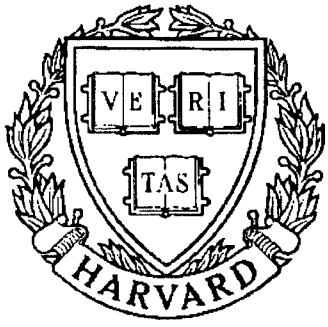
THESIS REPORT
Ph.D.

S Y S T E M S
R E S E A R C H
C E N T E R



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Heavy and Light Traffic Approximations for Queues with Synchronization Constraints

by S. Varma
Advisor: A.M. Makowski

Ph.D. 90-2

HEAVY AND LIGHT TRAFFIC APPROXIMATIONS FOR QUEUES WITH SYNCHRONIZATION CONSTRAINTS

by
Subir Varma

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1990

Advisory Committee:
Associate Professor Armand M. Makowski, Chairman/Advisor
Professor Ashok Agarwala
Assistant Professor Bernard L. Menezes
Associate Professor Prakash Narayan
Assistant Professor A. Yavuz Oruc

ABSTRACT

Title of Dissertation: Heavy and light traffic approximations for queues with synchronization constraints

Subir Varma, Doctor of Philosophy, 1990

Dissertation directed by: Armand M. Makowski
Associate Professor
Electrical Engineering Department

The aim of this dissertation is to develop approximations to performance measures for queues with synchronization constraints with the help of limit theorems. In particular we shall consider queues exhibiting the fork-join and resequencing constraints. These queues invariably exhibit non-product form behavior, and their analysis by any other method is extremely difficult. The limit theorems that we shall use come in two flavors, i.e. heavy traffic limit theorems and light traffic limit theorems. Heavy traffic limit theorems give estimates of the performance measures when the system is operating near its full capacity, while light traffic limit theorems give estimates of the performance measures when the system is very lightly loaded. By interpolating between these two limits it is possible to obtain estimates of the performance measures when the system operates at moderate loads.

DEDICATION

TO URVASHI

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to Dr. Armand Makowski for his guidance and support for the past five years. By his personal example, he sets a unique standard in research and presentation which I can only try to emulate in my professional carrier.

I would like to thank all my freinds, especially Tedijanto and Dimitris Tsakiris, for the many enjoyable discussions that we had over the years, on matters technical and otherwise.

I wish to thank my parents for their unending love, support and encouragement which made all this possible. Last but not the least, I would like to thank my wife Urvashi for her constant love and support and for creating the proper conditions at home, under which I was able to undertake the onerous task of researching and writing a PhD Dissertation.

TABLE OF CONTENTS

PART II

RESEQUENCING SYSTEMS

PART III

FORK-JOIN WITH RESEQUENCING

LIST OF FIGURES

CHAPTER I

1.1 Introduction

The aim of this dissertation is to develop approximations to performance measures for queues with synchronization constraints with the help of limit theorems. In particular we shall consider queues exhibiting the fork–join and resequencing constraints. These queues invariably exhibit non-product form behavior, and their analysis by any other method is extremely difficult. The limit theorems that we shall use come in two flavors, i.e. heavy traffic limit theorems and light traffic limit theorems. Heavy traffic limit theorems give estimates of the performance measures when the system is operating near its full capacity, while light traffic limit theorems give estimates of the performance measures when the system is very lightly loaded. By interpolating between these two limits it is possible to obtain estimates of the performance measures when the system operates at moderate loads.

Heavy traffic limit theorems are obtained by means of diffusion approximations. Intuitively this corresponds to replacing the discrete state or discrete time stochastic process under consideration by a diffusion process, with the understanding that in heavy traffic a scaled version of the original stochastic process behaves similarly to a diffusion process. The partial differential equations which are satisfied by the stationary distribution of the diffusion process can be obtained in some cases and then the burden of the analysis is transferred to finding the solutions to these equations.

A comprehensive theory for light traffic approximations has been developed recently by Reiman and Simon [59]. It is applicable to queueing systems the input of which is either a Poisson process or more generally, a process of the Phase type. Not only does the theory provide values of the performance measure in light traffic, but also more sensitive information in the form of the derivatives of the

performance measure with respect to the arrival rate. We can then combine the light and heavy traffic results to yield a polynomial expression (in the arrival rate) as an approximation to the performance measure. For example by using the heavy traffic limit, the light traffic limit, and the first derivative, we obtain a quadratic approximation. Some of the principal results of the light traffic theory are given in Appendix B.

The principle mathematical tool used in showing the convergence to a diffusion process is the theory of weak convergence of probability measures on the function spaces $C[0,1]$ or $D[0,1]$, an excellent treatment of which is given in [8] (also see Appendix A). Weak convergence in function spaces is the generalization of convergence in distribution, and is appropriate for stochastic processes.

Functional central limit theorems are readily available in the literature for basic processes like random walks and renewal processes [8]. With this in mind, the most commonly used method to obtain heavy-traffic limit theorems for queues has been to connect the stochastic processes which arise in queues to these basic processes. For example, limit theorems for the waiting time processes typically exploit recursive schemes such as Lindley's recursion, to connect them to random walks [65], while limit theorems for queue length processes make use of the connection between the queue lengths and the renewal processes generated by the arrivals and departures in a queue [29],[30].

A survey of the literature reveals that most of the work on diffusion approximations for queues has been limited to standard queueing systems. What we propose to do here is to extend the scope of this method to non-standard queueing systems which exhibit synchronization constraints in their behavior [1], [2], [3], [4], [5]. Such systems have assumed increasing importance in recent years in view of their applicability in modeling multi-processor architectures and distributed systems. In contrast to standard queues, it is difficult to solve these systems even under the assumptions of Poissonian arrivals and exponential service times. Therefore a theory of diffusion approximations for these queues has two virtues: not only will it provide a non-parametric body of results, but it will also enable

us to make estimates of performance measures, a task which seems otherwise to be difficult by any other method.

The aim of this introductory chapter is to acquaint the reader with some of the basic techniques used in obtaining limit theorem approximations. This is done by considering the special case of the $GI/GI/1$ queue. We consider this particular system for illustrative purposes since the theory for this queue is specially well developed owing to its simplicity. Moreover it was the first system to be analyzed using limit theorems.

The rest of the chapter is organized as follows: In Section 1.2 we consider heavy traffic approximations for the $GI/GI/1$ queue, and in Section 1.3 we discuss light traffic approximations for this queue as well as interpolations of the performance measures between heavy and light traffic. The final Section 1.4 contains a brief summary of the main results in the dissertation.

1.2 Heavy traffic for the $GI/GI/1$ queue

In this section we introduce the reader to the principle techniques of heavy traffic approximations by way of the $GI/GI/1$ queue. We begin the discussion by providing a recursive representation for the queueing delay process in this queue.

Let the following RVs be defined on a common probability space $(\Omega, \mathbb{F}, \mathbb{P})$. For $n = 0, 1 \ldots$, we set

u_{n+1} : Inter-arrival time between the $(n+1)^{rst}$ and n^{th} customers.

v_n : Service time of the n^{th} customer.

W_n : Waiting time of the n^{th} customer.

We shall assume that

(Ia): The sequences $\{u_{n+1}\}_0^\infty$ and $\{v_n\}_0^\infty$ are mutually independent sequences of iid RVs with

$$\begin{aligned} u &:= \mathbb{E}(u_n) < \infty, \quad \sigma_u^2 := Var(u_n) < \infty \\ v &:= \mathbb{E}(v_n) < \infty, \quad \sigma_v^2 := Var(v_n) < \infty. \end{aligned} \qquad n = 0, 1 \ldots$$

Assuming that the first customer arrives into an empty system at time $t = 0$, Lindley [48] gave the following recursive representation.

$$\begin{aligned} W_0 &= 0 \\ W_{n+1} &= [W_n + v_n - u_{n+1}]^+. \qquad n = 0, 1 \ldots \quad (2.1) \end{aligned}$$

We consider the system to be stable if the sequence of queueing delays $\{W_n\}_0^\infty$ converges weakly as $n \uparrow \infty$ to a proper RV W. It is well known [51] that the queue will be stable provided $v < u$.

We would like to obtain performance measures such as the average waiting time for this queue while it is operating in its stable regime. Unfortunately, in general there does not exist closed–form expressions for the average waiting time such as the Pollaczek-Khinchin formula for $M/GI/1$ queues. However Kingman [35],[36],[37] was able to prove the following interesting result which is independent of any distributional assumption on the inter-arrival or service times. Consider a

one parameter family of $GI/GI/1$ systems indexed by u and v and let

$$X_{n+1}(u,v) = v_n(v) - u_{n+1}(u) \qquad n = 0, 1 \ldots$$

Assume that

(Ib):

$$\sigma_u^2 + \sigma_v^2 \to \sigma^2, \text{ with } 0 < \sigma^2 < \infty, \text{ as } u \downarrow v$$

$$\sup_{u>v>0} [I\!E X_n^{2+\epsilon}(u,v)] < \infty \quad \text{for some} \quad \epsilon > 0$$

Under assumptions **(Ia)**–**(Ib)**, Kingman [36] considered the iterated limit, first letting $n \uparrow \infty$ to obtain the steady-state waiting time $W(u,v)$ for each $u > v$, and letting $u \downarrow v$ after normalization to obtain

$$\frac{2(u-v)}{\sigma_u^2 + \sigma_v^2} W(u,v) \xrightarrow{I\!D} E \tag{2.2a}$$

where $\xrightarrow{I\!D}$ denotes convergence in distribution and

$$I\!P(E \le x) = \begin{cases} 1 - e^{-x}, & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2b}$$

To use this result, assume that v is held fixed and that $u \downarrow v$. The following approximation

$$I\!E W(u,v) \approx \frac{\sigma_u^2 + \sigma_v^2}{2(u-v)}, \quad u \approx v \tag{2.3}$$

is then made plausible by the convergence (2.2).

Kingman concluded to (2.2) by making use of well-known results for the steady state waiting time for $GI/GI/1$ queues, e.g., Spitzer's identity [34]. However for more complicated systems, such results are usually not available and we have to find other techniques for obtaining heavy traffic approximations such as (2.2) and (2.3). One such technique is based on "diffusion limits" as explained below. These diffusion limits can be obtained by considering a sequence of appropriately normalized queueing processes instead of a sequence of appropriately normalized

steady-state distributions. Accordingly, consider a sequence of $GI/GI/1$ systems indexed by $r \geq 1$, each of which satisfies assumption **(Ia)**. Make the following additional assumptions **(Ic)**–**(Id)** where

(Ic): As $r \uparrow \infty$,

$$\sigma_u(r) \to \sigma_u$$

$$\sigma_v(r) \to \sigma_v$$

$$[u(r) - v(r)]\sqrt{r} \to c.$$

(Id): For some $\epsilon > 0$,

$$\sup_{r \geq 1} \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{\mid v_1(r) \mid^{2+\epsilon}\}\} < \infty.$$

In the next step we make use of the theory of weak convergence to obtain the diffusion limits. Define $D[0, \infty)$ to be the space of all real-valued right-continuous functions having left limits. This space is endowed with a metric defined by Lindvall [49], which makes it separable and complete (Appendix A).

For $r = 1, 2 \ldots$, the stochastic process $\mu_r \equiv \{\mu_r(t), t \geq 0\}$ with sample paths in $D[0, \infty)$ is defined by

$$\mu_t(r) = \frac{W_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0. \tag{2.4}$$

In order to obtain (2.2), Kingman first considered the limit of $\mu_t(r)$ as $t \uparrow \infty$ and then took the limit as $r \uparrow \infty$. In contrast, the methodology for obtaining diffusion limits directs us to take these two limits in the reverse order. We first let $r \uparrow \infty$ to obtain a diffusion process, and then let $t \uparrow \infty$ to obtain the stationary distribution of that diffusion process. It turns out that this stationary distribution is the same as the limiting distribution obtained by Kingmans's method. This interchange of limits has been justified by Prohorov [55] and Harrison [20].

Define a random process $\zeta \equiv \{\zeta_t, t \geq 0\}$ by

$$\zeta_t = \sigma_v \xi_t^v - \sigma_u \xi_t^u - ct, \quad t \geq 0 \tag{2.5}$$

where ξ^u and ξ^v are two independent Wiener processes defined on $[0,\infty)$. Next, define a mapping $g : D[0,\infty) \uparrow D[0,\infty)$ by

$$g(x)_t = x_t + \sup_{0\le s\le t} x_s^-, \quad t \ge 0. \tag{2.6}$$

For the subspace $D_0[0,\infty)$ of $D[0,\infty)$ for which $x_0 = 0$, the above definition simplifies to

$$g(x)_t = x_t - \inf_{0\le s\le t} x_s, \quad t \ge 0. \tag{2.7}$$

The mapping g is often referred to as the reflection mapping for the following reason. Consider an element in x in $D_0[0,\infty)$ such that $x_t \ge 0$ for all t in $[0,\infty)$. Then it is easy to see from (2.7) that $g(x) = x$ identically. However if the x becomes negative, then the mapping g acts in such a way that $g(x)$ is forced to stay positive. In this case x is said to have a normal reflection from the origin. Further discussion of the reflection mapping may be found in [24, Chap. 2] and [12, Chap. 8]

If we let $\Rightarrow$ denote weak convergence, then under assumptions **(Ia)** and **(Ic)-(Id)**, it can be shown [65] that

$$\mu(r) \Rightarrow g(\zeta) \tag{2.8}$$

as $r \uparrow \infty$, in $D[0,\infty)$.

The process $g(\zeta)$ is said to be the heavy traffic diffusion limit for the waiting time process in the $GI/GI/1$ queue. It is well known [24] that for all $t \ge 0$ and x in $\mathbb{R}$,

$$\mathbb{P}[g(\zeta)_t \le x] = \Phi\left(\frac{x+ct}{\sqrt{(\sigma_0^2+\sigma_1^2)t}}\right) - e^{-\frac{2cx}{\sigma_0^2+\sigma_1^2}}\Phi\left(\frac{-x+ct}{\sqrt{(\sigma_0^2+\sigma_1^2)t}}\right), \tag{2.9}$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \exp^{-\frac{x^2}{2}}\, dx, \quad z \text{ in } \mathbb{R}$$

We now take the limit in (2.9) as $t \uparrow \infty$ to obtain the stationary distribution of the diffusion limit $g(\zeta)$, namely

$$\lim_{t\uparrow\infty} I\!P[g(\zeta)_t \leq x] = 1 - e^{-\frac{2cx}{\sigma_0^2+\sigma_1^2}} \tag{2.10}$$

Hence by a different route, we have once again arrived at Kingman's exponential approximation (2.2).

1.3 Light traffic limits for the $GI/GI/1$ queue

The theory of light traffic approximations for queueing systems was initiated by Reiman and Simon [59]. We shall illustrate the main features of this theory in this section by applying it to the single server queue with arrival rate λ $(= \frac{1}{u})$ and service rate μ $(= \frac{1}{v})$. Appendix B contains a collection of some of the basic results of this theory.

Just as in heavy traffic theory we seek estimates of performance measures in the case when the arrival rate λ into the system approaches the maximum system utilization μ, in light traffic theory we seek estimates of performance measures when the arrival rate into the system approaches zero. In the Reiman-Simon theory, the arrival process is restricted to be either a Poisson process or a process driven by a Poisson process, e.g., a non-stationary Poisson process or a renewal process with a phase type renewal distribution. Hence we shall hereafter assume that the queue under consideration is of the $M/GI/1$ type so that the light traffic theory applies to it. Note that exact solutions already exist for the $M/GI/1$ queue [39], so that the limit theorem approximations are redundant for this case. However for other systems like those with fork–join or resequencing constraints, exact solutions are no longer available and limit theorem approximations are much more useful.

The performance measure we consider is the average waiting time in the $M/GI/1$ queue, since we have already developed heavy traffic limits for this measure in the last section. The un-normalized light traffic limit of the average waiting time is trivial, since in the limit as the arrival rate goes to zero, there are no customers in the system. However, as Reiman and Simon show, it is possible to obtain more sensitive information on light traffic behavior by calculating the derivatives of the average waiting time with respect to λ at $\lambda = 0$. Calculation of the n^{th} derivative requires consideration of the $M/GI/1$ queue in statistical equilibrium with a total of n arrivals in the interval $(-\infty, +\infty)$. Once we know the values of the average waiting time at $\lambda = 0$ and of its first n derivatives, a natural procedure would be to approximate it in the range $[0, \mu)$ by means of a Taylor expansion,

which has the form of an n^{th} degree polynomial. However a polynomial approximation would not be appropriate, since we know that typically some performance measures blow as $\lambda \uparrow \mu$. However heavy traffic approximation theory helps us get better estimates in the following way. If $\overline{W}(\lambda)$ is the average waiting time when the arrival rate is λ, then heavy traffic theory (2.3) shows that

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{W}(\lambda) = \frac{1 + \mu^2\sigma_1^2}{2}. \tag{3.1}$$

Hence if we normalize $\overline{W}(\lambda)$ by $(\mu - \lambda)$ and consider the function $w(\lambda) = (\mu - \lambda)\overline{W}(\lambda)$, then a polynomial approximation is more reasonable. Using the n derivatives at the origin and the heavy traffic approximation, our final approximation would be a $(n+1)^{rst}$ degree polynomial in λ.

On the space $(\Omega, I\!F)$ define a measure $I\!P_\lambda$ which renders the inter-arrival time sequence $\{u_n\}_0^\infty$ exponential with rate λ, so that the arrival process into the system is a Poisson process. For each ω in Ω we add a tagged customer which arrives at time $t = 0$, and whose service time v^* is independent of the sequence $\{v_n\}_0^\infty$ but has the same distribution. Let

$$W = \text{waiting time of the tagged customer entering at t} = 0. \tag{3.2}$$

Also let $\psi(\omega)$ be the tagged customer's expected waiting time for the sample ω, i.e. averaged over v^*,

$$\psi(\omega) = I\!E[W \mid I\!F] \tag{3.3}$$

The average waiting time is then given by the formula

$$\overline{W}(\lambda) = \int \psi dI\!P_\lambda. \tag{3.4}$$

Define the expectations

$$\overline{\psi}(\{t\}) = I\!E[\psi \mid \text{ one arrival at time } t], \quad t \text{ in } I\!R \tag{3.5a}$$

and

$$\overline{\psi}(\emptyset) = I\!E[\psi \mid \text{no arrivals}] \tag{3.5b}$$

Note that these expectations do not depend on λ. As shown in Appendix B, we have

$$\overline{W}(0) = \overline{\psi}(\emptyset) \tag{3.6}$$

and

$$\overline{W}'(0) = \int_{-\infty}^{+\infty} (\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset))dt. \tag{3.7}$$

From (3.6) it is clear that $\overline{W}(0) = 0$ and we now proceed to calculate $\overline{W}'(0)$. Let $w_t(s)$ denote the waiting time of a customer arriving at time $t = 0$ when another customer arrives at time t with service time s. Obviously, we have

$$w_t(s) = \begin{cases} 0 & \text{if } t > 0, \\ [s+t]^+ & \text{if } t \leq 0. \end{cases} \tag{3.8}$$

so that by (3.5a),

$$\overline{\psi}(\{t\}) = \int_0^{\infty} [s+t]^+ F(ds) \quad t \leq 0 \tag{3.9}$$

where F is the distribution function of the service time. Combining (3.9) with (3.7), we obtain

$$\overline{W}'(0) = \int_0^{\infty} \left[\int_s^{\infty} [s-t]^+ F(ds) \right] dt \tag{3.10}$$

$$= \frac{1}{2}(\sigma_V^2 + \frac{1}{\mu^2}) \tag{3.11}$$

after interchanging the order of integration and simplifying.

We now combine the results on light and heavy traffic to obtain an approximation which is valid for all values of the traffic intensity. As before let $w(\lambda) = (\mu - \lambda)\overline{W}(\lambda)$ for $0 \leq \lambda < \mu$. From the light traffic results (3.6) and (3.11), we have

$$w(0) = 0 \text{ and } w'(0) = \mu\overline{W}'(0) = \frac{\mu}{2}(\sigma_V^2 + \frac{1}{\mu^2}) \tag{3.12}$$

while from (3.1) we have

$$w(\mu) = \frac{1}{2}(1 + \mu^2\sigma_V^2). \tag{3.13}$$

Let $\hat{w}(\lambda)$ denote the quadratic interpolation of $w(\lambda)$ over the interval $[0, \mu]$ based on (3.12) and (3.13), i.e.,

$$\hat{w}(\lambda) = k_0 + k_1\lambda + k_2\lambda^2. \tag{3.14}$$

Using (3.12)–(3.13), we now come to the conclusion that

$$k_0 = k_2 = 0 \text{ and } k_1 = \frac{\mu}{2}(\sigma_V^2 + \frac{1}{\mu^2}), \tag{3.15}$$

whence

$$\hat{w}(\lambda) = \frac{\lambda\mu}{2}(\sigma_V^2 + \frac{1}{\mu^2}), \quad 0 \leq \lambda \leq \mu. \tag{3.16}$$

Finally, we undo the normalization to obtain an approximation $\hat{W}(\lambda)$ to the average waiting time as

$$\hat{W}(\lambda) = \frac{\lambda\mu}{2}\frac{(\sigma_V^2 + \frac{1}{\mu^2})}{(\mu - \lambda)}, \quad 0 \leq \lambda \leq \mu. \tag{3.17}$$

However note that (3.17) is just the Pollaczek-Khinchine formula for the average delay in the $M/GI/1$ queue. Hence for the simple case of the $M/G/1$ queue, this approximation method yields the exact answer.

1.4 Main results in the dissertation

The dissertation is divided into three parts. In Part I we consider limit theorem approximations for fork–join queues and in Part II limit theorem approximations for queues with resequencing. Lastly in Part III we analyze a model which exhibits both the fork–join and the resequencing constraints.

Part I is subdivided into six chapters. In Chapter 2 the convergence to diffusion processes is demonstrated for fork–join systems in heavy traffic. We first consider single stage fork–join queues and then general acyclic fork–join networks as defined by Baccelli, Massey and Towsley [5]. In Chapters 3 to 5 we concentrate on obtaining the invariant distribution of the limiting diffusion for the end-to-end delay in single stage fork–join queues. In Chapter 3 we obtain upper and lower bounds to this invariant distribution using ideas from stochastic ordering theory. In Chapter 4, following Harrison and Reiman [22], we obtain a partial differential equation for the invariant joint distribution of the K-dimensional delay process. This partial differential equation is solved in Chapter 5 for the special case when $K = 2$ and the two queues are identical. We thereby obtain an expression for the invariant distribution for the end-to-end delay, which we then use to obtain formulae for all its moments in heavy traffic. Chapter 6 is devoted to obtaining interpolation approximations for the fork–join queue by utilizing information about its light and heavy traffic limits. We also present a formula for the heavy traffic limit for a K dimensional system which agrees extremely well with experimental results.

In Part II, we obtain limit theorem approximations for several resequencing systems. In Chapter 7 we obtain the limiting diffusion in heavy traffic for a resequencing model possessing a disordering system with an infinite number of servers followed by resequencing and then service at a single server queue. We show that the normalized queueing delay sequence in the buffer of the single server queue converges to a reflected Wiener process in heavy traffic. We also show that this result continues to hold in the case when there are K infinite server disordering stages (with resequencing after each disordering), with the single server

queue as the final stage. We also obtain diffusion limits for the case when the disordering system consists of K single server queues operating in parallel. Chapter 8 is devoted to obtaining light traffic limits for some of the resequencing models described earlier as well as polynomial approximations which hold for moderate values of the traffic intensity.

Part III deals with a model which exhibits both fork–join and resequencing synchronization constraints. It is similar to the acylic fork–join network for which limit theorems were obtained in Chapter 2, except that every single server queue is preceded by an infinite server disordering system, followed by a resequencing box. This model is being introduced here for the first time, and it subsumes most of the different fork–join and resequencing models that we have analyzed so far. We obtain the basic recursions governing this model and derive its stability conditions. Our main result regarding this model, is that it has the same heavy traffic diffusion limit as the acyclic fork–join network from Chapter 2. Hence in heavy traffic the effect of resequencing on the queueing delays of this model is negligible. There is an interesting special case of this model for which we obtain polynomial approximations by interpolating between heavy traffic and light traffic limits. This was a model originally proposed by Baccelli [3] to model time stamp ordering in a distributed system.

CHAPTER II

2.1 Introduction

In this chapter our objective is to obtain heavy traffic diffusion limits for fork–join queueing systems. In Section 2.2 we deal with a single stage fork–join queue. The diffusion limits for this system are an easy extension of the limits for the $GI/GI/1$ queue. We show that the vector stochastic process generated by the queue delay sequences converges weakly to a K–dimensional correlated diffusion process in the non-negative orthant with normal reflections at the boundaries. Also the stochastic process generated by the end-to-end delay sequence converges weakly to a process which is the maximum of these K diffusions. Similar diffusion limits are shown to hold for the queue length processes in Section 2.3.

In Section 2.4 we prove heavy traffic diffusion limits for the class of acyclic fork-join networks introduced by Baccelli, Massey and Towsley [5]. We show that the vector stochastic process generated by the queue delay sequences converges weakly to a complicated function of a K–dimensional correlated diffusion process.

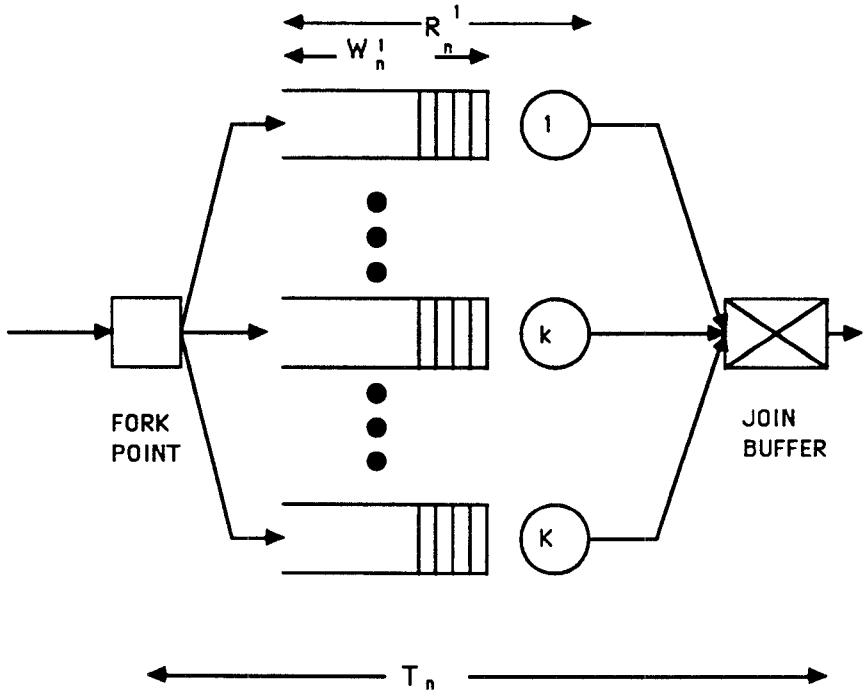


Fig. 2.1. A parallel fork–join queue.

2.2 Heavy traffic limit for parallel fork–join queue delays.

2.2.1 The model.

Consider a system of K single server queues operating in parallel. Each queue has an infinite capacity buffer and operates according to the FIFO discipline. There is a single stream of batch arrivals into the system so that every incoming batch splits into K distinct customers, with each customer entering a different buffer. This is known as the fork synchronization constraint. After a customer receives service, it may have to wait in another buffer until the other $(K-1)$ customers belonging to its batch have finished their service at the other queues, at which time the K customers all leave simultaneously. This is known as the join synchronization constraint.

Such queueing models arise in many application areas, including flexible manufacturing and parallel processing, with a wide variety of interpretations. The dif-

ficulty in analyzing such queueing models arises from the fact that the K queues are highly correlated due to the common arrival stream into their buffers. Previous approaches used complex variable theory [2], [14] and were confined to the case $K = 2$; even then the analysis was very tedious and involved.

As explained in the introductory chapter, we seek approximations to the moments of the response time (defined below) of the fork–join queue by means of limit theorems. In this chapter we initiate the process of obtaining a heavy traffic approximation for fork–join queues by establishing a heavy traffic diffusion limit for the delay processes, i.e., we show that an appropriately scaled and interpolated version of the delay vectors, converges in the heavy traffic limit to a K–dimensional correlated Wiener process with reflections.

The following RVs are defined on a common probability space $(\Omega, I\!F, I\!P)$. For $n = 0, 1 \ldots$ and $1 \leq k \leq K$,

u_{n+1} : Inter-arrival time between the $(n+1)^{rst}$ and n^{th} batch arrivals.

v_n^k : Service time of the customer belonging to the n^{th} batch which was sent to the k^{th} server.

W_n^k : Waiting time of the customer belonging to the n^{th} batch which was sent to the k^{th} server.

R_n^k : Response time of the customer belonging to the n^{th} batch which was sent to the k^{th} server.

T_n : System response time of the n^{th} batch.

We shall assume that

(IIa): The sequences $\{u_{n+1}\}_0^\infty$ and $\{v_n^k\}_0^\infty, 1 \leq k \leq K$, are iid with finite second moments, and mutually independent.

For $n = 0, 1 \ldots$, we set

$$u = I\!E(u_{n+1}) < \infty, \quad \sigma_0^2 = Var(u_{n+1}) < \infty$$

and

$$v^k = I\!E(v_n^k) < \infty, \quad \sigma_k^2 = Var(v_n^k) < \infty, \quad 1 \leq k \leq K$$

Under assumptions **(IIa)**, each queue in the fork–join system operates like a

$GI/GI/1$ queue. However the analysis of the queueing system is complicated by the fact that the K queues do not operate independently, owing to the common arrival stream.

2.2.2 Recursive representation for the delays

Assuming that the initial batch arrives into an empty system at time $t = 0$, we proceed to write down the Lindley recursion for the sequence of waiting times in the k^{th} queue, i.e., for each $1 \leq k \leq K$,

$$W_0^k = 0$$
$$W_{n+1}^k = [W_n^k + v_n^k - u_{n+1}]^+. \qquad n = 0, 1 \ldots (2.1)$$

The response time $R_n^k, 1 \leq k \leq K$, is given by

$$R_n^k = W_n^k + v_n^k, \qquad n = 0, 1 \ldots (2.2)$$

and the system response time T_n of the n^{th} batch is then given by

$$T_n = \max_{1 \leq k \leq K} R_n^k. \qquad n = 0, 1 \ldots (2.3)$$

We consider the system to be stable if the sequence of queueing delay vectors $\{(W_n^1, \ldots, W_n^K)\}_0^\infty$ converges in distribution as $n \uparrow \infty$ to a proper random vector $(W^1, \ldots, W^K)$. It is well known [4] that the condition

$$v^k < u, \quad 1 \leq k \leq K$$

is necessary and sufficient to insure stability.

2.2.3 The diffusion limit

We now proceed with the task of obtaining heavy traffic diffusion limits for the delay processes in the fork–join queue. For reasons explained in Chapter 1, we consider a sequence of fork–join systems approaching instability and show that a re–scaled K–dimensional stochastic process generated by the vector delay

sequence converges weakly to a K–dimensional correlated diffusion process in the non–negative orthant, with normal reflections at the boundaries. The attained convergence results give convergence over the interval $[0,\infty)$. However in the proofs we limit ourselves to proving convergence over any finite interval $[0,T]$, since the two cases are equivalent as long as the limiting process obtained has continuous sample paths [49, Thm. 3', pp. 120].

We now consider a sequence of fork–join systems indexed by $r = 1,2\ldots$, each of which satisfies assumption **(IIa)**. We make the following additional assumptions **(IIb)–(IIc)**, where

(IIb): As $r \uparrow \infty$,

$$\sigma_k(r) \to \sigma_k, \quad 0 \le k \le K,$$
$$[u(r) - v^k(r)]\sqrt{r} \to c_k, \quad 1 \le k \le K.$$

(IIc): For some $\epsilon > 0$,

$$\sup_{r,k} \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{\mid v_1^k(r) \mid^{2+\epsilon}\}\} < \infty.$$

For $r = 1,2\ldots$, define the following partial sums

$$V_0^k(r) = 0,$$
$$V_n^k(r) = v_0^k(r) + \ldots + v_{n-1}^k(r), \quad 1 \le k \le K, \ n = 1,2\ldots \quad (2.4a)$$

and

$$U_0(r) = 0,$$
$$U_n(r) = u_1(r) + \ldots + u_n(r). \qquad n = 1,2\ldots \quad (2.4b)$$

For $r = 1,2\ldots$, define the stochastic processes $\xi^k(r) \equiv \{\xi_t^k(r), t \ge 0\}, 0 \le k \le K$, with sample paths in $D[0,\infty)$ by

$$\xi_t^0(r) = \frac{U_{[rt]}(r) - u(r)[rt]}{\sqrt{r}}, \quad t \ge 0 \qquad (2.5a)$$

and

$$\xi_t^k(r) = \frac{V_{[rt]}^k(r) - v^k(r)[rt]}{\sqrt{r}}, \quad 1 \le k \le K, \quad t \ge 0. \tag{2.5b}$$

Let $\xi^k \equiv \{\xi_t^k, t \ge 0\}, 0 \le k \le K$, be $K+1$ independent Wiener processes. Lemma 2.2.1 shows that that the stochastic processes defined in (2.5) converge weakly to these Wiener processes.

Lemma 2.2.1. *As* $r \uparrow \infty$,

$$(\xi^0(r), \xi^1(r), \ldots, \xi^K(r)) \Rightarrow (\sigma_0\xi^0, \sigma_1\xi^1, \ldots, \sigma_K\xi^K) \tag{2.6}$$

in $D[0,\infty)^{K+1}$.

Proof. Equation (2.6) follows directly by Prohorov's functional central limit theorem for triangular arrays (see Appendix A, Theorem A3) under assumptions **(IIa)-(IIc)**. ∎

For $r = 1, 2 \ldots$, set

$$\begin{aligned} S_0^k(r) &= 0 \\ S_n^k(r) &= V_n^k(r) - U_n(r), \ 1 \le k \le K \qquad n = 1, 2 \ldots \end{aligned} \tag{2.7}$$

and define the stochastic processes $\zeta^k(r) \equiv \{\zeta_t^k(r), t \ge 0\}, 1 \le k \le K$, with sample paths in $D[0,\infty)$ by

$$\zeta_t^k(r) = \frac{S_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \le k \le K, \quad t \ge 0. \tag{2.8}$$

Also define the stochastic processes $\zeta^k \equiv \{\zeta_t^k, t \ge 0\}, 1 \le k \le K$, by

$$\zeta_t^k = \sigma_k\xi_t^k - \sigma_0\xi_t^0 - c_k t, \quad 1 \le k \le K, \quad t \ge 0. \tag{2.9}$$

The process $(\zeta^1, \ldots, \zeta^K)$ is a K–dimensional diffusion process with drift vector c and covariance matrix R given by

$$c = (-c_1, \ldots, -c_K) \tag{2.10}$$

and

$$R = \begin{pmatrix} \sigma_1^2 + \sigma_0^2 & \sigma_0^2 & \dots & \sigma_0^2 \\ \sigma_0^2 & \sigma_2^2 + \sigma_0^2 & \dots & \sigma_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_0^2 & \sigma_0^2 & \dots & \sigma_K^2 + \sigma_0^2 \end{pmatrix}. \tag{2.11}$$

Lemma 2.2.2 shows that the stochastic processes (2.8) generated by the random walk (2.7) converge to $(\zeta^1, \dots, \zeta^K)$ in the limit. The cross-correlation terms in the matrix R reflect the correlation between the K queues due to the common arrival process.

Lemma 2.2.2. *As* $r \uparrow \infty$,

$$(\zeta^1(r), \dots, \zeta^K(r)) \Rightarrow (\zeta^1, \dots, \zeta^K) \tag{2.12}$$

in $D[0, \infty)^K$.

Proof. Fix $r \geq 1$ and $t \geq 0$. For all $1 \leq k \leq K$, we see from (2.7) that

$$\begin{aligned} \zeta_t^k(r) &= \frac{V_{[rt]}^k(r) - U_{[rt]}}{\sqrt{r}} \\ &= \frac{V_{[rt]}^k(r) - v^k(r)[rt]}{\sqrt{r}} - \frac{U_{[rt]}^k(r) - u(r)[rt]}{\sqrt{r}} - \frac{[rt][u(r) - v^k(r)]}{\sqrt{r}} \\ &= \xi_t^k(r) - \xi_t^0(r) - \frac{[rt]}{r}[u(r) - v^k(r)]\sqrt{r} \end{aligned}$$

From assumption **(IIb)** it is clear that as $r \uparrow \infty$,

$$\frac{[rt]}{r}[u(r) - v^k(r)]\sqrt{r} \to c_k t, \quad 1 \leq k \leq K$$

and we conclude to (2.12) by invoking Lemma 2.2.1 and the continuous mapping theorem (Appendix A, Theorem A2).

∎

The Lindley recursion (2.1) for the queueing delays can be reformulated in the following way, which proves very useful in establishing limit theorems. For

$r = 1, 2 \ldots$ and $1 \leq k \leq K$, observe that

$$W_n^k(r) = \max\{S_n^k(r) - S_i^k(r) : i = 0, 1 \ldots, n\}$$
$$= S_n^k(r) - \min\{S_i^k(r) : i = 0, 1 \ldots, n\}, \; n = 0, 1 \ldots \quad (2.13)$$

For $r = 1, 2 \ldots$, we now define the stochastic processes $\mu^k(r) \equiv \{\mu_t^k(r), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$\mu_t^k(r) = \frac{W_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \quad (2.14)$$

We also define the stochastic processes $\eta^k \equiv \{\eta_t^k, t \geq 0\}, 1 \leq k \leq K$, by

$$\eta_t^k = g(\zeta^k)_t, \quad 1 \leq k \leq K, \quad t \geq 0. \quad (2.15)$$

where g denotes the reflection mapping (1.2.7).

In Lemma 2.2.3 we show that the vector process associated with (2.14), converges weakly to a K–dimensional diffusion process (2.15) with drift (2.10) and covariance (2.11). This limiting diffusion stays in the non–negative orthant of the K–dimensional space and exhibits normal reflections at the boundaries.

Lemma 2.2.3. *As* $r \uparrow \infty$,

$$(\mu^1(r), \ldots, \mu^K(r)) \Rightarrow (\eta^1, \ldots, \eta^K) \quad (2.16)$$

in $D[0, \infty)^K$.

Proof. From (2.13) and (2.14), we conclude for each $r = 1, 2, \ldots$, that

$$\mu^k(r) = g(\zeta^k(r)), \quad 1 \leq k \leq K.$$

Since g is a continuous mapping [69], the result follows by the continuous mapping theorem and Lemma 2.2.2. ∎

For $r = 1, 2 \ldots$, define the stochastic processes $\eta^k(r) \equiv \{\eta_t^k(r), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$\eta_t^k(r) = \frac{R_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \quad (2.17)$$

In Lemma 2.2.4 we show that the vector process defined by (2.17) converges weakly to the same limiting process as the vector process generated by the waiting times.

Lemma 2.2.4. *As* $r \uparrow \infty$,

$$(\eta^1(r), \ldots, \eta^K(r)) \Rightarrow (\eta^1, \ldots, \eta^K) \tag{2.18}$$

in $D[0,\infty)^K$.

Proof. Fix $T > 0$. We first show that

$$\sup_{0\le t\le T} \max_{1\le k\le K} \mid \eta_t^k(r) - \mu_t^k(r) \mid \xrightarrow{I\!P} 0 \quad \text{as } r \uparrow \infty.$$

For each $r = 1, 2 \ldots$, we see from (2.2) that

$$R_n^k(r) - W_n^k(r) = v_n^k(r), \quad 1 \le k \le K \qquad\qquad n = 0, 1 \ldots$$

whence

$$\sup_{0\le t\le T} \max_{1\le k\le K} \mid \eta_t^k(r) - \mu_t^k(r) \mid = \frac{1}{\sqrt{r}} \max_{1\le k\le K} \max\{v_n^k(r) : 0 \le n \le r\} \xrightarrow{I\!P} 0$$

as $r \uparrow \infty$, where this convergence ie validated by Theorem A5 in Appendix A. We conclude (2.18) from Lemma 2.2.3 and the converging together theorem. ∎

For $r = 1, 2 \ldots$, define the stochastic processes $\kappa(r) \equiv \{\kappa_t(r), t \ge 0\}$ with sample paths in $D[0,\infty)$ by

$$\kappa_t(r) = \frac{T_{[rt]}(r)}{\sqrt{r}}, \quad t \ge 0. \tag{2.19}$$

Also define the stochastic process $\kappa \equiv \{\kappa_t, t \ge 0\}$ with sample paths in $D[0,\infty)$ by

$$\kappa_t = \max_{1\le k\le K} \eta_t^k, \quad t \ge 0. \tag{2.20}$$

In Theorem 2.2.1 we show that the stochastic process (2.19) generated by the end-to-end delays, converges weakly to the process (2.20), which is the maximum of

K correlated Wiener processes with drift, in the non-negative orthant and normal reflection at the boundaries.

Theorem 2.2.1. *As* $r \uparrow \infty$,

$$\kappa(r) \Rightarrow \kappa \tag{2.21}$$

in $D[0,\infty)$.

Proof. From (2.2), (2.3) and (2.19) we conclude for each $r = 1, 2 \ldots$, that

$$\kappa_t(r) = \max_{1 \le k \le K} \eta_t^k(r), \quad t \ge 0.$$

Equation (2.21) now follows from Lemma 2.2.4 upon noting that $x \to \max_{1 \le k \le K} x_k$ is a continuous function on $\mathbb{R}^K$, and then applying the continuous mapping theorem. ∎

2.2.4 Interchange of limits

Recall that in Chapter I we made a distinction between the asymptotic distributions obtained depending upon the order in which the limits for r and t were taken for the single server queue waiting time process. Our objective is to show that for the fork–join queue, the stationary distribution for the normalized vector of response times is the same, regardless of the order in which these limits are taken. We begin by providing a sufficient condition for the vector diffusion process $(\eta^1, \ldots, \eta^K)$ to have a stationary distribution.

Proposition 2.2.1 *The condition* $c_k > 0, 1 \le k \le K$, *is necessary and sufficient to ensure that the* K*- dimensional process* $(\eta^1, \ldots, \eta^K)$ *converges in distribution to a proper vector* $(\eta_\infty^1, \ldots, \eta_\infty^K)$ *as* $t \uparrow \infty$.

Proof. For each $r = 1, 2 \ldots$, set

$$\tilde{W}_n^k(r) := \max_{0 \le i \le n} S_i^k(r), \quad 1 \le k \le K \tag{2.22}$$

and observe that

$$(W_n^1(r), \ldots, W_n^K(r)) =_{st} (\tilde{W}_n(r), \ldots, \tilde{W}_n^K(r)). \qquad n = 0, 1 \ldots \tag{2.23}$$

For $r = 1, 2 \ldots$, we now define the stochastic processes $\tilde{\mu}^k(r) \equiv \{\tilde{\mu}_t^k(r), t \geq 0\}, 1 \leq k \leq K$, by

$$\begin{aligned} \tilde{\mu}_t^k(r) &= \frac{\tilde{W}_{[rt]}^k(r)}{\sqrt{r}} \\ &= \max_{0 \leq s \leq t} \frac{S_{[rs]}^k(r)}{\sqrt{r}} \\ &= \max_{0 \leq s \leq t} \zeta_s^k(r), \quad 1 \leq k \leq K, \quad t \geq 0. \end{aligned} \tag{2.24}$$

Also define the stochastic processes $\tilde{\mu}^k \equiv \{\tilde{\mu}_t^k, t \geq 0\}, 1 \leq k \leq K$, by

$$\tilde{\mu}_t^k := \sup_{0 \leq s \leq t} \zeta_s^k, \quad 1 \leq k \leq K, \quad t \geq 0.$$

Using Lemma 2.2.2 we conclude that as $r \uparrow \infty$,

$$(\tilde{\mu}^1(r), \ldots, \tilde{\mu}^K(r)) \Rightarrow (\tilde{\mu}^1, \ldots, \tilde{\mu}^K) \tag{2.25}$$

in $D[0, \infty)^K$, and in particular

$$(\tilde{\mu}_t^1(r), \ldots, \tilde{\mu}_t^K(r)) \xrightarrow{\mathbb{D}} (\tilde{\mu}_t^1, \ldots, \tilde{\mu}_t^K), \quad t \geq 0. \tag{2.26}$$

However, (2.23) is equivalent to

$$(\tilde{\mu}_t^1(r), \ldots, \tilde{\mu}_t^K(r)) =_{st} (\mu_t^1(r), \ldots, \mu_t^K(r)), \quad r = 1, 2 \ldots, \quad t \geq 0 \tag{2.27}$$

and (2.26) thus implies that as $r \uparrow \infty$,

$$(\mu_t^1(r), \ldots, \mu_t^K(r)) \xrightarrow{\mathbb{D}} (\tilde{\mu}_t^1, \ldots, \tilde{\mu}_t^K), \quad t \geq 0. \tag{2.28}$$

Consequently,

$$(\eta_t^1, \ldots, \eta_t^K) =_{st} (\tilde{\mu}_t^1, \ldots, \tilde{\mu}_t^K), \quad t \geq 0 \tag{2.29}$$

upon invoking Lemma 2.2.3.

The monotonocity of the sample paths $t \to \tilde{\mu}_t^k, 1 \leq k \leq K$, yields the convergence

$$\tilde{\mu}_t^k \uparrow \tilde{\mu}_\infty^k := \sup_{t \geq 0} \zeta_t^k, \quad 1 \leq k \leq K \tag{2.30}$$

as $t \uparrow \infty$, hence by (2.29)

$$(\eta_t^1, \ldots, \eta_t^K) \xrightarrow{\mathbb{D}} (\tilde{\mu}_\infty^1, \ldots, \tilde{\mu}_\infty^K) \tag{2.31}$$

as $t \uparrow \infty$.

It is well known [24] that $\tilde{\mu}_\infty^k < \infty$ a.s. iff $c_k > 0, 1 \leq k \leq K$, so that $(\eta^1, \ldots, \eta^K)$ has a stationary distribution iff $c_k > 0, 1 \leq k \leq K$. ∎

Let us denote this stationary distribution of $(\eta^1, \ldots, \eta^K)$ by $(\eta_\infty^1, \ldots, \eta_\infty^K)$, so that

$$(\eta_\infty^1, \ldots, \eta_\infty^K) =_{st} (\tilde{\mu}_\infty^1, \ldots, \tilde{\mu}_\infty^K). \tag{2.32}$$

It is easy to deduce from (2.21) and (2.32) that under the condition $c_k > 0, 1 \leq k \leq K$, we have

$$\kappa_t \xrightarrow{\mathbb{D}} \kappa_\infty$$

as $t \uparrow \infty$, with

$$\kappa_\infty =_{st} \max_{1 \leq k \leq K} \sup_{t \geq 0} \zeta_t^k. \tag{2.33}$$

During the course of the preceding proof we have derived an expression for $(\eta_\infty^1, \ldots, \eta_\infty^K)$ which is the limiting process obtained from $(\eta_t^1(r), \ldots, \eta_t^K(r))$ by taking the limit in distribution as $r \uparrow \infty$ and then as $t \uparrow \infty$. Denote by $(\hat{\eta}_\infty^1, \ldots, \hat{\eta}_\infty^K)$ the limiting process obtained from $(\eta_t^1(r), \ldots, \eta_t^K(r))$ by taking the limit in distribution as $t \uparrow \infty$ and then as $r \uparrow \infty$. We now show that these two limits are in fact equal.

Theorem 2.2.2 *Under the conditions $c_k > 0, 1 \leq k \leq K$ and $v^k(r) < u(r), 1 \leq k \leq K, r = 1, 2 \ldots$, the equality*

$$(\eta_\infty^1, \ldots, \eta_\infty^K) =_{st} (\hat{\eta}_\infty^1, \ldots, \hat{\eta}_\infty^K) \tag{2.34}$$

holds where

$$(\tilde{\mu}_\infty^1(r), \ldots, \tilde{\mu}_\infty^K(r)) \xrightarrow{\mathbb{D}} (\hat{\eta}_\infty^1, \ldots, \hat{\eta}_\infty^K)$$

as $r \uparrow \infty$.

Proof. Fix $r = 1, 2 \ldots$, and recall the definition (2.24) of $\tilde{\mu}^k(r), 1 \leq k \leq K$. The monotonocity of the sample paths $t \to \tilde{\mu}^k_t(r), 1 \leq k \leq K$, yields the convergence

$$\tilde{\mu}^k_t(r) \uparrow \tilde{\mu}^k_\infty(r) := \sup_{t \geq 0} \zeta^k_t(r), \quad 1 \leq k \leq K, \tag{2.35}$$

as $t \uparrow \infty$. From standard results on the $GI/GI/1$ queue [39] we conclude that $\tilde{\mu}^k_\infty(r) < \infty, 1 \leq k \leq K$, since $v^k(r) < u(r), 1 \leq k \leq K$.

Define a subspace $\tilde{D}[0, \infty)$ of $D[0, \infty)$ as

$$\tilde{D}[0, \infty) = \{z \in D[0, \infty) : \sup_{t \geq 0} z_t < \infty\}$$

and define a mapping $T : \tilde{D}[0, \infty) \to I\!R$ by

$$Tz = \sup_{t \geq 0} z_t, \quad z \in \tilde{D}[0, \infty).$$

Under the assumptions $c_k > 0, 1 \leq k \leq K$ and $v^k(r) < u(r), 1 \leq k \leq K$, it is easy to see that the stochastic processes ζ and $\zeta(r), r = 1, 2 \ldots$, belong a.s. to $\tilde{D}[0, \infty)$. Moreover observe that

$$(\tilde{\mu}^1_\infty, \ldots, \tilde{\mu}^K_\infty) = (T\zeta^1, \ldots, T\zeta^K)$$

and

$$(\tilde{\mu}^1_\infty(r), \ldots, \tilde{\mu}^K_\infty(r)) = (T\zeta^1(r), \ldots, T\zeta^K(r)).$$

It can be shown [69] that T is a continuous mapping under Skohorkhod's topology on $\tilde{D}[0, \infty)$.

Lemma 2.2.2 yields that as $r \uparrow \infty$,

$$(\zeta^1(r), \ldots, \zeta^K(r)) \Rightarrow (\zeta^1, \ldots, \zeta^K) \tag{2.36}$$

in $\tilde{D}[0, \infty)^K$, so that by the continuous mapping theorem, we conclude that

$$(T\zeta^1(r), \ldots, T\zeta^K(r)) \xrightarrow{I\!D} (T\zeta^1, \ldots, T\zeta^K) \tag{2.37}$$

as $r \uparrow \infty$. But by definition

$$(\tilde{\mu}_{\infty}^{1}(r), \ldots, \tilde{\mu}_{\infty}^{K}(r)) \xrightarrow{\mathcal{D}} (\hat{\eta}_{\infty}^{1}, \ldots, \hat{\eta}_{\infty}^{K})$$

as $r \uparrow \infty$, so that

$$(\hat{\eta}_{\infty}^{1}, \ldots, \hat{\eta}_{\infty}^{K}) =_{st} (\tilde{\mu}_{\infty}^{1}, \ldots, \tilde{\mu}_{\infty}^{K}) \tag{2.38}$$

Combining (2.32) and (2.38), we now conclude (2.34). ∎

2.3 Heavy traffic limit for the number in the join buffer

In this section we develop heavy traffic diffusion limits for the number of customers at each queue and in the join buffer of the fork–join queue. Just as functional central limit theorems for random walk processes were crucial in obtaining heavy traffic limits for the queue delay processes, we shall see that functional central limit theorems for renewal processes are crucial in obtaining heavy traffic limits for the queue length processes. As before we assume that the queue is empty at $t = 0$ and that assumption **(IIa)** is in effect. For $1 \leq k \leq K$ and $t \geq 0$, we define

N_t : The number of customers in the join buffer of the fork–join queue.

Q_t^k : The number of customers in the k^{th} queueing system.

A_t : The total number of arrivals into the system in the interval $[0, t]$. Note that

$$A_t = \begin{cases} \max\{i : u_1 + \ldots + u_i \leq t\} & \text{if } u_1 \leq t \\ 0, & \text{if } u_1 > t \end{cases}$$

where the sequence $\{u_n\}_0^\infty$ was defined in Section 2.2.1.

S_t^k : The total number of potential service completions in the interval $[0, t]$ in the k^{th} queue. Note that

$$S_t^k = \begin{cases} \max\{i : v_1^k + \ldots + v_i^k \leq t\} & \text{if } v_1^k \leq t \\ 0, & \text{if } v_1^k > t \end{cases}$$

where the sequence $\{v_n^k\}_0^\infty, 1 \leq k \leq K$, was defined in Section 2.2.1.

D_t^k : The total number of departures in the interval $[0, t]$ from the k^{th} queue.

It is easy to see that the number of customers in the k^{th} queue is given by

$$Q_t^k = A_t - D_t^k, \quad t \geq 0. \tag{3.1}$$

Note that $A \equiv \{A_t, t \geq 0\}$ is a renewal process whereas $D^k \equiv \{D_t^k, t \geq 0\}, 1 \leq k \leq K$, are not renewal process since there are no departures from an empty

queue. Consequently it is difficult to prove functional central limit theorems for the $\mathbb{N}$–valued process $Q^k \equiv \{Q^k(t), t \geq 0\}, 1 \leq k \leq K$. To remedy this situation, a modified system is introduced in which the servers are not shut off when they become idle. This idea first originated with Borovkov [10], and was extensively used by Iglehart and Whitt [29],[30] from where the following description of the modified system is borrowed: We associate with each server a sequence of potential service times. If a server faces a continued demand for service, then its actual service times are just these potential service times; but if there is no demand during any potential service time, then the potential service time is ignored and there is no actual service and no departure. After a server has begun working in the absence of demand, then the next demand will occur in the middle of some potential service time. Let the remaining portion of that potential service time be that next customer's actual service time. Heavy traffic limit theorems are much easier to prove for the modified system, and the desired result for the original system is obtained by showing that the differences between the two systems is negligible in heavy traffic. In keeping with this program, for $1 \leq k \leq K$ and $t \geq 0$, we set

q_t^k : The number of customers at time t in the k^{th} modified queueing system.

With the definition

$$X_t^k = A_t - S_t^k, \quad 1 \leq k \leq K, \quad t \geq 0 \tag{3.2}$$

we easily see [29] that the relation

$$q^k = g(X^k), \quad 1 \leq k \leq K, \quad t \geq 0 \tag{3.3}$$

holds where g is the reflection mapping (2.15), and $q^k \equiv \{q_t^k, t \geq 0\}$ and $X^k \equiv \{X_t^k, t \geq 0\}, 1 \leq k \leq K$.

We now proceed to obtain the functional central limit theorems. Considering a sequence of modified queueing systems indexed by $r = 1, 2 \ldots$, we set

$$\lambda(r) = \frac{1}{\mathbb{E}u_1(r)} = \frac{1}{u(r)}, \quad \tau_0(r) = (\lambda^3(r)\sigma_0^2(r))^{\frac{1}{2}} \tag{3.4}$$

and for $1 \leq k \leq K$,

$$\mu^k(r) = \frac{1}{I\!E v_1^k(r)} = \frac{1}{v^k(r)}, \quad \tau_k(r) = ((\mu^k(r))^3 \sigma_k^2(r))^{\frac{1}{2}} \tag{3.5}$$

so that $\lambda(r)$ and $\mu^k(r), 1 \leq k \leq K$, are the arrival and service rates respectively in the r^{th} system. We assume that

(IId): As $r \uparrow \infty$,

$$\begin{aligned} \lim \tau_0(r) &\to \tau_0, \\ \lim \tau_k(r) &\to \tau_k, \quad 1 \leq k \leq K \\ [\lambda(r) - \mu^k(r)]\sqrt{r} &\to d_k, \quad 1 \leq k \leq K. \end{aligned}$$

For $r = 1, 2 \ldots$, define the stochastic processes $A_r \equiv \{A_r(t), t \geq 0\}$ and $S_r^k \equiv \{S_r^k(t), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$A_t(r) = \frac{A_{[rt]}(r) - \lambda(r)[rt]}{\sqrt{r}}, \quad t \geq 0 \tag{3.6}$$

and

$$S_t^k(r) = \frac{S_{[rt]}^k(r) - \mu^k(r)[rt]}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{3.7}$$

Recall that $\xi^k \equiv \{\xi_t^k, t \geq 0\}, 0 \leq k \leq K$, are $K+1$ independent Wiener processes. Lemma 2.3.1 shows that the stochastic processes (3.6)–(3.7) converge to these Wiener processes.

Lemma 2.3.1. *As $r \uparrow \infty$,*

$$(A(r), S^1(r), \ldots, S^K(r)) \Rightarrow (\tau_0 \xi^0, \tau_1 \xi^1, \ldots, \tau_K \xi^K) \tag{3.8}$$

in $D[0, \infty)^{K+1}$.

Proof. Equations (3.8) is an immediate consequence of the functional central limit theorem for renewal processes (Appendix A, Theorem A4) under assumptions **(IIa), (IIc)** and **(IId)**. ∎

For $r = 1, 2 \ldots$, define the stochastic processes $X^k(r) \equiv \{X_t^k(r), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$X_t^k(r) = \frac{X_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{3.9}$$

Moreover define the processes $\phi^k \equiv \{\phi_t^k, t \geq 0\}, 1 \leq k \leq K$, by

$$\phi_t^k = \tau_0 \xi_t^0 - \tau_k \xi_t^k + d_k t, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{3.10}$$

The process $(\phi^1, \ldots, \phi^K)$ is a K–dimensional diffusion process with drift d and covariance matrix Q given by

$$d = (d_1, \ldots, d_K) \tag{3.11}$$

and

$$Q = \begin{pmatrix} \tau_1^2 + \tau_0^2 & \tau_0^2 & \ldots & \tau_0^2 \\ \tau_0^2 & \tau_2^2 + \tau_0^2 & \ldots & \tau_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0^2 & \tau_0^2 & \ldots & \tau_K^2 + \tau_0^2 \end{pmatrix}. \tag{3.12}$$

Lemma 2.3.2 shows that the vector stochastic processes induced by (3.9) weakly converge to $(\phi^1, \ldots, \phi^K)$. The cross-correlation terms in the above matrix reflect the correlation between the K queues due to the common arrival process.

Lemma 2.3.2. *As* $r \uparrow \infty$,

$$(X^1(r), \ldots, X^K(r)) \Rightarrow (\phi^1, \ldots \phi^K) \tag{3.13}$$

in $D[0, \infty)^K$.

Proof. Fix $r = 1, 2 \ldots$ and $t \geq 0$. For all $1 \leq k \leq K$, we see from (3.2) that

$$\begin{aligned} X_t^k(r) &= \frac{X_{[rt]}^k(r)}{\sqrt{r}} \\ &= \frac{A_{[rt]}(r) - S_{[rt]}^k(r)}{\sqrt{r}} \\ &= \frac{A_{[rt]}(r) - \lambda(r)[rt]}{\sqrt{r}} - \frac{S_{[rt]}^k(r) - \mu^k(r)[rt]}{\sqrt{r}} - \frac{[rt][\lambda(r) - \mu^k(r)]}{\sqrt{r}} \\ &= A_t(r) - S_t^k(r) - \frac{[rt]}{r}[\lambda(r) - \mu^k(r)]\sqrt{r}. \end{aligned}$$

From assumption **(IId)** it is clear that as $r \uparrow \infty$,

$$\frac{[rt]}{r}[\lambda(r) - \mu^k(r)]\sqrt{r} \to d_k t, \quad t \geq 0$$

and we conclude to (3.13) from Lemma 2.3.1 and the continuous mapping theorem.

∎

For $r = 1, 2 \ldots$, we define the stochastic processes $q^k(r) \equiv \{q_t^k(r), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$q_t^k(r) = \frac{q_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{3.14}$$

Lemma 2.3.3. *As $r \uparrow \infty$,*

$$(q^1(r), \ldots, q^K(r)) \Rightarrow (g(\phi^1), \ldots, g(\phi^K)) \tag{3.15}$$

in $D[0, \infty)^K$.

Proof. From (3.3) we see that for all $1 \leq k \leq K$,

$$q_t^k(r) = g(X^k(r))_t, \quad t \geq 0.$$

Now using Lemma 2.3.2 and the continuous mapping theorem we obtain (3.15).

∎

For $r = 1, 2 \ldots$, we define the stochastic processes $Q^k(r) \equiv \{Q_t^k(r), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$Q_t^k(r) = \frac{Q_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{3.16}$$

In Lemma 2.3.4 we show that the vector process associated with (3.16) and generated by the queue length processes in the original system, converges to the same limit as the vector process (3.14) generated by the queue length process of the modified system.

Lemma 2.3.4. *As* $r \uparrow \infty$,

$$(Q^1(r), \ldots, Q^K(r)) \Rightarrow (g(\phi^1), \ldots, g(\phi^K)) \tag{3.17}$$

in $D[0,\infty)^K$.

Proof. Fix $T > 0$. The result (3.17) follows from the fact that as $r \uparrow \infty$,

$$\sup_{0 \le t \le T} \max_{1 \le k \le K} \mid Q_t^k(r) - q_t^k(r) \mid \xrightarrow{I\!P} 0, \quad 1 \le k \le K \tag{3.18}$$

and a simple application of the converging together theorem. Proofs of (3.18) were given by Borovkov [10] and by Iglehart and Whitt [29]. ∎

For $r = 1, 2 \ldots$, we define the stochastic process $N(r) \equiv \{N_t(r), t \ge 0\}$ with sample paths in $D[0,\infty)$ by

$$N_t(r) = \frac{N_{[rt]}(r)}{\sqrt{r}}, \quad t \ge 0. \tag{3.19}$$

We conclude this section by obtaining a functional central limit theorem for the number of customers in the join buffer.

Theorem 2.3.1. *As* $r \uparrow \infty$,

$$N(r) \Rightarrow \sum_{k=1}^{K} [\max\{g(\phi^1), \ldots, g(\phi^K)\} - g(\phi^k)] \tag{3.20}$$

in $D[0,\infty)$.

Proof. For each $r = 1, 2 \ldots$, the identity

$$N_r(t) = \sum_{k=1}^{K} [\max\{Q_t^1(r), \ldots, Q_t^K(r)\} - Q_t^k(r)], \quad t \ge 0 \tag{3.21}$$

holds true, so that (3.20) follows from Lemma 2.3.4 and the continuous mapping theorem. ∎

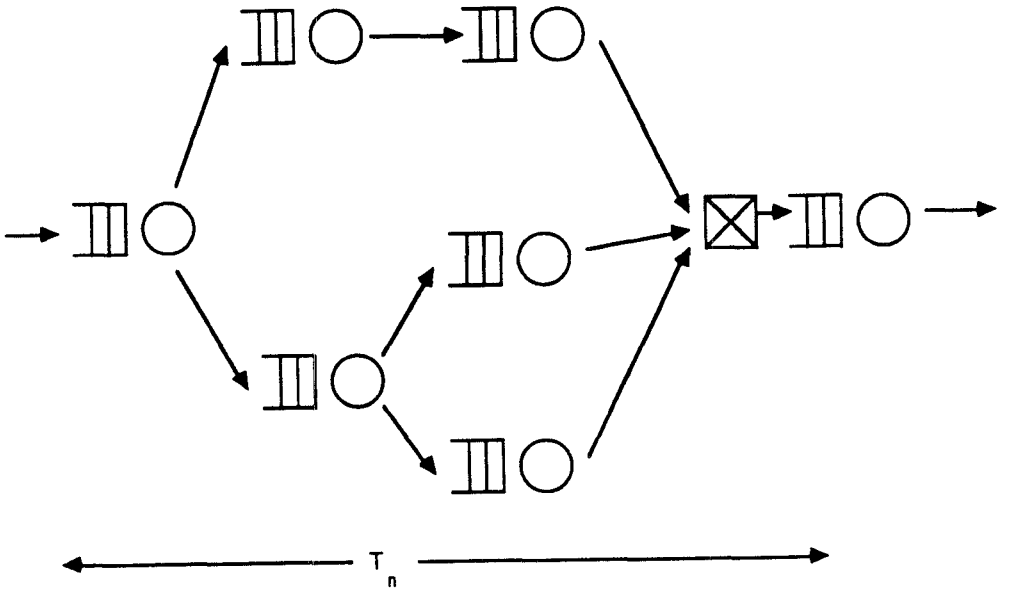


Fig. 2.2. An example of an acyclic fork–join network

2.4 Heavy traffic limit for acyclic fork–join networks

2.4.1 The model

Baccelli, Massey and Towsley [5] extended the notion of a single stage fork–join queue to acyclic fork–join networks. The single stage fork–join queue analyzed in Section 2.2 is a special case of these networks. To see the motivation for introducing these more general networks, the reader may consult the survey paper [6]. We now introduce the notation and definitions associated with this network, most of which is borrowed from [6].

The acyclic fork–join network under consideration is represented by an acyclic graph $G = (V, E)$ where V is a set of B FIFO queues labeled $i = 1, \ldots, B$ and E is a set of links such that (i, j) in E implies $j > i$. Also add for the sake of convenience fictitious queues 0 and $B + 1$, which act respectively as source and sink for the network.

For $1 \leq i \leq B$, we define the set of immediate predecessors $p(i)$ of queue i as the set of queues that have direct link to queue i, i.e.,

$$p(i) = \{j \in (1, \ldots, B) \mid (j, i) \in E\} \tag{4.1}$$

and the set of immediate successors $s(i)$ of queue i, as the set of queues to which i has a direct link, i.e.,

$$s(i) = \{j \in (1, \ldots, B) \mid (i, j) \in E\}. \tag{4.2}$$

We also denote as $s(0)$, the set of queues with no incoming links and as $p(B+1)$, the set of queues with no outgoing links. It will be assumed that the numbering of the queues is such that

$$s(0) = \{1, \ldots, B'\}, \quad B' \leq B$$

and

$$p(B+1) = \{B'', \ldots, B\}, \quad B'' \leq B.$$

We now describe the operation of the network. We assume that customers are being created at the source which acts as the outside world for the network. These exogenous customers enter the network through the queues in $s(0)$ and traverse it upon following certain synchronization rules $(SR_1) - (SR_3)$ described below. Finally customers leave the network from the queues in $p(B+1)$ by being absorbed into the network sink and disappearing. We now specify the synchronization rules that govern the network.

(SR_1): The exogenous customers created at the source are routed instantaneously to the queues in $s(0)$ under the constraint of a Fork primitive, i.e., the n^{th} arrival date to each one of the queues in $s(0)$ coincides with the n^{th} date of customer creation. An alternate way of viewing this constraint is to assume that upon its creation, a customer creates B' replicas of itself which are then dispatched at the same time and instantaneously to the queues in $s(0)$, one replica per queue.

(SR_2): A service completion in some queue i in $s(0)$ will not systematically trigger an arrival to a queue in $s(i)$. In fact, more generally, the arrivals to queue j, with $B' < j \leq B$, are generated as follows: Assume the sequence of service completions to be known for all queues i, with $1 \leq i \leq j$ and $B' < j \leq B$. The n^{th} arrival date to queue j coincides with the latest date among all the n^{th} service completions at the queues in $p(j)$. Due to the acyclic structure of (V, E), this mechanism will successively define the arrival patterns to queues $B'+1, B'+2, \ldots, B$.

(SR_3): Customers leave the network through the queues in $p(B+1)$ in the form of a single output stream by imposing the following synchronization of the join type: The n^{th} network departure is defined as the latest date among the dates of n^{th} service completions in the queues $B'', B''+1, \ldots, B$.

2.4.2 Recursive representation of the delays

In this section a recursive representation for the delays in the network is provided. The material of this section is borrowed from [6].

Given an acyclic graph $G = (V, E)$, the performance measures associated with the corresponding network are fully specified by $B+1$ sequences of $\mathbb{R}_+$-valued RVs with the interpretation that for all $n = 0, 1 \ldots$, and $1 \leq j \leq B$,

τ_n : Arrival epoch of the n^{th} customer into the network.

v_n^j : Service time requirement of the n^{th} customer to be served in queue j.

We assume the system to be initially empty and adopt the convention that the 0^{th} exogenous customer is created at time $t = 0$, so that $\tau_0 = 0$. In terms of these RVs we define the following quantities for all $n = 0, 1 \ldots$ and all $1 \leq j \leq B$,

u_n : Inter-arrival time between the $(n+1)^{rst}$ and n^{th} exogenous customers $(= \tau_{n+1} - \tau_n)$.

D_n^j : Delay between the arrival of the n^{th} exogenous customer in the network and the beginning of the n^{th} service in queue j.

W_n^j : Waiting time of the n^{th} exogenous customer in the buffer of queue j.

T_n : End-to-end delay or network response time of the n^{th} exogenous customer.

The following recursion between these variables was established by Baccelli, Massey and Towsley [5].

Lemma 2.4.1. *Consider the acyclic fork–join network defined above. If the system is initially empty, then for* $1 \leq j \leq B$, *the recursions*

$$D_0^j = \max_{i \in p(j)} \{D_0^i + v_0^i\}$$

$$D_{n+1}^j = \max\{\max_{i \in p(j)} \{D_{n+1}^i + v_{n+1}^i\}, D_n^j + v_n^j - u_{n+1}\}, \quad n = 0, 1 \ldots (4.3)$$

and

$$W_0^j = 0$$

$$W_{n+1}^j = \max\{0, W_n^j + \max_{i \in p(j)} \{D_n^i + v_n^i\} - \max_{i \in p(j)} \{D_{n+1}^i + v_{n+1}^i\} + v_n^j - u_{n+1}\},$$
$$n = 0, 1 \ldots (4.4)$$

hold where the maximum over an empty set is zero by convention. Moreover the network response time of the n^{th} *customer is given by*

$$T_n = \max_{i \in p(B+1)} \{D_n^i + v_n^i\}. \qquad n = 0, 1 \ldots (4.5)$$

Proof. Since the system is initially empty, the boundary conditions (4.3)–(4.4) are immediate from the synchronization rules (SR_1)–(SR_2). Customers arriving to queue j in $s(0)$ do so according to the pattern of exogenous arrivals, so that D_n^j corresponds to the n^{th} waiting time in a FIFO queue generated by the sequences of interarrivals $\{u_{n+1}\}_0^\infty$ and service requirements $\{v_n^j\}_0^\infty, 1 \leq j \leq B'$. Writing the corresponding Lindley equation, we get

$$D_{n+1}^j = \max\{0, D_n^j + v_n^j - u_{n+1}\}, \quad 1 \leq j \leq B' \qquad n = 0, 1 \ldots (4.6)$$

and this reduces to (4.3), since $p(j) = \emptyset$ for j in $s(0)$.

For $B' < j \leq B$, we fix $n = 0, 1 \ldots$. The $(n+1)^{rst}$ service completion at queue i in $p(j)$ takes place at time $\tau_{n+1} + D^i_{n+1} + v^i_{n+1}$, so that by applying the synchronization rule (SR_2), we see that the $(n+1)^{rst}$ arrival to queue j takes place at time $\tau_{n+1} + \max_{i \in p(j)} \{D^i_{n+1} + v^i_{n+1}\}$. Since the server at queue j becomes available for service at time $\tau_n + D^j_n + v^j_n$, we readily obtain (4.3).

In order to derive (4.4) we just have to note the relations

$$W^j_n = D^j_n - \max_{i \in p(j)} \{D^i_n + v^i_n\}, \quad 1 \leq j \leq B. \qquad n = 0, 1 \ldots$$

∎

We now state a result regarding the stability of these networks. First we make the assumption **(IIe)** where

(IIe): The sequences $\{u_{n+1}\}_0^\infty$ and $\{v^j_n\}_0^\infty, j = 1, \ldots, B$, are iid with finite second moments and mutually independent.

For $n = 0, 1 \ldots$, we set

$$u = I\!E(u_{n+1}) < \infty, \quad \sigma_0^2 = Var(u_{n+1}) < \infty$$

and

$$v^j = I\!E(v^j_n) < \infty, \quad \sigma_j^2 = Var(v^j_n) < \infty, \quad 1 \leq j \leq B$$

Again, as for the simple fork–join queue, we consider the system to be stable if the vector of delays $\{(D^1_n, \ldots, D^B_n)\}_0^\infty$ converges jointly in distribution as $n \uparrow \infty$ to a proper random vector $(D^1, \ldots, D^B)$. The stability conditions for this system were given in [5], and are reproduced below.

Lemma 2.4.2. *Assume that condition* **(IIe)** *holds. The system is stable iff*

$$v^j < u, \quad 1 \leq j \leq B. \tag{4.7}$$

2.4.3 The diffusion limit

In the last section we saw that the acyclic fork–join network will be stable provided $v^j < u, 1 \leq j \leq B$. The system is said to be in heavy traffic if $v^j \approx u$ for at least one of the queues. In this section our objective is to develop heavy traffic diffusion limits for the delay processes in these networks. The methodology that we employ is the same as the one used in Section 2.3.3. In short we shall use the recursions (4.3)–(4.4) to connect the delay processes to partial sums of iid RVs and then use well-known functional central limit theorems for these partial sums in order to deduce the corresponding limit theorems for the delay processes by means of the continuous mapping theorem. However, since the recursions in this case are much more involved than those in Section 2.3.3, the limiting process is correspondingly more complex.

We now consider a sequence of these networks indexed by $r = 1, 2 \ldots$, each of which satisfies condition **(IIe)**. Moreover assume that

(IIf): As $r \uparrow \infty$,

$$\sigma_j(r) \to \sigma_j, \quad 0 \leq j \leq B$$

$$[u(r) - v^j(r)]\sqrt{r} \to c_j. \quad 1 \leq j \leq B$$

(IIg): For some $\epsilon > 0$,

$$\sup_{r,j} \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{\mid v_1^j(r) \mid^{2+\epsilon}\}\} < \infty.$$

For $1 \leq j \leq B$ and $r = 1, 2 \ldots$, define the partial sums

$$V_0^j(r) = 0,$$

$$V_n^j(r) = v_0^j(r) + \ldots + v_{n-1}^j(r), \qquad n = 1, 2 \ldots \quad (4.8a)$$

and

$$U_0(r) = 0,$$

$$U_n(r) = u_1(r) + \ldots + u_n(r). \qquad n = 1, 2 \ldots \quad (4.8b)$$

For $r = 1, 2 \ldots$, define the stochastic processes $\xi^j(r) \equiv \{\xi_t^j(r), t \geq 0\}, 0 \leq j \leq B$, with sample paths in $D[0, \infty)$ by

$$\xi_t^0(r) = \frac{U_{[rt]}(r) - u(r)[rt]}{\sqrt{r}}, \quad t \geq 0 \qquad (4.9a)$$

and

$$\xi_t^j(r) = \frac{V_{[rt]}^j(r) - v^j(r)[rt]}{\sqrt{r}}, \quad 1 \le j \le B, \quad t \ge 0. \tag{4.9b}$$

Let $\xi^j \equiv \{\xi_t^j, t \ge 0\}, 0 \le j \le B$, be $B+1$ independent Wiener processes. Lemma 2.4.3 shows that the stochastic processes defined in (4.9) converge weakly to these Wiener processes.

Lemma 2.4.3 *As* $r \uparrow \infty$,

$$(\xi^0(r), \xi^1(r), \ldots, \xi^B(r)) \Rightarrow (\sigma_0\xi^0, \sigma_1\xi^1, \ldots, \sigma_B\xi^B) \tag{4.10}$$

in $D[0,\infty)^{B+1}$.

Proof. The proof is exactly the same as for Lemma 2.2.1., with assumptions **(IIa)–(IIc)** now replaced by assumptions **(IIe)–(IIg)**. ∎

For $r = 1, 2 \ldots$, we set

$$\begin{aligned} S_0^j(r) &= 0 \\ S_n^j(r) &= V_n^j(r) - U_n(r), \qquad n = 1, 2 \ldots \end{aligned} \tag{4.11}$$

and define the stochastic processes $\zeta^j(r) \equiv \{\zeta_t^j(r), t \ge 0\}, 1 \le j \le B$, with sample paths in $D[0,\infty)$ by

$$\zeta_t^j(r) = \frac{S_{[rt]}^j(r)}{\sqrt{r}}, \quad 1 \le j \le B, \quad t \ge 0. \tag{4.12}$$

We also define the stochastic processes $\zeta^j \equiv \{\zeta_t^j, t \ge 0\}, 1 \le j \le B$, by

$$\zeta_t^j = \sigma_j\xi_t^j - \sigma_0\xi_t^0 - c_j t, \quad 1 \le j \le B, \quad t \ge 0. \tag{4.13}$$

The next result shows that the stochastic process $(\zeta^1(r), \ldots, \zeta^B(r))$ converges weakly to $(\zeta^1, \ldots, \zeta^B)$. As noted in the discussion preceding Lemma 2.2.2, the stochastic process $(\zeta^1, \ldots \zeta^B)$ is a K–dimensional diffusion process with drift given by (2.16) and covariance given by (2.17).

Lemma 2.4.4 *As* $r \uparrow \infty$,

$$(\zeta^1(r), \ldots, \zeta^B(r)) \Rightarrow (\zeta^1, \ldots, \zeta^B) \tag{4.14}$$

in $D[0,\infty)^B$.

Proof. The proof is exactly the same as for Lemma 2.2.2. ∎

For $r = 1, 2 \ldots$, we define the stochastic processes $\eta^j(r) \equiv \{\eta_t^j(r), t \geq 0\}$ and $\mu^j(r) \equiv \{\mu_t^j(r), t \geq 0\}, 1 \leq j \leq B$, with sample paths in $D[0,\infty)$, by setting

$$\eta_t^j(r) = \frac{D_{[rt]}^j(r)}{\sqrt{r}}, \quad 1 \leq j \leq B, \quad t \geq 0 \tag{4.15}$$

and

$$\mu_t^j(r) = \frac{W_{[rt]}^j(r)}{\sqrt{r}}, \quad 1 \leq j \leq B, \quad t \geq 0. \tag{4.16}$$

The processes $\eta^j \equiv \{\mu_t^j, t \geq 0\}, 1 \leq j \leq B$, and $\mu^j \equiv \{\eta_t^j, t \geq 0\}, 1 \leq j \leq B$, are now defined by

$$\eta^j = g(\zeta^j - \max_{i \in p(j)} \eta^i) + \max_{i \in p(j)} \eta^i, \quad 1 \leq j \leq B \tag{4.17}$$

and

$$\mu^j = g(\zeta^j - \max_{i \in p(j)} \eta^i), \quad 1 \leq j \leq B. \tag{4.18}$$

In contrast with the situation for single stage fork–join queues, we note that the limiting processes (4.17)–(4.18) for acyclic fork–join networks are much more complicated.

Theorem 2.4.1. *As* $r \uparrow \infty$,

$$(\eta^1(r), \ldots, \eta^B(r)) \Rightarrow (\eta^1, \ldots, \eta^B) \tag{4.19}$$

in $D[0,\infty)^B$.

Before providing a proof for Theorem 2.4.1, we present the following two corollaries which identify the diffusion limit for the waiting times and the end-to-end delay of the system respectively.

Corollary 2.4.1. *As* $r \uparrow \infty$,

$$(\mu^1(r), \ldots, \mu^B(r)) \Rightarrow (\mu^1, \ldots, \mu^B) \tag{4.20}$$

in $D[0,\infty)^B$.

Proof. Note that for all $r = 1, 2 \ldots$,

$$W_n^j(r) = D_n^j(r) - \max_{i \in p(j)} \{D_n^i(r) + v_n^i(r)\}, \quad 1 \leq j \leq B \qquad\qquad n = 0, 1 \ldots$$

so that for all $r = 1, 2 \ldots$,

$$\mu_t^j(r) = \eta_t^j(r) - \max_{i \in p(j)} \{\eta_t^i(r) + \frac{v_{[rt]}^i(r)}{\sqrt{r}}\}, \quad 1 \leq j \leq B, \quad t \geq 0 \tag{4.21}$$

We obtain (4.20) from (4.19) and (4.21) by applying the continous mapping theorem and the converging together theorem. ∎

For $r = 1, 2 \ldots$, we introduce the stochastic processes $\kappa(r) \equiv \{\kappa_t(r), t \geq 0\}$ with sample paths in $D[0,\infty)$ by

$$\kappa_t(r) = \frac{T_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0. \tag{4.22}$$

Corollary 2.4.2. *As* $r \uparrow \infty$,

$$\kappa(r) \Rightarrow \max_{i \in p(B+1)} \eta^i \tag{4.23}$$

in $D[0,\infty)$.

Proof. Using the fact that for all $r = 1, 2 \ldots$

$$\kappa_t(r) = \max_{i \in p(B+1)} \{\eta_t^i(r) + \frac{v_{[rt]}^i}{\sqrt{r}}\}, \quad t \geq 0 \tag{4.24}$$

we obtain (4.23) from (4.19) and (4.24) by applying the continuous mapping theorem and the converging together theorem. ∎

We now proceed with the proof for Theorem 2.4.1. For $1 \leq i \leq B$, we define the level $l(i)$ of queue i by

$$l(i) = \max_{j \in p(i)} l(j) + 1 \tag{4.15}$$

where by definition $l(i) = 1$ if $p(i) = \emptyset$. The level N of the graph is defined by $N := \max_{i \in V} l(i)$.

We denote the set of queues on level l as $q(l), 1 \leq l \leq N$ and assume that the cardinality of $q(l)$ is B_l. The queues are numbered in such a way that

$$\begin{aligned} q(1) &= \{1, \ldots, B_1\} \\ q(2) &= \{B_1 + 1, \ldots, B_1 + B_2\}, \\ &\vdots \\ q(N) &= \{B_1 + \ldots + B_{N-1} + 1, \ldots, B\} \end{aligned} \tag{4.26}$$

Note that the sets $q(1)$ and $q(N)$ have already been defined earlier as $s(0)$ and $p(B+1)$ respectively, with $B' = B_1$ and $B'' = B_1 + \ldots + B_{N-1} + 1$

Proof. Our proof proceeds by induction on the levels of the acyclic graph which underlies the queueing network. First consider the queues belonging to the set $q(1)$, i.e., queues j such that $l(j) = 1$. Recall that for these queues $p(j) = \emptyset$, so that for $r = 1, 2 \ldots$ we have that

$$\begin{aligned} D^j_{n+1}(r) &= \max\{0, D^j_n(r) + v^j_n(r) - u_{n+1}(r)\} \\ &= S^j_{n+1}(r) - \min_{0 \leq k \leq n+1} S^j_k(r) \end{aligned} \qquad n = 0, 1 \ldots$$

so that

$$\eta^j_t(r) = g(\zeta^j(r))_t, \quad t \geq 0, \quad j = 1, \ldots, B_1, \tag{4.27}$$

upon taking note of (4.13) and (4.15). From (4.27) and (4.14) it follows that

$$(\eta^1(r), \ldots, \eta^{B_1}(r), \zeta^1(r), \ldots, \zeta^B(r)) \Rightarrow (\eta^1, \ldots, \eta^{B_1}, \zeta^1, \ldots, \zeta^B) \tag{4.28}$$

as $r \uparrow \infty$, so that (4.21) is verified for the queues belonging to the set $q(1)$.

As the induction hypothesis, assume that

$$(\eta^1(r),\ldots,\eta^{B_1+\ldots+B_l}(r),\zeta^1(r),\ldots,\zeta^B(r)) \Rightarrow (\eta^1,\ldots,\eta^{B_1+\ldots+B_l},\zeta^1,\ldots,\zeta^B) \tag{4.29}$$

as $r \uparrow \infty$, which implies that (4.21) holds for the queues belonging to the first l levels. Using (4.29) we shall prove that (4.21) holds for queues belonging to the first $l+1$ levels, thus completing the induction step.

Consider queue j such that $l(j) = l+1$. Expanding the recursion in Lemma 2.4.1 for $r = 1,2\ldots$, $n = 0,1\ldots$ and $j = B_l+1,\ldots,B_{l+1}$, we obtain

$$\begin{aligned} D^j_{n+1}(r) &= \max_{i\in p(j)}\{D^i_{n+1}(r)+v^i_{n+1}(r)\} \\ &\quad + \max\{0, D^j_n(r) - \max_{i\in p(j)}\{D^i_{n+1}(r)+v^i_{n+1}(r)\} + v^j_n(r) - u_{n+1}(r)\}, \\ &= \max_{i\in p(j)}\{D^i_{n+1}(r)+v^i_{n+1}(r)\} + S^j_{n+1}(r) - \max_{i\in p(j)}\{D^i_{n+1}(r)+v^i_{n+1}(r)\} \\ &\quad - \min_{0\le k\le n+1}\{S^j_k(r) - \max_{i\in p(j)}\{D^i_k(r)+v^i_k(r)\}\} \end{aligned} \tag{4.30}$$

Note that by (4.30), we have for $j = B_l+1,\ldots,B_{l+1}$ and $t \ge 0$,

$$\begin{aligned} &\eta^j_t(r) \\ &= \max_{i\in p(j)}\{\eta^i_t(r) + \frac{v^i_{[rt]}}{\sqrt{r}}\} + g\left(\zeta^j(r) - \max_{i\in p(j)}\{\eta^i(r) + \frac{v^i_{[r\cdot]}}{\sqrt{r}}\}\right)_t \end{aligned} \tag{4.31}$$

From (4.29), (4.31), the continuous mapping theorem and the converging together theorem, we conclude that as $r \uparrow \infty$,

$$(\eta^1(r),\ldots,\eta^{B_1+\ldots+B_{l+1}}(r),\zeta^1(r),\ldots,\zeta^B(r)) \Rightarrow (\eta^1,\ldots,\eta^{B_1+\ldots+B_{l+1}},\zeta^1,\ldots,\zeta^B) \tag{4.32}$$

as $r \uparrow \infty$, which completes the induction step.

∎

CHAPTER III

3.1 Introduction

In the last chapter we obtained heavy traffic diffusion limits for single stage fork–join queues. Our ultimate goal is to obtain heavy traffic approximations for the end–to–end delay of this system. Hence according to the methodology sketched in Chapter 1, our next step should be to obtain the stationary distribution of the limiting diffusion for the end–to–end delay. However, as we now illustrate evaluating this distribution is not as simple as was the case for the single server queue in Chapter 1. Recall from Section 2.2.3 that the stochastic process $\kappa(r)$ generated by the end–to–end delay sequence, converges as $r \uparrow \infty$ over the interval $[0, \infty)$, to a stochastic process κ given by

$$\kappa = \max_{1 \leq k \leq K} (g(\zeta^1), \ldots, g(\zeta^K))$$

Here, g is the reflection mapping, and the processes $\zeta^k \equiv \{\zeta_t^k, t \geq 0\}, 1 \leq k \leq K$, are given by

$$\zeta_t^k = \sigma_k \xi_t^k - \sigma_0 \xi_t^0 - c_k t, \quad 1 \leq k \leq K, \quad t \geq 0.$$

where the processes $\xi^0, \ldots, \xi^K$ are $K+1$ independent standard Wiener processes over the interval $[0, \infty)$.

For all $t \geq 0$ and $1 \leq k \leq K$, it is known [24] that the marginal distribution of each RV $\eta_t^k = g(\zeta^k)_t$ is given by

$$I\!P(\eta_t^k \leq x) = \Phi\left(\frac{x + c_k t}{\sqrt{\sigma_k^2 + \sigma_0^2 t}}\right) - e^{-\frac{2c_k x}{\sigma_k^2 + \sigma_0^2}} \Phi\left(\frac{-x + c_k t}{\sqrt{\sigma_k^2 + \sigma_0^2 t}}\right), \quad x \geq 0.$$

However we do not know the joint distribution of the vector $(\eta_t^1, \ldots, \eta_t^K)$ due to the correlation that exists between the different components. Hence, since the

distribution of κ_t depends upon this joint distribution, we are unable to evaluate it directly.

The traditional method of overcoming this difficulty is by deriving a partial differential equation (with appropriate boundary conditions) that the joint distribution satisfies. We explore this option in the next chapter. In the present chapter we obtain diffusions that bound the limiting diffusion for the end–to–end delay from above and from below in the sense of stochastic ordering. The significant fact is that the stationary distributions for the bounding diffusions can be easily obtained and they serve to bound the stationary distribution of the original diffusion. The basic methodology for carrying out this plan was first presented by Baccelli and Makowski and Shwartz [4] .

This chapter is organized as follows: In Section 3.2 we give the definitions and some basic properties of the stochastic orderings that we shall use. In Section 3.3 a diffusion that bounds the diffusion for the end–to–end delay from below is obtained by using convex increasing stochastic ordering. In Sections 3.4 an upper bound is obtained by using the idea of associated RVs. Lastly in Section 3.5 we explicitly compute the stationary distributions of the bounding diffusions for both the transient as well the steady-state cases.

3.2 Some preliminaries

We first give a definition of convex–increasing and strong stochastic orderings for continuous time stochastic processes.

Definition 3.2.1. *Let X and Y be two real–valued RVs. The RV X is said to be smaller than the RV Y in the sense of strong stochastic ordering if*

$$\mathbb{E}f(X) \leq \mathbb{E}f(Y)$$

for all non–decreasing functions $f : \mathbb{R} \to \mathbb{R}$. This is denoted as $X \leq_{st} Y$.

The RV X is smaller than the RV Y in the sense of convex increasing stochastic ordering if

$$\mathbb{E}f(X) \leq \mathbb{E}f(Y)$$

for all convex non–decreasing functions $f : \mathbb{R} \to \mathbb{R}$. This is denoted as $X \leq_{icx} Y$.

Let the symbol $\prec$ denote one of the stochastic orderings $\leq_{st}$ or $\leq_{icx}$. Let $X \equiv \{X_t, t \geq 0\}$ and $Y \equiv \{Y_t, t \geq 0\}$ be two real–valued stochastic processes. The process X is smaller than Y with respect to $\prec$, denoted as $X \prec Y$, if

$$X_t \prec Y_t, \quad t \geq 0.$$

We now introduce the concept of associated stochastic processes.

Definition 3.2.2. *The real-valued RVs $\{X^1, \ldots, X^K\}$, are associated if and only if, the inequality*

$$\mathbb{E}[f(X)h(X)] \geq \mathbb{E}[f(X)]\mathbb{E}[h(X)]$$

holds for all pairs of monotone non–decreasing mappings $f, h : \mathbb{R}^K \to \mathbb{R}$ for which these expectations exist.

The real–valued stochastic processes $X^k \equiv \{X_t^k, t \geq 0\}, 1 \leq k \leq K$, are associated if and only if, for all $t \geq 0$, the RVs $\{X_t^1, \ldots, X_t^K\}$ are associated.

Definition 3.2.3 *The stochastic processes* $\overline{X} \equiv \{\overline{X}_t^k, t \geq 0\}, 1 \leq k \leq K$, *are said to form independent versions of the stochastic processes* $X \equiv \{X_t^k, t \geq 0\}, 1 \leq k \leq K$, *if*

(i) : *For all* $t \geq 0$, *the RVs* $\{\overline{X}_t^1, \ldots, \overline{X}_t^K\}$ *are mutually independent, and*

(ii) : *For every* $1 \leq k \leq K$ *and* $t \geq 0$, *the RVs* X_t^k *and* $\overline{X}_t^k$ *have the same probability distribution.*

The following result [7] is an immediate consequence of this definition.

Lemma 3.2.1. *If the stochastic processes* $X \equiv \{X_t^k, t \geq 0\}, 1 \leq k \leq K$, *are associated, then the inequality*

$$\max_{1 \leq k \leq K} X_t^k \leq_{st} \max_{1 \leq k \leq K} \overline{X}_t^k \quad t \geq 0$$

holds true.

The following lemma [7] is very useful.

Lemma 3.2.2.

(i) : Independent RVs are always associated.

(ii) : The union of independent collections of associated RVs forms a set of associated RVs.

(iii) : Any subset of a family of associated RVs forms a set of associated RVs.

(iv) : A monotone non-decreasing function of associated RVs generates a set of associated RVs.

3.3 A lower bound

It is a well–known fact that for certain queueing systems operating in their stable regime, determinism in either the arrival or the service processes minimizes queueing delays. Our results in this section imply that this property continues to hold for the limiting diffusion of the end–to–end delay of the fork–join queue in heavy traffic. We prove this result by working directly with the limiting diffusion.

Recall that for each $1 \leq k \leq K$, $t \geq 0$, we have

$$\zeta_t^k = \sigma_k \xi_t^k - \sigma_0 \xi_t^0 - c_k t, \quad \eta_t^k = g(\zeta^k)_t \tag{3.1a}$$

and

$$\kappa_t = \max_{1 \leq k \leq K} \eta_t^k. \tag{3.1b}$$

We now construct a new limiting diffusion for the fork–join system which is the same as the original one, except for the Wiener process ξ^0, which no longer appears in the equations. The intuitive reason for this may be understood as follows: The stochastic process $\xi^0(r)$ obtained after appropriately scaling a deterministic input sequence converge to 0 as $r \uparrow \infty$, instead of to a Wiener process as was formerly the case. We shall use the same notation to denote quantities in the new system except that we shall underline them. For $1 \leq k \leq K$, $t \geq 0$, we define

$$\underline{\zeta}_t^k = \sigma_k \xi_t^k - c_k t, \quad \underline{\eta}_t^k = g(\underline{\zeta}^k)_t \tag{3.2a}$$

and

$$\underline{\kappa}_t = \max_{1 \leq k \leq K} \underline{\eta}_t^k. \tag{3.2b}$$

We now present our first result.

Lemma 3.3.1. *Let $\underline{\mathcal{F}}$ be the $\sigma-$ field of events generated on the sample space Ω by the stochastic process $(\xi^1, \ldots \xi^K)$. The inequalities*

$$\underline{\eta}_t^k \leq I\!E[\eta_t^k \mid \underline{\mathcal{F}}], \quad 1 \leq k \leq K, \quad t \geq 0 \tag{3.3}$$

hold, whence

$$\underline{\kappa}_t \leq I\!E[\kappa_t \mid \underline{\mathcal{F}}], \quad t \geq 0 \tag{3.4}$$

Proof. For each $1 \leq k \leq K$ and $t \geq 0$, we have

$$\begin{aligned}\eta_t^k &= g(\zeta^k)_t \\ &= \sup_{0 \leq s \leq t} (\zeta_t^k - \zeta_s^k).\end{aligned}$$

Since

$$\eta_t^k = \sup_{0 \leq s \leq t} (\zeta_t^k - \zeta_s^k) \geq \zeta_t^k - \zeta_s^k, \quad 1 \leq k \leq K, \quad 0 \leq s \leq t,$$

we readily conclude that

$$I\!E(\eta_t^k \mid \underline{I\!S}) \geq I\!E(\zeta_t^k \mid \underline{I\!S}) - I\!E(\zeta_s^k \mid \underline{I\!S}), \quad 1 \leq k \leq K, \quad 0 \leq s \leq t$$

so that

$$\begin{aligned}I\!E(\eta_t^k \mid \underline{I\!S}) &\geq \sup_{0 \leq s \leq t} \left[I\!E(\zeta_t^k \mid \underline{I\!S}) - I\!E(\zeta_s^k \mid \underline{I\!S})\right] \\ &= \sup_{0 \leq s \leq t} (\sigma_k \xi_t^k - \sigma_0 I\!E\xi_t^0 - c_k t - \sigma_k \xi_s^k + \sigma_0 I\!E\xi_s^0 + c_k s).\end{aligned}$$

Since $I\!E\xi_t^0 = 0$ for all $t \geq 0$, we get

$$\begin{aligned}I\!E(\eta_t^k \mid \underline{I\!S}) &\geq \sup_{0 \leq s \leq t} (\sigma_k \xi_t^k - c_k t - \sigma_k \xi_s^k + c_k s) \\ &= \sup_{0 \leq s \leq t} (\underline{\zeta}_t^k - \underline{\zeta}_s^k) \\ &= \underline{\eta}_t^k, \quad 1 \leq k \leq K, \quad t \geq 0\end{aligned}$$

and this proves (3.3). In order to prove (3.4) we note that

$$I\!E(\kappa_t \mid \underline{S}) \geq \max_{1 \leq k \leq K} I\!E(\eta_t^k \mid \underline{S}), \quad t \geq 0$$

and (3.3) now implies

$$I\!E(\kappa_t \mid \underline{S}) \geq \max_{1 \leq k \leq K} \underline{\eta}_t^k = \underline{\kappa}_t, \quad t \geq 0.$$

∎

Theorem 3.3.1. *The following inequality holds*

$$\underline{\kappa}_t \leq_{icx} \kappa_t, \quad t \geq 0. \tag{3.5}$$

Proof. Let $f : I\!R \to I\!R$ be a convex non-decreasing function. By Lemma 3.3.1, we have

$$f(I\!E(\kappa_t \mid \underline{I\!S})) \geq f(\underline{\kappa}_t), \quad t \geq 0.$$

Therefore, upon applying Jensen's inequality, we see that

$$I\!E f(\kappa_t) \geq I\!E f(\underline{\kappa}_t), \quad t \geq 0$$

and (3.5) now follows. ∎

According to Proposition 2.2.1, the condition $c_k > 0, 1 \leq k \leq K$, is sufficient to ensure that the RVs $\underline{\kappa}_t$, and κ_t converge as $t \uparrow \infty$ to proper RVs $\underline{\kappa}_\infty$, and κ_∞ respectively. Our next result shows that the the RVs $\underline{\kappa}_\infty$ and κ_∞ continue to satisfy (3.5).

Theorem 3.3.2 *Under the condition $c_k > 0, 1 \leq k \leq K$, the following inequality holds*

$$\underline{\kappa}_\infty \leq_{icx} \kappa_\infty. \tag{3.6}$$

Proof. Recall from the discussion in Section 2.2.4 that

$$\kappa_t =_{st} \tilde{\kappa}_t \text{ and } \underline{\kappa}_t =_{st} \underline{\tilde{\kappa}}_t, \quad t \geq 0 \tag{3.7}$$

where

$$\tilde{\kappa}_t = \max_{1 \leq k \leq K} \tilde{\eta}_t^k \text{ with } \tilde{\eta}_t^k = \sup_{0 \leq s \leq t} \zeta_s^k, \quad 1 \leq k \leq K, \quad t \geq 0$$

and

$$\underline{\tilde{\kappa}}_t = \max_{1 \le k \le K} \underline{\tilde{\eta}}_t^k \quad \text{with} \quad \underline{\tilde{\eta}}_t^k = \sup_{0 \le s \le t} \zeta_s^k, \quad 1 \le k \le K, \quad t \ge 0.$$

From (3.5) and (3.7) we conclude that for all convex non-decreasing functions f, we have

$$I\!E f(\underline{\tilde{\kappa}}_t) \le I\!E f(\tilde{\kappa}_t), \quad t \ge 0. \tag{3.8}$$

Since the RVs $\tilde{\eta}_t^k$ and $\underline{\tilde{\eta}}_t^k$ for $1 \le k \le K$, are non-decreasing with t, it follows that the RVs $\tilde{\kappa}_t$ and $\underline{\tilde{\kappa}}_t$ are also non-decreasing with t. An application of the monotone convergence theorem now ensures that

$$I\!E f(\underline{\tilde{\kappa}}_\infty) \le I\!E f(\tilde{\kappa}_\infty), \tag{3.9}$$

from which (3.6) is now immediate. ∎

3.4 Upper bounds by association

By using the concept of associated stochastic processes, we exhibit a family of diffusions that bound the diffusion for the end–to–end delay in the sense of strong stochastic ordering. For the case of the fork–join queue operating in its stable regime, upper bounds based on association have been found to be tighter than the upper bounds obtained by convex ordering arguments [6].

Lemma 3.4.1. *For each $t \geq 0$, the RVs $\{\eta_t^k, \ldots, \eta_t^K\}$, are associated.*

Proof. In order to prove this property we use Lemma 3.2.2 of Section 3.2. First note that the RVs

$$\{\xi_t^1 - \xi_s^1, \ldots, \xi_t^K - \xi_s^K, -(\xi_t^0 - \xi_s^0)\}, \quad 0 \leq s \leq t$$

are independent and hence are associated by property (i). By property (iv) the RVs

$$\{\sigma_k[\xi_t^k - \xi_s^k] - \sigma_A[\xi_t^0 - \xi_s^0] - c_k(t-s), 1 \leq k \leq K\}, \quad 0 \leq s \leq t$$

are associated, i.e., the RVs

$$\{\zeta_t^1 - \zeta_s^1, \ldots, \zeta_t^K - \zeta_s^K\}, \quad 0 \leq s \leq t \tag{4.1}$$

are associated.

Fix $t \geq 0$. Define the set $\mathbb{D}_t$ by

$$\mathbb{D}_t = \{\frac{kt}{n}, 0 \leq k \leq n, n = 1, 2 \ldots\}$$

and note that $\mathbb{D}_t$ is a countable dense subset of $[0, t]$. Since the process ζ is separable, it follows [Bi2, p. 468] that

$$\sup_{0 \leq s \leq t} (\zeta_t^k - \zeta_s^k) = \max_{s \in \mathbb{D}_t} (\zeta_t^k - \zeta_s^k), \quad 1 \leq k \leq K. \tag{4.2}$$

For each $n = 1, 2 \ldots$, define the sets of RVs $A_k^n, 1 \leq k \leq n$, by

$$A_k^n = \{\zeta_{\frac{kt}{n}}^1 - \zeta_{\frac{(k-1)t}{n}}^1, \ldots, \zeta_{\frac{kt}{n}}^K - \zeta_{\frac{(k-1)t}{n}}^K\}, \quad 1 \leq k \leq n.$$

By (4.1) it follows that the RVs in each $A_k^n, 1 \leq k \leq n$, are associated. But since the processes ζ has independent increments, it follows from property (ii) that the RVs in $A^n = A_1^n \cup \ldots \cup A_n^n$ are associated. By taking sums of RVs in A^n it follows that the RVs

$$\{\zeta_t^1, \ldots, \zeta_t^K,$$

$$\zeta_t^1 - \zeta_{\frac{t}{n}}^1, \ldots, \zeta_t^K - \zeta_{\frac{t}{n}}^K, \ldots$$

$$\ldots, \zeta_t^1 - \zeta_{\frac{n-1}{n}t}^1, \ldots, \zeta_t^K - \zeta_{\frac{n-1}{n}t}^K\}, \quad n = 1, 2 \ldots, \quad t \geq 0$$

are associated. Another application of property (iv) assures us that the RVs

$$\{\max_{0 \leq k \leq n}(\zeta_t^1 - \zeta_{\frac{kt}{n}}^1), \ldots, \max_{0 \leq k \leq n}(\zeta_t^K - \zeta_{\frac{kt}{n}}^K)\}, \quad n = 1, 2 \ldots, \quad t \geq 0$$

are associated. Letting $n \uparrow \infty$ it follow that the RVs

$$\{\max_{s \in \mathbb{D}_t}(\zeta_t^1 - \zeta_s^1), \ldots, \max_{s \in \mathbb{D}_t}(\zeta_t^K - \zeta_s^K)\}, \quad t \geq 0$$

are associated. Finally from (4.2) we conclude that the RVs

$$\{\max_{0 \leq s \leq t}(\zeta_t^1 - \zeta_s^1), \ldots, \max_{0 \leq s \leq t}(\zeta_t^K - \zeta_s^K)\}, \quad t \geq 0$$

are associated, just another way of saying that the RVs $\{\eta_t^1, \ldots, \eta_t^K\}$ are associated.

∎

We now define the stochastic processes $\overline{\eta}^k \equiv \{\overline{\eta}_t^k, t \geq 0\}, 1 \leq k \leq K$, which form independent versions of the stochastic processes $(\eta^1, \ldots \eta^K)$ in the sense of Definition 3.2.3. For this purpose define K additional independent Wiener processes $\xi^{0,1}, \ldots, \xi^{0,K}$. For $1 \leq k \leq K$, and $t \geq 0$, define

$$\overline{\zeta}_t^k = \sigma_k \xi_t^k - \sigma_0 \xi_t^{0,k} - c_k t, \tag{4.3a}$$

$$\overline{\eta}_t^k = g(\overline{\zeta}^k)_t \quad \text{and} \quad \overline{\kappa}_t = \max_{1 \leq k \leq K} \overline{\eta}_t^k. \tag{4.3b}$$

From Lemma 3.4.1 and Lemma 3.2.1, it directly follows that

$$\kappa_t = \max_{1 \leq k \leq K} \eta_t^k \leq_{st} \max_{1 \leq k \leq K} \overline{\eta}_t^k = \overline{\kappa}_t$$

This is stated in the next result.

Theorem 3.4.1. *The following relation holds true for* $t \geq 0$,

$$\kappa_t \leq_{st} \overline{\kappa}_t. \tag{4.4}$$

It was shown in the last chapter that the condition $c_k > 0, 1 \leq k \leq K$, is necessary and sufficient to ensure that the RVs κ_t and $\overline{\kappa}_t$ converge weakly as $t \uparrow \infty$ to proper RVs κ_∞ and $\overline{\kappa}_\infty$, respectively. The following result is then an immediate consequence of Theorem 3.4.1 and Proposition 1.2.3 in [60].

Theorem 3.4.2. *Under the condition* $c_k > 0, 1 \leq k \leq K$, *the following inequality holds*

$$\kappa_\infty \leq_{st} \overline{\kappa}_\infty. \tag{4.5}$$

3.5 Some computations

In this section we carry out explicit calculations of the distributions of the bounding diffusions in the transient as well the stationary case. We shall denote as a symmetric fork–join queue, the one in which all the K service times have identical probability distribution functions, so that

$$c_1 = c_2 = \ldots = c_K = c, \text{ and } \sigma_1 = \sigma_2 = \ldots = \sigma_K = \sigma$$

We also use the notation $H_K, K = 1, 2 \ldots$ for the partial sums of the Harmonic series, i.e.,

$$H_K = \sum_{k=1}^{K} \frac{1}{k}. \qquad K = 1, 2 \ldots$$

Lower bounds are computed in Section 3.5.1, while upper bounds are computed in Section 3.5.2.

3.5.1 Lower Bounds

The results of Theorem 3.3.1 and Theorem 3.3.2 imply that

$$I\!E\underline{\kappa}_t \leq I\!E\kappa_t, \quad t \geq 0$$

and

$$I\!E\underline{\kappa}_\infty \leq I\!E\kappa_\infty.$$

Our objective in this section is to give explicit formulae for $I\!E\underline{\kappa}_t$ and $I\!E\underline{\kappa}_\infty$.

We first proceed with the calculation of $I\!E\underline{\kappa}_t$. Recall that

$$\underline{\kappa}_t = \max_{1 \leq k \leq K} \underline{\eta}_t^k, \quad t \geq 0$$

where $\underline{\eta}_t^k = g(\sigma_k \xi^k - c_k)_t, t \geq 0, 1 \leq k \leq K$. Note that since the stochastic processes $\xi^k, 1 \leq k \leq K$, are independent, it follows that the stochastic processes $\underline{\eta}^k, 1 \leq k \leq K$ are also independent, and therefore

$$I\!P(\underline{\kappa}_t \leq z) = \prod_{k=1}^{K} I\!P(\underline{\eta}_t^k \leq z), \quad t \geq 0 \tag{5.1}$$

for all $z \geq 0$. Note that $\underline{\eta}^k, 1 \leq k \leq K$, are diffusion processes with drift $-c_k$ and variance σ_k, which are reflected from the origin. The transient distribution for these processes is well known [24]. For all $z \geq 0$,

$$I\!P(\underline{\eta}_t^k \leq z) = \Phi\left(\frac{z + c_k t}{\sigma_k \sqrt{t}}\right) - e^{-\frac{2c_k z}{\sigma_k^2}} \Phi\left(\frac{-z + c_k t}{\sigma_k \sqrt{t}}\right), \quad t \geq 0. \tag{5.2}$$

From (5.1) and (5.2) it follows that for all $z \geq 0$,

$$I\!P(\underline{\kappa}_t \leq z) = \prod_{k=1}^{K} \left[\Phi\left(\frac{z + c_k t}{\sigma_k \sqrt{t}}\right) - e^{-\frac{2c_k z}{\sigma_k^2}} \Phi\left(\frac{-z + c_k t}{\sigma_k \sqrt{t}}\right)\right], \quad t \geq 0. \tag{5.3}$$

Since

$$I\!E\underline{\kappa}_t = \int_0^\infty [1 - \prod_{k=1}^{K} I\!P(g(\underline{\zeta}^k)_t \leq z)] dz, \quad t \geq 0 \tag{5.4}$$

it follows that

$$I\!E\underline{\kappa}_t = \int_0^\infty \left[1 - \prod_{k=1}^{K} \left(\Phi\left(\frac{z + c_k t}{\sigma_k \sqrt{t}}\right) - e^{-\frac{2c_k z}{\sigma_k^2}} \Phi\left(\frac{-z + c_k t}{\sigma_k \sqrt{t}}\right)\right)\right] dz, \quad t \geq 0. \tag{5.5}$$

In the symmetric case, equations (5.3) and (5.5) reduce for all $z \geq 0$ to

$$I\!P(\underline{\kappa}_t \leq z) = \left[\Phi\left(\frac{z + ct}{\sigma\sqrt{t}}\right) - e^{-\frac{2cz}{\sigma^2}} \Phi\left(\frac{-z + ct}{\sigma\sqrt{t}}\right)\right]^K, \quad t \geq 0 \tag{5.6}$$

and

$$I\!E\underline{\kappa}_t = \int_0^\infty \left[1 - \left(\Phi\left(\frac{z + ct}{\sigma\sqrt{t}}\right) - e^{-\frac{2cz}{\sigma^2}} \Phi\left(\frac{-z + ct}{\sigma\sqrt{t}}\right)\right)^K\right] dz, \quad t \geq 0. \tag{5.7}$$

We now proceed with the calculation of $I\!E\underline{\kappa}_\infty$ under the condition $c_k > 0, 1 \leq k \leq K$. Recalling that

$$\underline{\kappa}_\infty = \max_{1 \leq k \leq K} \underline{\eta}_\infty^k,$$

we see from (5.2) that

$$I\!P(\underline{\eta}_{\infty}^{k} \leq z) = 1 - e^{-\frac{2c_k z}{\sigma_k^2}}, \quad z \geq 0 \tag{5.8}$$

so that

$$I\!P(\underline{\kappa}_{\infty} \leq z) = \prod_{k=1}^{K} [1 - e^{-\frac{2c_k z}{\sigma_k^2}}], \quad z \geq 0 \tag{5.9}$$

Hence $I\!E\underline{\kappa}_{\infty}$ is given by

$$I\!E\underline{\kappa}_{\infty} = \int_0^{\infty} [1 - \prod_{k=1}^{K} (1 - e^{-\frac{2c_k z}{\sigma_k^2}})] dz \tag{5.10}$$

It is clear that

$$1 - \prod_{k=1}^{K} (1 - e^{-\frac{2c_k z}{\sigma_k^2}}) = \sum_{k=1}^{K} (-1)^{k+1} \sum_{I \in I\!I_k} e^{-\sum_{k \in I} \frac{2c_k z}{\sigma_k^2}} \tag{5.11}$$

where

$$I\!I_k = \{I \subseteq \{1, \ldots, K\} :| I |= k\}, \quad 1 \leq k \leq K.$$

For any non-empty subset I of $\{1, \ldots, K\}$, we see that

$$\int_0^{\infty} e^{-\sum_{k \in I} \frac{2c_k z}{\sigma_k^2}} dz = (\sum_{k \in I} \frac{2c_k}{\sigma_k^2})^{-1} \tag{5.12}$$

so that

$$I\!E\underline{\kappa}_{\infty} = \sum_{k=1}^{K} (-1)^{k+1} \sum_{I \in I\!I_k} (\sum_{k \in I} \frac{2c_k}{\sigma_k^2})^{-1}. \tag{5.13}$$

In the symmetric case, equations (5.9) and (5.10) reduce to

$$I\!P(\underline{\kappa}_{\infty} \leq z) = [1 - e^{-\frac{2cz}{\sigma^2}}]^K, \quad z \geq 0 \tag{5.14}$$

and

$$I\!E\underline{\kappa}_{\infty} = \int_0^{\infty} [1 - e^{-\frac{2cz}{\sigma^2}}]^K dz. \tag{5.15}$$

Taking note of the fact that $| \mathcal{I} | = \binom{K}{k}$ and of the identity

$$\sum_{k=1}^{K} \frac{(-1)^{k+1}}{k} \binom{K}{k} = \sum_{k=1}^{K} \frac{1}{k} = H_K, \qquad K = 1, 2 \ldots$$

equation (5.15) reduces to

$$\mathbb{E}\underline{\kappa}_\infty = \frac{\sigma^2}{2c} H_K. \tag{5.16}$$

3.5.2 Upper Bounds

The results of Theorem 3.4.1 and Theorem 3.4.2 imply that

$$\mathbb{E}\overline{\kappa}_t \geq \mathbb{E}\kappa_t, \quad t \geq 0$$

and

$$\mathbb{E}\overline{\kappa}_\infty \geq \mathbb{E}\kappa_\infty.$$

Our objective in this section is to give explicit formulae for $\mathbb{E}\overline{\kappa}_t$ and $\mathbb{E}\overline{\kappa}_\infty$. Since all the calculations involved are exactly the same as in the last section, we only give the final formulae in each case.

Proceeding exactly as in the last section, it is possible to show that

$$\mathbb{E}\overline{\kappa}_t = \int_0^\infty \left[1 - \prod_{k=1}^{K} \left(\Phi\left(\frac{z + c_k t}{\sqrt{(\sigma_k^2 + \sigma_0^2)t}} \right) - e^{-\frac{2c_k z}{\sigma_k^2 + \sigma_0^2}} \Phi\left(\frac{-z + c_k t}{\sqrt{(\sigma_k^2 + \sigma_0^2)t}} \right) \right) \right] dz,$$

$$t \geq 0 \quad (5.17)$$

and in the symmetric case,

$$\mathbb{E}\overline{\kappa}_t = \int_0^\infty \left[1 - \left(\Phi\left(\frac{z + ct}{\sqrt{(\sigma^2 + \sigma_0^2)t}} \right) - e^{-\frac{2cz}{\sigma^2 + \sigma_0^2}} \Phi\left(\frac{-z + ct}{\sqrt{(\sigma^2 + \sigma_0^2)t}} \right) \right)^K \right] dz,$$

$$t \geq 0. \quad (5.18)$$

Under the condition $c_k > 0, 1 \leq k \leq K$, we further have that

$$I\!E\overline{\kappa}_\infty = \sum_{k=1}^{K}(-1)^{k+1} \sum_{I \in \mathcal{I}_k} (\sum_{k \in I} \frac{2c_k}{\sigma_k^2 + \sigma_0^2})^{-1} \tag{5.19}$$

and in the symmetric case

$$I\!E\overline{\kappa}_\infty = \frac{\sigma^2 + \sigma_0^2}{2c} H_K. \tag{5.20}$$

Equations (5.16) and (5.20) imply that in the symmetric case

$$\frac{\sigma^2}{2c} H_K \leq I\!E\kappa_\infty \leq \frac{\sigma^2 + \sigma_0^2}{2c} H_K, \tag{5.21}$$

and since

$$\log(K+1) \leq H_K \leq \log K$$

it follows that the expectation of the normalized end–to–end delay of a symmetric fork–join queue in heavy traffic, increases logarithmically with K.

Equation (5.21) reveals an interesting difference between the asymptotic behavior in K, for fork–join queues operating in heavy traffic with those operating in their stable regime. It was shown in [4] that moments of the end–to–end delay of a stable fork–join queue increase logarithmically in K provided the following condition is satisfied;

Let $A^*(s)$ and $B^*(s)$ denote the Laplace-Stieltjes transform of the interarrival and service times. The transform $B^*(s)$ is assumed to be rational so that the function $s \to f(s)$ which is initially defined for $\mathbf{R}e(s) = 0$ by

$$f(s) = A^*(s)B^*(-s)$$

is continuable in the region $\mathbf{R}e(s) \geq 0$.

Under this assumption it is shown in [4] that the response time of each queue has an exponential tail, which leads to the logarithmic behavior. However in heavy traffic the response times *always* have an exponential tail provided they satisfy

assumptions **(IIa)–(IIc)** from Chapter 2. Hence even those fork–join queues whose end–to–end delay does not grow logarithmically with K when they are in their stable regime, (since they do not satisfy the above assumption), exhibit logarithmic growth of their end–to–end delay with K, once they are in heavy traffic.

CHAPTER IV

4.1 Introduction

In Chapter 2 we obtained a diffusion limit for the delay processes in a single stage fork–join queue. We saw that this diffusion was a correlated Wiener process with drift in the non–negative orthant, with normal reflection at each boundary. Our objective in this chapter is to obtain a PDE for the stationary distribution of this diffusion.

The most general theory for multi–dimensional diffusions with reflections is the one given by Stroock and Varadhan [61]. However this theory is applicable only if the domain within which the diffusion is confined is bounded and has smooth boundaries. Our reflected diffusion does not satisfy these conditions, and therefore lies outside the scope of the Stroock–Varadhan theory.

In Section 4.2, we show that our diffusion is in fact a strong Markov process, by taking advantage of its sample path structure. The next step is to obtain the equation satisfied by the stationary density of this diffusion. Harrison and Williams [25], have given a integral equation that is satisfied by the stationary density. Using this integral equation, it is possible to derive the PDE that the stationary density satisfies [22], [25]. There is another way by which this PDE can be obtained [22], and that is by first writing down the forward PDE that is satisfied by the transition density for the process (with the help of the Ito formula). The PDE for the stationary distribution is this forward PDE at steady state. In this chapter we follow that latter method.

The rest of the chapter is organized as follows: In section 4.2 we give a path–by–path construction of diffusion process obtained as a weak limit of of the queue delay processes and show that it is strong Markov process with continuous sample paths. Following Harrison and Reiman [22], in Section 4.3 we give a heuristic

derivation of the PDE that is satisfied by the stationary distribution of the diffusion process. Finally in Section 4.4 we give the PDE that is satisfied by the stationary distribution of the diffusion for the queue length process.

4.2 The Markov property

For every $x = (x^1, \ldots, x^K)$ in $\mathbb{R}_+^K$, let P_x be probability measure on the function space $C^K[0,\infty)$, such that under this measure, $(\zeta^1, \ldots, \zeta^K)$ is a diffusion process satisfying $\zeta_0 = x$. Further more it has a covariance matrix R and drift vector c given by

$$c = -(c_1, c_2, \ldots, c_K) \tag{2.1a}$$

and

$$R = \begin{pmatrix} \sigma_1^2 + \sigma_0^2 & \sigma_0^2 & \ldots & \sigma_0^2 \\ \sigma_0^2 & \sigma_2^2 + \sigma_0^2 & \ldots & \sigma_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_0^2 & \sigma_0^2 & \ldots & \sigma_K^2 + \sigma_0^2 \end{pmatrix}. \tag{2.1b}$$

We now define a mapping $h : D[0,\infty) \to D[0,\infty)$, by

$$h(x)_t = -\inf_{0 \le s \le t} x_s, \quad t \ge 0. \tag{2.2}$$

Define a K–dimensional process $\gamma^k \equiv \{\gamma_t^k, t \ge 0\}, 1 \le k \le K$, by

$$\gamma_t^k = h(\zeta^k)_t, \quad 1 \le k \le K, \quad t \ge 0. \tag{2.3}$$

The K–dimensional reflected diffusion process, $(\eta^1, \ldots, \eta^K)$, is then given by

$$\eta_t^k = \zeta_t^k + \gamma_t^k, \quad 1 \le k \le K, \quad t \ge 0. \tag{2.4}$$

Define $\mathcal{F}_t = \mathcal{F}(\zeta_s^k, 0 \le s \le t, 1 \le k \le K), t \ge 0$. We now present the main result of this section.

Proposition 4.2.1

(a) *For each $t \ge 0$, the processes η_t and γ_t are measurable with respect to $\mathcal{F}_t$.*

(b) *The process η is a Markov process with stationary transition probabilities.*

Proof.

(a) From (2.3) and (2.4), it is clear that η_t and γ_t depend only on the restriction of ζ_t to the interval $[0,t]$. This fact implies Part (a) of the Proposition.

(b) We first show that (ζ_t, γ_t) jointly form a Markov process under the measure P_x, through the following sequence of equalities. Note that

$$E_x[e^{iy'\zeta_{t+h}+iz'\gamma_{t+h}} \mid \mathcal{F}_t]$$

$$= e^{iy'\zeta_t+iz'\gamma_t} E_x[e^{iy'(\zeta_{t+h}-\zeta_t)+iz'(\gamma_{t+h}-\gamma_t)} \mid \mathcal{F}_t] \tag{2.6}$$

Furthermore, for $1 \leq k \leq K$, we have that

$$\begin{aligned}
\gamma^k_{t+h} - \gamma^k_t &= h(\zeta^k)_{t+h} - h(\zeta^k)_t \\
&= -\inf\{-h(\zeta^k)_t, \inf_{t\leq s\leq t+h} \zeta^k_s\} - h(\zeta^k)_t \\
&= \inf\{0, \inf_{t\leq s\leq t+h} \zeta^k_s + h(\zeta^k)_t\} \\
&= \inf\{0, \inf_{t\leq s\leq t+h} \zeta^k_s - \zeta^k_t + \zeta^k_t + h(\zeta^k)_t\} \\
&= -\inf\{0, \inf_{t\leq s\leq t+h} \zeta^k_s - \zeta^k_t + \eta^k_t\}
\end{aligned} \tag{2.7}$$

From (2.6)–(2.7), it is clear that

$$E_x[e^{iy'\zeta_{t+h}+iz'\gamma_{t+h}} \mid \mathcal{F}_t]$$

$$= e^{iy'\zeta_t+iz'\gamma_t} E_0[e^{iy'\zeta_h} e^{-iz'\inf\{0,\inf_{0\leq s\leq h}\zeta_s+u\}}], \quad P_x \text{ a.s.} \tag{2.8}$$

on subsets of $\mathcal{F}_t$ in which $\eta_t = u$. This implies the Markov property for (ζ_t, γ_t) under the the measure P_x. To conclude the Markov proporty for the process η_t, simply put $y = z$ in (2.8).

∎

We have already shown in Chapter 2 that the process η has a stationary distribution iff $c_k > 0, 1 \leq k \leq K$. In the next section we obtain a PDE whose solution gives this stationary distribution.

4.3 A PDE for the stationary distribution

In this section our objective is to obtain a PDE that is satisfied by the stationary distribution of the Markov process η. Harrison and Williams [25] gave an integral relation which they called the *basic adjoint relation* (BAR) that must be satisfied by any stationary distribution for the process. However the BAR has been shown to be only a necessary and not sufficient condition for the stationary distribution for the process. It was used them to give necessary and sufficient conditions for the process to have a product form stationary distribution. If we apply these results to our problem, it can be shown that the process η will have a product form stationary distribution iff $\sigma_0 = 0$, i.e., the arrival stream into the system is deterministic which is clearly not a very interesting case. Hence we proceed heuristically and assume that the BAR is a necessary as well as a sufficient condition for the stationary distribution of the process.

Using the BAR it is possible to obtain the PDE that the stationary distribution satisfies [26]. However since the calculations involved are quite tedious, we adopt an indirect way of deriving these PDE's in this chapter. We proceed by writing down the forward PDE's that the stationary transition density satisfies, and then assuming that the process is positive recurrent, the PDE satisfied by the stationary distribution is simply this forward PDE in steady state. This assumption has been shown to be true by Harrison and Reiman [22] for the diffusion obtained by taking the heavy traffic limit of two queues in tandem.

Consider the following K correlated Wiener processes,

$$\zeta_t^k = \sigma_k \xi_t^k - \sigma_0 \xi_t^0 - c_k t, \quad 1 \leq k \leq K, \quad t \geq 0 \tag{3.1}$$

Recall that the drift vector and the covariance matrix for these diffusions are given by c and R defined in (2.6).

Define the following matrices

$$\Gamma = \begin{pmatrix} \sigma_1^2 + \sigma_0^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 + \sigma_0^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_K^2 + \sigma_0^2 \end{pmatrix} \tag{3.2}$$

and

$$\Pi = \begin{pmatrix} \sigma_1^2 + \sigma_0^2 & 2\sigma_0^2 & \dots & 2\sigma_0^2 \\ 2\sigma_0^2 & \sigma_2^2 + \sigma_0^2 & \dots & 2\sigma_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ 2\sigma_0^2 & 2\sigma_0^2 & \dots & \sigma_K^2 + \sigma_0^2 \end{pmatrix}. \tag{3.3}$$

Let $P_x(.)$ be a distribution on the path space of ζ corresponding to starting state $x = (x^1, \dots, x^K) \in I\!R_+^K$. Recall that $\gamma = h(\zeta)$ and $\eta = g(\zeta)$, so that

$$\eta_t^k = \zeta_t^k + \gamma_t^k, \quad 1 \le k \le K, \quad t \ge 0. \tag{3.4}$$

We now write $\zeta_t = \beta_t + ct$ where $\beta = \{\beta_t^k, t \ge 0\}, 1 \le k \le K$, is a K dimensional Wiener process with covariance matrix R, zero drift and $\beta_0 = \zeta_0 = \eta_0$. This implies that

$$\eta_t^k = \beta_t^k + \upsilon_t^k, \quad 1 \le k \le K, \quad t \ge 0, \tag{3.5}$$

where

$$\upsilon_t^k = -c_k t + \gamma_t^k, \quad 1 \le k \le K, \quad t \ge 0. \tag{3.6}$$

Note that β^k is a martingale over the σ-fields $I\!F_t$ and υ^k is a continuous adapted process of bounded variation. Thus each η^k is a continuous semimartingale and hence we can apply the Ito-formula to it. Let $I\!C_1$ be the class of functions $f(t,x,y); x, y \in I\!R_+^K$ that are continuously differentiable in t and $y^k, 1 \le k \le K$, and are twice continuously differentiable in $x^k, 1 \le k \le K$. Suppose that $f(t,x,y) \in I\!C_1$ and let

$$L_x f = \frac{1}{2} \sum_{i=1}^{K} \sum_{j=1}^{K} R_{ij} \frac{\partial^2 f}{\partial x^i \partial x^j}. \tag{3.7}$$

We also use the following convention: If θ is a K dimensional vector, then the directional derivative of a function f in the direction of θ is denoted by

$$\theta \nabla_x f(t,x,y) = \sum_{j=1}^{K} \theta_j \frac{\partial}{\partial x^j} f(t,x,y).$$

Then one has the following result.

Lemma 4.3.1. *If $f \in \mathcal{C}_1$ then*

$$\begin{aligned} f(t,\eta_t,\gamma_t) - f(0,\eta_0,0) = & \int_0^t (L_x + c\nabla_x + \frac{\partial}{\partial u}) f(u,\eta_u,\gamma_u) du \\ & + \sum_{i=1}^{K} \int_0^t \frac{\partial}{\partial x^i} f(u,\eta_u,\gamma_u) d\beta_u^i \\ & + \sum_{i=1}^{K} \int_0^t (\frac{\partial}{\partial x^i} + \frac{\partial}{\partial y^i}) f(u,\eta_u,\gamma_u) d\gamma_u^i, \quad t \geq 0. \end{aligned} \tag{3.8}$$

Here integrals involving $d\beta_u^i$ are of the Ito type, and those involving γ_u^j are defined path-by-path as ordinary Riemann-Stieltjes integrals.

Proof. This is a direct consequence of the multi-dimensional generalized Ito formula (see Appendix C, Theorem C1). ∎

We now proceed to obtain the backward and forward Kolmogorov equations for the reflected diffusion process η. The approach that we shall follow is basically the same as given by Harrison and Reiman [22] and proceeds by the following two steps:

(1) Using the Girsanov transformation [24], we express the transition probability density for the diffusion η with drift in terms of an augmented transition probability density (defined below) for this diffusion without drift (i.e. $c = 0$).

(2) Using the Ito formula we obtain the partial differential equations satisfied by the augmented transition probability density for the diffusion η without drift.

Let P_x^c be the probability measure on the path space of the process ζ, corresponding to the initial state $\zeta_0 = x$ and drift vector c and let E_x^c be the corresponding expectation. Let P_x and E_x be the corresponding quantities for the process ζ with drift $c = 0$.

We now define the transition density of diffusion η with initial position x and drift c, as a function $p^c(t, x, z)$ such that

$$P_x^c(\eta_t \in B) = \int_B p^c(t, x, z)dz, \tag{3.9}$$

for any Borel subset B of $\mathbb{R}_+^K$. This can also be expressed as

$$p^c(t, x, z)dz = P_x^c(\eta_t \in dz) \tag{3.10}$$

or

$$p^c(t, x, z)dz = E_x^c[\delta(\eta_t - z)] \tag{3.11}$$

We now proceed to define the augmented transition density for η without drift, as a function $g(t, x, z, \alpha)$ such that

$$E_x[e^{\alpha\gamma_t'} 1_B(\eta_t)] = \int_B g(t, x, z, \alpha)dz \tag{3.12}$$

for any Borel subset B of $\mathbb{R}_+^K$. This can also be written as

$$g(t, x, z, \alpha) = E_x[e^{\alpha\gamma_t'}\delta(\eta_t - z)] \tag{3.13}$$

In the next Lemma from Harrison and Reiman [22], we express $p^c(t, x, z)$ in terms of $g(t, x, z, \alpha)$ with an appropriately chosen α.

Lemma 4.3.2 *The following formula holds true*

$$p^c(t, x, z) = e^{-cM(z-x)' + \frac{1}{2}(cMc')t} g(t, x, z, cM) \tag{3.14}$$

where $M = R^{-1}$.

Proof. By applying Girsanov's formula for the process ζ, we obtain

$$P_x^c(G) = \int_G e^{-cM(\zeta_t - x)' - \frac{1}{2}(cMc')t} dP_x \tag{3.15}$$

for all sets $G \in \mathcal{F}_t$. But $\zeta_t = \eta_t - \gamma_t$, so

$$dP_x^c = e^{-cM(\eta_t - \gamma_t - x)' - \frac{1}{2}(cMc')t} dP_x \tag{3.16}$$

Hence

$$\begin{aligned} P_x^c(\eta_t \in dz) &= E_x^c[\delta(\eta_t - z)] \\ &= E_x[e^{-cM(\eta_t - \gamma_t - x)' - \frac{1}{2}(cMc')t}\delta(\eta_t - z)] \\ &= E_x[e^{-cM(z - \gamma_t - x)' - \frac{1}{2}(cMc')t}\delta(\eta_t - z)] \\ &= e^{-cM(z-x)' - \frac{1}{2}(cMc')t} E_x[e^{cM\gamma_t'}\delta(\eta_t - z)] \\ &= e^{-cM(z-x)' - \frac{1}{2}(cMc')t} g(t, x, z, cM) \end{aligned} \tag{3.17}$$

so that

$$p^c(t, x, z) = e^{-cM(z-x)' + \frac{1}{2}(cMc')t} g(t, x, z, cM). \tag{3.18}$$

∎

Once again we follow Harrison and Reiman [22] in obtaining the PDE's for the augmented transition density g. Note that by (3.14) it is sufficient to obtain the PDE's satisfied by the augmented transition density g in order to obtain the PDE's satisfied by the transition density p^c.

Consider a function $\phi(t, x, \alpha)$ which is continuously in t and twice continuously differentiable in $x^k, 1 \le k \le K$ and further satisfies

$$L_x\phi - \frac{\partial}{\partial t}\phi = 0, \tag{3.19a}$$

$$\frac{\partial \phi}{\partial x^i} + \alpha_i \phi = 0 \quad \text{if} \quad x^i = 0, \tag{3.19b}$$

$$\phi(0, x, \alpha) = \psi(x), \tag{3.19c}$$

where $\psi(x)$ is bounded and continuous on S. For a fixed $t > 0$ let

$$f(u, x, y) = e^{y\alpha'}\phi(t - u, x, \alpha), \quad 0 \le u \le t. \tag{3.20}$$

Then

$$\begin{aligned} f(t,\eta_t,\gamma_t) - f(0,x,0) &= e^{\gamma_t \alpha'}\phi(0,\eta_t,\alpha) - \phi(t,x,\alpha) \\ &= e^{\gamma_t \alpha'}\psi(\eta_t) - \phi(t,x,\alpha) \end{aligned} \tag{3.21}$$

by (3.19c). From (3.19a), (3.19b) and (3.20)

$$(L_x + \frac{\partial}{\partial u})f = 0 \tag{3.22}$$

and

$$(\frac{\partial}{\partial x^i} + \frac{\partial}{\partial y^i})f = 0 \text{ if } x_i = 0. \tag{3.23}$$

Substituting this f into the Ito formula (3.8) with $c = 0$ and taking expectations we obtain

$$\phi(t,x,\alpha) = E_x[e^{\gamma_t \alpha'}\psi(\eta_t)]. \tag{3.24}$$

Now suppose that their exists a function $g \in \mathcal{C}_1$ satisfying the augmented backward equation

$$L_x g - \frac{\partial}{\partial t}g = 0, \tag{3.25a}$$

$$\frac{\partial g}{\partial x^i} + \alpha_i g = 0 \quad \text{if} \quad x_i = 0, \tag{3.25b}$$

$$g(0,x,z,\alpha) = \delta(x-z), \tag{3.25c}$$

where δ is the Dirac delta function. Define

$$\phi(t,x,\alpha) = \int_{\mathbb{R}^K_+} \psi(z)g(t,x,z,\alpha)dz. \tag{3.26}$$

From (3.25a)–(3.25c) it follows that ϕ satisfies (3.19a)–(3.19c) and thus by (3.24)

$$E_x[e^{\gamma_t \alpha'}\psi(\eta_t)] = \int_{\mathbb{R}^K_+} \psi(z)g(t,x,z,\alpha)dz. \tag{3.27}$$

Choosing ψ to be the indicator function $1_B(.)$ for any Borel set B of $\mathbb{R}^K_+$, we obtain

$$E_x[e^{\gamma_t \alpha'}1_B(\eta_t)] = \int_B g(t,x,z,\alpha)dz. \tag{3.28}$$

Hence the function $g(t,x,z,\alpha)$ is in fact the augmented transition density for the diffusion η without drift that was defined earlier in (3.12) and (3.25a)–(3.25c) are the backward partial differential equations which this function satisfies. By means of complex calculations [22], it is possible to show that that $g(t,x,z,\alpha)$ satisfies (3.25a)–(3.25c) if and only if it satisfies the following forward PDE's,

$$L_z g - \frac{\partial}{\partial t} g = 0, \tag{3.29a}$$

$$\Pi_i \nabla_z g + \alpha_i R_{ii} = 0 \text{ if } x_i = 0, \tag{3.29b}$$

$$g(0,x,z,\alpha) = \delta(x-z). \tag{3.29c}$$

Finally we combine Lemma 3.3.2 with (3.29a)–(3.29c) to obtain the backward and forward equations for the diffusion η with drift c. These equations are stated in the next theorem [22].

Theorem 4.3.1 *The transition density $p^c(t,x,z)$ satisfies backward equations*

$$L_x p^c + c\nabla_x p^c - \frac{\partial}{\partial t} p^c = 0, \tag{3.30a}$$

$$\frac{\partial p^c}{\partial x^i} = 0 \quad \text{if} \quad x^i = 0, \tag{3.30b}$$

$$p^c(0,x,z) = \delta(x-z). \tag{3.30c}$$

It also satisfies the forward equations

$$L_z p^c - c\nabla_z p^c - \frac{\partial}{\partial t} p^c = 0, \tag{3.31a}$$

$$\Pi_i \nabla_z p^c - 2c_i p^c = 0 \quad \text{if} \quad z^i = 0, \tag{3.31b}$$

$$p^c(0,x,z) = \delta(x-z). \tag{3.31c}$$

∎

Conditions for the Markov process η to be positive recurrent are not known in general. However if that is the case, and if $\pi^c(z)$ is the the stationary density function of the process (which we know exists since $c_k > 0, 1 \le k \le K$), then

$$p^c(t,x,z) \to \pi^c(z) \quad \text{as} \quad t \uparrow \infty \quad \text{for all} \quad x \quad \text{in} \quad I\!R_+^K$$

In this case equations (3.31a)-(3.31c) suggest that $\pi^c(z)$ satisfies the following equations.

$$L_z\pi^c - c\nabla_z\pi^c = 0, \tag{3.32a}$$

$$\Pi_i\nabla_z\pi^c - 2c_i\pi^c = 0 \quad \text{if} \quad z_i = 0, \tag{3.32b}$$

$$\int_{\mathbb{R}^K_+} \pi^c(z)dz = 1. \tag{3.32c}$$

Equations (3.32a)–(3.32c) will solved in the next chapter to obtain a formula for π^c.

4.1 The queue length processes

An argument similar to the one outlined in the last section can be used to obtain the PDE's satisfied by the stationary distribution of the queue lengths in the fork–join queue. Recall from Chapter 2 that the diffusions for the queue lengths are given by

$$Q^k = g(\phi^k), \quad 1 \leq k \leq K \tag{4.1}$$

where

$$\phi_t^k = \tau_0 \xi_t^0 - \tau_k \xi_t^k + d_k t, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{4.2}$$

The co-variance matrix Q and the drift vector d for ϕ are given by

$$d = (\, d_1, d_2, \ldots, d_K \,), \tag{4.3a}$$

and

$$Q = \begin{pmatrix} \tau_1^2 + \tau_0^2 & \tau_0^2 & \ldots & \tau_0^2 \\ \tau_0^2 & \tau_2^2 + \tau_0^2 & \ldots & \tau_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau_0^2 & \tau_0^2 & \ldots & \tau_K^2 + \tau_0^2 \end{pmatrix}. \tag{4.3b}$$

Also define the following matrices,

$$\Gamma_q = \begin{pmatrix} \tau_1^2 + \tau_0^2 & 0 & \ldots & 0 \\ 0 & \tau_2^2 + \tau_0^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tau_K^2 + \tau_0^2 \end{pmatrix} \tag{4.4}$$

and

$$\Pi_q = \begin{pmatrix} \tau_1^2 + \tau_0^2 & 2\tau_0^2 & \ldots & 2\tau_0^2 \\ 2\tau_0^2 & \tau_2^2 + \tau_0^2 & \ldots & 2\tau_0^2 \\ \vdots & \vdots & \ddots & \vdots \\ 2\tau_0^2 & 2\tau_0^2 & \ldots & \tau_K^2 + \tau_0^2 \end{pmatrix}. \tag{4.5}$$

Let P_x^d be the probability measure on the path space of the process ϕ, corresponding to the initial state $\phi_0 = x$ and drift vector d. We now define the

transition density of diffusion Q with drift d and initial position x as a function $p_q^d(t,x,z)$ such that

$$P_x^d(Q \in B) = \int_B p_q^d(t,x,z)dz, \tag{4.6}$$

for any Borel subset B of $\mathbb{R}_+^K$. Then the following result holds.

Lemma 4.4.1. *The transition density $p_q^d(t,x,z)$ satisfies the backward equations*

$$L_x p_q^d - d\nabla_x p_q^d - \frac{\partial}{\partial t} p_q^d = 0, \tag{4.7a}$$

$$\frac{\partial p^d}{\partial x^i} = 0 \quad \text{if} \quad x^i = 0, \tag{4.7b}$$

$$p_q^d(0,x,z) = \delta(x-z). \tag{4.7c}$$

It also satisfies the forward equations

$$L_z p_q^d + d\nabla_z p_q^d - \frac{\partial}{\partial t} p_q^d = 0, \tag{4.8a}$$

$$\Pi_{q,i}\nabla_z p_q^d + 2dp_q = 0 \quad \text{if} \quad z^i = 0, \tag{4.8b}$$

$$p_q^d(0,x,z) = \delta(x-z). \tag{4.8c}$$

Finally we let $t \uparrow \infty$ in the forward equations (4.8a)-(4.8c) to obtain the PDE's for the stationary distribution of the diffusion for the queue lengths in the fork–join queue.

$$L_z \pi_q + d\nabla_z \pi_q = 0, \tag{4.9a}$$

$$\Pi_{q,i}\nabla_z \pi_q^d + 2d\pi_q^d = 0 \quad \text{if} \quad z_i = 0, \tag{4.9b}$$

$$\int_S \pi_q^d dz = 1. \tag{4.9c}$$

CHAPTER V

5.1 Introduction

In this chapter our objective is to obtain a solution to the PDE for the stationary density of the diffusion process for the queue delays in a fork–join queue, which was derived in Chapter 4. From this stationary density we can then recover some heavy traffic information about the queue. We only consider the solution of the PDE for the case of two independent variables, so that the results of this chapter are applicable to two dimensional fork–join systems. However, as illustrated in the next chapter, the solution that we obtain for this case helps us in obtaining some information about heavy traffic behavior for general K, without having to solve PDE's.

The technique that we shall use for solving the PDE is similar to the one used by Harrison [21] and Foschini [16] in the context of a system of single server queues in tandem. For this technique to be applicable, it is necessary to assume that $\sigma_0 = \sigma_1 = \sigma_2$. In Section 5.3 we obtain formulae for all the moments of the diffusion for the end–to–end delay for this case. These moments are combined with light traffic results in Chapter 6 in order to obtain interpolation approximations. Lastly in Section 5.4 we give the corresponding solutions for the diffusion due to the queue length processes in the fork–join queue.

We use the following notation. As in the last chapter, the non–negative quadrant in the (x, y)–plane will be reffered to as $\mathbb{R}^2_+$. For each $0 \leq \beta \leq 1$, the region in the first and fourth quadrants that is bounded by the lines $y = \sqrt{\frac{1+\beta}{1-\beta}}x$ and $y = -\sqrt{\frac{1+\beta}{1-\beta}}x$ will be referred to as $\mathbb{R}^2_\beta$, i.e.,

$$\mathbb{R}^2_\beta = \{(x, y) \in \mathbb{R}^2_+ : -\sqrt{\frac{1+\beta}{1-\beta}}x \leq y \leq \sqrt{\frac{1+\beta}{1-\beta}}x\}$$

5.2 The queue delay processes: Symmetrical case

We consider the problem of determining the stationary density for the waiting time processes of the fork–join queue in the case when $K = 2$. We start by writing down the PDE obtained in the Chapter 4, which the stationary density satisfies.

$$\frac{1}{2}(\sigma_1^2+\sigma_0^2)\frac{\partial^2\pi(x,y)}{\partial x^2}+\sigma_0^2\frac{\partial^2\pi(x,y)}{\partial x\partial y}+\frac{1}{2}(\sigma_2^2+\sigma_0^2)\frac{\partial^2\pi(x,y)}{\partial y^2}$$

$$+c_1\frac{\partial\pi(x,y)}{\partial x}+c_2\frac{\partial\pi(x,y)}{\partial y}=0,\quad (x,y)\in I\!R_+^2 \tag{2.1a}$$

$$BC(x=0):\frac{1}{2}(\sigma_1^2+\sigma_0^2)\frac{\partial\pi(0,y)}{\partial x}+\sigma_0^2\frac{\partial\pi(0,y)}{\partial y}+c_1\pi(0,y)=0 \tag{2.1b}$$

$$BC(y=0):\sigma_0^2\frac{\partial\pi(x,0)}{\partial x}+\frac{1}{2}(\sigma_2^2+\sigma_0^2)\frac{\partial\pi(x,0)}{\partial y}+c_2\pi(x,0)=0. \tag{2.1c}$$

We further make the assumption that the two queues are identical with $\sigma_1 = \sigma_2 = \sigma$ and $c_1 = c_2 = c$, and we set $\alpha^2 = \sigma_0^2+\sigma^2$ in what follows. The equilibrium equations then simplify to the following.

$$\frac{1}{2}\alpha^2\frac{\partial^2\pi(x,y)}{\partial x^2}+\sigma_0^2\frac{\partial^2\pi(x,y)}{\partial x\partial y}+\frac{1}{2}\alpha^2\frac{\partial^2\pi(x,y)}{\partial y^2}$$

$$+c\frac{\partial\pi(x,y)}{\partial x}+c\frac{\partial\pi(x,y)}{\partial y}=0,\quad (x,y)\in I\!R_+^2 \tag{2.2a}$$

$$BC(x=0):\frac{1}{2}\alpha^2\frac{\partial\pi(0,y)}{\partial x}+\sigma_0^2\frac{\partial\pi(0,y)}{\partial y}+c\pi(0,y)=0 \tag{2.2b}$$

$$BC(y=0):\sigma_0^2\frac{\partial\pi(x,0)}{\partial x}+\frac{1}{2}\alpha^2\frac{\partial\pi(x,0)}{\partial y}+c\pi(x,0)=0. \tag{2.2c}$$

We now scale the co–ordinates according to the transformation $T_0 : (x,y) \to (x_1,y_1)$, so that so that $(x_1,y_1) = (ax,ay)$ where $a = \frac{2c}{\alpha^2}$, and we set $\beta = \frac{\sigma_0^2}{\alpha^2}$ in what follows. Denoting $\pi(\frac{x_1}{a},\frac{y_1}{a})$ by $\pi_a(x_1,y_1)$, (2.2a)–(2.2c) can then be written as

$$\frac{\partial^2 \pi_a(x_1, y_1)}{\partial x_1^2} + 2\beta \frac{\partial^2 \pi_a(x_1, y_1)}{\partial x_1 \partial y_1} + \frac{\partial^2 \pi_a(x_1, y_1)}{\partial y_1^2}$$

$$+ \frac{\partial \pi_a(x_1, y_1)}{\partial x_1} + \frac{\partial \pi_a(x_1, y_1)}{\partial y_1} = 0, \quad (x_1, y_1) \in I\!R_+^2 \tag{2.3a}$$

$$BC(x_1 = 0) : \frac{\partial \pi_a(0, y_1)}{\partial x_1} + 2\beta \frac{\partial \pi_a(0, y_1)}{\partial y_1} + \pi_a(0, y_1) = 0 \tag{2.3b}$$

$$BC(y_1 = 0) : 2\beta \frac{\partial \pi_a(x_1, 0)}{\partial x_1} + \frac{\partial \pi_a(x_1, 0)}{\partial y_1} + \pi_a(x_1, 0) = 0. \tag{2.3c}$$

Since

$$\beta = \frac{\sigma_0^2}{\alpha^2} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}, \tag{2.4}$$

the parameter β is constrained to lie in the set $[0, 1]$, and we shall therefore seek solutions to (2.3a)–(2.3c) with β constrained to lie in this set.

5.3 The solution in polar co–ordinates

In this section we recast our basic equation (2.3a)–(2.3c) into the form $\nabla^2\phi = \phi$, where ∇^2 is the two–dimensional Laplacian in polar form. This is accomplished by several transformations as shown below. The development of this section is inspired by Foschini [16] and Harrison [21].

The transformation is achieved in the following five steps:

(1): We start with an exponential substitution to eliminate the drift terms. Let us introduce a new function π_1 defined by

$$\pi_1(x_1, y_1) = \pi_a(x_1, y_1)e^{-b(x_1+y_1)}, \quad (x_1, y_1) \in I\!R^2_+ \tag{3.1}$$

where $b = -\frac{1}{2(1+\beta)}$. The PDE can then be re–written as

$$\frac{\partial^2\pi_1}{\partial x_1^2} + 2\beta\frac{\partial^2\pi_1}{\partial x_1 \partial y_1} + \frac{\partial^2\pi_1}{\partial y_1^2} = \frac{\pi_1}{2(1+\beta)}, \quad (x_1, y_1) \in I\!R^2_+ \tag{3.2a}$$

$$BC(x_1 = 0) : \frac{\partial\pi_1}{\partial x_1} + 2\beta\frac{\partial\pi_1}{\partial y_1} + \frac{\pi_1}{2(1+\beta)} = 0 \tag{3.2b}$$

$$BC(y_1 = 0) : 2\beta\frac{\partial\pi_1}{\partial x_1} + \frac{\partial\pi_1}{\partial y_1} + \frac{\pi_1}{2(1+\beta)} = 0. \tag{3.2c}$$

(2): The term with the mixed derivatives can be removed by the orthogonal transformation $T_1 : (x_1, y_1) \to (x_2, y_2)$, defined by

$$\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \quad (x_1, y_1) \in I\!R^2_+. \tag{3.3}$$

This transformation maps the quadrant $I\!R^2_+$ in the (x_1, y_1)–plane into a region $I\!R^2_0$ in the first and fourth quadrants of the (x_2, y_2)–plane that is bounded by the lines $x_2 = y_2$ and $x_2 = -y_2$.

Denoting $\pi_1(T_1(x_1, y_1))$ by $\pi_2(x_2, y_2)$, we obtain the following PDE.

$$(1+\beta)\frac{\partial^2\pi_2}{\partial x_2^2} + (1-\beta)\frac{\partial^2\pi_2}{\partial y_2^2} = \frac{\pi_2}{2(1+\beta)}, \quad (x_2, y_2) \in I\!R^2_0 \tag{3.4a}$$

$$BC(x_2 = y_2) : (2\beta+1)\frac{\partial \pi_2}{\partial x_2} + (2\beta-1)\frac{\partial \pi_2}{\partial y_2} + \frac{\pi_2}{\sqrt{2}(1+\beta)} = 0 \qquad (3.4b)$$

$$BC(x_2 = -y_2) : (2\beta+1)\frac{\partial \pi_2}{\partial x_2} - (2\beta-1)\frac{\partial \pi_2}{\partial y_2} + \frac{\pi_2}{\sqrt{2}(1+\beta)} = 0. \qquad (3.4c)$$

(3): The next transformation $T_2 : (x_2, y_2) \to (x_3, y_3)$ is defined by

$$\begin{pmatrix} x_3 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1+\beta}} & 0 \\ 0 & \frac{1}{\sqrt{1-\beta}} \end{pmatrix} \begin{pmatrix} x_2 \\ y_2 \end{pmatrix}, \quad (x_2, y_2) \in I\!R_0^2 \qquad (3.5)$$

This transformation maps the region $I\!R_0^2$ in the (x_2, y_2)–plane into a region $I\!R_\beta^2$ in the first and fourth quadrants of the (x_3, y_3)–plane that is bounded by the lines $y_3 = \sqrt{\frac{1+\beta}{1-\beta}}x_3$ and $y_3 = -\sqrt{\frac{1+\beta}{1-\beta}}x_3$. Denoting $\pi_2(T_2(x_2, y_2))$ by $\pi_3(x_3, y_3)$, we obtain the PDE

$$\nabla^2 \pi_3 = \frac{\pi_3}{2(1+\beta)}, \quad (x_3, y_3) \in I\!R_\beta^2 \qquad (3.6a)$$

$$BC(y_3 = \sqrt{\frac{1+\beta}{1-\beta}}x_3) : \frac{(2\beta+1)}{\sqrt{1+\beta}}\frac{\partial \pi_3}{\partial x_3} + \frac{(2\beta-1)}{\sqrt{1-\beta}}\frac{\partial \pi_3}{\partial y_3} + \frac{\pi_3}{\sqrt{2}(1+\beta)} = 0 \qquad (3.6b)$$

$$BC(y_3 = -\sqrt{\frac{1+\beta}{1-\beta}}x_3) : \frac{(2\beta+1)}{\sqrt{1+\beta}}\frac{\partial \pi_3}{\partial x_3} - \frac{(2\beta-1)}{\sqrt{1-\beta}}\frac{\partial \pi_3}{\partial y_3} + \frac{\pi_3}{\sqrt{2}(1+\beta)} = 0. \qquad (3.6c)$$

(4): The next transformation $T_3 : (x_3, y_3) \to (x_4, y_4)$ is given by

$$\begin{pmatrix} x_4 \\ y_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2(1+\beta)}} & 0 \\ 0 & \frac{1}{\sqrt{2(1+\beta)}} \end{pmatrix} \begin{pmatrix} x_3 \\ y_3 \end{pmatrix}, \quad (x_3, y_3) \in I\!R_\beta^2 \qquad (3.7)$$

Denoting $\pi_3(T_3(x_3, y_3))$ by $\pi_4(x_4, y_4)$, we obtain

$$\nabla^2 \pi_4 = \pi_4, \quad (x_4, y_4) \in I\!R_\beta^2 \qquad (3.8a)$$

$$BC(y_4 = \sqrt{\frac{1+\beta}{1-\beta}}x_4) : (2\beta+1)\frac{\partial \pi_4}{\partial x_4} + \sqrt{\frac{1+\beta}{1-\beta}}(2\beta-1)\frac{\partial \pi_4}{\partial y_4} + \pi_4 = 0 \tag{3.8b}$$

$$BC(y_4 = -\sqrt{\frac{1+\beta}{1-\beta}}x_4) : (2\beta+1)\frac{\partial \pi_4}{\partial x_4} - \sqrt{\frac{1+\beta}{1-\beta}}(2\beta-1)\frac{\partial \pi_4}{\partial y_4} + \pi_4 = 0. \tag{3.8c}$$

(5): Finally we recast this equation into polar co-ordinates with the the transformation $T_4 : (x_4, y_4) \rightarrow (r, \theta)$ given by

$$x_4 = r\cos\theta \quad \text{and} \quad y_4 = r\sin\theta$$

We retain the notation $I\!R^2_\beta$ for the region in the (r, θ) plane that is bounded by the straight lines $\theta = \tan^{-1}\sqrt{\frac{1+\beta}{1-\beta}}$ and $\theta = -\tan^{-1}\sqrt{\frac{1+\beta}{1-\beta}}$. Denoting $\pi_4(T(x_4, y_4)) = \phi(r, \theta)$, we finally obtain

$$\frac{\partial^2\phi}{\partial r^2} + \frac{1}{r}\frac{\partial\phi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\phi}{\partial\theta^2} = \phi, \quad (r,\theta) \in I\!R^2_\beta \tag{3.9a}$$

$$BC(\theta = \tan^{-1}\sqrt{\frac{1+\beta}{1-\beta}}) : \frac{2\beta}{\sqrt{1-\beta}}\frac{\partial\phi}{\partial r} - 2\frac{\sqrt{1+\beta}}{r}\frac{\partial\phi}{\partial\theta} + \sqrt{2}\phi = 0 \tag{3.9b}$$

$$BC(\theta = -\tan^{-1}\sqrt{\frac{1+\beta}{1-\beta}}) : \frac{2\beta}{\sqrt{1-\beta}}\frac{\partial\phi}{\partial r} + 2\frac{\sqrt{1+\beta}}{r}\frac{\partial\phi}{\partial\theta} + \sqrt{2}\phi = 0. \tag{3.9c}$$

We shall find a solution to this equation in the case when $\beta = \frac{1}{2}$. In this case the equation becomes,

$$\nabla^2\phi = \phi, \quad (r,\theta) \in I\!R^2_{0.5} \tag{3.10a}$$

$$BC(\theta = \frac{\pi}{3}) : \frac{\partial\phi}{\partial r} - \frac{\sqrt{3}}{r}\frac{\partial\phi}{\partial\theta} + \phi = 0 \tag{3.10b}$$

$$BC(\theta = -\frac{\pi}{3}) : \frac{\partial\phi}{\partial r} + \frac{\sqrt{3}}{r}\frac{\partial\phi}{\partial\theta} + \phi = 0. \tag{3.10c}$$

The case $\beta = \frac{1}{2}$ is of importance because it corresponds to the situation when $\sigma_0 = \sigma$, i.e., when the inter-arrival and service distributions have the same limiting variance. This will always be the case if the service and inter–arrival time

distributions are taken from the same family. For example consider the case of a fork–join queue with exponential inter–arrival and service distributions with rate $\frac{1}{\lambda}$ an $\frac{1}{\mu}$ respectively. Then as $\lambda \uparrow \mu$ in heavy traffic,

$$\sigma_0^2 = \frac{1}{\lambda^2} \to \frac{1}{\mu^2} = \sigma^2.$$

Equations similar to (3.10) have been encountered earlier by Harrison [21] and Foschini [16] in the context of the diffusion limit for queues in tandem. Guided by their work, we try a solution of the form

$$\phi(r,\theta) = \frac{1}{\sqrt{r}} e^{-r} \cos(\frac{\theta}{2}), \quad (r,\theta) \in I\!R^2_{0.5}. \tag{3.11}$$

Note that it satisfies the PDE as well as the boundary conditions (3.10a)–(3.10c), as can be verified by a direct substitution.

Our next objective is to obtain an expression for the density in terms of (x,y). Note that the tranformation $T : (r,\theta) \to (x,y)$, which is a composition of the transformations $T = T_0^{-1}T_1^{-1}T_2^{-1}T_3^{-1}T_4^{-1}$, can be written as

$$\begin{aligned} ax &= (1+\beta)r\cos\theta - \sqrt{1-\beta^2}r\sin\theta \\ ay &= (1+\beta)r\cos\theta + \sqrt{1-\beta^2}r\sin\theta, \quad (r,\theta) \in I\!R^2_{0.5}. \end{aligned} \tag{3.12}$$

If we undo the transformation which corresponded to a multiplication by $e^{-\frac{1}{3}(x_1+y_1)}$, we obtain the function $\psi(r,\theta)$, where

$$\psi(r,\theta) = \frac{1}{\sqrt{r}} e^{-r(1+\cos\theta)} \cos(\frac{\theta}{2}), \quad (r,\theta) \in I\!R^2_{0.5}. \tag{3.13}$$

Letting $\psi(T(r,\theta)) = \varphi(x,y)$, the final solution is of the form $K\varphi(x,y)$ where the constant K is chosen so that

$$\int_0^\infty \int_0^\infty K\varphi(x,y)dxdy = 1. \tag{3.14}$$

We shall evaluate this integral on the (r,θ) plane where the calculations are much easier. It can easily be checked that the Jacobian J for the transformation (3.12)

is given by $J = \frac{2r}{a^2}(1+\beta)\sqrt{1-\beta^2} = \frac{3r}{2a^2}\sqrt{3}$. The integral (3.14) then transforms to

$$3\sqrt{3}K \int_{\theta=0}^{\frac{\pi}{3}} \int_{r=0}^{\infty} \sqrt{r} e^{-r(1+\cos\theta)} \cos(\frac{\theta}{2}) dr d\theta = a^2. \tag{3.15}$$

Making the substitution $\gamma(\theta) = 1 + \cos\theta$, it follows that

$$\begin{aligned}
\int_0^{\infty} \sqrt{r} e^{-\gamma(\theta)r} dr &= \gamma^{-\frac{3}{2}}(\theta) \int_0^{\infty} \sqrt{u} e^{-u} du \\
&= \gamma^{-\frac{3}{2}}(\theta)\Gamma(\frac{3}{2}) \\
&= \frac{1}{4}\sqrt{\frac{\pi}{2}} \cos^{-3}\frac{\theta}{2}
\end{aligned}$$

Substituting this back into (3.15), it follows that

$$\begin{aligned}
a^2 &= \frac{3\sqrt{3}}{4}\sqrt{\frac{\pi}{2}} K \int_{\theta=0}^{\frac{\pi}{3}} \sec^{-2}(\frac{\theta}{2}) d\theta \\
&= \frac{3\sqrt{3}}{4}\sqrt{\frac{\pi}{2}} K [\tan(\frac{\theta}{2})]_0^{\frac{\pi}{3}} \\
&= \frac{3K}{2}\sqrt{\frac{\pi}{2}}
\end{aligned}$$

so that

$$K = \frac{2a^2}{3}\sqrt{\frac{2}{\pi}}$$

Hence the final solution is

$$\pi(x,y) = \frac{2a^2}{3}\sqrt{\frac{2}{\pi}}\varphi(x,y), \quad (x,y) \in I\!R_+^2. \tag{3.16}$$

Making use of the fact that

$$r = \frac{2a}{3}\sqrt{x^2 - xy + y^2}, \quad \cos\theta = \frac{x+y}{2\sqrt{x^2 - xy + y^2}} \tag{3.17}$$

and substituting for r and $\cos\theta$ in (3.13), we finally conclude to the following result.

Theorem 5.3.1 *The stationary density $\pi(x,y)$ of the diffusion for the waiting times in a symmetric two dimensional fork–join queue in heavy traffic, which satisfies $\sigma = \sigma_0$ is given by*

$$\pi(x,y) = a\sqrt{\frac{a}{3}}\frac{\sqrt{2\sqrt{x^2-xy+y^2}+x+y}e^{-\frac{2a}{3}\sqrt{x^2-xy+y^2}-\frac{a}{3}(x+y)}}{\sqrt{x^2-xy+y^2}},$$

$$(x,y) \in I\!R_+^2 \quad (3.18)$$

where $a = \frac{c}{\sigma^2}$.

Knessel [41] has also considered the problem of solving an equation similar to (2.3) from the point of view of the theory of singular perturbations, and obtained an expression for $\pi(x,y)$ in the case when x and y are very large. As expected, our solution (3.18) agrees with his for large x, y.

5.3.1 Calculations of the moments of the end–to–end delay

In this sub–section our objective is to obtain some information regarding the stationary density and the moments of the diffusion for the end–to–end delay of the fork–join queue. Our first objective is to find an expression for the density of the equilibrium response time κ_∞, where

$$\kappa_\infty = \max\{\eta_\infty^1, \eta_\infty^2\}.$$

It is clear that

$$F(z) = I\!P(\kappa_\infty \le z) = I\!P(\eta_\infty^1 \le z, \eta_\infty^2 \le z) = \int_0^z \int_0^z \pi(x,y)dxdy, \quad z \ge 0. \quad (3.19)$$

We make a change of co–ordinates from (x,y) to (r,θ), by using the transformation (3.13). The square $[0,z]\times[0,z]$ in the (x,y) plane maps into the rhombus

with sides of length $\frac{2az}{3}$ and vertices at $(0,0),(\frac{2az}{3},\frac{\pi}{3}),(\frac{2az}{3},0)$ and $(\frac{2az}{3},-\frac{\pi}{3})$ in the (r,θ) plane. Using the law of sines for triangles, it is clear that the limits of integration for r are from $r=0$ to $r=\frac{2az}{3}\frac{\sin\frac{\pi}{3}}{\sin(\frac{2\pi}{3}-\theta)}=\frac{az}{\sqrt{3}}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}$. Hence

$$F(z)=2\sqrt{\frac{6}{\pi}}\int_{\theta=0}^{\frac{\pi}{3}}\int_{r=0}^{\frac{az}{\sqrt{3}}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}\sqrt{r}e^{-r(1+\cos\theta)}\cos\frac{\theta}{2}drd\theta,\quad z\geq 0. \tag{3.20}$$

As before, let $\gamma(\theta)=1+\cos\theta$, so that with the substitution $u=\gamma(\theta)r$ we obtain

$$\int_{r=0}^{\frac{az}{\sqrt{3}}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}\sqrt{r}e^{-\gamma(\theta)r}dr=\gamma^{-\frac{3}{2}}(\theta)\int_{u=0}^{\frac{az\gamma(\theta)}{\sqrt{3}}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}\sqrt{u}e^{-u}du.$$

The resultant integral above is known as the incomplete Gamma function and occurs frequently in analysis. It is well known that

$$\int_0^x\sqrt{t}e^{-t}dt+\int_x^\infty\sqrt{t}e^{-t}dt=\Gamma(\frac{3}{2})=\frac{\sqrt{\pi}}{2}. \tag{3.21}$$

Hence (3.20) can be re-written as

$$\begin{aligned}F(z)&=2\sqrt{\frac{6}{\pi}}\int_{\theta=0}^{\frac{\pi}{3}}\cos\frac{\theta}{2}\gamma^{-\frac{3}{2}}(\theta)[\frac{\sqrt{\pi}}{2}-\int_{\frac{az\gamma(\theta)}{\sqrt{3}}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}^{\infty}\sqrt{u}e^{-u}du]d\theta\\&=1-2\sqrt{\frac{6}{\pi}}\int_{\theta=0}^{\frac{\pi}{3}}\int_{u=\frac{az\gamma(\theta)}{\sqrt{3}}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}^{\infty}\gamma^{-\frac{3}{2}}(\theta)\cos\frac{\theta}{2}\sqrt{u}e^{-u}dud\theta\\&=1-\sqrt{\frac{3}{\pi}}\int_{\theta=0}^{\frac{\pi}{3}}\int_{u=\frac{2az}{\sqrt{3}}\cos^2\frac{\theta}{2}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}^{\infty}\frac{\sqrt{u}e^{-u}}{\cos^2\frac{\theta}{2}}dud\theta,\quad z\geq 0.\end{aligned} \tag{3.22}$$

We now use the fact that

$$I\!E\kappa_\infty=\int_{z=0}^{\infty}[1-F(z)]dz$$

so that

$$I\!E\kappa_\infty=\sqrt{\frac{3}{\pi}}\int_{z=0}^{\infty}\int_{\theta=0}^{\frac{\pi}{3}}\int_{u=\frac{2az}{\sqrt{3}}\cos^2\frac{\theta}{2}\frac{1}{\sin(\frac{2\pi}{3}-\theta)}}^{\infty}\frac{\sqrt{u}e^{-u}}{\cos^2\frac{\theta}{2}}dud\theta dz. \tag{3.23}$$

Interchanging the limits of integration, we obtain

$$\mathbb{E}\kappa_\infty = \sqrt{\frac{3}{\pi}} \int_{\theta=0}^{\frac{\pi}{3}} \int_{u=0}^{\infty} \int_{z=0}^{\frac{\sqrt{3}u}{2a} \frac{1}{\cos^2 \frac{\theta}{2}} \sin(\frac{2\pi}{3}-\theta)} \frac{\sqrt{u}e^{-u}}{\cos^2 \frac{\theta}{2}} dz du d\theta$$

$$= \frac{3\Gamma(\frac{5}{2})}{4a} \sqrt{\frac{3}{\pi}} \frac{1}{\sin \frac{\pi}{3}} \int_{\theta=0}^{\frac{\pi}{3}} \frac{\sin(\frac{2\pi}{3}-\theta)}{\cos^4 \frac{\theta}{2}} d\theta \tag{3.24}$$

Using the fact that $\Gamma(\frac{5}{2}) = \frac{3}{4}\sqrt{\pi}$ and $\int_0^{\frac{\pi}{3}} \frac{\sin(\frac{2\pi}{3}-\theta)}{\cos^4 \frac{\theta}{2}} d\theta = \frac{11}{9}$, we finally obtain

$$\mathbb{E}\kappa_\infty = \frac{11}{8a} = \frac{11}{8} \frac{\sigma^2}{c}. \tag{3.25}$$

We can check the correctness of (3.25), by noting that in the special case when the arrivals are Poisson and the service times are identically exponentially distributed, $\mathbb{E}\kappa_\infty$ can also be calculated using the results of Flatto and Hahn [14]. This was done by Nelson and Tantawi [52], and their formula for $E\kappa_\infty$ exactly matches ours in this special case. This result is also consistent with the bounds for $\mathbb{E}\kappa_\infty$ obtained in Chapter 3.

Using (3.22), the density function $f(z)$ of the response time is given by

$$f(z) = \frac{dF(z)}{dz} = \sqrt{\frac{3}{\pi}} \int_{\theta=0}^{\frac{\pi}{3}} \frac{1}{\cos \frac{\theta}{2}} \delta(\theta) \sqrt{\delta(\theta) z} e^{-\delta(\theta) z} d\theta, \quad z \geq 0 \tag{3.26}$$

where $\delta(\theta) = \frac{2a}{\sqrt{3}} \frac{\cos^2 \frac{\theta}{2}}{\sin(\frac{2\pi}{3}-\theta)}$.

Using this expression it is possible to obtain a formula for the n^{th} moment of the response time, as is done next. Note that

$$\mathbb{E}\kappa_\infty^n = \int_{z=0}^{\infty} z^n f(z) dz$$

$$= \sqrt{\frac{3}{\pi}} \int_{\theta=0}^{\frac{\pi}{3}} \delta(\theta) \sqrt{\delta(\theta)} \frac{d\theta}{\cos^2 \frac{\theta}{2}} \int_{z=0}^{\infty} z^{n+\frac{1}{2}} e^{-\delta(\theta) z} dz$$

$$= \Gamma(n+\frac{3}{2}) \sqrt{\frac{3}{\pi}} \int_{\theta=0}^{\frac{\pi}{3}} \frac{1}{\cos^2 \frac{\theta}{2}} \frac{1}{\delta^n(\theta)} d\theta. \tag{3.27}$$

Substituting for $\delta(\theta)$ and a, we finally obtain the following result.

Proposition 5.3.1 *The n^{th} moment of the stationary density of the diffusion for the end-to-end delay in a symmetrical two dimensional fork-join queue in heavy traffic which satisfies $\sigma = \sigma_0$, is given by*

$$I\!E\kappa_\infty^n = \Gamma(n+\frac{3}{2})\sqrt{\frac{3}{\pi}}(\frac{\sqrt{3}}{2}\frac{\sigma^2}{c})^n P_n \tag{3.28}$$

where

$$\Gamma(n+\frac{3}{2}) = \frac{(2n+1)(2n-1)\ldots 3.1}{2^{n+1}}\sqrt{\pi} \tag{3.29}$$

and

$$P_n = \int_{\theta=0}^{\frac{\pi}{3}} \frac{\sin^n(\frac{2\pi}{3}-\theta)}{\cos^{2n+2}\frac{\theta}{2}}d\theta. \tag{3.30}$$

The reader may check that for the case $n = 1$, this formula agrees with the expression for $E\kappa_\infty$ derived earlier. The integral (3.30) is evaluated for some values of n in Section 5.5.

In the following corollary we obtain the heavy traffic limit for the fork–join system. The proof is a direct consequence of Theorem 2.2.2 and Proposition 5.3.1 and is omitted.

Corollary 5.3.1 *Consider a symmetric fork–join queue governed by an arrival process with mean $\frac{1}{\lambda}$ and variance $\sigma_0^2(\lambda)$, and a service time distribution with mean $\frac{1}{\mu}$ and variance σ^2. Further assume that $\lim_{\lambda\uparrow\mu}\sigma_0^2(\lambda) = \sigma^2$. The heavy traffic limit for the n^{th} moment of the end–to–end delay of this queue $\overline{T}^n(\lambda)$, is given by*

$$\lim_{\lambda\uparrow\mu}(\mu-\lambda)^n\overline{T}^{(n)}(\lambda) = \Gamma(n+\frac{3}{2})\sqrt{\frac{3}{\pi}}(\frac{\sqrt{3}}{2}\sigma^2\mu^2)^n P_n \tag{3.31}$$

where $\Gamma(n+\frac{3}{2})$ and P_n are defined in (3.29)-(3.30).

We now provide a formula for the normalized correlation between the delay

processes of the two queues in heavy traffic. This is given by

$$I\!E(\eta_\infty^1 \eta_\infty^2) = \int_0^\infty \int_0^\infty xy\pi(x,y)dxdy.$$

Making the usual change of co–ordinates from (x,y) to (r,θ), we obtain after some calculations that

$$I\!E(\eta_\infty^1 \eta_\infty^2) = \frac{11}{8}(\frac{\sigma^2}{c})^2. \tag{3.32}$$

Note that the two queues by themselves behave like $GI/GI/1$ queues, so that

$$I\!P(\eta_\infty^k \leq x) = 1 - e^{-\frac{c}{\sigma^2}x}, \quad k = 1,2$$

and

$$I\!E\eta_\infty^k = \frac{\sigma^2}{c}, \quad k = 1,2.$$

It follows that

$$Cov(\eta_\infty^1 \eta_\infty^2) = I\!E(\eta_\infty^1 \eta_\infty^2) - I\!E\eta_\infty^1 I\!E\eta_\infty^2 = \frac{3}{8}(\frac{\sigma^2}{c})^2. \tag{3.33}$$

This implies the following result.

Proposition 5.3.1 *Consider a symmetric K–dimensional fork–join queue governed by an arrival process with mean $\frac{1}{\lambda}$ and variance $\sigma_0^2(\lambda)$, and a service time distribution with mean $\frac{1}{\mu}$ and variance σ^2. Further assume that $\lim_{\lambda\uparrow\mu} \sigma_0^2(\lambda) = \sigma^2$. If $(W^1(\lambda),\ldots,W^K(\lambda))$ represents the steady state vector of queueing delays in the system, then*

$$\lim_{\lambda\uparrow\mu} Corr(W^i(\lambda)W^j(\lambda)) = \frac{3}{8}, \quad 1 \leq i,j \leq K, \quad i \neq j. \tag{3.34}$$

It is a remarkable coincidence that the asymptotic correlation $\frac{3}{8}$, almost equals the constant $\frac{11}{4}$ that was obtained by Nelson and Tantawi (see Section 6.4.2), as part of their heuristic approximation. We also note that the correlation betwen two

queues in the system is crucially dependent on the parameter β. We just showed that for the case $\beta = \frac{1}{2}$, the coefficient of variation is given by $\frac{3}{8}$. For the case $\beta = 0$, it is given by zero, since in this case the two queues are independent, while in the case $\beta = 1$ it is given by one, since in this case both queues are perfectly synchronized with one another. Hence we observe that as the service times become more deterministic, i.e., as β increases, the two queues become more correlated with one another.

5.4 The queue length processes: Symmetrical case

In this section we concern ourselves with the task of finding the statistics of the diffusion connected with the queue length processes of the fork–join queue and particularly the statistics of the diffusion due to the number in the join buffer.

We start by writing down the PDE with the appropriate boundary conditions which the stationary density satisfies.

$$\frac{1}{2}(\tau_0^2+\tau_1^2)\frac{\partial^2\pi_q(x,y)}{\partial x^2}+\tau_0^2\frac{\partial^2\pi_q(x,y)}{\partial x\partial y}+\frac{1}{2}(\tau_0^2+\tau_2^2)\frac{\partial^2\pi_q(x,y)}{\partial y^2}$$

$$-d_1\frac{\partial\pi_q(x,y)}{\partial x}-d_2\frac{\partial\pi_q(x,y)}{\partial y}=0,\quad (x,y)\in I\!R_+^2 \tag{4.1a}$$

$$BC(x=0):\frac{1}{2}(\tau_0^2+\tau_1^2)\frac{\partial\pi_q(0,y)}{\partial x}+\tau_0^2\frac{\partial\pi_q(0,y)}{\partial y}-d_1\pi(0,y)=0 \tag{4.1b}$$

$$BC(y=0):\tau_0^2\frac{\partial\pi_q(x,0)}{\partial x}+\frac{1}{2}(\tau_0^2+\tau_2^2)\frac{\partial\pi_q(x,0)}{\partial y}-d_2\pi_q(x,0)=0. \tag{4.1c}$$

We further make the assumption that the two queues are identical with $\tau_1=\tau_2=\tau$ and $d_1=d_2=d$ and we set $\gamma^2=\tau_0^2+\tau^2$ in what follows. The equilibrium equations then simplify to the following.

$$\frac{1}{2}\gamma^2\frac{\partial^2\pi_q(x,y)}{\partial x^2}+\tau_0^2\frac{\partial^2\pi_q(x,y)}{\partial x\partial y}+\frac{1}{2}\gamma^2\frac{\partial^2\pi_q(x,y)}{\partial y^2}$$

$$-d\frac{\partial\pi_q(x,y)}{\partial x}-d\frac{\partial\pi_q(x,y)}{\partial y}=0,\quad (x,y)\in I\!R_+^2 \tag{4.2a}$$

$$BC(x=0):\frac{1}{2}\gamma^2\frac{\partial\pi_q((0,y)}{\partial x}+\tau_0^2\frac{\partial\pi_q(0,y)}{\partial y}-d\pi_q(0,y)=0 \tag{4.2b}$$

$$BC(y=0):\tau_0^2\frac{\partial\pi_q(x,0)}{\partial x}+\frac{1}{2}\gamma^2\frac{\partial\pi_q(x,0)}{\partial y}-d\pi_q(x,0)=0. \tag{4.2c}$$

We now scale the co–ordinates so that $(x_1,y_1)=(bx,by)$ where $b=-\frac{2d}{\gamma^2}$, and set $\beta'=\frac{\tau_0^2}{\gamma^2}$ in what follows. Denoting $\pi_q(\frac{x_1}{b},\frac{y_1}{b})$ by $\pi_q^b(x_1,y_1)$, (4.2a)–(4.2c) can be written as

$$\frac{\partial^2 \pi_q^b(x_1,y_1)}{\partial x_1^2} + 2\beta' \frac{\partial^2 \pi_q^b(x_1,y_1)}{\partial x_1 \partial y_1} + \frac{\partial^2 \pi_q^b(x_1,y_1)}{\partial y_1^2}$$

$$+ \frac{\partial \pi_q^b(x_1,y_1)}{\partial x_1} + \frac{\partial \pi_q^b(x_1,y_1)}{\partial y_1} = 0, \quad (x_1,y_1) \in I\!R_+^2 \tag{4.3a}$$

$$BC(x_1 = 0) : \frac{\partial \pi_q^b((0,y_1)}{\partial x_1} + 2\beta' \frac{\partial \pi_q^b(0,y_1)}{\partial y_1} + \pi_q^b(0,y_1) = 0 \tag{4.3b}$$

$$BC(y_1 = 0) : 2\beta' \frac{\partial \pi_q^b(x_1,0)}{\partial x_1} + \frac{\partial \pi_q^b(x_1,0)}{\partial y_1} + \pi_q(x_1,0) = 0. \tag{4.3c}$$

Observe that the PDE for the queue length process has the same structure as the PDE for the queue delay processes. Hence, guided by our analysis in the last section, we shall write down the solution directly in the next section.

5.4.1 The solution to the PDE

Proceeding as in the last section, the following result is immediate.

Theorem 5.4.1 *The stationary density $\pi_q(x,y)$ of the diffusion for the queue lengths in a symmetric two dimensional fork–join queue in heavy traffic, which satisfies $\tau = \tau_0$ is given by*

$$\pi_q(x,y) = b\sqrt{\frac{b}{3}} \frac{\sqrt{2\sqrt{x^2-xy+y^2}+x+y}\, e^{-\frac{2b}{3}\sqrt{x^2-xy+y^2}-\frac{b}{3}(x+y)}}{\sqrt{x^2-xy+y^2}},$$

$$(x,y) \in I\!R_+^2 \tag{4.4}$$

where $b = -\frac{d}{\tau^2}$.

Our next objective is to find an expression for the expectation of the equilibrium number of customers in the join buffer. Let the RV N_∞ possess the stationary distribution of the diffusion for the number of customers in the join buffer of a two dimensional fork–join queue in heavy traffic. Note that

$$I\!E N_\infty = \int_{y=0}^{\infty} \int_{x=0}^{\infty} \mid x - y \mid \pi_q(x,y) dx dy \tag{4.5}$$

We make a change of variables according to the transformation

$$bx = (1+\beta)r\cos\theta - \sqrt{1-\beta^2}r\sin\theta$$
$$by = (1+\beta)r\cos\theta + \sqrt{1-\beta^2}r\sin\theta, \quad (r,\theta) \in I\!R^2_{0.5}$$

and evaluate (4.5) on the (r,θ) plane, to obtain

$$I\!EN_\infty = \frac{3}{4b} = -\frac{3}{4}\frac{\tau^2}{d}. \tag{4.6}$$

As we now demonstrate, this expression for $I\!EN_\infty$ along with a heavy traffic version of Little's law leads us to the expression for $I\!E\kappa_\infty$ in (3.25). Note that the two queues by themselves act as $GI/GI/1$ queues, so that $I\!P(Q^k_\infty \leq x) = 1 - \exp^{-bx}, k = 1,2$. Hence if n_∞ denotes the expectation for the stationary distribution of the diffusion for the total number of customers in the system in steady state, then

$$I\!En_\infty = \frac{2}{b} + \frac{3}{4b} = \frac{11}{4b}.$$

Next note that the total input rate into the system in heavy traffic is given by 2μ, so that by Little's law we have

$$\frac{I\!En_\infty}{2\mu} = \frac{11}{8b}$$

so that taking note of the fact that $\tau^2 = \mu^3\sigma^2$ and

$$\lim_{r\uparrow\infty}\sqrt{r}(a(r)-b) = \lim_{r\uparrow\infty}\sqrt{r}\frac{(\mu-\lambda(r))}{\mu^2}$$

we finally obtain that

$$I\!E\kappa_\infty = \frac{I\!En_\infty}{2\mu}\frac{1}{\mu^2} = \frac{11}{8a}.$$

We now present a formula for the n^{th} moment of the diffusion for the number of customers in the join buffer, i.e.,

$$I\!EN^n_\infty = \int_0^\infty \int_0^\infty \mid x-y \mid^n \pi_q(x,y)dxdy. \tag{4.7}$$

A straightforward evaluation of this integral yields the following result.

Proposition 5.4.1 *The n^{th} moment of the stationary density of the diffusion for the number of customers in the join buffer in a two dimensional fork–join queue in heavy traffic is given by*

$$I\!E N_\infty^n = \Gamma(n+\frac{3}{2})\sqrt{\frac{3}{\pi}}(-\frac{\sqrt{3}}{2}\frac{\tau^2}{d})^n R_n \tag{4.8}$$

where $\Gamma(n+\frac{3}{2})$ was defined in (3.29) and

$$R_n = \int_{\theta=0}^{\frac{\pi}{3}} \frac{\sin^n\theta}{\cos^{2n+2}\frac{\theta}{2}} d\theta. \tag{4.9}$$

The integral (4.9) is evaluated for some values of n in Section 5.5. The correlation between the normalized queue lengths in heavy traffic is given by

$$\begin{aligned} I\!E q_\infty^1 q_\infty^2 &= \int_0^\infty \int_0^\infty xy\pi_q(x,y)dxdy \\ &= \frac{11}{8}(-\frac{\tau^2}{d})^2. \end{aligned} \tag{4.10}$$

5.5 Tables for P_n and R_n

The co-efficients P_n and R_n defined in (3.30) and (4.8) have been calculated for $n = 1, \ldots, 4$, with the help of the symbolic computation language MACSYMA, and set down in the table below.

n	P_n	$\mathbb{E}\kappa_\infty^n$	R_n	$\mathbb{E}N_\infty^n$
1	$\frac{11}{9}$	$\frac{11}{8a}$	$\frac{2}{3}$	$\frac{3}{4b}$
2	$\frac{3}{5}\sqrt{3}$	$\frac{81}{32a^2}$	$\frac{8}{27}\sqrt{3}$	$\frac{5}{4b^2}$
3	$\frac{1759}{1260}$	$\frac{10.3}{a^3}$	$\frac{4}{9}$	$\frac{105}{32b^3}$
4	$\frac{59123}{68040}\sqrt{3}$	$\frac{43.3}{a^3}$	$\frac{32}{135}\sqrt{3}$	$\frac{189}{16b^3}$

CHAPTER VI

6.1 Introduction

In the last few chapters we developed heavy traffic limits for fork–join queues. These provided good approximations to the performance measures of the queue in the case when the utilization of the queue is close to unity. In this chapter we concentrate on light traffic limits for the fork–join queue. By combining heavy and light traffic results it is possible to obtain good approximations for the case of moderate traffic.

This chapter is organized as follows. In Section 6.2 we show that the response time of the fork–join queue is an admissible RV in the sense of Definition B1 (Appendix B), so that the light traffic theory may be applicable to it. Section 6.3 is devoted to light traffic approximations for the response time of a two–dimensional symmetric fork–join queue with Poisson arrivals and exponential service times. These approximations are developed for all moments of the response time.

So far we have developed approximations only for two–dimensional fork–join queues due to the fact that we were able to solve the basic PDE for the stationary distribution of the diffusion limit for the queue delay processes, in the case $K = 2$. Hence, even though light traffic limits are available for the case $K > 2$, our ignorance of the corresponding heavy traffic limits prevents us from giving approximations for this case. However in Section 6.4 we make a crucial observation which enables us to obtain formulae for the heavy traffic limit even for the case $K > 2$. In particular we observe that, in the cases $K = 1, 2$ when the inter–arrival and service times are exponential, the first derivative of the expected value of the average response time in light traffic is equal to its heavy traffic limit. Postulating a similar behavior for the case $K > 2$ we recover the heavy traffic limits for the average response time for the general case. This conjecture is borne out by

extremely good agreement with simulation results, as well as with the so–called scaling approximation of Nelson and Tantawi.

In Section 6.5 we obtain approximations for the average response time for a K–dimensional fork–join queue with Poisson arrivals and Erlangian service times. In order to obtain the heavy traffic approximations for this system we use the conjecture that the first derivative of the average response time in light traffic is equal to the heavy traffic limit, as long as the arrivals are Poisson. Good agreement with experimental results is observed. Section 6.6 is devoted to light traffic approximations for all the moments of the response time of a two–dimensional symmetric fork–join queue with Erlang type arrivals and service distributions. A similar conjecture to that in Sections 6.4 and 6.5 is made in Section 6.7 to obtain heavy traffic limits for the first moment of the response time for general K, in the case when the inter–arrival and service times have the same distribution (not necessarily exponential). Again we are guided by the exact results available for the case $K = 2$ in making this conjecture. Using it, we give approximations for general K for the average response time, when the inter–arrival and service are Erlang distributed. These approximations also agree extremely well with simulation results.

In Section 6.8 we give a formula for the heavy traffic limit for K–dimensional fork–join queues with general inter–arrival and service times. The approximations obtained agree very well with experimental results. In particular we give approximations for the average response time of a K–dimensional fork–join queue with second order Erlangian inter–arrival time and exponential service times. We observe that this class of approximations does not work well for the case when the service time distributions have large coefficients of variation. In particular we present simulation results for the case of Poisson arrivals and hyper–exponential service times with coefficient of variation equal to ten. Lastly, in Section 6.9 we present an heuristic simulation based procedure for obtaining approximations for the response times of general acyclic fork–join networks. These approximations are also shown to agree well with simulations.

6.2 Admissibility

In this section our objective is to prove the admissibility of the average response time measure for general acyclic fork–join networks. Consider the following sample space $(\Omega, I\!F)$, where Ω is the set of infinite sequences $\{(\tau_n, v_n^1, \ldots, v_n^B)\}_0^\infty$. Here τ_n has the interpretation of the arrival time of the n^{th} batch, while $v_n^j, 1 \leq j \leq B$ has the interpretation of the service time of the customer that is sent to queue j. We introduce a measure $I\!P_\lambda$ on $(\Omega, I\!F)$ such that the arrival process under this measure is a Poisson process with parameter $\lambda > 0$. For each ω in Ω we add a tagged batch which arrives at time zero and whose service times $\hat{v}^j, 1 \leq j \leq B$, are independent of $\{v_n^j\}_0^\infty, 1 \leq j \leq B$, but have the same distribution. In order to do so, we define an augmented probability space $(\Omega', I\!F', Q_\lambda)$, such that for each ω' in Ω', we have that $\omega' = (\omega, (\hat{v}^1, \ldots, \hat{v}^B))$, where ω is an element of Ω. Let

$$T = \text{response time of batch entering at } t = 0 \tag{2.1}$$

and set

$$\psi^{(n)} = I\!E_{Q_\lambda}[T^n \mid I\!F]. \qquad n = 1, 2 \ldots (2.2)$$

The n^{th} moment $\overline{T}^{(n)}(\lambda)$ of the response time is then given by the formula

$$\overline{T}^{(n)}(\lambda) = I\!E_{Q_\lambda}[T^n] = \int \psi^{(n)} dQ_\lambda. \qquad n = 1, 2 \ldots (2.3)$$

We define $\overline{\psi}^{(n)}(\emptyset)$ and $\overline{\psi}^{(n)}(\{t\})$, t in $I\!R$ by

$$\overline{\psi}^{(n)}(\emptyset) = I\!E_{Q_\lambda}[\psi^{(n)} \mid \text{no arrivals }] \qquad n = 1, 2 \ldots (2.4)$$

and

$$\overline{\psi}^{(n)}(\{t\}) = I\!E_{Q_\lambda}[\psi^{(n)} \mid \text{arrival at time } t]. \qquad n = 1, 2 \ldots (2.5)$$

We conclude from Theorem B1 that

$$\overline{T}^{(n)}(0) := \lim_{\lambda \downarrow 0} \overline{T}^{(n)}(\lambda) = \overline{\psi}^{(n)}(\emptyset) \qquad n = 1, 2 \ldots (2.6)$$

$$\frac{d\overline{T}^{(n)}(0)}{d\lambda} := \lim_{\lambda \downarrow 0} \frac{d\overline{T}^{(n)}(\lambda)}{d\lambda} = \int_{-\infty}^{\infty} (\overline{\psi}^{(n)}(\{t\}) - \overline{\psi}^{(n)}(\emptyset))dt \qquad n = 1, 2 \ldots (2.7)$$

provided we can show that the RV $\psi^{(n)}$ is admissible. In order to do so, let

$$M_j(\theta) = I\!E[e^{\theta v_n^j}], \quad 1 \leq j \leq B, \quad \theta \in I\!R \qquad n = 1, 2 \ldots (2.8)$$

and introduce assumption (**VIa**), where

(VIa): There exists $\theta^* > 0$ such that

$$\prod_{j=1}^{B} M_j(\theta) < \infty, \quad \theta < \theta^*$$

Theorem 6.2.1 *If assumption* **(VIa)** *is satisfied, then* $\psi^{(n)}$ *as defined in (2.2) is admissible.*

Proof. We introduce an admissible $M/GI/1$ queue (defined on the same probability space as the network) which 'upper bounds' the fork–join network in the sense that as long as there is work remaining in the network, it works at least as fast as the $M/GI/1$ queue. The arrival and service sequences in the $M/GI/1$ queue are given by $\{u_n\}_0^\infty$ and $\{\sum_{j=1}^{B} v_n^j\}_0^\infty$ respectively. According to Theorem B2 (Appendix B), the $M/GI/1$ queue is admissible provided assumption **(VIa)** is satisfied. Now proceeding as in Theorem B3 (Appendix B), it can easily be shown that the fork–join network is admissible under assumption **(VIa)**.

∎

6.3 Markovian symmetric fork–join queue: The case $K = 2$

Consider a symmetric two–dimensional fork–join queue subject to Poisson arrivals with rate λ and exponential service times with rate μ. In the present section we develop light traffic estimates for the response time statistics of this queue, and combine it with the heavy traffic estimates of Chapter 5 to yield an estimate that is also valid for moderate traffic. The approximations of this section are zero order approximations in the sense that we shall only consider the limiting value of the response time as λ approaches zero (for light traffic approximations) and as λ approaches μ (for heavy traffic approximations). The approximations that we develop in this section hold for all moments of the response time.

We now proceed to calculate $\overline{T}^{(n)}(0)$ using (2.6). Since the batch arriving at $t = 0$ does not experience interference from any other customers, its queueing delay will be zero, and therefore its response time will simply be the maximum of two identically distributed exponential RVs. With F denote the distribution of the maximum of two exponential RVs with rate μ, we have

$$F(z) = (1 - e^{-\mu z})^2 = 1 - 2e^{-\mu z} + e^{-2\mu z}, \quad z \geq 0 \tag{3.1}$$

so that the corresponding density function f is given by

$$f(z) = 2\mu(e^{-\mu z} - e^{-2\mu z}), \quad z \geq 0. \tag{3.2}$$

Hence

$$\begin{aligned} \overline{T}^{(n)}(0) &= \int_0^\infty z^n f(z) dz \\ &= \frac{2\Gamma(n+1)}{\mu^n}[1 - \frac{1}{2^{n+1}}]. \qquad n = 1, 2 \ldots \end{aligned} \tag{3.3}$$

This concludes the calculation of the n^{th} moment of the steady–state response time at $\lambda = 0$. The next objective is to calculate these measures at $\lambda = \mu$, i.e., in heavy traffic. This was carried out in Sections 5.3 of the last chapter. If we

specialize (V.3.31) to the case when both the service times and inter–arrival times are exponential, we obtain

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)^n \overline{T}^n(\lambda) = \Gamma(n + \frac{3}{2})\sqrt{\frac{3}{\pi}}(\frac{\sqrt{3}}{2})^n P_n \qquad n = 1, 2 \ldots (3.4)$$

where $\Gamma(n + \frac{3}{2})$ and P_n are as given in (V.4.31) and (V.4.32) respectively.

Combining the light traffic result (3.3) with the heavy traffic result (3.4), we obtain as the 0^{th} order approximation to the n^{th} moment of the response time of a two dimensional fork–join queue in the form

$$\begin{aligned}\hat{T}^{(n)}(\lambda) =& \frac{\Gamma(n+1)}{(\mu-\lambda)^n}(2 - \frac{1}{2^n}) \\ &+ \frac{\lambda}{\mu}\frac{1}{(\mu-\lambda)^n}\left[\Gamma(n + \frac{3}{2})\sqrt{\frac{3}{\pi}}(\frac{\sqrt{3}}{2})^n P_n - \Gamma(n+1)(2 - \frac{1}{2^n})\right].\end{aligned}$$

$$0 \leq \lambda < \mu, \; n = 1, 2 \ldots (3.5)$$

6.3.1 Simulation results

All the simulation results presented in this and succeeding sections are for $\mu = 1s^{-1}$, while λ is varied from $0.1s^{-1}$ to $0.9s^{-1}$ in steps of $0.1s^{-1}$.The 95% confidence levels have been obtained in all cases using the method of batch means [LaKe, p. 296]. The % error is calculated by using the formula

$$\% \text{ Error } = \frac{\overline{T}^{(1)}(\lambda) - \hat{T}^{(1)}(\lambda)}{\overline{T}^{(1)}(\lambda)} \times 100.$$

λ	$\overline{T}^{(1)}(\lambda)$	$\hat{T}^{(1)}(\lambda)$	% Error
0.1	1.65 ± 0.007	1.65	0.30
0.2	1.85 ± 0.011	1.84	0.43
0.3	2.09 ± 0.0158	2.09	0.43
0.4	2.44 ± 0.024	2.41	1.22
0.5	2.91 ± 0.037	2.87	1.37
0.6	3.63 ± 0.061	3.56	1.93
0.7	4.80 ± 0.109	4.71	1.87
0.8	7.16 ± 0.23	7.0	2.23
0.9	13.91 ± 0.32	13.87	0.29

6.4 Markovian symmetric fork–join queue: The case $K > 2$

In this section we initially develop first order light traffic approximations for the average response time of a two–dimensional symmetric fork–join queue with Poisson arrivals with rate λ and exponential service times with rate μ. The approximations are first order approximations in the sense that we shall consider the limiting value of the average response time as λ approaches zero and μ, as well as the value of the derivative of the response time at $\lambda = 0$. In Section 6.4.1 we make a conjecture regarding heavy traffic limits for the case $K > 2$. Using this conjecture we give polynomial approximations for the first moment of the response time for a K–dimensional fork–join queue, which agree extremely well with experimental results.

Since we only consider approximations for the first moments of the average response time, the super–script n will be omitted from $\overline{T}^{(n)}$ for the rest of this section. Note that by (3.3) we have

$$\overline{T}(0) = \frac{3}{2\mu} \tag{4.1}$$

while by (3.4) we have

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{T}(\lambda) = \frac{11}{8}. \tag{4.2}$$

We now proceed to calculate $\overline{T}'(0)$ with the help of (2.7). To that end, let $T(t, s_1, s_2, v_1, v_2)$ denote the response time of the batch which enters the system at time 0 with service times v_1 and v_2, given that another batch arrives at time t with service times s_1 and s_2. It is clear that

$$T(t, s_1, s_2, v_1, v_2) = \begin{cases} \max(v_1, v_2) & \text{if } t \geq 0 \\ \max(v_1 + (s_1 + t)^+, v_2 + (s_2 + t)^+) & \text{if } t < 0. \end{cases} \tag{4.3}$$

It is plain from (4.3) that for $t \geq 0$

$$\overline{\psi}(\{t\}) = I\!E_{Q_\lambda}[\max\{v_1, v_2\}] = \overline{\psi}(\emptyset)$$

so that (2.7) now reduces to

$$\frac{d\overline{T}(0)}{d\lambda} = \int_{-\infty}^{0} (\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset))dt$$

Define the RVs X_1, X_2, Y_1 and Y_2 as

$$X_i = (s_i + t)^+, \quad Y_i = v_i + X_i, \quad i = 1, 2.$$

The RVs $X_i, i = 1, 2$ are iid with common distribution F_X given by

$$F_X(x) = 1 - e^{\mu t} e^{-\mu x}, \quad x \geq 0. \tag{4.4}$$

while the RVs $Y_i, i = 1, 2$ are iid with common distribution F_Y given by

$$F_Y(y) = 1 - e^{-\mu y} - \mu y e^{\mu t} e^{-\mu y}, \quad y \geq 0.$$

Note that

$$T := T(t, s_1, s_2, v_{1,2}) = \max(Y_1, Y_2) \quad \text{if } t < 0$$

Since the RVs Y_1 and Y_2 are independent, we obtain

$$\begin{aligned} Q_\lambda(T \leq x) =& Q_\lambda(Y_1 \leq x) Q_\lambda(Y_2 \leq x), \quad x \geq 0 \\ =& 1 + e^{-2\mu x} + \mu^2 x^2 e^{2\mu t} e^{-2\mu x} - 2e^{-\mu x} \\ & + 2\mu x e^{\mu t} e^{-2\mu x} - 2\mu x e^{\mu t} e^{-\mu x}. \end{aligned} \tag{4.5}$$

Using the fact that

$$\overline{\psi}(\{t\}) = \int_0^\infty (1 - Q_\lambda(T \leq x))dx, \quad \text{if } t < 0$$

and using (2.5), (4.1) and (4.5), we see that

$$\overline{\psi}(\{t\}) = \begin{cases} \frac{3}{2\mu}, & \text{if } t \geq 0, \\ \frac{2e^{\mu t}}{\mu} - \frac{e^{\mu t}}{2\mu} - \frac{e^{2\mu t}}{4\mu} + \frac{3}{2\mu} & \text{if } t < 0. \end{cases} \tag{4.6}$$

Finally combining (2.7) with (4.6) and (4.1), we obtain

$$\overline{T}'(0) = -\frac{1}{8\mu^2} - \frac{1}{2\mu^2} + \frac{2}{\mu^2} = \frac{11}{8\mu^2}. \tag{4.7}$$

Using (4.1), (4.2) and (4.7), we now obtain a first order approximation to the average response time of the fork–join queue. Setting,

$$t(\lambda) = (\mu - \lambda)\overline{T}(\lambda), \quad 0 \leq \lambda \leq \mu \tag{4.8}$$

we readily see that

$$t(0) = \frac{3}{2}, \quad t'(0) = -\frac{1}{8\mu}, \quad t(\mu) = \frac{11}{8}. \tag{4.9}$$

Let $\hat{t}(\lambda)$ denote the quadratic interpolation of $t(\lambda)$ over the range $[0, \mu]$, say

$$\hat{t}(\lambda) = k_0 + k_1\lambda + k_2\lambda^2, \quad 0 \leq \lambda \leq \mu. \tag{4.10}$$

Using (4.9) we come to the conclusion that

$$k_0 = \frac{3}{2}, \quad k_1 = -\frac{1}{8\mu}, \quad k_2 = 0 \tag{4.11}$$

so that

$$\hat{t}(\lambda) = \frac{3}{2} - \frac{1}{8}\frac{\lambda}{\mu}, \quad 0 \leq \lambda \leq \mu. \tag{4.12}$$

Finally undoing the normalization, we obtain the first order approximation $\hat{T}(\lambda)$ to the average response time in steady state in the form

$$\hat{T}(\lambda) = \frac{3}{2(\mu - \lambda)} - \frac{\lambda}{8\mu}\frac{1}{(\mu - \lambda)}, \quad 0 \leq \lambda < \mu. \tag{4.13}$$

Note that the first order approximation to the average response time of the two–dimensional fork–join queue is the same as the 0^{th} order approximation due to the fact that $k_2 = 0$.

6.4.1 A conjecture

So far we have developed approximations only for two dimensional fork–join queues due to the fact that only in the case $K = 2$, were we able solve the basic PDE for the stationary distribution of the diffusion limit of the end–to–end delay. Because of the complexity involved, it is unlikely that we shall be able to obtain heavy traffic limits for the case $K > 2$ by solving PDE's. Hence even though light traffic limits are available for $K > 2$, our ignorance of the corresponding heavy traffic limits prevents us from obtaining approximations.

We now make the following crucial observation. For the case $K = 1$, when the system consists of a single $M/M/1$ queue, we observe that

$$\mu^2 \overline{T}'(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda) = 1.$$

Moreover, for the case $K = 2$ and exponential inter–arrival and service times, we note that

$$\mu^2 \overline{T}'(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda) = \frac{11}{8}. \tag{4.14}$$

Our conjecture is that such an equality is true in general for K–dimensional fork–join queues, as long as the arrival process is Poisson. Hence, if $\overline{T}_K(\lambda)$ denotes the average response time of a K–dimensional fork–join queue, then we conjecture the equality

$$\mu^2 \overline{T}'_K(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda), \qquad K = 3, 4 \ldots \tag{4.15}$$

This conjecture is extended to cover fork–join queues having the same inter–arrival and service distributions in Section 6.6. The approximations obtained by using this conjecture agree extremely well with simulation results as well as with the so–called "scaling approximation" of Nelson and Tantawi [52].

The rest of this section is devoted to obtaining approximations for $T_K(\lambda), K = 3, 4 \ldots$ with the help of (4.15). We first obtain a formula for $\overline{T}_K(0)$ with the help of (2.6). Since the batch arriving at $t = 0$ does not experience interference from any other customer, its queueing delay will be zero. Hence its response time will

be the maximum of K independent identically distributed exponential RVs. Let F_K denote the distribution of the maximum of K exponential RVs with rate μ, then

$$\begin{aligned} F_K(x) &= [1 - e^{-\mu x}]^K \\ &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r e^{-r\mu x}. \end{aligned}$$

so that

$$\begin{aligned} \overline{T}_K(0) &= \int_0^\infty [1 - F_K(x)]dx \\ &= \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \int_0^\infty (e^{-r\mu x} \\ &= \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{r\mu} \end{aligned}$$

With the help of the identity given after (III.5.15), this simplifies to

$$\overline{T}_K(0) = \frac{H_K}{\mu}. \qquad K = 2, 3, \ldots \quad (4.16)$$

We now proceed to calculate $\overline{T}'_K(0)$ with the help of (2.7). To that end, let $T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K)$ denote the response time of the batch that enters the system at time 0 with service times $v_1, \ldots, v_K$, given that another batch arrives at time t with service times $s_1, \ldots, s_K$. It is clear that

$$\begin{aligned} & T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K) \\ &= \begin{cases} \max(v_1, \ldots, v_K), & \text{if } t \geq 0 \\ \max(v_1 + (s_1 + t)^+, \ldots, v_K + (s_K + t)^+), & \text{if } t < 0. \end{cases} \end{aligned} \quad (4.17)$$

It is plain from (4.17) that for $t \geq 0$,

$$\overline{\psi}(\{t\}) = I\!E_{Q_\lambda}[\max\{v_1, \ldots, v_K\}] = \overline{\psi}(\emptyset)$$

so that (2.7) now reduces to

$$\overline{T}'_K(0) = \int_{-\infty}^{0} (\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset))dt.$$

Define the RVs $X_k, 1 \leq k \leq K$ and $Y_k, 1 \leq k \leq K$ by

$$X_k = (s_k + t)^+, \quad Y_k = v_k + X_k, \quad 1 \leq k \leq K. \tag{4.18}$$

Then each $X_k, 1 \leq k \leq K$, has the distribution F_X given in (4.4) while each $Y_k, 1 \leq k \leq K$, has the distribution F_Y given in (4.5). Note that

$$T_K := T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K) = \max(Y_1, \ldots, Y_K) \quad \text{if} \quad t < 0$$

Since the RVs $Y_k, 1 \leq k \leq K$, are independent, we obtain

$$\begin{aligned} Q_\lambda(T_K \leq x) &= \prod_{k=1}^{K} Q_\lambda(Y_k \leq x), \quad x \geq 0 \\ &= [1 - e^{-\mu x} - \mu x e^{\mu t} e^{-\mu x}]^K \\ &= [1 - (1 + \mu x e^{\mu t}) e^{-\mu x}]^K \\ &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r (1 + \mu x e^{\mu t})^r e^{-r\mu x} \\ &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r e^{-r\mu x} \sum_{m=0}^{r} \binom{r}{m} (\mu x e^{\mu t})^m. \end{aligned} \tag{4.19}$$

and therefore

$$\begin{aligned} \overline{\psi}(\{t\}) &= \int_0^\infty [1 - Q_\lambda(T_K \leq x)]dx \\ &= \int_0^\infty [1 - \sum_{r=0}^{K} \binom{K}{r} (-1)^r e^{-r\mu x} \sum_{m=0}^{r} \binom{r}{m} (\mu x e^{\mu t})^m]dx \\ &= \int_0^\infty \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} e^{-r\mu x} \sum_{m=0}^{r} \binom{r}{m} (\mu x e^{\mu t})^m dx \\ &= \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} (\mu e^{\mu t})^m \int_0^\infty x^m e^{-r\mu x} dx \end{aligned}$$

From the well known identity for the exponential distribution

$$\int_0^\infty x^m e^{-x} dx = m!, \qquad n = 0, 1 \ldots (4.20)$$

we conclude that

$$\overline{\psi}(\{t\}) = \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} e^{m\mu t} \frac{m!}{r^{m+1}}. \qquad (4.21)$$

Finally using (2.7) and the fact that

$$H_K = \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{r}$$

we obtain

$$\begin{aligned} \overline{T}'_K(0) &= \frac{1}{\mu} \int_{-\infty}^{0} \left[\sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} e^{m\mu t} \frac{m!}{r^{m+1}} - H_K \right] dt \\ &= \frac{1}{\mu} \int_{-\infty}^{0} \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=1}^{r} \binom{r}{m} e^{m\mu t} \frac{m!}{r^{m+1}} dt \\ &= \frac{1}{\mu^2} \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=1}^{r} \binom{r}{m} \frac{(m-1)!}{r^{m+1}}. \qquad (4.22) \end{aligned}$$

We shall write (4.22) as

$$\overline{T}'_K(0) = \frac{V_K}{\mu^2}, \qquad K = 2, 3 \ldots$$

where

$$V_K = \sum_{r=1}^{K} \binom{K}{r} (-1)^{r-1} \sum_{m=1}^{r} \binom{r}{m} \frac{(m-1)!}{r^{m+1}}. \qquad K = 2, 3 \ldots (4.23)$$

We have tabulated $V_K, 1 \leq K \leq 20$, in Section 6.10.

We now give the heavy traffic limit by using the conjecture (4.15), i.e.,

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda) = V_K. \qquad K = 2, 3 \ldots (4.24)$$

Finally combining (4.16), (4.23) and (4.24) we obtain a 1^{st} order approximation to the average response time,

$$\hat{T}_K(\lambda) = \left[H_K + (V_K - H_K)\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu. \qquad K = 2, 3 \ldots (4.25)$$

Nelson and Tantawi [52] gave the following formula $\hat{T}_K^{NT}(\lambda)$ for the average response time of a K–dimensional fork–join queue with exponential service and inter–arrival times.

$$\hat{T}_K^{NT}(\lambda) = \left[\frac{H_K}{H_2} + \frac{4}{11}(1 - \frac{H_K}{H_2})\frac{\lambda}{\mu}\right]\hat{T}, \quad 0 \leq \lambda < \mu \qquad K = 2, 3 \ldots (4.26)$$

where

$$\hat{T} = \frac{12 - \frac{\lambda}{\mu}}{8(\mu - \lambda)}, \quad 0 \leq \lambda < \mu$$

is the 1^{st} order approximation to the average response time of a two–dimensional fork–join queue. They arrived at this formula by using both experimental as well as theoretical considerations. They showed that the relative error of their approximation as compared to simulation results was less than 5 percent for systems where $K \leq 32$.

We have checked our approximation against that of Nelson and Tantawi for $K \leq 15$, and our approximation seems to perform just as well (see Section 6.4.2). The two approximations are closely related as can be seen by taking the heavy traffic limit of $\hat{T}_K^{NT}$,

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\hat{T}_K^{NT}(\lambda) = \frac{1}{2} + \frac{7}{12}H_K. \qquad K = 2, 3 \ldots (4.27)$$

The numbers V_K and H_K are tabulated in Section 6.10 and as the reader may see, the right hand side of (4.27) agrees quite closely with V_K.

The advantages of our approximation to that of Nelson and Tantawi are two–fold,

(1): Nelson and Tantawi resorted to experimental results to obtain the values of the constants in their approximation, while we give exact closed–form expressions for all the constants appearing in our approximation.

(2): Nelson and Tantawi's approximation is only valid for fork–join queues with exponential inter–arrival and service distributions. On the other hand as we show in Sections 6.5, 6.6 and 6.7, our approximation procedure can be extended to cover fork–join queues with general inter–arrival and service distributions.

6.4.2 Simulation results

In this section, approximation (4.25) is compared with simulation results for the case when $\mu = 1$ while $K = 5, 10$ and 15.

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error	$\hat{T}_5^{NT}(\lambda)$
0.1	2.49 ± 0.008	2.49	0.04	2.48
0.2	2.75 ± 0.012	2.75	0.07	2.73
0.3	3.09 ± 0.017	3.08	0.32	3.06
0.4	3.55 ± 0.027	3.52	0.84	3.49
0.5	4.19 ± 0.042	4.14	1.09	4.10
0.6	5.16 ± 0.074	5.07	1.74	5.01
0.7	6.80 ± 0.154	6.62	2.64	6.54
0.8	9.85 ± 0.09	9.72	1.3	9.59
0.9	19.30 ± 0.43	19.02	1.45	18.74

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error	$\hat{T}_{10}^{NT}(\lambda)$
0.1	3.17 ± 0.009	3.17	0.09	3.16
0.2	3.48 ± 0.013	3.48	0.06	3.47
0.3	3.88 ± 0.018	3.86	0.51	3.86
0.4	4.42 ± 0.026	4.39	0.68	4.38
0.5	5.18 ± 0.042	5.12	1.16	5.11
0.6	6.34 ± 0.072	6.22	1.89	6.21
0.7	8.23 ± 0.137	8.05	2.18	8.05
0.8	11.92 ± 0.30	11.71	1.76	11.72
0.9	23.49 ± 0.41	22.68	3.45	22.76

λ	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error	$\hat{T}_{15}^{NT}(\lambda)$
0.1	3.58 ± 0.009	3.58	0.03	3.58
0.2	3.91 ± 0.013	3.91	0.13	3.91
0.3	4.35 ± 0.020	4.34	0.23	4.34
0.4	4.95 ± 0.031	4.90	1.01	4.92
0.5	5.78 ± 0.050	5.70	1.38	5.72
0.6	7.03 ± 0.086	6.88	2.13	6.94
0.7	9.14 ± 0.166	8.87	3.04	8.96
0.8	13.44 ± 0.392	12.83	4.54	13.01
0.9	25.90 ± 0.47	24.73	4.52	25.18

6.5 Approximations for queues with Poisson arrivals

Our objective in this section is to obtain approximations for K–dimensional fork–join queues with Poisson arrivals. Since these queues do not satisfy the condition $\beta = \frac{1}{2}$ (except when the service times are exponential), we do not have their heavy traffic limit even for the case $K = 2$. However, notice that that in the case $K = 1$, when the system consists of a single $M/GI/1$ queue, the following equality holds

$$\mu^2 \overline{T}'(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda) = \frac{1 + \mu^2 \sigma_V^2}{2}.$$

Hence, we assume that such an equality is true in general as K increases, so that

$$\mu^2 \overline{T}'_K(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda) \qquad K = 2, 3 \ldots (5.1)$$

holds for all fork–join queues with Poisson arrivals. Our experimental results indicate that the heavy traffic approximations so obtained are quite good.

We illustrate the general methodology by considering the special case of Poisson arrivals and second order Erlangian service times, even though any other service time distribution would have sufficed.

We now proceed to obtain a formula for $\overline{T}_K(0)$ with the help of (2.6). The response time of the batch arriving at $t = 0$ into an empty system will be the maximum of K identically distributed RVs with density f given by

$$f(x) = 4\mu^2 x e^{-2\mu x}, \quad x \geq 0.$$

Let F_K denote the distribution of the maximum of K of these RVs with rate μ, then for $x \geq 0$, we see that

$$\begin{aligned} F_K(x) &= [1 - (1 + 2\mu x) e^{-2\mu x}]^K \\ &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r (1 + 2\mu x)^r e^{-2r\mu x} \\ &= \sum_{r=0}^{K} \binom{K}{r} (-1)^r \sum_{m=0}^{r} \binom{r}{m} (2\mu x)^m e^{-2r\mu x} \end{aligned}$$

$$K = 2, 3 \ldots (5.2)$$

Therefore

$$\overline{T}_K(0) = \int_0^\infty \Big[1 - \sum_{r=0}^{K} \binom{K}{r}(-1)^r \sum_{m=0}^{r} \binom{r}{m}(2\mu x)^m e^{-2r\mu x}\Big] dx$$

$$= \sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m}(2\mu)^m \int_0^\infty x^m e^{-2r\mu x} dx$$

$$= \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} \frac{m!}{(2r)^{m+1}}. \qquad K = 2, 3 \ldots (5.3)$$

We shall write (5.3) as

$$\overline{T}_K(0) = \frac{F_K}{\mu}, \qquad K = 2, 3 \ldots (5.4)$$

where

$$F_K = \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r}(-1)^{r-1} \sum_{m=0}^{r} \binom{r}{m} \frac{m!}{2r^{m+1}}. \qquad K = 2, 3 \ldots 5)$$

The numbers $F_K, 1 \leq k \leq 20$, are tabulated in Section 6.10. Our next objective is to obtain a formula for $\overline{T}'_K(0)$. Let $T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K)$ be the response time of the batch that enters the system at time $t = 0$ with service times $v_1, \ldots, v_K$ given that another batch arrives at time t with service times $s_1, \ldots, s_K$. Then $T_K(t, s_1, \ldots, s_K, v_1, \ldots, v_K)$ satisfies equation (4.20). For $1 \leq k \leq K$, define the RVs X_k and Y_k as in (4.21)–(4.22). Here each RV $X_k, 1 \leq k \leq K$, has distribution F_X where

$$F_X(x) = 1 - e^{-2\mu x} - 2\mu x e^{-2\mu x}, \quad x \geq 0 \qquad (5.6)$$

and each RV $Y_k, 1 \leq k \leq K$, has distribution F_Y where

$$F_Y(x) = 1 - (1 + 2\mu x)e^{-2\mu x} - \Big[(2\mu^2 - 4\mu^3 t)x^2 + \frac{4}{3}\mu^3 x^3\Big] e^{2\mu t} e^{-2\mu x}, \; x \geq 0 \quad (5.7)$$

Note that

$$T_K = \max(Y_1, \ldots, Y_K) \quad \text{if} \quad t \leq 0$$

Since the RVs $Y_k, 1 \leq k \leq K$, are iid, we obtain

$$
\begin{aligned}
&Q_\lambda(T_K \leq x) \\
&\quad = \left[1 - (1+2\mu x)e^{-2\mu x} - \left[(2\mu^2 - 4\mu^3 t)x^2 + \frac{4}{3}\mu^3 x^3\right]e^{2\mu t}e^{-2\mu x}\right]^K \\
&\quad = \sum_{r=0}^{K} \binom{K}{r}(-1)^r \left[1 - (1+2\mu x)e^{-2\mu x}\right]^{K-r} \\
&\qquad \times \left[(2\mu^2 - 4\mu^3 t)x^2 + \frac{4}{3}\mu^3 x^3\right]^r e^{2r\mu t}e^{-2r\mu x} \\
&\quad = \sum_{r=0}^{K} \binom{K}{r}(-1)^r \left[\sum_{q=0}^{K-r}\binom{K-r}{q}(-1)^q \sum_{m=0}^{q}\binom{q}{m}(2\mu x)^m e^{-2q\mu x}\right] \\
&\qquad \times \left[\sum_{n=0}^{r}\binom{r}{n}(2\mu^2 x^2)^n (\frac{4}{3}\mu^3 x^3)^{r-n}\sum_{p=0}^{n}\binom{n}{p}(-1)^p (2\mu t)^p\right] e^{2r\mu t}e^{-2r\mu x}
\end{aligned}
$$

Hence it follows that

$$\overline{\psi}(\{t\})$$

$$
\begin{aligned}
&= \int_0^\infty [1 - Q_\lambda(T_K \leq x)]dx \\
&= \int_0^\infty [\sum_{r=0}^{K}\binom{K}{r}(-1)^{r-1}\sum_{q=1}^{K-r}\binom{K-r}{q}(-1)^q \sum_{m=0}^{q}\binom{q}{m}(2\mu x)^m e^{-2q\mu x} \\
&\qquad \times \sum_{n=0}^{r}\binom{r}{n}(2\mu^2 x^2)^n (\frac{4}{3}\mu^3 x^3)^{r-n}\sum_{p=0}^{n}\binom{n}{p}(-1)^p (2\mu t)^p e^{2r\mu t}e^{-2r\mu x}]dx \\
&= \sum_{r=0}^{K}\binom{K}{r}(-1)^{r-1}\sum_{q=1}^{K-r}\binom{K-r}{q}(-1)^q \sum_{m=0}^{q}\binom{q}{m}(2\mu)^m \\
&\qquad \times \sum_{n=0}^{r}\binom{r}{n}(2\mu^2)^n (\frac{4}{3}\mu^3)^{r-n}\sum_{p=0}^{n}\binom{n}{p}(-1)^p (2\mu t)^p e^{2r\mu t}\int_0^\infty x^{m+3r-n}e^{-2(r+q)\mu x}dx \\
&= \sum_{r=0}^{K}\binom{K}{r}(-1)^{r-1}\sum_{q=1}^{K-r}\binom{K-r}{q}(-1)^q \sum_{m=0}^{q}\binom{q}{m}(2\mu)^m
\end{aligned}
$$

$$\times \sum_{n=0}^{r} \binom{r}{n} (2\mu^2)^n (\frac{4}{3}\mu^3)^{r-n} \sum_{p=0}^{n} \binom{n}{p} (-1)^p (2\mu t)^p e^{2r\mu t} \frac{(m+3r-n)!}{[2(r+q)\mu]^{m+3r-n+1}}$$

$$= \frac{1}{\mu} \sum_{r=0}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=1}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$

$$\times \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \sum_{p=0}^{n} \binom{n}{p} (-1)^p (2\mu t)^p e^{2r\mu t} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \tag{5.8}$$

Finally using (4.4) we obtain

$$\overline{T}'_K(0) = \frac{1}{\mu} \int_{-\infty}^{0} [\sum_{r=0}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$

$$\times \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} (-1)^p (2\mu t)^p e^{2r\mu t} - F_K] dt$$

$$= \frac{1}{\mu} \int_{0}^{\infty} [\sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$

$$\times \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} (2\mu t)^p e^{-2r\mu t}] dt$$

$$= \frac{1}{\mu} \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$

$$\times \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} (2\mu)^p \int_{0}^{\infty} t^p e^{-2r\mu t} dt$$

$$= \frac{1}{\mu^2} \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$

$$\times \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} \frac{p!}{(2r)^{p+1}} \tag{5.9}$$

We shall write (5.9) as

$$\overline{T}'_K(0) = \frac{G_K}{\mu^2}, \qquad K = 2, 3 \ldots \tag{5.10}$$

where

$$G_K = \sum_{r=1}^{K} \binom{K}{r} \frac{(-1)^{r-1}}{2^{r+1}} \sum_{q=0}^{K-r} \binom{K-r}{q} (-1)^q \sum_{m=0}^{q} \binom{q}{m}$$

$$\times \sum_{n=0}^{r} \binom{r}{n} \frac{1}{3^{r-n}} \frac{(m+3r-n)!}{(r+q)^{m+3r-n+1}} \sum_{p=0}^{n} \binom{n}{p} \frac{p!}{(2r)^{p+1}}$$

$$K = 2, 3 \ldots (5.11)$$

We have tabulated $G_K, 1 \leq K \leq 20$, in Section 6.10.

Now using the conjecture (5.1) we conclude that

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda) = G_K. \qquad K = 2, 3 \ldots (5.12)$$

Combining (5.4), (5.10) and (5.12) we obtain the following 1^{rst} order approximation $\hat{T}_K(\lambda)$, for the average response time in K–dimensional fork–join queue with Poisson arrivals and Erlangian service times,

$$\hat{T}_K(\lambda) = \left[F_K + (G_K - F_K) \frac{\lambda}{\mu} \right] \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu. \qquad K = 2, 3 \ldots (5.13)$$

This approximation is in very good agreement with simulation results (see Section 6.5.1).

6.5.1 Simulation results

In this section, approximation (5.13) is compared with simulation results for the case when $\mu = 1$ while $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error
0.1	1.48 ± 0.005	1.48	0.13
0.2	1.62 ± 0.007	1.61	0.37
0.3	1.79 ± 0.009	1.78	0.50
0.4	2.02 ± 0.014	2.01	0.49
0.5	2.35 ± 0.022	2.33	0.86
0.6	2.85 ± 0.039	2.81	1.4
0.7	3.65 ± 0.07	3.61	1.09
0.8	5.30 ± 0.19	5.20	1.88
0.9	10.08 ± 0.20	9.98	0.99

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	2.03 ± 0.005	2.04	0.24
0.2	2.20 ± 0.007	2.21	0.45
0.3	2.43 ± 0.011	2.42	0.41
0.4	2.73 ± 0.015	2.71	0.73
0.5	3.15 ± 0.022	3.11	1.26
0.6	3.80 ± 0.041	3.72	2.10
0.7	4.88 ± 0.079	4.73	3.07
0.8	6.98 ± 0.06	6.74	3.43
0.9	13.34 ± 0.23	12.79	4.12

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
0.1	2.46 ± 0.005	2.46	0.08
0.2	2.66 ± 0.007	2.66	0.007
0.3	2.92 ± 0.011	2.91	0.31
0.4	3.27 ± 0.017	3.23	1.22
0.5	3.75 ± 0.027	3.70	1.33
0.6	4.49 ± 0.047	4.39	2.22
0.7	5.75 ± 0.10	5.54	3.65
0.8	8.30 ± 0.07	7.85	5.42
0.9	15.90 ± 0.27	14.77	7.10

λ	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	2.71 ± 0.005	2.71	0.07
0.2	2.92 ± 0.008	2.92	0.07
0.3	3.19 ± 0.012	3.18	0.31
0.4	3.57 ± 0.021	3.53	1.12
0.5	4.10 ± 0.037	4.02	1.95
0.6	4.90 ± 0.062	4.76	2.86
0.7	6.26 ± 0.114	5.99	4.31
0.8	9.03 ± 0.29	8.44	6.53
0.9	17.34 ± 0.30	15.81	8.82

6.6 Erlangian symmetric fork–join queue: The case $K = 2$

In the present section we obtain approximations for all the moments of the response time of a two–dimensional fork–join queue with Erlang distributed service and inter–arrival times. Exact heavy traffic results for this system were obtained in Chapter 5. In the next section we obtain approximations for the average response time of a K–dimensional fork–join queue with Erlang distributions.

Consider a two–dimensional symmetric fork–join queue with second order Erlangian arrivals with rate λ and second order Erlangian service times with rate μ. In the present section we develop light traffic estimates for the response time statistics of this queue, and then combine it with the heavy traffic estimates of Chapter 5 to yield an estimate that is also valid for moderate traffic. The approximations that we develop in this section hold for all moments of the response time.

The density f of a second order Erlangian distribution with rate μ is given by

$$f(x) = 4\mu^2 x e^{-2\mu x}, \quad x \geq 0. \tag{6.1}$$

and its mean and variance are given by $\frac{1}{\mu}$ and $\frac{1}{2\mu^2}$ respectively.

We now proceed to calculate $\overline{T}^{(n)}(0)$ as follows. Since the batch arriving at $t = 0$ into an empty system does not experience interference from any other customers, its queueing delay will be zero, and hence its response time will simply be the maximum of two identically distributed Erlangian RVs. If F denotes the distribution of the maximum of two second order Erlangian RVs with rate μ, then

$$F(z) = \left[1 - e^{-2\mu z} - 2\mu z e^{-2\mu z}\right]^2, \quad z \geq 0 \tag{6.2}$$

so that its density function f is given by

$$f(z) = \left[1 - e^{-2\mu z} - 2\mu z e^{-2\mu z}\right] 8\mu^2 z e^{-2\mu z}, \quad z \geq 0. \tag{6.3}$$

Hence, $\overline{T}^{(n)}(0)$ is given by the formula

$$\overline{T}^n(0) = \int_0^\infty z^n f(z) dz$$

$$= \frac{8}{\mu^n}\left[\frac{(n+1)!}{2^{n+2}} - \frac{(n+1)!}{4^{n+2}} - \frac{2(n+2)!}{4^{n+3}}\right] \quad n = 1, 2 \ldots (6.4)$$

We shall write (6.4) as

$$\overline{T}^{(n)}(0) = \frac{Q(n)}{\mu^n}, \qquad n = 1, 2 \ldots (6.5)$$

where

$$Q(n) = 8\left[\frac{(n+1)!}{2^{n+2}} - \frac{(n+1)!}{4^{n+2}} - \frac{2(n+2)!}{4^{n+3}}\right]. \qquad n = 1, 2 \ldots (6.6)$$

This concludes the calculation of the n^{th} moment of the steady response time at $\lambda = 0$. The next objective is to calculate this measure at $\lambda = \mu$, i.e., in heavy traffic. If we specialize the results of Sections 5.4 to the case when both the service times and inter–arrival times are second order Erlangian, we obtain

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)^n \overline{T}^{(n)}(\lambda) = \Gamma(n + \frac{3}{2})\sqrt{\frac{3}{\pi}}(\frac{\sqrt{3}}{2})^n \frac{P_n}{2^n} \qquad n = 1, 2 \ldots (6.7)$$

where $\Gamma(n + \frac{3}{2})$ and P_n are given in (V.4.31)-(V.4.32).

Finally combining (6.5) with (6.7), we obtain as the 0^{th} order approximation $\hat{T}^{(n)}(\lambda)$ to the n^{th} moment of the response time as

$$\hat{T}^{(n)}(\lambda) = \frac{Q(n)}{(\mu - \lambda)^n} + \frac{\lambda}{\mu}\frac{1}{(\mu - \lambda)^n}\left[\Gamma(n + \frac{3}{2})\sqrt{\frac{3}{\pi}}(\frac{\sqrt{3}}{2})^n \frac{P_n}{2^n} - Q(n)\right].$$

$$0 \leq \lambda < \mu, \; n = 1, 2 \ldots (6.8)$$

6.6.1 Simulation results

In this section, approximation (6.8) is compared with simulation results for the case when $\mu = 1$ and $n = 1$.

λ	$\overline{T}^{(1)}(\lambda)$	$\hat{T}^{(1)}(\lambda)$	% Error
0.1	1.40 ± 0.004	1.45	3.57
0.2	1.46 ± 0.005	1.54	5.48
0.3	1.55 ± 0.006	1.67	7.74
0.4	1.69 ± 0.008	1.83	8.28
0.5	1.91 ± 0.012	2.06	7.85
0.6	2.23 ± 0.021	2.41	8.07
0.7	2.80 ± 0.041	2.98	6.42
0.8	4.03 ± 0.11	4.12	2.23
0.9	7.44 ± 0.14	7.56	1.61

6.7 Erlangian symmetric fork–join queue: The case $K > 2$.

The main difficulty in obtaining limit theorem approximations for the case $K > 2$ is that heavy traffic approximations are no longer available. We were able to overcome this problem for the case when the inter–arrival times are exponential by postulating that

$$\mu^2 \overline{T}'_K(0) = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda). \qquad K = 3, 4 \ldots (7.1)$$

However simulation results suggest that such an equality is no longer true, even for the case of single server queues if the assumption about exponential inter–arrival times is relaxed. For example for a single server queue with second order Erlangian inter–arrival times and exponential service times, we found

$$\mu^2 \overline{T}'(0) = 0 \neq \frac{3}{4} = \lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda).$$

However, we are able to recover the heavy traffic approximation for the case $K > 2$ and the inter–arrival and service times have the same distribution, by making the following crucial observation. For the case $K = 2$, and exponential inter–arrival and service distributions, we have

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda) = \frac{11}{8}.$$

More generally, if the arrivals and services have the same distribution for $K = 2$, then by (V.4.25) we have

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}(\lambda) = \frac{11}{8} \sigma^2 \mu^2. \qquad (7.2)$$

For the case $K > 2$, and exponential inter–arrival and service distributions, our conjecture (4.18) states that

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda) \overline{T}_K(\lambda) = V_K.$$

By observing the form of the expression (7.2) when $K = 2$, one would expect that for the case $K > 2$ and the arrivals and services have the same distributions, the following holds

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = V_K \sigma^2 \mu^2. \qquad K = 3, 4 \ldots (7.3)$$

Equation (7.3) does indeed provide a very good heavy traffic approximation as our simulation results indicate (in Section 6.7.1). In the remainder of this section we use (7.3) to obtain a first order approximation to the average end–to–end delay of a K–dimensional Erlangian fork–join queue. An application of (7.3) to this queue yields

$$\lim_{\lambda \uparrow \mu} \overline{T}_K(\lambda) = \frac{V_K}{2}. \qquad K = 2, 3 \ldots (7.4)$$

Using results from Section 6.5, we can write down the following light traffic limit

$$\overline{T}_K(0) = \frac{F_K}{\mu} \qquad K = 2, 3 \ldots (7.5)$$

where F_K was defined in (5.5).

Finally, by combining (7.4) and (7.5), we obtain an approximation for the average response time of a K–dimensional fork–join queue with Erlangian inter–arrival and service distributions. This approximation, denoted by $\hat{T}_K(\lambda)$, is given by

$$\hat{T}_K(\lambda) = \left[F_K + (\frac{V_K}{2} - F_K)\frac{\lambda}{\mu}\right] \frac{1}{\mu - \lambda} \quad 0 \leq \lambda < \mu \qquad K = 2, 3 \ldots (7.6)$$

and agrees extremely well with simulation results (see Section 6.7.1).

6.7.1 Simulation results

In this section, approximation (7.6) is compared with simulation results for the cases $K = 5, 10$ and 15.

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	1.93 ± 0.004	2.01	4.14
0.2	2.01 ± 0.005	2.13	5.97
0.3	2.14 ± 0.007	2.30	7.47
0.4	2.33 ± 0.010	2.52	8.15
0.5	2.62 ± 0.015	2.83	8.01
0.6	3.06 ± 0.024	3.29	7.51
0.7	3.83 ± 0.046	4.07	6.26
0.8	5.43 ± 0.11	5.62	3.49
0.9	10.11 ± 0.15	10.27	1.58

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
0.1	2.34 ± 0.004	2.43	3.84
0.2	2.43 ± 0.005	2.58	6.17
0.3	2.58 ± 0.007	2.78	7.75
0.4	2.81 ± 0.010	3.04	8.18
0.5	3.15 ± 0.014	3.41	8.25
0.6	3.68 ± 0.022	3.96	7.60
0.7	4.59 ± 0.043	4.87	6.10
0.8	6.45 ± 0.09	6.70	3.87
0.9	12.09 ± 0.15	12.19	0.83

λ	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	2.58 ± 0.003	2.68	3.87
0.2	2.68 ± 0.004	2.85	6.34
0.3	2.84 ± 0.006	3.06	7.74
0.4	3.08 ± 0.009	3.34	8.44
0.5	3.43 ± 0.014	3.74	9.03
0.6	33.99 ± 0.025	4.33	8.52
0.7	4.94 ± 0.052	5.32	7.69
0.8	6.87 ± 0.12	7.31	6.40
0.9	12.89 ± 0.47	13.25	2.79

6.8 A formula for the heavy traffic limit for general β and K

Consider a K–dimensional fork–join queue. Recall that the parameter β was defined by

$$\beta = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}$$

where σ_0 is the variance of the inter–arrival times and σ is the variance of the service times. Heavy traffic limits for this queue are known for $\beta = 0$ and $\beta = 1$, while for $\beta = \frac{1}{2}$, we have our conjecture (7.3). In particular we see that,

(1): $\beta = 0$: In this case the K queues are decoupled from each other since the arrival stream is deterministic, and using (III.5.31), we can write

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = H_K \frac{\sigma^2 \mu^2}{2}. \qquad K = 2, 3 \ldots (8.1)$$

(2): $\beta = \frac{1}{2}$: In this case $\sigma_0 = \sigma$ and according to our conjecture (7.3),

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = V_K \sigma^2 \mu^2. \qquad K = 2, 3 \ldots (8.2)$$

(3): $\beta = 1$: In this case $\sigma = 0$ and the system behaves like a $GI/D/1$ queue so that

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = \frac{\sigma_0^2 \mu^2}{2}. \qquad K = 2, 3 \ldots (8.3)$$

Observing the structure of (8.1)–(8.3), we may venture to write down the following expression for the heavy traffic limit for a general value of β.

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = M_K(\beta) \frac{\sigma^2 + \sigma_0^2}{2} \mu^2, \quad 0 \leq \beta \leq 1, \qquad K = 2, 3 \ldots (8.4)$$

where

$$M_K(0) = H_K, \quad M_K(\frac{1}{2}) = V_K, \quad M_K(1) = 1. \qquad (8.5)$$

In the absence of any further information we may use a quadratic approximation for $M(\beta)$. This leads to

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda)$$

$$= \left[H_K + (4V_K - 3H_K - 1)\beta + 2(1 + H_K - 2V_K)\beta^2\right] \frac{\sigma^2 + \sigma_0^2}{2}\mu^2, \quad 0 \leq \beta \leq 1.$$

$$K = 2, 3 \ldots (8.6)$$

We may verify that (8.6) is indeed a very good approximation by making the following observation: Consider the case when the arrivals are Poisson and the service times are second order Erlang distributed, so that $\beta = \frac{2}{3}$. In this case (8.6) leads to

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = \frac{1}{6} - \frac{H_K}{12} + \frac{2}{3}V_K. \qquad K = 2, 3 \ldots (8.7)$$

If we compare the right hand side of (8.7) with the values of G_K given in the table in Section 6.10, then we observe that they indeed match very closely. Moreover, simulation results in Section 6.5.1 indicated that the approximations obtained by using G_K as the heavy traffic limit, matched closely with simulation results. This implies that the approximation obtained by using (8.7) as the heavy traffic limit would also match closely with simulation.

As further evidence that (8.6) provides a good heavy traffic approximation, consider the case when the arrivals are second order Erlangian and the service times are exponential, in which case $\beta = \frac{1}{3}$ and (8.6) leads to

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = -\frac{1}{12} + \frac{1}{6}H_K + \frac{2}{3}V_K. \qquad K = 2, 3 \ldots (8.8)$$

Also it is obvious that for this system

$$\overline{T}_K(0) = \frac{H_K}{\mu}. \qquad K = 2, 3 \ldots (8.9)$$

Combining (8.8) with (8.9) we obtain the following approximation $\hat{T}_K(\lambda)$, for the average response time,

$$\hat{T}_K(\lambda) = \left[H_K + (\frac{2}{3}V_K - \frac{5}{6}H_K - \frac{1}{12})\frac{\lambda}{\mu}\right] \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu. \quad K = 2, 3 \ldots (8.10)$$

This approximation agrees extremely well with simulation results (see Section 6.8.1).

We now present a general methodology for obtaining limit theorem approximations for fork–join queues. Assume that the K–dimensional fork–join queue is governed by inter–arrival times possessing a distribution function A and service times possessing a distribution function B. Also assume that the distribution A has mean $\frac{1}{\lambda}$ and variance σ_0, while the distribution B has mean $\frac{1}{\mu}$ and variance σ. The following procedure then gives the heavy and light traffic limits for the system.

(1): Calculate the value of $\beta = \frac{\sigma_0^2}{\sigma_0^2+\sigma^2}$. Substitute the values of $H_K, V_K, \sigma_0, \sigma, \mu$ and β into (8.6) to obtain the heavy traffic limit a, for the system.

(2): Obtain the light traffic b for the system by using the formula

$$b = \int_0^\infty [1-(1-B(x))^K]dx.$$

This integration may be done numerically if necessary.

The limit theorem approximation for the system $\hat{T}(\lambda)$, is then given by

$$\hat{T}(\lambda) = \left[b\mu + (a-b\mu)\frac{\lambda}{\mu}\right]\frac{1}{\mu-\lambda}, \quad 0 \leq \lambda < \mu. \tag{8.11}$$

We now present an application of (8.11) to the case when the coefficient of variation of the service time distribution exceeds one. In this case we observe that the approximation works well in heavy traffic, but the relative error is quite large (in the region of 30 percent) for moderate or light traffic.

Consider a RV S possessing the following hyper–exponential density function f_S,

$$f_S(x) = p_1\mu_1 e^{-\mu_1 x} + p_1\mu_2 e^{-\mu_2 x}, \quad x \geq 0. \tag{8.12}$$

Assume that the condition

$$\frac{p_1}{\mu_1} = \frac{p_2}{\mu_2} = \frac{1}{2} \tag{8.13}$$

is satisfied. Then note that

$$ES = 1 \text{ and } ES^2 = \frac{1}{\mu_1} + \frac{1}{\mu_2} \tag{8.14}$$

so that the square of the coefficient of variation, c_S^2 is given by

$$c_S^2 = \frac{1}{\mu_1} + \frac{1}{\mu_2} - 1 = \frac{2}{\mu_1 \mu_2} - 1 \tag{8.15}$$

The light traffic limit for the average response time of a K–dimensional fork–join queue subject to Poisson arrivals, and service times with density function given by f_S can be obtained with the help of (2.6), and is reproduced below.

$$\overline{T}_K(0) = \sum_{r=1}^{K} \binom{K}{r} (-1)^{r+1} \sum_{m=0}^{r} \binom{r}{m} \frac{p_1^m p_2^{r-m}}{m\mu_1 + (r-m)\mu_2} \tag{8.16}$$

Let us consider the special case when $\mu_1 = 0.1, \mu_2 = 1.9, p_1 = 0.05$ and $p_2 = 0.95$. In this case $ES = 1, ES^2 = 10.526$, $c_s^2 = 9.526$ and $\beta = 0.095$. Substituting into (8.6), we come to the conclusion that the heavy traffic limit is given by

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{T}_K(\lambda) = 3.85H_K + 1.81V_K - 0.4. \tag{8.17}$$

Combining (8.16) with (8.17), it is possible to obtain interpolation approximations for this system, which are written down below.

$$\hat{T}_2(\lambda) = \frac{1.704 + 6.176\lambda}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{8.18}$$

$$\hat{T}_5(\lambda) = \frac{3.32 + 8.44\lambda}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{8.19}$$

$$\hat{T}_{10}(\lambda) = \frac{5.45 + 9.4\lambda}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{8.20}$$

and

$$\hat{T}_{15}(\lambda) = \frac{7.23 + 9.45\lambda}{1 - \lambda}, \quad 0 \leq \lambda \leq 1 \tag{8.21}$$

These approximations are compared with simulation results in Section 6.8.2. We observe that the relative error is quite small in heavy traffic, but large in light and moderate traffic. Hence it seems that the interpolation approximation technique gives good results only for the case when the coefficient of variation of the service times is less than one.

6.8.1 Simulation results

In this section, approximation (8.10) is compared with simulation results for the case when $\mu = 1$ while $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error
0.1	1.55 ± 0.007	1.62	4.51
0.2	1.65 ± 0.009	1.77	7.27
0.3	1.81 ± 0.011	1.96	8.28
0.4	2.05 ± 0.015	2.22	8.29
0.5	2.39 ± 0.024	2.58	7.95
0.6	2.91 ± 0.042	3.12	7.21
0.7	3.81 ± 0.088	4.03	5.77
0.8	5.65 ± 0.203	5.83	3.18
0.9	10.96 ± 0.24	11.25	2.64

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	2.34 ± 0.007	2.45	4.70
0.2	2.49 ± 0.008	2.67	7.23
0.3	2.71 ± 0.011	2.94	8.48
0.4	3.03 ± 0.016	3.31	9.24
0.5	3.50 ± 0.026	3.82	9.14
0.6	4.22 ± 0.044	4.59	8.77
0.7	5.47 ± 0.086	5.87	7.31
0.8	8.01 ± 0.21	8.43	5.24
0.9	15.75 ± 0.27	16.12	2.34

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
0.1	3.00 ± 0.007	3.13	4.33
0.2	3.18 ± 0.009	3.39	6.60
0.3	3.47 ± 0.013	3.73	7.49
0.4	3.88 ± 0.019	4.17	7.47
0.5	4.49 ± 0.03	4.79	6.68
0.6	5.41 ± 0.05	5.73	5.91
0.7	6.99 ± 0.09	7.29	4.29
0.8	10.21 ± 0.22	10.40	1.86
0.9	19.51 ± 0.28	19.74	1.18

λ	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	3.39 ± 0.007	3.54	4.30
0.2	3.59 ± 0.009	3.83	6.68
0.3	3.89 ± 0.011	4.20	7.97
0.4	4.34 ± 0.017	4.69	8.06
0.5	4.99 ± 0.027	5.37	7.61
0.6	5.99 ± 0.045	6.40	6.84
0.7	7.69 ± 0.008	8.11	5.46
0.8	11.08 ± 0.19	11.53	4.06
0.9	21.03 ± 0.85	21.80	3.67

6.8.2 Simulation results

In this section, approximations (8.18)–(8.21) are compared with simulation results for the case when $\mu = 1$ while $K = 2, 5, 10$ and 15.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error
0.1	2.78 ± 0.021	2.58	7.19
0.2	4.09 ± 0.042	3.67	10.26
0.3	5.71 ± 0.067	5.08	11.03
0.4	7.83 ± 0.105	6.96	11.11
0.5	10.73 ± 0.172	9.58	10.72
0.6	14.95 ± 0.289	13.52	9.56
0.7	21.70 ± 0.503	20.09	7.42
0.8	34.71 ± 1.50	33.22	4.29
0.9	71.14 ± 2.59	72.62	2.04

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	5.59 ± 0.030	4.63	17.26
0.2	8.22 ± 0.054	6.26	23.84
0.3	11.32 ± 0.086	8.36	26.15
0.4	15.10 ± 0.130	11.16	26.09
0.5	19.93 ± 0.198	15.08	24.33
0.6	26.70 ± 0.332	20.96	21.50
0.7	37.40 ± 0.609	30.76	17.75
0.8	57.94 ± 1.30	46.14	20.36
0.9	114.03 ± 3.11	109.16	4.27

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
0.1	9.18 ± 0.038	7.10	22.66
0.2	13.16 ± 0.066	9.16	30.39
0.3	17.56 ± 0.099	11.81	32.74
0.4	22.67 ± 0.147	15.35	32.29
0.5	28.99 ± 0.219	20.30	29.97
0.6	37.64 ± 0.34	27.72	26.35
0.7	50.89 ± 0.59	40.10	21.20
0.8	79.95 ± 1.19	64.85	18.88
0.9	146.72 ± 2.95	139.10	5.19

λ	$\overline{T}_{15}(\lambda)$	$\hat{T}_{15}(\lambda)$	% Error
0.1	11.99 ± 0.041	9.08	24.27
0.2	16.79 ± 0.069	11.40	32.10
0.3	21.92 ± 0.106	14.38	34.40
0.4	27.76 ± 0.157	18.35	33.90
0.5	34.94 ± 0.234	23.91	31.57
0.6	44.72 ± 0.363	32.25	27.88
0.7	59.97 ± 0.630	46.15	23.04
0.8	89.09 ± 1.29	73.95	16.99
0.9	174.26 ± 3.28	157.35	9.70

6.9 Approximations for acyclic fork–join networks

The reader may recall that we obtained the heavy traffic diffusion limit for acyclic fork–join networks in Chapter 2. However the form of the limiting diffusion was quite complicated and we were unable to say anything about the stationary distribution of this diffusion. Basing ourselves on the experience gained in the process of solving the single stage fork–join queue, we now present a simulation based approach for obtaining approximations for these systems. These approximations agree quite well with experimental results.

Consider the homogeneuous fork–join network depicted in Fig 2.2. Asume that the arrivals into the system constitute a Poisson stream with rate λ, while the service time distribution at each queue has rate μ and variance σ^2. Let $\overline{T}(\lambda)$ be the average response time of the network when the arrival rate is λ. Based on our experience with the fork–join queue, we would expect that the light traffic limits of the network are given by

$$\overline{T}(0) = \frac{A}{\mu} \tag{9.1}$$

$$\overline{T}'(0) = \frac{B}{\mu^2} \tag{9.2}$$

where A and B are constants. Moreover we use the conjecture (4.15) for this system so that

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{T}(\lambda) = B \tag{9.3}$$

Combining (9.1), (9.2) and (9.3) we obtain the following approximation $\hat{T}(\lambda)$ for the average response time of this system.

$$\hat{T}(\lambda) = \left[A + (B - A)\frac{\lambda}{\mu}\right] \frac{1}{\mu - \lambda} \tag{9.4}$$

The only task left is to identify the value of the constants A and B. Even though it is possible to calculate them using the light traffic equations, it is likely that the calculations would be quite tedious for arbitrary networks. A more viable

method would be automate the calculations using some symbolic computation tool such as MACSYMA. However, this is a subject for future research and here we only give a simulation based technique for obtaining A and B. The main idea in this technique is to simulate the system with $\mu = 1$, and use the simulation output for obtaining A and B. For e.g., for the network in Fig 2.2, for the case when the services are exponential with rate $\mu = 1$, we found that

$$\overline{T}(0.0005) = 5.05685 \text{ and } \overline{T}(0.01) = 5.100354$$

so that

$$A = 5.05685 \text{ and } B = 4.35$$

so that

$$\hat{T}(\lambda) = \left[5.05685 - 0.70685\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda} \tag{9.5}$$

This approximation is compared with simulations in Section 6.9.1, and as the reader may note, the agreement is quite good. For the case of Poisson arrivals and Erlang-2 services, the approximation is given by

$$\hat{T}(\lambda) = \left[4.759 - 2.176\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda} \tag{9.6}$$

while for the case of Poisson arrivals and deterministic services, the approximation is given by

$$\hat{T}(\lambda) = \left[4 - 3.467\frac{\lambda}{\mu}\right]\frac{1}{\mu - \lambda} \tag{9.7}$$

These approximations also compare extremely well with simulation results as shown in Section 6.9.1.

The technique that we have proposed here for obtaining approximations for acyclic fork–join networks is very efficient, since by incurring the cost of just two simulation runs, we are able to obtain a formula that holds for all values of λ and μ. Moreover, the simulations are made for $\lambda \approx 0$, so that a good confidence level can be obtained from a relatively short run.

6.9.1 Simulation results

Approximations (9.5), (9.6) and (9.7) are compared with simulation for the case when $\mu = 1$.

λ	$\overline{T}(\lambda)$	$\hat{T}(\lambda)$	% Error
0.1	5.56 ± 0.018	5.54	0.28
0.2	6.19 ± 0.026	6.141	0.73
0.3	7.02 ± 0.039	6.926	1.38
0.4	8.11 ± 0.058	7.96	1.91
0.5	9.63 ± 0.090	9.41	2.31
0.6	11.92 ± 0.152	11.59	2.87
0.7	15.73 ± 0.285	15.21	3.33
0.8	23.71 ± 0.215	22.46	5.29
0.9	46.93 ± 0.76	44.21	5.80

λ	$\overline{T}(\lambda)$	$\hat{T}(\lambda)$	% Error
0.1	5.06 ± 0.010	5.04	0.28
0.2	5.44 ± 0.017	5.40	0.66
0.3	5.93 ± 0.024	5.87	1.01
0.4	6.57 ± 0.034	6.48	1.35
0.5	7.47 ± 0.054	7.34	1.74
0.6	8.84 ± 0.093	8.63	2.37
0.7	11.19 ± 0.189	10.78	3.66
0.8	15.89 ± 0.123	15.09	5.03
0.9	30.13 ± 0.496	28.00	7.06

λ	$\overline{T}(\lambda)$	$\hat{T}(\lambda)$	% Error
0.1	4.05 ± 0.001	4.06	0.09
0.2	4.12 ± 0.002	4.13	0.19
0.3	4.21 ± 0.003	4.23	0.35
0.4	4.33 ± 0.005	4.35	0.53
0.5	4.50 ± 0.010	4.53	0.73
0.6	4.75 ± 0.017	4.79	1.02
0.7	5.18 ± 0.034	5.24	1.14
0.8	5.99 ± 0.023	6.13	2.33
0.9	8.44 ± 0.096	8.80	4.06

6.10 Tables

K	H_K	V_K	F_K	G_K
2	1.5	1.375	1.375	0.957
3	1.833	1.594	1.606	1.072
4	2.083	1.745	1.773	1.151
5	2.283	1.860	1.904	1.210
6	2.449	1.951	2.011	1.258
7	2.593	2.027	2.101	1.297
8	2.717	2.091	2.180	1.330
9	2.829	2.147	2.249	1.359
10	2.929	2.195	2.313	1.384
11	3.019	2.240	2.367	1.407
12	3.103	2.280	2.418	1.427
13	3.180	2.316	2.465	1.446
14	3.251	2.349	2.508	1.463
15	3.318	2.379	2.549	1.474
16	3.380	2.408	2.587	1.494
17	3.439	2.434	2.622	1.507
18	3.495	2.460	2.658	1.520
19	3.547	2.478	2.688	1.532
20	3.597	2.510	2.734	1.547

CHAPTER VII

7.1 Introduction

The next two chapters in this thesis are devoted to approximations for queues exhibiting the resequencing synchronization constraint. As was the case for fork–join queues, the presence of resequencing renders the analysis to be very complex and very few results are known [63]. Our work on heavy traffic diffusion limits for resequencing systems leads to the following advances:

(1): We have obtained good estimates for the queueing delays for several models that were previously intractable analytically;

(2): We have identified a class of models in which resequencing can be ignored in heavy traffic.

In the present chapter we obtain heavy traffic diffusion limits for a variety of resequencing systems possessing the following generic structure: Customers enter a disordering system which they leave (after being served) in an order different from the one in which they entered it. This necessitates resequencing which takes place in a so–called resequencing buffer. After leaving the resequencing buffer, the customers enter the buffer of a single server queue from where they leave the system. This generic model is introduced in Section 7.2, where we also give the recursive equations governing its delays. In Section 7.3 we obtain the heavy traffic diffusion limit for the generic model from Section 7.2 from which diffusion limits for specific models can be easily recovered.

In Section 7.4 we specialize the results of Section 7.3 for the important special case when the disordering system is an infinite server queue. We show that the queue delay process of this system has the same heavy traffic limit as an ordinary single server queue, i.e., in heavy traffic the resequencing delay has negligible influence on the operation of the system. We also extend this result to the case

when there may be more than one disordering and resequencing stages before the single server queue.

In Sections 7.5 and 7.6 we obtain the heavy traffic limit for finite server disordering systems. When the disordering system is a $GI/GI/K$ queue, we show that the normalized resequencing delay converges to zero in heavy traffic. For the case when the disordering system is composed of K single server queues operating in parallel, we use an alternate representation for the end–to–end delay of the system than the one given in Section 7.2, to obtain the heavy traffic diffusion limit. In this case, our results show that the resequencing delay constitutes the major portion of the total delay, in heavy traffic.

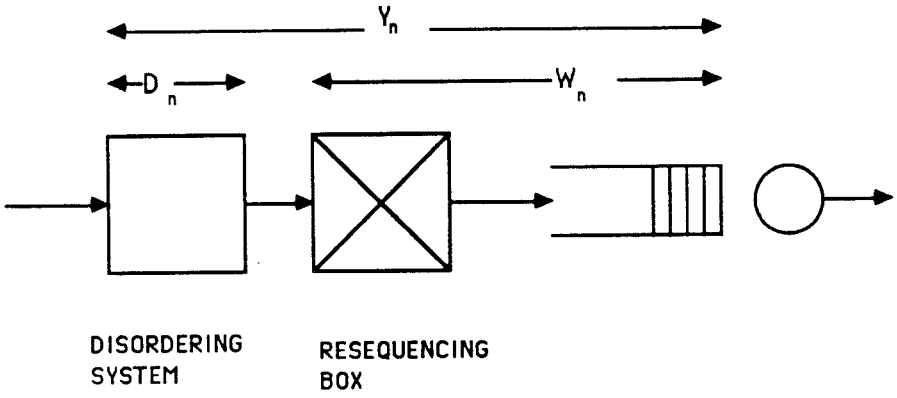


Fig. 7.1. A generic resequencing model.

7.2 The model

In this section we introduce a generic resequencing model, from which specific resequencing structures can be recovered as special cases. There is a stream of customers which enter a disordering system, and leave in an order different than the one in which they entered it. After leaving the disordering system, they wait in a resequencing buffer until all customers which entered the disordering system prior to them have left it. After leaving the resequencing box, these customers are served by a single server queue, before finally leaving the system. The model of Baccelli, Gelenbe and Plateau [1] is special case of this model when the disordering system corresponds to an infinite server queue.

We now define some RVs that are useful in discussing the properties of this system. Let the sequences of RVs $\{D_n\}_0^\infty$, $\{v_n\}_0^\infty$ and $\{\tau_n\}_0^\infty$ be defined on some probability space $(\Omega, I\!F, I\!P)$. Here τ_n represents the time of arrival of the n^{th} customer into the system, D_n represents its disordering delay and v_n represents its service time in the single server queue. In terms of these RVs define the following quantities for all $n = 0, 1 \ldots$,

u_{n+1} : Inter-arrival time between the $(n+1)^{rst}$ and the n^{th} customers $(= \tau_{n+1} -$

τ_n).

W_n : Delay of the n^{th} customer in the resequencing box and in the buffer of the single server queue.

Y_n : The end–to–end delay of the n^{th} customer ($= D_n + W_n$).

Various kinds of disordering systems can be realized by assuming different statistical structures on the sequence $\{D_n\}_0^\infty$. For example, if the delay sequence $\{D_n\}_0^\infty$ is an iid sequence which is independent of the inter-arrival sequence $\{u_n\}_0^\infty$ then the disordering system corresponds to a $GI/GI/\infty$ queue. Similarly we can realize the disordering system as $GI/GI/K$ queue or a system of K parallel $GI/GI/1$ queues by imposing a particular structure on $\{D_n\}_0^\infty$.

The analysis of this model is very difficult, one of the reasons being that the output stream from the resequencing buffer is a complicated process with batch departures and correlations between the batch sizes and inter–departure times. For the special case when the disordering system has an infinite number of servers, and the the sequences $\{u_n\}_0^\infty$, $\{D_n\}_0^\infty$ and $\{v_n\}_0^\infty$ are all exponentially distributed, Baccelli, Gelenbe and Plateau [1] were able to derive a complicated expression for the Laplace transform of the end–to–end delay Y_n. We now derive a recursion first given by Baccelli, Gelenbe and Plateau [1], governing the sequences $\{Y_n\}_0^\infty$ and $\{W_n\}_0^\infty$.

Lemma 7.2.1. *Consider the resequencing system defined above. If the system is initially empty, then the recursions*

$$Y_0 = D_0$$
$$Y_{n+1} = \max\{D_{n+1}, Y_n + v_n - u_{n+1}\}, \qquad n = 0, 1 \ldots (2.1)$$

and

$$W_0 = 0$$
$$W_{n+1} = \max\{0, W_n + D_n - D_{n+1} + v_n - u_{n+1}\}, n = 0, 1 \ldots (2.2)$$

hold.

Proof. Since there is no initial load in the system by assumption, the first customer in the system will not undergo any resequencing delay and the initial conditions are therefore immediate.

In order to prove (2.1) consider the $(n+1)^{rst}$ customer. Its resequencing delay plus queueing delay will be zero if the n^{th} customer has left the system at the time it leaves the disordering subsystem, i.e.,

$$Y_{n+1} = D_{n+1} \text{ if } \tau_{n+1} + D_{n+1} > \tau_n + Y_n + v_n, \qquad n = 0, 1 \ldots (2.3)$$

If the n^{th} customer has not left the system at the time the $(n+1)^{rst}$ leaves the disordering subsystem, then the $(n+1)^{rst}$ customer will experience a resequencing delay of duration $\tau_n + Y_n + v_n - (\tau_{n+1} + D_{n+1})$, hence

$$Y_{n+1} = D_{n+1} + [\tau_n + Y_n + v_n - (\tau_{n+1} + D_{n+1})], \text{ if } \tau_{n+1} + D_{n+1} < \tau_n + Y_n + v_n \qquad n = 0, 1 \ldots (2.4)$$

By combining (2.3) and (2.4) it is plain that

$$Y_{n+1} = \max\{D_{n+1}, Y_n + v_n - (\tau_{n+1} - \tau_n)\}. \qquad n = 0, 1 \ldots (2.5)$$

and (2.1) follows since $u_{n+1} = \tau_{n+1} - \tau_n$.

In order to derive (2.2), simply take note of the fact that $W_n = Y_n - D_n$ for $n = 0, 1 \ldots$. ∎

Throughout we shall assume that

(VIIa): The sequences $\{u_n\}_0^\infty$ and $\{v_n\}_0^\infty$ are iid with finite second moments and mutually independent.

For all $n = 0, 1 \ldots$, we set

$$u = I\!E(u_{n+1}) < \infty, \quad \sigma_U^2 = Var(u_{n+1}) < \infty$$

and

$$v = I\!E(v_n) < \infty, \quad \sigma_V^2 = Var(v_n) < \infty.$$

7.3 The heavy traffic limit for general resequencing systems

In this section we obtain heavy traffic diffusion limits for resequencing systems possessing a disordering system with an arbitrary structure. Disordering systems possessing specific structures are discussed in Sections 7.4–7.6.

We now consider a sequence of resequencing systems indexed by $r = 1, 2 \ldots$, each of which satisfies assumptions **(VIIa)**. Moreover assume that

(VIIb): As $r \uparrow \infty$,

$$\sigma_U(r) \to \sigma_U,$$

$$\sigma_V(r) \to \sigma_V,$$

$$[u(r) - v(r)]\sqrt{r} \to c.$$

(VIIc): For some $\epsilon > 0$,

$$\sup_r \{I\!E\{| u_1(r) |^{2+\epsilon}\}, I\!E\{v_1(r) |^{2+\epsilon}\} < \infty.$$

For $r = 1, 2 \ldots$, define the following partial sums

$$V_0(r) = 0,$$

$$V_n(r) = v_0(r) + \ldots + v_{n-1}(r), \qquad n = 1, 2 \ldots \quad (3.1)$$

and

$$U_0(r) = 0,$$

$$U_n(r) = u_0(r) + \ldots + u_{n-1}(r). \qquad n = 1, 2 \ldots \quad (3.2)$$

For $r = 1, 2 \ldots$ define the stochastic processes $\xi^j \equiv \{\xi_t^j(r), t \geq 0\}, j = 0, 1,$ with sample paths in $D[0, \infty)$ by

$$\xi_t^0(r) = \frac{U_{[rt]}(r) - u(r)[rt]}{\sqrt{r}}, \quad t \geq 0 \quad (3.3)$$

and

$$\xi_t^1(r) = \frac{V_{[rt]}(r) - v(r)[rt]}{\sqrt{r}}, \quad t \geq 0. \quad (3.4)$$

Let $\xi^j \equiv \{\xi_t^j, t \geq 0\}, j = 0, 1$, be two independent Wiener processes. Lemma 8.3.1 shows that the stochastic processes defined in (3.1)-(3.2) converge weakly to these Wiener processes.

Lemma 7.3.1. *As* $r \uparrow \infty$,

$$(\xi^0(r), \xi^1(r)) \Rightarrow (\sigma_U \xi^0, \sigma_V \xi^1) \tag{3.5}$$

in $D[0, \infty)^2$.

Proof. The proof is the same as for Lemma 2.2.1 in Chapter 2. ∎

For $r = 1, 2 \ldots$, we set

$$S_0(r) = 0$$

$$S_n(r) = V_n(r) - U_n(r), \qquad n = 1, 2 \ldots \tag{3.6}$$

and define the stochastic processes $\zeta \equiv \{\zeta_t(r), t \geq 0\}$, with sample paths in $D[0, \infty)$, by

$$\zeta_t(r) = \frac{S_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0. \tag{3.7}$$

We also define the stochastic process $\zeta \equiv \{\zeta_t, t \geq 0\}$, by

$$\zeta_t = \sigma_V \xi_t^1 - \sigma_U \xi_t^0 - ct, \quad t \geq 0. \tag{3.8}$$

Lemma 8.3.2 shows that the stochastic processes (3.7) generated by the random walk (3.6) converge weakly to ζ.

Lemma 7.3.2. *As* $r \uparrow \infty$,

$$\zeta(r) \Rightarrow \zeta \tag{3.9}$$

in $D[0, \infty)$.

Proof. The proof is the same as for Lemma 2.2.2 in Chapter 2. ∎

For $r = 1, 2 \ldots$, we define the stochastic process $\mu \equiv \{\mu_t(r), t \geq 0\}$ and $\delta \equiv \{\delta_t(r), t \geq 0\}$ with sample paths in $D[0, \infty)$ by

$$\mu_t(r) = \frac{W_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0 \tag{3.10}$$

and

$$\delta_t(r) = \frac{D_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0. \tag{3.11}$$

Theorem 7.3.1.

(a): *Assume that as* $r \uparrow \infty$,

$$D_0(r) \xrightarrow{\mathbb{D}} D_0 \tag{3.12a}$$

and

$$(\delta(r), \zeta(r)) \Rightarrow (\delta, \zeta) \tag{3.12b}$$

in $D[0,\infty)^2$. *Further assume that* $\mid c \mid < \infty$, *then*

$$\mu(r) \Rightarrow g(\zeta - \delta) \tag{3.13}$$

in $D[0,\infty)$ *as* $r \uparrow \infty$.

(b): *Assume that as* $r \uparrow \infty$,

$$(\delta(r), \xi^0(r), \xi^1(r)) \Rightarrow (\delta, \xi^0, \xi^1) \tag{3.14a}$$

in $D[0,\infty)^3$ *and*

$$u(r) \to u \text{ and } v(r) \to v \text{ with } u(r) > v(r) \text{ and } u > v, \tag{3.14b}$$

then

$$\mu(r) \Rightarrow 0 \tag{3.15}$$

in $D[0,\infty)$ *as* $r \uparrow \infty$.

Proof. We first prove Part (a). Fix $r = 1, 2 \ldots$. We can write the recursion (2.2) for the waiting time sequence as

$$\begin{aligned} W_0(r) &= 0, \\ W_{n+1}(r) &= \max\{0, W_n(r) + X_{n+1}(r)\}, \qquad n = 0, 1, \ldots \end{aligned} \tag{3.16}$$

where

$$X_{n+1}(r) = D_n(r) - D_{n+1}(r) + v_n(r) - u_{n+1}(r). \qquad n = 0, 1 \ldots \tag{3.17}$$

By successive substitutions, we obtain

$$W_n(r) = \max\{0, X_n(r), X_n(r) + X_{n-1}(r), \ldots, X_n(r) + \ldots + X_1(r)\}.$$
$$n = 0, 1 \ldots (3.18)$$

Let

$$Z_0(r) = 0,$$
$$Z_n(r) = \sum_{i=1}^{n} X_i(r). \qquad n = 1, 2 \ldots (3.19)$$

It follows that

$$W_n(r) = Z_n(r) - \min_{0 \le k \le n} Z_k(r). \qquad n = 0, 1 \ldots (3.20)$$

Note that

$$Z_n(r) = D_0(r) - D_n(r) + S_n(r). \qquad n = 0, 1 \ldots (3.21)$$

For $r = 1, 2 \ldots$ we introduce the stochastic process $\rho(r) \equiv \{\rho_t(r), t \ge 0\}$ with sample paths in $D[0, \infty)$ by

$$\rho_t(r) = \frac{Z_{[rt]}(r)}{\sqrt{r}}, \quad t \ge 0. \tag{3.22}$$

From (3.20) and (3.22) it follows that

$$\mu_t(r) = g(\rho(r))_t, \quad t \ge 0. \tag{3.23}$$

Hence by the continuous mapping theorem, in order to prove (3.13), it is sufficient to show that as $r \uparrow \infty$,

$$\rho(r) \Rightarrow \zeta - \delta \tag{3.24}$$

in $D[0, \infty)$. From (3.21), we see that

$$\rho(r) = \frac{D_0(r)}{\sqrt{r}} + \zeta(r) - \delta(r). \tag{3.25}$$

As a consequence of (3.12a), it follows that

$$\frac{D_0(r)}{\sqrt{r}} \Rightarrow 0 \text{ as } r \uparrow \infty$$

so that (3.24) follows from (3.12b), (3.25) and the converging together theorem.

We now provide a proof for Part (b), the main idea of which is borrowed from Iglehart and Whitt [30]. For $r = 1, 2 \ldots$ introduce the stochastic processes $\rho'(r) \equiv \{\rho'_t(r), t \geq 0\}$, with sample paths in $D[0, \infty)$, by

$$\rho'_t(r) = \frac{Z_{[rt]}(r) - [v(r) - u(r)][rt]}{\sqrt{r}}, \quad t \geq 0. \tag{3.26}$$

Then proceeding as in Part (a), it can be easily shown with the help of (3.14a) that

$$\rho'(r) \Rightarrow \sigma_V \xi^1 - \sigma_U \xi^0 - \delta \tag{3.27}$$

in $D[0, \infty)$ as $r \uparrow \infty$.

In order to prove that $\mu(r) \Rightarrow 0$, it is sufficient to show for each $T > 0$, that

$$\sup_{0 \leq t \leq T} \mid \mu_t(r) \mid \xrightarrow{I\!P} 0 \tag{3.28}$$

as $r \uparrow \infty$. It is intuitive to expect that (3.28) would be true, since

$$\mu_t(r) = \rho_t(r) - \inf_{0 \leq s \leq t} \rho_s(r), \quad t \geq 0$$

and as a consequence of (3.14b)

$$\lim_{r \uparrow \infty} [u(r) - v(r)] \sqrt{r} = \infty$$

so that $\rho_t(r) \downarrow -\infty$ as $r \uparrow \infty$.

Fix $T > 0$, a value d in $[0, T]$ and $0 < \epsilon < 1$. We first show that as $r \uparrow \infty$, with probability greater that $1 - \epsilon$, we have

$$\inf_{0 \leq s \leq t-d} \frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}} \geq \frac{S_{[rt]}(r) - D_{[rt]}(r)}{\sqrt{r}}, \quad 0 \leq t \leq T \tag{3.29}$$

which is a justification for the intuitive fact that $\inf_{0\le s\le t}\rho_s(r) \approx \rho_t(r)$ for a sufficiently large value of r. For $d \le t \le T$, we note that

$$\inf_{0\le s\le t-d} \frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}}$$

$$= \inf_{0\le s\le t-d} \left(\frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}} - \frac{[v(r)-u(r)][rs]}{\sqrt{r}} + \frac{[v(r)-u(r)][rs]}{\sqrt{r}} \right)$$

$$\ge \inf_{0\le s\le t-d} \left(\frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}} - \frac{[v(r)-u(r)][rs]}{\sqrt{r}} \right)$$

$$+ \frac{[v(r)-u(r)][r(t-d)]}{\sqrt{r}}$$

$$= \inf_{0\le s\le t-d} \left(\frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}} - \frac{[v(r)-u(r)][rs]}{\sqrt{r}} \right)$$

$$- \left(\frac{S_{[rt]}(r) - D_{[rt]}(r)}{\sqrt{r}} - \frac{[v(r)-u(r)][rt]}{\sqrt{r}} \right)$$

$$+ \frac{S_{[rt]}(r) - D_{[rt]}(r)}{\sqrt{r}} + \frac{[u(r)-v(r)][rd]}{\sqrt{r}}$$

$$\ge \frac{S_{[rt]}(r) - D_{[rt]}(r)}{\sqrt{r}}$$

with probability greater that $1-\epsilon$ for sufficiently large $r \ge r_0$. The second inequality follows from assumption (3.14b), while the last inequality follows from the fact that the terms in the first two brackets have weak limits while the last term blows to infinity as r increases.

As a result of (3.29), it follows that for $r \ge r_0$, with probability greater than $1-\epsilon$, we have

$$\inf_{0\le s\le t} \frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}} = \inf_{t-d\le s\le t} \frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}}$$

for a fixed value of d. Hence, for $r \ge r_0$, with probability greater that $1-\epsilon$, we see that

$$\sup_{0\le t\le T} \mid \mu_t(r) \mid$$

$$= \sup_{0\le t\le T} \mid \frac{S_{[rt]}(r) - D_{[rt]}(r)}{\sqrt{r}} - \inf_{t-d\le s\le t} \frac{S_{[rs]}(r) - D_{[rs]}(r)}{\sqrt{r}} \mid$$

$$\le \sup_{0\le s,t\le T} \sup_{|s-t|<d} \mid \frac{Z_{[rt]}(r) - Z_{[rs]}(r)}{\sqrt{r}} - \frac{[v(r) - u(r)][r(t-s)]}{\sqrt{r}} \mid$$

$$= \omega_{\rho'(r)}(d) \tag{3.30}$$

where the modulus of continuity $\omega_{\rho'(r)}$ for the process $\rho'(r)$ is given by

$$\omega_{\rho'(r)}(d) = \sup_{0\le s,t\le T} \sup_{|s-t|\le d} \mid \rho'_t(r) - \rho'_s(r) \mid, \quad d > 0.$$

Also as a result of (3.27) and of the continous mapping theorem, we get

$$\omega_{\rho'(r)}(d) \Rightarrow \omega_{\rho'}(d) \tag{3.31}$$

as $r \uparrow \infty$. Since we can make the value of d as small as we please, and since

$$\omega_{\rho'}(d) \xrightarrow{I\!P} 0 \tag{3.32}$$

as $d \downarrow 0$, it follows from (3.30) that

$$\mu(r) \Rightarrow 0 \tag{3.33}$$

in $D[0,T]$ as $r \uparrow \infty$, and this proves the theorem. ∎

Part (b) of Lemma 7.3.1 implies the surprising fact that the normalized resequencing delay of the customers will always be zero if the single server queue is operating in its stable regime, irrespective of whether the disordering system is in heavy traffic or not. However this result hinges upon the crucial condition (3.14a), that $\delta(r) \Rightarrow \delta$ in $D[0,\infty)$ as $r \uparrow \infty$. This condition is satisfied for disordering systems of the $GI/GI/K$ type as well as for infinite server disordering systems. Unfortunately it is not satisfied for disordering systems with probabilistic routing of customers. For example, as elaborated in Section 7.6, this condition is not satisfied for disordering systems which are made up of parallel queues with

Bernoulli switching of arriving customers, or disordering systems which involve probabilistic feedback from the output to the input. For such disordering systems, the conclusion of Part (b) does not hold. In fact we show in Section 7.6, that rather than going to zero, the resequencing delay constitutes the major portion of the total delay of such systems, in heavy traffic.

7.4 The heavy traffic limit for infinite server systems

In this section we specialize the results of Theorem 7.3.1 to the case when the disordering system is an infinite server queue. This queueing system was first analyzed in detail by Baccelli, Gelenbe and Plateau [1], and consequently we will refer to it in the remainder of this dissertation as the BGP model. We shall assume that

(VIId): The sequences $\{u_n\}_0^\infty$, $\{D_n\}_0^\infty$ and $\{v_n\}_0^\infty$ are iid with finite second moments and independent.

For all $n = 0, 1 \ldots$, we set

$$u = I\!E(u_n) < \infty, \quad \sigma_U^2 = Var(u_n) < \infty,$$

$$v = I\!E(v_n) < \infty, \quad \sigma_V^2 = Var(v_n) < \infty$$

and

$$d = I\!E(D_n) < \infty, \quad \sigma_D^2 = Var(D_n) < \infty.$$

We now consider a sequence of resequencing systems indexed by $r = 1, 2 \ldots$, each of which satisfies assumption **(VIId)**. Moreover assume that

(VIIe): As $r \uparrow \infty$,

$$\sigma_U(r) \to \sigma_U,$$

$$\sigma_V(r) \to \sigma_V,$$

$$\sigma_D(r) \to \sigma_D,$$

$$[u(r) - v(r)]\sqrt{r} \to c.$$

(VIIf): For some $\epsilon > 0$,

$$\sup_r \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{v_1(r) \mid^{2+\epsilon} I\!E\{\mid D_1(r) \mid^{2+\epsilon}\}\} < \infty$$

Under these assumptions, Theorem 7.3.1 immediately yields the following corollary.

Corollary 7.4.1. *As* $r \uparrow \infty$

$$\mu(r) \Rightarrow g(\zeta) \tag{4.1}$$

in $D[0,\infty)$.

Proof. This follows from Part (a) of Theorem 7.3.1 owing to the fact that $\delta(r) \Rightarrow 0$ in $D[0,\infty)$ as $r \uparrow \infty$, since

$$\delta_t(r) = \frac{D_{[rt]}(r)}{\sqrt{r}}.$$

∎

For infinite server disordering systems the sequence $\{W_n\}_0^\infty$ has the same traffic limit as the sequence of waiting times in an ordinary single server queue. This means that asymptotically resequencing as a negligible effect on the operation of the single server queue in heavy traffic. This result is surprising if seen from the following viewpoint: Kingman [38] has shown that the diffusion limit for a single server queue depends on the particular discipline chosen to serve the customers, for example it is different if the customers are served in LCFS order rather than in FCFS order. Resequencing may be viewed as a special type of service discipline (if the resequencing buffer and the single server queue buffer are regarded as a single buffer of an equivalent single server), because customers are served in the order in which they entered the infinite server queue, rather than the order in which they enter the equivalent buffer. Also note that this effect remains unchanged in heavy traffic. Hence in this case even though we change the service discipline of the single server queue, we nevertheless obtain the same diffusion limit.

7.4.1 Generalization to a tandem system

We now proceed to extend the result of the previous sub–section to the case when there are an arbitrary number of disordering and resequencing systems preceding the single server queue. The system under consideration operates as follows: Each customer is disordered by an infinite server queue and resequenced K successive times before it enters the buffer of a single server queue. After getting served there, it leaves the system (Fig 7.2). Let the sequences $\{u_n\}_0^\infty$ and $\{v_n\}_0^\infty$ be defined as before, and for each $1 \leq k \leq K$ and $n = 0, 1 \ldots$, define the following,

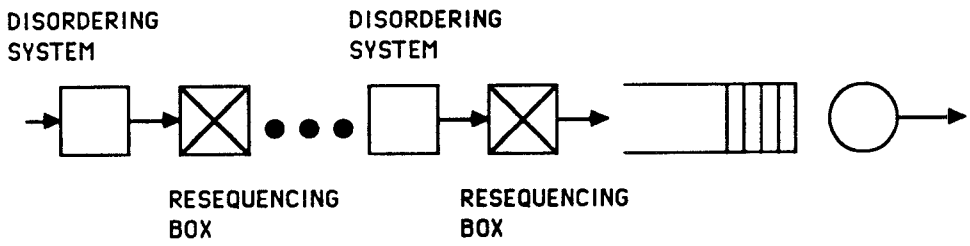


Fig. 7.2. A multi–stage resequencing model.

D_n^k : Delay of the n^{th} customer at the k^{th} disordering system;

W_n^k : For $1 \leq k \leq K-1$, this RV represents the delay of the n^{th} customer in the k^{th} resequencing box;

W_n : Delay of the n^{th} customer in the K^{th} resequencing box plus the delay in the buffer of the single server queue; and

D_n : $= D_n^1 + W_n^1 + \ldots + D_n^{K-1} + W_n^{K-1} + D_n^K$, i.e., the total disordering delay of the n^{th} customer, before it is resequenced and sent to the buffer of the single server queue.

We shall assume that

(VIIg): The sequences $\{v_n\}_0^\infty$, $\{u_n\}_0^\infty$ and $\{D_n^k\}_0^\infty, 1 \leq k \leq K$, are iid with finite second moments and mutually independent.

For all $n = 0, 1 \ldots$, we set

$$u = I\!E(u_n) < \infty, \quad \sigma_U^2 = Var(u_n) < \infty$$

$$v = I\!E(v_n) < \infty, \quad \sigma_V^2 = Var(v_n) < \infty$$

and

$$d_k = I\!E(D_n^k) < \infty, \quad \sigma_k^2 = Var(D_n^k) < \infty, \quad 1 \leq k \leq K.$$

Now consider a sequence of resequencing systems indexed by $r = 1, 2 \ldots$ each of which satisfies assumptions **(VIIg)**. Moreover assume that

(VIIh): As $r \uparrow \infty$,

$$\sigma_U(r) \to \sigma_U,$$

$$\sigma_V(r) \to \sigma_V,$$

$$\sigma_k(r) \to \sigma_k, \quad 1 \le k \le K$$

$$[u(r) - v(r)]\sqrt{r} \to c.$$

(VIIi): For some $\epsilon > 0$,

$$\sup_{r,k} \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{v_1(r) \mid^{2+\epsilon}\} I\!E\{\mid D_1^k(r) \mid^{2+\epsilon}\}\} < \infty$$

Define the partial sums $\{V_n\}_0^\infty$, $\{U_n\}_0^\infty$, $\{X_n\}_0^\infty$ and $\{Z_n\}_0^\infty$ as in (3.1), (3.2), (3.17) and (3.19) respectively. Also define the stochastic processes $\xi^0(r), \xi^1(r)$, $\zeta(r)$, $\mu(r)$, $\delta(r)$ and $\rho(r)$ as in (3.1), (3.4), (3.7), (3.10), (3.11) and (3.22). It is easy to see that Lemma 7.3.1 and Lemma 7.3.2 continue to hold for this model.

For $r = 1, 2 \ldots$, define the stochastic processes $\mu^k \equiv \{\mu_t^k, t \ge 0\}, 1 \le k \le K-1$ with sample paths in $D[0, \infty)$ by

$$\mu_t^k(r) = \frac{W_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \le k \le K-1. \tag{4.2}$$

Theorem 7.4.1. *As* $r \uparrow \infty$

$$\delta(r) \Rightarrow 0 \tag{4.3}$$

in $D[0, \infty)$.

Proof. Fix $r = 1, 2 \ldots$. Note that

$$\delta_t(r) = \frac{D_{[rt]}^1}{\sqrt{r}} + \mu_t^1(r) + \ldots + \frac{D_{[rt]}^{K-1}}{\sqrt{r}} + \mu_t^{K-1}(r) + \frac{D_{[rt]}^K}{\sqrt{r}}, \quad t \ge 0. \tag{4.4}$$

Hence in order to prove (4.3), it is sufficient to show that as $r \uparrow \infty$

$$(\mu^1(r), \ldots, \mu^{K-1}(r)) \Rightarrow (0, \ldots, 0) \tag{4.5}$$

in $D[0,\infty)^{K-1}$.

We propose to prove (4.5) by induction on the number of levels in the system. We first show that as $r \uparrow \infty$

$$\mu^1(r) \Rightarrow 0 \tag{4.6}$$

in $D[0,\infty)$.

Note that

$$\begin{aligned} W^1_{n+1}(r) &= \max\{0, W^1_n(r) + D^1_n(r) - D^1_{n+1}(r) - -u_{n+1}(r)\} \\ &= Z^1_{n+1}(r) - \min_{0\le i\le n+1} Z^1_i(r), \qquad n = 0,1\ldots \end{aligned} \tag{4.7}$$

where

$$Z^1_n(r) = D^1_0(r) - D^1_n(r) - u_1(r) - \ldots - u_n(r). \qquad n = 0,1\ldots \tag{4.8}$$

From (4.7)–(4.8) it follows that

$$\mu^1_t(r) = \frac{D^1_0(r)}{\sqrt{r}} - \frac{D^1_{[rt]}(r)}{\sqrt{r}} - \frac{U_{[rt]}(r)}{\sqrt{r}} - \inf_{0\le s\le t}\{\frac{D^1_0(r)}{\sqrt{r}} - \frac{D^1_{[rs]}(r)}{\sqrt{r}} - \frac{U_{[rs]}(r)}{\sqrt{r}}\}, \quad t \ge 0 \tag{4.9}$$

For $r = 1,2\ldots$, define the stochastic processes $\hat{\mu}^1(r) \equiv \{\hat{\mu}^1_t(r), t \ge 0\}$ in $D[0,\infty)$ by

$$\hat{\mu}^1_t(r) = -\frac{U_{[rt]}(r)}{\sqrt{r}} - \inf_{0\le s\le t}\{-\frac{U_{[rs]}(r)}{\sqrt{r}}\}, \quad t \ge 0 \tag{4.10}$$

and note that $\mu^1(r)$ and $\hat{\mu}^1(r)$ both have the same limit due to the converging together theorem. Note that in this case we do not require an additional condition such as (3.12a) to conclude that $\frac{D^1_0(r)}{\sqrt{r}}$ or $\frac{D^1_{[rt]}(r)}{\sqrt{r}}$ converges to zero, since by assumption they form sequences of iid RVs.

Hence in order to prove (4.6), it is sufficient to show that $\hat{\mu}^1(r) \Rightarrow 0$ as $r \uparrow \infty$. In order to prove this, it is sufficient to show that

$$\sup_{0\le t\le 1} \mid \hat{\mu}^1_t(r) \mid \xrightarrow{I\!P} 0 \tag{4.11}$$

as $r \uparrow \infty$. The proof for (4.11) is similar to the proof given for Part (b) of Theorem 7.3.1, and is therefore omitted.

As the induction step, assume that as $r \uparrow \infty$

$$(\mu^1(r), \ldots, \mu^k(r)) \Rightarrow (0, \ldots, 0) \tag{4.12}$$

in $D[0,\infty)^k$, for some $2 \leq k \leq K-2$. We shall show that as $r \uparrow \infty$

$$(\mu^1(r), \ldots, \mu^{k+1}(r)) \Rightarrow (0, \ldots, 0) \tag{4.13}$$

in $D[0,\infty)^{k+1}$.

For $1 \leq k \leq K-1$ and $r = 1, 2 \ldots$, define the RVs $T_n^k(r)$ by

$$T_n^k(r) = D_n^1(r) + W_n^1(r) + \ldots + D_n^k(r) + W_n^k(r)$$

Note that as a consequence of (4.12), it follows that as $r \uparrow \infty$,

$$\frac{T_{[r.]}^k(r)}{\sqrt{r}} \Rightarrow 0 \tag{4.14}$$

in $D[0,\infty)$.

Note that

$$\begin{aligned} W_{n+1}^{k+1}(r) &= \max\{0, W_n^{k+1}(r) + T_n^k(r) + D_n^{k+1}(r) - T_{n+1}^k(r) - D_{n+1}^{k+1}(r) - u_{n+1}(r)\} \\ &= Z_{n+1}^{k+1}(r) - \min_{0 \leq i \leq n+1} Z_i^{k+1}(r), \qquad n = 0, 1 \ldots \end{aligned} \tag{4.15}$$

where

$$Z_n^{k+1}(r) = T_0^k(r) + D_0^{k+1}(r) - T_n^k(r) - D_n^{k+1}(r) - u_1(r) - \ldots - u_n(r) \qquad n = 0, 1 \ldots \tag{4.16}$$

From (4.15)–(4.16) it follows that

$$\begin{aligned} \mu_t^{k+1}(r) = &\frac{T_0^1(r)}{\sqrt{r}} + \frac{D_0^{k+1}(r)}{\sqrt{r}} - \frac{T_{[rt]}^k(r)}{\sqrt{r}} - \frac{D_{[rt]}^{k+1}(r)}{\sqrt{r}} - \frac{U_{[rt]}(r)}{\sqrt{r}} \\ &- \inf_{0 \leq s \leq t} \{\frac{T_0^1(r)}{\sqrt{r}} + \frac{D_0^{k+1}(r)}{\sqrt{r}} - \frac{T_{[rs]}^k(r)}{\sqrt{r}} - \frac{D_{[rs]}^{k+1}(r)}{\sqrt{r}} - \frac{U_{[rs]}(r)}{\sqrt{r}}\} \end{aligned} \tag{4.17}$$

From equation (4.14) (which is a consequence of the induction hypothesis (4.12)) and (4.17), and the fact that $T_n^k(r)$ is independent of $D_n^{k+1}(r)$, it is clear that that (4.13) holds, and this completes the proof.

∎

Equation (4.3) in combination with Part (a) of Theorem 7.3.1, implies that

Theorem 7.4.2. *As* $r \uparrow \infty$

$$\mu(r) \Rightarrow g(\zeta) \tag{4.18}$$

in $D[0, \infty)$.

7.5 The heavy traffic limit for finite server resequencing systems: Multiserver disordering systems

The resequencing systems analyzed in the next two sections deviate slightly from the general model introduced in Section 7.2 due to the fact that the single server queue after the resequencing buffer is omitted from the system. In this section we shall consider the case when the disordering system is a $GI/GI/K$ queue, while in Section 7.6 we shall consider the case when the disordering system consists of K single server queues operating in parallel.

The resequencing system under consideration operates as follows: Customers enter a $GI/GI/K$ queue, after obtaining service from which they are resequenced in a resequencing buffer and leave the system. Our basic heavy traffic result about this system is stated next.

Theorem 7.5.1 *The end–to–end delay in the $GI/GI/K$ resequencing system has the same heavy traffic limit as the end–to–end delay of a $GI/GI/K$ queue.*

Proof. This system satisfies the conditions in Part (b) of Theorem 7.3.1, so that the conclusion is a direct consequence of (3.15).

∎

Let the average end–to–end delay for the system be denoted by $\overline{T}_K(\lambda)$. Then Theorem 7.5.1 and results regarding heavy traffic limits for $GI/GI/K$ queues in Kollerstrom [44] imply that

$$\lim_{\lambda \uparrow K\mu} (K\mu - \lambda)\overline{T}_K(\lambda) = [\sigma_U^2 + \frac{\sigma_V^2}{K^2}]\frac{K^2\mu^2}{2} \qquad K = 2, 3 \ldots (5.1)$$

where λ, σ_U and μ, σ_V are the rates and variances of the arrival and service processes respectively.

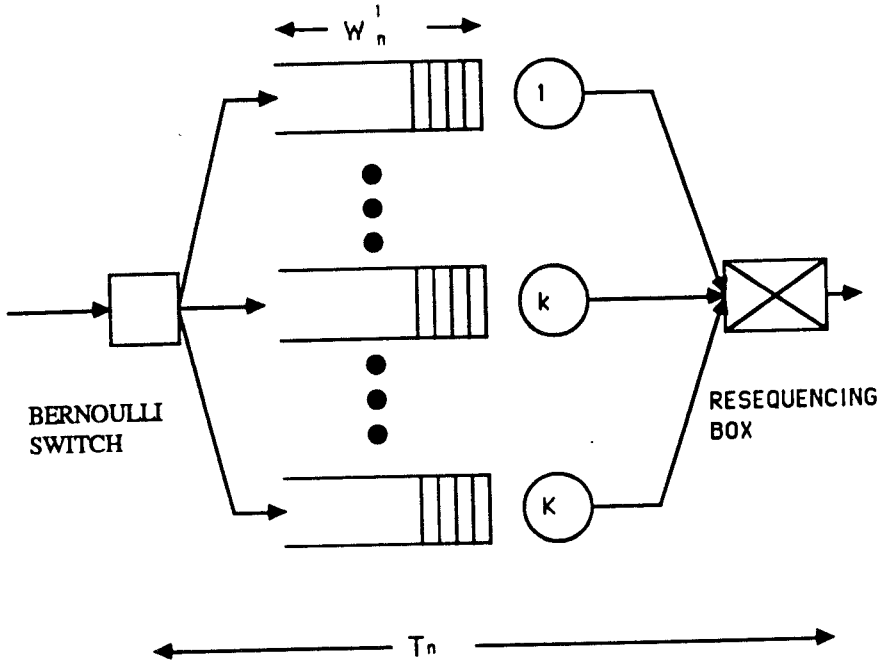


Fig. 7.3. Parallel queues with resequencing.

7.6 The heavy traffic limit for finite server resequencing systems: Disordering due to parallel queues

The system discussed here has a disordering system composed of K parallel single server queues. We assume that customers are routed to the different queues according to a Bernoulli switch with switching probability $p_k, 1 \leq k \leq K$. After receiving service, they are resequenced in a resequencing buffer before leaving the system.

This system was analyzed by Gün and Jean-Marie [17] when the arrival process into the system is Poissonian. They gave a complicated expression for the average end–to–end delay involving the virtual waiting time in the system. However, since for most systems it is difficult to obtain a formula for the virtual waiting time, we expect that the limit theorem approximations to be of practical computational value.

The following RVs are defined on a common probability space $(\Omega, I\!F, I\!P)$. For $n = 0, 1 \ldots$ and $1 \leq k \leq K$,

u_{n+1} : Inter–arrival time between the $(n+1)^{rst}$ and n^{th} customers.

v_n^k : Service time of the n^{th} customer to enter the system, if it were to join the k^{th} queue.

a_n^k : This is a $\{0,1\}$–valued RV, with $a_n^k = 1$ if the n^{th} customer joins the k^{th} queue.

W_n^k : Waiting time of the n^{th} customer to enter the system, if it were to join the k^{th} queue.

T_n : End–to–end delay of the n^{th} customer to enter the system (including the resequencing delay).

We shall assume that

(VIIj): The sequences $\{u_{n+1}\}_0^\infty$, $\{a_n^k\}_0^\infty$ and $\{v_n^k\}_0^\infty, 1 \leq k \leq K$, are iid with finite second moments, and mutually independent.

For $n = 0, 1 \ldots$, we set

$$I\!P(a_n^k = 1) = p_k,$$

$$u = I\!E(u_{n+1}) < \infty, \quad \sigma_0^2 = Var(u_{n+1}) < \infty$$

and

$$v_k = I\!E(v_n^k) < \infty, \quad \sigma_k^2 = Var(v_n^k) < \infty, \quad 1 \leq k \leq K$$

7.6.1 Recursive representation for the delays

The delays in the system obey the recursions given in Lemma 7.2.1 with D_n replaced by the reponse time of the n^{th} customer in the system of parallel queues. However we give another set of recursions for the system which have the advantage of facilitating the proof of the heavy traffic limit theorems. Assuming that the initial customer arrives into an empty system at time $t = 0$, it is easy to see that for each $1 \leq k \leq K$,

$$W_0^k = 0$$

$$W_{n+1}^k = \max\{0, W_n^k + a_n^k v_n^k - u_{n+1}\}. \qquad n = 0, 1 \ldots (6.1)$$

The end–to–end delay $T_n, 1 \leq k \leq K$, is given by

$$T_n = \max_{1 \leq k \leq K} \{W_n^k + a_n^k \sigma_n^k\}. \qquad n = 0, 1 \ldots (6.2)$$

It is well known [1] that the stability condition of a system with resequencing is the same as the system without resequencing. Therefore the system is stable iff each queue is stable, i.e.,

$$p_k v_k < u, \quad 1 \leq k \leq K. \qquad (6.3)$$

7.6.2 The diffusion limit

We now proceed with the task of obtaining heavy traffic diffusion limits for the delay processes in the resequencing system. We consider a sequence of resequencing systems indexed by $r = 1, 2 \ldots$, each of which satisfies assumption **(VIIj)**. We make the following additional assumptions **(VIIk)**–**(VIIl)**, where

(VIIk): As $r \uparrow \infty$,

$$\begin{aligned} \sigma_k(r) &\to \sigma_k, \quad 0 \leq k \leq K, \\ p_k(r) &\to p_k, \quad 1 \leq k \leq K, \\ v_k(r) &\to v_k, \quad 1 \leq k \leq K, \\ [u(r) - p_k(r) v_k(r)]\sqrt{r} &\to c_k, \quad 1 \leq k \leq K. \end{aligned}$$

(VIIl): For some $\epsilon > 0$,

$$\sup_{r,k} \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{\mid v_1^k(r) \mid^{2+\epsilon}\}\} < \infty.$$

For $r = 1, 2 \ldots$, define the following partial sums

$$\begin{aligned} V_0^k(r) &= 0, \\ V_n^k(r) &= a_0^k(r) v_0^k(r) + \ldots + a_{n-1}^k(r) v_{n-1}^k(r), \quad 1 \leq k \leq K, \; n = 1, 2 \ldots (6.4) \\ U_0(r) &= 0, \\ U_n(r) &= u_1(r) + \ldots + u_n(r). \qquad n = 1, 2 \ldots (6.5) \end{aligned}$$

and

$$S_0^k(r) = 0$$

$$S_n^k(r) = V_n^k(r) - U_n(r),\ 1 \le k \le K \qquad n = 1, 2 \ldots (6.6)$$

We also define the stochastic processes $\zeta^k(r) \equiv \{\zeta_t^k(r), t \ge 0\}, 1 \le k \le K$, with sample paths in $D[0,\infty)$ by

$$\zeta_t^k(r) = \frac{S_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \le k \le K, \quad t \ge 0. \tag{6.7}$$

With $\xi \equiv \{\xi_t^k, t \ge 0\}, 1 \le k \le K$, as K independent Wiener processes, we define the stochastic processes $\zeta^k \equiv \{\zeta_t^k, t \ge 0\}, 1 \le k \le K$, by

$$\zeta_t^k = \sum_{j=1}^{K} Q_{kj}\xi_t^j - c_k t, \quad 1 \le k \le K, \quad t \ge 0 \tag{6.8}$$

where the matrix $Q \equiv \{Q_{ij}\}_{i,j=1}^K$ is such that the covariance matrix R for the diffusion is given by

$$R = QQ^T$$

$$= \begin{pmatrix} \sigma_0^2 + p_1\sigma_1^2 + p_1\overline{p}_1 v_1^2 & \sigma_0^2 - p_1 p_2 v_1 v_2 & \ldots & \sigma_0^2 - p_1 p_K v_1 v_K \\ \sigma_0^2 - p_2 p_1 v_2 v_1 & \sigma_0^2 + p_2\sigma_2^2 + p_2\overline{p}_2 v_2^2 & \ldots & \sigma_0^2 - p_2 p_K v_2 v_K \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_0^2 - p_K p_1 v_K v_1 & \sigma_0^2 - p_K p_2 v_K v_2 & \ldots & \sigma_0^2 + p_K\sigma_K^2 + p_K\overline{p}_K v_K^2 \end{pmatrix}. \tag{6.9}$$

with $\overline{p}_k = 1 - p_k, 1 \le k \le K$. The process $(\zeta^1, \ldots, \zeta^K)$ is thus a K–dimensional diffusion process with drift vector $c = (-c_1, \ldots, -c_K)$ and covariance matrix R.

Theorem 7.6.2 shows that the stochastic processes (6.7) generated by the random walk (6.6) converge weakly to $(\zeta^1, \ldots, \zeta^K)$.

Theorem 7.6.2. *As* $r \uparrow \infty$,

$$(\zeta^1(r), \ldots, \zeta^K(r)) \Rightarrow (\zeta^1, \ldots, \zeta^K) \tag{6.10}$$

in $D[0,\infty)^K$.

Before providing a proof of Theorem 7.6.2, we present the following two corollaries.

For $r = 1, 2 \ldots$ and $1 \leq k \leq K$, observe that

$$\begin{aligned} W_n^k(r) &= \max\{S_n^k(r) - S_i^k(r) : i = 0, 1 \ldots, n\} \\ &= S_n^k(r) - \min\{S_i^k(r) : i = 0, 1 \ldots, n\}, \ n = 0, 1 \ldots \end{aligned} \tag{6.11}$$

For $r = 1, 2 \ldots$, we now define the stochastic processes $\mu^k(r) \equiv \{\mu_t^k(r), t \geq 0\}, 1 \leq k \leq K$, with sample paths in $D[0, \infty)$ by

$$\mu_t^k(r) = \frac{W_{[rt]}^k(r)}{\sqrt{r}}, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{6.12}$$

We also define the stochastic processes $\mu^k \equiv \{\mu_t^k, t \geq 0\}, 1 \leq k \leq K$, by

$$\mu_t^k = g(\zeta^k)_t, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{6.13}$$

In Corollary 7.6.1 we show that the vector process associated with (6.12), converges weakly to a K–dimensional diffusion process (6.13) with drift c and covariance (6.9). This limiting diffusion stays in the non–negative orthant of the K–dimensional space and exhibits normal reflections at the boundaries.

Corollary 7.6.1. *As* $r \uparrow \infty$,

$$(\mu^1(r), \ldots, \mu^K(r)) \Rightarrow (\mu^1, \ldots, \mu^K) \tag{6.14}$$

in $D[0, \infty)^K$.

Proof. From (6.11) and (6.12), we conclude for each $r = 1, 2, \ldots$, that

$$\mu^k(r) = g(\zeta^k(r)), \quad 1 \leq k \leq K.$$

and the result follows by the continuous mapping theorem and Theorem 7.6.2.

∎

For $r = 1, 2 \ldots$, define the stochastic processes $\kappa(r) \equiv \{\kappa_t(r), t \geq 0\}$ with sample paths in $D[0, \infty)$ by

$$\kappa_t(r) = \frac{T_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0. \tag{6.15}$$

Also define the stochastic process $\kappa \equiv \{\kappa_t, t \geq 0\}$ with sample paths in $D[0, \infty)$ by

$$\kappa_t = \max_{1 \leq k \leq K} \mu_t^k, \quad t \geq 0. \tag{6.16}$$

In Corollary 7.6.2 we show that the stochastic process (6.15) generated by the end-to-end delays, converges weakly to the process (6.16), which is the maximum of K correlated Wiener processes with drift, in the non–negative orthant and normal reflection at the boundaries.

Corollary 7.6.2. *As* $r \uparrow \infty$,

$$\kappa(r) \Rightarrow \kappa \tag{6.17}$$

in $D[0, \infty)$.

Proof. From (6.2), (6.15) and (6.16) we conclude for each $r = 1, 2 \ldots$, that

$$\kappa_t(r) = \max_{1 \leq k \leq K} \{\mu_t^k(r) + \frac{a_{[rt]}^k v_{[rt]}^k}{\sqrt{r}}\}, \quad t \geq 0.$$

Equation (6.17) now follows from Corollory 6.7.1 by the continous mapping theorem and the converging together theorem ∎

We now proceed with the proof of Theorem 7.6.2.

Proof. We write (6.7) as

$$\zeta_t^k(r) = \frac{S_{[rt]}^k(r) - (p_k(r)v_k(r) - u(r))[rt]}{\sqrt{r}} + (p_k(r)v_k(r) - u(r))\frac{[rt]}{\sqrt{r}},$$

$$1 \leq k \leq K, \quad t \geq 0 \tag{6.18}$$

As a result of assumption **(VIIk)**, we have

$$\lim_{r\uparrow\infty}(p_k(r)v_k(r)-u(r))\frac{[rt]}{\sqrt{r}}=-c_kt. \tag{6.19}$$

By a multi–dimensional version of Prohorov's theorem, it follows that as $r\uparrow\infty$,

$$\left(\frac{S^1_{[rt]}(r)-(p_1(r)v_1(r)-u(r))[rt]}{\sqrt{r}},\ldots,\frac{S^K_{[rt]}(r)-(p_K(r)v_K(r)-u(r))[rt]}{\sqrt{r}}\right)$$

$$\Rightarrow(\sum_{k=1}^{K}Q_{1k}\xi^1_t,\ldots,\sum_{j=k}^{K}Q_{Kk}\xi^K_t) \tag{6.20}$$

in $D[0,\infty)^K$, and it now remains for us to identify the components of the matrix Q. This can be done by observing that for each $t\geq 0$

$$\begin{aligned}(QQ^T)_{ij}t=&R_{ij}t\\ &=\lim_{r\uparrow\infty}I\!E\left(\frac{S^i_{[rt]}(r)-(p_i(r)v_i(r)-u(r))[rt]}{\sqrt{r}}\right)\\ &\quad\left(\frac{S^j_{[rt]}(r)-(p_j(r)v_j(r)-u(r))[rt]}{\sqrt{r}}\right).\end{aligned} \tag{6.21}$$

A straightforward computation of the right hand side in (6.21) leads to the conclusion that

$$R_{ij}=\begin{cases}\sigma_0^2+p_i\sigma_i^2+p_i(1-p_i)v_i^2, & \text{if } i=j\\ \sigma_0^2-p_ip_jv_iv_j, & \text{if } i\neq j\end{cases}$$

which proves the theorem. ∎

Recall that in Chapter I we made a distinction between the asymptotic distributions obtained depending upon the order in which the limits for r and t were taken for the single server queue waiting time process. In the next result we show that for the parallel queue resequencing system, the stationary distribution for the normalized vector of response times is the same, regardless of the order in

which these limits are taken. We also provide a sufficient condition for the vector diffusion process $(\mu^1, \ldots, \mu^K)$ to have a stationary distribution. Since the proof is exactly the same as for Proposition 2.2.1 and Theorem 2.2.2 in Chapter 2, it is omitted.

Theorem 7.6.3

(a): *The condition $c_k > 0, 1 \leq k \leq K$, is necessary and sufficient to ensure that the K- dimensional process $(\mu^1, \ldots, \mu^K)$ converges in distribution to a proper vector $(\mu^1_\infty, \ldots, \mu^K_\infty)$ as $t \uparrow \infty$.*

(b): *Denote by $(\hat{\mu}^1_\infty, \ldots, \hat{\mu}^K_\infty)$ the limiting process obtained from $(\mu^1_t(r), \ldots, \mu^K_t(r))$ by taking the limit in distribution as $t \uparrow \infty$ and then as $r \uparrow \infty$. Under the conditions $c_k > 0, 1 \leq k \leq K$ and $p_k(r)v_k(r) < u(r), 1 \leq k \leq K, r = 1, 2 \ldots$,the equality*

$$(\mu^1_\infty, \ldots, \mu^K_\infty) =_{st} (\hat{\mu}^1_\infty, \ldots, \hat{\mu}^K_\infty) \tag{6.22}$$

holds.

7.6.3 Homogeneous queues with Poisson arrivals

Consider the case when each queue has identical parameters so that $v = v_k$, $\sigma = \sigma_k$ and $c = c_k, 1 \leq k \leq K$. Further assume that $p_k = \frac{1}{K}, 1 \leq k \leq K$. In this homogeneous case, we have

$$R_{ij} = \begin{cases} \sigma_0^2 + \frac{\sigma^2}{K} + \frac{1}{K}(1 - \frac{1}{K})v^2, & \text{if } i = j \\ \sigma_0^2 - (\frac{v}{K})^2, & \text{if } i \neq j. \end{cases}$$

Therefore under the the assumption

$$\sigma_0^2 = \frac{v^2}{K^2} \tag{6.23}$$

the cross–corelation terms in R cancel and we get

$$R = \begin{pmatrix} \sigma_0^2 + \frac{\sigma^2}{K} + \frac{K-1}{K^2}v^2 & 0 & \ldots & 0 \\ 0 & \sigma_0^2 + \frac{\sigma^2}{K} + \frac{K-1}{K^2}v^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_0^2 + \frac{\sigma^2}{K} + \frac{K-1}{K^2}v^2 \end{pmatrix}. \tag{6.24}$$

One of the most common inter–arrival distributions that satisfies (6.23) is the exponential, since in this case

$$\sigma_0^2 = \lim_{\lambda \uparrow K\mu} \frac{1}{\lambda^2} = \frac{1}{K^2\mu^2}$$

where as usual, we have set $v = \frac{1}{\mu}$. Making these substitutions in (6.24), we obtain

$$R = \begin{pmatrix} \frac{1}{K}(\sigma^2 + \frac{1}{\mu^2}) & 0 & \dots & 0 \\ 0 & \frac{1}{K}(\sigma^2 + \frac{1}{\mu^2}) & \dots & v^2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{K}(\sigma^2 + \frac{1}{\mu^2}) \end{pmatrix}. \tag{6.25}$$

Thus under the condition that the arrivals are Poisson and $p_k = \frac{1}{K}, 1 \leq k \leq K$, (6.8) simplifies to

$$\zeta_t^k = \sqrt{\frac{1}{K}(\sigma^2 + \frac{1}{\mu^2})}\xi_t^k - ct, \quad 1 \leq k \leq K, \quad t \geq 0. \tag{6.26}$$

Note that now the stochastic processes $\zeta^k, 1 \leq k \leq K$, are independent, so that we have reduced the diffusion to a form from which it is easy to obtain the stationary distribution. Carrying out the calculations for the case $c > 0$ as in Section 3.5.2, it can be shown that the RV κ_t converges in distribution to a RV κ_∞, such that

$$I\!E\kappa_\infty = (\sigma^2 + \frac{1}{\mu^2})\frac{H_K}{2Kc} \tag{6.27}$$

where H_K as usual is the Harmonic series. In general, the n^{th} moment is given by

$$I\!E\kappa_\infty^n = n!\left[(\sigma^2 + \frac{1}{\mu^2})\frac{1}{2Kc}\right]^n \sum_{k=1}^{K} \binom{K}{k} \frac{(-1)^{k+1}}{k^n}. \tag{6.28}$$

Let us denote the average end–to–end delay of a K–dimensional resequencing system with Poisson arrivals by $\overline{T}_K(\lambda)$ and its n^{th} moment by $\overline{T}_K^{(n)}(\lambda)$. Owing to (6.27)–(6.28) and Theorem 7.6.3, it can be shown that

$$\lim_{\lambda \uparrow K\mu} (K\mu - \lambda)\overline{T}_K(\lambda) = (\sigma^2 + \frac{1}{\mu^2})\frac{KH_K\mu^2}{2} \tag{6.29}$$

and

$$\lim_{\lambda \uparrow K\mu} (K\mu - \lambda)^n \overline{T}_K^{(n)}(\lambda) = n! \left[(\sigma^2 + \frac{1}{\mu^2}) \frac{K\mu^2}{2} \right]^n \sum_{k=1}^{K} \binom{K}{k} \frac{(-1)^{k+1}}{k^n}. \quad (6.30)$$

Note that in this case, in contrast to the case of disordering by $GI/GI/K$ queues, the resequencing delay grows logarithmically with K.

It is interesting to contrast the behavior of parallel queues operating under the fork–join and resequencing constraints. From (3.5.31) and (6.27) we come to the conclusion that the average response time in heavy traffic for these queues under both the synchronization constraints varies logarithmically with the number of queues K. However in light traffic, the fork–join synchronization still leads to logarithmic increase of the average response time with K (see Chapter 6), while the resequencing synchronization leads to a constant light traffic limit for the average response time, i.e. $\overline{T}(0) = \frac{1}{\mu}$. From this we come to the conclusion that while the fork–join constraint leads to an equal degradation of the average response time in both light and heavy traffic, the resequencing constraint becomes important in the calculation of the average response time only when the system is heavily loaded and the number of parallel queues K is large.

Before closing this chapter, we would like to give another example of a system in which the resequencing constraint leads to a significant degradation of the average response time in heavy traffic. The disordering system is composed of a single server queue with feedback in which the customers may be routed back to the end of the queue with probability q, or they may enter the resequencing box with probability p after receiving service. We have been unable to carry out a heavy traffic analysis of this system, however an exact analysis was carried out by Horlatt and Mailles [27] for the special case when the inter-arrival and service times are exponential with rate λ and μ respectively. They showed that the average end-to-end delay $\overline{T}(\rho)$ as a function of $\rho = \frac{\lambda}{p\mu}$ satisfies

$$\overline{T}(\rho) = \frac{p}{\mu} \sum_{k=1}^{\infty} \frac{k q^{k-1}}{[1 - \rho(1 - q^k)][1 - \rho(1 - q^{k-1})]}. \quad (6.31)$$

From (6.31) it is possible to obtain the following heavy traffic limit

$$\lim_{\rho \uparrow 1}(1-\rho)^2 \overline{T}(\rho) = \sum_{k=1}^{\infty} \frac{k q^{k-1}}{(1+q^k+q^{k-1}-q^{2k-1})}. \tag{6.32}$$

Hence as a result of the resequencing constraint, the average response time grows at rate $\frac{1}{(1-\rho)^2}$ as $\rho \uparrow 1$, rather than at rate $\frac{1}{(1-\rho)}$.

CHAPTER VIII

8.1 Introduction

In the last chapter we developed heavy traffic approximations for a variety of queueing systems with resequencing. In the present chapter our objective is to provide polynomial approximations for some of those models. To that end we calculate the light traffic limits using the Reiman-Simon theory and combine them with the heavy traffic limits of Chapter 7.

In Section 8.2 we obtain polynomial approximations for the BGP model discussed in Section 7.4. An earlier analysis of this model using complex analytic methods [1] yielded a complicated expression for the Laplace transform of the end–to–end delay, from which it was very difficult to obtain explicit formulas. However using our methods, we obtain a quadratic approximation to the average waiting time which agrees extremely well with simulation results. In Sections 8.3 and 8.4 we obtain polynomial approximations for the resequencing model from Section 7.6, in which the disordering is due to K single server queues operating in parallel. This model was analyzed by Gün and Jean–Marie [17], who gave an expresion for the average end–to–end delay for the case when the arrivals are Poisson. However, this expression is difficult to evaluate except in the case when the services are exponential. Using, heavy and light traffic theory, we give simple but good approximation for non–exponential service times, such as deterministic service times (in Section 8.3) and r^{th} order Erlangian service times (in Section 8.4). Lastly in Section 8.5 we obtain polynomial approximations for a generalized BGP model, in which the disordering system, which is composed of two single server queues operating in parallel, is followed by a single server queue.

Before we can obtain the light traffic limits for the various performance measures that we are interested in, we have to verify that these measures are admissible

in the sense of Definition B1 (Appendix B). This verification was carried out for the case of the fork–join queue in Chapter 6, and since the procedure for doing so for queues with resequencing is very similar, we have omitted it.

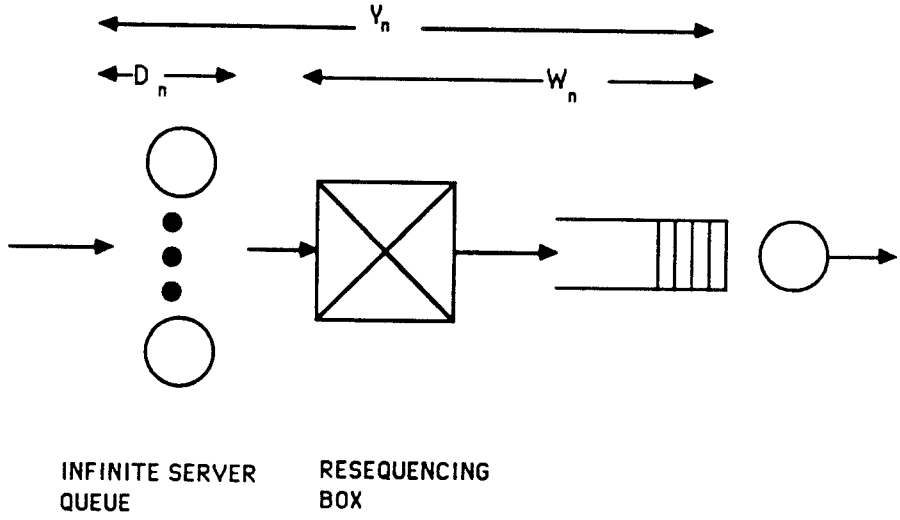


Fig. 8.1. The BGP model

8.2 Polynomial approximations for the BGP model

The reader may recall that the BGP model operates as follows: customers arrive into an infinite server disordering system after receiving service from which, they are resequenced and sent into a single server queue. In this section we develop light traffic approximations for the BGP model in the special case when the arrival process is Poisson with rate λ, the disordering distribution is exponential with rate ν and the service distribution is also exponential with rate μ. The methodology of finding the light traffic limits is the same as the one employed for fork-join queues in Chapter 6. We combine the heavy traffic limit of Section 7.4 with the light traffic limits of this section, to obtain an approximation that provides good estimates over the entire range of λ. Note also that even though we restrict our attention to exponential service and disordering distributions, this methodology can be applied to arbitrary service and disordering distributions.

Let $\overline{W}(\lambda)$ be the sum of the average waiting times in the resequencing box and the buffer of the single server queue. We now proceed to obtain formulae for

$\overline{W}(0)$ and $\overline{W}'(0)$ with the help of (VI.2.6) and (VI.2.7). It is trivial to see that

$$\overline{W}(0) = 0 \tag{2.1}$$

since if only one customer arrives over the entire time interval, then the only delay it encounters before getting served is the disordering delay.

We now proceed to calculate $\overline{W}'(0)$. Let $W(t, d_0, d_1, s_1)$ be the waiting time of the customer that arrives at time zero with disordering delay d_0, given that another customer arrives at time t with disordering delay d_1 and service time s_1 at the single server queue. It is clear that

$$W(t, d_0, d_1, s_1) = \begin{cases} 0, & \text{if } t > 0 \\ \max(0, t + d_1 - d_0 + s_1), & \text{if } t \leq 0. \end{cases} \tag{2.2}$$

Define the RVs X and Y by

$$X = d_1 + s_1, \tag{2.3}$$

$$Y = X - d_0. \tag{2.4}$$

Then it can be easily shown that X has the density function f_X given by

$$f_X(x) = \frac{\mu\nu}{\nu - \mu}(e^{\mu x} - e^{-\nu x}), \quad x \geq 0. \tag{2.5}$$

while Y has the distribution function F_Y given by

$$F_Y(x) = 1 + \frac{\mu e^{-\nu x}}{2(\nu - \mu)} - \frac{\nu^2 e^{-\mu x}}{(\nu^2 - \mu^2)}, \quad x \in I\!R. \tag{2.6}$$

Note that

$$W := W(t, d_0, d_1, s_1) = \max(0, t + Y), \quad t \leq 0$$

so that the RV W has distribution F_W given by

$$\begin{aligned} F_W(x) &= I\!P(Y + t \leq x), \\ &= 1 + \frac{\mu e^{-\nu x} e^{\nu t}}{2(\nu - \mu)} - \frac{\nu^2 e^{-\mu x} e^{\mu t}}{(\nu^2 - \mu^2)}, \quad x \geq 0. \end{aligned} \tag{2.7}$$

Using the fact that

$$\overline{\psi}(\{t\}) = \int_0^\infty (1 - F_W(x))dx, \quad t < 0$$

it follows that

$$\overline{\psi}(\{t\}) = \frac{\nu^2 e^{\mu t}}{\mu(\nu^2 - \mu^2)} - \frac{\mu e^{\nu t}}{2\nu(\nu - \mu)}, \quad t < 0. \tag{2.8}$$

Finally combining (2.8) with (VI.2.7), we obtain

$$\overline{W}'(0) = \frac{\nu^2}{\mu^2(\nu^2 - \mu^2)} - \frac{\mu}{2\nu^2(\nu - \mu)}. \tag{2.9}$$

For the case $\nu = \mu$, if we use L'Hospitals rule and take the limit as $\nu \to \mu$ in (2.8), we obtain,

$$\overline{W}'(0) = \frac{7}{4\mu^2}. \tag{2.10}$$

We now combine the light traffic estimates with the heavy traffic estimates of the last section to obtain a first order approximation to the average waiting time of the resequencing model. If we specialize the heavy traffic result of Section 7.4 to the case when all RVs all exponentially distributed, we obtain

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{W}(\lambda) = 1. \tag{2.11}$$

Finally from (2.1), (2.9) and (2.11), we obtain the first order approximation to the average waiting time in steady state in the form

$$\begin{aligned}\hat{W}(\lambda) = &\frac{\lambda\nu^2}{\mu(\nu^2 - \mu^2)(\mu - \lambda)} - \frac{\lambda\mu^2}{2\nu^2(\nu - \mu)(\mu - \lambda)} + \frac{\lambda^2}{\mu^2(\mu - \lambda)} \\ &- \frac{\lambda^2\nu^2}{\mu^2(\nu^2 - \mu^2)(\mu - \lambda)} + \frac{\lambda^2\mu}{2\nu^2(\nu - \mu)(\mu - \lambda)}, \quad \mu \neq \nu, \\ &\qquad\qquad 0 \leq \lambda < \mu \end{aligned} \tag{2.12}$$

and

$$\hat{W}(\lambda) = \frac{\lambda}{\mu - \lambda} + \frac{3}{4(\mu - \lambda)}[\frac{\lambda}{\mu} - (\frac{\lambda}{\mu})^2], \quad \mu = \nu, \quad 0 \leq \lambda < \mu. \tag{2.13}$$

This approximation agrees extremely well with simulation results (see Section 8.3.1).

Note that in (2.12)

$$\lim_{\nu \uparrow \infty} \hat{W}(\lambda) = \frac{\lambda}{\mu} \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu$$

which is the average waiting time in a $M/M/1$ queue, as expected, because as $\nu \uparrow \infty$ the disordering delay goes to zero and the system behaves like an ordinary $M/M/1$ queue.

8.2.1 Simulation results

The approximation (2.12) is compared with simulation results in the case $\nu = 2$ and $\mu = 1$. Substituting these values into (3.12) we obtain

$$\hat{W}(\lambda) = \frac{\lambda}{24(1-\lambda)}(29 - 5\lambda), \quad 0 \leq \lambda < 1.$$

The 95% confidence levels have been obtained in all cases.

λ	$\overline{W}(\lambda)$	$\hat{W}(\lambda)$	% Error
0.1	0.133 ± 0.003	0.13	2.25
0.2	0.29 ± 0.006	0.29	1.69
0.3	0.49 ± 0.011	0.49	1.80
0.4	0.76 ± 0.017	0.75	1.83
0.5	1.12 ± 0.029	1.10	1.78
0.6	1.65 ± 0.049	1.62	1.82
0.7	2.51 ± 0.097	2.48	1.19
0.8	4.21 ± 0.24	4.16	1.19
0.9	9.50 ± 0.31	9.19	3.26

8.3 Approximations for the parallel queue resequencing model: The case of deterministic services

Our objective in this section is to obtain interpolation approximations for the average response time $\overline{T}(\lambda)$ of the resequencing model in which the disordering is due to K single server queues operating in parallel. We further assume that the input into the system is a Poisson process and the customers are routed to the various queues with a Bernoulli switch with equi–probable switching probability.

In this section we assume that the service times are deterministic and equal to $\frac{1}{\mu}$. In the next section we treat the case in which the service times possess the Erlangian distribution. The heavy traffic limit for this system was given in (8.4.10), which in the case of deterministic service times reduces to

$$\lim_{\lambda \to K\mu} (K\mu - \lambda)\overline{T}_K(\lambda) = \frac{KH_K}{2}. \tag{3.1}$$

We now proceed to obtain the light traffic limits for this system. It is easy to that

$$\overline{T}_K(0) = \frac{1}{\mu} \tag{3.2}$$

since if only one customer arrives over the entire time interval, then it does not encounter any queueing or resequencing delay in the system.

We now proceed with the calculation of $\overline{T}'_K(0)$. Let $T(t)$ be the response time of a customer that arrives at time zero, given that another customer arrived at time t. We have

$$T(t) = \begin{cases} \frac{1}{\mu} & \text{if } t \geq 0 \text{ or if } t < 0 \text{ and the two customers} \\ & \text{go to different queues,} \\ \frac{1}{\mu} + \max(0, t + \frac{1}{\mu}) & \text{if } t < 0 \text{ and both customers} \\ & \text{go to the same queue.} \end{cases} \tag{3.3}$$

It is plain from (3.3) that for the case $t \geq 0$

$$T(t) = \overline{\psi}(\{t\}) = \frac{1}{\mu} = \overline{\psi}(\emptyset)$$

so that (VI.2.7) now reduces to

$$\overline{T}'(0) = \int_{-\infty}^{0} (\overline{\psi}(\{t\}) - \overline{\psi}(\emptyset))dt. \tag{3.4}$$

Taking note of the fact that both customers go to the same queue with probability $\frac{1}{K}$, while they go to different queues with probability $\frac{K-1}{K}$, and substituting (3.3) into (3.4), we obtain

$$\begin{aligned} \overline{T}'_K(0) &= \frac{1}{K} \int_{t=-\frac{1}{\mu}}^{0} (t + \frac{1}{\mu})dt \\ &= \frac{1}{2K\mu^2}. \end{aligned} \tag{3.5}$$

Combining (3.1), (3.2) and (3.5) we obtain an approximation $\hat{T}_K(\lambda)$ for the average response time of the system in the form

$$\hat{T}_K(\lambda) = \frac{K}{(K\mu - \lambda)} - \frac{\lambda}{2\mu(K\mu - \lambda)} + \frac{H_K - 1}{2K(K\mu - \lambda)}(\frac{\lambda}{\mu})^2, \quad 0 \leq \lambda < K\mu. \tag{3.6}$$

This approximation agrees extremely well with simulation results (see Section 8.3.1).

8.3.1 Simulation results

Approximation (3.5) is compared with simulation for the case when $\mu = 1$, while $K = 2, 5$ and 10.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	%Error
0.2	1.05 ± 0.001	1.05	0.01
0.4	1.13 ± 0.002	1.14	0.88
0.6	1.24 ± 0.004	1.25	0.81
0.8	1.38 ± 0.007	1.40	1.45
1.0	1.60 ± 0.012	1.62	1.25
1.2	1.95 ± 0.024	1.97	1.02
1.4	2.53 ± 0.048	2.57	1.58
1.6	3.70 ± 0.111	3.80	2.70
1.8	7.53 ± 0.16	7.52	0.13

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.5	1.06 ± 0.001	1.06	0.11
1.0	1.16 ± 0.003	1.16	0.44
1.5	1.31 ± 0.006	1.30	1.42
2.0	1.52 ± 0.011	1.50	1.31
2.5	1.85 ± 0.020	1.81	2.16
3.0	2.36 ± 0.039	2.31	2.16
3.5	3.23 ± 0.072	3.18	1.55
4.0	5.01 ± 0.167	4.98	0.60
4.5	10.68 ± 0.21	10.51	1.59

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
1	1.07 ± 0.002	1.06	0.65
2	1.20 ± 0.004	1.17	2.50
3	1.41 ± 0.009	1.34	4.96
4	1.73 ± 0.016	1.59	8.06
5	2.19 ± 0.028	1.98	9.58
6	2.90 ± 0.053	2.62	9.65
7	4.08 ± 0.104	3.74	8.33
8	6.43 ± 0.257	6.08	5.44
9	13.69 ± 0.22	13.31	2.77

8.4 Approximations for the parallel queue resequencing model: The case of Erlang services

The model to be analysed is the same as in the last section, except for the fact that now the parallel queues are assumed to possess a r^{th} order Erlang service distribution with rate μ. The heavy traffic limit (VII.4.10) reduces to

$$\lim_{\lambda \to K\mu} (K\mu - \lambda)\overline{T}_K(\lambda) = \frac{r+1}{2r} K H_K. \tag{4.1}$$

As in the last section, we have

$$\overline{T}_K(0) = \frac{1}{\mu} \tag{4.2}$$

and we now proceed to calculate $\overline{T}'_K(0)$.

Let $T(t, s_0, s_1)$ be the response time of a customer that arrives at time zero with service time s_0, given that another customer arrived at time t with service time s_1. We see that

$$T(t, s_0, s_1) = \begin{cases} s_0, & \text{if } t \geq 0 \\ s_0 + \max(0, t + s_1), & \text{if } t < 0 \text{ and the two customers join the same queue,} \\ \max(s_0, t + s_1), & \text{if } t < 0 \text{ and the two customers join different queues.} \end{cases} \tag{4.3}$$

As before, the two customers join the same queue with probability $\frac{1}{K}$, while they join different queues with probability $\frac{K-1}{K}$. Let $\overline{T}'_{K,I}(0)$ be the first derivative of the average response time which is obtained under the assumption that the two customers join the same queue, and let $\overline{T}'_{K,II}(0)$ be this derivative obtained under the assumption that the two customers join different queues. Then it is clear that

$$\overline{T}'_K(0) = \frac{1}{K}\overline{T}'_{K,I}(0) + \frac{K-1}{K}\overline{T}'_{K,II}(0)$$

When both customers join the same queue, the calculation of $\overline{T}'_{K,I}(0)$ reduces to the calculation of the corresponding quantity in a single server queue, since resequencing does not play any role. From the light traffic limits for the single server queue obtained in Chapter 1, we conclude that

$$\overline{T}'_{K,I}(0) = \frac{r+1}{2r\mu^2}. \tag{4.4}$$

We now treat the case where the two customers join different queues. Note that

$$\overline{T}'_{K,II}(0) = \int_{t=0}^{\infty}\int_{s_0=0}^{\infty}\int_{s_1}^{\infty} \max(0, s_1 - s_0 - t)dt\ h_r(s_1)h_r(s_0)ds_1\ ds_0 \tag{4.5}$$

where h_r, is the density function of a r^{th} order Erlang distribution, given by

$$h_r(x) = \frac{r\mu(r\mu x)^{r-1}e^{-r\mu x}}{(r-1)!}, \quad x \geq 0. \qquad r = 1, 2 \ldots \tag{4.6}$$

Interchanging the order of integration in (4.5), we then get

$$\overline{T}'_{K,II}(0) = \int_{s_1=0}^{\infty}\int_{s_0=0}^{s_1}\int_{t=0}^{s_1-s_0} (s_1 - s_0 - t)dt\ h_r(s_0)h_r(s_1)ds_0 ds_1. \quad r = 1, 2 \ldots$$

Carrying out the integration with respect to t and simplifying, we conclude that

$$\overline{T}'_{K,II}(0) = \frac{r+1}{2r\mu^2} - \int_{s_1=0}^{\infty}\int_{s_0=0}^{s_1} s_0 s_1 h_r(s_0)h_r(s_1)ds_0 ds_1. \qquad r = 1, 2 \ldots \tag{4.7}$$

Substituting the expression (4.6) for h_r into (4.7) and making some further simplifications, we obtain

$$\overline{T}'_{K,II}(0) = \frac{1}{\mu^2}\left(\frac{r+1}{2r} - Q\right), \qquad r = 1, 2 \ldots \tag{4.8}$$

where

$$Q = \frac{1}{(r!)^2}\int_{v=0}^{\infty} v^r e^{-v}dv \int_{u=v}^{\infty} u^r e^{-u}du = \frac{1}{2}. \qquad r = 1, 2 \ldots \tag{4.9}$$

Combining (4.4) and (4.8), we obtain

$$\overline{T}'_K(0) = \frac{1}{\mu^2}(\frac{r+1}{2r} - \frac{K-1}{2K}). \qquad r = 1, 2 \ldots (4.10)$$

Finally combining (4.1), (4.2) and (4.10) we obtain an approximation $\hat{T}_K(\lambda)$ for the average response time of the system in the form

$$\hat{T}_K(\lambda) = \frac{K}{K\mu - \lambda} + \left[K(\frac{r+1}{2r} - \frac{K-1}{2K}) - 1\right] \frac{\lambda}{\mu(K\mu - \lambda)} + \left[\frac{r+1}{2r}\frac{H_K}{K} - (\frac{r+1}{2r} - \frac{K-1}{2K})\right] (\frac{\lambda}{\mu})^2 \frac{1}{(K\mu - \lambda)},$$
$$0 \leq \lambda < K\mu, \; r = 1, 2 \ldots (4.11)$$

This approximation agrees extremely well with simulation results (see Section 8.4.1).

8.4.1 Simulation results

Approximation (4.11) is compared with simulation for the case when $r = 2, \mu = 1$, while $K = 2, 5$ and 10.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error
0.2	1.11 ± 0.005	1.11	0.09
0.4	1.25 ± 0.007	1.25	0.08
0.6	1.43 ± 0.010	1.44	0.70
0.8	1.69 ± 0.016	1.74	2.96
1.0	2.05 ± 0.028	2.06	0.49
1.2	2.59 ± 0.048	2.61	0.77
1.4	3.49 ± 0.089	3.54	1.43
1.6	5.31 ± 0.232	5.40	1.69
1.8	10.90 ± 0.28	11.01	1.03

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.5	1.19 ± 0.005	1.19	0.25
1.0	1.43 ± 0.007	1.43	0.35
1.5	1.74 ± 0.017	1.74	0.28
2.0	2.15 ± 0.028	2.15	0.30
2.5	2.72 ± 0.047	2.73	0.37
3.0	3.57 ± 0.083	3.59	0.56
3.5	5.08 ± 0.174	5.02	1.18
4.0	7.87 ± 0.121	7.88	0.13
4.5	16.14 ± 0.372	16.44	1.86

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
1	1.31 ± 0.007	1.32	0.76
2	1.67 ± 0.014	1.71	2.39
3	2.11 ± 0.023	2.18	3.32
4	2.67 ± 0.038	2.79	4.49
5	3.43 ± 0.062	3.60	4.95
6	4.56 ± 0.110	4.78	4.82
7	6.50 ± 0.225	6.69	2.92
8	10.49 ± 0.574	10.44	0.47
9	20.97 ± 0.414	21.52	2.62

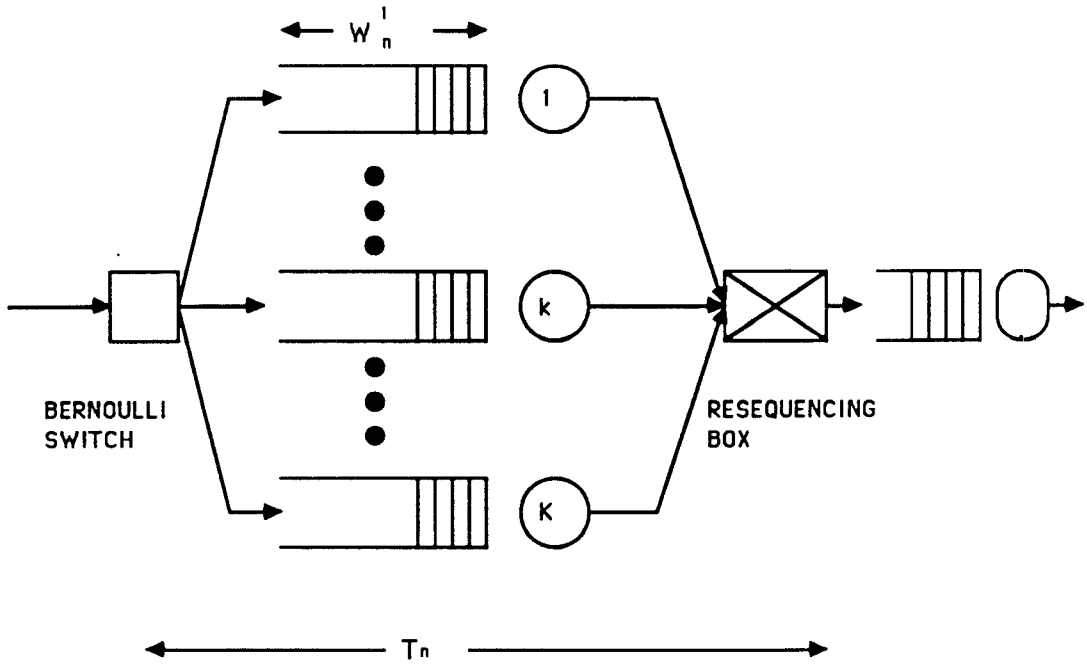


Fig. 8.2. A generalized BGP model

8.5 Light traffic approximations for a generalized BGP model

The model analyzed in this section operates as follows: Customers arriving according to a Poisson process with rate λ are routed to one of K single server queues operating in parallel with identical exponential service rates ν. The routing decision is made with a Bernoulli switch, with equal routing probability. After they leave this system they are resequenced in a resequencing buffer and sent to the buffer of single server queue with exponential service rate μ. After getting served in this queue they leave the system. Note that the parallel $M/M/1$ queues can also be in heavy traffic, in addition to the single server queue. Hence we shall assume that $K\nu > \mu$, so that as the the arrival rate λ increases from zero, the single server queue goes into heavy traffic earlier than the K $M/M/1$ queues.

We now proceed to find the light traffic limits for the average waiting time $\overline{W}_K(\lambda)$, in this system, which is defined as the sum of the average waiting times in the resequencing box and the buffer of the single server queue.

It is trivial to see that

$$\overline{W}_K(0) = 0 \tag{5.1}$$

since if only one customer arrives over the entire time interval, then it does not encounter any resequencing or queueing delay.

We now proceed to calculate $\overline{W}'_K(0)$. Let $W(t, c_0, c_1, s_1)$ be the waiting time of the customer that arrives at time zero with service time c_0 at one of the $M/M/1$ queues, given that another customer arrives at time t with service times c_1 at one of the $M/M/1$ queues and s_1 at the single server queue. Then it clear that

$$W(t, c_0, c_1, s_1) = \begin{cases} 0, & \text{if } t > 0; \\ \max(0, t + c_1 + s_1 - c_0), & \text{if } t \leq 0 \text{ and the customers are routed to different queues.} \\ \max(0, t + c_1 + s_1 - c_0 - \max(0, t + c_1)), & \text{if } t \leq 0 \text{ and both customers are routed to the same queue.} \end{cases} \tag{5.2}$$

Note that the customers are routed to the same queue with probability $\frac{1}{K}$, while they are routed to different queues with probability $\frac{K-1}{K}$. When the customers go to different queues, the waiting time is exactly the same as for the case when the disordering system is an infinite server queue. Hence

$$\begin{aligned} \overline{W}'_K(0) =& \frac{K-1}{K} \int_{t=-\infty}^{0} \int_{c_0} \int_{c_1} \int_{s_1} \max(0, t + c_1 + s_1 - c_0) H_1(dc_0) H_1(dc_1) H_2(ds_1) \\ &+ \frac{1}{K} \int_{t=-\infty}^{0} \int_{c_0} \int_{c_1} \int_{s_1} \max(0, t + c_1 + s_1 - c_0 - \max(0, t + c_1)) \\ & H_1(dc_0) H_1(dc_1) H_2(ds_1) \end{aligned} \tag{5.3}$$

where H_1 and H_2 are exponential distributions with rate ν and μ respectively.

The first of these integrals was already calculated in Section 8.2 so that

$$\overline{W}'_K(0) = \frac{(K-1)\nu^2}{K\mu^2(\nu^2 - \mu^2)} - \frac{(K-1)\mu}{2K\nu^2(\nu - \mu)}$$

$$+\frac{1}{K}\int_{t=-\infty}^{0}\int_{c_0}\int_{c_1}\int_{s_1}\max(0,t+c_1+s_1-c_0-\max(0,t+c_1))$$

$$H_1(dc_0)H_1(dc_1)H_2(ds_1). \tag{5.4}$$

We now proceed to calculate the second integral which we denote as I. The RV X defined by

$$X = c_1 + t \tag{5.5}$$

has the density function.

$$f_X(x) = \nu e^{\nu t}e^{-\nu x}, \quad x \geq t. \tag{5.6}$$

Also, the RV Y defined as

$$Y = s_1 - c_0 \tag{5.7}$$

has distribution function

$$F_Y(x) = \begin{cases} 1 - \frac{\nu}{\nu+\mu}e^{-\mu x}, & \text{if } x \geq 0 \\ e^{\nu x} - \frac{\nu}{\nu+\mu}e^{(\nu+2\mu)x}, & \text{if } x < 0\ . \end{cases} \tag{5.8}$$

With the help of (5.5) and (5.7), (5.2) simplifies to

$$W(X,Y) = \max(0, X + Y - \max(0,X)) \tag{5.9}$$

for the case $t \geq 0$. Our next objective is to find the distribution function of the RV $W(X,Y)$. Note that for $z \geq 0$, we have

$$\begin{aligned} I\!P(W(X,Y) \leq z) &= I\!P(X + Y - \max(0,X) \leq z) \\ &= \int_{x=t}^{\infty} I\!P(x + Y - \max(0,x) \leq z \mid X = x) f_X(x)dx \end{aligned} \tag{5.10}$$

Since that the RVs X and Y are independent, (5.10) simplifies to

$$\begin{aligned} I\!P(W(X,Y) \leq z) &= \int_{x=t}^{0} I\!P(x + Y \leq z) f_X(x)dx + \int_{x=0}^{\infty} I\!P(Y \leq z) f_X(x)dx \\ &= (1 - \frac{\nu}{\nu+\mu}e^{-\mu z}) \int_{x=0}^{\infty} \nu \exp^{\nu t} e^{-\nu x}dx \\ &\quad + \int_{x=t}^{0} (1 - \frac{\nu}{\nu+\mu}e^{-\mu(z-x)})\nu \exp^{\nu t} e^{-\nu x}dx \\ &= 1 - \frac{\nu}{\nu+\mu}e^{\nu t}e^{-\mu z} - \frac{\nu^2}{\nu^2-\mu^2}e^{-\mu z}(e^{\mu t} - e^{\nu t}), \quad z \geq 0 \end{aligned} \tag{5.11}$$

and we get

$$\mathbb{E}W(X,Y) = \int_{z=0}^{\infty} [1 - \mathbb{P}(W(X,Y) \leq z)]dz$$

$$= \frac{\nu}{\mu(\nu+\mu)} e^{\nu t} + \frac{\nu^2}{\mu(\nu^2-\mu^2)}(e^{\mu t} - e^{\nu t}). \tag{5.12}$$

Integrating over t we conclude that

$$I = \frac{1}{\mu^2}. \tag{5.13}$$

Whence, upon combining (5.4) and (5.13),

$$\overline{W}'_K(0) = \frac{(K-1)\nu^2}{K\mu^2(\nu^2-\mu^2)} - \frac{(K-1)\mu}{2K\nu^2(\nu-\mu)} + \frac{1}{K\mu^2}. \tag{5.14}$$

Note that if we specialize the heavy traffic result of Section 7.3 to the case when all RVs all exponentially distributed, we obtain

$$\lim_{\lambda \uparrow \mu} (\mu - \lambda)\overline{W}_K(\lambda) = 1. \tag{5.15}$$

Combining (5.1), (5.14) and (5.15), we obtain the first order approximation $\hat{W}_K(\lambda)$ to the average waiting time for the case when $K\nu > \mu$ and $\nu \neq \mu$, in the form

$$\hat{W}_K(\lambda) = \frac{\lambda}{K\mu(\mu-\lambda)} + \frac{(K-1)\lambda\nu^2}{K\mu(\nu^2-\mu^2)(\mu-\lambda)} - \frac{(K-1)\lambda\mu^2}{2K\nu^2(\nu-\mu)(\mu-\lambda)}$$

$$+ \frac{(K-1)\lambda^2}{K\mu^2(\mu-\lambda)} - \frac{(K-1)\lambda^2\nu^2}{K\mu^2(\nu^2-\mu^2)(\mu-\lambda)} + \frac{(K-1)\lambda^2\mu}{2K\nu^2(\nu-\mu)(\mu-\lambda)},$$

$$0 \leq \lambda < \mu. \tag{5.16}$$

Note that

$$\lim_{\nu \uparrow \infty} \hat{W}_K(\lambda) = \frac{\lambda}{\mu} \frac{1}{\mu - \lambda}, \quad 0 \leq \lambda < \mu$$

which is the average waiting time in a $M/M/1$ queue, again as expected, because as $\nu \uparrow \infty$ the disordering delay goes to zero.

In the case $\nu = \mu$, a similar calculation shows that

$$\hat{W}_K(\lambda) = \frac{7(K-3)}{4K}\frac{\lambda}{\mu}\frac{1}{\mu-\lambda} - \frac{(3-3K)}{4K}(\frac{\lambda}{\mu})^2\frac{1}{\mu-\lambda}, \quad 0 \leq \lambda < \mu. \qquad (5.17)$$

Even though approximation (6.16) is valid in the range $K\nu > \mu$, simulation results suggest that it performs quite poorly when $K\nu$ is close to μ. It performs best when $K\nu \gg \mu$ (see Section 8.5.2), and we suggest that the reader who is interested in applying (5.16), choose ν such that at least $\nu \geq \mu$.

8.5.2 Simulation results

Approximation (5.16) is compared with simulation for the case when $K = 2, \nu = 2$ and $\mu = 1$. Substituting these values into (5.16) we obtain

$$\hat{W}_2(\lambda) = \frac{\lambda}{1-\lambda}(1.104 - 0.104\lambda), \quad 0 \leq \lambda < 1.$$

λ	$\overline{W}_2(\lambda)$	$\hat{W}_2(\lambda)$	% Error
0.1	0.124 ± 0.004	0.121	2.42
0.2	0.27 ± 0.006	0.271	0.37
0.3	0.46 ± 0.010	0.456	0.22
0.4	0.72 ± 0.016	0.71	1.39
0.5	1.07 ± 0.028	1.05	1.86
0.6	1.60 ± 0.050	1.56	2.50
0.7	2.46 ± 0.104	2.41	2.03
0.8	4.12 ± 0.07	4.08	0.97
0.9	8.91 ± 0.25	9.09	2.02

CHAPTER IX

9.1 Introduction

So far in this dissertation, we have considered queueing systems with either the fork–join synchronization constraint or the resequencing constraint. Heavy traffic limit theorems were given for fork–join systems in Chapter 2, while light traffic results were presented in Chapter 6. The corresponding results for resequencing systems were given in Chapters 7 and 8 respectively.

We now consider queueing models that exhibit both fork–join as well as resequencing synchronization constraints. In Section 9.2 we introduce a new model which is a generalization of the acyclic fork–join network analyzed in Chapter 2. It is similar to the acylic fork–join network, except that every single server queue is preceded by an infinite server disordering system, followed by a resequencing box. This model is being introduced here for the first time, and it subsumes most of the different fork–join and resequencing models analyzed so far. We obtain the basic recursions governing this model and give the stability conditions. Our main result regarding this model is that it has the same heavy traffic diffusion limit as the acyclic fork–join network from Chapter 2. Hence in effect we have identified a class of queueing models in which the resequencing constraint can be ignored in heavy traffic. There is an interesting special cases of this model for which we shall obtain polynomial approximations by interpolating between heavy traffic and light traffic limits. This is a model originally proposed by Baccelli [3] to model time–stamp ordering in a distributed system.

This chapter is organized as follows: In Section 9.2.1 we introduce the acyclic fork–join network with resequencing, while in Section 9.2.2 we derive the recursions for the delays in the network as well as its stability conditions. Heavy traffic diffusion limits for this network are presented in Section 9.2.3. In Section 9.3 we

prove the admissibility of the queueing systems whose light traffic limits are to be obtained. Sections 9.4 is devoted to obtaining polynomial approximations for the time–stamp ordering model.

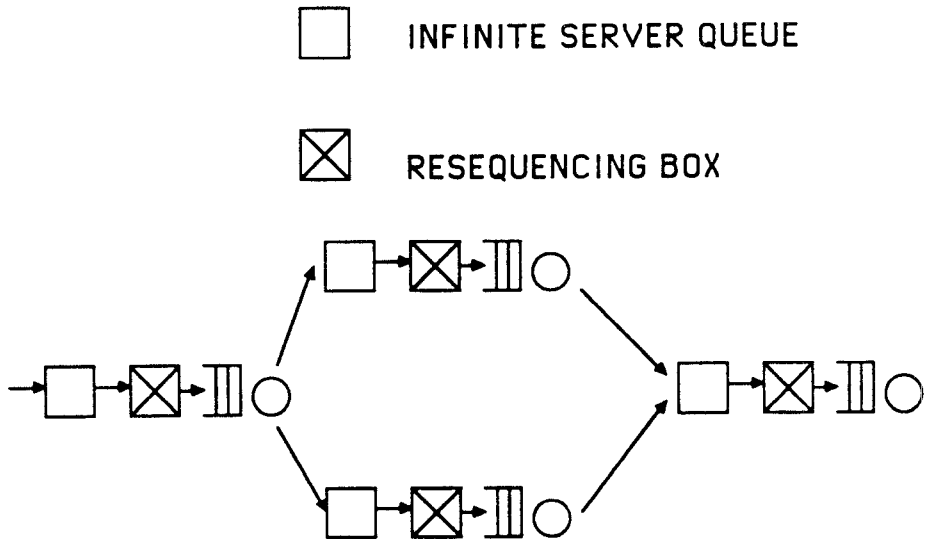


Fig. 9.1. An example of an acyclic synchronized network.

9.2.1 The model

In this section we generalize the notion of an acyclic fork–join network by introducing an additional synchronization constraint, i.e., resequencing, into its framework. In order to do so, we precede every single server queue in the acyclic fork–join network by an infinite server disordering system, followed by a resequencing box (Fig. 9.1). For convenience this network will be referred to as an *acyclic synchronized network.* Our principle result is that the acyclic synchronized network has the same heavy traffic diffusion limit as the usual acyclic fork–join network.

We now introduce the notations and definitions associated with the acyclic synchronized network. Just as the basic building block of an acyclic fork–join network was a single server queue, the basic building block of an acyclic synchronized network is a single server queue preceded by an infinite server disordering system which is followed by a resequencing box. We shall refer to this unit as a *resequenced queue.*

The acyclic synchronized network under consideration is represented by an acyclic graph $G = (V, E)$ where V is a set of B resequenced queues labeled $i =$

$1, \ldots, B$ and E is a set of links such that $(i,j) \in E$ implies $j > i$. Also add for the sake of convenience fictitious single server queues 0 and $B+1$, which act respectively as source and sink for the network.

Define the sets $s(i), s(0), p(i), p(B+1)$ and the queues B', B'', as in (2.4.1)–(2.4.2) after substituting the phrase resequenced queue for all occurrences of the word queue.

We now describe the operation of the network. We assume that customers are being created at the source which acts as the outside world for the network. These exogenous customers enter the network through the resequenced queues in $s(0)$ and traverse it upon following certain synchronization rules (SR_1)–(SR_4) described below. Finally customers leave the network from the resequenced queues in $p(B+1)$ from where they are absorbed into the network sink and disappear. We now specify the synchronization rules that govern the network.

(SR_1): The exogenous customers created at the source are routed instantaneously to the resequenced queues in $s(0)$ under the constraint of a Fork primitive, i.e., the n^{th} arrival date to each one of the queues in $s(0)$ coincides with the n^{th} date of customer creation. An alternate way of viewing this constraint is to assume that upon its creation, a customer creates B' replicas of itself which are then dispatched at the same time and instantaneously to the resequenced queues in $s(0)$, one replica per queue.

(SR_2): Each resequenced queue j, for $1 \leq j \leq B$, processes customers according to the BGP model of resequencing (Section 8.4), with the disordering being carried out by an infinite server queue.

(SR_3): The service completion in some resequenced queue in $s(0)$ will not systematically trigger an arrival to a resequenced queue in $s(i)$. In fact, more generally, the arrivals to resequenced queue j, with $B' < j \leq B$, are generated as follows: Assume the sequence of service completions to be known for all resequenced queues i, with $1 \leq i \leq j$ where $B' < j \leq B$. The n^{th} arrival date to resequenced queue j co–incides

with the latest date among all the n^{th} service completions at the resequenced queues in $p(j)$. Due to the acyclic structure of (V, E), this mechanism will successively define the arrival patterns to resequenced queues $B'+1, B'+2, \ldots, B$.

(SR_4): Customers leave the network through the queues in $p(B+1)$ in the form of a single output stream by imposing the following synchronization of the join type: The n^{th} network departure is defined as the latest date among the dates of n^{th} service completions in the resequenced queues $B'', B''+1, \ldots, B$.

9.2.2 Recursive representation of the delays

In this section a recursive representation for the delays in the network is provided. The single server queue and the infinite server queue associated with the j^{th} resequenced queue will be referred to as the j^{th} single server queue and the j^{th} infinite server queue respectively. Given an acyclic graph $G = (V, E)$, the performance measures associated with the corresponding network are fully specified by $(B+1)$ sequences of $\mathbb{R}_+$–valued RVs with the interpretation that for all $n = 0, 1 \ldots$, and $1 \leq j \leq B$,

τ_n : Arrival epoch of the n^{th} customer into the network.

v_n^j : Service time requirement of the n^{th} customer to be served in the single server queue j.

d_n^j : Service time requirement of the n^{th} customer to be served in infinite server queue j.

We assume the system to be initially empty and adopt the convention that the 0^{th} exogenous customer is created at time $t = 0$, so that $\tau_0 = 0$. In terms of these RVs we define the following quantities for all $n = 0, 1 \ldots$ and all $1 \leq j \leq B$,

u_{n+1} : Inter-arrival time between the $(n+1)^{rst}$ and n^{th} exogenous customers $(= \tau_{n+1} - \tau_n)$.

D_n^j : Delay between the arrival of the n^{th} exogenous customer and the beginning of the n^{th} service in resequenced queue j.

W_n^j : Waiting time of the n^{th} exogenous customer in the the resequencing box associated with the infinite server queue j, as well as the buffer of queue j.

T_n : End–to–end delay or network response time of the n^{th} exogenous customer.

The following recursion holds between these variables.

Lemma 9.2.1. *Consider the acyclic resequenced network defined above. If the system is initially empty, then for* $1 \leq j \leq B$, *the recursions*

$$D_0^j = d_0^j + \max_{i \in p(j)} \{D_0^i + v_0^i\}$$

$$D_{n+1}^j = \max\{d_{n+1}^j + \max_{i \in p(j)} \{D_{n+1}^i + v_{n+1}^i\}, D_n^j + v_n^j - u_{n+1}\},$$

$$n = 0, 1 \ldots (2.1)$$

and

$$W_0^j = 0$$

$$W_{n+1}^j = \max\{0, W_n^j + d_n^j + \max_{i \in p(j)} \{D_n^i + v_n^i\} - d_{n+1}^j - \max_{i \in p(j)} \{D_{n+1}^i + v_{n+1}^i\}$$

$$+ v_n^j - u_{n+1}\}, \qquad n = 0, 1 \ldots (2.2)$$

hold where the maximum over an emptyset is zero by convention. Moreover the network response time of the n^{th} *exogeneous customer is given by*

$$T_n = \max_{i \in p(B+1)} \{D_n^i + v_n^i\}. \qquad n = 0, 1 \ldots (2.3)$$

Proof. The system being initially empty, the boundary conditions (2.1)–(2.2) are thus immediate from the synchronization rules (SR_1)–(SR_3). Customers arriving to the resequenced queue j in $s(0)$ do so according to the pattern of exogenous arrivals, so that D_n^j corresponds to the n^{th} delay in resequenced queue, generated by the sequences of interarrivals $\{u_{n+1}\}_0^\infty$, infinite server queue delays $\{d_n^j\}_0^\infty, 1 \leq$

$j \leq B$ and single server service requirements $\{v_n^j\}_0^\infty, 1 \leq j \leq B$. Writing the corresponding recursion which was given in Lemma 7.2.1, we get

$$D_{n+1}^j = \max\{d_{n+1}^j, D_n^j + v_n^j - u_{n+1}\}, \quad 1 \leq j \leq B_0 \qquad n = 0, 1 \ldots (2.4)$$

and this reduces to (2.1), since $p(j) = \emptyset$ for $j \in s(0)$.

For $B' < j \leq B$, we fix $n = 0, 1 \ldots$. The $(n+1)^{rst}$ service completion at queue i in $p(j)$ takes place at time $\tau_{n+1} D_{n+1}^i + v_{n+1}^i$, so that by applying the synchronization rule (SR_3), we see that the $(n+1)^{rst}$ arrival to the resequencing buffer in queue j takes place at time $\tau_{n+1} + d_{n+1}^j + \max_{i \in p(j)}\{D_{n+1}^i + v_{n+1}^i\}$. Since the server in queue j becomes available for service at time $\tau_n + D_n^j + v_n^j$, we readily obtain (2.1).

In order to derive (2.2) we just have to note that

$$W_n^j = D_n^j - d_n^j - \max_{i \in p(j)}\{D_n^i + v_n^i\}, \quad 1 \leq j \leq B. \qquad n = 0, 1 \ldots$$

∎

We now state a result regarding the stability of these networks. First we make the following assumption.

(IXa): The sequences $\{u_{n+1}\}_0^\infty$, $\{d_n^j\}_0^\infty$ and $\{v_n^j\}_0^\infty, j = 1, \ldots, B$, are iid with finite second moments and mutually independent.

For all $n = 0, 1 \ldots$, we set

$$u = I\!E(u_n) < \infty, \quad \sigma_0^2 = Var(u_n) < \infty$$
$$v^j = I\!E(v_n^j) < \infty, \quad \sigma_j^2 = Var(v_n^j) < \infty, \quad 1 \leq j \leq B$$
$$\overline{d}^j = I\!E(d_n^j) < \infty, \quad \overline{\sigma}_j^2 = Var(d_n^j) < \infty, \quad 1 \leq j \leq B.$$

The next lemma provides conditions for stability of the system.

Lemma 9.2.2. *Assume that condition* **(IXa)** *holds. If*

$$v^j < u, \quad 1 \leq j \leq B \qquad n = 0, 1 \ldots (2.5)$$

holds, the system is stable in the sense that the vector of delays $(D_n^1, \ldots, D_n^B)$ *converges jointly in distribution as* $n \uparrow \infty$ *to a proper random vector* $(D^1, \ldots, D^B)$.

Proof. The proof is a simple extension of the argument given for acyclic fork–join networks in [5], and is left to the interested reader.

9.2.3 The diffusion limit

In the last section we saw that the acyclic synchronized network will be stable provided $v^j < u, 1 \leq j \leq B$. The system is said to be in heavy traffic if $v^j \approx u$ for one or more queues. In this section our objective is to develop heavy traffic diffusion limits for the delay processes in these networks. The methodology that we shall employ is the same as the one used in Section 2.4.3, i.e., we shall use the recursions (2.1)–(2.2) to connect the delay processes to partial sums of iid RVs and then use the well–known results regarding functional central limit theorems for these partial sums in order to deduce the corresponding limit theorems for the delay processes by means of the continuous mapping theorem. The main result that we obtain is that acyclic synchronized networks have the same diffusion limit as acyclic fork–join networks.

We now consider a sequence of these networks indexed by $r = 1, 2 \ldots$, each of which satisfies condition **(IXa)**. Moreover assume that:

(IXb): As $r \uparrow \infty$,

$$\sigma_j(r) \to \sigma_j, \quad 0 \leq j \leq B$$

$$\overline{\sigma}_j(r) \to \overline{\sigma}_j, \quad 0 \leq j \leq B$$

$$[u(r) - v^j(r)]\sqrt{r} \to c_j. \quad 1 \leq j \leq B$$

(IXc): For some $\epsilon > 0$,

$$\sup_{r,j} \{I\!E\{\mid u_1(r) \mid^{2+\epsilon}\}, I\!E\{\mid v_1^j(r) \mid^{2+\epsilon} I\!E\{\mid d_1^j(r) \mid^{2+\epsilon}\} < \infty.$$

For $1 \leq j \leq B$ and $r = 1, 2 \ldots$, define the partial sums

$$V_0^j(r) = 0,$$

$$V_n^j(r) = v_0^j(r) + \ldots + v_{n-1}^j(r), \qquad n = 1, 2 \ldots (2.6a)$$

and

$$U_0(r) = 0,$$
$$U_n(r) = u_1(r) + \ldots + u_n(r). \qquad n = 1, 2 \ldots (2.6b)$$

For $r = 1, 2 \ldots$, define the stochastic processes $\xi^j(r) \equiv \{\xi_t^j(r), t \geq 0\}, 0 \leq j \leq B$, with sample paths in $D[0, \infty)$ by

$$\xi_t^0(r) = \frac{U_{[rt]}(r) - u(r)[rt]}{\sqrt{r}}, \quad t \geq 0 \qquad (2.7a)$$

$$\xi_t^j(r) = \frac{V_{[rt]}^j(r) - v^j(r)[rt]}{\sqrt{r}}, \quad 1 \leq j \leq B, \quad t \geq 0. \qquad (2.7b)$$

Let $\xi^j \equiv \{\xi_t^j, t \geq 0\} 0 \leq j \leq B$, be $B + 1$ independent Wiener processes. Lemma 9.2.3 shows that the random functions defined in (2.7) converge weakly to these Wiener processes.

Lemma 9.2.3 *As* $r \uparrow \infty$,

$$(\xi^0(r), \xi^1(r), \ldots, \xi^B(r)) \Rightarrow (\sigma_0 \xi^0, \sigma_1 \xi^1, \ldots, \sigma_B \xi^B) \qquad (2.8)$$

in $D[0, \infty)^B$.

Proof. The proof is exactly the same as for Lemma 2.2.1., with assumptions **(IIe)–(IIg)** now replaced by assumptions **(IXa)–(IXc)**. ∎

For $r = 1, 2 \ldots$, we set

$$S_0^j(r) = 0$$
$$S_n^j(r) = V_n^j(r) - U_n(r), \qquad n = 1, 2 \ldots (2.9)$$

and define the following stochastic processes $\{\zeta^j(r) \equiv \{\zeta_t^j(r), t \geq 0\}, 1 \leq j \leq B$,with sample paths on $D[0, \infty)$, by

$$\zeta_t^j(r) = \frac{S_{[rt]}^j(r)}{\sqrt{r}}, \quad 1 \leq j \leq B, \quad t \geq 0. \qquad (2.10)$$

We also define the stochastic processes $\zeta^j \equiv \{\zeta^j, t \geq 0\}, 1 \leq j \leq B$, by

$$\zeta_t^j = \sigma_j \xi_t^j - \sigma_0 \xi_t^0 - c_j t, \quad 1 \leq j \leq B, \quad t \geq 0. \tag{2.11}$$

The next result shows that the stochastic processes $(\zeta^1(r), \ldots, \zeta^B(r))$ converge weakly to $(\zeta^1, \ldots, \zeta^B)$. As we noted in the discussion preceding Lemma 2.2.2, the random process $\zeta^j, 1 \leq j \leq B$ form a K–dimensional diffusion process with drift given by (2.2.16) and co–variance given by (2.2.17).

Lemma 9.2.4 *As* $r \uparrow \infty$,

$$(\zeta^1(r), \ldots, \zeta^B(r)) \Rightarrow (\zeta^1, \ldots, \zeta^B) \tag{2.12}$$

in $D[0, \infty)^B$.

Proof. The proof is exactly the same as for Lemma 2.2.2. ∎

For $r = 1, 2 \ldots$, we define the stochastic processes $\eta^j(r) \equiv \{\eta_t^j(r), t \geq 0\}$ and $\mu^j(r) \equiv \{\mu_t^j(r), t \geq 0\}, 1 \leq j \leq B$, with sample paths in $D[0, \infty)$, by setting

$$\eta_t^j(r) = \frac{D_{[rt]}^j(r)}{\sqrt{r}}, \quad 1 \leq j \leq B, \quad t \geq 0 \tag{2.13}$$

and

$$\mu_t^j(r) = \frac{W_{[rt]}^j(r)}{\sqrt{r}}, \quad 1 \leq j \leq B, \quad t \geq 0. \tag{2.14}$$

The processes $\eta^j \equiv \{\eta_t^j, t \geq 0\}, 1 \leq j \leq B$, and $\mu^j \equiv \{\mu_t^j, t \geq 0\}, 1 \leq j \leq B$, are now defined by

$$\eta^j = g(\zeta^j - \max_{i \in p(j)} \eta^i) + \max_{i \in p(j)} \eta^i, \quad 1 \leq j \leq B \tag{2.15}$$

and

$$\mu^j = g(\zeta^j - \max_{i \in p(j)} \eta^i), \quad 1 \leq j \leq B. \tag{2.16}$$

We now present the main result of this section.

Theorem 9.2.1. *As* $r \uparrow \infty$,

$$(\eta^1(r), \ldots, \eta^B(r)) \Rightarrow (\eta^1, \ldots, \eta^B) \tag{2.17}$$

in $D[0,\infty)^B$.

The reader may note that the limiting process obtained for the acyclic synchronized network in heavy traffic is identical to the limiting process obtained for the acyclic fork–join network in heavy traffic. Before providing a proof for Theorem 9.2.1, we present the following two corollaries which identify the diffusion limit for the waiting times and the end–to–end delay of the system respectively.

Corollary 9.2.1. *As* $r \uparrow \infty$,

$$(\mu^1(r), \ldots, \mu^B(r)) \Rightarrow (\mu^1, \ldots, \mu^B) \tag{2.18}$$

in $D[0,\infty)^B$.

Proof. Note that for all $r = 1, 2 \ldots$,

$$W_n^j(r) = D_n^j(r) - d_n^j(r) - \max_{i \in p(j)} \{D_n^i(r) + v_n^i(r)\}, \quad 1 \leq j \leq B \qquad n = 0, 1 \ldots$$

so that for all $r = 1, 2 \ldots$,

$$\mu_t^j(r) = \eta_t^j(r) - \frac{d_{[rt]}^j(r)}{\sqrt{r}} - \max_{i \in p(j)} \{\eta_t^i(r) + \frac{v_{[rt]}^i(r)}{\sqrt{r}}\}, \quad 1 \leq j \leq B, \quad t \geq 0 \tag{2.19}$$

We obtain (2.18) from (2.17) and (2.19) by applying the continous mapping theorem and the converging together theorem. ∎

For $r = 1, 2 \ldots$, we introduce the stochastic processes $\kappa(r) \equiv \{\kappa_t(r), t \geq 0\}$ with sample paths in $D[0,\infty)$ by

$$\kappa_t(r) = \frac{T_{[rt]}(r)}{\sqrt{r}}, \quad t \geq 0. \tag{2.20}$$

Corollary 9.2.2. *As* $r \uparrow \infty$,

$$\kappa(r) \Rightarrow \max_{i \in p(B+1)} \eta^i \tag{2.21}$$

in $D[0,\infty)$.

Proof. Using the fact that for all $r = 1, 2 \ldots$

$$\kappa_t(r) = \max_{i \in p(B+1)} \{\eta_t^i(r) + \frac{v^i_{[rt]}}{\sqrt{r}}\}, \quad t \geq 0 \tag{2.22}$$

we obtain (2.21) from (2.17) and (2.22) by applying the continuous mapping theorem and the converging together theorem. ∎

We now proceed with the proof for Theorem 9.2.1. For $1 \leq i \leq B$, we define the level $l(i)$ of queue i and the set $q(l), 1 \leq l \leq N$, of queues on level as in (4.15)–(4.16).

Proof. Our proof proceeds by induction on the levels of the acyclic graph which underlies the queueing network. First consider the queues belonging to the set $q(1)$, i.e., queues j such that $l(j) = 1$. Recall that for these queues $p(j) = \emptyset$, so that for $r = 1, 2 \ldots$ we have

$$\begin{aligned} D^j_{n+1}(r) &= \max\{0, D^j_n(r) + d^j_n(r) - d^j_{n+1}(r) + v^j_n(r) - u_{n+1}(r)\} \\ &= S^j_{n+1}(r) - d^j_{n+1}(r) - \min_{0 \leq k \leq n+1} \{S^j_k(r) - d^j_{n+1}(r)\} \end{aligned}$$

$$n = 0, 1 \ldots \tag{2.23}$$

From (2.12), (2.23), the continuous mapping theorem and the converging together theorem, it follows that

$$(\eta^1(r), \ldots, \eta^{B_1}(r), \zeta^1(r), \ldots, \zeta^B(r)) \Rightarrow (\eta^1, \ldots, \eta^{B_1}, \zeta^1, \ldots, \zeta^B) \tag{2.24}$$

as $r \uparrow \infty$, so that (2.17) is verified for the queues belonging to the set $q(1)$.

As the induction hypothesis, assume that

$$(\eta^1(r), \ldots, \eta^{B_1+\ldots+B_l}(r), \zeta^1(r), \ldots, \zeta^B(r)) \Rightarrow (\eta^1, \ldots, \eta^{B_1+\ldots+B_l}, \zeta^1, \ldots, \zeta^B) \tag{2.25}$$

as $r \uparrow \infty$, which implies that (2.17) holds for the queues belonging to the first l levels. Using (2.25) we shall prove that (2.17) holds for queues belonging to the first $l+1$ levels, thus completing the induction step.

Consider queue j such that $l(j) = l + 1$. Expanding the recursion in Lemma 9.2.1 for $r = 1, 2 \ldots$, $n = 0, 1 \ldots$ and $j = B_l + 1, \ldots, B_{l+1}$, we obtain

$$D^j_{n+1}(r)$$

$$= d^j_{n+1}(r) + \max_{i \in p(j)} \{D^i_{n+1}(r) + v^i_{n+1}(r)\}$$

$$+ \max\{0, D^j_n(r) - d^j_{n+1}(r) - \max_{i \in p(j)} \{D^i_{n+1}(r) + v^i_{n+1}(r)\} + v^j_n(r) - u_{n+1}(r)\},$$

$$= d^j_{n+1}(r) + \max_{i \in p(j)} \{D^i_{n+1}(r) + v^i_{n+1}(r)\} + S^j_{n+1}(r) - d^j_{n+1}(r)$$

$$- \max_{i \in p(j)} \{D^i_{n+1}(r) + v^i_{n+1}(r)\} - \min_{0 \le k \le n+1} \{S^j_k(r) - d^j_k(r) - \max_{i \in p(j)} \{D^i_k(r) + v^i_k(r)\}\} \tag{2.26}$$

Note that by (2.26), we have for $j = B_l + 1, \ldots, B_{l+1}$ and $t \ge 0$,

$$\eta^j_t(r)$$

$$= \frac{d^j_{[rt]}(r)}{\sqrt{r}} + \max_{i \in p(j)} \{\eta^i_t(r) + \frac{v^i_{[rt]}}{\sqrt{r}}\} + g\left(\zeta^j(r) - \frac{d^j_{[r.]}(r)}{\sqrt{r}} - \max_{i \in p(j)} \{\eta^i(r) + \frac{v^i_{[r.]}}{\sqrt{r}}\}\right)_t \tag{2.27}$$

From (2.25), (2.27), the continuous mapping theorem and the converging together theorem, we conclude that as $r \uparrow \infty$,

$$(\eta^1(r), \ldots, \eta^{B_1 + \ldots + B_{l+1}}(r), \zeta^1(r), \ldots, \zeta^B(r)) \Rightarrow (\eta^1, \ldots, \eta^{B_1 + \ldots + B_{l+1}}, \zeta^1, \ldots, \zeta^B) \tag{2.28}$$

as $r \uparrow \infty$, which completes the induction step. ∎

9.3 Admissibility

In this section our objective is to prove the admissibility of the average response time measure for acyclic synchronized networks. Consider the following sample space $(\Omega, I\!F)$, where Ω is the set of infinite sequences $\{(\tau_n, v_n^1, \ldots, v_n^B, d_n^1, \ldots, d_n^B)\}_0^\infty$. Here τ_n has the interpretation of the arrival time of the n^{th} batch, $v_n^j, 1 \leq j \leq B$ has the interpretation of the service time of the n^{th} customer that is sent to the single server queue j, while $d_n^j, 1 \leq j \leq B$ has the interpretation of the service time of the n^{th} customer that is sent to the infinite server queue j. We introduce a measure $I\!P_\lambda$ on $(\Omega, I\!F)$ such that the arrival process under this measure is a Poisson process with parameter $\lambda > 0$. For each ω in Ω we add a tagged batch which arrives at time zero and whose service times $\hat{v}^j, \hat{d}^j, 1 \leq j \leq B$, are independent of $\{v_n^j\}_0^\infty, \{d_n^j\}_0^\infty, 1 \leq j \leq B$, but have the same distribution. In order to do so, we define an augmented probability space $(\Omega', I\!F', Q_\lambda)$, such that for each ω' in Ω', we have $\omega' = (\omega, (\hat{v}^1, \ldots, \hat{v}^B, \hat{d}^1, \ldots, \hat{d}^B))$, where ω is an element of Ω. Let

$$T = \text{response time of batch entering at } t = 0 \tag{3.1}$$

and set

$$\psi^{(n)} = I\!E_{Q_\lambda}[T^n \mid I\!F]. \qquad n = 1, 2 \ldots \tag{3.2}$$

We now show that show that the RV $\psi^{(n)}$ is admissible so that one may obtain the light traffic limits for the system by using the formulae in Theorem B1 (Appendix B). For θ in $I\!R$, set

$$M_j(\theta) = I\!E[e^{\theta v_n^j}], \quad \overline{M}_j(\theta) = I\!E[e^{\theta d_n^j}] \quad 1 \leq j \leq B. \tag{3.3}$$

Assume that

(IXd): There exists $\theta^* > 0$ such that

$$\prod_{j=1}^{K} M_j(\theta)\overline{M}_j(\theta) < \infty, 1 \leq j \leq B, \theta < \theta^*.$$

Theorem 9.3.1 *If Assumption* **(IXd)** *is satisfied, then* $\psi^{(n)}$ *as defined in (3.2) is admissible.*

Proof. We introduce an $M/GI/1$ queue (defined on the same probability space as the network) which upper bounds the acyclic synchronized network and is itself admissible. The arrival and service sequences in the $M/GI/1$ queue are given by $\{u_n\}_0^\infty$ and $\{\sum_{j=1}^{B}(v_n^j + d_n^j)\}_0^\infty$ respectively. It is clear that as long as there is work remaining in the system, it works at least as fast as the bounding $M/GI/1$ queue. The $M/GI/1$ queue is admissible provided assumption **(IXd)** is satisfied. Now proceeding as in Theorem B3 (Appendix B), it can be easily shown that the synchronized network is admissible under assumption **(IXd)**. ∎

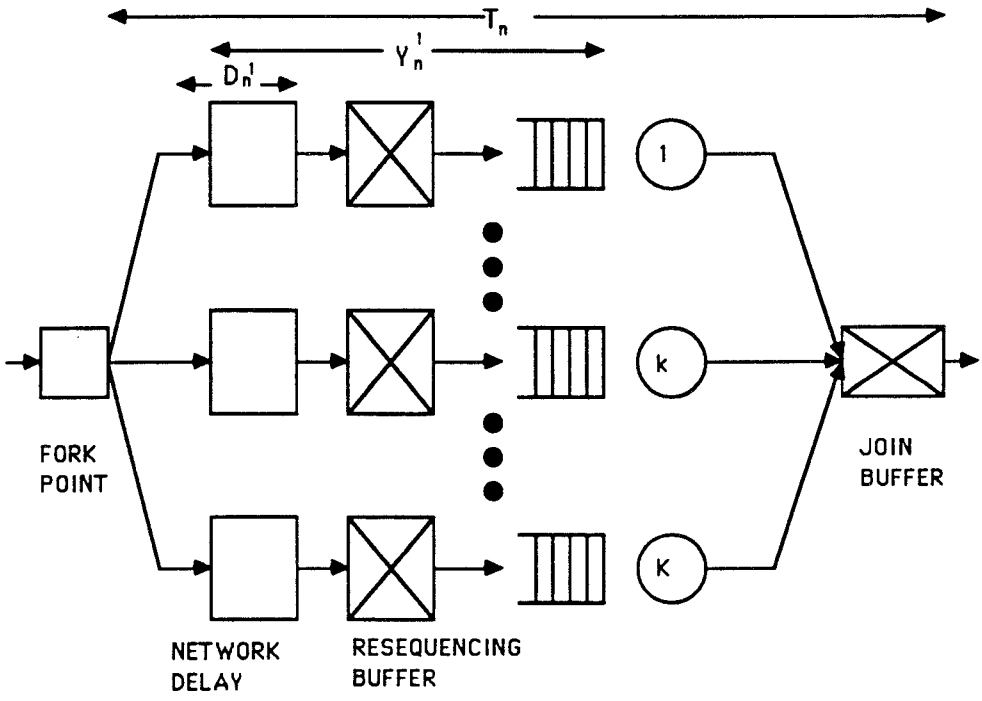


Fig. 9.2. The time–stamp ordering system.

9.4 Approximations for the time–stamp ordering model

The following queueing system (Fig. 9.2), was introduced by Baccelli [3] to model the time–stamp ordering, consistency preserving scheme in a distributed database. The noteworthy feature of this model is that it exhibits fork–join as well as resequencing mechanisms.

The model operates as follows: Consider K single server queues operating in parallel, which are fed by renewal process that sends customers to each queue simultaneously, i.e., by a fork primitive. However the customers do not enter the single server queues directly, but first enter a disordering system after which they are resequenced and are finally sent to the buffer of the single server queue. After a customer belonging to the n^{th} arrival batch finishes service in one of the single server queues, it waits in a join buffer until all the other customers from that batch have completed service, at which point the n^{th} batch leaves the system, thereby

realizing the join primitive. Note that this model is a special case of the acyclic synchronized network from Section 9.2.

We shall assume that the batches arrive into the system according to a renewal process with rate λ, the K infinite server queues have service times with the same rate ν, and the K single server queues also have service times with the same rate μ. Results from Section 9.2 suggest when $\lambda \approx \mu$, i.e., when the system is in heavy traffic, it has the same limiting diffusion limit as the K–dimensional fork–join system which was analyzed in detail in Part I of this thesis. This fact can be used to obtain a heavy traffic approximation for the average end-to-end delay of the time stamp ordering model in the following way: Consider a K dimensional fork–join system with arrival rate λ and service rate μ in all the queues. The following result was given in Chapter 6. If we denote the average end–to–end delay for this system by $\bar{t}_K(\lambda)$, then

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\bar{t}_K(\lambda)$$

$$= \left[H_K + (4V_K - 3H_K - 1)\beta + 2(1 + H_K - 2V_K)\beta^2\right] \frac{\sigma^2 + \sigma_0^2}{2}\mu^2, \quad 0 \leq \beta \leq 1.$$

$$K = 2, 3 \ldots (4.1)$$

where β and V_K are defined in (5.2.4) and (6.4.26) respectively, and σ and σ_0 are the limiting variances of the service and inter-arrival distributions, respectively. Now if we precede each queue with disordering and resequencing, then using results from Section 9.2, the heavy traffic limit for the average end-to-end delay (denoted by $\overline{T}_K(\lambda)$) is unchanged so that for the time stamp ordering model the average end–to–end delay satisfies

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda)$$

$$= \left[H_K + (4V_K - 3H_K - 1)\beta + 2(1 + H_K - 2V_K)\beta^2\right] \frac{\sigma^2 + \sigma_0^2}{2}\mu^2,$$

$$0 \leq \beta \leq 1 \quad K = 2, 3 \ldots (4.2)$$

For the special case when the inter–arrival and service times are exponentially

distributed, (4.2) simplifies to

$$\lim_{\lambda \uparrow \mu}(\mu - \lambda)\overline{T}_K(\lambda) = V_K. \qquad K = 2, 3 \ldots (4.3)$$

We now proceed to obtain light traffic limits for $\overline{T}_K(\lambda)$ for the case when the inter–arrival, disordering and service distributions are all exponential with rates λ, ν and μ respectively. Admissibility for this system was proven in Theorem 9.3.1. We first calculate $\overline{T}_K(0)$. Consider the batch arriving at $t = 0$ into an empty system. Let $d_1, \ldots, d_K$ be its disordering delays and $s_1, \ldots, s_K$ be its service times at the K queues. Since this batch does not experience interference from any other customer, i.e., it does not experience any queueing or resequencing delay, it is clear that

$$\overline{T}_K(0) = I\!E \max_{1 \le k \le K}(d_k + s_k). \qquad (4.4)$$

Each of the RVs $d_k + s_k, 1 \le k \le K$, has a common distribution F, given by

$$F(z) = 1 + \frac{\mu}{\nu - \mu}e^{-\nu z} - \frac{\nu}{\nu - \mu}e^{-\mu z}, \quad z \ge 0 \qquad (4.5)$$

so that

$$\begin{aligned}
\overline{T}_K(0) &= \int_0^\infty \left[1 - (1 + \frac{\mu}{\nu - \mu}e^{-\nu z} - \frac{\nu}{\nu - \mu}e^{-\mu z})^K\right] dz \\
&= \int_0^\infty \left[1 - \sum_{r=0}^{K}\binom{K}{r}[\frac{\mu}{\nu - \mu}e^{-\nu z} - \frac{\nu}{\nu - \mu}e^{-\mu z}]^r\right] dz \\
&= \sum_{r=1}^{K}\binom{K}{r}\sum_{m=0}^{r}\binom{r}{m}(-1)^{m+1}(\frac{\mu}{\nu - \mu})^{r-m}(\frac{\nu}{\nu - \mu})^m \\
&\quad \times \int_0^\infty e^{-(m\mu + (r-m)\nu)z}dz \\
&= \sum_{r=1}^{K}\binom{K}{r}\sum_{m=0}^{r}\binom{r}{m}(-1)^{m+1}(\frac{\mu}{\nu - \mu})^{r-m}(\frac{\nu}{\nu - \mu})^m \frac{1}{m\mu + (r-m)\nu}.
\end{aligned}$$

Let $L_K(\mu, \nu)$ denote the right hand side of this last equation, so that

$$\overline{T}_K(0) = L_K(\mu, \nu). \qquad (4.6)$$

Tables for $L_K(1,2), 1 \leq K \leq 10$ are given in Section 9.5.

We now proceed to calculate $\overline{T}'_K(0)$. Let

$$T(t, d_1, \ldots, d_K, s_1, \ldots, s_K, \overline{d}_1, \ldots, \overline{d}_K, \overline{s}_1, \ldots, \overline{s}_K)$$

be the response time of the batch that arrives at time $t = 0$ with service times $s_1, \ldots, s_K$ and disordering delays $d_1, \ldots, d_K$ given that another customer arrives at time t with disordering delays $\overline{d}_1, \ldots, \overline{d}_K$ and service time $\overline{s}_1, \ldots, \overline{s}_K$. It is not difficult to see that

$$\begin{aligned} & T(t, d_1, \ldots, d_K, s_1, \ldots, s_K, \overline{d}_1, \ldots, \overline{d}_K, \overline{s}_1, \ldots, \overline{s}_K) \\ & = \begin{cases} \max_{1 \leq k \leq K}(d_k + s_k), & \text{if } t > 0 \\ \max_{1 \leq k \leq K}[\max(d_k, t + \overline{d}_k + \overline{s}_k) + s_k] & \text{if } t \leq 0. \end{cases} \end{aligned} \tag{4.7}$$

Define the RVs $X_k, 1 \leq k \leq K$, by

$$X_k = t + \overline{d}_k + \overline{s}_k. \tag{4.8}$$

Then it can be shown that each $X_k, 1 \leq k \leq K$, has distribution F_X given by

$$F_X(x) = 1 - \frac{\nu}{\nu - \mu} e^{\mu(t-x)} + \frac{\mu}{\nu - \mu} e^{\nu(t-x)}, \quad x \geq t. \tag{4.9}$$

Next define the RVs $Y_k, 1 \leq k \leq K$, by

$$Y_k = \max(d_k, X_k). \tag{4.10}$$

Since the RVs d_k and X_k are independent for $1 \leq k \leq K$, each RV $Y_k, 1 \leq k \leq K$, has the distribution F_Y given by

$$\begin{aligned} F_Y(x) =& [1 - \frac{\nu}{\nu - \mu} e^{\mu(t-x)} + \frac{\mu}{\nu - \mu} e^{\nu(t-x)}](1 - e^{-\nu x}) \\ =& 1 - \frac{\nu}{\nu - \mu} e^{\mu(t-x)} + \frac{\mu}{\nu - \mu} e^{\nu(t-x)} - e^{-\nu x} \\ & + \frac{\nu}{\nu - \mu} e^{\mu t} e^{-(\nu+\mu)x} - \frac{\mu}{\nu - \mu} e^{\nu t} e^{-2\nu x}, \quad x \geq 0. \end{aligned} \tag{4.11}$$

Lastly define the RVs $R_k, 1 \leq k \leq K$, by

$$R_k = Y_k + s_k. \tag{4.12}$$

Taking into account that the RVs s_k and Y_k are independent for $1 \leq k \leq K$, it can be shown that each $R_k, 1 \leq k \leq K$, has the distribution F_R given by

$$\begin{aligned} F_R(x) =& 1 + \frac{\mu}{\nu-\mu}e^{-\nu x} - \frac{\nu}{\nu-\mu}e^{-\mu x} - \frac{\nu\mu}{\nu-\mu}e^{\mu t}xe^{-\mu x} \\ &+ \frac{\mu^2(3\nu-2\mu)}{(\nu-\mu)^2(2\nu-\mu)}e^{\nu t}e^{-\mu x} - \frac{\mu^2}{(\nu-\mu)^2}e^{\nu t}e^{-\nu x} \\ &+ \frac{\mu}{\nu-\mu}e^{\mu t}e^{-\mu x} - \frac{\mu}{\nu-\mu}e^{\mu t}e^{-(\nu+\mu)x} \\ &- \frac{\mu^2}{(2\nu-\mu)(\nu-\mu)}e^{\nu t}e^{-2\nu x}, \quad x \geq 0. \end{aligned} \tag{4.13}$$

Note (4.7), (4.8), (4.10) and (4.12), that

$$T = \max_{1 \leq k \leq K} R_k, \quad t \leq 0 \tag{4.14}$$

where the left hand side of (4.7) has been abbreviated to T. Since the RVs $R_k, 1 \leq k \leq K$, are independent, we obtain that

$$I\!P(T \leq x) = \prod_{k=1}^{K} I\!P(R_k \leq x) = F_R^K(x), \quad x \geq 0. \tag{4.15}$$

Proceeding as in Section 6.5, we can show after some calculations that

$$\overline{T}'_K(0) = -\int_{t=-\infty}^{0}\int_{x=0}^{\infty}\sum_{r=1}^{K}\binom{K}{r}U^{K-r}V^r dxdt \tag{4.16}$$

where

$$U = 1 + \frac{\mu}{\nu-\mu}e^{-\nu x} - \frac{\nu}{\nu-\mu}e^{-\mu x} \tag{4.17}$$

and

$$V = \frac{\mu^2(3\nu-2\mu)}{(\nu-\mu)^2(2\nu-\mu)}e^{\nu t}e^{-\mu x} - \frac{\nu\mu}{\nu-\mu}e^{\mu t}xe^{-\mu x}$$
$$- \frac{\mu^2}{(\nu-\mu)^2}e^{\nu t}e^{-\nu x} + \frac{\mu}{\nu-\mu}e^{\mu t}e^{-\mu x}$$
$$- \frac{\mu}{\nu-\mu}e^{\mu t}e^{-(\nu+\mu)x} - \frac{\mu^2}{(2\nu-\mu)(\nu-\mu)}e^{\nu t}e^{-2\nu x}. \tag{4.18}$$

It can be shown that

$$U^{K-r} = \sum_{m_1=0}^{K-r}\binom{K-r}{m_1}\sum_{m_2=0}^{m_1}\binom{m_1}{m_2}(-1)^{m_2}(\frac{\mu}{\nu-\mu})^{m_1-m_2}(\frac{\nu}{\nu-\mu})^{m_2}$$
$$\times e^{-(\nu(m_1-m_2)+\mu m_2)x} \tag{4.19}$$

and

$$V^r = \sum_{k_1=0}^{r}(-1)^{k_1}\binom{r}{k_1}\sum_{k_2=0}^{r-k_1}\binom{r-k_1}{k_2}(\frac{\mu}{\nu-\mu})^{r-k_1-k_2}$$
$$\times [\frac{\mu^2(3\nu-2\mu)}{(\nu-\mu)^2(2\nu-\mu)}]^{k_2}e^{(\mu(r-k_1-k_2)+\nu k_2)t}e^{-\mu x(r-k_1)}$$
$$\times \sum_{k_3=0}^{k_1}\binom{k_1}{k_3}\sum_{k_4=0}^{k_1-k_3}\binom{k_1-k_3}{k_4}(\frac{\nu\mu}{\nu-\mu})^{k_4}[\frac{\mu^2}{(\nu-\mu)^2}]^{k_1-k_3-k_4}$$
$$\times e^{(\mu k_4+\nu(k_1-k_3-k_4)t}x^{k_4}e^{-(\mu k_4+\nu(k_1-k_3-k_4))x}$$
$$\times \sum_{k_5=0}^{k_3}\binom{k_3}{k_5}(\frac{\mu}{\nu-\mu})^{k_5}[\frac{\mu^2}{(\nu-\mu)(2\nu-\mu)}]^{k_3-k_5}$$
$$\times e^{(\mu k_5+\nu(k_3-k_5))t}e^{-(\mu k_5+\nu(2k_3-k_5))x}. \tag{4.20}$$

From (4.16), (4.19) and (4.20) it follows that

$$\overline{T}'_K(0)$$
$$= -\sum_{r=1}^{K}\binom{K}{r}\sum_{m_1=0}^{K-r}\binom{K-r}{m_1}\sum_{m_2=0}^{m_1}\binom{m_1}{m_2}(-1)^{m_2}(\frac{\mu}{\nu-\mu})^{m_1-m_2}(\frac{\nu}{\nu-\mu})^{m_2}$$

$$\times \sum_{k_1=0}^{r} (-1)^{k_1} \binom{r}{k_1} \sum_{k_2=0}^{r-k_1} \binom{r-k_1}{k_2} (\frac{\mu}{\nu-\mu})^{r-k_1-k_2} [\frac{\mu^2(3\nu-2\mu)}{(\nu-\mu)^2(2\nu-\mu)}]^{k_2}$$

$$\times \sum_{k_3=0}^{k_1} \binom{k_1}{k_3} \sum_{k_4=0}^{k_1-k_3} \binom{k_1-k_3}{k_4} (\frac{\nu\mu}{\nu-\mu})^{k_4} [\frac{\mu^2}{(\nu-\mu)^2}]^{k_1-k_3-k_4}$$

$$\times \sum_{k_5=0}^{k_3} \binom{k_3}{k_5} (\frac{\mu}{\nu-\mu})^{k_5} [\frac{\mu^2}{(\nu-\mu)(2\nu-\mu)}]^{k_3-k_5}$$

$$\times \int_0^{\infty} x^{k_4} e^{-(\mu(r+m_2-k_1+k_4+k_5)+\nu(m_1-m_2+k_1+k_3-k_4-k_5))x} dx$$

$$\times \int_{-\infty}^{0} e^{(\mu(r-k_1-k_2+k_4+k_5)+\nu(k_1+k_2-k_4-k_5))t} dt$$

$$= -\sum_{r=1}^{K} \binom{K}{r} \sum_{m_1=0}^{K-r} \binom{K-r}{m_1} \sum_{m_2=0}^{m_1} \binom{m_1}{m_2} (-1)^{m_2} (\frac{\mu}{\nu-\mu})^{m_1-m_2} (\frac{\nu}{\nu-\mu})^{m_2}$$

$$\times \sum_{k_1=0}^{r} (-1)^{k_1} \binom{r}{k_1} \sum_{k_2=0}^{r-k_1} \binom{r-k_1}{k_2} (\frac{\mu}{\nu-\mu})^{r-k_1-k_2} [\frac{\mu^2(3\nu-2\mu)}{(\nu-\mu)^2(2\nu-\mu)}]^{k_2}$$

$$\times \sum_{k_3=0}^{k_1} \binom{k_1}{k_3} \sum_{k_4=0}^{k_1-k_3} \binom{k_1-k_3}{k_4} (\frac{\nu\mu}{\nu-\mu})^{k_4} [\frac{\mu^2}{(\nu-\mu)^2}]^{k_1-k_3-k_4}$$

$$\times \sum_{k_5=0}^{k_3} \binom{k_3}{k_5} (\frac{\mu}{\nu-\mu})^{k_5} [\frac{\mu^2}{(\nu-\mu)(2\nu-\mu)}]^{k_3-k_5}$$

$$\times \frac{k_4!}{[\mu(r+m_2-k_1+k_4+k_5)+\nu(m_1-m_2+k_1+k_3-k_4-k_5)]^{k_4+1}}$$

$$\times \frac{1}{\mu(r-k_1-k_2+k_4+k_5)+\nu(k_1+k_2-k_4-k_5)}. \tag{4.21}$$

We shall denote the right hand side of (4.21) as $G_K(\mu,\nu)$ so that

$$\overline{T}'_K(0) = G_K(\mu,\nu). \tag{4.22}$$

Tables for $G_K(1,2), 2 \leq K \leq 10$ are given in Section 9.5.

Finally, combining (4.3), (4.6) and (4.22), we obtain the following first order

approximation to the average response time of the time stamp ordering model,

$$\hat{T}_K(\lambda) = \frac{\mu L_K(\mu,\nu)}{\mu-\lambda} + [\mu G_K(\mu,\nu) - L_K(\mu,\nu)]\,\frac{\lambda}{\mu-\lambda} + \left[V_K - \mu^2 G_K(\mu,\nu)\right](\frac{\lambda}{\mu})^2 \frac{1}{\mu-\lambda}, \quad 0 \le \lambda < \mu. \tag{4.23}$$

This approximation agrees extremely well with simulation results (see Section 9.4.2).

9.4.2 Simulation results

In this section, approximation (4.23) is compared with simulation results for the case $\mu = 1, \nu = 2$ and $K = 2, 3, 5$ and 10.

λ	$\overline{T}_2(\lambda)$	$\hat{T}_2(\lambda)$	% Error
0.1	2.25 ± 0.008	2.23	0.89
0.2	2.46 ± 0.012	2.43	1.22
0.3	2.73 ± 0.016	2.67	2.19
0.4	3.08 ± 0.017	3.00	2.59
0.5	3.56 ± 0.036	3.46	2.80
0.6	4.27 ± 0.062	4.15	2.81
0.7	5.40 ± 0.112	5.30	1.88
0.8	7.59 ± 0.24	7.59	0.00
0.9	14.54 ± 0.31	14.47	0.48

λ	$\overline{T}_3(\lambda)$	$\hat{T}_3(\lambda)$	% Error
0.1	2.64 ± 0.007	2.63	0.38
0.2	2.88 ± 0.011	2.85	1.04
0.3	3.17 ± 0.015	3.14	0.95
0.4	3.57 ± 0.022	3.52	1.40
0.5	4.11 ± 0.036	3.89	5.35
0.6	4.92 ± 0.059	4.86	1.22
0.7	6.31 ± 0.111	6.19	1.90
0.8	9.13 ± 0.295	8.85	3.07
0.9	17.63 ± 1.23	16.82	4.59

λ	$\overline{T}_5(\lambda)$	$\hat{T}_5(\lambda)$	% Error
0.1	2.64 ± 0.007	2.63	0.38
0.1	3.14 ± 0.009	3.14	0.06
0.2	3.41 ± 0.012	3.40	0.32
0.3	3.76 ± 0.017	3.73	0.80
0.4	4.22 ± 0.026	4.18	0.99
0.5	4.87 ± 0.041	4.88	0.20
0.6	5.83 ± 0.068	5.73	1.71
0.7	7.39 ± 0.12	7.28	1.48
0.8	10.44 ± 0.28	10.39	0.57
0.9	19.96 ± 0.39	19.69	1.35

λ	$\overline{T}_{10}(\lambda)$	$\hat{T}_{10}(\lambda)$	% Error
0.1	3.84 ± 0.008	3.84	0.08
0.2	4.16 ± 0.011	4.14	0.48
0.3	4.57 ± 0.016	4.53	0.87
0.4	5.11 ± 0.019	5.05	1.17
0.5	5.86 ± 0.026	5.78	1.36
0.6	6.99 ± 0.076	6.88	1.57
0.7	8.89 ± 0.15	8.70	2.13
0.8	12.73 ± 0.37	12.34	3.06
0.9	24.25 ± 0.43	23.26	4.08

9.5 Tables

K	$L_K(1,2)$	$G_K(1,2)$
2	2.08	1.39
3	2.45	1.62
4	2.72	1.77
5	2.93	1.88
6	3.10	1.97
7	3.25	2.04
8	3.38	2.10
9	3.49	2.15
10	3.60	2.19

APPENDIX A

The principle results regarding the weak convergence of probability measures are given in this appendix. Most of these are results are borrowed from Billingsley [8] to which the reader is referred to for proofs and further details. Let (S, m) be a metric space and let $\mathcal{I\!I}$ be the $\sigma-$ field generated by the open sets in S.

Definition A1. *Consider a sequence of probability measures* $\{P_n\}_0^\infty$ *as well as a single probability measure* P *defined on* $\mathcal{I\!I}$. *If these probability measures satisfy*

$$\int_S f dP_n \to \int_S f dP$$

for every bounded, continous fuction f *on* S, *we say that* P_n *converges weakly to* P *and write* $P_n \Rightarrow P$.

Definition A2. *A probabilty measure* P *on* $(S, \mathcal{I\!I})$ *is tight if for each positive* ϵ *there exists a compact set* K *such that* $P(K) > 1 - \epsilon$.

A stochastic process X is a measurable mapping from probability space $(\Omega, \mathcal{I\!F}, I\!P)$ into S. The distribution of X is the probability measure $P = I\!PX^{-1}$ on $(S, \mathcal{I\!I})$. We shall say that a sequence of stochastic processes $\{X_n\}_0^\infty$ converges weakly to a stochastic process X, and write

$$X_n \Rightarrow X$$

if the distribution P_n of X_n converges weakly to the distribution P of X.

Definition A3. *A sequence of stochastic processes* $\{X_n\}_0^\infty$ *converges in probability to* X *if* X_n *and* X *are defined on a common probability space and for all* $\epsilon > 0$.

$$I\!P\{m(X_n, X) \geq \epsilon\} \to 0.$$

When X is a constant process (non-random), convergence in probability is equivalent to weak convergence. In such casej we shall write $m(X_n, X) \Rightarrow 0$ or $X_n \Rightarrow X$. If X_n and Y_n have a common domain, we also write $m(X_n, Y_n) \Rightarrow 0$ when for all $\epsilon > 0$, $I\!P\{m(X_n, Y_n) \geq \epsilon\} \to 0$.

The following result is Theorem 4.1 in Billingsley [8].

Theorem A1. (The converging together theorem.) *If* $X_n \Rightarrow X$ *and* $m(X_n, Y_n) \Rightarrow 0$, *then* $Y_n \Rightarrow X$.

Suppose h is a measurable mapping of S into S', a second metric space with Borel sets $I\!I'$. Each probability measure P on $(S, I\!I)$ induces on $(S', I\!I')$ a unique probability measure $Ph^{-1}(A) = P(h^{-1}A)$ for A in $I\!I'$. Let D_h be the set of discontinuities of h.

The following result is Theorem 5.1 in Billingsley [8].

Theorem A2. (The continous mapping theorem). *If* $X_n \Rightarrow X$ *and* $I\!P\{X \in D_h\} = 0$, *then* $h \circ X_n \Rightarrow h \circ X$.

Two function spaces have received the greatest attention in th weak convergence literature; the space of continous functions on $[0, 1]$ and the space of functions on $[0, 1]$ having only jump discontinuities. These spaces are the natural ones for the sample paths of most processes which arise in applied probability.

Let $C[0, 1]$ denote the space of all continous real-valued functions on $[0, 1]$ with the metric of uniform convergence,

$$\rho(x, y) = \sup\{|\, x(t) - y(t)\, |\colon 0 \leq t \leq 1\},$$

and Borel sets $I\!C$. With this metric $C[0, 1]$ is a complete separable metric space.

Let $D[0, 1]$ be the space of all real valued, right continous functions on $[0, 1]$ having left limits. In order to describe the metric for $D[0, 1]$ we let Λ denote the class of strictly increasing, continous maps of $[0, 1]$ onto itself. For λ in Λ, $\lambda(0) = 0$

and $\lambda(1) = 1$. Think of λ as being a new time scale. For λ in Λ let

$$|| \lambda ||= \sup_{s \neq t} | \log \frac{\lambda(t) - \lambda(s)}{t - s} |$$

and define $d(x, y)$ as follows:

$$d(x, y) = \inf\{\epsilon > 0 :|| \lambda ||\leq \epsilon \text{ and } \rho(x, y.\lambda) \leq \epsilon, \text{ for some } \lambda \in \Lambda\}.$$

The function d is a metric on $D[0, 1]$ which renders it a complete separable metric space. This metric generates the Skorokhod topology [8] which relativized to $C[0, 1]$ coincides with the uniform topology. Let the Borel sets of $D[0, 1]$ be denoted by $\mathbb{D}$.

We now state Prohorovs functional central limit theorem which first appeared as Theorem 3.1 in Prohorov [54]. Given a double sequence $\{X_r(i)\}, i, r = 1, 2 \ldots$, of RVs assume that

(A1).

(i): $X_1(i), X_2(i), \ldots$ are independent and identically distributed for each $i \geq 1$, and defined on some probability space $(\Omega, \mathbb{F}\mathbb{P})$.

(ii): $\mathbb{E}[X_r(i)] = m_X(i) \to m_X < \infty$ as $i \to \infty$. If $m_X(i) \neq 0$ and $m_X \neq 0$, we let $\frac{1}{\mu_X(i)} = m_X(i)$ and $\frac{1}{\mu_X} = m_X$.

(iii): $0 < Var[X_r(i)] = \sigma_X^2(i) \to \sigma_X^2, 0 < \sigma_X^2 < \infty$, as $i \to \infty$.

(iv): $\mathbb{E}[| X_r(i) |^{2+\epsilon}]$ is bounded in i for some positive ϵ.

Theorem A3. (Prohorovs Theorem). *If the double sequence of RVs* $\{X_r(i)\}, i, r = 1, 2 \ldots$, *satisfies Assumption* **(A1)** *and if for each* $t \in [0, 1]$ *we define in* $D[0, 1]$

$$\xi_r^i(t) = \frac{S_{[rt]}(i) - m_X(i)[rt]}{\sqrt{r}}$$

where $S_j(i) = X_1(i) + \ldots + X_j(i), S_0(i) = 0$ *and* $[.]$ *is the largest integer function, then*

$$\xi_r^i \Rightarrow \sigma_X \xi$$

in $D[0,1]$ *as* $i,r \to \infty$ *in general manner. Here* ξ *denotes the Wiener process.*

For a sequence $\{X_r(i)\}, i,r = 1,2\ldots$, of non-negative RVs we define a sequence of counting renewal processes $\{N^i(t); t \geq 0, i = 1,2\ldots\}$ as follows

$$N^i(t) = \begin{cases} \max\{n > 0 \mid \sum_{k=1}^{n} X_k(i) \leq t\} & \text{if } X_1(i) \geq t, \\ 0 & \text{if } X_1(i) < t. \end{cases}$$

We then have the following result whoose proof may be found in Kyprianou [45; Theorem 1].

Theorem A4. (Functional central limit theorem for renewal processes). *If the double sequence of non-negative RVs* $\{X_r(i)\}, i,r = 1,2\ldots$, *satisfies Assumption* **(A1)** *and if for each* t *in* $[0,1]$ *we define in* $D[0,1]$

$$\hat{N}_r^i(t) = \frac{N^i(rt) - rt\mu_X(i)}{\sqrt{r}}$$

then

$$\hat{N}_r^i \Rightarrow \sqrt{\sigma_X^2 \mu_X^3}\xi$$

as $i,r \to \infty$ *in general manner.*

The proof of the following useful result that was used repeatedly throughout the thesis may be found in Kyprianou [45; Lemma 3].

Theorem A5. *If the sequence* $\{X_r(i), i,r = 1,2\ldots$ *of RVs satisfies Assumption* **(A1)** *then as* $i,r \to \infty$ *in a general manner*

$$\sup_{0 \leq n \leq r} \frac{X_n(i)}{\sqrt{r}} \xrightarrow{I\!P} 0, \quad 0 \leq t \leq 1.$$

The following discussion that clarifies the relationship between the function spaces $C[0,1]$ and $D[0,1]$ is taken from Iglehart and Whitt [29]. Analysis is much easier in $C[0,1]$ as compared to $D[0,1]$, but many processes of interest are not continous and must be regarded as elements of $D[0,1]$. The standard procedure

has been to consider linearly interpolated versions of such processes which will be in $C[0,1]$, but the following result due to Ligget and Rosen [29] shows that the analysis in $C[0,1]$ may often be used for processes in $D[0,1]$.

Theorem A6. *Let $\{X_n\}_0^\infty$ be a sequence of stochastic processes in $D[0,1]$, $\{Y_n\}_0^\infty$ a sequence of stochastic processes in $C[0,1]$. If $d(X_n, Y_n) \Rightarrow 0$, then $X_n \Rightarrow X$ in $D[0,1]$ iff $Y_n \Rightarrow X$ in $C[0,1]$.*

As a consequence of this Theorem $A6$ the functional central limit theorems for stocastic processes induced in $D[0,1]$ by sequences of partial sums or renewal processes are equivalent to the corresponding theorems for the linearly-interpolated stochastic processes in $C[0,1]$.

We now state a result that gives $C[0,1]-$ tightness for sequences of stochastic processes in $D[0,1]$. Knowing that a sequence of stochastic processes converges weakly in $D[0,1]$, we often want to use the resulting tightness for other arguments. The main condition for $C[0,1]-$ tightness is expressed in terms of the modulus of continuity, $w(\delta) : C[0,1] \rightarrow I\!R$, defined for any x in $C[0,1]$ by

$$w_x(\delta) = \sup_{0 \leq s,t \leq 1} \sup_{|s-t|<\delta} \mid x(t) - x(s) \mid .$$

Theorem A7. *Let $\{X_n\}_0^\infty$ be a sequence of stochastic processes in $D[0,1]$, and let X be a stochastic process such that $P(X \in C[0,1]) = 1$. If $X_n \Rightarrow X$, then $\{X_n\}_0^\infty$ is $C[0,1]-$ tight: for all positive ϵ and η, there exists a δ $(0 < \delta < 1)$ and an integer n_0 such that*

$$P\{w_{X_n}(\delta) \geq \epsilon\} \leq \eta$$

for all $n \geq n_0$.

APPENDIX B

This appendix contains some of the principle results from the theory of light traffic approximations for queueing systems. For proofs and further details the reader may consult Reiman and Simon [59].

In this theory, we are concerned with functions of a marked Poisson process, $\{T_n, X_n)\}$, where the markings, $\{X_n\}$, chosen from a mark space $(X, \mathbb{B})$, are iid, and are independent of the Poisson process $\{T_n\}$. In a queueing context, the Poisson process will usually constitute the arrival process, and x in X is the description of a customer, which would include its service times, priorty levels, or any other information needed to specify its behaviour in the system. The marked Poisson process is defined on $(\Omega, \mathbb{F})$, where Ω is the set of all finite and infinite sequences $\{(\tau_n, x_n), -M_1 < n < M_2\}, 0 \leq M_1, M_2 \leq \infty$, that satisfy x_n belongs to X, and $\ldots < \tau_{-2} < \tau_{-1} \leq 0 < \tau_0 < \tau_1 < \ldots$. The counting process associated with ω in Ω is given by $N_t(\omega) = \inf\{n : T_n(\omega) \geq t\}$. Thus $N_{t_2} - N_{t_1}$ is the number of points in $[t_1, t_2)$.

The $\sigma-$ algebra, $\mathbb{F}$, is generated by subsets of the form $\{N_{t_2} - N_{t_1} = k\}, -\infty < t_1 < t_2 < \infty, k \geq 0$ and $\{x_k \in B\}, B \in \mathbb{B}, -\infty < k\infty$. We can decompose $\mathbb{F}$ into $\hat{\mathbb{F}}$ and $\mathbb{G}$, where $\hat{\mathbb{F}}$ is generated by $\{N_{t_2} - N_{t_1} = k\}$ and $\mathbb{G}$ is generated by $\{x_k \in B\}$. Let $\mathbb{F}_T \subset \mathbb{F}$ be generated by subsets of the form $\{N_{t_2} - N_{t_1} = k\}, t \in (-T, T), k \geq 0$ and $\{x_n \in B\} \cap \{|\ t_n\ |, T\}, B \in \mathbb{B}$, $-\infty < n < \infty$; i.e. $\mathbb{F}_T$ measures the process on the interval $(-T, T)$. In addition, $\hat{\mathbb{F}}_T \subset \mathbb{F}_T$ will denote the family of $\sigma-$ algebras generated by $\{N_t - N_{-T} = k\}$. Note that $\hat{F}$ measures the Poisson process and $\hat{\mathbb{F}}_T$ measures the Poisson process on the interval $(-T, T)$.

Let $\mathbb{P}_\lambda$ be a measure on $(\Omega, \mathbb{F})$ such that $\{N_t, -\infty < t < \infty\}$ is a Poisson

process under $I\!P_\lambda$, and

$$I\!P_\lambda(A_1 \cap A_2) = I\!P_\lambda(A_1) I\!P_\lambda(A_2), \quad A_1 \in \hat{I\!F}, A_2 \in I\!G.$$

In other words, $(\Omega, \hat{I\!F}, I\!P_\lambda)$ is a probability space for a Poisson process intensity λ.

For $\omega \in \Omega$, define ω_T to be ω excluding all points outside the interval $(-T, T)$. Thus ω_T satisfies

$$N_t(\omega_T) = \begin{cases} N_{-T}(\omega), & t \leq -T, \\ N_t(\omega), & \mid t \mid < T, \\ N_T(\omega), & t \geq T. \end{cases}$$

Let $\psi : \Omega \to I\!R$ be $I\!F$ measurable and define $\psi_T : \Omega \to I\!R$ by $\psi_T(\omega) = \psi(\omega_T)$. Note that ψ_T is $I\!F_T-$ measurable. Let

$$\overline{\psi} = I\!E(\psi \mid \hat{I\!F}) \text{ and } \overline{\psi}_T = I\!E(\psi_T \mid \hat{I\!F}_T).$$

We are concerned with functions, $f(\lambda)$, which are constructed as limits as $T \to \infty$ of $f_T(\lambda)$, where

$$f_T(\lambda) = \int \overline{\psi}_T dI\!P_\lambda = \int \psi_T dI\!P_\lambda.$$

To insure that $f(\lambda)$ is well defined, with derivatives at $\lambda = 0$, we require that the RV ψ, has a certain type of regularity. In order to define this 'admissibility' condition, introduce the following additional notation. For $0 < x < y, l \leq 0$, we define $A_{(x,y]}(l) \in \hat{I\!F}$ by

$$A_{(x,y]}(l) = \{\omega : l \text{ arrivals in the set } [-y, -x) \cup (x, y]\}.$$

Definition B1. *The RV ψ is admissable if there exist constants $K, N < \infty, 1 < a < \infty$, and $\theta > 0$ such that for any $0 < T < S$,*

$$I\!E\big(\mid \psi_T - \psi_S \mid\mid A_{(0,T]}(j), A_{(T,S]}(l)\big) \leq K(j+l)^N a^{j+l} \exp(-\theta t). \qquad (B1)$$

Admissability assures that $f(\lambda)$ and all its derivatives can be obtained as limits as $T \to \infty$ of $f_T(\lambda)$ and its derivatives for λ in a neighborhood of zero.

Let $\overline{\psi}(\{t_1, \ldots, t_n\})$ correspond to $\overline{\psi}(\omega)$ when ω consists of n points $\{t_1, \ldots, t_n\}$. Similarly, let $\psi(\emptyset)$ be the value of $\overline{\psi}(\omega)$ when $\omega \in \Omega_0$ (i.e. no points at all).

Let $\{\binom{n}{j}\}$ be the set of $j-$ tuples chosen from $\{1, 2 \ldots, n\}$, and for $\pi = \{i_1, \ldots, i_k\} \in \{\binom{n}{k}\}$, let $t_\pi = \{t_{i_1}, \ldots, t_{i_k}\}$. Define the function

$$\Psi(\{t_1, \ldots, t_n\}) = \sum_{j=0}^{n} (-1)^{n-j} \sum_{\pi \in \{\binom{n}{j}\}} \overline{\psi}(\{t_\pi\}).$$

The formulas for the derivatives are given by the following result.

Theorem B1. *If ψ is abmissible, then $f(0) = \hat{\psi}(\emptyset)$, and for $n \geq 1$,*

$$f^{(n)}(0) = \int_{t_1=-\infty}^{\infty} \cdots \int_{t_n=-\infty}^{\infty} \Psi(\{t_1, \ldots, t_n\}) dt_1 \ldots dt_n. \qquad (B2)$$

Reiman and Simon proved that several important functions of the $M/G/1$ queue are admissible, and using this fact they were able to prove admisibility of these functions in general open queueing networks. The procedure that we employ in Chapters 6, 8 and 9 for proving admissabilty for the response times for fork-join and resequencing systems is very similar, since we use the fact that certain queueing networks (which are special cases of general queueing networks) are admissable in order to obtain our results. We now proceed to give the statement and detailed proof of admissibility for the $M/G/1$ case since some of the constructs in the proof are used in the thesis to prove admissibility for fork-join and resequencing systems.

The mark space X for the $M/G/1$ queue is $\mathbb{R}_+$, the non-negative reals. For each $\omega \in \Omega$ we add an extra (tagged) customer who arrives at time zero, and whose mark X^*, is independent of the X_n's but has the same distribution. Let

$$M(\theta) = \mathbb{E}[\exp(\theta X_n)], \quad \theta \in \mathbb{R}.$$

Let the RV Q denote the number of customers in the system at $t = 0$ and let W denote the sojourn time of a customer who enters at $t = 0$. If $\psi = Q^N$, then $f(\lambda)$ is the N^{th} moment of the queue length distribution and likewise for $\psi = E[W^N \mid X^*]$.

Theorem B2. *If there exists a $\theta^* > 0$ such that $M(\theta) < \infty$ for $\theta \leq \theta^*$, and ψ is defined as*

(a) $\psi = Q^N$,

(b) $\psi = I\!E[W^N \mid X^*]$,

for $1 \leq N \leq \infty$, then ψ is admissable.

The following lemma is needed in the proof of Theorem B2. Since the proof of the lemma is not important for the discussion in the thesis, we omit it.

Lemma B1. *Let X be non-negative RV such that there exists a $\theta^* > 0$ with $I\!E[\exp(\theta x)] < \infty$ for $\theta \leq \theta^*$. Let Z be a RV having an exponential distribution with parameter θ^*, let $C = \frac{\log I\!E[\exp(\theta^* X)]}{\theta^*}$, and define $Y = C + Z$. Then, for $t \geq 0$,*

$$X \leq_{st} Y.$$

In addition it is possible to construct a probability space $(\Omega, I\!F, P)$ on which $\overline{X}$ and $\overline{Y}$ having the same distributions as x and Y respectively are defined, such that $\overline{X} \leq \overline{Y}$ for every $\omega \in \Omega$.

Proof of Theorem B2. Fix $0 < T < S$. Let

$$Z_S(T) = \{\omega : \text{with arrivals turned off outside } [-S, S] \text{ the server is idle at}$$

$$\text{some point during both } [-T, 0) \text{ and } (0, T]\}. \tag{B3}$$

We can write

$$I\!E(\mid \psi_T - \psi_S \mid\mid A_{(0,T]}(j), A_{(T,S]}(l))$$

$$= I\!E(\mid \psi_T - \psi_S \mid\mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S(T)) \times I\!P(Z_S(T) \mid A_{(0,T]}(j), A_{(T,S]}(l))$$

$$+ I\!E(\mid \psi_T - \psi_S \mid\mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S^C(T)) \times I\!P(Z_S^C(T) \mid A_{(0,T]}(j), A_{(T,S]}(l)). \tag{B4}$$

Note that

$$I\!E(\mid \psi_T - \psi_S \mid\mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S(T)) = 0$$

so that only the second term in (B4) remains.

Renumber customers so that the first to enteron or after $-S$ is 1. Let $D(t) = N_t - N_{-S}, -S \leq t \leq S$, and set $V_n = \sum_{i=1}^{n} X_i, n \geq 1$. Since the worst case for keeping a server busy continualy during an interval is to have all arrivals during that interval enter at the beginning, we have

$$I\!P(Z_S^C(T) \mid A_{(0,T]}(j), A_{(T,S]}(l)) \leq 2I\!P\{V_{D(S)} > T \mid A_{(0,T]}(j), A_{[T,S)}(l)\}$$

and

$$I\!P\{V_{D(S)} > T \mid A_{(0,T]}(j), A_{[T,S)}(l)\} = \begin{cases} I\!P\{V_{j+l} > T\} & \text{for } (a), \\ I\!P\{V_{j+l+1} > T\} & \text{for } (b). \end{cases}$$

For $\theta \leq \theta^*$, we have $I\!E[\exp(\theta V_K)] = [M(\theta)]^k$. By Chebychevs inequality we have

$$I\!P\{V_k > T\} \leq [M(\theta^*)]^k \exp(-\theta^* T),$$

so that

$$I\!P\{Z_S^C(T) \mid A_{(0,T]}(j), A_{[T,S)}(l)\} \leq \begin{cases} [M(\theta^*)]^{j+l} \exp(-\theta^* T) & \text{for } (a), \\ [M(\theta^*)]^{j+l+1} \exp(-\theta^* T) & \text{for } (b). \end{cases} \quad (B5)$$

Equation (B1) is satisfied for case (a) due to the bound

$$\mid \psi_T - \psi_S \mid \leq D(S), \quad (B6)$$

so that Q^N is admissable.

Case (b) is more involved since the conditioning can change the expectation. By Lemma B1, we can construct a 'companion' system on the same probability space as the original system which has service times $\{Y_i, i = 1, 2 \ldots\}$ such that $X_i \leq Y_i, i = 1, 2 \ldots$, and $Y_i = C + Z_i, C = \frac{\log M(\theta^*)}{\theta^*}$, and Z_i is an exponentially distributed RV with parameter θ^*. Let $\overline{\mu}_N$ denote the N^{th} moment of Y_1.

Let $\overline{V}_n = \sum_{i=1}^{n} Y_i$. For case (b) we have

$$\psi_S \leq [V_{D(S)+1}]^N, \tag{B7}$$

so that

$$I\!E[\psi_S \mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S^C(T)] \leq I\!E[V_{D(S)+1}^N \mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S^C(T)]$$

$$\leq I\!E[\overline{V}_{D(S)+1}^N \mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S^C(T)]$$

The memoryless property for the Z_i's yields

$$\begin{aligned}
&I\!E[\overline{V}_{D(S)+1}^N \mid A_{(0,T]}(j), A_{(T,S]}(l), Z_S^C(T)] \\
&\leq I\!E[(2T + \overline{V}_{D(S)+1})^N \mid A_{(0,S]}(j+l)] \\
&\leq \sum_{k=0}^{N} \binom{N}{k} (2T)^{N-k} E(\overline{V}_{j+l})^k \qquad (B8) \\
&\leq 2^{2N} T^N [j+l+1]^N \overline{\mu}_N \leq (8T)^N [j+l]^N \overline{\mu}_N \\
&= K_N T^N (j+l)^N.
\end{aligned}$$

Since $T^N \exp(-\theta T) \leq \overline{K}_N \exp(-\frac{\theta T}{2})$ for some $\overline{K} < \infty$, the result follows. ∎

In the next step the admissibility property is extended to general queueing networks with priorties and feedback. The greatest complication involved in doing so is the burdensome notation involved in describing the system.

Consider a queueing network with $1 \leq K < \infty$ stations, each with an infinite waiting room, where station k has L_k servers, $1 \leq k \leq K$. Server l at station k works at deterministic rate μ_{kl}. Cuustomers arrrive to the network in a Poisson process with rate λ. There are $J \geq 1$ classes of customers; the probability that an arriving customer is from class j is $p_j, 1 \leq j \leq J$. If we let j_i denote the class of the i^{th} customer to enter the system, then $\{j_i, i = 1, 2 \ldots\}$ is iid, with $P(j_i = l) = p_l$.

A customers behavior in the network is determined by an itinerary, which is a random vector. For each class, itineraries form an iid sequence. The sequences

corresponding to different classes are independent. The i^{th} class j customer's itinerary is

$$\{R_j^i;(k_{j1}^i,\nu_{j1}^i,r_{j1}^i),\ldots,(k_{jR_j^1}^i,\nu_{jR_j^i}^i,r_{jR_j^i}^i)\},\quad 1\le j\le J \qquad i=1,2\ldots$$

The number of 'stops' in the itinerary is given by R_j^i. The m^{th} stop is described by $(K_{jm}^i,\nu_{jm}^i r_{jm}^i)$, where k_{jm}^i is the station, ν_{jm}^i is the customers service requirement, and r_{jm}^i is the customers priorty. The priorty discipline can be either pre-emptive resume or non-pre-emptive, with cuszomers of the same priorty level being served in order of arrival to the station. Customers at the head of the queue are served by the next available server, with an arbitrary choice being allowed when more than one server is free.

Let

$$V_i^j=\sum_{m=1}^{R_j^i}\nu_{jm}^i,\quad 1\le j\le J \qquad i=1,2\ldots$$

denote the total service time of the i^{th} class j customer, and define $M_j(\theta)=\mathbb{E}[\exp(\theta V_j^1]$. Let Q denote a generic queue length at $t=0$, that is, the number if a specific type at a specific station, or a sum over a set of types and/or stations. Similarly, W will denote a generic sojourn time, which may consist of any subset of a customers itinerary. In addition W may be conditioned on a specific type, or a set of types.

Theorem B3. *If there exists a $\theta^*>0$ such that $M_j(\theta)<\infty$ for $1\le j\le J,\theta\le\theta^*$, and ψ is defined as*

(a) $\psi=Q^N$,

(b) $\psi=\mathbb{E}[W^N\mid X^*]$,

for $1\le N\le\infty$, then ψ is admissable.

Proof. We introduce an $M/G/1$ queue (defined on the same probability space as the network) which 'bounds' the network in an appropriate manner, and which is itself admissable. Let

$$\overline{\mu}=\min_{1\le k\le K}\min_{1\le l\le L_k}\mu_{kl}.$$

Define

$$\overline{V}_i^j = \frac{V_i^j}{\overline{\mu}}, \quad 1 \le j \le J, \qquad\qquad i = 1, 2 \ldots$$

We define J(dependent) discrete renewal processes by

$$P_l(n) = \sum_{i=1}^{n} 1_{j_i = l}, \quad 1 \le l \le J \qquad\qquad n = 1, 2 \ldots$$

so that $P_l(n)$ denotes the number of type l customers among the first n customers entering the system. Finally we define $\overline{V}_i = \overline{V}_{j_i}^{P_{j_i}(i)}$ as the service time of the i^{th} customer to enter the bounding $M/G/1$ queue. Defining $\overline{M}(\theta) = I\!E[\exp(\theta \overline{V}_1)]$, we have

$$\overline{M}(\theta) = \sum_{j=1}^{J} p_j M_j(\frac{\theta}{\overline{\mu}}).$$

The bounding $M/G/1$ queue is thus admissable with $\overline{\theta}^* = \overline{\mu}\theta^*$.

Let $\overline{Z}_S(T)$ be as defined in (B3) for the bounding $M/G/1$ queue, and define

$$Z_S(T) = \{\omega : \text{with arrivals turned off outside } [-S, S] \text{ all servers are idle at}$$

$$\text{some point during both } [-T, 0) \text{ and } (0, T]\}. \qquad (B9)$$

When there is any work remaining in the network, it is working at least as fast as the bounding $M/G/1$ queue. Hence , $\overline{Z}_S(T) \subset Z_S(T)$. Thus (B5) holds for the network (with θ^* replaced by $\overline{\theta}^*$, and M replaced by $\overline{M}$). The bound (B6) still holds and (B8) goes through as before. ∎

APPENDIX C

The following result known as the multi-dimensional Ito formula for semi-martingales is Theorem 5.10 in Chung and Williams [12].

Theorem C1. *Let $m, n \in \mathbb{N}$. Let M^i be a continous local martingale for $1 \leq i \leq m$, and V^k be a continous process which is locally of bounded variation for $1 \leq k \leq n$. Suppose that D is a domain in $\mathbb{R}^{m+n}$ such that a.s.*

$$Z_t = (M_t^1, \ldots, M_t^m, V_t^1, \ldots, V_t^n)$$

takes values in D for all t. Let $f(x,y)$ be a continous real-valued function of $(x,y) \in D$ such that $\frac{\partial f}{\partial x_i}, \frac{\partial^2 f}{\partial x_i \partial x_j}, 1 \leq i,j \leq m$, and $\frac{\partial f}{\partial y_k}, 1 \leq k \leq n$, exist and are continous in D. Then a.s. we have for all t:

$$\begin{aligned} f(Z_t) - f(Z_0) = & \sum_{i=1}^{m} \int_0^t \frac{\partial f}{\partial x_i}(Z_s) dM_s^i \\ & + \sum_{k=1}^{n} \int_0^t \frac{\partial f}{\partial y_k}(Z_s) dV_s^k \\ & + \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \int_0^t \frac{\partial^2 f}{\partial x_i \partial x_j}(Z_s) d < M^i, M^j >_s . \end{aligned}$$

REFERENCES

[1] F. Baccelli, E Gelenbe and B. Plateau,"An end-to-end approach to the resequencing problem," *JACM*, Vol. 31, No. 3, pp. 474-485 (1984).

[2] F. Baccelli,"Two parallel queues created by arrivals with two demands," *INRIA Rapport de Recherche*, No. 426, (1985).

[3] F. Baccelli,"A queueing model for timestamp ordering in a distributed system," *Performance '87*, Brussels, pp. 413-431 (1987).

[4] F. Baccelli, A.M. Makowski and A. Shwartz, "Fork-Join queue and related systems with synchronization constraints: Stochastic ordering, approximations and computable bounds," Electrical Engineering Technical Report, No. TR-87-01 University of Maryland, College Park (1987).

[5] F. Baccelli, W.A. Massey and D. Towsley, "Acyclic fork-join queueing networks," *INRIA Rapport de Recherche*, No. 688 (1987).

[6] F. Baccelli and A.M. Makowski,"Queueing systems with synchronization constraints," *Proceeding of the IEEE*, pp. 138–161, (1989).

[7] R.E. Barlow and F. Proschan, *Statistical theory of reliability and life testing*, Holt, Rinehart and Winston, Reading, MA.

[8] P. Billingsley, *"Convergence of probability measures,"* J. Wiley and Sons, New York, (1968).

[9] P. Billingsley, *"Probability and measure,"* J. Wiley and Sons, New York, (1979).

[10] A.A. Borovkov,"Some limit theorems in the theory of mass service II," *Theor. Probability Appl*, Vol. 10, pp. 375-400 (1965).

[11] A.A. Borovkov, *Asymptotic methods in queueing theory*, J. Wiley and Sons, New York (1984).

[12] K.L. Chung and R.J. Williams, *Introduction to stochastic integration*,

Birkhauser, Boston (1983).

[13] E.G. Coffman Jr. and M.I. Reiman,"Diffusion approximations for computer/
communication systems," *Mathematical Computer Performance and Reliability*, G. Iazeolla, P.J. Courtois and A. Hordijk (editors), North Holland (1984).

[14] L. Flatto and S. Hahn,"Two parallel queues created by arrivals with two demands I," *SIAM J. Appl. Math.*, Vol. 44, pp. 1041-1053 (1984).

[15] G.J. Foschini and J. Salz, "A basic dynamic routing problem and diffusion," *IEEE Trans. on Comm.*, Vol. 26, No. 3, pp. 320-327 (1978).

[16] G.J. Foschini, "Equilibra for diffusion models of pairs of communicating computers
-Symmetric case," *IEEE Trans. on IT*, Vol. 28, No. 2, pp. 273-284 (1982).

[17] L. Gün and A. Jean-Marie,"Parallel queues with resequencing," Manuscript, University of Maryland, (1989).

[18] B. Hajek,"Mean stochastic comparison of diffusions," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, Vol. 68, pp. 315-329 (1985).

[19] J.M. Harrison, "The heavy traffic approximation for single server queues in series," *J. Appl. Prob.*, No. 10, pp. 613-629 (1973).

[20] J.M. Harrison,"A limit theorem for priority queues in heavy traffic," *J. Appl. Prob.*, No. 10, pp. 907-912 (1973).

[21] J.M. Harrison,"The diffusion approximation for tandem queues in heavy traffic," *Adv. Appl. Prob.*, No. 10, pp. 886-905 (1978).

[22] J.M. Harrison and M.I. Reiman,"On the distribution of multidimensional reflected brownian motion," *SIAM J. Appl. Math.*, Vol. 41, No. 2, pp. 345-361 (1981).

[23] J.M. Harrison and M.I. Reiman, "Reflected Brownian motion on an orthant," *Ann. Prob.*, Vol. 9, No. 2, pp. 302-308 (1981).

[24] J.M. Harrison, *Brownian motion and stochastic flow systems*, J. Wiley and Sons, New York (1985).

[25] J.M. Harrison and R.J. Williams,"Brownian models of open queueing networks with homogenous customer population," *Stochastics*, Vol. 22, pp. 77-115 (1987).

[26] J.M. Harrison and R.J. Williams,"Multidimensional reflected Brownian motions having exponential stationary distributions," *Ann. Prob.*, Vol. 15, pp. 115-137 (1987).

[27] E. Horlatt and D. Mailles,"Etude du resequencement dans un reseau de files d'attente" *Technical Report No. 125, Universitie P. et M. Curie*, (1986).

[28] D.L. Iglehart, "Limiting diffusion approximations for the many server queue and the repairman problem," *J. Appl. Prob.* No 2, pp. 429-441 (1965).

[29] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic. I," *Adv. Appl. Prob.*, No. 2, pp. 150-177 (1970).

[30] D.L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic. II: Sequences, networks and batches," *Adv. Appl. Prob.*, No. 2, pp.355-369 (1970).

[31] D.L. Iglehart, "Weak convergence in queueing theory," *Adv. Appl. Prob.*, Vol. 5, pp. 570-594 (1973).

[32] A. Jean-Marie,"Load balancing in a system of two queues with resequencing," *Performance '87*, Brussels, pp. 75-88 (1988).

[33] S. Karlin and H.M. Taylor, *A first course in stochastic processes*, Academic Press, New York (1975).

[34] S. Karlin and H.M. Taylor, *A second course in stochastic processes*, Academic Press, New York (1981).

[35] J.F.C. Kingman, "The single server queue in heavy traffic," *Proc. Cambridge Philos. Soc.* No. 57, pp. 902-904 (1961).

[36] J.F.C. Kingman, "On queues in heavy traffic," *J. Roy. Statist. Soc. B.* Vol. 25, pp. 383-392 (1962).

[37] J.F.C. Kingman, "The heavy traffic approximation in the theory of

queues," *Proceedings of the Symposium on Congestion Theory*, W.L. Smith and W.E. Wilkinson, eds. University of North Carolina Press. pp. 137-159 (1965).

[38] J.F.C. Kingman, "Queue disciplines in heavy traffic," *Math. of Oper. Res.*, Vol. 7, No. 2, pp. 262-271 (1982).

[39] L. Kleinrock, *Queueing Systems Vol I: Theory*, J. Wiley and Sons, New York (1975).

[40] L. Kleinrock, *Queueing Systems Vol II: Computer Applications*, J. Wiley and Sons, New York (1975).

[41] C. Knessl," On the diffusion approximation to a fork and join queueing model," *Technical Report No. AM88-02*, Dept. of Mathematics, Statistics and Computer Science, University of Chicago at Illinois (1988).

[42] H. Kobayashi, "Applications of the diffusion approximation to queueing networks I: Equilibrium queue distributions," *JACM*, Vol. 21, No. 2, pp. 316-328 (1974).

[43] H. Kobayashi, "Applications of the diffusion approximation to queueing networks II: Nonequilibrium distributions and applications to computer modelling," *JACM*, Vol. 21, No. 3, pp. 459-469 (1974).

[44] J. Kollerstrom,"Heavy traffic theory for queues with several servers. I," *J. Appl. Prob.*, Vol. 11, pp. 544-552 (1974).

[45] E. Kyprianou, "The virtual waiting time of the $GI/G/1$ ueue in heavy traffic," *Adv. Appl. Prob.*, No. 3, pp. 249-268 (1971).

[46] A.M. Law and W.D. Kelton, *Simulation modelling and analysis*, McGraw Hill, New York (1982).

[47] A.J. Lemoine, "Networks of queues-A survey of weak convergence results," *Manag. Sci.*, Vol. 24, No. 11, pp. 1175-1193 (1978).

[48] D.V. Lindley,"The theory of queues with a single server," *J. Appl. Prob.*, Vol. 10, pp. 109-121 (1973).

[49] T. Lindvall,"Weak convergence of probability measures and random functions in the function space $D[0,\infty)$," *J. Appl. Prob.*, Vol. 10, pp. 109-121

(1973).

[50] R. Loulou," On the extension of some heavy traffic theorems to multiple channel systems," Lecture Notes in Economics and Mathematical Systems, No. 98, Springer-Verlag, Berlin, pp. 185-197 (1974).

[51] R.M. Loynes,"The stability of a queue with non-dependent inter-arrival and service times," *Proc. Cambridge Philos. Soc* No. 58, pp. 497-520 (1962).

[52] R. Nelson and A.N. Tantawi,"Approximate analysis of fork-join synchronization in parallel queues," *IEEE Trans. on Computers*, Vol. 37, No. 6, pp. 739-743 (1988).

[53] K.R. Parthasarthy, *Probability measures on metric spaces*, Academic Press, New York (1967).

[54] Y. Prohorov,"Convergence of random processes and limit theorems of probability theory," *Theor. Probability Appl.*, Vol. 1, pp. 157-214 (1956).

[55] Y. Prohorov,"Transient phenomena in processes of mass service," *Litovsk. Mat. Sb.*, Vol. 3, pp. 199-205 (1963).

[56] M.I. Reiman, "Open queueing networks in heavy traffic," *Maths. of Oper. Res.*, Vol. 9, No. 3, pp. 441-458 (1984).

[57] M.I. Reiman and B. Simon, "An interpolation approximation for queueing systems with Poisson input," *Oper. Res.*, Vol. 36, No. 3, pp. 454-469 (1988).

[58] M.I. Reiman and B. Simon, "Light traffic limits of sojourn time distributions in Markovian queuing networks," *Commun. Statist.-Stochastic Models*, Vol. 4, No. 2, pp. 191-233 (1988).

[59] M.I. Reiman and B. Simon, "Open queueing systems in light traffic," *Maths. of Oper. Res.*, Vol. 14, No. 1, pp. 26-59 (1989).

[60] D. Stoyan, *Comparison methods for queues and other stochastic models*, English Translation (D.J. Daley, Editor), J. Wiley and Sons, New York, (1984).

[61] D.W. Stroock and S.R.S. Varadhan,"Diffusion processes with boundary

conditions," *Comm. in Pure and Appl. Math*, Vol. XXIV, pp. 147-225 (1971).

[62] S.R.S. Varadhan and R.J. Williams,"Brownian motian in a wedge with oblique refleczion," *Comm. in Pure and Appl. Math*, Vol. XXXVIII, pp. 405-443 (1985).

[63] S. Varma, "Some problems in queueing systems with resequencing," MS Thesis, University of Maryland; also available as SRC Technical Report No. TR-87-192 (1987).

[64] A. Weiss,"Invariant measures of diffusion processes on domains with boundaries," Ph.D. Dissertation, Dept. of Mathematics, Ney York University.

[65] W.Whitt,"Weak convergence theorems for queues in heavy traffic," Ph.D. Thesis, Cornell University (1968).

[66] W. Whitt,"Weak convergence of probability measures on the function space $C[0,\infty)$." *Ann. Math. Stat.*, Vol. 41, No. 3, pp. 939-944 (1970).

[67] W. Whitt, "Multiple channel queues in heavy traffic. III: Random server selection," *Adv. Appl. Prob.*, No. 2, pp. 370-375 (1970).

[68] W. Whitt,"Heavy traffic limit theorems for queues: A survey," Lecture Notes in Economics and Mathematical Systems, No. 98, Springer-Verlag, Berlin, pp. 307-350 (1974).

[69] W. Whitt,"Some useful functions for functional limit theorems," *Maths. of Oper. Res*, Vol. 5, No. 1, pp. 67-85 (1980).

[70] W. Whitt,"Performance of the queueing network analyzer," *The Bell Syst. Tech. Journ.*, Vol. 62, No. 9, pp. 2817-2843 (1983).

[71] W. Whitt,"On approximations for queues, I: Extremal distributions," *AT&T Bell Lab. Tech. Journ.*, Vol. 63, No. 1, pp. 115-138 (1984).

[72] W. Whitt,"On approximations for queues, II: Shape constraints," *AT&T Bell Lab. Tech. Journ.*, Vol. 63, No. 1, pp. 139-161 (1984).

[73] W. Whitt,"On approximations for queues, III: Exponential distributions," *AT&T Bell Lab. Tech. Journ.*, Vol. 63, No. 1, pp. 163-

175 (1984).

[74] R.J. Williams,"Recurrence classification and invariant measure for reflected Brownian motion in a wedge," *Ann. of Prob.*, Vol. 13, No. 3, pp. 758-778 (1985).

[75] R.J. Williams,"Reflected Brownian motion with skew symmetric data in a polyhedral domain," *Probab. Th. Rel. Fields*, Vol. 75, pp. 459-485 (1987).