Proceedings of the 37th IEEE
Conference on Decision & Control
Tampa, Florida USA • December 1998

Approximations for Periodic Load Balancing

Gísli Hjálmtýsson and Ward Whitt
AT&T Labs - Research, Florham Park, NJ 07932-0971
{gisli,wow}@research.att.com

Abstract

We consider multiple queues in parallel with unlimited waiting space to which jobs come either in separate independent streams or by assignment (either in random or round robin order) from a single stream. Resource sharing is achieved by periodically redistributing the jobs 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 heavy load and a large number of homogeneous queues. With these approximations, we show how performance depends on the assumed arrival pattern of jobs and the model parameters. We 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 [5]. 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. There is typically substantially more overhead with migration of jobs in process, but it has been shown to yield significant performance improvement [5].

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, 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 this study is to lend support for a notion of lightweight call setup, supporting connection and connectionless services in communication networks [6], [7]. 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 among the queues to balance the loads. We assume that the service discipline 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). When redistributing the jobs, we order all the waiting jobs according to their arrival times and assign the jobs to the queues in a round robin order, assigning the older jobs first. 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 only consider periodic load balancing with homogeneous queues, but we have also considered the case in which a proportion of the queues are temporarily down (arrivals come but no service is provided); see Sec. 10 of [8], which is a longer version of this paper. Consistent with intuition, load balancing is even more important in unbalanced scenarios.

Our main contributions are analytical models and formulas describing the performance of periodic load balancing. We describe the distribution of the number of jobs at each queue as a function of time, especially just before and just after each balancing. We 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

0-7803-4394-8/98 $10.00

a 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 [10], [11]. 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

$$EW \approx \frac{\rho}{s(1-\rho)} \frac{(c_a^2 + c_s^2)}{2} , \qquad (1)$$

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 [13]. Formula (1) shows that the mean EW is inversely proportional to s for fixed ρ. 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.

Unfortunately, however, it is not always possible to fully share resources. 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 use the shortest queue; (SQ) rule [12]. 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 $\rho \to 1$ [14].

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. 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 establish a heavy-traffic diffusion approximation for the case of general arrival and service processes. In Section 3 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 4 we make comparisons between the approximation and simulations of M/G/1 queues coupled by periodic load balancing. We consider exponential and Pareto service-time distributions (with finite variance). We focus more on long-tail service-time distributions in [8].

2 The Diffusion Approximation

We start with m separate independent and identically distributed (i.i.d.) G/GI/s queueing systems. Each queue has s servers, unlimited waiting room, the FCFS service discipline and its own general arrival process with finite arrival rate λ. We assume that the service times are independent of the arrival process, coming from a sequence of i.i.d. random variables with a general distribution having mean 1 and finite SCV c_s^2. The traffic intensity is $\rho = \lambda/s$. We assume that $\rho < 1$, so that the queue is stable, but we will let $\rho \uparrow 1$ to obtain a diffusion approximation.

Let $\{A(t) : t \geq 0\}$ denote an arrival process to any one queue, i.e., $A(t)$ counts the number of arrivals in $[0,t]$. 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.,

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

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 $D \equiv D[0,\infty)$ [4]. If $\{A(t) : t \geq 0\}$ is a renewal process, then to satisfy (2) it is necessary and sufficient for the time between renewals to have a finite SCV c_a^2.

For each ρ with $0 < \rho < 1$, let a queueing model with traffic intensity ρ be defined by letting the arrival rate be $\lambda = s\rho$, in particular by scaling a rate-1 arrival process $\{A(t) : t \geq 0\}$ by $A_\rho(t) = A(\rho st)$, $t \geq 0$. In this setting, the normalized queue length process in the standard G/GI/s model converges to reflected Brownian motion (RBM) as $\rho \to 1$ [9]. In particular, if $Q_\rho(t)$ denotes the queue length (number in system) at time t in system ρ, then

$$\frac{(1-\rho)}{\rho(c_a^2 + c_s^2)} Q_\rho(t(c_a^2 + c_s^2)/s(1-\rho)^2) \Rightarrow R(t) \qquad (3)$$

as $\rho \to 1$, 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 (3) 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.)

We now state the analog for periodic load balancing. (Proofs of all theorems appear in [8].) We assume that the queues are balanced every T_ρ time units, so that after balancing the queue lengths differ by at most one. To obtain an interesting nondegenerate limit, it is essential to let the length T_ρ of the intervals between balancing depend on the traffic intensity ρ. 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 2.1 *Consider m G/GI/s queues controlled by periodic load balancing. Make the assumptions above on the arrival and service processes. If $\rho \to 1$ with the redistribution intervals T_ρ satisfying*

$$s(1-\rho)^2 T_\rho/(c_a^2 + c_s^2) \;\to\; T \qquad (4)$$

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

$$(1-\rho)x_{0\rho}/\rho(c_a^2 + c_s^2) \;\to\; x_0 , \qquad (5)$$

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

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

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

$$P(Y_{n+1} \leq y | Y_n = x) =$$

$$P(\textstyle\sum_{i=1}^m R_i(T) \leq y | R_i(0) = x/m \ , \ 1 \leq i \leq m)$$

and conditional Laplace transform

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

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

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

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

The evolution of the limiting stochastic process $\{(X_1(t), \ldots, X_m(t)) : t \geq 0\}$ in Theorem 2.1 can be described by first calculating the distribution of the variables Y_n. 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 displayed transform, exploiting the two-dimensional Laplace transform

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

which is given explicitly in (9.3) of [1]. The numerical transform inversion algorithm in [3] can be used to calculate the transition kernel.

A more elementary approximation can be obtained by considering the 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 below.

Theorem 2.2 *In the setting of Theorem 2.1, if $m \to \infty$ after $\rho \to 1$, then $x_n \equiv X_1(nT)$ evolves deterministically as $x_{n+1} = f_T(x_n)$, where*

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

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

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

The approximation based on Theorem 2.1 is load-balanced canonical RBM using a redistribution interval T. The associated approximate redistribution interval T_ρ and levels $x_{\rho n}$ in the queueing system with traffic intensity ρ are

$$T_\rho \approx \frac{(c_a^2+c_s^2)T}{s(1-\rho)^2} \quad \text{and} \quad x_{\rho n} \approx \frac{\rho(c_a^2+c_s^2)x_n}{1-\rho} \ . \qquad (6)$$

Theorem 2.1 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 generates the approximation

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

where $(X_1(t),\ldots,X_m(t))$ is controlled canonical RBM, as indicated in Theorem 2.1. Thus, invoking Theorem 2.2 as well, the queue length just before and after the n^{th} redistribution have the approximate form

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

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

Theorems 2.1 and 2.2 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 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.)

We now develop normal distribution refinement to the deterministic sequence $\{x_n\}$.

Theorem 2.3 *In the setting of Theorem 2.1,*

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

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 is as in Theorem 2.2,

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

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

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

It is significant that the deterministic values $x_n \equiv X_1(nT)$ converge to a limit $x^*(T)$, which is the unique fixed point of the function f_T.

Theorem 2.4 *The function f_T 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 $T \to 0$.*

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. A refined approximation is a normal distribution, where the mean and variance σ^2 are the solutions of a pair of equations describing the mean μ and variance for RBM immediately after balancing. 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 $x^*(T)$.

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 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 were solved iteratively using numerical integration 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 $\gamma \equiv \sqrt{V(T, x^*(T))}$. 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.)

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

Table 1: A comparison between the steady-state mean content of each queue with m independent RBM processes, using the normal iteration, and the deterministic fixed point.

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 $x^*(T)$; for moderate m it is preferable to use the normal fixed point pair (μ, σ^2); 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.

3 Performance Comparisons

In this section we apply the diffusion approximation in Section 2 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.

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	$m = 2$	$m = 4$	$m = 16$	$m = 64$	γ
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

Table 2: A comparison between the approximate standard deviation of the steady-state content with m independent RBM processes, using the normal iteration, and the variance approximation.

scheme	distrib.	mean	st. dev.
m separate M/M/1 queues	expon.	$\frac{\rho}{1-\rho}$	$\frac{\rho}{1-\rho}$
a single M/M/m queue	normal	ρ	$\frac{\rho}{\sqrt{m}}$
m M/M/1 queues with load balancing	normal	$\gamma_1 \left(\frac{\rho}{1-\rho}\right)$	$\frac{\gamma_2}{\sqrt{m}} \left(\frac{\rho}{1-\rho}\right)$

Table 3: Approximations for the distribution of the steady-state number of jobs in the system per server.

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) 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 $\rho/\sqrt{m}$.

Now consider the case of load balancing, where the balancing intervals T_ρ in the queues are chosen consistently with the scaling in (4) for some reasonable T, e.g., with $.02 < T < 2$. Our analysis in Section 2 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 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 $\gamma_2 = 2\sqrt{V(T, x^*(T))} < 1$.

In summary, in what we regard as the typical case (consistent with the scaling in (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 is substantiated by simulation results.

4 Comparisons with Simulations

In this section we compare the RBM approximations developed in Section 2 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 2, we scale so that each is to be approximated by canonical RBM (drift -1, diffusion coefficient 1). For the results we 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 (6). Since we are considering M/M/1 queues, $s = c_a^2 = c_s^2 = 1$ and

$$T_\rho = 2T/(1-\rho)^2 = 2/(1-\rho)^2 \ .$$

We first consider the case $m = 64$ for three values of ρ : $\rho = 0.8, 0.9$ 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 1. (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

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

These two approximations are also shown in Figure 1. From Figure 1, 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 pair of equations using the RBM conditional mean and variance functions $M(t,x)$ and $V(t,x)$. 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 1 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 2 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. 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

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

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

$$c_s^2 = 1 + 2\left((\alpha-1)^2/(\alpha-2) - \alpha\right) \ .$$

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 (6), so that

$$T_\rho = (c_a^2 + c_s^2)T/(1-\rho)^2 = 4/(1-\rho)^2 \ .$$

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, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho -)/4\rho$. We compare the M/G/1 Pareto and exponential service-time-distribution cases with $\rho = 0.95$, $T = 1.0$ and $m = 64$ in Figure 3. The Pareto and exponential cases were scaled differently, so that the approximation for both involves canonical RBM. In Figure 3 we include the normal approximation $N(\mu, m\sigma^2) \approx N(x^*(1), V(1, x^*(1)))$. The close agreement between the exponential and Pareto simulation results shows the remarkable power of the heavy-traffic scaling.

References

[1] Abate, J. and Whitt, W. (1987) Transient behavior of regulated Brownian motion, I and II. *Adv. Appl. Prob.* 19, 560–631.

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

[3] Choudhury, G. L., Lucantoni, D. M. and Whitt, W. (1994) Multidimensional transform inversion with applications to the transient M/G/1 queue. *Ann. Appl. Prob.* 4, 719–740.

[4] Ethier, S. N. and Kurtz, T. G. (1986) *Characterization and Approximation of Markov Processes*, Wiley, New York.

[5] Harchol-Balter, M. and Downey, A. B. (1996) Exploiting process lifetime distributions for dynamic load balancing. *Proceedings SIGMETRICS '96.*

[6] Hjálmtýsson, G. (1997) 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, pp. 35–45.

[7] Hjálmtýsson, G. and Ramakrishnan, K. K. (1997) UNITE — An architecture for lightweight signalling in ATM networks. *Proceedings IEEE Infocom '98.*

[8] Hjálmtýsson, G. and Whitt, W. (1998) Periodic load balancing. *Queueing Systems*, to appear.

[9] Iglehart, D. L. and Whitt, W. (1970) Multiple channel queues in heavy traffic, I and II. *Adv. Appl. Prob.* 2, 150–177 and 355–369.

[10] Laws, C. N. (1992) Resource pooling in queueing networks with dynamic routing. *Adv. Appl. Prob.* 24, 699–726.

[11] Mandelbaum, A. and Reiman, M. I. (1996) On pooling in queueing networks. *Management Sci.*, to appear.

[12] Weber, R. W. (1978) On the optimal assignment of customers to parallel servers. *J. Appl. Prob.* 15, 406–413.

[13] Whitt, W. (1993) Approximations for the GI/G/m queue. *Production and Operations Management* 2, 114–160.

[14] Zhang, H., Hsu, G. and Wang, R. (1995) Heavy traffic limit theorems for a sequence of shortest queueing systems. *Queueing Systems* 21, 217–238.

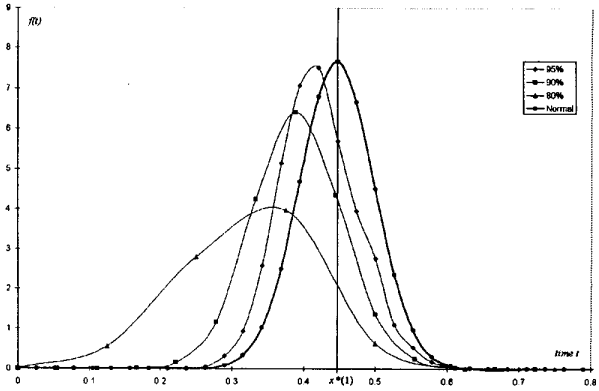

Figure 1: A comparison between the RBM approximations and histograms of the normalized queue lengths after load balancing, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/2\rho$, in 64 M/M/1 queues for $\rho = 0.80$, 0.90 and 0.95 and T_ρ scaled from $T = 1.0$.

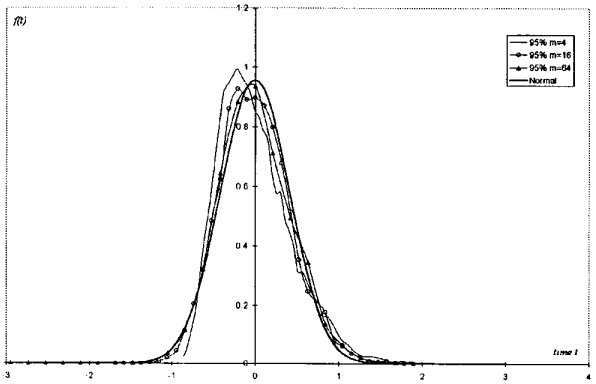

Figure 2: A comparison between the RBM approximations and histograms of the centered and normalized queue lengths after load balancing, $m^{1/2}[(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/\rho - \bar{n}_{i\rho}^{(m)}]$, in m M/M/1 queues with $\rho = 0.95$ for $m = 4$, 16 and 64 and T_ρ scaled from $T = 1.0$.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

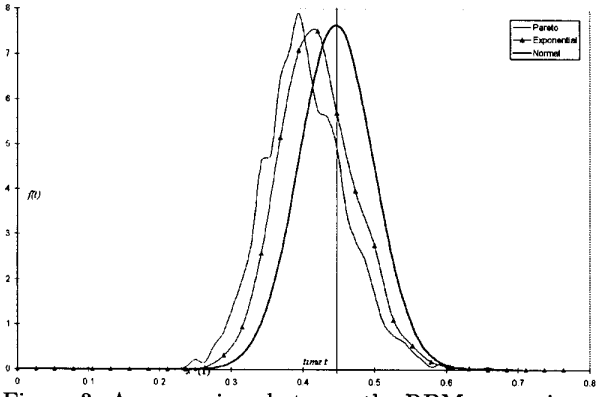

Figure 3: A comparison between the RBM approximation and histograms of the normalized queue lengths after load balancing, $(1-\rho)N_{i\rho}^{(m)}(nT_\rho)/(1+c_s^2)\rho$, in 64 M/G/1 queues with $\rho = 0.95$ and $T = 1$ for exponential ($c_s^2 = 1$) and Pareto ($\alpha = c_s^2 = 3$) service-time distributions.