L. Accardi
C.C. Heyde (Editors)

Probability Towards 2000
PREFACE

This volume is the end product of a Symposium titled *Probability Towards 2000* held at Columbia University, New York from October 2-6, 1995. The Symposium was generously sponsored by the Istituto dell’Enciclopedia Italiana in Rome and organized in New York through the cooperation of the Centro Vito Volterra, University of Rome, Tor Vergata, the Italian Academy for Advanced Study in America at Columbia and the Center for Applied Probability at Columbia.

A key objective of the Symposium was to obtain a broad view of probability and where the subject is heading. This matter was addressed both through 34 talks and at a round table discussion on the last afternoon, at which it was decided to produce this volume of perspectives. The intention is to chart a course ahead for probability and versions of selected conference talks plus some additional commissioned material are included.

To elucidate the perspective of the Symposium and this volume, the Manifesto of the Symposium, written by L. Accardi, and an Opening Address to the Symposium by J.L. Teugels, then President of the Bernoulli Society for Mathematical Statistics and Probability are included in the Preface.

GOAL OF THE SYMPOSIUM: MANIFESTO

Luigi ACCARDI

The fact that nowadays there exists no scientific field, from biology to economics, from physics to social sciences, from medicine to complexity theory, from meteorology to decision theory,... in which probability theory does not play a major role, should not let one forget that only the period between the two world wars marks the definitive entrance of probability theory among the fundamental mathematical disciplines such as geometry, analysis, algebra, .... In these years P. Lévy, A.N. Kolmogorov and N. Wiener opened the way to the establishment of strong connections of probability theory with several branches of classical mathematics: combinatorial theory, classical analysis, in particular measure theory, elliptic and parabolic equations, potential
## CONTENTS

Preface ........................................................................................................... v

ACCARDI, L., LU, Y.-G., and VOLOVICH, I. Non-linear extensions of classical and quantum stochastic calculus and essentially infinite dimensional analysis ......................... 1

ARNOLD, L. Trends and open problems in the theory of random dynamical systems .......... 34

BARNDORFF-NIELSEN, O.E. Probability and Statistics: Selfdecomposability, finance and turbulence ................................................................. 47

BIERLANT, J., TRÜGELS, J.L. and VYNCKIER, P. Some thoughts on extreme values .... 58

BELAVKIN, V.P. Stochastic positive flows and quantum filtering dynamics ................. 74

CHEN, L.H.Y. Stein's method: Some perspectives with applications .......................... 97

CHEN, M.-F. Trilogy of couplings and general formulas for lower bound of spectral gap 123

COLLET, P. Stochasticity and chaos .................................................................. 137

DE LA PEÑA, V.H. Decoupling inequalities: A second generation of martingale inequalities 151

ELWORTHY, K.D. Stochastic differential equations on manifolds ........................... 165

FAGNOLLA, F. Extending flows of classical Markov processes to quantum flows in Fock space 179

FUDSON, R.L. Quantum stochastic analysis after four decades ............................ 194

KENDALL, W.S. Perfect simulation for the area interaction point-process ................. 218

MEYER, P.A. Quantum probability seen by a classical probabilist ......................... 235

NUALART, D. Stochastic anticipating calculus .................................................. 249

OHYA, M. Foundation of entropy, complexity and fractals in quantum systems ....... 263

RÖCKNER, M. Dirichlet forms on infinite dimensional 'manifold-like' state spaces: A survey of recent results and some prospects for the future .......................... 287

SINHA, K.B. Quantum stochastic calculus and applications – a review .................. 307

THORISSON, H. Coupling .......................................................... 319

WILLIAMS, R.J. Some recent developments for queueing networks ...................... 340
SOME RECENT DEVELOPMENTS FOR QUEUEING NETWORKS

R. J. WILLIAMS,* University of California, San Diego

1. Introduction

Early investigations in queueing theory provided detailed analysis of the behavior of a single queue and of networks that in a sense could be decomposed into a product of single queues.Whilst insights from these early investigations are still used, more recent investigations have focused on understanding how network components interact.

In particular, queueing network models are of current relevance for analyzing congestion and delay in computer systems, communication networks and complex manufacturing systems (see e.g., [1, 54, 66]). Many of these systems have stations that can process more than one class of customer or job (so-called multiclass networks) and/or have complex feedback structures. For example, in a computer communication network one may have voice, video and data being transmitted through each node in the network, and in the manufacture of semiconductor wafers, a job may return to the same machine several times for different stages of processing. Generally such systems cannot be analyzed in closed form and frequently are heavily loaded. One method for analyzing the performance of such systems is to approximate them by more tractable objects.

A certain class of diffusion processes, known as reflecting Brownian motions, have been shown to approximate normalized versions of the queue length or workload processes in single class queueing networks and some multiclass queueing networks under conditions of heavy traffic, i.e., when the networks are heavily loaded (see [65] for an overview). There is now a substantial theory for these diffusion processes which are generally more tractable than the original queueing networks, although some open problems remain (see [64] for a recent survey).

One of the outstanding challenges in contemporary research on approximate models for queueing networks is to understand which multiclass networks can be approximated by reflecting Brownian motions in heavy traffic and to prove limit theorems justifying such approximations. In the last few years there have been some surprises both with regard to conditions for the stability of multiclass queueing networks and to the behavior of such networks in heavy traffic. The aim of this paper is to describe some of the recent developments in this area.

The paper is organized as follows. In section 2 a heavy traffic approximation for a single class tandem queue is described. The purpose of this is to illustrate in a

* Postal address: Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, La Jolla CA 92037-0112, USA.
Research supported in part by NSF Grant GER 5923335.

Some recent developments for queueing networks 341

concrete setting (a) how stability is quantified, and (b) the form of the heavy traffic approximation. In section 3, a brief synopsis of heavy traffic limit theory for single class and some multiclass queueing networks is given. One may extrapolate from this, as was done in [22, 24, 25, 31], to conjecture a general form for the heavy traffic limit of a wide variety of multiclass networks. In section 4, an example similar to one given by Dai and Wang [16] illustrates that care is required with such an extrapolation, in the sense that not all multiclass queueing networks have a heavy traffic approximation of the form conjectured in [22, 24, 25, 31]. A variety of explanations might be proposed for the failure of the approximation in this case. Two possible explanations that might be proposed are (a) the example is not stable, i.e., one has the wrong notion of heavy traffic, or (b) the example may satisfy a different kind of limit theorem than the "conventional" one proposed in [22, 24, 25, 31]. Indeed, contemporaneous work on stability of multiclass networks (see e.g., [42, 44, 53, 3, 4, 55]), illustrates that the problem of determining conditions for the stability of multiclass networks with feedback is more complex than previously supposed. Furthermore, work of Harrison and Williams [32] (see also section 4 of [65]), shows that not all multiclass networks with feedback have conventional heavy traffic behavior. The paper concludes with a final section on open problems.

2. A Single Class Tandem Queue

Consider the tandem queueing network pictured in Fig. 1 under the following assumptions. Customers (or jobs) arrive at station 1 from outside the system according to a renewal process where the i.i.d. interarrival times are assumed to have positive finite mean $1/\alpha$ and finite variance $\sigma^2$ (a is thus the (long run average) external arrival rate). There is a single server at each of the two stations and the service times at station 1 are assumed to be i.i.d. with positive finite mean $m_1$ and variance $\sigma_1^2$, for $i = 1, 2$. The sequences of interarrival and service times are assumed to be mutually independent. Customers are served on a first-in-first-out (FIFO) basis at each station. After receiving service at station 1, a customer goes next to station 2. Upon completing service there, with probability $p \in (0, 1)$, the customer is routed back to join the end of the queue at station 1, and with probability $1 - p$ the customer exits the system. Such a tandem queue might be used to model a simple processing facility where a completed job requires total rework with probability $p$.

Congestion is measured through the behavior of the two-dimensional queue length process $Q(t) = (Q_1(t), Q_2(t))$, where for $i = 1, 2$, $Q_i(t)$ is the number of customers at station $i$ (waiting and being served) at time $t$. With the general distributional assumptions described above, this system cannot be analyzed exactly. However, when it is stable, but heavily loaded, one can approximate a normalized version of the queue length process by a reflecting Brownian motion living in the positive two-dimensional quadrant.

To describe this approximation, the first question that arises is what does "stable and heavily loaded" mean? Here it will be taken to mean that the system is stable and near the boundary between stability and instability, where "stable" means that the mean queue lengths are bounded for all time. Stability can be quantified in terms of the station level traffic intensity parameters $\rho_i, i = 1, 2$, defined as follows. Let $\lambda$
be the unique solution of the traffic equation

$$\lambda = \alpha + p\lambda.$$  

The traffic intensity parameters are defined by

$$\rho_i = \lambda m_i, \quad i = 1, 2.$$  

The system is stable if and only if

$$\rho_i < 1, \quad \text{for } i = 1, 2,$$

(see e.g., [47]). Assuming this holds, one can interpret $\lambda$ as the long run average rate at which customers visit stations 1 and 2 (it is the same for each station because of the tandem structure of the network). Then the flow balance equation (1) is a natural consistency condition. Also, the traffic intensity parameter $\rho_i$ can be interpreted as the long run average rate at which work (measured in units of required service time) arrives at station $i$, $i = 1, 2$.

A heavy traffic limit theorem for this system may be described as follows. Consider a sequence of systems indexed by $n$ (with associated parameters and processes having a superscript of $(n)$), all with the same common structure as described above except that the exogenous arrival rate $\alpha^{(n)}$ for the $n^{th}$ system tends to a value $\alpha$ in such a way that the traffic intensity vector $\rho^{(n)} = (\rho_1^{(n)}, \rho_2^{(n)})$ for the $n^{th}$ system tends to the vector $(1, 1)$ in the following manner:

$$\sqrt{n}(\rho_i^{(n)} - 1) \rightarrow c_i \quad \text{as} \quad n \rightarrow \infty,$$

where $c_i$ is a finite negative constant for $i = 1, 2$. (The negativity of the $c_i$ guarantees that the heavy traffic limit will be positive recurrent [29].) Normalize the two-dimensional queue length process $Q^{(n)}$ for the $n^{th}$ system using a central limit theorem (or diffusion) type of scaling:

$$Q^{(n)}(n \cdot) = \frac{Q^{(n)}(n \cdot)}{\sqrt{n}}.$$  

Some recent developments for queuing networks

The process $Q^{(n)}$ takes values in the space of two-dimensional r.c.l.l. (right continuous with finite left limits) paths defined on $(0, \infty)$. When this space is endowed with the usual Skorokhod $J_1$-topology [56, 20], $Q^{(n)}$ converges in distribution as $n \rightarrow \infty$ to a reflecting Brownian motion $Z$ that lives in the positive quadrant and which has a semimartingale decomposition of the form

$$Z = X + RY,$$

where $X$ is a two-dimensional Brownian motion with constant drift $c$ and non-degenerate covariance matrix

$$\Gamma = \begin{bmatrix} \sigma_1^2 + \mu_1^2 \sigma_1^2 + \mu_2(1 - p + p\mu_2 \sigma_2^2) & -\mu_1^2 \sigma_1^2 + \mu_2^2 \sigma_2^2 \\ -\mu_1^2 \sigma_1^2 + \mu_2^2 \sigma_2^2 & \mu_1^2 \sigma_1^2 + \mu_2^2 \sigma_2^2 \end{bmatrix},$$

for $\mu_i = 1/m_i$, $i = 1, 2$.

$$R = \begin{bmatrix} 1 & -p \\ -1 & 1 \end{bmatrix}$$

is called the reflection matrix, and $Y$ is a two-dimensional continuous non-decreasing process that starts from the origin and is such that $Y_i$ can increase only when $Z_i$ is zero (in fact, $m_i Y_i$ is the limit in distribution of the normalized cumulative idletime process for station $i$, where the normalization is the same central limit theorem type of scaling (5) used for the queue length processes). Informally, the behavior of the reflecting Brownian motion $Z$ may be described as follows. In the interior of the quadrant, $Z$ behaves like the Brownian motion $X$. The erratic movements of this Brownian motion are the limit of the up and down movements of the queue length processes corresponding to arrivals and departures,
respectively. The process $Z$ is confined to the quadrant by "pushing" at the boundary in the fixed directions shown in Fig. 2. These directions of control are given by the columns of the matrix $R$, and for historical reasons stemming from the one-dimensional case, are called directions of "reflection", though one should not think of constructing the process by any kind of mirror reflection. These directions may be loosely interpreted as follows. Consider the case where $Z_1 = 0$. This corresponds to the first queue being empty. Imagine that when the queuing network is in this situation, the server at queue 1 continues working even though there are no customers to serve. The server thereby generates "potential" services. To prevent $Q_1$ from becoming negative (due to the completion of such a service), each potential service performed by server 1 needs to be corrected by a unit step in the positive $Q_1$ direction to keep $Q_1$ zero. This lost potential service also has an effect on the second queue in the sense that there is "lost potential flow" to the second queue. Accordingly, for each corrective unit step in the positive $Q_1$ direction, there is a corresponding downward unit step in the $Q_2$ direction. In the heavy traffic limit, this behavior translates to instantaneous pushing at the boundary $Z_2 = 0$ in the direction $(1,-1)$ indicated in Fig. 2. In an analogous manner, on the boundary $Z_2 = 0$ one has pushing in the direction $(-1,1)$. The term $-p$ comes from the lost potential flow from queue 2 back to queue 1.

The limit result cited above is justified by the heavy traffic theorem of Reiman [50]. Besides varying the arrival rate as $n \to \infty$, one can also allow suitable variations in the initial queue lengths, service rates, arrival and service time variances, and the routing probability $p$. The reader is referred to [50] for such refinements. Existence and uniqueness of the limiting diffusion process $Z$ follows from a path-by-path construction due to Harrison and Reiman [27].

This tandem queue is a single class network in the sense that at each station there is just one class of customer, i.e., the customers are indistinguishable from one another. The result of Reiman [50] was cited to justify the approximation in this two-station case, his result applies to a general $d$-station ($d \geq 1$) single class FIFO network. Similarly, the existence and uniqueness theorem of Harrison and Reiman [27] applies to the associated reflecting Brownian motion which lives in the positive $d$-dimensional orthant. This has a form that is the $d$-dimensional analogue of (6) and in particular the reflection matrix $R = I - P^t$, where $P$ is the transition matrix for a transient Markov chain on $d$ states (corresponding to the Markovian routing matrix for the queuing network).

A brief synopsis of the extent heavy traffic limit theory for single and some multiclasses networks is given in the next section. For more details the reader is referred to the paper [65] and references therein.

3. Heavy Traffic Limit Theorems: A Brief Synopsis

To facilitate comparison of different results, in the sequel, queuing networks will be assumed to have the following common features. There is a single server at each station, the arrival process for each customer class is a renewal process for which the interarrival times have finite means and variances, the service times for each class are given by a sequence of i.i.d. random variables having finite means and variances, the

Some recent developments for queuing networks

routing is Markovian, and the arrival processes, service times and customer routing are mutually independent. For single class networks there is just one customer class per station, whereas for multiclass networks there may be several different classes of customers served at a single station and the mapping from classes to stations is many-to-one.

Heavy traffic limit theory has been concerned largely with open networks in which customers arrive from outside the system, receive a finite number of services at various stations and then exit the system. Furthermore the theory is most well-developed for single class networks with FIFO (first-in-first-out) service discipline. Let us consider open networks first.

The heavy traffic limit theorem of Iglehart and Whitt [35, 36], for a single class FIFO station, is the prototype for conventional heavy traffic limit theorems. Following their work, Harrison [21] proved a heavy traffic limit theorem for two stations in tandem (without feedback) and first identified a sample path representation for the two-dimensional limit process, which is a reflecting Brownian motion living in the positive quadrant. His limit theorem was generalized by Reiman [50] to single class FIFO networks with Markovian routing which may have feedback. These networks are sometimes referred to as generalized Jackson networks.

In multiclass networks, different classes of customers, perhaps having different service distributions or different routing requirements, may be served at a station. For such networks, it is natural to consider other service disciplines besides FIFO. The extent heavy traffic limit theorems for such networks have mainly focused on those with (static) priority service across classes and FIFO service within a priority class. In multiclass networks, the dimension of the queue length process is equal to the number of classes served in the system (recall that the mapping from classes to stations is many-to-one). Another process of interest is the (immediate) workload process $W$ whose dimension is equal to the number of stations and is such that $W_i(t)$ represents the amount of work (measured in units of service time) embodied in the customers at station $i$ at time $t$.

Whitt [60] proved a heavy traffic limit theorem for a single multiclass station with two priority classes (high and low). A notable feature here is that the two-dimensional queue length process, normalized with a central limit theorem type of scaling, converges in distribution to a process in which the component corresponding to the high priority class is identically zero and that corresponding to the low priority class is a one-dimensional reflecting Brownian motion. Johnson [37] combined the features of the Reiman [50] and Whitt [60] results to prove a heavy traffic limit theorem for multiclass networks having two types of customers, those of high priority and those of low priority, where a customer retains the same priority designation during its entire sojourn through the network, i.e., once a high priority customer, always a high priority customer etc. A network of this kind is said to have separated priorities. Peterson [49] proved a heavy traffic limit theorem for multiclass networks with priority service (two levels, not separated), but he restricted to the case of feedforward routing. In this, the $d$-dimensional workload process $W^{(n)}$, when normalized with the a central
limit theorem type of scaling:

\[ W(n)(\cdot) = \frac{W(n)(n\cdot)}{\sqrt{n}}. \]

converges in distribution to a reflecting Brownian motion \( B \) that lives in the positive \( d \)-dimensional orthant and has a semimartingale decomposition of the form (6) where \( X, Y \) are now \( d \)-dimensional and \( R \) is a \( d \times d \) matrix of the form \((I + G)^{-1}\) where \( G \) is non-negative and upper triangular. The class level queue length processes, with the same normalization as for the workload processes, also converge in distribution. As in the single station case [60], for each station, the limit for the high priority queue length processes is identically zero and for the low priority queue length processes is proportional to the limit of the normalized workload processes for the station. Although heavy traffic limit theorems have been proved for a single multiclass station with feedback and certain service disciplines such as round-robin [31] and FIFO [13], there is currently no general heavy traffic limit theorem for open multiclass networks with feedback.

Closed queueing networks, in which a fixed number of customers or jobs circulate perpetually in the system, are natural models for some manufacturing systems. Analogues of the open network heavy traffic limit theorems of Reiman [50] and Johnson [37] have been proved by Chen and Mandelbaum [8] for closed networks. In particular, they considered closed networks that are single class or have separated priorities. For a closed network, the heavy traffic parameter \( n \) has a natural interpretation as the fixed number of customers in the system. Then a natural central limit theorem type of scaling for the queue length process \( Q(n) \) is given by

\[ Q(n)(\cdot) = \sqrt{n} \mathcal{N}(0, n^2) \]

Under suitable conditions, the (low priority) \( d \)-dimensional queue length process normalized as in (10) converges in distribution as \( n \to \infty \) to a reflecting Brownian motion that lives in the \( d \)-dimensional simplex (see Harrison, Williams and Chen [33] for some analysis of this limit process), and the normalized high priority queue length process vanishes in the heavy traffic limit. Despite these positive results, as in the open network case, there is currently no general heavy traffic limit theorem for closed multiclass queueing networks.

In contrast to the lack of a general heavy traffic limit theorem for multiclass networks with feedback, there is a rigorous existence and uniqueness theory for semimartingale reflecting Brownian motions (SRBMs). These diffusion processes have a semimartingale form as in (6) where the reflection matrix \( R \) need only satisfy a natural feasibility condition that it be completely-S, i.e., for each principal submatrix \( R \) of \( R \) there is a positive vector \( \hat{y} \) such that \( \hat{R} \hat{y} > 0 \). Reiman and Williams [52] established the necessity of this condition and Taylor and Williams [57] established its sufficiency for existence and uniqueness in law of a SRBM, provided one adds to the definition the mild condition that \( X \) minus its drift process is a martingale relative to the filtration generated by \( X, Y, Z \). For an extension of these results to convex polyhedral state spaces, which is relevant to closed network approximations, see Dai and Williams [18].

Some recent developments for queueing networks

![Figure 3. A Multiclass Open Network with FIFO Service](image)

Extrapolating from the extant limit theorems, Harrison [22] and Harrison and Nguyen [24, 25] conjectured a reflecting Brownian motion approximation (or an approximate "Brownian model") for open multiclass queueing networks with FIFO service in heavy traffic. An analogous approximation was proposed in the Appendix to Harrison and Williams [31], where static priority and processor sharing service disciplines are included in addition to FIFO. Despite the appeal of these approximations, apart from the situations covered by [35, 36, 21, 50, 60, 37, 49, 51, 13], no heavy traffic limit theorem has been proved to justify them. However, it was still a surprise when Dai and Wang [16] produced an example which showed that not all open multiclass queueing networks can have such an approximation. A variant of this example and possible explanations for it are described in the next section.

4. Dai-Wang-Type Example and Stability of Open Networks

The following variant of the Dai-Wang example appears in the paper of Dai and Nguyen [14].

Consider the two-station network pictured in Fig. 3. Arrivals to this network are assumed to be given by a Poisson process with arrival rate \( \alpha \in (0, 1) \). There is a single server at each of the two stations. Customers are routed through the network in a deterministic manner making visits to the two stations in the following order \( 1, 2, 2, 1 \). A customer awaiting or undergoing its fourth service is called a class \( k \) customer, \( k = 1, 2, 3, 4 \). Thus, classes 1, 2, 3 are served at station 1 and classes 3, 4 are served at station 2. Customers at a station, regardless of class, are served on a first-in-first-out basis. Thus, for example, after receiving service as a class 3 customer at station 2, a customer changes to class 4 and goes to the end of the line at station 2. The service times are assumed to be independent and exponentially distributed with mean \( m_k \) for class \( k = 1, 2, 3, 4, 5 \), where

\[ m = \begin{pmatrix} 1 & 1 & 23 & 4 & 4 \\ 10 & 10 & 27 & 27 & 5 \end{pmatrix}. \]

Interarrival and service times are mutually independent.
The traffic intensity parameters for the two stations are given by

\[ \rho_1 = \alpha(m_1 + m_2 + m_5) \quad \text{and} \quad \rho_2 = \alpha(n_3 + m_4). \]

Traditionally, \( \rho_i \) has been interpreted as the long run average rate at which work arrives at station \( i \) (implicit here is an assumption that such long run average behavior exists). Extrapolating from existing theory, it natural to define heavy traffic for this example as occurring when these traffic intensity parameters are close to one. Indeed, the approximation proposed in [22, 24, 25, 31] would say that with \( \alpha = 1 - \frac{1}{n^2} \) (and so \( \rho_i = 1 - \frac{1}{n^2}, i = 1, 2 \)), for \( n \) sufficiently large one can approximate the two-dimensional workload process (normalized with the central limit theorem scaling as in (9)), by a reflecting Brownian motion that lives in the positive quadrant and has the form (6), where the reflection matrix is given by

\[ R = \begin{bmatrix} -20 & 16 \\ 20 & -27 \end{bmatrix}. \]

However, this reflection matrix corresponds to directions of reflection that point out of the quadrant. There is no semimartingale reflecting Brownian motion living in the positive quadrant with such directions of reflection. Consequently, the conjectured limit theorem cannot hold in this case. Indeed, investigation by Dai and Nguyen [14] of this example showed that the normalized workload processes cannot converge in distribution to a continuous limit.

Around the time that the Dai-Wang counterexample was produced, there was a separate growing interest in the stability of open multiclass networks. For a single class, \( d \)-station, open queueing network satisfying the conditions described at the beginning of section 3, the condition for stability is that

\[ \rho_i < 1 \quad \text{for} \quad i = 1, \ldots, d, \]

(see e.g., Meyn and Down [47]). Here the traffic intensity parameters \( \rho_i \) are defined by

\[ \rho_i = \lambda_i m_i, \]

where \( \lambda = (\lambda_1, \ldots, \lambda_d)' \) is the solution of the vector traffic equation

\[ \lambda = \alpha + P \lambda, \]

\( \alpha = (\alpha_1, \ldots, \alpha_d)' \) is the vector of exogenous arrival rates (one component for each station), \( P \) is the \( d \times d \) matrix of Markovian routing probabilities (so that \( P_{ij} \) denotes the probability that a customer completing service at station \( i \) goes next to station \( j \)), and \( m_i \) is the mean service time per customer at station \( i \). It is natural to conjecture that (14) is also the condition for stability of multiclass networks, provided \( \rho_i \) is now defined by

\[ \rho_i = \sum_{i \in C_i} \lambda_i m_i. \]

Some recent developments for queueing networks

Figure 4. A Multiclass Open Network with Priority Service

where \( C_i \) denotes the constituency of station \( i \), i.e., it consists of those customer classes \( k \) that are served at station \( i \), \( m_k \) denotes the mean service time for class \( k \) customers \( k = 1, \ldots, K \), and \( \lambda = (\lambda_1, \ldots, \lambda_K)' \) is the solution of the traffic equation (16) now considered to be written at the class level, so that \( \alpha \) is the vector of class level exogenous arrival rates and \( P \) is the class level matrix of Markovian routing probabilities. Despite some cases [37, 49] where this conjecture is true, Kumar and coworkers [42, 44] have given simple two-station deterministic examples with priority service which show that it is false in general for multiclass networks with feedback, i.e., they have given examples of networks in which \( \rho_i < 1 \) for each \( i \), but the networks are unstable. Rybko and Stolyar [53] gave the first stochastic counterexample. This is a two-station network with priority service which has similar structure to the Lu-Kumar [44] example.

Another counterexample can be given by considering the network pictured in Fig. 4. This is a hybrid of the Lu-Kumar and Rybko-Stolyar examples. Here the arrivals are given by a Poisson process of rate \( \alpha \). Customers are routed through the network in a deterministic manner, making visits to the two stations in the order 1, 2, 2, 1. Customers that are awaiting or undergoing their \( k \)-th service will be called class \( k \) customers, \( k = 1, 2, 3, 4 \). All services are independent and class \( k \) service times are exponentially distributed with mean \( m_k > 0 \), \( k = 1, 2, 3, 4 \). Each server follows a preemptive resume priority discipline where classes 2 and 4 have priority over classes 3 and 1, respectively. The traffic intensities for the two stations are given by

\[ \rho_1 = m_1 + m_4 \quad \text{and} \quad \rho_2 = m_3 + m_4. \]

Dai and Weiss [17] have shown that this network is stable if

\[ \rho_i < 1 \quad \text{for} \quad i = 1, 2, \text{and} \quad m_2 + m_4 < 1. \]

Furthermore, Dai and Van de Vate [15] have recently shown that if any of the inequalities in (19) is violated with strict inequality, then the network is unstable. Thus, for example, if \( m_1 = m_3 = 1/4, m_2 = m_4 = 2/3 \), then the network is unstable, but \( \rho_i < 1 \) for \( i = 1, 2 \).

Whilst it might be argued that one can see (with hindsight) that the priorities in the Lu-Kumar/Rybko-Stolyar type examples are "bad" for stability, it was a
Some recent developments for queuing networks

from (or at least a refinement of) that used by Chen and Mandelbaum for closed networks; indeed, the network parameters are chosen so that (19) holds with equality in place of inequality (the motivation here is that since the closed network with a fixed number of customers is automatically stable, it is natural to consider parameters that correspond to the boundary between stability and instability in the open network), (b) the normalization of the queue length processes is different from the usual central limit theorem type of scaling, and (c) the limit of the normalized queue length processes is not obtained from a reflecting Brownian motion, although it is related to Brownian motion. Moreover, to fully describe the limit process, non-trivial limits of all of the normalized queue length processes are needed and for the convergence in distribution, the topology on path space is weaker for some components than the usual Skorokhod $J_1$-topology. For further details of this example, the reader is referred to the survey paper [65] or the full paper [32].

5. Open Problems

Some of the open problem areas for heavy traffic analysis of queuing networks are described below. For details the reader is referred to the papers cited.

(i) Stability. As mentioned in section 4, one of the very active areas of current research on queuing networks is concerned with determining sufficient conditions for the stability of open multiclass networks. Whilst some progress has been made (e.g., [5, 6, 15, 17, 41]) on identifying service disciplines and network structures for which the conventional condition “$\rho_i < 1$ for all $i$” is sufficient for stability, we are still a considerable distance from a general classification.

(ii) Heavy Traffic Approximation of Queuing Networks. As described in section 3, there is no general heavy traffic limit theorem for multiclass queuing networks, the extant limit theorems being restricted to single class networks (or ones with separated priorities) or to multiclass priority networks with a feedforward structure. It is a compelling problem to identify a “good” class of multiclass networks for which conventional heavy traffic limit theorems can be proved. Conversely, it would be helpful to know the size of the collection of networks that do not have conventional heavy traffic behavior, for example, is it a small set in some measure theoretic sense? In the same vein, it would be interesting to identify the spectrum of possible unconventional heavy traffic behavior. For instance, the example of Harrison and Williams [32] illustrates several ways in which a heavy traffic limit theorem can be unconventional, but are there others?

There are many possible variations on the queuing network models described here. For instance, heavy traffic approximations for state- and/or time-dependent networks with Markovian assumptions have been established by some authors (see e.g., [40, 45, 46, 48]). Also, heavy traffic approximations have been proposed (but no limit theorem has been proved) for queuing networks that incorporate server breakdown and repair [26]. Such models are especially relevant for manufacturing applications.

(iii) Semimartingale Reflecting Brownian Motions. Although there is a solid existence and uniqueness theory for semimartingale reflecting Brownian motions (SRBMs), several problems associated with the analysis of these processes remain. A necessary and sufficient condition for positive recurrence is known in the two-dimensional case

R. J. WILLIAMS

further surprise when Bramson [3, 4] gave a two-station stochastic counterexample having FIFO service discipline. (Also, Seidman [55] gave a deterministic FIFO counterexample.)

The appearance of these counterexamples was followed by a burst of activity concerned with the stability of open multiclass networks. Frequently Lyapunov functions have been used as a tool for establishing sufficient conditions for stability. Early work used mathematical programming to determine such Lyapunov functions for the queuing networks (see for example, Kumar and Meyn [41] and Bertsimas et al. [2]). A significant advance was made when Dai [10] showed that the stability of associated fluid limits (obtained as limits of a Markovian state descriptor for the queuing network under a law of large numbers type of scaling), was sufficient for stability of the original queuing network. (An analogue of this result for semimartingale reflecting Brownian motions was proved a little earlier by Dupuis and Williams [19].) Thus to determine sufficient conditions for stability, one can seek Lyapunov functions for the often simpler fluid limits rather than for the original queuing networks. This idea has been exploited by a number of authors, especially in combination with piecewise linear Lyapunov functions, to prove sufficient conditions for the stability of open multiclass networks (see Dai [11] for a survey). Networks with FIFO service disciplines can be especially difficult to analyze because of the need to keep track of the order in which customers arrive at each queue. Since the writing of [11], by establishing the stability of associated fluid limits, Bramson [5] has proved the stability of open Kelly-type networks with FIFO service discipline, provided $\rho_i < 1$ for each $i$. Here Kelly means that the service rate is the same for all customers at a given station, i.e., $m_k$ is the same for all $k \in C_i$. Bramson [6] has also shown stability for networks with a processor sharing service discipline and $\rho_i < 1$ for all $i$. Here processor sharing means that service at each station is equally divided amongst all customers present at the station. In a very recent paper, Chen and Zhang [9] claim to prove stability of open multiclass networks with FIFO service when $\rho_i < 1$ for each $i$, assuming a spectral radius condition on one of the data matrices. At this time, it is not known if there is a large natural class of multiclass networks satisfying this condition.

In light of this work on stability, one might be tempted to conjecture that the Dai-Wang type examples fail to have Brownian approximations because they are not stable, i.e., one has the wrong notion of heavy traffic. However, a deterministic network example of Whitt [61], which exhibits large oscillations of the queue length processes, suggests that one might also entertain other possible explanations, such as the wrong scaling, the wrong topology on path space, or the wrong limit process. In general, one might consider one or more of the above as explanations for why a multiclass network does not follow a “conventional” heavy traffic limit theorem of the form proposed in [22, 24, 25, 31].

Indeed, in a recent work, Harrison and Williams [32] proved a heavy traffic limit theorem for a queuing network which is a closed network analogue of the Lu-Kumar/Ryboko-Stolyar hybrid shown in Fig. 4 and which incorporates all of the possible exceptional features mentioned above. The network is obtained simply by turning off the exogenous arrival process and closing the loop on the left of Fig. 4 so that customers return to class 1 after completing service in class 4. Unconditional features of this limit theorem are as follows: (a) the notion of heavy traffic is different
[34, 62], but there is no general recurrence classification in dimensions three or more, although some sufficient conditions for positive recurrence are known [29, 30, 63]. The result of Dupuis and Williams [19] implies that one can obtain sufficient conditions for positive recurrence by study of a simpler deterministic dynamical system (see a recent paper of Chen [7] for an illustration of how this can be applied to simplify the proofs of previously known sufficient conditions for positive recurrence). When an SRBM is positive recurrent, its stationary distribution is characterized as the solution of a certain integral relation (see e.g., [64]). Whilst there is a numerical method [12] for analyzing this relation, improvements of this method would facilitate analysis of larger networks. Furthermore, the numerical method would be enhanced by knowledge of the tail behavior of the stationary distribution. A final problem is that of justifying the interchange of limits inherent in using the stationary distribution of an SRBM as an approximation to the equilibrium distribution of the original queueing network (see [38] for a discussion of this in the case of closed networks).

(iv) Optimization. The discussion in this paper has been directed to the problem of performance analysis for heavily loaded networks with a fixed structure. However, in some network applications one may have control over some of the system parameters, e.g., service discipline or routing, and want to choose them so as to optimize a measure of performance. Such control problems frequently cannot be analyzed exactly. One possible solution is to optimize an approximate model. This kind of approach, using approximate Brownian models, has been pursued by a few authors (see e.g., [28, 39, 43, 58, 59]), but this is an area with potential for much further development.

References


Some recent developments for queueing networks


Lecture Notes in Statistics
For information about Volumes 1 to 53 please contact Springer-Verlag