SRC Technical Note
1998 - 001

February 8, 1998

On the Analysis of Randomized Load Balancing Schemes

Michael Mitzenmacher

digital

Systems Research Center
130 Lytton Avenue
Palo Alto, California 94301

http://www.research.digital.com/SRC/

On the Analysis of Randomized Load Balancing Schemes

Michael Mitzenmacher *

Abstract

It is well known that simple randomized load balancing schemes can balance load effectively while incurring only a small overhead, making such schemes appealing for practical systems. In this paper, we provide new analyses for several such dynamic randomized load balancing schemes.

Our work extends a previous analysis of the *supermarket model*, a model that abstracts a simple, efficient load balancing scheme in the setting where jobs arrive at a large system of parallel processors. In this model, customers arrive at a system of n servers as a Poisson stream of rate λn, $\lambda < 1$, with service requirements exponentially distributed with mean 1. Each customer chooses d servers independently and uniformly at random from the n servers, and is served according to the First In First Out (FIFO) protocol at the choice with the fewest customers. For the supermarket model, it has been shown that using $d = 2$ choices yields an exponential improvement in the expected time a customer spends in the system over $d = 1$ choice (simple random selection) in equilibrium. Here we examine several variations, including constant service times and *threshold models*, where a customer makes up to d successive choices until finding one below a set threshold.

Our approach involves studying limiting, deterministic models representing the behavior of these systems as the number of servers n goes to infinity. Results of our work include useful general theorems for showing that these deterministic systems are stable or converge exponentially to fixed points. We also demonstrate that allowing customers two choices instead of just one leads to exponential improvements in the expected time a customer spends in the system in several of the related models we study, reinforcing the concept that just two choices yields significant power in load balancing.

1 Introduction

Distributed computing systems continue to rise in prevalence; networks of workstations and clusters of personal computers hold the promise of increased power and price/performance ratios. It has long been known that in distributed systems, redistributing the workload through load balancing can lead to significant performance improvements, in terms of both the mean and standard deviation of the time jobs spend in the system (for example, see [7, 35]). Moreover, simple randomized schemes with low overhead have proven effective in simulations; however, analyzing such schemes is often difficult. In this paper, we provide new analyses for several dynamic randomized load balancing models. Unlike previous similar analyses, we do not assume that in equilibrium each server is stochastically independent from other servers.

One example of the type of problem we consider, previously studied in [27], is the following natural dynamic model: customers arrive as a Poisson stream of rate λn, where $\lambda < 1$, at a collection of n servers. The service times for the customers are independent and exponentially distributed with mean 1. Each customer chooses some constant number d of servers independently and uniformly at random from the n servers, and waits for service at the one currently containing the fewest customers (ties being broken arbitrarily), according to the First In First Out (FIFO) protocol. We call this model the *supermarket model*, or the *supermarket*

*This work was supported in part by the ONR and in part by NSF Grant CCR-9505448. Much of this work was done while the author was a student at U.C. Berkeley. A previous version of this work appeared in the *9th ACM Symposium on Parallel Algorithms and Architectures*, 1997.

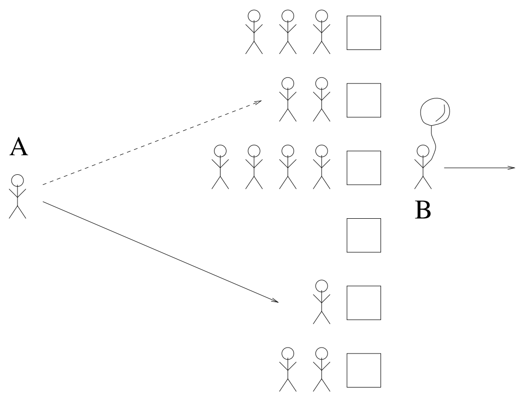

Figure 1: The supermarket model. Incoming customer A chooses two random servers, and queues at the shorter one. Customer B has recently been served and leaves the system.

system (see Figure 1). We are interested in the behavior of this system in equilibrium. Note that the average arrival rate per queue is less than service rate ($\lambda < 1$), we expect the system to be *stable*, in the sense that the expected number of customers per queue remains finite in equilibrium.

Standard queueing theory does not directly apply to the supermarket model, because the server loads are dependent: the arrival rate at any server depends on the loads at the other servers. This dependency complicates the analysis dramatically.

Many variations on the supermarket model exist. For example, in a *threshold* system, an incoming customer successively chooses queues at random until either finding one with a load below a fixed threshold or using d choices. A threshold scheme may be more efficient than giving each customer d choices in practice, since each choice will generally require some communication, and threshold schemes reduce the amount of necessary communication. As another example, service times might not be eponentially distributed, but constant, or given by another distribution. In this paper, we introduce new analyses for these and other variations. Our approach, following that of [27], has two main components:

- We define an idealized process, corresponding to a system with an infinite number of servers. We then analyze this process, which is cleaner and easier because its behavior is completely deterministic.
- We relate the idealized system to the finite system, bounding the error between them.

Our analysis of the limiting system (as the number of servers grows to infinity) focuses on finding the *fixed point* (or *equilibrium point*) to which the system tends. If the system converges to its fixed point, then we can use it to determine such quantities as the expected time a customer spends in the system. For most of the idealized systems we consider, we show *exponential convergence* to the fixed point, which demonstrates that the system approaches the fixed point very quickly. Indeed, besides determining the behavior of several interesting systems, a major contribution of this work is a simple, general theorem that gives appropriate conditions for convergence; we expect this theorem will prove useful in other settings as well. We also demonstrate through simulations that the method provides accurate numerical estimates of performance, even when the actual number of servers is relatively small.

For ease of presentation, we have made several assumptions to simplify the models we consider. For example, we assume that the time for a customer to obtain information about server loads and move to a server is zero, and that the servers are homogeneous. Many of our techniques, however, generalize to more complex systems, such as systems where transferring a customer incurs a delay (see [24]). Moreover, even the simple systems we study demonstrate remarkably interesting behavior. In particular, we emphasize throughout that there is often a qualitative difference between systems where customers choose a single destination randomly and systems where customers have two or more choices available, leading to exponential improvement in measures such as the expected time in the system. Hence our work extends a great deal of previous work demonstrating the power of two choices in load balancing to several new settings, providing further evidence of the significance of this idea in the design of distributed systems.

1.1 Previous work

Distributed load balancing strategies where individual customer decisions are based on information about a limited number of other processors have been studied analytically by Eager *et al.* [7, 8, 9] and through trace-driven simulations by Zhou [35]. In fact, Eager *et al.* also use Markovian models for their analysis [7, 8, 9]; however, the authors derive their results assuming that the state of each queue is stochastically independent of the state of any other queue. This approach is exact in the asymptotic limit as the number of queues grows to infinity. Our work avoids these assumptions and introduces several new directions in the analysis of these systems. Zhou's work examines the effectiveness of the load balancing strategies proposed by Eager *et al.* as well as others in practice using a trace-driven simulation. Both Eager *et al.* and Zhou suggest that simple randomized load balancing schemes, based on choosing from a small subset of processors, perform extremely well.

In another well-studied model, incoming customers join the shortest queue; see, for example, the work by Adan, van Houtum, and van der Wal [1] and by Adan, Wessels, and Zijm [2, 3] for results and further references. The shortest queue model appears more applicable to *centralized* systems, whereas the limited coordination enforced by our model corresponds nicely to models of *distributed* systems.

Randomized load balancing schemes have also been analyzed in the static case, where there are a fixed number of customers to be permanently distributed, as in a static hash table. For example, Karp, Luby, and Meyer auf der Heide showed that using two hash functions instead of one could provide an exponential improvement in the maximum load of a hash bucket [13]; this idea was further developed and aanalyzed by Azar, Broder, Karlin, and Upfal [5]. Our work demonstrates that making two choices leads to a similar exponential improvement in the dynamic setting as well.

The justification of the relationship between the finite and limiting systems relies on Kurtz's work on *density dependent jump Markov processes* [10, 19, 20, 21, 22]. Because Kurtz's work is rather technical, we only briefly describe it here, focusing instead on examining a variety of models and attempting to gain insight into the load balancing problem. More details regarding the application of Kurtz's work these models can be found in [28]. This approach has been used similarly in several other works (for example, see [4, 11, 14, 15, 27, 31, 33, 34]).

The rest of the paper proceeds as follows: in Section 2, we briefly review the work of [27] by examining the limiting system for the supermarket model. This allows us to introduce the necessary terminology and keeps this paper essentially self-contained. To demonstrate the applicability of our methods to more realistic systems, we consider alternative service distributions in Section 3, focusing on the example of constant service times. In Sections 4 and 5, we explore some variations on the supermarket model that may also prove useful in practice, including threshold models. Section 4 includes general theorems for proving the stability or exponential convergence of the limiting systems. We specialize these theorems to handle threshold systems in Section 5. We conclude with some final comments and open questions. The main points of Kurtz's work are summarized in an appendix for the interested reader.

2 The supermarket model

In this section, we review results for the supermarket model from [27]. This review allows us to introduce the necessary terminology and methodology that we will use to study other systems.

2.1 The limiting system

Recall the definition of the supermarket model: customers arrive as a Poisson stream of rate λn, where $\lambda < 1$, at a collection of n FIFO servers. Each customer chooses some constant $d \geq 2$ servers independently and uniformly at random with replacement[1] and queues at the server currently containing the fewest customers. The service time for a customer is exponentially distributed with mean 1.

We define $m_i(t)$ to be the number of queues with *at least* i customers at time t, and $s_i(t) = m_i(t)/n$ to be fraction of queues with *at least* i customers. We drop the reference to t in the notation where the meaning is clear. In an *empty system*, which corresponds to one with no customers, $s_0 = 1$ and $s_i = 0$ for $i \geq 1$. We can represent the state of the system at any given time by an infinite dimensional vector $\vec{s} = (s_0, s_1, s_2, \ldots)$. It is clear that for each value of n, the supermarket model can be considered as a Markov chain on the above state space.

We now introduce a deterministic *limiting system* related to the finite supermarket system, given by the following set of differential equations:

$$\begin{cases} \dfrac{ds_i}{dt} & = \lambda(s_{i-1}^d - s_i^d) - (s_i - s_{i+1}) \ \text{ for } \ i \geq 1\,; \\ s_0 & = 1. \end{cases} \tag{1}$$

To explain the reasoning behind the system (1), we determine the expected change in the number of servers with at least i customers over a small period of time of length dt. The probability a customer arrives during this period is $\lambda n\, dt$, and the probability an arriving customer joins a queue of size $i-1$ is $s_{i-1}^d - s_i^d$. (This is the probability that all d servers chosen by the new customer are of size at least $i-1$, but not all are of size at least i.) Thus the expected change in m_i due to arrivals is exactly $\lambda n(s_{i-1}^d - s_i^d)dt$. Similarly, the probability a customer leaves a server of size i in this period is $n_i\, dt = n(s_i - s_{i+1})dt$. Hence, if the system behaved according to these expectations, we would have

$$\frac{ds_i}{dt} = \frac{1}{n} \cdot \frac{dm_i}{dt} = \lambda(s_{i-1}^d - s_i^d) - (s_i - s_{i+1}).$$

It should be intuitively clear that as $n \to \infty$ the behavior of the supermarket system approaches that of this deterministic system; this is justified by Kurtz's theorem, as explained in Appendix A. For now, we simply take this set of differential equations to be the appropriate limiting process.

2.2 The fixed point

Given a reasonable condition on the initial point $\vec{s}(0)$, the infinite process described by the system (1) converges to a *fixed point* $\vec{\pi}$ such that if $\vec{s}(t) = \vec{\pi}$ then $\vec{s}(t') = \vec{\pi}$ for all $t' \geq t$. For the supermarket model a necessary and sufficient condition for $\vec{s}$ to be a fixed point is that for all i, $\frac{ds_i}{dt}|_{\vec{\pi}} = 0$.

Lemma 1 [[27], Lemma 1.] *The system (1) with $d \geq 2$ has a unique fixed point with $\sum_{i=1}^{\infty} \pi_i < \infty$ given by $\pi_i = \lambda^{\frac{d^i-1}{d-1}}$.*

[1] We note that our results also hold with minor variations if the d queues are chosen without replacement.

Definition 2 *A sequence* $(x_i)_{i=0}^{\infty}$ *is said to* decrease doubly exponentially *if and only if there exist positive constants* $N, \alpha < 1, \beta > 1$, *and* γ *such that for* $i \geq N$, $x_i \leq \gamma \alpha^{\beta^i}$.

It is worth contrasting the result of Lemma 1 with the case where $d = 1$ (*i.e.*, all servers are M/M/1 queues), for which the fixed point is given by $\pi_i = \lambda^i$. For $d = 2$, the fixed point is given by $\pi_i = \lambda^{2^i - 1}$. The key feature of the supermarket system is that for $d \geq 2$ the tails π_i decrease doubly exponentially, while for $d = 1$ the tails decrease only geometrically (or singly exponentially).

2.3 Convergence to the fixed point

The deterministic differential equations (1), along with an initial point, define a *trajectory* of the system in the infinite dimensional space. In [27] it was shown that every trajectory of the limiting model of the supermarket system converges to the fixed point $\vec{\pi} = (\pi_i)$ of Lemma 1 in an appropriate metric. We review the main points here. In what follows we assume that $d \geq 2$ unless otherwise specified.

To show convergence, we find a suitable *potential function* (also called a *Lyapunov function* in the dynamical systems literature) $\Phi(t)$. The potential function must be related to the distance between the current point on the trajectory and the fixed point; by showing the potential function decreases quickly over time, we may show the trajectory heads towards the fixed point. A natural potential function to consider is $D(t) = \sum_{i=1}^{\infty} |s_i(t) - \pi_i|$, which measures the L_1-distance (or Manhattan distance) between the two points. The potential function used in [27] is actually a weighted variant of this, namely $\Phi(t) = \sum_{i=1}^{\infty} w_i |s_i(t) - \pi_i|$ for suitably chosen weights w_i.

The supermarket system not only converges to its fixed point, but that it does so *exponentially*.

Definition 3 *The potential function* Φ *is said to* converge exponentially to 0, *or simply to converge exponentially, if* $\Phi(0) < \infty$ *and* $\Phi(t) \leq c_0 e^{-\delta t}$ *for some constant* $\delta > 0$ *and a constant* c_0 *which may depend on the state at* $t = 0$.

Exponential convergence implies not only that the limiting system approaches the fixed point, but that it does so rapidly, making it a suitable reference point for system performance in practice.

Theorem 4 [[27], Theorem 6] *Let* $\Phi(t) = \sum_{i=1}^{\infty} w_i |s_i(t) - \pi_i|$, *where for* $i \geq 1$, $w_i \geq 1$ *are appropriately chosen constants. If* $\Phi(0) < \infty$, *then* Φ *converges exponentially to* 0. *In particular, if there exists a* j *such that* $s_j(0) = 0$, *then* Φ *converges exponentially to* 0.

The condition of Theorem 4 that there exists a j such that $s_j(0) = 0$ is a natural one. It can be interpreted as saying initially there is an upper bound on the maximum queue size.

Corollary 5 [[27], Corollary 7] *Under the conditions of Theorem 4, the* L_1*-distance from the fixed point* $D(t) = \sum_{i=1}^{\infty} |s_i(t) - \pi_i|$ *converges exponentially to* 0.

Corollary 5 shows that the L_1-distance to the fixed point converges exponentially quickly to 0. Given this convergence, we may now ask what the expected time in the system looks like. It is interesting to compare the case where $d \geq 2$ to the case of $d = 1$ (for which the expected time is well known).

Theorem 6 [[27], Theorem 8] *The expected time a customer spends in the limiting model of an initially empty supermarket system for* $d \geq 2$ *converges as* $t \to \infty$ *to* $T_d(\lambda) \equiv \sum_{i=1}^{\infty} \lambda^{\frac{d^i - d}{d-1}}$. *If* $T_1(\lambda) \equiv \frac{1}{1-\lambda}$, *then for* $\lambda \in [0, 1]$, $T_d(\lambda) \leq c_d(\ln T_1(\lambda))$ *for some constant* c_d *dependent only on* d. *Furthermore,* $\lim_{\lambda \to 1^-} \frac{T_d(\lambda)}{\ln T_1(\lambda)} = \frac{1}{\log d}$.

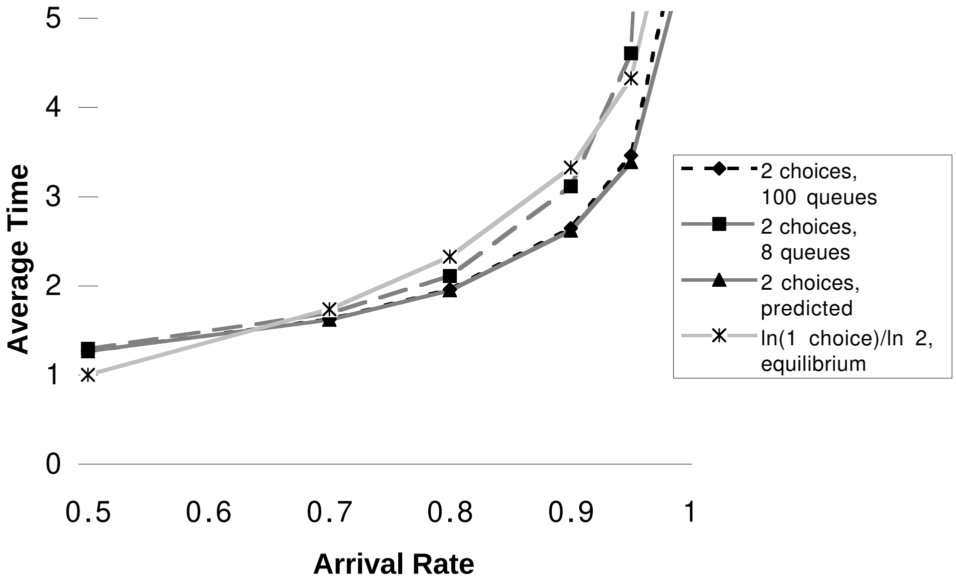

Figure 2: The graph compares the expected time in the system from simulations of 8 and 100 queues with the limiting system prediction when two choices are made and the *logarithm* of the expected time in equilibrium when one choice is made under various arrival rates (λ).

Choosing from $d > 1$ queues hence yields an exponential improvement in the expected time a customer spends in the limiting system, and as $\lambda \to 1^-$ the choice of d affects the time only by a small constant factor (dependent on d). These results are remarkably similar to those for the static load balancing problem studied in [5].

Simulations verify that this behavior is apparent even in small systems; for example, see Figure 2. More details are given [27] or [28].

3 Constant service times

The assumptions underlying the supermarket model, namely that the arrival process is Poisson and that the service times are exponentially distributed, do not accurately describe many (and probably most) real systems, although they are useful because they lead to a simple Markovian system. In this section, we demonstrate how to modify our approach to handle more general service and arrival times. We focus on the example where the service time is a fixed constant. The approach we use is based on *Erlang's method of stages*, which we shall describe briefly here. For a more detailed explanation see [17, Sections 4.2 and 4.3]. We approximate the constant service time with a *gamma distribution*: a single service will consists of r stages of service, where each stage is exponentially distributed with mean $1/r$. As r becomes large, the expected service time remains 1 while the variance falls like $1/r$, so that the service time behaves like a constant random variable in the limit as $r \to \infty$.

The state of a queue will now be the total number of stages remaining that the queue has to process, rather than the number of customers; that is, the state of a queue is [r(# of waiting customers) + stages of the customer being served]. Since r determines the size of the state space, numerical calculations will be easier if we choose r to be a reasonably small finite number. Our simulations suggest that for $r \approx 20$ the approximations for constant service times are quite accurate.

There is some ambiguity in the meaning of a customer choosing the shortest queue. If the number of customers in two queues are the same, can an incoming customer distinguish which queue has fewer stages of service remaining? Let us first consider the case where we have *aware* incoming customers, who can tell how many stages are left for each of their d choices and choose accordingly. Let s_j be the fraction of queues with at least j stages left to process (where we take $s_j = 1$ whenever $j \leq 0$). Then s_j increases whenever an arrival comes to a queue with at least $j - r$ and fewer than j stages left to complete. Similarly, s_j decreases whenever a queue with j stages completes a stage, which happens at rate r. The corresponding system of differential equations is thus

$$\frac{ds_j}{dt} = \lambda(s_{j-r}^d - s_j^d) - r(s_j - s_{j+1}).$$

(When $r = 1$, this corresponds exactly to the standard supermarket model.)

We can identify a unique fixed point $\vec{\pi}$ for this system (using $\frac{ds_j}{dt} = 0$ at the fixed point). We must have $\pi_1 = \lambda$ (intuitively because the arrival rate and exit rate of customers must be equal), and $\pi_i = 1$ for $i \leq 0$. From these initial conditions one can find successive values of π_j from the recurrence

$$\pi_{j+1} = \pi_j - \frac{\lambda(\pi_{j-r}^d - \pi_j^d)}{r}. \tag{2}$$

Unfortunately, we have not found a convenient closed form for π_j.

We say that the system has *unaware* customers if customers learn only the queue size of their choices, and not the number of stages. If more than one server chosen by an incoming customer has the shortest queue, then the customer chooses randomly from those servers. The differential equations are slightly more complicated than in the aware case. Again, let s_j be the fraction of queues with at least j stages left to process. For notational convenience, let $S_i = s_{(i-1)r+1}$ be the fraction of queues with at least i customers (where $S_0 = 1$ always), and let $\phi(j) = \lceil \frac{j}{r} \rceil$ be the number of customers in a queue with j stages left to process. The corresponding differential equations are:

$$\begin{aligned}\frac{ds_j}{dt} &= \lambda(S_{\phi(j)-1}^d - S_{\phi(j)}^d)\frac{s_{j-r} - S_{\phi(j)}}{S_{\phi(j)-1} - S_{\phi(j)}} + \\ &\quad \lambda(S_{\phi(j)}^d - S_{\phi(j)+1}^d)\frac{S_{\phi(j)} - s_j}{S_{\phi(j)} - S_{\phi(j)+1}} - r(s_j - s_{j+1}).\end{aligned}$$

Note that the fixed point cannot be determined by a simple recurrence, as the derivative of s_j depends on $S_{\phi(j)}$,$S_{\phi(j)-1}$, and $S_{\phi(j)+1}$. One can find the fixed point to a suitable degree of accuracy by standard numerical methods, however.

3.1 Constant versus exponential service times

The question of whether constant service times reduce the expected delay in comparison to exponential service times often arises when one tries to use standard queueing theory results to find performance bounds on networks. (See, for example, [12, 25, 26, 29, 32].) Generally, results comparing various service times are achieved using stochastic comparison techniques. Here, we instead compare the fixed points of the corresponding limiting systems.

We show that at the fixed points, the fraction of servers with at least k customers is greater when service times are exponential than when service times have a gamma distribution (with $r \geq 2$) with the same mean. Since gamma distributed random variables become constant in the limiting case, we can conclude that constant service times are better than exponential service times in supermarket systems in terms of measures such as the expected time in the system. (We note that to formally compare constant service times with exponential service times with this approach requires technical arguments regarding changing the order in which the limits as $n \to \infty$ and $r \to \infty$ are taken; for example, see [31, Chapter 14]. We have not completed such a formal justification. However, the theorem below is the key step in the argument, and moreover it is interesting in its own right.)

We consider the case of aware customers where service times have a gamma distribution corresponding to r stages. Recall that the fixed point was given by the recurrence (2) as $\pi_{j+1} = \pi_j - \lambda(\pi_{j-r}^d - \pi_j^d)/r$, with $\pi_1 = \lambda$ and $\pi_i = 1$ for $i \leq 0$. The fixed point for the standard supermarket model, as found in Lemma 1, satisfies $\pi_{i+1} = \lambda\pi_i^d$. Since π_1 is λ in both the standard supermarket model and the model with gamma distributed service times, to show that the tails are larger in the standard supermarket model, it suffices to show that $\pi_{\phi(j)+1} \leq \lambda\pi_{\phi(j)}^d$ in the aware customer model. Inductively it is easy to show the following stronger fact:

Theorem 7 *In the system with aware customers, for $j \geq 1$,*

$$\pi_j = \frac{\lambda}{r} \sum_{i=j-r}^{j-1} \pi_i^d.$$

Proof: The equality can easily be verified for $1 \leq j \leq r$. For $j > r$, the following induction yields the theorem:

$$\begin{aligned}
\pi_j &= \pi_{j-1} - \frac{\lambda}{r}(\pi_{j-r-1}^d - \pi_{j-1}^d) \\
&= \pi_{j-2} - \frac{\lambda}{r}(\pi_{j-r-1}^d + \pi_{j-r-2}^d - \pi_{j-1}^d - \pi_{j-2}^d) \\
&\vdots \\
&= \pi_{j-r} - \frac{\lambda}{r}\left(\sum_{i=j-2r}^{j-r-1} \pi_i^d - \sum_{k=j-r}^{j-1} \pi_k^d\right) \\
&= \frac{\lambda}{r} \sum_{k=j-r}^{j-1} \pi_k^d.
\end{aligned}$$

Here the last step follows from the inductive hypothesis, and all other steps follow from the recurrence equation (2) for the fixed point. ■

An entirely similar proof holds even in the case of *unaware* customers [28, Theorem 4.7].

3.2 Simulations and other service times

We show with simulations that small values for the number of stages r yield good approximations for constant service times. Table 1 compares the value of the expected time a customer spends in a limiting system with unaware customers and $d = 2$ choices per customer obtained using various values of r against the results from simulations with constant service times for 100 queues. The simulation results are the average of ten runs, each for 100,000 time units, with the first 10,000 time units excluded to account for the fact that the system begins empty. In all cases except $\lambda = 0.99$ increasing r yields a better match between the simulation and the prediction from the fixed point; this discrepancy is because the predictions for $\lambda = 0.99$ are not sufficiently accurate for systems of only one hundred queues.

In principle, this approach could be used to develop deterministic differential equations that approximate the behavior of any service time distribution. This follows from the fact that the distribution function of any

λ	Simulation	$r = 10$	$r = 20$	$r = 30$
0.50	1.1352	1.1478	1.1412	1.1390
0.70	1.3070	1.3355	1.3200	1.3148
0.80	1.4654	1.5090	1.4847	1.4766
0.90	1.7788	1.8492	1.8065	1.7923
0.95	2.1427	2.2355	2.1714	2.1500
0.99	3.2678	3.2461	3.1243	3.0644

Table 1: Simulations versus estimates for constant service times: 100 queues.

positive random variable can be approximated arbitrarily closely by a mixture of countably many gamma distributions [16, Lemma 3.9]. In practice, for the solution of this problem to be computable in a reasonable amount of time, both the number of distributions in the mixture and the number of stages for each distribution must be small in order to keep the total number of states reasonably small. Although these limitations appear severe, many service distributions can still be handled easily. For example, as we have seen, in the case of constant service times one only needs to use a single gamma distribution with a reasonable number of stages r to get a very good approximation. This increases the state space, and hence approximately the time to determine the behavior of the linear equations, by a factor of r over the case where service times are exponential. Distributions where the service time takes on one of a small finite number of values can be handled similarly.

4 Other dynamic models

In this section, we shall develop limiting systems for some variations on the supermarket model and show that many of these systems also converge exponentially to their fixed points. (As all of the systems we examine have a unique fixed point where the average number of customers per queue is finite, we shall simply refer to *the* fixed point for these systems.)

4.1 Customer types and errors

One way to extend the supermarket model is to consider what happens when different customers can have different numbers of choices. We will observe that giving even a small fraction of customers an extra choice can have a dramatic effect on load distribution, especially in a heavily loaded system. This fact has important practical ramifications; for example, since obtaining load information typically requires sending messages through the system, one may wish to reduce the average number of messages per customer by only giving a fraction of the customers additional choices.

We examine the specific case where there are two types of customers. One type chooses only one queue; each customer is of this type with probability $1 - p$. The more privileged customer chooses two queues; each customer is of this type with probability p. The corresponding limiting system is governed by the following set of differential equations:

$$\frac{ds_i}{dt} = \lambda p(s_{i-1}^2 - s_i^2) + \lambda(1-p)(s_{i-1} - s_i) - (s_i - s_{i+1}). \qquad (3)$$

The fixed point is given by $\pi_0 = \lambda$, $\pi_i = \lambda\pi_{i-1}(1 - p + p\pi_{i-1})$. Note that this matches the supermarket model for $d = 1$ and $d = 2$ in the cases where $p = 0$ and $p = 1$, respectively. There does not appear to be a convenient closed form for the fixed point for other values of p.

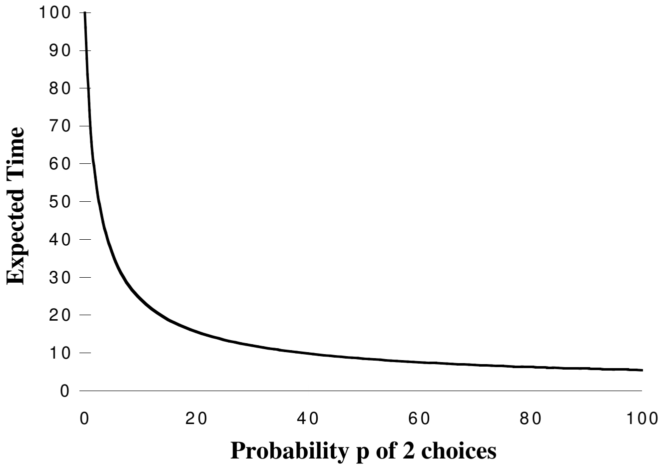

Figure 3: Expected time in the system versus probability (p) of that a customer chooses two locations ($\lambda = 0.99$).

As shown in Figure 3, which demonstrates the results for the limiting system, the effect of increasing the fraction of customers with two choices has a non-linear effect on the expected time that is dramatic at high loads; at $\lambda = 0.99$, most of the gain occurs when only 20% of the customers have two choices. Our simulation results verify that the behavior of finite systems accurately matches the behavior predicted by our limiting model.

This model has an interesting alternative interpretation. A customer who only has one choice is equivalent to a customer who has two choices, but erroneously goes to the wrong queue half of the time. Hence, the above system is equivalent to a two-choice system where customers make errors and go to the wrong queue with probability $\frac{1-p}{2}$. A model of this sort may therefore also be useful in the case where the information available to the customers from the chosen servers is unreliable or approximate. This analysis suggests that as long as this approximate load information reflects server loads with some reasonable accuracy between updates, choosing from two servers should still perform quite well. (See also [24] for similar ideas in other scenarios.)

4.2 Closed models

In the *closed* supermarket model, at each time step exactly one non-empty queue, chosen uniformly at random, completes service, and the customer is immediately recycled back into the system by again choosing the shortest of d random queues. Let the number of customers that cycle through the system be αn. Note that the average number of customers per queue is α; this corresponds to the invariant $\sum_{i=1}^{\infty} s_i = \alpha$.

The limiting system is again very similar to that of the original supermarket model. An important difference is that at each step, the probability that a customer leaves a server with i customers is $\frac{s_i - s_{i+1}}{s_1}$, since a random queue with at least one customer loses a customer. The corresponding differential equations are thus

$$\frac{ds_i}{dt} = s_{i-1}^d - s_i^d - \frac{s_i - s_{i+1}}{s_1}. \tag{4}$$

To find the fixed point, assume $\pi_1 = \beta$. Then inductively, we can solve to find $\pi_i = \beta^{\frac{d^i - 1}{d-1}}$; the correct value of β can be found by using the constraint $\sum_{i=1}^{\infty} \pi_i = \sum_{i=1}^{\infty} \beta^{\frac{d^i - 1}{d-1}} = \alpha$.

4.3 Bounded buffers

In practice, we may have a system where the queue size has a maximum limit, say b. For example, if customers are processes with associated data, then the queue size may be limited by the amount of memory in a server's buffer. In this case, we assume that arriving customers that find queues filled are turned away. That is, for the supermarket model, if an arriving customer chooses d queues that all have b customers already waiting, the customer leaves the system unserved immediately.

The state can be represented by a finite dimensional vector $(s_0, s_1, \ldots, s_b)$. The long-term probability that a customer is turned away can be determined by finding from the point, and is π_b^d. The limiting system is given by the following equations:

$$\begin{aligned}\frac{ds_i}{dt} &= \lambda(s_{i-1}^d - s_i^d) - (s_i - s_{i+1}), \quad i < b\,;\\ \frac{ds_b}{dt} &= \lambda(s_{b-1}^d - s_b^d) - s_b.\end{aligned}$$

Note that at the fixed point for this problem, $\pi_1 \neq \lambda$. The total arrival rate of customers into the queues at the fixed point is $\lambda(1-\pi_b^d)$, as some customers do not enter the system. Since at the fixed point the total rate at which customers arrive must equal the rate at which they leave, we have $\pi_1 = \lambda(1-\pi_b^d)$. Using the differential equations, we can develop a recurrence for the values of the fixed point π_i. This recurrence yields a polynomial equation for π_b, which can be shown to have a unique root between 0 and 1. Solving for π_b then allows us to compute the fixed point numerically.

4.4 Convergence and stability of limiting systems

In this section, we provide a general theorem (similar to Theorem 4) that can be used to show that several systems we have considered converge exponentially to their fixed point. In some cases, however, proving convergence is difficult. Instead of proving convergence, it is often easier to prove the weaker property of *stability* of the fixed point. We will say that a fixed point is stable if the L_1-distance to the fixed point is non-increasing along every trajectory (this is actually stronger than the standard definition). We also give a general theorems with conditions for stability. We believe these results are interesting in their own right and will be useful in the future for studying other systems. (For another approach to proving convergence for these problems, see [33].)

We consider general systems governed by the equations $\frac{ds_i}{dt} = f_i(\vec{s})$ for $i \geq 1$, with fixed point $\vec{\pi} = (\pi_i)$. Let $\epsilon_i(t) = s_i(t) - \pi_i$, with the understanding that for $i < 1$ or i larger than the dimension of the state space we fix $\epsilon_i = 0$. We shall drop the explicit dependence on t when the meaning is clear. For convenience, we shall consider only systems where $s_i(t) \in [0, 1]$ for all t, and hence $\epsilon_i(t) \in [-\pi_i, 1-\pi_i]$ for all t. This restriction simplifies the statements of our theorems and can easily be removed; however, all the systems described in this section meet this condition.

We examine the L_1-distance $D(t) = \sum_{i \geq 1} |\epsilon_i(t)|$. In the case where our state space is countably infinite dimensional, the upper limit of the summation is infinity, and otherwise it is the dimension of the state space. For technical reasons, we let $\frac{dD}{dt}$ denote the right-hand derivative (this will be explained in the last paragraph of the proof). We shall prove that $\frac{dD}{dt} \leq 0$ everywhere; this implies that $D(t)$ is non-increasing over time, and hence the fixed point is stable.

For many of the systems we have examined, the functions f_i have a convenient form: they can be written as sums of polynomial functions of the individual s_j, with no product terms $s_j s_k$ for $j \neq k$. This allows us to group together terms in dD/dt containing only ϵ_i, and consider them separately. By telescoping the terms of the derivative appropriately, we can show the system is stable by showing that the sum of the terms containing ϵ_i are at most 0.

Theorem 8 *Suppose we are given a system* $d\epsilon_i/dt = \sum_j g_{i,j}(\epsilon_j)$*, where the functions* $g_{i,j}$ *satisfy the following conditions:*

1. $g_{i,i}(x) = -\sum_{j \neq i} g_{j,i}(x)$ *for* $x \in [-\pi_i, 1 - \pi_i]$*;*

2. *for all* $i \neq j$, $\mathrm{sgn}(g_{j,i}(x)) = \mathrm{sgn}(x)$ *for* $x \in [-\pi_i, 1 - \pi_i]$.

Then for $D(t) = \sum_{i=1}^{\infty} |\epsilon_i(t)|$ *we have* $dD/dt \leq 0$*, and hence the fixed point is stable.*

Proof:

For each i, we group the terms in ϵ_i of dD/dt, and show that the sum of all terms involving ϵ_i is at most 0. Note that, technically, dD/dt is not well-defined when some $\epsilon_i = 0$; we shall clarify this problem subsequently and temporarily we assume that all ϵ_i are non-zero.

The terms containing ϵ_i in dD/dt sum to $h(\epsilon_i) = g_{i,i}(\epsilon_i)\,\mathrm{sgn}(\epsilon_i) + \sum_{j \neq i} g_{j,i}(\epsilon_i)\,\mathrm{sgn}(\epsilon_j)$. By condition 2 of the statement of the theorem, $h(\epsilon_i)$ is maximized when $\mathrm{sgn}(\epsilon_j) = \mathrm{sgn}(\epsilon_i)$ for all $j \neq i$. Hence $h(\epsilon_i) \leq \mathrm{sgn}(\epsilon_i) \sum_j g_{j,i}(\epsilon_i) = 0$, where the last equaity follows from condition 1 of the theorem. Hence $dD/dt \leq 0$, and this suffices to show that the fixed point is stable.

We now consider the technical problem of defining dD/dt when $\epsilon_i(t) = 0$ for some i. Since we are interested in the forward progress of the system, it is sufficient to consider the upper right-hand derivatives of ϵ_i. (See, for instance, [23, p. 16].) That is, we may define

$$\left.\frac{d|\epsilon_i|}{dt}\right|_{t=t_0} \equiv \lim_{t \to t_0^+} \frac{|\epsilon_i(t)|}{t - t_0},$$

and similarly for dD/dt. Note that this choice has the following property: if $\epsilon_i(t) = 0$, then $\left.\frac{d|\epsilon_i|}{dt}\right|_{t=t_0} \geq 0$, as it intuitively should be. The above proof applies unchanged with this definition of dD/dt, with the understanding that with regard to the sgn function the case $\epsilon_i > 0$ includes the case where $\epsilon_i = 0$ and $d\epsilon_i/dt \geq 0$, and similarly the case $\epsilon_i < 0$ includes the case where $\epsilon_i = 0$ and $d\epsilon_i/dt < 0$. ■

It is simple to check that the conditions of Theorem 8 hold for several of the systems we have studied. Hence we immediately have the following corollary:

Corollary 9 *The limiting systems for the following systems have stable fixed points: gamma distributed service times with aware customers (Section 3), customer types (Section 4.1), and bounded buffers (Section 4.3).*

Proof: We consider only the system with customer types described in Section 4.1 and whose behavior is given by equation (3), as the argument is entirely similar for the other models stated.

With the substitution $\epsilon_i = s_i - \pi_i$, equation (3) becomes

$$\frac{d\epsilon_i}{dt} = -2\lambda p \pi_i \epsilon_i - \lambda p \epsilon_i^2 - \lambda(1-p)\epsilon_i - \epsilon_i + 2\lambda\pi_{i-1}\epsilon_{i-1} + \lambda\epsilon_{i-1}^2 + \lambda(1-p)\epsilon_{i+1} + \epsilon_{i+1}. \quad (5)$$

(Note that all terms without some ϵ_j factor sum to 0 by definition of the fixed point.)

Condition 1 of Theorem 8 clearly holds from equation (5). Condition 2 is also easily checked– note that $\mathrm{sgn}(\epsilon_{i-1} = \mathrm{sgn}(\lambda\epsilon_{i-1}^2 + 2\lambda\pi_{i-1}\epsilon_{i-1})$ over the appropriate interval. Hence the conditions of Theorem 8 hold, proving the corollary. ■

A simple generalization of Theorem 8 allows us to prove convergence, using a weighted form of the potential function as in Theorem 4.

Theorem 10 *Suppose we are given a system $d\epsilon_i/dt = \sum g_{i,j}(\epsilon_j)$, and suppose also that there exists an increasing sequence of real numbers w_i (with $w_0 = 0$) and a positive constant δ such that the w_i and the functions $g_{i,j}$ satisfy the following conditions:*

1. $\operatorname{sgn}(x)\sum_j w_j g_{j,i}(x) \leq -\delta w_i |x|$ *for* $x \in [-\pi_i, 1-\pi_i]$*;*
2. *for all* $i \neq j$, $\operatorname{sgn}(g_{j,i}(x)) = \operatorname{sgn}(x)$ *for* $x \in [-\pi_i, 1-\pi_i]$.

Then for $\Phi(t) = \sum_{i=1}^{\infty} w_i|\epsilon_i(t)|$, we have that $d\Phi/dt \leq -\delta\Phi$, and hence from any initial point where $\sum_i w_i|\epsilon_i| < \infty$ the process converges exponentially to the fixed point in L_1-distance.

Proof: We group the terms in ϵ_i from $d\Phi/dt$ as in Theorem 8. By the assumptions of the theorem, the sum of all the terms involving ϵ_i is at most $-\delta w_i|\epsilon_i|$. We may conclude that $d\Phi/dt \leq -\delta\Phi(t)$ and hence $\Phi(t)$ converges exponentially to 0. Also, note that we may assume without loss of generality that $w_1 = 1$, since we may scale the w_i. Hence we may take $\Phi(t)$ to be larger than the L_1-distance to the fixed point $D(t)$, and thus the process converges exponentially to the fixed point in L_1-distance. ■

Proving convergence thus reduces to showing that a suitable sequence of weights w_i satisfying Condition 1 of Theorem 10 exist, which is quite often straightforward. In fact, Theorem 10 applies directly to several of the models we have mentioned. For these models we will assume, as in Theorem 4, that in our intial state there exists an upper bound on the initial queue size, to guarantee that the system begins in a well-defined state.

Corollary 11 *The limiting systems for the following systems converge exponentially to their fixed points: gamma distributed service times with aware customers (Section 3), customer types (Section 4.1), and bounded buffers (Section 4.3).*

Proof: Again we consider only the system with customer types given by equation (3), as the argument for other models is similar. That Condition 2 of Theorem 10 holds was shown in Corollary 9. Hence we need only show that a δ and a sequence w_i that satisfies Condtion 1 of Theorem 10 exist. We set $w_0 = 0$ and $w_1 = 1$ and show how to define the other w_i and the δ accordingly.

Using equation (5), Condition 1 of Theorem 10 becomes the following:

$$\operatorname{sgn}(\epsilon_i)\left[w_{i+1}(2\lambda p\pi_i\epsilon_i + \lambda p\epsilon_i^2) - w_i(2\lambda\pi_i\epsilon_i + \lambda\epsilon_i^2 + \lambda(1-p)\epsilon_i + \epsilon_i) + w_{i-1}(\lambda(1-p)\epsilon_i + \epsilon_i)\right] \leq -\delta w_i|\epsilon_i|$$

As $|\epsilon_i| = \operatorname{sgn}(\epsilon_i)\epsilon_i$, and the condition trivially holds if $\epsilon_i = 0$, we may divide through by $|\epsilon_i|$ to restate the condition as

$$(w_i - w_{i-1})(1+\lambda(1-p)) + (2\lambda p\pi_i + \lambda p\epsilon_i)(w_i - w_{i+1}) \geq \delta w_i\,;$$

or, using the fact that $|\epsilon_i| \leq 1$,

$$w_{i+1} \leq w_i + \frac{w_i(1+\lambda(1-p)-\delta) - w_{i-1}(1+\lambda(1-p))}{\lambda p(2\pi_i + 1)}.$$

It is simple to check inductively that one can choose an increasing sequence of w_i (starting with $w_0 = 0$, $w_1 = 1$) and a δ such that the w_i satisfy the above restriction. For example, we break the terms up into two subsequences. The first subsequence consists of all w_i such that π_i satisfies $\lambda p(2\pi_i + 1) \geq \frac{1+\lambda}{2}$. For these i we can choose $w_{i+1} = w_i + \frac{w_i(1-\delta)-w_{i-1}}{3}$. Because this subsequence has only finitely many terms, we can choose a suitably small δ so that this sequence is increasing. For sufficiently large i, we must have $\lambda p(2\pi_i + 1) < \frac{1+\lambda}{2} < 1$, and for these i we may set $w_{i+1} = w_i + \frac{2w_i(1+\lambda(1-p)-\delta)-2(1+\lambda(1-p))w_{i-1}}{1+\lambda}$. This

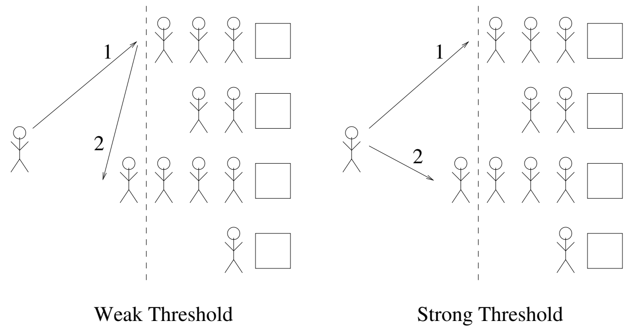

Figure 4: Weak and strong threshold models. A customer rechooses if and only if they would start behind the dashed line. In the weak model, the customer jumps to a second server, and may go to a longer line (2). In the strong model, the customer goes to the shorter of the two lines (1).

simple recurrence for the w_i is easily solved and clearly increasing for suitably small δ. Hence, by taking a δ small enough, both sequences of w_i will be increasing.

Technically, we should choose a sequence of w_i so that the the corresponding $\Phi(0) = \sum_{i=1}^{\infty} w_i |\epsilon_i(0)|$ is finite. We can easily modify the tail of the w_i sequence above so that it is dominated by a geometrically increasing sequence, where the ratio of successive terms is less than $1/\lambda$. If we assume that in the initial state $s_j(0) = 0$ for some j, then ϵ_j is eventually dominated by geometric series where the ratio of successive terms is at most λ. Hence we may find a suitable sequence of w_i such that $\sum_{i=1}^{\infty} w_i |\epsilon_i(0)|$ is finite. From this it is clear that the conditions of Theorem 10 holds, proving the corollary. ■

For the closed model and the model with unaware customers, Theorems 8 and 10 do not immediately apply. However, the technique of examining the terms in each ϵ_i separately can still prove effective; for example, it can be used to prove that the fixed point for the closed model given by the equations (4) is stable.

5 Threshold models

In practice, it may often be more efficient not to give all customers several choices, as each choice may have a corresponding cost (for example, a cost corresponding to communication). A threshold system reduces the number of choices by only allowing a customer a second random choice if the load at its first choice exceeds a fixed threshold. The customer begins by choosing a single queue uniformly at random: if the queue length at this first choice (excluding the incoming customer) is at most T, the customer queues there; otherwise, the customer chooses a second queue uniformly at random (with replacement). Two variations are now possible. In the *weak threshold model*, the customer waits at the second queue, regardless of whether it is longer or shorter than the first. In the *strong threshold model*, the customer queues at the shorter of its two choices. (See Figure 4.) One could also expand both models so that a customer has several successive choices, with a different threshold set for each choice, up to any fixed number of choices; here we model only the case where a customer has at most two choices. Although threshold systems have been shown to perform well in practice [7, 18, 35], our results distinguishing these two models are new.

5.1 Limiting systems

We consider the limiting system for the weak threshold model. The rate at which a queue changes size depends on whether it has more or fewer than T customers. We first calculate $\frac{ds_i}{dt}$ in the case $i \leq T+1$. Let

$p_i = s_i - s_{i+1}$ be the fraction of queues with exactly i customers. An arriving customer becomes the ith customer in a queue if one of two events happen: either its first choice has $i-1$ customers, or its first choice has $T+1$ or more customers and its second choice has $i-1$ customers. Hence over a time interval dt the expected number of jumps from queues of size $i-1$ to i is $\lambda n(p_{i-1} + s_{T+1}p_{i-1})$. Similarly, the expected number of jumps from queues of size i to $i-1$ is $np_i dt$. Hence we find

$$\begin{aligned}\frac{ds_i}{dt} &= \lambda(p_{i-1} + s_{T+1}p_{i-1}) - p_i \ , \ i \leq T+1, \ \text{ or} \\ \frac{ds_i}{dt} &= \lambda(s_{i-1} - s_i)(1 + s_{T+1}) - (s_i - s_{i+1}) \ , \ i \leq T+1. \end{aligned} \tag{6}$$

The case where $i \geq T+1$ can be calculated similarly, yielding

$$\frac{ds_i}{dt} = \lambda(s_{i-1} - s_i)s_{T+1} - (s_i - s_{i+1}) \ , \ \ i > T+1. \tag{7}$$

We now determine the fixed point. As usual, $\pi_0 = 1$ and, because at the fixed point the rate at which customers arrive must equal the rate at which they leave, $\pi_1 = \lambda$. In this case we also need to find the value of π_{T+1} to be able to calculate further values of π_i. Using the fact that $\frac{ds_i}{dt} = 0$ at the fixed point yields that for $2 \leq i \leq T+1$,

$$\pi_i = \pi_{i-1} - \lambda(\pi_{i-2} - \pi_{i-1})(1 + \pi_{T+1}). \tag{8}$$

Recursively plugging in, we find

$$\pi_{T+1} = 1 - \frac{(1-\lambda)[((1+\pi_{T+1})\lambda)^{T+1} - 1]}{(1+\pi_{T+1})\lambda - 1}.$$

Given the threshold T, π_{T+1} can be computed effectively by finding the unique root between 0 and 1 of the above equation. (The root is unique as the left hand side is increasing in π_{T+1}, while the right hand side is decreasing in π_{T+1}.) Note that in this system the π_i *do not* decrease doubly exponentially, although they can decrease very quickly if π_{T+1} is sufficiently small.

The strong threshold model is given by the following differential equations:

$$\frac{ds_i}{dt} = \lambda(s_{i-1} - s_i)(1 + s_{T+1}) - (s_i - s_{i+1}) \ , \ \ i \leq T+1 \, ; \tag{9}$$

$$\frac{ds_i}{dt} = \lambda(s_{i-1}^2 - s_i^2) - (s_i - s_{i+1}) \ , \ \ i > T+1. \tag{10}$$

As equations (6) and (9) are the same, thre recurrence (8) also holds for the fixed point of the strong threshold system, so π_{T+1} for the strong threshold system is calculated similarly.

For small thresholds, the behavior of this system is very similar to that of the supermarket system, as has been noted empirically previously in [7] and [35]. In fact, the strong threshold model is double exponentially decreasing.

Lemma 12 *The fixed point for the strong threshold model decreases doubly exponentially.*

Proof: To show that the fixed point decreases doubly exponentially, we note that it is sufficient to show that $\pi_{T+j+1} = \lambda\pi_{T+j}^2$ for all $j \geq 1$, from which the lemma follows by a simple induction. Moreover, to prove that $\pi_{T+j+1} = \lambda\pi_{T+j}^2$ for all $j \geq 1$, it is sufficient to show that $\pi_{T+2} = \lambda\pi_{T+1}^2$. That this is sufficient follows from equation (10) and the fact that $\frac{ds_i}{dt} = 0$ at the fixed point, from which we obtain

$$\lambda\pi_{i-1}^2 - \pi_i = \lambda\pi_i^2 - \pi_{i+1}$$

for $i \geq T+2$.

Hence, to prove the lemma, we now need only show that $\pi_{T+2} = \lambda\pi_{T+1}^2$. From equation (9) we have

$$\pi_{T+2} \quad = \quad \pi_{T+1} - \lambda(\pi_T - \pi_{T+1})(1 + \pi_{T+1}),$$

which can be written in the form

$$\pi_{T+2} - \lambda\pi_{T+1}^2 \quad = \quad (1+\lambda)\pi_{T+1} - \lambda(1+\pi_{T+1})\pi_T. \tag{11}$$

We show that the right hand side of equation (11) is 0.

The recurrence (8) yields that

$$\lambda(\pi_{i-2} - \pi_{i-1})(1 + \pi_{T+1}) = \pi_{i-1} - \pi_i.$$

Summing the left and right hand sides of the above equation for all values of i in the range $2 \leq i \leq T+1$ yields

$$\lambda(1 - \pi_T)(1 + \pi_{T+1}) = \lambda - \pi_{T+1},$$

or more conveniently,

$$\lambda(1 + \pi_{T+1})\pi_T = (1+\lambda)\pi_{T+1}.$$

Hence the right hand side of equation (11) is 0 and the lemma is proved. ■

5.2 Convergence and stability

For the strong threshold model, we can show that the infinite system converges exponentially to the fixed point, as we have done for the supermarket model. Unfortunately, for the weak threshold model, we have only been able to prove stability. We present both proofs here, beginning with the stability of the weak model.

It is convenient to write the derivatives $d\epsilon_i/dt$ obtained from equations (6) and (7) in the following form:

$$\frac{d\epsilon_i}{dt} \quad = \quad \lambda(\epsilon_{i-1} - \epsilon_i)(1 + \pi_{T+1}) - (\epsilon_i - \epsilon_{i+1}) + \lambda\epsilon_{T+1}(s_{i-1} - s_i)\,, \quad i \leq T+1\,; \tag{12}$$

$$\frac{d\epsilon_i}{dt} \quad = \quad \lambda(\epsilon_{i-1} - \epsilon_i)\pi_{T+1} - (\epsilon_i - \epsilon_{i+1}) + \lambda\epsilon_{T+1}(s_{i-1} - s_i)\,, \quad i > T+1. \tag{13}$$

Notice that we have made all the terms appear linear in ϵ_i by leaving terms of the form $\lambda\epsilon_{T+1}(s_{i-1} - s_i)$ unexpanded.

Theorem 13 *The fixed point of the weak threshold model is stable.*

Proof: We shall assume the ϵ_i are non-zero; the case $\epsilon_i = 0$ can be handled as in Theorem 8. We examine the potential function given by the L_1-distance $D(t) = \sum_{i=1}^{\infty} |\epsilon_i(t)|$, and show that $\frac{dD}{dt} \leq 0$. As in Theorem 8 we collect all terms with a factor of ϵ_i. For $i \neq T+1$, it is simple to verify that all terms are linear in ϵ_i, and that the coefficient of sum of all such terms is at most 0. For example, for $i < T+1$, the sum of the terms in ϵ_i is

$$(-\lambda(1 + \pi_{T+1}) - 1)\epsilon_i \operatorname{sgn}(\epsilon_i) + \lambda(1 + \pi_{T+1})\epsilon_i \operatorname{sgn}(\epsilon_{i+1}) + \epsilon_i \operatorname{sgn}(\epsilon_{i-1}),$$

which is at most 0. The case $i > T+1$ is similarly straightforward.

The only difficulty arises in the ϵ_{T+1} term. Note the different form of the first expression on the right hand side of (12) and (13) : one has a factor of π_{T+1}, and one has a factor of $1+\pi_{T+1}$. Hence, in gathering the terms in ϵ_{T+1}, we have the following sum:

$$
\begin{aligned}
&(-\lambda(1+\pi_{T+1})-1)\epsilon_{T+1}\,\mathrm{sgn}(\epsilon_{T+1})+\lambda\pi_{T+1}\epsilon_{T+1}\,\mathrm{sgn}(\epsilon_{T+2})\\
&+\epsilon_{T+1}\,\mathrm{sgn}(\epsilon_T)+\epsilon_{T+1}\sum_{j=1}^{\infty}\lambda(s_{j-1}-s_j)\,\mathrm{sgn}(\epsilon_j).
\end{aligned}
$$

Let us suppose that ϵ_T, ϵ_{T+1}, and ϵ_{T+2} are all strictly positive; all other cases are similar. Then the above summation reduces to

$$
-\lambda\epsilon_{T+1}+\epsilon_{T+1}\sum_{j=1}^{\infty}\lambda(s_{j-1}-s_j)\,\mathrm{sgn}(\epsilon_j).
$$

The largest value the second expression can take is when $\mathrm{sgn}(\epsilon_j)=1$ for all j, in which case it is $\lambda\epsilon_{T+1}$. Hence, regardless of the signs of the remaining ϵ_i, we find that the coefficient of the sum of the terms in ϵ_{T+1} is also at most 0. ■

For the weak threshold model, proving convergence to the fixed point appears possible using the technique of [33], although their methods do not appear to provide bounds on the rate of covergence. (Note that stability does not imply convergence, nor does convergence imply our strong notion of stability, namely that the L_1 distance is non-increasing.)

We can, however, show that the strong threshold model does converge exponentially. As in Theorem 13, it will help us to rewrite the derivatives $\frac{d\epsilon_i}{dt}$ for the infinite system of the strong threshold model obtained from the equations (9) and (10) in the following form:

$$
\begin{aligned}
\frac{d\epsilon_i}{dt} &= \lambda(\epsilon_{i-1}-\epsilon_i)(1+\pi_{T+1})-(\epsilon_i-\epsilon_{i+1})+\lambda\epsilon_{T+1}(s_{i-1}-s_i)\,,\ \ i\le T+1\,; && (14)\\
\frac{d\epsilon_i}{dt} &= \lambda(\epsilon_{i-1}^2+2\pi_{i-1}\epsilon_{i-1}-\epsilon_i^2-2\pi_i\epsilon_i)-(\epsilon_i-\epsilon_{i+1})\,,\ \ i>T+1. && (15)
\end{aligned}
$$

Theorem 14 *The strong threshold model converges exponentially to its fixed point from any initial state where there exists a k such that $s_k(0)=0$.*

Proof: We shall find an increasing sequence w_i and $\delta>0$ such that for $\Phi(t)=\sum_i w_i|\epsilon_i(t)|$, we have $d\Phi/dt=-\delta\Phi$. As in Theorem 10, the proof will depend on finding a sequence w_i such that the terms of $d\Phi/dt$ in ϵ_i sum to at most $-\delta w_i|\epsilon_i|$. In fact, any sequence satisfying

$$
\begin{aligned}
w_{i+1} &\le w_i+\frac{w_i(1-\delta)-w_{i-1}}{\lambda(1+\pi_{T+1})}\,,\ \ i<T+1 && (16)\\
w_{i+1} &\le w_i+\frac{w_i(1-\delta)-w_{i-1}}{\lambda(1+2\pi_i)}\,,\ \ i\ge T+1 && (17)
\end{aligned}
$$

will suffice, and it is easy to verify that such sequences exist, as in Theorem 10. That this condition suffices can be easily checked by grouping all the ϵ_i terms from equations (14) and (15) for all ϵ_i except ϵ_{T+1}. The difficulty for the ϵ_{T+1} terms lies in the extraneous $\lambda\epsilon_{T+1}(s_{i-1}-s_i)$ terms in equation (14).

We now bound the sum of the terms in ϵ_{T+1}. We consider here only the case where all ϵ_i are positive; other cases are similar. The sum of all the terms in ϵ_{T+1} is

$$
\begin{aligned}
&(-\lambda(1+\pi_{T+1})-1)w_{T+1}\epsilon_{T+1}\,\mathrm{sgn}(\epsilon_{T+1})+\lambda(2\pi_{T+1}+\epsilon_{T+1})w_{T+2}\epsilon_{T+1}\,\mathrm{sgn}(\epsilon_{T+2})+\\
&w_T\epsilon_{T+1}\,\mathrm{sgn}(\epsilon_T)+\epsilon_{T+1}\sum_{j=1}^{T+1}w_j\lambda(s_{j-1}-s_j)\,\mathrm{sgn}(\epsilon_j).
\end{aligned}
$$

If all ϵ_i are positive this reduces to

$$(-\lambda(1+\pi_{T+1})-1)w_{T+1}\epsilon_{T+1}+\lambda(2\pi_{T+1}+\epsilon_{T+1})w_{T+2}\epsilon_{T+1}+w_T\epsilon_{T+1}+\epsilon_{T+1}\sum_{j=1}^{T+1}w_j\lambda(s_{j-1}-s_j).$$

As the w_i are increasing, the term $\epsilon_{T+1}\sum_{j=1}^{T+1}w_j\lambda(s_{j-1}-s_j)$ can be bounded above by

$$\epsilon_{T+1}\sum_{j=1}^{T+1}w_{T+1}\lambda(s_{j-1}-s_j)=\epsilon_{T+1}w_{T+1}\lambda(1-\pi_{T+1}-\epsilon_{T+1}).$$

Hence the sum of the terms in ϵ_{T+1} is bounded above by

$$(-\lambda(2\pi_{T+1}+\epsilon_{T+1})-1)w_{T+1}\epsilon_{T+1}+\lambda(2\pi_{T+1}+\epsilon_{T+1})w_{T+2}\epsilon_{T+1}+w_T\epsilon_{T+1},$$

and it is easily checked that equation (17) is sufficient to guarantee that this sum is at most $-\delta w_{T+1}\epsilon_{T+1}$.

Finally, we note that we may choose the w_i so that they are eventually dominated by a geometric series, as in Theorem 10. Since the tail of the fixed point for the strong threshold model decreases doubly exponentially by Lemma 12, we have

$$\Phi(0)=\sum_{i=1}^{\infty}w_i|\epsilon_i|=\sum_{i=1}^{k}w_i|\epsilon_i|+\sum_{i=k}^{\infty}w_i\pi_i$$

is finite. ∎

5.3 Simulations of threshold schemes

We first demonstrate the accuracy of the differential equations in describing system behavior. We consider the weak threshold scheme of Section 5 (where customers who make a second choice always queue at their second choice) with 100 queues at various arrival rates in Table 2. As before, simulations were done for 100,000 units of time with the first 10,000 thrown out for calculation purposes. For arrival rates up to 95% of the service rate, the predictions are within approximately 2% of the simulation results; with smaller arrival rates, the prediction is even more accurate. These results again demonstrates the accuracy of this approach.

We also compare the strong threshold scheme and the weak threshold scheme to the standard supermarket model where each customer always has two choices. Since the performance of the weak threshold scheme depends on the threshold chosen, we graph the best choice and second best choice for specific arrival rates λ. (Note the strong threshold scheme with the threshold set to 0 is equivalent to the supermarket model.) As one might expect, threshold schemes do not perform as well as the supermarket model (See Figure 5). It is worth noting, however, that even the weak threshold scheme performs almost as well for reasonable arrival rates (say $\lambda < 0.9$), despite the proven difference in the behavior of the tails (exponential versus doubly exponential dropoff). In many applications threshold schemes may be suitable, or even preferable, because they reduce the overall amount of communication that is necessary. Even though the threshold must be chosen appropriately to match the load, small thresholds are adequate over a large range of arrival rates.

6 Concluding remarks

We have demonstrated techniques for studying large decentralized systems that use simple, effective load balancing strategies, based on analyzing the corresponding infinite system. We have applied our methods

λ	Threshold	Simulation	Prediction	Relative Error (%)
0.50	0	1.3360	1.3333	0.2025
	1	1.4457	1.4444	0.0900
0.70	0	1.9635	1.9608	0.1377
	1	1.8144	1.8074	0.3873
	2	2.0150	2.0109	0.2039
0.80	0	2.7868	2.7778	0.3240
	1	2.2493	2.2346	0.6578
	2	2.3518	2.3387	0.5601
0.90	1	3.5322	3.4931	1.1194
	2	3.1497	3.1067	1.3841
	3	3.2903	3.2580	0.9914
0.95	2	4.5767	4.4464	2.9305
	3	4.2434	4.1274	2.8105
	4	4.3929	4.3061	2.0158
0.99	4	8.1969	7.4323	10.2875
	5	7.5253	6.8674	9.5800
	6	7.6375	6.9369	10.0996

Table 2: Simulations versus estimates for the weak threshold model: 100 queues.

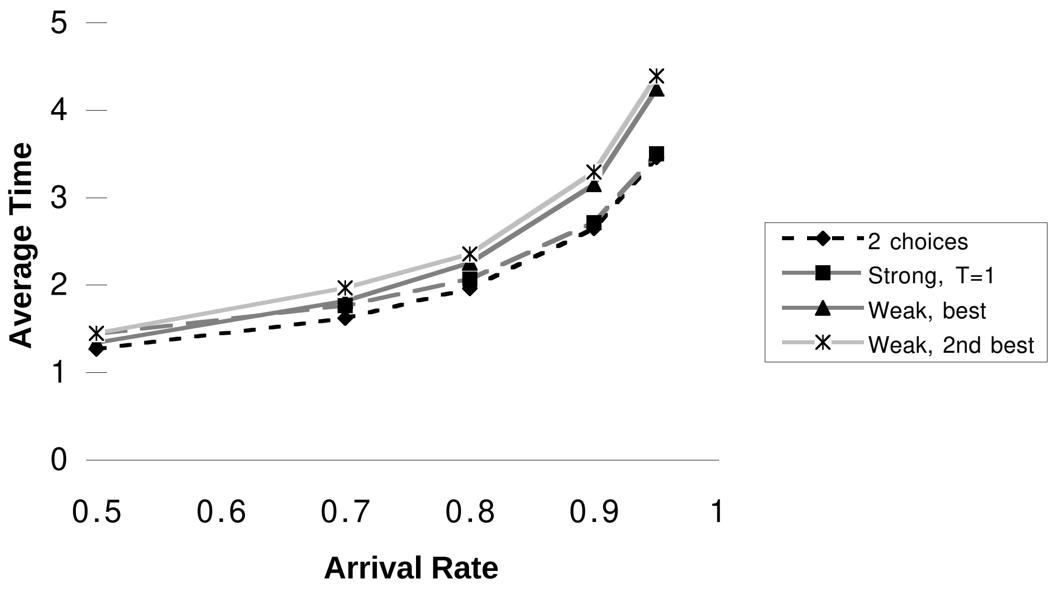

Figure 5: Comparison of the threshold models with two choices.

to the supermarket model and several variations, including the case of fixed service times and threshold systems. Besides allowing an analysis of these systems, our work demonstrates that there are important behavioral differences between systems where customers have one choice and systems where they have more than one choice. In particular, we have shown that using two choices can lead to an exponential improvement in the expected time in the system over using one choice; using more choices leads to much less substantial improvements.

Extrapolating from our results, we believe that the paradigm of using load information from a small random sample of possible destinations will prove effective in many load balancing scenarios. Indeed, the effectiveness of this general approach has been noted recently in practical load balancing scenarios [30] as well as for load profiling in real-time systems [6].

Although our methodology has been successful for several models, there remain several open questions. We conjecture that the closed model and the weak threshold model converge exponentially, although a proof appears to require different techniques than given here. The problem of analyzing the behavior of these simple randomized strategies on small systems and systems with fixed network topologies also appears to lie outside the range of our techniques. Finally, it would be interesting to test the performance of these methods in the context of more complex service and arrival distributions, such as heavy-tailed distributions.

References

[1] I. J. B. F. Adan, G. van Houtum, and J. van der Wal. Upper and lower bounds for the waiting time in the symmetric shortest queue system. *Annals of Operations Research*, 48:197–217, 1994.

[2] I. J. B. F. Adan, J. Wessels, and W. H. M. Zijm. Analysis of the symmetric shortest queue problem. *Stochastic Models*, 6:691–713, 1990.

[3] I. J. B. F. Adan, J. Wessels, and W. H. M. Zijm. Analysis of the asymmetric shortest queue problem. *Queueing Systems*, 8:1–58, 1991.

[4] M. Alanyali and B. Hajek. Analysis of simple algorithms for dynamic load balancing. In *INFOCOM '95*, 1995. To appear in Math. Oper. Res.

[5] Y. Azar, A. Broder, A. Karlin, and E. Upfal. Balanced allocations. In *Proceedings of the 26th ACM Symposium on the Theory of Computing*, pages 593–602, 1994.

[6] A. Bestavros. Load profiling: A methodology for scheduling real-time tasks in a distributed system. In *Proceedings of ICDCS'97: The IEEE International Conference on Distributed Computing Systems*, 1997.

[7] D. L. Eager, E. D. Lazokwska, and J. Zahorjan. Adaptive load sharing in homogeneous distributed systems. *IEEE Transactions on Software Engineering*, 12:662–675, 1986.

[8] D. L. Eager, E. D. Lazokwska, and J. Zahorjan. A comparison of receiver-initiated and sender-initiated adaptive load sharing. *Performance Evaluation Review*, 16:53–68, March 1986.

[9] D. L. Eager, E. D. Lazokwska, and J. Zahorjan. The limited performance benefits of migrating active processes for load sharing. *Performance Evaluation Review*, 16:63–72, May 1988. Special Issue on the 1988 SIGMETRICS Conference.

[10] S. N. Ethier and T. G. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley and Sons, 1986.

[11] B. Hajek. Asymptotic analysis of an assignment problem arising in a distributed communications protocol. In *Proceedings of the 27th Conference on Decision and Control*, pages 1455–1459, 1988.

[12] M. Harchol-Balter and D. Wolfe. Bounding delays in packet-routing networks. In *Proceedings of the Twenty-Seventh Annual ACM Symposium on the Theory of Computing*, pages 248–257, 1995.

[13] R. M. Karp, M. Luby, and F. Meyer auf der Heide. Efficient PRAM simulation on a distributed memory machine. In *Proceedings of the 24th ACM Symposium on the Theory of Computing*, pages 318–326, 1992.

[14] R. M. Karp and M. Sipser. Maximum matchings in sparse random graphs. In *Proceedings of the 22nd IEEE Symposium on Foundations of Computer Science*, pages 364–375, 1981.

[15] R. M. Karp, U. V. Vazirani, and V. V. Vazirani. An optimal algorithm for on-line bipartite matching. In *Proceedings of the 22nd ACM Symposium on the Theory of Computing*, pages 352–358, 1990.

[16] F. P. Kelly. *Reversibility and Stochastic Networks*. John Wiley and Sons, 1979.

[17] L. Kleinrock. *Queueing Systems, Volume I: Theory*. John Wiley and Sons, 1976.

[18] T. Kunz. The influence of different workload descriptions on a heuristic load balancing scheme. *IEEE Transactions on Software Engineering*, 17:725–730, 1991.

[19] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7:49–58, 1970.

[20] T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8:344–356, 1971.

[21] T. G. Kurtz. Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Applications*, 6:223–240, 1978.

[22] T. G. Kurtz. *Approximation of Population Processes. CBMS-NSF Regional Conf. Series in Applied Math*. SIAM, 1981.

[23] A. N. Michel and R. K. Miller. *Qualitative Analysis of Large Scale Dynamical Systems*. Academic Press, Inc., 1977.

[24] M. Mitzenmacher. How useful is old information? Submitted to PODC '97.

[25] M. Mitzenmacher. Bounds on the greedy routing algorithm for array networks. In *Proceedings of the Sixth Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 248–259, 1994. To appear in the *Journal of Computer Systems and Science*.

[26] M. Mitzenmacher. Constant time per edge is optimal on rooted tree networks. In *Proceedings of the Eighth Annual ACM Symposium on Parallel Algorithms and Architectures*, pages 162–169, 1996.

[27] M. Mitzenmacher. Density dependent jump Markov processes and applications to load balancing. In *Proceedings of the 37th IEEE Symposium on Foundations of Computer Science*, pages 213–222, 1996.

[28] M. Mitzenmacher. *The Power of Two Choices in Randomized Load Balancing*. PhD thesis, University of California, Berkeley, September 1996.

[29] R. Righter. and J. Shanthikumar. Extremal properties of the FIFO discipline in queueing networks. *Journal of Applied Probability*, 29:967–978, November 1992.

[30] J. R. Santos and R. Muntz. Design of the RIO (Randomized I/O) storage server. Computer science division, University of California, Los Angeles, May 1997.

[31] A. Shwartz and A. Weiss. *Large Deviations for Performance Analysis*. Chapman & Hall, 1995.

[32] G. D. Stamoulis and J. N. Tsitsiklis. The efficiency of greedy routing in hypercubes and butterflies. *IEEE Transactions on Communications*, 42(11):3051–3061, November 1994. An early version appeared in the *Proceedings of the Second Annual ACM Symposium on Parallel Algorithms and Architectures*, p. 248-259, 1991.

[33] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmission*, 32:15–27, 1996.

[34] N. C. Wormald. Differential equations for random processes and random graphs. *Annals of Appl. Prob.*, 17:1217–1235, 1995.

[35] S. Zhou. A trace-driven simulation study of dynamic load balancing. *IEEE Transactions on Software Engineering*, 14:1327–1341, 1988.

A From infinite to finite: Kurtz's theorem

In this section, we briefly describe the formal theory that connects the limiting system with systems of finite size, based on the work of Kurtz. As even stating an appropriate theorem requires a great deal of background and notation, we here provide only an informal argument; further explication with regard to load balancing problems is available in [28] or [33]; more general works covering the appropriat theory include [10, 31]. The supermarket model is an example of a *density dependent family of jump Markov processes*. Informally, such a family is a one parameter family of Markov processes, where the parameter n corresponds to the total population size (or, in some cases, area or volume). The states can be normalized and interpreted as measuring population densities, so that the transition rates depend only on these densities. As we have seen in Section 2.1, for the supermarket model the transition rates between states depend only upon the densities s_i. Hence the supermarket model fits our informal definition of a density dependent family. The limiting system corresponding to a density dependent family is the limiting model as the population size grows arbitrarily large.

Kurtz's work provides a basis for relating the limiting system for a density dependent family to the corresponding finite systems. Essentially, Kurtz's theorem provides a law of large numbers and Chernoff-like bounds for density dependent families. The primary differences between the limiting system and the finite system are:

- The limiting system is deterministic; the finite system is random.
- The limiting system is continuous; the finite system has jump sizes that are discrete values.

Imagine starting both systems from the same point for a small period of time. Since the jump rates for both processes are initially the same, they will have nearly the same behavior. Now suppose that if two points are close in the infinite dimensional space then their transition rates are also close; this is called the *Lipschitz condition*, and it is a precondition for Kurtz's theorem. Then even after the two processes separate, if they remain close, they will still have nearly the same behavior. Continuing this process inductively over time, we can bound how far the processes separate over any interval $[0, T]$.

The following theorem, which we state without proof, is derived from an application of Kurtz's results to the finite supermarket model to obtain bounds on the expected time a customer spends in the system.

Theorem 15 *For any fixed T, the expected time a customer spends in an initially empty supermarket system of size n over the interval $[0, T]$ is bounded above by*

$$\sum_{i=1}^{\infty} \lambda^{\frac{d^i - d}{d-1}} + o(1),$$

where the $o(1)$ is understood as $n \to \infty$ and may depend on T. ■

The $o(1)$ term in Theorem 15 is the correction for the finite system, while the main term is the expected time in the limiting system from Corollary 6.

Of course, similar theorems bounding the deviation of the infinite and finite processes hold for the other systems we have studied as well. Essentially, whenever the limiting system converges to a fixed point, the equilibrium distribution of the corresponding finite system is concentrated around the fixed point. Hence the fixed point may be used to give good approximations for such quantities as the average time a customer spends in the system. The discrepancy between the finite and limiting system is generally $o(1)$. In practice, as we have seen, for load balancing problem the discrepancy is small even when the number of queues n is relatively small.