary 12. In the setting of Theorem 9,

$$\frac{N_n^{(m)} - mx_n}{\sqrt{m}} \Rightarrow N(0, \sum_{k=1}^{n} v_k) \quad \text{as } m \to \infty$$

for each n, where v_k is defined in (3.5).

We remark that Theorem 10 and Corollary 12 imply that the standard deviation of the number assigned to each queue after load balancing with m queues is of order $1/\sqrt{m}$. Since $Var[Q(t)|Q(0) = j]$ becomes small as t decreases, the deterministic approximation tends to be more accurate when m is larger and when T is smaller.

Intuitively, it is apparent that the limiting case as $m \to \infty$ is optimistic (serves a a lower bound). This can be made precise by a stochastic comparison. The following is proved by a minor modification of the proof of Theorem 7.

Theorem 13. In the setting of Theorem 9, if $s = 1$ and $N_0^{(m)} = x_0 m$, then

$$Eg(N_n^{(m)}/m) \geq g(x_n) \quad \text{for all } n \geq 1$$

and all nondecreasing convex real-valued functions g, where $\{x_n\}$ satisfies (3.2).

We now want to describe the evolution of the limiting deterministic sequence $\{x_n : n \geq 0\}$ defined by (3.2).

Theorem 14. The function f_T in (3.2) characterizing the evolution of $\{x_n\}$ is strictly increasing and continuous. If $\rho < 1$, then there is a unique fixed point $x^*(T)$ of the equation $f_T(x) = x$ for each T and $x_n \to x^*(T)$ as $n \to \infty$ for each x_0.

Proof. First fix T. From (3.2) it is immediate that $x_{n+1} = f(x_n)$ for a function $f \equiv f_T$ independent of n. Continuity of f follows from the M/M/s model structure; i.e., changes in state in an interval of length h occur with probability of order $O(h)$ as $h \to 0$; we omit the details. Monotonicity of f can be shown by artificially constructing two M/M/s processes on the same sample space so that the sample paths are ordered strictly until they couple, as in Theorem 3. This is achieved by letting the two processes have the same arrival process. The higher process has the same departures as the lower one plus possibly additional ones until they couple (the sample paths coincide). This yields $(Q(t)|Q(0) = n)$ stochastically increasing in n for each t. Hence, $E[Q(t)|Q(0) = n]$ is strictly increasing in n for each t, so that f is strictly increasing. Since $f(0) > 0$, successive iterates $f^{(n)}(x) \equiv f(f^{(n-1)}(x))$ increase as $n \to \infty$ to a limit x^* starting in 0. Since f is continuous as well, $f^{(n)}(x^*) = f(f^{(n-1)}(x^*)) \to f(x^*)$ as $n \to \infty$, so that x^* is a fixed point of f. For all x with $0 < x < x^*$, $f^{(n)}(0) < f^{(n)}(x) < f^{(n)}(x^*) = x^*$, so that $f^{(n)}(x) \to x^*$ as $n \to \infty$ too. As x increases, $f(x) - x$ approaches $(\lambda - s)T$. Since $\rho < 1$, $(\lambda - s)T < 0$. Hence, for all sufficiently large x, $f(x) < x$. Hence, for such x, $f^{(n)}(x)$ decreases to a limit $\hat{x}$ as $n \to \infty$. Since f is continuous, $\hat{x}$ must also be a fixed point of f. Since coupling is always possible in the special construction above, we must have $f(\hat{x}) - f(x^*) < \hat{x} - x^*$ if $\hat{x} > x^*$. Hence, we must have $\hat{x} = x^*$, so that there is a unique fixed point. Moreover, $f^{(n)}(x) \to x^*$ as $n \to \infty$ for all x. (Monotonicity can be used for $x > x^*$, just as it was for $x < x^*$.) □

As a consequence of Theorem 13, we can deduce that the fixed point $x^*(T)$ is a lower bound for the steady-state random variable $N_0^{(m)}/m$ in the increasing convex stochastic order.

Corollary 15. In the setting of Theorem 13, $N_\infty^{(m)} \geq_{ic} mx^*(T)$ for each m.

We now describe how the fixed point $x^*(T)$ depends on upon T.

Theorem 16. The fixed point $x^*(T)$ of the equation $x = f_T(x)$ for f_T in Theorem 9 is a strictly increasing continuous function of T with $x^*(T) \to x_U$ as $T \to \infty$ and $x^*(T) \to x_L$ as $T \to 0$, where

$$EQ(\infty) \leq x_U \leq \lfloor EQ(\infty) \rfloor + 1 , \tag{3.6}$$

while

$$\lfloor x_Z \rfloor \leq x_L \leq x_Z \tag{3.7}$$

where x_Z is the unique value of x such that $M'(0,x) = 0$ for $M(t,x)$ in (3.3).

Proof. Note that

$$M'(0,x) = (x - \lfloor x \rfloor)(\lambda - \lfloor x \rfloor + 1) + (1 - x + \lfloor x \rfloor)(\lambda - \lfloor x \rfloor) .$$

Hence $M'(0,x)$ so can be 0 only if $\lfloor x \rfloor < \lambda < \lfloor x \rfloor + 1$. Suppose that is the case. Since $M'(0, \lfloor x \rfloor) > 0$ and $M'(0, \lfloor x \rfloor + 1) < 0$ and since $M'(0,x)$ is continuous and monotone, there is one and only one x for which $M'(0,x) = 0$; let it be denoted by x_Z. For each x satisfying $x_Z < x < EQ(\infty)$, $M(t,x)$ initially decreases and then eventually converge to $EQ(\infty)$. Since $x < EQ(\infty)$ and $M(t,x)$ is continuous in t, there must be an intermediate T yielding a fixed point. Thus, each of these x is a fixed point for some T, denoted by $x^*(T)$. By the coupling construction used in Theorems 3 and 14 if $x_1 < x_2$, then $M(t,x_2) - M(t,x_1)$ must be strictly decreasing in t. Thus, if $x < x^*(T)$, then

$$M(T, x^*(T)) - M(T,x) < x^*(T) - x$$

or, since $M(T, x^*(T)) = x^*(T)$, $M(T,x) > x$, which implies that the time yielding the fixed point for x, denoted by T_x^*, must be less than T. Hence, the fixed point times T_x must be strictly increasing in x. (The strict order is also implied by Theorem 13.) By continuity of $M(t,x)$ in x, T_x^* must be continuous in x as well, which implies that the inverse of T_x^*, $x^*(T)$, is continuous and strictly increasing as well. Let x_U and x_L be the limiting fixed points, which must be defined, since a fixed point $x^*(T)$ exists for all positive T. We have noted that $x_L \leq x_Z$ and $x_U \geq EQ(\infty)$. We now show that $x_L \geq \lfloor x_Z \rfloor$ and $x_U \leq \lfloor EQ(\infty) \rfloor + 1$, as in (3.6) and (3.7). For this step, we exploit known properties of the mean function $M(t,n) \equiv E[Q(t)|Q(0)=n]$ in M/M/s queues, as given in Lemma 9.4.1 (ii) and Theorem 9.4.3 (ii) of van Doorn [40]: First if $M'(t,n) \geq 0$, then $M'(t+u,n) \geq 0$ for all $u > 0$. Second if $M'(t,n) \leq 0$, then also $M''(t,n) \geq 0$. By the first property, no integer n can be a fixed point if $n \geq EQ(\infty)$, because $M(t,n)$ must first decrease and then eventually increase to $EQ(\infty)$. It cannot go above $EQ(\infty)$, because it must converge to $EQ(\infty)$ and remain nondecreasing after it first becomes nondecreasing. Similarly, no integer $n \leq x_Z$ can be a fixed point, because $M(t,n)$ is always increasing, again by the first property. □

Remark. We conjecture that the limiting fixed points in Theorem 16 are $x_U = EQ(\infty)$ and $x_L = x_Z$, but that remains to be proven. The difficulty is in treating non-integer x. It is not clear whether the properties of the mean function in van Doorn [40] used in the proof extend to convex combinations $pM(t,n+1)+(1-p)M(t,n)$. However, we can cover a subset of cases: Let $n = \lfloor EQ(\infty) \rfloor$ and let T_n be the fixed point time. If $M'(T_n,n+1) > 0$, then $x_U = EQ(\infty)$. This is so, because if $M'(T_n,n+1) > 0$, then $M(t,n+1)$ must go below $EQ(\infty)$ and increase to it. Hence $M(t,x) < EQ(\infty)$ for all $t > T_n$, but we must have $T_x > T_n$ for $x > n$. Hence there can be no fixed point x for $x \geq EQ(\infty)$.

Given a desired level after redistribution, x, or a desired redistribution interval T, we can find the associated fixed point $T_x = x^*(T)$ by solving the fixed point equation $f_T(x) = x$ for the free variable. Computation is aided by the monotonicity of the function $x^*(T)$. To compute $M(t,x)$ as a function of t, we need to compute the transient mean $E[Q(t)|Q(0)=n]$ in the M/M/s model. We can do so numerically by imposing a suitably large finite waiting room and solving a finite system of ordinary differential equations, as in Taaffe and Ong [38]. A related alternative algorithm is given in Davis, Massey and Whitt [13]. It is applied for the M/M/s delay model with time-dependent arrival rate in Massey and Whitt [30].

For the special case of $s=1$, it is especially convenient to use numerical integration with integral representations, as indicated in Abate and Whitt [3]. For example, for $s=1$, the mean function is given by

$$M(t,n) = \frac{\rho}{1-\rho} - \frac{2\rho^{-n/2}}{\pi} \int_0^{\pi} \frac{e^{-\gamma(y)t} \sin y (\sin(n+1)y - \rho^{-1/2}\sin(ny))}{\gamma(y)^2} dy \quad (3.8)$$

where

$$\gamma(y) = 1 + \rho - 2\sqrt{\rho}\cos y \; ; \tag{3.9}$$

see p. 27 of Takács [39]. (Formula (3.8) is expressed slightly differently in [39].)

Alternatively, for $s = 1$, the mean can be calculated by numerical transform inversion. For the M/M/1 model, the Laplace transform of the conditional mean $E[Q(t)|Q(0) = n]$ with respect to time t is given explicitly on p. 162 of Abate and Whitt [2] (see also pp. 148 and 157 there). For example, the Fourier-series method for numerically inverting Laplace transforms can be applied; see Abate and Whitt [4], [5]. In [2] time is scaled, so that the normalized mean $M(t,n)/M(\infty)$ converges to a nondegenerate limit as $\rho \to 1$. This nondegenerate limit is the RBM limit discussed in the next section. For further discussion about the connection between M/M/1 and RBM characteristics, see Section 10 of [2]. As indicated after Theorem 1 on p. 148 of [2], the transform of the second moment function can be obtained in the same way.

By Corollary 15, the normalized fixed point $mx^*(T)$ is a lower bound in the increasing convex stochastic order for the steady-state quantity $N_\infty^{(m)}$ for finite m. The quantity $mx^*(T)$ also serves as a first order approximation to $N_\infty^{(m)}$ for finite m. As in Theorem 10, we can invoke the central limit theorem to generate a refinement, in particular, the normal distribution. Assuming that the steady-state number of jobs in the system at redistribution times is approximately normally distributed, we can solve a pair of equations to calculate the mean and variance. (Recall that a normal distribution is fully characterized by its mean and variance.) Let μ and σ^2 denote the approximating steady-state mean and variance of the number of jobs in each queue after balancing. Let $\phi(x|\mu,\sigma^2)$ and $\Phi(x|\mu,\sigma^2)$ denote the probability density function (pdf) and cumulative distribution functions (cdf) of a $N(\mu,\sigma^2)$ random variable, i.e., normally distributed with mean μ and variance σ^2. Then, for any balancing interval T and any number of queues m, the parameter μ and σ^2 can be obtained as the solution to the pair of equations

$$\mu = \Phi(0|\mu,\sigma^2)M(T,0) + \int_0^\infty \phi(x|\mu,\sigma^2)M(T,x)dx \tag{3.10}$$

and

$$\sigma^2 = m^{-1}\left[\Phi(0|\mu,\sigma^2)M_2(T,0) + \int_0^\infty \phi(x|\mu,\sigma^2)M_2(T,x)dx - \mu^2\right] \; , \tag{3.11}$$

where $M_2(T,x)$ is the second moment, i.e.,

$$M_2(T,x) = V(T,x) + M(T,x)^2 \tag{3.12}$$

for $M(T,x)$ and $V(T,x)$ in (3.3) and (3.5). The approximate distributions just before and after load balancing are thus $N(\mu, m\sigma^2)$ and $N(\mu,\sigma^2)$, respectively.

In words, we calculate the mean and second moment as normal mixtures over the initial state x of the means $M(T,x)$ and second moments $M_2(T,x)$.

The first terms on the right in (3.10) and (3.11) account for the possibility that the normal approximation can have positive mass at negative values. The factor m^{-1} in (3.11) is present because the variance of an average of m i.i.d. terms is $1/m$ times the variance of one term. (We act as if the terms are identically distributed even though they are not quite because of the integrality condition.) The equations (3.10) and (3.11) can be solved iteratively for the desired pair (μ, σ^2). We apply this approximation scheme with the RBM approximation in the next section.

The deterministic fixed point $x^*(T)$ serves as a convenient approximation for the mean μ in the normal iteration above. When m is large, the variance σ^2 should be reasonably well approximated by the variance in one interval starting from $x^*(T)$ divided by m; i.e., we can use the simple approximation

$$N(\mu, \sigma^2) \approx N(x^*(T), V(T, x^*(T))/m) \ . \tag{3.13}$$

for $V(t,x)$ in (3.5).

The performance is only partly determined by the random quantity $N_\infty^{(m)}$. Thus, even in the limit as $m \to \infty$, the performance is only partly determined by the fixed point function $x^*(T)$. The queues evolve randomly in each redistribution interval. Thus, to describe the performance, we also want to calculate the probability distribution of the queue length in between redistribution points and appropriate summary statistics. We can calculate the transition probability function by similar methods. For the M/M/1 case, we can use either integral representations or numerical transform inversion. The time-dependent cumulative distribution function is readily calculated by two-dimensional numerical inversion, as in Choudhury, Lucantoni and Whitt [11]. For the M/M/1 case, the busy-period transform is available explicitly, so it is not necessary to obtain it by iterating the Kendall functional equation for the busy period as in [11].

The random number of jobs that must be moved from one queue becomes more elementary as $m \to \infty$. Considering the worst case in which we start with $\lfloor x^*(T) \rfloor + 1$ and end with $\lfloor x^*(T) \rfloor$, this random number is

$$J_{i\infty} = [(Q(T)|Q(0) = \lfloor x^*(T) \rfloor + 1) - \lfloor x^*(T) \rfloor]^+ \ . \tag{3.14}$$

We can calculate the distribution of $J_{i\infty}$ by first solving for the deterministic fixed point $x^*(T)$ and then calculating the cdf of $(Q(T)|Q(0) = \lfloor x^*(T) \rfloor + 1)$, as indicated above.

4. A Heavy-Traffic Diffusion Approximation

We now see how the balancing interval T and the fixed-point function $x^*(T)$ developed in Section 3 should scale with the traffic intensity ρ. We also develop more tractable approximations for both the M/M/s case considered before and the G/GI/s case with more general arrival and service processes. A major simplification resulting from the scaling as $\rho \uparrow 1$ is the elimination of the integrality constraint; i.e., the queue lengths no longer need be integers. Thus formulas such as (3.3) and (3.5) simplify.

For each ρ with $0 < \rho < 1$, let a queueing model with traffic intensity ρ be defined by scaling a rate-1 arrival process $\{A(t) : t \geq 0\}$ by $A_\rho(t) = A(\rho t), t \geq 0$. Let $\{A(t) : t \geq 0\}$ denote an arrival process to any one queue. Assume that the arrival processes to different queues are mutually independent. Assume that each arrival process satisfies a functional central limit theorem (FCLT), i.e.,

$$\frac{A(nt) - \lambda nt}{\sqrt{n\lambda c_a^2}} \Rightarrow B(t) \text{ in } D \text{ as } n \to \infty , \tag{4.1}$$

where $\{B(t) : t \geq 0\}$ is standard (drift 0, diffusion coefficient 1) Brownian motion (BM) and $\Rightarrow$ denotes weak convergence (convergence in distribution) in the function space $D \equiv D[0, \infty)$; see Billingsley [9], Ethier and Kurtz [15] and Whitt [43]. If $\{A(t) : t \geq 0\}$ is a renewal process, then to satisfy (4.1) it is necessary and sufficient for the time between renewals to have a finite second moment. Then its SCV is c_a^2 in (4.1). The form of (4.1) allows dependence among successive arrivals. We assume that the service times are independent of the arrival process, coming from an i.i.d. sequence with a general distribution having mean 1 and finite second moment. Let c_s^2 denote the SCV of a service time. The independence assumed for the service times is not strictly needed. It would suffice for the partial sums of the service times at each queue to satisfy a FCLT; see Iglehart and Whitt [24]. then c_s^2 should be determined by the normalization in the FCLT as in (4.1). It is important is recognize that dependence can influence the parameters c_a^2 and c_s^2. In general there could even be a term c_{as}^2 reflecting the dependence between arrival times and service times, see Fendick, Saksena and Whitt [17], but we assume that $c_{as}^2 = 0$.

In this setting, the normalized queue length process in the standard G/GI/s model converges to RBM as $\rho \to 1$ by Iglehart and Whitt [24]. In particular, as reviewed in Whitt [46], if $Q_\rho(t)$ denotes the queue length (number in system) at time t in a standard G/GI/s system indexed by ρ, then

$$\frac{(1-\rho)}{\rho(c_a^2 + c_s^2)} Q_\rho(t(c_a^2 + c_s^2)/s(1-\rho)^2) \Rightarrow R(t) \text{ in } D \text{ as } \rho \to 1 , \tag{4.2}$$

where $\{R(t) : t \geq 0\}$ is canonical (drift -1 and diffusion coefficient 1) RBM. We insert the extra ρ in the denominator of the initial multiplicative factor in (4.2) as a heuristic refinement to make the formula exact for the M/M/1 steady-state mean $\rho/(1-\rho)$. (The steady-state RBM variable $R(\infty)$ is exponentially distributed with mean 1/2.) Of course, the ρ is asymptotically negligible as $\rho \to 1$. As a consequence of the limit in (4.2), we have the associated approximation

$$Q_\rho(t) \approx \frac{\rho(c_1^2 + c_s^2)}{1-\rho} R(s(1-\rho)^2 t/(c_a^2 + c_s^2)) . \tag{4.3}$$

We now state the analog for periodic load balancing. We will only sketch the proof since the heavy-traffic limit follows by essentially the same argument as in Iglehart and Whitt [24] and Kella and Whitt [25]. Let $N_{i\rho}^{(m)}(t)$ denote the queue length in the i^{th} queue at time t with m queues and traffic intensity ρ. Let Φ be the cdf of the standard (mean 0 and variance 1) normal distribution and let ϕ be its density. Let Φ^c be the complementary cdf, i.e., let $\Phi^c(x) = 1 - \Phi(x)$.

Theorem 17. Consider m G/GI/s queues controlled by periodic load balancing. Make the assumptions above on the arrival and service processes. If $\rho \rightarrow 1$ with the redistribution intervals ρ satisfying

$$\frac{s(1-\rho)^2 T_\rho}{(c_a^2 + c_s^2)} \quad \rightarrow \quad T \tag{4.4}$$

and the initial queue lengths $x_{0\rho}$ satisfying

$$\frac{(1-\rho)}{\rho(c_a^2 + c_s^2)} x_{0\rho} \quad \rightarrow \quad x_0 \ , \tag{4.5}$$

then the queue-length processes converge to load-balanced RBM, i.e.,

$$\frac{(1-\rho)}{\rho(c_a^2 + c_s^2)} (N_{i\rho}^{(m)}(t(c_a^2 + c_s^2)/s(1-\rho)^2) : 1 \le i \le m)) \Rightarrow \\ (X_i(t) : 1 \le i \le m) \ in \ D^m, \tag{4.6}$$

where $\{X_i(t) : t \ge 0\}$ are conditionally i.i.d. processes given $\{(X_1(nT), \ldots, X_m(nT)) : n \ge 0\}$, $Y_n \equiv X_1(nT) + \ldots + X_m(nT)$, $n \ge 0$, is a stochastically monotone, irreducible, aperiodic Markov process on $\mathbb{R}$ with transition probabilities

$$P(Y_{n+1} \le y | Y_n = x) = P(\sum_{i=1}^{m} R_i(T) \le y | R_i(0) = x/m \ , \quad 1 \le i \le m) \tag{4.7}$$

and conditional Laplace transform

$$E(e^{-sY_{n+1}} | Y_n = x) = (E(e^{-(s/m)R(T)} | R(0) = x/m))^m \ , \tag{4.8}$$

and $\{R_i(t) : t \ge 0\}$ are m i.i.d. canonical RBMs, with

$$P(R(t) > y | R(0) = x) = \Phi\left(\frac{-y + x - t}{\sqrt{t}}\right) + e^{-2y}\Phi\left(\frac{-y - x + t}{\sqrt{t}}\right) \ , \tag{4.9}$$

and

$$X_i(nT + t) \stackrel{d}{=} (R(t) | R(0) = Y_n/m) \ , \quad 0 \le t < T \ . \tag{4.10}$$

Proof. Proceed by induction over successive redistribution intervals. At each redistribution point the residual interarrival times and service times are asymptotically negligible. For the arrival times this follows from the FCLT (4.1). That

FCLT implies a corresponding FCLT for the inverse partial sum process and the normalized maximum jump in it over any interval is 0. The remaining argument follows Iglehart and Whitt [24]. The result is a Markov process as in Theorem 1 with the individual queues evolving as canonical RBM instead of the M/M/s queue length process. The conditional complementary cdf in (4.9) is standard; see (1.1) of [1]. □

The evolution of the limiting stochastic process $\{(X_1(t), \ldots, X_m(t) : t \geq 0\}$ in Theorem 17 can be described by first calculating the distribution of the variables Y_n and then applying (4.10). The Markov chain kernel (transition probability density function) giving the conditional density of Y_{n+1} given Y_n can be found by numerically inverting the transform in (4.8), exploiting the two-dimensional Laplace transform

$$\hat{\psi}(s, \sigma|x) \equiv \int_0^\infty e^{-st} E(e^{-\sigma R(t)}|R(0) = x)dt \ , \tag{4.11}$$

which is given explicitly in (9.3) of [1]. The numerical transform inversion algorithm in Choudhury, Lucantoni and Whitt [11] can be used to calculate the transition kernel. The steady-state distribution of the Markov chain $\{Y_n\}$ can be calculated by making a finite-state approximation. However, we will use other approximations below.

A more elementary approximation can be obtained by considering the double limit as $\rho \to 1$ and then $m \to \infty$. An attractive feature of the following RBM limit is the explicit form for the mean function in (4.13) below.

Theorem 18. In the setting of Theorem 17, if $m \to \infty$ after $\rho \to 1$, then (4.6) holds, $x_n \equiv X_1(nT)$ evolves deterministically as

$$x_{n+1} = f_T(x_n) \tag{4.12}$$

where

$$\begin{aligned} f_T(x) \equiv M(t, x) \quad &\equiv \quad E[R(t)|R(0) = x] \\ = \frac{1}{2} + \sqrt{t}\phi\left(\frac{t-x}{\sqrt{t}}\right) &- (t - x + \frac{1}{2})\Phi^c\left(\frac{t-x}{\sqrt{t}}\right) - \frac{1}{2}e^{2x}\Phi^c\left(\frac{t+x}{\sqrt{t}}\right), \end{aligned} \tag{4.13}$$

$\{R(t) : t \geq 0\}$ is canonical RBM, and

$$X_i(nT + t) \stackrel{d}{=} (R(t)|R(0) = x_n), \quad 0 \leq t < T, \quad i \geq 1 \ . \tag{4.14}$$

Proof. The additional limiting argument for $m \to \infty$ is as in Theorem 9. The mean function in (4.13) comes from Theorem 1.1 of [1]. □

The approximation based on Theorem 17 is load-balanced canonical RBM using an redistribution interval T. By (4.4) and (4.6), the associated approximate redistribution interval T_ρ and levels $x_{\rho n}$ in the queueing system with traffic intensity ρ are

$$T_\rho \approx \frac{(c_a^2 + c_s^2)T}{s(1-\rho)^2} \tag{4.15}$$

and

$$x_{\rho n} \approx \frac{\rho(c_a^2 + c_s^2)x_n}{1-\rho} \ . \tag{4.16}$$

Theorem 18 implies that we can study periodic load balancing for canonical RBM and apply the results to generate approximations for the general G/GI/s queueing model, provided that ρ and m are suitably large. The limit (4.6) generates the approximation

$$N_{1\rho}(t) \approx \left(\frac{\rho(c_a^2 + c_s^2)}{1-\rho}\right) X_1(s(1-\rho)^2 t/(c_a^2 + c_s^2)) \ , \tag{4.17}$$

where $(X_1(t), \ldots, X_m(t))$ is controlled canonical RBM, as indicated in Theorem 17. Thus, invoking Theorem 18 as well, the queue length just before and after the n^{th} redistribution has the approximate form

$$N_{1\rho}(nT_\rho-) \approx \frac{\rho(c_a^2 + c_s^2)}{1-\rho} X_1(nT-) \stackrel{d}{=} \frac{\rho(c_a^2 + c_s^2)}{1-\rho}(R(T)|R(0) = x_{n-1}) \tag{4.18}$$

and

$$N_{1\rho}(nT_\rho) \approx \frac{\rho(c_a^2 + c_s^2)}{1-\rho} X_1(nT) = \frac{\rho(c_a^2 + c_s^2)}{1-\rho} x_n \tag{4.19}$$

for $\{x_n\}$ in (4.12). For ease of application, it is significant that the conditional mean function for RBM, $M(t,x)$ in (4.13), and the conditional complementary cdf, $P(R(t) > y|R(0) = x)$ in (4.9), are available explicitly. Unlike with Section 3 and Theorem 17, no numerical integration or numerical transform inversion is needed. The standard normal cdf Φ is usually available on computers often via the error function. It can be computed directly using rational approximations for the error function; see p. 299 of Abramowitz and Stegun [6].

Remark. Theorem 17 and approximations (4.15)–(4.19) also have important implications for customer waiting times. Since the cycle lengths are of order $(1-\rho)^{-2}$ while the queue lengths and waiting times are of order $(1-\rho)^{-1}$ as $\rho \to 1$, we see that arrivals will tend to be served in the same cycle they arrive in. The waiting time can thus be approximated by a random sum of i.i.d. service times, where the random numer at time t is $N_{1\rho}(t)$ in (4.17). The expected waiting time is thus just $EN_{1\rho}(t)$.

By Theorem 18, the transient behavior of the approximate system at (just after) balancing points for any balancing interval T is described by the conditional

RBM mean $M(T, x)$ in (4.13). We display the mean function $M(t, x)$ as a function of t for several x in Figure 1. We can see that $M(t, x)$ approaches the steady-state mean $ER(\infty) = 1/2$ as $t \to \infty$. We also see that $M'(0, x) = -1$ for all $x > 0$, because canonical RBM behaves initially like canonical BM with drift -1, since it starts at the point x away from the reflecting barrier at 0.

Theorems 17 and 18 allow us to describe the impact of the arrival pattern. If each queue has its own arrival process initially, then the parameter c_a^2 in (4.1) is just the one associated with the arrival process. On the other hand, suppose that there is a single arrival process to the system (with stationary increments), with jobs assigned to the queues upon arrival. As noted before, if the assignment is random, then $c_a^2 = 1$, because the split processes to individual queues become independent Poisson processes as $m \to \infty$. On the other hand, if the assignment is round robin, then $c_a^2 = 0$, because the split processes to individual queues become deterministic as $m \to \infty$. For finite m, we would let $c_a^2(m) \approx c_a^2/m$, because that is what happens with a renewal arrival process. (The new interarrival time is the sum of m i.i.d. original interarrival times.) Hence, the three possible arrival patterns are reflected by the single parameter c_a^2. Since the total impact of the variability of the arrival and service processes is reflected by the term $(c_a^2 + c_s^2)$, the arrival pattern makes a bigger (relative) difference when c_s^2 is small. When $c_a^2 = c_s^2 = 0$, the normalized queue lengths are asymptotically negligible in the limit. (It is an open problem to determine if there is a nondegenerate limit with a different normalization.)

The limit (4.4) in Theorem 17 and the approximate formula (4.15) show how the redistribution interval T_ρ should grow with ρ in order to obtain a nondegenerate RBM limit. If T_ρ grows more slowly, then the normalized queue lengths are asymptotically negligible. Similarly, if T_ρ grows more quickly, then the queue reaches steady state before the redistribution. We formalize this behavior below.

Corollary 19. Consider the setting of Theorem 18. If, instead of (4.4), $(1 - \rho)^2 T_\rho \Rightarrow 0$ as $\rho \to 1$, then

$$(1-\rho) N_{1\rho}^{(m)}(t) \Rightarrow 0 \quad as \quad \rho \to 1 \quad for\ each\ t\ . \tag{4.20}$$

On the other hand, if $(1 - \rho)^2 T_\rho \Rightarrow \infty$ as $\rho \to 1$, then

$$\frac{(1-\rho)}{\rho(c_a^2 + c_s^2)} N_{1\rho}^{(m)}(kT_\rho -) \Rightarrow R(\infty) \quad as \quad \rho \to 1 \ \text{ and then } m \to \infty \text{ for each } k, \tag{4.21}$$

where

$$P(R(\infty) > y) = e^{-2y}\ , \quad y \geq 0\ . \tag{4.22}$$

We now state an analog of Corollary 12, providing a normal distribution refinement to the deterministic sequence $\{x_n\}$.

Theorem 20. In the setting of Theorem 17,

$$\sqrt{m}(\eta(\rho)N_{1\rho}^{(m)}(nT_\rho) - x_n) \to N(0, \sum_{k=1}^{n} v_k) \tag{4.23}$$

as $\rho \to 1$ and then $m \to \infty$ for each n, where $\eta(\rho) = (1-\rho)/\rho(c_a^2 + c_s^2)$, x_n satisfies (4.12),

$$v_k \equiv V(T, x_{k-1}) \equiv Var(R(t)|R(0) = x_{k-1}) \ , \quad k \geq 1 \ . \tag{4.24}$$

$V(t,x) = M_2(t,x) - M(t,x)^2$, $M(t,x)$ as in (4.13) and

$$\begin{aligned} M_2(t,x) = & \frac{1}{2} + ((x-1)\sqrt{t} - t^{3/2})\phi\left(\frac{t-x}{\sqrt{t}}\right) \\ & + ((t-x)^2 + t - \frac{1}{2})\Phi^c\left(\frac{t-x}{\sqrt{t}}\right) + e^{2x}(t + x - \frac{1}{2})\Phi^c\left(\frac{t+x}{\sqrt{t}}\right) . \end{aligned} \tag{4.25}$$

Proof. The argument is essentially the same as for Theorem 10 and Corollary 12. Indeed, with the scaling in (4.6) RBM is contained as a special case of M/M/1 with $\rho = 1$; see Abate and Whitt [1], [2] for further discussion. The conditional second moment function in (4.25) comes from Theorem 1.1 of Abate and Whitt [1]. □

To show the form of the conditional RBM variance $V(t,x)$, we display it as a function of t for several values of x in Figure 2. To show different regions, we display it in two scales, over the intervals $[0,5]$ and $[0,0.25]$. Note that $V(t,x) \to Var\ R(\infty) = 1/4$ as $t \to \infty$. Note that $V(t,x) \approx t$, for suitably small t, which is the variance of ordinary BM, $B(t)$. A crude upper bound is $V(t,x) \leq \min\{t, 1/4\}$.

Paralleling Theorem 14, there is a unique fixed point for the RBM function f_T in (4.12). Indeed, results for RBM can be obtained directly from previous M/M/1 results by regarding RBM as the limit (after scaling) as $\rho \to 1$. Here are properties of the RBM fixed point function $x^*(T)$.

Theorem 21. The function f_T in (4.12) is strictly increasing and continuous. There is a unique fixed point $x^*(T)$ of the equation $x = f_T(x)$ for each T and $x_n \to x^*(T)$ as $n \to \infty$. The fixed point $x^*(T)$ is a strictly increasing continuous function of T with $x^*(T) \to 1/2$ as $T \to \infty$ and $x^*(T) \to 0$ as $T \to 0$.

Proof. The proof is essentially the same as for Theorems 14 and 16. By-taking the heavy traffic limit after scaling in Theorem 16, we obtain $x_U = ER(\infty) = 1/2$ and $x_L = 0$. The shape of the RBM first moment function was previously established in Section 8 of Abate and Whitt [1]. □

From Figure 1, we can see that for each x with $0 < x < 1/2 = ER(\infty)$, there is a unique T such that x is a fixed point, i.e., $x = M(T,x)$, and we can see how this fixed point $x^*(T)$ depends upon x or T. We display the RBM fixed point $x^*(T)$ as a function of T and x in Figure 3. Parts (a) and (b) of Figure 3 give separate displays over the intervals $[0,4]$ and $[0,0.5]$. The longer interval shows that the fixed point $x^*(T)$ gets quite close to the limit $1/2$ occurring as $T \to \infty$ for $1 \leq T \leq 4$. The shorter interval $[0,0.5]$ shows that the region where $x^*(T)$ ranges from 20% to 80% of the limit $1/2$ is about $0.01 \leq T \leq 0.5$. The cases of $T = 1.0$ and 0.1 are highlighted because we use them in our simulation experiments in Section 6.

We can also combine Theorems 18 and 21 to describe the asymptotic behavior of the fixed point equation. We need to have the means converge to the mean of the limit in (4.6). This holds under an additional uniform integrability assumption; see p. 32 of Billingsley [9]. (We regard this as a minor technical regularity condition.)

Theorem 22. If, in addition to the conditions of Theorem 17, the normalized queue-length variables $N_{1\rho}^{(m)}(t)$ are uniformly integrable, then the associated fixed-point levels satisfy

$$\frac{(1-\rho)x_\rho^*(T_\rho)}{\rho(c_a^2+c_s^2)} \to x^*(T) \quad as \quad \rho \to 1 \ .$$

Proof. By (4.6) in Theorem 17 the transient mean functions converge, implying that the normalized fixed points converge as well. □

The first order approximation for the level in one queue after balancing in the RBM model is $x^*(T)$ computed from the fixed point equation associated with f_T in (4.12) and (4.13). Just as in equations (3.10)–(3.12), a refined approximation is a normal distribution, where the mean and variance σ^2 are the solutions of a pair of equations. With RBM, the equations are still (3.10)–(3.12) but with RBM the mean and second-moment functions $M(T,x)$ and $M_2(T,x)$ are greatly simplified, being as in (4.13) and (4.25). Just as in (3.13), an approximation for this stochastic normal fixed point is the normal distribution $N(x^*(T), V(T,x^*(T))/m)$, which is the normal distribution we obtain after balancing at the end of a single interval of length T, starting at $x^*(T)$. The simple normal approximation $N(x^*(T), V(T,x^*(T))/m)$ motivates displaying $V(T,x^*(T))$, the variance function starting at the fixed point $x^*(T)$. We do so in Figure 4.

We compare these approximation schemes in Tables 1 and 2. In Table 1 we compare the deterministic fixed point $x^*(T)$ to the mean $\mu \equiv \mu(T)$ in the pair (μ, σ^2) obtained from the normal iteration in (3.10)–(3.12) for RBM for six values of T ($T = 0.01, 0.05, 0.10, 0.50, 1.00, 5.00$) and four values of m ($m = 2, 4, 16, 64$). The equations (3.10) and (3.11) were solved iteratively using numerical integration with (4.13) and (4.25) to calculate the integrals. The iteration tended to converge relatively quickly (3–20 iterations), starting from an initial pair $(\mu, \sigma^2) = (0, \epsilon)$ for a small positive ϵ.

As illustrated by the cases with $m = 64$ in Table 1, $\mu \approx x^*(T)$ when m is suitably large. The agreement in these cases also confirms that both calculations can be performed with sufficient accuracy. When m is not large, $x^*(T)$ underestimates μ.

In Table 2 we compare the corresponding approximations for the standard deviation of the steady-state queue content just before load balancing with m independent RBM processes. In particular, we compare $\sqrt{m}\sigma$ from the normal iteration to $\sqrt{V(T, x^*(T))}$. As with the mean, when m is suitably large, e.g., when $m = 64$, $\sqrt{m}\sigma \approx \sqrt{V(T, x^*(T))}$, but the more elementary approximation $\sqrt{V(T, x^*(T)}$ underestimates $\sqrt{m}\ \sigma$ when m is small.

Table 1
A comparison between μ, the approximation for the steady-state mean content of each queue just before (and after) load balancing with m independent RBM processes, using the normal iteration, and the deterministic fixed point $x^*(T)$.

	μ from normal iteration				
T	$m = 2$	$m = 4$	$m = 16$	$m = 64$	$x^*(T)$
0.01	0.1756	0.1526	0.1358	0.1347	0.1336
0.05	0.2850	0.2504	0.2314	0.2274	0.2260
0.10	0.3321	0.2999	0.2812	0.2771	0.2758
0.50	0.4159	0.4139	0.4035	0.4009	0.4000
1.00	0.4638	0.4547	0.4484	0.4469	0.4464
5.00	0.4985	0.4982	0.4979	0.4979	0.4979
∞	0.5000	0.5000	0.5000	0.5000	0.5000

From our numerical experience, we conclude that for large m (e.g., $m \geq 64$), it suffices to use the simple normal approximation based on $x^*(T)$ in (3.13); for moderate m it is preferable to use the normal fixed point pair (μ, σ^2) based on (3.10)–(3.12); and for very small m (e.g., for $m \leq 4$), it may be better not to use the normal approximation. We can interpolate from Tables 1 and 2 to obtain good estimates of the pair $(\mu, \sqrt{m}\sigma)$ for any m and T.

Given the explicit RBM cdf formula in (4.9), it is also possible to approximately describe the distribution of the number of jobs that must be moved away from any one queue, say $J_{1\rho}$, when ρ and m are suitably large:

$$P\left(J_{1\rho} \geq \frac{\rho(c_a^2 + c_s^2)x}{1-\rho}\right) \approx P(R(T) > x^*(T) + x | R(0) = x^*(T)) \ , \tag{4.26}$$

Table 2
A comparison between $\sqrt{m}\sigma$, the approximate standard deviation of the steady-state content of each queue just before load balancing with m independent RBM processes, using the normal iteration, and the approximation $\sqrt{V(T, x^*(T))}$.

	$\sqrt{m}\sigma$ from normal iteration				
T	$m = 2$	$m = 4$	$m = 16$	$m = 64$	$\sqrt{V(T, x^*(T))}$
0.01	0.1105	0.0964	0.0881	0.0864	0.0858
0.05	0.2076	0.1842	0.1713	0.1686	0.1677
0.10	0.2597	0.2354	0.2213	0.2182	0.2172
0.50	0.3810	0.3719	0.3608	0.3581	0.3572
1.00	0.4383	0.4271	0.4194	0.4176	0.4170
5.00	0.4967	0.4962	0.4957	0.4956	0.4956
∞	0.5000	0.5000	0.5000	0.5000	0.5000

with the right side being computed by (4.9) after obtaining the fixed point $x^*(T)$.

We now show that the second moment grows during the interval between balancing.

Theorem 23. For periodic load balancing of RBM with $m \to \infty$, in steady state (starting from a fixed point $x^*(T)$, the second moment $M_2(t, x^*(T))$ is increasing in t in the interval $(0, T)$. The derivative of the variance as a function of time is

$$V'(t, x) = 1 - M(t, x)g(0; t, x) \ , \tag{4.27}$$

where $g(0; t, x)$ is the density of the cdf in (4.9) evaluated at 0, i.e.,

$$g(0; t, x) = \frac{1}{\sqrt{t}}\phi\left(\frac{x - t}{\sqrt{t}}\right) + 2e^{-2y}\Phi\left(\frac{-x + t}{\sqrt{t}}\right) + \frac{e^{-2y}}{\sqrt{t}}\phi\left(\frac{-x + t}{\sqrt{t}}\right) \ . \tag{4.28}$$

Proof. By Theorem 8.3 of [1], $M_2(t, x^*(T))$ is strictly increasing in the interval $(0, T)$ because $x(T) < 1/2$. Combining Theorems 8.1 and 8.3 of [1], we obtain

$$\begin{aligned} V'(t, x) &= M_2'(t, x) - 2M(t, x)M'(t, x) \\ &= 1 - 2M(t, x)(1 + M'(t, x)) \\ &= 1 - M(t, x)g(0; t, x) \ . \end{aligned}$$

□

We now give the asymptotic form for the RBM fixed-point $x^*(T)$ as $T \to \infty$. We write $f(x) \sim g(x)$ as $x \to \infty$ if $f(x)/g(x) \to 1$ as $x \to \infty$.

Theorem 24. For periodic load balancing with canonical RBM,

$$\frac{1}{2} - x^*(T) \sim \frac{e^{(1-T)/2}}{\sqrt{2\pi T^3}} \quad \text{as} \quad T \to \infty \; . \tag{4.29}$$

Proof. This follows directly from the asymptotic form of the mean given in Corollary 1.1.2(a) in Abate and Whitt [1], in particular,

$$\frac{1}{2} - M(t,x) \sim \frac{2(1-x)}{\sqrt{2\pi t^3}} e^{x-t/2} \quad \text{as} \quad t \to \infty \; , \tag{4.30}$$

noting that $x^*(T) \to 1/2$ as $T \to \infty$. □

5. Performance Comparisons

In this section we apply the diffusion approximation in Section 4 to make comparisons between load balancing and two natural alternatives: m separate s-server queues and 1 combined ms-server. For simplicity, we now focus on the case of M/M/1 queues, so that $s = 1$. (The advantage of resource sharing is larger when the systems being combined have fewer servers.) We develop approximations for the distribution of the steady-state number of jobs in the system per server with each scheme. We display our conclusions in Table 3. As indicated in Section 1, the differences can be great.

Table 3
Approximations for the distribution of the steady-state number of jobs in the system per server just after load balancing. (The parameters γ_1 and γ_2 are constants less than 1.)

scheme	distribution	mean	standard deviation
m separate M/M/1 queues	exponential	$\frac{\rho}{1-\rho}$	$\frac{\rho}{1-\rho}$
a single M/M/m queue	normal	ρ	$\frac{\rho}{\sqrt{m}}$
m M/M/1 queues with load balancing	normal	$\gamma_1 \left(\frac{\rho}{1-\rho}\right)$	$\frac{\gamma_2}{\sqrt{m}} \left(\frac{\rho}{1-\rho}\right)$

Intuitively, it is evident that load balancing can achieve both alternatives as well as a range of performance behavior in between. Clearly, if the balancing interval T_ρ is very short, then load balancing is the same as the combined M/M/m system. Indeed, for sufficiently small T_ρ, periodic load balancing outperforms joining the shortest queue. On the other hand, if the balancing interval T_ρ is very large, then except after the infrequent balancing times, the queues behave like separate M/M/1 queues. We focus on the intermediate case, which can be characterized by the scaling in (4.4) as $\rho \to 1$.

Using heavy-traffic diffusion approximations, as described at the beginning of Section 4, we conclude that the steady-state number of jobs in a single M/M/1 queue for suitably high traffic intensity ρ has approximately an exponential distribution with mean (and thus also standard deviation) $\rho/(1-\rho)$.

For any fixed ρ, when m is suitably large, a single M/M/m queue behaves like an infinite-server queue. Thus the steady-state number of jobs in an M/M/m queue with traffic intensity ρ and suitably large m has approximately a Poisson distribution with mean (and thus variance) $m\rho$. (More elaborate approximations were described in Section 1.) The Poisson distribution in turn can be approximated by a normal distribution. The steady-state number of jobs per server in an M/M/m queue is the steady-state number in the system divided by m. Thus, the steady-state number of jobs per server in an M/M/m system is approximately normally distributed with mean ρ and standard deviation $\rho/\sqrt{m}$.

From the above analysis, we see that under heavy loads the multi-server system has a much smaller mean per server than the simple-server queue because of the factor $1-\rho$ in the denominator (ρ versus $\rho/(1-\rho)$). The chance of large values above the mean is also much smaller in the multi-server queue. First, the tail of a normal distribution decays more rapidly than the tail of an exponential distribution. Second, the standard deviation in the multi-server queue has the extra factor $\sqrt{m} > 1$ in the denominator, while the standard deviation in the single-server queue has the extra factor $(1-\rho) < 1$ in the denominator.

Now we consider the case of load balancing, where the balancing intervals T_ρ in the queues are chosen consistently with the scaling in (4.4) for some reasonable T, e.g., with $.02 < T < 2$. Our analysis in Section 4 leads us to conclude that the steady-state number of jobs in one queue after load balancing has approximately a normal distribution with mean $\gamma_1\rho/(1-\rho)$ and standard deviation $\gamma_2\rho/(1-\rho)\sqrt{m}$ for some constants γ_1 and γ_2. We draw this conclusion because the scaling in the heavy-traffic limit theorem in (4.6) is the same as in the heavy-traffic limit theorem for a single M/M/1 queue. For a single M/M/1 queue, the steady-state number after normalization is approximated by the exponentially-distributed random variable $R(\infty)$. Thus the constant γ_1 is the ratio of the realized mean, approximately $x^*(T)$, to the mean $ER(\infty) = 1/2$; i.e., $\gamma_1 = 2x^*(T) < 1$. Similarly, the variance after normalization is approximately $V(T, x^*(T))/m$ instead of $V(\infty, x) = 1/4$, so that $\gamma_2 = 2\sqrt{V(T, x^*(T))} < 1$ (see Figure 4).

More formally, we can conclude that the ratio of the two steady-state means in the load-balancing case to the separate-single-server case converges to $2x^*(T)$ as $\rho \to 1$, when the balancing intervals T_ρ grow as in (4.4). In contrast, if the load balancing intervals T_ρ where fixed independent of ρ, then the ratio would converge to 0, as noted in (4.20). Indeed if T is suitably small, then the constants γ_1 and γ_2 can be $1-\rho$, so that load balancing can perform just as well as the multi-server queue. On the other hand, if T is large, then there remains a benefit for load balancing in the standard deviation. However, the distribution just before load balancing is then approximately the same as in a single-server queue.

In summary, in what we regard as the typical case (consistent with the scaling in (4.4) with high ρ and large m), load balancing provides a modest gain over separate M/M/1 queues in the mean by a factor $2x^*(T)$ and a substantial gain in the standard deviation by a factor of $2\sqrt{V(T,x^*(T))}/\sqrt{m} \approx 1/\sqrt{m}$ and in the distribution — going from exponential to normal. Thus, we conclude that load balancing should be very effective for reducing the likelihood of large queue lengths. This conclusion will be substantiated by the simulation results in the next section.

6. Comparisons with Simulations

In this section we compare the RBM approximations developed in Section 4 to simulations. We first simulated m M/M/1 queues coupled by periodic load balancing for a range of values of m and ρ. To dramatically show the advantage of the heavy-traffic limit and associated scaling in Section 4, we scale so that each is to be approximated by canonical RBM (drift -1, diffusion coefficient 1).

For the results we will display, we start by picking a single time point for canonical RBM, $T = 1.0$. We then choose balancing times T_ρ as a function of ρ to satisfy (4.15). Since we are considering M/M/1 queues, $s = c_a^2 = c_s^2 = 1$ and

$$T_\rho = \frac{2T}{(1-\rho)^2} = \frac{2}{(1-\rho)^2} \ . \tag{6.1}$$

We first consider the case $m = 64$ for three values of $\rho : \rho = 0.8, 0.9$ and 0.95. For $\rho = 0.8$, 0.9 and 0.95, $T_\rho = 50$, 200 and 800, respectively. For each value of ρ, the simulation was based on three independent replications of 64×10^6 arrivals (10^6 arrivals per queue). The histograms of the normalized queue lengths just after redistribution, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/2\rho$, are displayed for $\rho = 0.8$, 0.9 and 0.95 in Figure 5. (When plotted, the histograms for the three replications were barely distinguishable, demonstrating that the run length was more than adequate to achieve high statistical precision.) Since the scaling was applied, the RBM fixed point $x^*(1) = 0.446$ becomes the initial approximation to the normalized number at each queue after balancing. A second refined approximation is the normal approximation

$$\frac{(1-\rho)}{2\rho} N_{i\rho}^{(m)}(nT_\rho) \approx N(x^*(T), V(T,x^*(T))/m) \ . \tag{6.2}$$

These two approximations are also shown in Figure 5. From Figure 5, we see that the two RBM approximations perform quite well, with both slightly overestimating the true distributions. Convergence toward the approximations as $\rho \to 1$ is also evident. For smaller values of ρ, the queue lengths tend to be very small, and the heavy-traffic approximation is not very accurate.

A third approximation is the normal approximation $N(\mu, \sigma^2)$, where the pair (μ, σ^2) are obtained by iteratively solving the equations (3.10)–(3.11) using the RBM conditional mean and variance functions $M(t, x)$ and $V(t, x)$ in (4.13), (3.12) and (4.25). However, as shown in Tables 1 and 2, the fixed point (μ, σ^2) of the normal iteration agrees closely with the pair $(x^*(T), V(T, x^*(T))/m)$ in this case. The differences present in Figure 5 thus seem to primarily represent the error in the heavy-traffic approximation.

Next, to describe the dependence upon m, we consider the cases of $m = 4$, 16 and 64 with $\rho = 0.95$ for the same case $T = 1.0$. The deterministic fixed point $x^*(T)$ is again 0.446. The sample means of the normalized queue lengths after load balancing when $m = 4$, 16 and 64 were 0.4274, 0.4210 and 0.4209, respectively.

To describe the rest of the distribution beyond the mean, we display in Figure 6 histograms of the normalized and centered variables, $\sqrt{m}[(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/2\rho - \bar{n}_{i\rho}^{(m)}]$, where $\bar{n}_{i\rho}^{(m)}$ is the sample mean of $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/2\rho$ given above. We add the factor $\sqrt{m}$ so that three cases should have approximately the same variance $V(T, x^*(T))$ using the normal approximation in (6.2). The estimated sample standard deviations for $m = 4$, 16 and 64 were 0.4440, 0.4279 and 0.4434, respectively, while $\sqrt{V(1, x^*(1))} = 0.4170$.

Finally to consider non-Markovian queues, we consider M/G/1 queues with a Pareto service-time distribution. We let the service-time complementary cdf have the specific form

$$G^c(t) = (1 + bt)^{-\alpha}, \quad t \geq 0 , \tag{6.3}$$

where $b = 1/(\alpha - 1)$ to give the distribution mean 1. The associated SCV is

$$c_s^2 = 1 + 2\left(\frac{(\alpha-1)^2}{\alpha-2} - \alpha\right) . \tag{6.4}$$

To keep within the heavy-traffic limit framework in Section 4, we need $\alpha > 2$, so that $c_s^2 < \infty$. In particular, we choose $\alpha = 3$, which makes $c_s^2 = 3$. We then scale as in (4.15), so that

$$T_\rho = \frac{(c_a^2 + c_s^2)T}{(1-\rho)^2} = \frac{4}{(1-\rho)^2} . \tag{6.5}$$

When we balance, we do not move the customers in service, so that all customers have their original service times. We then consider the normalized queue lengths just before redistribution, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho-)/4\rho$, as in (4.6). We compare the M/G/1 Pareto and exponential service-time-distribution cases with $\rho = 0.95$, $T = 1.0$, and $m = 64$ and 4 in Figures 7 and 8. The Pareto and exponential cases were scaled differently, so that the approximation for both involves canonical RBM. In Figures 7 and 8 we include the normal approximation $N(\mu, m\sigma^2) \approx N(x^*(1), V(1, x^*(1)))$. As should be expected, the normal approximation is more accurate for $m = 64$ than for $m = 4$. The close agreement between the exponential and Pareto simulation results in both cases shows the remarkable power of the heavy-traffic scaling.

7. Redistributing Work with a Work-Conserving Discipline

In Sections 2–4 we assumed that each queue has s-servers and uses the FCFS service discipline. However, for computer system applications it is usually more appropriate to assume a single server with the round robin (RR) or processor-sharing (PS) discipline. Indeed, these disciplines are traditionally used in the dynamic load balancing literature, e.g., see Harchol-Balter and Downey [21] and references cited there. In this section we discuss periodic load balancing with such alternative service disciplines.

First, one might use the results in Sections 2–4 as approximations for these other service disciplines. This can be motivated by the fact that the PS discipline has the insensitivity property. In particular, in the M/G/1 (PS) model the steady-state queue length distribution has the same geometric distribution as in the M/M/1 (FCFS) model (with the same interarrival time and service time means). However, with periodic load balancing, we apply the transient distribution over intervals of length T, not the steady-state distribution. Unfortunately, the transient distributions do not have this insensitivity property. Nevertheless, as a rough approximation, it should be reasonable to use the M/M/1 (FIFO) model as an approximation for periodic load balancing with M/G/1 (PS) queues. For the RBM approximation in Section 4, the resulting variability parameters are $c_a^2 = c_s^2 = 1$. Using the insensitivity logic, we would apply the M/M/1 FIFO results to any service-time distribution having a finite mean.

It is also of interest to consider redistribution of remaining work (in required service time) instead of jobs. Work redistribution is directly applicable in systems where the full service requirements are known in advance, and jobs can be split up with pieces sent to different queues. More generally, a work redistribution model is interesting as a lower bound on what can be achieved by other periodic load balancing algorithms.

If we focus on periodic redistributions of work in single-server queues, then the behavior is the same for any work-conserving discipline. In particular, then the behavior is the same for RR, PS and FCFS. Moreover, it is known that the heavy-traffic limit for the workload process in a G/GI/s queue coincides with the heavy-traffic limit for the queue-length process, providing that the mean service time is 1. By similar reasoning, the same limit involving RBM with periodic load balancing holds for the workload process. Hence, the approximation in Section 4 applies directly to the workload process with periodic load balancing of remaining work in G/GI/1 queues with a work-conserving service discipline. Direct heavy-traffic limits for the workload process in a single-server queue are contained in Whitt [42]. We state a workload limit theorem in Section 9 that also covers additional extra long service times.

8. Reducing the Likelihood of Severe Congestion

One of the main goals of load balancing is to reduce the likelihood of large queue lengths or large workloads. To show that periodic load balancing achieves this goal, we show that the tail probabilities of conditional RBM $P(R(t) > y|R(0) = x)$ decay more rapidly than the exponential steady-state. (Recall that $P(R(\infty) > y) = e^{-2y}$.) In fact, we show that the RBM conditional tail probability is of order $e^{-y^2/2t}$ as $y \to \infty$.

Theorem 25. The RBM conditional tail probability satisfies

$$P(R(t) > y|R(0) = x) \sim \alpha(y,x,t)e^{-\left(\frac{y^2}{2t}+y\left(1-\frac{x}{t}\right)\right)} \quad as \quad y \to \infty \ , \tag{8.1}$$

where

$$\begin{aligned}\alpha(y,x,t) &= \frac{1}{\sqrt{2\pi}}e^{-(x-t)^2/2t}\left(\frac{\sqrt{t}}{y-x+t}+e^{-2yx/t}\left(\frac{\sqrt{t}}{y+x-t}\right)\right) \\ &\sim \frac{1}{y}\sqrt{\frac{t}{2\pi}}e^{-(x-t)^2/2t} \quad for\ x > 0\ ,\end{aligned} \tag{8.2}$$

so that

$$\left(\log P(R(t) > y|R(0) = x) - \frac{y^2}{2t}\right) \sim y\left(1-\frac{x}{t}\right) \quad as \quad y \to \infty\ . \tag{8.3}$$

Proof. Use (4.9) with the asymptotic relation $\Phi(-y) \sim y^{-1}\phi(y)$ as $y \to \infty$; see p. 175 of Feller [16]. □

Formulas (8.1)–(8.3) show that the RBM conditional tail probability decays rapidly (of order $e^{-y^2/2t}$) if x and t are not large. We have seen that the fixed points must satisfy $x^*(T) < ER(\infty) = 1/2$, so that $x^*(T)$ will not be large. However, T_x^* increases as $x \to 1/2$, so that T_x^* can be large. If we keep $x^*(T)$ well below 1/2, then we will not encounter large values of t, and the system behavior should be well described by Theorem 25.

However, if T is allowed to grow, then the control of large queue lengths and workloads weakens. To describe the behavior for larger t, we consider the limit as $y \to \infty$ and $t \to \infty$ with $y = ct$ or $(y-t)/\sqrt{t} \to c$. The following theorem shows that conditional tail probabilities decay more slowly in this regime (but still more rapidly than the steady-state tail probabilities).

Theorem 26. (a) If $t \to \infty$ with $y = ct$ for $c > 1$, then

$$P(R(t) > y | R(0) = x) \sim \alpha(x,t) e^{-\frac{(c+1)^2}{2} t} , \tag{8.4}$$

where

$$\begin{aligned}\alpha(x,t) &= \frac{1}{\sqrt{2\pi}} e^{-(\frac{x^2}{2t} - (c+1)x)} \left(\frac{\sqrt{t}}{(c+1)t - x} + e^{-2cx} \frac{\sqrt{t}}{(c-1)t + x} \right) \\ &\sim \frac{1}{\sqrt{2\pi t}} \left(\frac{e^{(c+1)x}}{c+1} + \frac{e^{(1-c)x}}{c-1} \right) .\end{aligned} \tag{8.5}$$

(b) If $t \to \infty$ with $y = ct$ for $c < 1$, then

$$P(R(t) > y | R(0) = x) \sim e^{-2y} . \tag{8.6}$$

(c) If $t \to \infty$ with $(y - t)/\sqrt{t} \to c$, then

$$P(R(t) > y | R(0) = x) \sim e^{-2y} \Phi(-c) . \tag{8.7}$$

Proof. As in Theorem 25, we apply the asymptotic relation $\Phi(-y) \sim y^{-1}\phi(y)$ as $y \to \infty$. To have $\Phi(-y)$ with $y \to \infty$ in both terms of (4.9), we need $y = ct$ with $c > 1$. Cases (b) and (c) of Theorem 26 follow directly from (4.9). □

In this section we have considered the limits $\rho \to 1$ and $y \to \infty$ in that order. If instead we fixed $\rho < 1$ and let $y \to \infty$, then the asymptotic behavior depends on more of the fine structure of the queueing system. The transient workload will have a tail decaying no more rapidly than the service-time distribution. (Consider the case of a single arrival in the interval $(0, t)$.)

9. Exceptionally Long Service Times

In addition to requiring heavy loads, the RBM approximation requires that the job arrival and service processes be not too bursty. The RBM approximation depends critically on the sums of the interarrival times and service times converging to standard normal distribution after the usual $\sqrt{n}$ normalization, e.g., as in (4.1). For i.i.d. service times, this means that the service-time distribution must have a finite second moment.

However, measurements of computer systems by Leland and Ott [28] and Harchol-Balter and Downey [21] indicate that service requirements may often come from long-tail distributions, without a finite second moment. Indeed, Harchol-Balter and Downey found that the cdf $1 - c/t$ is often appropriate. As they indicate, this distribution has no mean. We first point out that such a distribution rules out conventional steady-state queueing analysis, with or without load balancing. With the standard models having unlimited waiting room, a service time with an infinite mean implies that the queue length and workload processes will diverge to $+\infty$ with probability one as time evolves, e.g., see Theorem 8 on p. 18 of Borovkov [10]. Hence, in that context it makes no sense to talk about long-run average performance. Thus, there can be no counterpart to the fixed-point equations in Sections 3 and 4. However, it is of course possible to use transient analysis, but then some care should be given to formulating realistic initial conditions.

In this section we briefly indicate some possible approaches to represent unusually long service times. First, a similar heavy-traffic limit theorem can be obtained when the arrival and service processes are bursty. Then these processes may converge to other processes, such as stable processes, with a different normalization. For example, instead of (4.1) we might have

$$\frac{A(nt) - \lambda nt}{n^{1/\alpha}} \Rightarrow S_\alpha(t) \quad \text{as } n \to \infty$$

where $0 < \alpha < 2$ and $\{S(t) : t \geq 0\}$ is a stable process with index α. As indicated in Whitt [43] the heavy-traffic limit theorems easily extend to these different normalizations. The difficulty for our application to periodic load balancing is obtaining useful descriptions of the transient behavior of these alternative limit processes, i.e., analogs of (4.9) and (4.13) here.

A promising alternative approach to rare exceptionally long service times is to apply the reasoning used to establish the heavy-traffic limit for queues with rare long server vacations in Kella and Whitt [25]. The limit here as $\rho \to 1$ also applies with such additional long service times. Instead of the limit process in Section 4, we obtain a limit process that is a reflection of Brownian motion plus an extra jump process.

We now present the framework for this alternative limit theorem. We do so for the workload in the setting of Section 7. After the theorem, we indicate how it can be applied to generate alternative approximations, which do not require that the service times have finite second moments.

We modify the setting of Section 4 to allow for additional rare long service times. Consider a single queue. Let the arrival time of the n^{th} special service time in the system with traffic intensity ρ occur at time $U_{\rho n}$, where

$$(1-\rho)^2(U_{\rho 1}, \ldots, U_{\rho n}) \Rightarrow (U_1, \ldots, U_n) \quad \text{in } \mathbb{R}^n \text{ as } \rho \to 1 \text{ for each } n \ . \tag{9.1}$$

Let $\{C_\rho(t) : t \geq 0\}$ and $\{C(t) : t \geq 0\}$ be the counting processes associated with $\{U_{\rho n}\}$ and $\{U_n\}$, e.g.,

$$C_\rho(t) = \max\{n : U_{\rho n} \leq t\} \ , t \geq 0 \ , \tag{9.2}$$

where $U_{\rho 0} = 0$. Let the n^{th} special service time in system ρ be $V_{\rho n}$, where

$$(1-\rho)(V_{\rho 1}, \ldots, V_{\rho n}) \Rightarrow (V_1, \ldots, V_n) \quad \text{in } \mathbb{R}^n \text{ as } \rho \to 1 \text{ for each } n \ . \tag{9.3}$$

The scaling in (9.1) and (9.2) is explained by the fact that time is scaled by $(1-\rho)^2$ while space is scaled by $(1-\rho)$ in the usual heavy-traffic limit theorem, i.e., as in (4.2). The associated total input process of special work in system ρ is

$$I_\rho(t) = \sum_{i=1}^{C_\rho(t)} V_{\rho i} \ , \quad t \geq 0 \ . \tag{9.4}$$

We are now ready to state a limit theorem.

Theorem 27. Consider m G/GI/1 queues with work-conserving service disciplines, controlled by periodic redistribution of remaining work, as in Section 7. Let the basic arrival and service processes satisfy the assumptions of Theorem 4.1. Assume that the redistribution intervals satisfy (4.4). Assume that the initial workloads satisfy (4.5). Assume that extra long service times arrive independently according to the input process $I_\rho(t)$ in (9.4), satisfying (9.1)–(9.3). Then the individual workload processes satisfy

$$\frac{(1-\rho)}{(c_a^2+c_s^2)} W_{1\rho}^{(m)}(t(c_a^2+c_s^2)/(1-\rho)^2) \Rightarrow X(t) \ \ in \ \ D \tag{9.5}$$

as first $\rho \to 1$ and then $m \to \infty$, where $x_n \equiv X(nT)$ evolves deterministically as

$$x_{n+1} = f_T(x_n) \tag{9.6}$$

with $f_T(x) = M(T,x)$, where

$$M(t,x) = E[Y(t)|Y(0)=x] \tag{9.7}$$

$$Y(t) = R_x(Z)(t) \tag{9.8}$$

$$Z(t) = B(t) - t + \sum_{i=1}^{C(t(c_a^2+c_s^2))} V_i/(c_a^2+c_s^2) \ , \tag{9.9}$$

$\{B(t) : t \geq 0\}$ is standard (drift 0, diffusion coefficient 1) Brownian motion and R_x is the reflection map, defined by

$$R_x(z)(t) = \max\{x + z(t) \ , \ \ z(t) - \inf_{0\leq s\leq t} z(s)\} \ . \tag{9.10}$$

If in addition $V_n, n \geq 1$, are i.i.d. and $\{C(t) : t \geq 0\}$ is a Poisson process, then

$$X(nT+t) \stackrel{d}{=} (Y(t)|Y(0)=x_n) \ , \quad 0 \leq t \leq T \ . \tag{9.11}$$

Proof. First, it is elementary (using basic properties of the function space D [9], [43]) that

$$\frac{(1-\rho)}{c_a^2+c_s^2} I_\rho(t(c_a^2+c_s^2)/(1-\rho)^2) \Rightarrow \sum_{i=1}^{C(t(c_a^2+c_s^2))} V_i/(c_a^2+c_s^2) \quad \text{in } D \text{ as } \rho \to 1 \ ,$$

because the limit process has finitely many jumps in a bounded interval and, by (9.1) and (9.3), the normalized times and sizes of the jumps converge. The rest of the proof is a workload analog of that in Theorem 18, closely paralleling Theorem 3.1 of Kella and Whitt [25]. (The result here is actually somewhat more elementary. Since there is a single jump at each discontinuity point, it is not necessary to use the M_1 topology here.) The limiting net input process between redistributions in (9.9) is Brownian motion minus t plus the jump process, just as in (3.3) of Kella and Whitt [25]. Finally, (9.11) holds under the extra conditions, because then the net input process $\{Z(t) : t \geq 0\}$ has stationary independent increments. □

We now indicate how we can apply Theorem 27 to generate approximations for long-tail service-time distributions. If the service times are i.i.d. with c.d.f G, then we can truncate the distribution at some large value z. We then let a new basic service-time distribution have cdf $H(x) = G(x)/G(z)$, $0 \leq x \leq z$. Then with probability $G^c(z)$ each arrival has a special service time with cdf $F(x) = G(x)/G^c(z)$, $x > z$ and $F(z) = 0$, and with probability $G(z)$ there is no extra service time. The extra arrival process of additional service times is a thinned version of the original arrival process. As the truncation point z increases, the selection probability $G^c(z)$ decreases and the thinned process approaches a Poisson process independent of the original arrival process, e.g., see Çinlar [12] or Serfozo [35]. Even if the original service time cdf G had no finite moments, the truncated cdf H has all moments finite.

Based on (9.5), we can use the approximation

$$W_{1\rho}^{(m)}(t) \approx \frac{(c_a^2+c_s^2)}{1-\rho} X(1-\rho)^2 t/(c_a^2+c_s^2)) \ , \tag{9.12}$$

for X characterized in (9.6)–(9.11), where ρ and c_s^2 are the traffic intensity and service-time SCV based on the truncated cdf H.

It remains to specify the jump process. At traffic intensity ρ, each interarrival time has mean $1/\rho$. The number of arrivals between each special arrival is geometrically distributed with mean $1/G^c(z)$. Hence, $EU_{\rho 1} = 1/\rho G^c(z)$, so that with (9.1) we let

$$EU_1 = \frac{(1-\rho)^2}{\rho G^c(z)} \ . \tag{9.13}$$

The rate of the Poisson process $\{C(t) : t \geq 0\}$ is $1/EU$. Thus, the rate of the Poisson process $\{C(t(c_a^2 + c_s^2)) : t \geq 0\}$ is $\rho(c_a^2 + c_s^2)G^c(z)/(1-\rho)^2$. Similarly, the limiting special service time can be obtained from the cdf F, making appropriate adjustments for the scaling in (9.2). By (9.2), we let $P(V \leq x) = P(V_\rho \leq x/(1-\rho))$. Hence

$$P\left(\frac{V}{c_a^2 + c_s^2} \leq x\right) = F((c_a^2 + c_s^2)x/(1-\rho)) \ . \tag{9.14}$$

As a consequence of (9.13) and (9.14), the limiting jump process

$$\sum_{i=1}^{C(t(c_a^2+c_s^2))} \frac{V_i}{(c_a^2 + c_s^2)}$$

has drift $\rho G^c(z)m(F)/(1-\rho)$, where $m(F)$ is the mean of the cdf F. (We require that $m(F) < \infty$ to have finite drift.) Thus, the net input process $\{Z(t) : t \geq 0\}$ in (9.9) has drift

$$EZ(1) = \frac{\rho G^c(z)m(F)}{1-\rho} - 1 \ . \tag{9.15}$$

It is elementary to show that a proper steady state exists for the approximating process $\{X(t) : t \geq 0\}$ if and only if $EZ(1) < 0$.

The remaining problem for applications is to compute the mean function $M(t,x)$ in (9.7). We suggest an approximation based on choosing the Poisson rate to be small compared with the redistribution interval length T. (This is achieved in the approximation following Theorem 27 by making the truncation point z sufficiently high.) As an approximation, we can then assume that there is at most one special service time in each redistribution interval. To be more specific, let γ be the rate of the Poisson process, which is $\rho(c_a^2 + c_s^2)G^c(z)/(1-\rho)^2$ with the truncation at z. Let $\hat{V}$ be distributed as $V/(c_a^2 + c_s^2)$, which has the cdf $F((c_a^2 + c_s^2)x/(1-\rho))$ as in (9.14) with the truncation at z. Then

$$\begin{aligned} M(T,x) = {} & e^{-\gamma T}\tilde{M}(T,x) \\ & + \frac{(1-e^{-\gamma T})}{T} \int_0^T \int_0^\infty \int_0^\infty \tilde{M}(T-y, z+v) dP(\hat{V} \leq v) \\ & \times P(R(y) = dz | R(0) = x) dy \ , \end{aligned} \tag{9.16}$$

where $\tilde{M}(t,x)$ is the RBM mean function in (4.13). The first term in (9.16) corresponds to no special arrival, while the second term corresponds to at least one special arrival, which we treat as exact one. Conditional on one Poisson arrival, it is uniformly distributed over $[0,T]$. Note that, by (4.13) and (4.9), the integrand in the second term of (9.16) can be expressed in closed form. Hence, the calculation in (9.16) can be performed. We suggest applying (9.16) to analyze the behavior with long-tail service-time distributions. Recall that the drift $EZ(1)$

in (9.15) must be negative in order for a proper steady state to exist (and the fixed point equation $x = f_T(x)$ to have a solution).

If $\hat{V}$ is large compared to T, we might use the alternate approximate

$$M(T,x) \approx e^{-\gamma T}\tilde{M}(T,x) + (\gamma T)EV \ . \tag{9.17}$$

for

$$\gamma = \frac{\rho G^c(z)(c_a^2 + c_s^2)}{(1-\rho)^2} \quad \text{and} \quad E\hat{V} = \frac{(1-\rho)}{c_a^2 + c_s^2} m(F) \ . \tag{9.18}$$

Approximation (9.17) ignores the RBM component in the second term.

10. Periodic Load Balancing in Unbalanced Systems

In our analysis so far, we have assumed that the queues are homogeneous. However, in practice, the arrival processes and service-time distributions may be different at different queues. As an extreme case, service may be temporarily unavailable at some queues, e.g., because of queue failure. Periodic load balancing provides a way to address this problem without having to know which queues are down.

In this section we consider periodic load balancing when a proportion p of the queues are down during each redistribution interval. We assume that we do not know which queues are down. We thus redistribute to all queues. Down queues generate substantial congestion, because we assume that arrivals come to all queues in i.i.d. arrival processes. Thus there are arrivals but no service completions at down queues.

During the interval between redistributions, the number of jobs at a down queue grows like the arrival process there. Given that the arrival process satisfies a FCLT as in (4.1), that FCLT describes the growth of the queue length process as the length of the redistribution interval grows. If we consider a sequence of models in which the redistribution intervals grow, then the FCLT for the arrival process describes the behavior at the down queue.

It is significant that the presence of down queues alters the form of the heavy-traffic limit theorems in Section 4. Even as $\rho \to 1$ in the up queues, the growth of jobs or work within each redistribution interval tends to be dominated by the queues that are down. A limit holds with the law-of-large-numbers scaling instead of the central-limit-theorem scaling.

Theorem 28. Consider the setting of Theorem 17 modified by having a proportion p of the queues down (for all time). Let 1 index an up queue and 2 index a down queue. If $\rho \to 1$ and $m \to \infty$ with (4.4) and (4.5) holding, then

$$\frac{(1-\rho)^2}{c_a^2 + c_s^2} N_{i\rho}^{(m)}(t(c_a^2 + c_s^2)/s(1-\rho)^2) \Rightarrow X_i(t) \quad in \quad D \quad for \quad i = 1,2 \tag{10.1}$$

where $x_n \equiv X_1(nT) = X_2(nT)$ evolves deterministically as $x_{n+1} = f_T(x_n)$ with

$$f_T(x) = p(x+T) \ , \tag{10.2}$$

while

$$X_1(kT+t) = 0 \ , \quad 0 < t < T \ , \tag{10.3}$$

and

$$X_2(kT+t) = x_k + t \ , \quad 0 < t < T \ . \tag{10.4}$$

Proof. Given (4.1) and (4.4), it follows that

$$\frac{(1-\rho)^2}{(c_a^2+c_s^2)} A_\rho(T_\rho -) \Rightarrow T \quad \text{as } \rho \ \rightarrow \ 1 \ . \tag{10.5}$$

Given the space normalization by $(1-\rho)^2$ in (10.1) instead of by $(1-\rho)$ in (4.5), the queue lengths in the up queues are asymptotically negligible. □

The behavior of the function f_T in (10.2) is elementary.

Theorem 29. The function f_T in (10.2) has a unique fixed point $x^*(T) = pT/(1-p)$ and $f_T^{(n)}(x_0) \ \rightarrow \ x^*(T)$ as $n \ \rightarrow \ \infty$ for each x_0.

In this unbalanced scenario the approximation is quite simple; we would approximate the normalized queue length after balancing, $(1-\rho)^2 N(kT_\rho)/(c_a^2+c_s^2)$ by the fixed point $pT/(1-p)$. Note that a large proportion of the jobs must be moved in each balancing though.

More interesting limiting behavior occurs if we assume that the proportion of down queues decreases with ρ, in particular, if $p_\rho = (1-\rho)\alpha$. Then both the up and down queues contribute to the limit behavior as $\rho \ \rightarrow \ 1$.

Theorem 30. In the setting of Theorem 28, if the proportion p of down queues is a function of ρ satisfying $p_\rho = (1-\rho)\alpha$, then

$$\frac{(1-\rho)}{c_a^2+c_s^2} N_{1\rho}^{(m)}(t(c_a^2+c_s^2)/s(1-\rho)^2) \Rightarrow X_1(t) \ \ in \ \ D \ \ for \ \ i = 1,2 \ , \tag{10.6}$$

where $x_n \equiv X_1(nT)$ evolves deterministically as $x_{n+1} = f_T(x_n)$ with

$$f_T(x) = M(T,x) + \alpha T \tag{10.7}$$

for $M(t,x)$ in (4.13) and

$$X_1(nT+t) \stackrel{d}{=} (R(t)|R(0) = x_n) \tag{10.8}$$

as in (4.14).

Remark. In the scaling of (10.6), the content of the individual down queues explodes in the limit. The nondegenerate limit in (10.6) is obtained because the proportion of down queues is asymptotically negligible.

Proof. Given that $p_\rho = (1-\rho)\alpha$,

$$\frac{p_\rho(1-\rho)}{(c_a^2 + c_s^2)} A_\rho(T_\rho) \Rightarrow \alpha T \quad \text{as } \rho \to 1$$

by (10.5). (However, as noted in the Remark above, the left side diverges for individual down queues.) Given that the initial level satisfies (4.5), after normalization, the totality of down queues maps x_0 into αT at time T. The up queues behave just as in Theorem 4.1. Given that $p_\rho = (1-\rho)\alpha$, the proportion of up queues approaches 1. □

We now characterize when the function (10.7) has a fixed point.

Theorem 31. The function f_T in (10.7) has a proper fixed point $x^*(T)$ if and only if $\alpha < 1$. If $\alpha \geq 1$, then $f_T^{(n)}(x_0) \to \infty$ for all x_0. If $\alpha < 1$, then equation $x = f_T(x)$ has a unique solution $x^*(T)$ and $f_T^{(n)}(x_0) \to x^*(T)$ as $n \to \infty$ for all x_0.

Proof. If $\alpha \geq 1$, then $f_T(x) > x$ for all x, so that there can be no fixed point. To see this, note that $M(T,x) > x - T$. Hence $f_T(x) > x - (1-\alpha)T$. Now suppose that $\alpha < 1$. For any ϵ with $0 < \epsilon < (1-\alpha)T$, there is an x sufficiently large that $M(T,x) < x - T + \epsilon$, so that

$$f_T(x) < x - (1-\alpha)T + \epsilon < x \; . \tag{10.9}$$

Moreover, the inequality holds for all higher x. Since f_T is continuous with $f_T(0) > 0$, there necessarily is a fixed point. The argument of Theorem 14 can be applied to derive the remaining properties. □

References

[1] J. Abate and W. Whitt, Transient behavior of regulated Brownian motion, I and II, *Adv. Appl. Prob.* 19 (1987) 560–631.
[2] J. Abate and W. Whitt, Transient behavior of the M/M/1 queue via Laplace transforms, *Adv. Appl. Prob.* 20 (1988) 145–178.
[3] J. Abate and W. Whitt, Calculating time-dependent performance measures for the M/M/1 queue, *IEEE Trans. Commun.* 37 (1989) 1102-1104.
[4] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems* 10 (1992) 5–88.
[5] J. Abate and W. Whitt, Numerical inversion of Laplace transforms of probability distributions, *ORSA J. Computing* 7 (1995) 36–43.
[6] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.
[7] F. Baccelli, and P. Brémaud, *Elements of Queueing Theory*, Springer-Verlag, New York, 1994.
[8] A. Barak, G. Shai and R. G. Wheeler, *The MOSIX Distributed Operating System: Load Balancing for UNIX*, Springer-Verlag, Berlin, 1993.
[9] P. Billingsley, *Convergence of Probability Measures*, Wiley, New York, 1968.
[10] A. A. Borovkov, *Stochastic Processes in Queueing Theory*, Springer-Verlag, New York, 1976.
[11] G. L. Choudhury, D. M. Lucantoni and W. Whitt, Multidimensional transform inversion with applications to the transient M/G/1 queue, *Ann. Appl. Prob.* 4 (1994) 719–740.
[12] E. Çinlar, Superpositions of point processes. In *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, P. A. W. Lewis (ed.) Wiley, New York, 1972, 549–606.
[13] J. L. Davis, W. A. Massey and W. Whitt, Sensitivity to the service-time distribution in the nonstationary Erlang loss model, *Management Science* 41 (1995) 1107–1116.
[14] D. L. Eager, E. D. Lazowska and J. Zahorjan, Adaptive load balancing in homogeneous distributed systems, *IEEE Trans. Software Eng.* 12 (1986) 662–675.
[15] S. N. Ethier and T. G. Kurtz, *Characterization and Approximation of Markov Processes*, Wiley, New York, 1986.
[16] W. Feller, *An Introduction to Probability Theory and its Applications*, vol. I, Wiley, New York, 1968.
[17] K. W. Fendick, V. R. Saksena and W. Whitt, Dependence in packet queues, *IEEE Trans. Commun.* 37 (1989) 1173–1183.
[18] G. J. Foschini, Unobtrusive communication of status in a packet network in heavy traffic, *AT&T Tech. J.* 64 (1985) 463–479.
[19] G. J. Foschini and J. Salz, A basic dynamic routing problem and diffusion, *IEEE Trans. Commun.* 26 (1978) 320–327.
[20] B. Hajek, Performance of global load balancing by local adjustment, *IEEE Trans. Inf. Thy.* 36 (1990) 1398–1414.
[21] M. Harchol-Balter and A. B. Downey, Exploiting process lifetime distributions for dynamic load balancing. *Proceedings SIGMETRICS '96*, 1996.
[22] G. Hjálmtýsson, Lightweight call setup — supporting connection and connectionless services. *Teletraffic Contributions for the Information Age, Proceedings 15th International Teletraffic Congress*, V. Ramaswami and P. E. Wirth (eds.), Elsevier, Amsterdam, 1997, pp. 35–45.
[23] G. Hjálmtýsson and K. K. Ramakrishnan, UNITE — An architecture for lightweight signalling in ATM networks, *IEEE Infocom '98*, 1998, to appear.
[24] D. L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic, I and II, *Adv. Appl. Prob.* 2 (1970) 150–177 and 355–369.

[25] O. Kella and W. Whitt, Diffusion approximations for queues with server vacations, *Adv. Appl. Prob.* 22 (1990) 706–729.
[26] J. Köllerström, Heavy traffic theory for queues with several servers, I, *J. Appl. Prob.* 11 (1974) 544–552.
[27] C. N. Laws, Resource pooling in queueing networks with dynamic routing, *Adv. Appl. Prob.* 24 (1992) 699–726.
[28] W. E. Leland and T. J. Ott, Load balancing heuristics and process behavior, *Sigmetrics 86*, 14 (1986) 54–69.
[29] A. Mandelbaum and M. I. Reiman, On pooling in queueing networks, *Management Sci.* (1996), to appear.
[30] W. A. Massey and W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, *Queueing Systems*, 25 (1997) 157–172.
[31] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, Springer-Verlag, New York, 1993.
[32] M. I. Reiman, Some diffusion approximate with state space collapse, in *Modelling and Performance Evaluation Methodology*, F. Baccelli and G. Fayolle (eds.), Springer-Verlag, 1984, pp. 209–240.
[33] M. H. Rothkopf and P. Rech, Perspectives on queues: combining queues is not always beneficial, *Operations Research* 35 (1987) 906–909.
[34] H. Sakasegawa, An approximate formula $L_q = \alpha\beta^{\rho}/(1-\rho)$, *Ann. Inst. Stat. Math.* 29 (1977) 67–75.
[35] R. F. Serfozo, Partitions of point processes: multivariate Poisson approximations, *Stoch. Proc. Appl.* 20 (1985) 281–294.
[36] D. R. Smith and W. Whitt, Resource sharing for efficiency in traffic systems, *Bell System Tech. J.* 60 (1981) 39–55.
[37] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Wiley, New York, 1983.
[38] M. R. Taaffe and K. L. Ong, Approximating Ph(t)/M(t)/S/C queueing systems, *Ann. Oper. Res.* 8 (1987) 103–116.
[39] L. Takács, *Introduction to the Theory of Queues*, Oxford University Press, New York, 1962.
[40] E. van Doorn, *Stochastic Monotonicity and Queueing Applications of Birth-Death Processes*, Springer-Verlag, New York, 1981.
[41] R. W. Weber, On the optimal assignment of customers to parallel servers, *J. Appl. Prob.* 15 (1978) 406–413.
[42] W. Whitt, Weak convergence theorems for priority queues: preemptive-resume discipline, *J. Appl. Prob.* 8 (1971) 74–94.
[43] W. Whitt, Some useful functions for functional limit theorems, *Math. Oper. Res.* 5 (1980) 67–85.
[44] W. Whitt, Comparing counting processes and queues, *Adv. Appl. Prob.* 13 (1981) 207–220.
[45] W. Whitt, Deciding which queue to join: some counterexamples, *Opns. Res.* 34 (1986) 55–62.
[46] W. Whitt, Planning queueing simulations, *Management Science* 35 (1989) 1341–1366.
[47] W. Whitt, Understanding the efficiency of multi-server service systems, *Management Sci.* 38 (1992) 708–723.
[48] W. Whitt, Approximations for the GI/G/m queue, *Production and Operations Management* 2 (1993) 114–160.
[49] M. H. Willebeck-LeMair and A. P. Reeves, Strategies for dynamic load balancing on highly parallel computers, *IEEE Trans. Parallel and Distrib. Systems* 9 (1993) 979–993.
[50] W. Winston, Optimality of the shortest line discipline, *J. Appl. Prob.* 14 (1977) 181–189.
[51] R. W. Wolff, An upper bound for multi-channel queues, *J. Appl. Prob.* 14 (1977) 884–888.

[52] H. Zhang, G. Hsu and R. Wang, Heavy traffic limit theorems for a sequence of shortest queueing systems, *Queueing Systems* 21 (1995) 217–238.
[53] S. Zhou, A trace-driven simulation study of dynamic load balancing. *IEEE Trans. Software Engr.* 14 (1988) 1327–1341.

Figure 1. The conditional mean of RBM, $M(t,x) \equiv E[R(t)|R(0) = x]$, as a function of t for several values of $x \equiv M(0,x)$.

Figure 2. The conditional RBM variance, $V(t,x) \equiv Var(R(t)|R(0) = x)$, as a function of t for several values of x (in two scales).

Figure 3. The RBM fixed point $x^*(T)$ as a function of T (in two scales).

Figure 4. The approximate variance for RBM with load balancing, $V(T, x^*(T)) \equiv Var(R(T)|R(0) = x^*(T))$, as a function of T.

Figure 5. A comparison between the RBM approximations and histograms of the normalized queue lengths after load balancing, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/2\rho$, in 64 M/M/1 queues for $\rho = 0.80$, 0.90 and 0.95 and T_ρ scaled from $T = 1.0$.

Figure 6. A comparison between the RBM approximations and histograms of the centered and normalized queue lengths after load balancing, $\sqrt{m}[(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/2\rho - \bar{n}_{i\rho}^{(m)}]$, in m M/M/1 queues with $\rho = 0.95$ for $m = 4$, 16 and 64 and T_ρ scaled from $T = 1.0$. The approximating normal density, for the RBM approximation is $N(0, V(1.0, x^*(1.0)))$.

Figure 7. A comparison between the RBM approximation and histograms of the normalized queue lengths after load balancing, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/(1+c_s^2)\rho$, in 64 M/G/1 queues with $\rho = 0.95$ and $T = 1$ for exponential ($c_s^2 = 1$) and Pareto ($\alpha = c_s^2 = 3$) service-time distributions.

Figure 8. A comparison between the RBM approximation and histograms of the normalized queue lengths after load balancing, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/(1+c_s^2)\rho$, in 4 M/G/1 queues with $\rho = 0.95$ and $T = 1$ for exponential ($c_s^2 = 1$) and Pareto ($\alpha = c_s^2 = 3$) service-time distributions.