Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling

F.P. Kelly and C.N. Laws*

Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England

Received 10 October 1991; revised 13 April 1992

We present an introductory review of recent work on the control of open queueing networks. We assume that customers of different types arrive at a network and pass through the system via one of several possible routes; the set of routes available to a customer depends on its type. A route through the network is an ordered set of service stations: a customer queues for service at each station on its route and then leaves the system. The two methods of control we consider are the routing of customers through the network, and the sequencing of service at the stations, and our aim is to minimize the number of customers in the system. We concentrate especially on the insights which can be obtained from heavy traffic analysis, and in particular from Harrison's Brownian network models. Our main conclusion is that in many respects dynamic routing *simplifies* the behaviour of networks, and that under good control policies it may well be possible to model the aggregate behaviour of a network quite straightforwardly.

Keywords: Brownian network models, resource pooling, threshold strategies, generalized cut constraints, heavy traffic analysis, pathwise solution, dynamic sequencing, shortest delay routing.

1. Introduction

The success of product-form queueing networks in modelling complex systems probably owes less to their ability to represent accurately the various detailed features of a system than to the simple framework they provide for such fundamental features as mean arrival rates and traffic intensities. Even when the dynamics of individual queues are far from those necessary for a product-form solution to hold precisely, the framework of such solutions can be used to suggest approximations and asymptotics (see, for example, Gelenbe and Pujolle [11], Kelly [26], Whitt [48]), and simple routing and capacity allocation algorithms (see, for example, Gallager [10], Kelly [25], Kleinrock [28]).

Dynamic routing is, however, an important feature of many systems that is *not*

* Supported by SERC grant GR/F 94194. Current address: Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, England.

well represented in product-form queueing networks. If customers can choose their route through the network on the basis of information about the current states of various queues, then exercise of this choice may produce significant dependencies between different parts of the network. Further, such dependencies may be a major determinant of system performance. Our aim in this paper is to illustrate these aspects of dynamic routing, through a systematic discussion of a sequence of examples. We concentrate especially on the study of networks in heavy traffic, where the important features of good control policies are displayed in sharpest relief. In particular, we make extensive use of the Brownian network models developed in a major and sustained campaign by Harrison, and his coauthors Reiman, Williams and Wein [15,16,19,20,22,23]. Our main conclusion is that in many respects dynamic routing *simplifies* the behaviour of networks, and that under good control policies it may well be possible to model the aggregate behaviour of a network quite straightforwardly.

In section 2 we consider our first example, that of a collection of parallel queues, where an arriving customer can be routed to any one of the queues: see fig. 1 for the case of two queues. This problem has received much attention, and we begin section 2 with a brief review of the literature. We then describe the heavy traffic analysis of Foschini, Salz and Reiman [7,9,38]. They show that if customers are routed to the shortest queue, then in heavy traffic there is *state space collapse*: the limiting diffusion process representing queue lengths collapses to one dimension. One consequence of this collapse is a *resource pooling* effect: customer delay is distributed as in a system where there is a single queue with multiple servers. To illustrate the magnitude of this effect consider the special case where the arrival process is Poisson, service times are independent and exponential, and there are K identical servers. Then the resource pooling effect reduces mean delay by a factor K over a system where customers are allocated randomly on arrival to servers. Resource pooling can also be achieved by many strategies which do not cause state space collapse, and as an example we describe in section 2 a simple threshold strategy: the key feature is that all servers be kept busy if there is substantial work in the system for any of them.

In general the control of open queueing networks involves not just *routing*, but also *sequencing*. A routing policy determines which route a customer should take through the network; a sequencing policy specifies which type of customer to serve at each station, at each point in time. In section 3 we consider perhaps the simplest

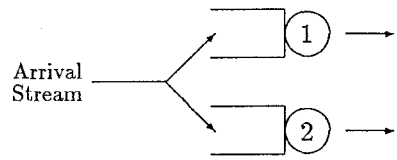

Fig. 1. Two parallel queues.

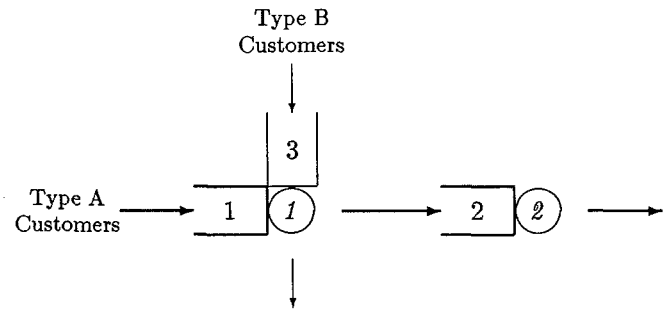

Fig. 2. A network with sequencing.

example of dynamic sequencing, that illustrated in fig. 2 and studied by Harrison and Wein [21]. In this network there are no routing decisions. Rather, the problem is to decide, at each point in time, which type of customer to serve at station 1 in order to minimize the mean delay of customers. In heavy traffic an optimal policy is to give priority to type B customers at station 1 unless the number of customers at station 2 is less than a threshold, in which case priority is given to type A customers. This dynamic sequencing policy is able, in heavy traffic, to achieve two objectives: it exits customers from the system as quickly as possible, and also manages to keep server 2 busy whenever there is substantial work in the system for that server.

In section 4 we consider our first example that involves *both* routing and sequencing, taken from Wein [46]. In fig. 3 type B customers require service at either station 1 or station 2: at the time of arrival a type B customer is routed to one or other of queues 3 and 4. In addition to these routing decisions, a control policy specifies which type of customer to serve at stations 1 and 2. The optimal control in this network causes the two servers to act as a pooled resource, with substantial work building up only in queue 1. An explicit formula, expression (4.8), describes the mean system population in heavy traffic, and indicates the considerable benefits available from the optimal control. As illustration, we describe a simple example where in heavy traffic the optimal routing and sequencing policy reduces the

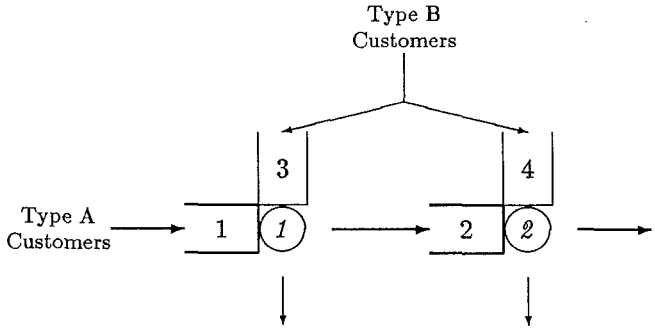

Fig. 3. Routing and sequencing.

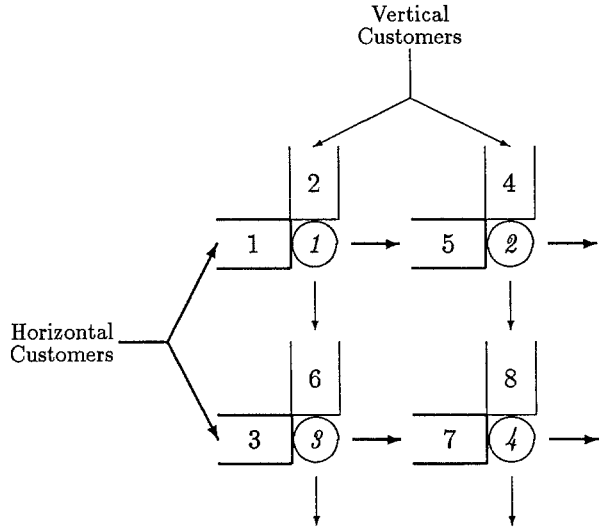

Fig. 4. The four station network.

mean number of customers in the system, or equivalently the mean delay, by a factor of 3.

Section 5 is devoted to the four station network illustrated in fig. 4 and studied by Laws and Louth [34]. The system is used by two different types of customer: horizontal customers pass through the system by using the top row, queueing at station 1 and then at station 2 before leaving, or the bottom row, queueing at station 3 then station 4. Similarly vertical customers use either the left or right hand column. The heavy traffic behaviour of this network under optimal control can be interpreted in terms of the reduced fork–join queueing system illustrated in fig. 5. Consider two queues, each with two servers. Let the first queue have the servers from stations 1 and 4 of the original network, and let the second queue have the servers from stations 2 and 3 of the original network. Customers arrive as for the original network. Each arriving customer sends a token to both queues simultaneously, and these progress independently. When both tokens have been served the customer leaves the system. In heavy traffic and under the optimal routing and se-

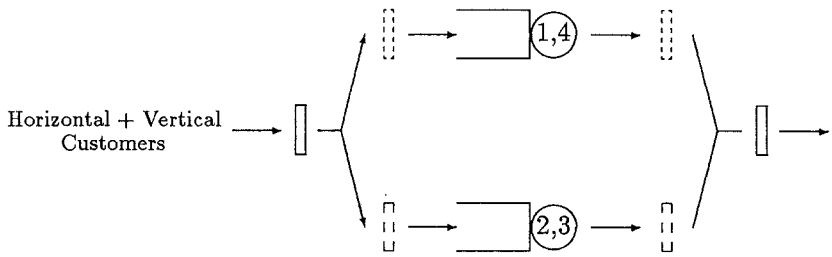

Fig. 5. The reduced fork–join system.

quencing policies it is *as if* servers 1 and 4, and servers 2 and 3, are combined to form *two* pooled resources. Also the network behaves *as if* customers queue for their service from these resources in parallel rather than in series. The reader should note that servers 1 and 4 form a *cut* for the original network: every customer entering the network, whether horizontal or vertical, must be served by one of these servers. Similarly servers 2 and 3 form a second cut.

The analysis of the four station network has been generalized in various directions by Laws [32]. In section 6 we briefly describe a network with two pooled resources each comprising four or more servers, and a network with three pooled resources, each comprising two servers. In each case the pooled resources correspond to disjoint cuts of the network, and the heavy traffic behaviour of the network under optimal control can again be interpreted in terms of a reduced fork–join queueing system.

Must pooled resources correspond to disjoint cuts of the network? The answer to this question is *no*: a pooled resource can be more general than just a cut, and pooled resources may overlap. We demonstrate this in sections 6 and 7 with a discussion of the 2×3 network illustrated in fig. 6. Let λ_1, λ_2 label the arrival rates of type 1 and type 2 customers respectively, and let $\mu_1, \mu_2, \ldots, \mu_6$ label the service rates of servers $1, 2, \ldots, 6$ respectively. Then for stability we must have

$$3\lambda_1 + 2\lambda_2 \leqslant 2\mu_1 + 2\mu_5 + \mu_3 + \mu_6 , \tag{1.1}$$

as the following argument makes clear. Suppose that customers are charged 2, 2, 1 and 1 units at stations 1, 5, 3 and 6 respectively. To pay to get through the network, type 1 and type 2 customers enter the system with 3 and 2 units respectively. If the network is stable, the total arrival rate of revenue, $3\lambda_1 + 2\lambda_2$, cannot exceed

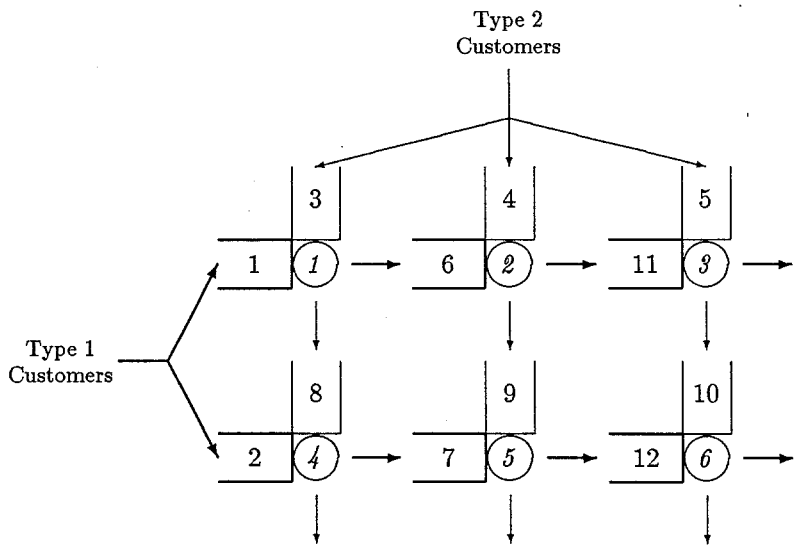

Fig. 6. The 2×3 network.

the maximal charging rate, $2\mu_1 + 2\mu_5 + \mu_3 + \mu_6$: but this is precisely condition (1.1). Observe that servers 1, 5, 3 and 6 form a pooled resource more general than a cut; we call condition (1.1) a *generalized cut constraint*. By symmetry there are another five generalized cut constraints of the form (1.1), and no two involve disjoint sets of servers.

In many of the examples we discuss, the optimal control in heavy traffic produces a *pathwise solution* (Harrison [16]): the total number of customers in the system is simultaneously minimized for all times $t \geqslant 0$. A pathwise solution is not always possible, and in section 6 we describe two examples. In the first example there is a conflict between the objective of keeping a pooled resource busy whenever there is work in the system for it, and the objective of exiting customers from the system as fast as possible; in the second example, the symmetric version of the 2×3 network illustrated in fig. 6, there is a conflict between two pooled resources over which should be kept busy the longer. In examples such as these the problem of minimizing long-run expected averages is, even in heavy traffic, quite delicate. For instance the optimal control may depend upon second moments of interarrival and service times distributions. Wein [44,45] has considered in detail an example where the problem can be reduced to a singular control problem for a one-dimensional Brownian motion, and hence an explicit solution obtained. In Wein [47] a more general example is tackled using the method of finite difference approximations developed by Kushner [29,30]. The associated analytical problems involve partial differential equations with oblique derivative boundary conditions: for a discussion of the issues involved and a survey of recent progress the interested reader is referred to Dai, Harrison and Nguyen [3,17]. In this paper we concentrate on cruder insights: while the performance of the policies we discuss may depend on second moments, the policies themselves do not.

How reasonable is it to assume that multiple pooled resources will approach heavy traffic together? This clearly depends on the context. If arrival and service rates are carefully balanced, as may be the case if the network represents a manufacturing system, then it is possible that several pooled resources may be simultaneously heavily loaded. If there are more customer types than stations and if arrival rates themselves vary over time, as may be the case if the network represents a communications network, then we might expect to find at most a single pooled resource approaching overload. In section 7 we concentrate on this case, following the work of Laws [33]. We consider a quite general network structure, and describe the form of the optimal pathwise solution. Under this control the system population is a reflected Brownian motion, and a simple explicit formula, expression (7.18), describes its mean. We also consider performance under shortest delay routing (SDR), a policy which sends customers via the route with the shortest expected delay. (Routing algorithms based on delay estimates are of considerable practical importance in communication networks: see Bertsekas and Gallager [2], Schwartz [41].) In general SDR is not optimal, but it does have desirable properties. In particular, we describe how it achieves the important resource pooling property of the pathwise solution.

We make no attempt to justify the heavy traffic control policies we discuss via weak convergence results. The construction and analysis by Harrison and Reiman [19,20] of multidimensional reflected Brownian motion, and the weak convergence results of Reiman [38,39] for open queueing networks, parallel servers and priority queues, provide substantial motivation for the Brownian network model, but stop short of providing a complete justification (see Dai and Wang [4], and Harrison and Nguyen [18] in this issue). An approach to weak convergence in the control context is provided by Martins, Kushner and Ramachandran [31,35]. We simply note our view that the eventual rigorous treatment of pathwise solutions may well fall within a simpler framework, not involving the distributional properties of Brownian motion.

Although a general weak convergence underpinning is currently lacking, there is by now a reassuring body of numerical evidence for the asymptotic optimality of the routing and scheduling schemes we describe in this paper. Often pathwise lower bounds on the total system population are available, and the scaled performance of the schemes is observed to approach the bounds as the load on the network increases (Harrison and Wein [21], Laws [33], Laws and Louth [34]). Indeed it is possible to devise simple schemes that perform well even in light or moderate traffic, although, as we would expect, the percentage improvement over other schemes (involving, for example, random or alternate routing, with fixed or no priorities) generally improves as the network load increases. These simple schemes often involve choices not completely specified by the heavy traffic analysis, for example the precise levels of threshold, or routing decisions when no part of a pooled resource is threatened with idleness; but the heavy traffic analysis does give clear insight into the important features to which routing and sequencing controls should be directed.

2. Parallel queues

Consider the two station system of fig. 1. There is a single stream of arriving customers and on arrival each customer is routed to one of the two stations where it queues for service. Such routing decisions are irrevocable: customers may not move between queues at a later stage. Each server operates a first come first served discipline and after its service a customer leaves the system. This simple system has received much attention in the literature; for early work see Haight [13] and Kingman [27]. Optimal control of the system has been analyzed as has the performance of the system under given routing policies.

Suppose that there is a single exponential server of rate μ at each station, that the arrival process is Poisson with rate λ, where the traffic intensity $\rho = \lambda/2\mu < 1$, and that arriving customers join the shorter queue. Winston [52] has shown that this dynamic policy minimizes the mean delay of customers in the system, and indeed that it is optimal in the stronger sense of stochastic order. Weber [43] extended

this result to arbitrary arrival streams and service times with nondecreasing failure rate. The case of arbitrary arrivals and exponential servers was also considered by Ephremides et al. [5]. However, Whitt [49] has shown via light traffic analysis that the shorter queue policy is *not* optimal for general service times, even when the arrival process is Poisson: there exists a service time distribution such that when the difference between the numbers in the two queues is small, the longer queue is likely to have a sudden series of departures and hence be a better choice. Kingman [27] found that for ρ near 1 the distribution of a customer's waiting time is approximately the same as for a single server with traffic intensity ρ^2. Flatto and McKean [6] also compared the performance of the system to that of a single server, with a service rate of 2μ, and showed that the mean population of the two server system is only slightly greater than that of the single-server system when ρ is near 1. Halfin [14] obtained bounds for the probability distribution of the number of customers in the system and its expected value in equilibrium; these bounds are tight as $\rho \uparrow 1$ and again show that, in the limit, the two-server system behaves like a system with a single-server of rate 2μ. Recently, Adan et al. [1] have shown that the stationary queue length distribution of the system can be represented by an infinite sum of product-form solutions and that this allows efficient numerical calculation of the distribution.

When there are multiple servers (but only one queue) at each station the natural extension of the above policy is to join the queue with the shorter expected delay. Whitt [49] has shown that this policy is *not* optimal in general, even for Poisson arrivals and exponential service times. However, Houck [24] has shown via simulation that it is close to being optimal in many cases.

Returning to the case of identical single-server stations, the behaviour of the shorter queue policy has been analyzed in the heavy traffic limit ($\rho \uparrow 1$) by Foschini and Salz [9] for Poisson arrivals and exponential service times. The two separate servers act as a single pooled resource in the limit: that is, when ρ is near 1 the two-server system behaves as if it were an $M/M/2$ queue with arrival rate λ and servers of rate μ. The limiting mean delay is half of that found when routing arrivals to each queue with probability 1/2 (the policy minimizing mean delay amongst the class of quasi-static policies described by Gallager [10], Kelly [25]). Reiman [38] has generalized the heavy traffic analysis to the case of renewal arrivals and general service times.

We now review the result of Reiman [38], established for the slightly more general system illustrated in fig. 7. Arrival stream A_j exclusively feeds server j, for

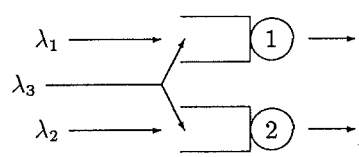

Fig. 7. Parallel queues.

$j = 1, 2$, while customers in arrival stream A_3 join the shorter of the two queues at the moment of arrival. Assume these three arrival streams form independent renewal processes, with interarrival times in stream j having mean λ_j^{-1} and variance a_j. Service times at server $k, k = 1, 2$, are independent random variables with mean μ^{-1} and variance s_k.

Consider now a sequence of such systems, indexed by n, such that $\lambda_j(n) \to \lambda_j$, $\mu(n) \to \mu$, and

$$n^{1/2}(\lambda_1(n) + \lambda_2(n) + \lambda_3(n) - 2\mu(n)) \to \theta \quad \text{as } n \to \infty, \tag{2.1}$$

where $\theta < \infty$. Assume also $a_j(n) \to a_j, s_k(n) \to s_k$ and that all interarrival and service times have a uniformly bounded moment of order $2 + \epsilon$ for some $\epsilon > 0$. Condition (2.1) defines heavy traffic; we also require $|\lambda_1 - \lambda_2| < \lambda_3$, so that there are enough customers with a choice to balance arrival rates at the queues. Let $Q_k^{(n)}(t)$ be the number of customers in the queue for server k at time t, including the customer in service, if any. Let $\mathbf{Q}^{(n)}(t) = (Q_1^{(n)}(t), Q_2^{(n)}(t))$, and define the scaled queue length process

$$\mathbf{Z}^{(n)}(t) = \frac{\mathbf{Q}^{(n)}(nt)}{n^{1/2}}.$$

Reiman [38] shows that

$$\sup_{0 \leqslant t \leqslant 1} |Z_1^{(n)}(t) - Z_2^{(n)}(t)| \xrightarrow{P} 0 \quad \text{as } n \to \infty \tag{2.2}$$

and that

$$Z_1^{(n)}(t) + Z_2^{(n)}(t) \Rightarrow Z(t), \tag{2.3}$$

where $Z(t)$ is a certain reflected Brownian motion. The result (2.2) shows that the stream A_3 is able to keep the two queue lengths $Q_1(t), Q_2(t)$ approximately equal. The reflected Brownian motion $Z(t)$ is that which would obtain if all arrivals joined a single queue served by the two servers: namely a Brownian motion with drift θ and variance $(\sum_{j=1}^{3} \lambda_j^3 a_j + \mu^3 \sum_{k=1}^{2} s_k)$, reflected at the origin.

Assuming $\theta < 0$, the mean of the stationary distribution of the reflected Brownian motion $Z(t)$ is

$$m_S = \tfrac{1}{2}\left(\sum_{j=1}^{3} \lambda_j^3 a_j + \mu^3 \sum_{k=1}^{2} s_k\right) |\theta|^{-1},$$

and Reiman [38] has shown how this expression can be used to compare the performance of routing strategies. Suppose $\lambda_1 = \lambda_2, a_1 = a_2, s_1 = s_2$ and consider two other routing strategies, the coin-toss strategy and the alternating strategy. Under the coin-toss strategy customers of type 3 flip a fair coin to determine which queue to join, while under the alternating strategy they alternate deterministically between the two queues. Under these strategies the means of the normalized number

in the system, $Z_1(t) + Z_2(t)$, are, in the heavy traffic limit,

$$m_C = \tfrac{1}{2}(4\lambda_1^3 a_1 + \lambda_3 + \lambda_3^3 a_3 + 4\mu^3 s_1)|\theta|^{-1},$$
$$m_A = \tfrac{1}{2}(4\lambda_1^3 a_1 + \lambda_3^3 a_3 + 4\mu^3 s_1)|\theta|^{-1},$$

respectively, while that for the shorter queue policy is

$$m_S = \tfrac{1}{2}(2\lambda_1^3 a_1 + \lambda_3^3 a_3 + 2\mu^3 s_1)|\theta|^{-1}.$$

Clearly $m_S \leqslant m_A \leqslant m_C$. For the special case where all distributions are exponential, $a_3 = \lambda_3^{-2}$, and so $m_S = \frac{1}{2}m_C$. Thus queue lengths are *halved* by using the shorter queue policy (Foschini and Salz [9]).

Let us now consider further the special case where all distributions are exponential. The key feature of the shorter queue policy, that allows it to halve mean queue lengths, is that is keeps both servers busy when there is substantial work to be done in the system. To show that this is the key feature, rather than the stronger property (2.2) that both queue lengths are held equal, and to quantify the term substantial, we consider a threshold strategy. Suppose that each queue has a *threshold parameter r* and that customers from arrival stream A_3 are routed randomly, in accordance with the tosses of a fair coin, *except* when one queue is below threshold and the other queue is above threshold, in which case the customers from stream A_3 are routed to the former queue. The aim of this threshold strategy is to prevent a server becoming idle when there is substantial work in the other queue.

Figure 8 describes some of the transition rates of the Markov chain $(Q_1(t), Q_2(t))$ under the threshold strategy. Observe that while a sample path remains off the lines $\{Q_1 = 0\}$ and $\{Q_2 = 0\}$, the number in the system, $Q_1(t) + Q_2(t)$, behaves as the number in a coupled $M/M/2$ queue. However, if a sample path hits the line $\{Q_k = 0\}$ while $Q_{3-k} > 1$ then there is the possibility that a customer might be served by server k in the coupled $M/M/2$ queue but not in the system under considera-

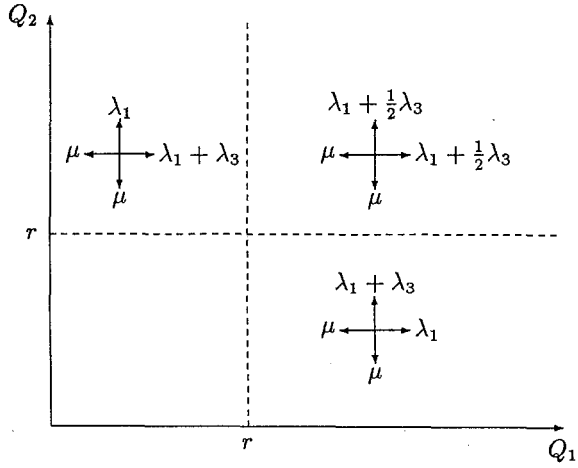

Fig. 8. Transition rates for threshold routing strategy.

tion. This would leave the number in the system, $Q_1(t) + Q_2(t)$, above that in the coupled $M/M/2$ queue. We shall see that we can bound the probability P of such an event occurring, for a sample path starting at $(Q_1(0), Q_2(0)) = (r, r)$ in the period before the sample path next returns to the set $\{Q_1 \leqslant r, Q_2 \leqslant r\}$.

Note first that while the sample path remains in the region $\{Q_1 \geqslant r, Q_2 \geqslant r\}$, the pair $(Q_1(t) - r, Q_2(t) - r)$ behave as two independent $M/M/1$ queues each with traffic intensity $\rho = (\lambda_1 + \frac{1}{2}\lambda_3)/\mu$. The expected number of times that a pair of such independent $M/M/1$ queues hits one or other axis between visits to (0, 0) is $(2\rho + 1)/(1 - \rho)$, by an analysis of the associated jump chain. Consider now a sample path $(Q_1(t), Q_2(t))$ that leaves the region $\{Q_1 > r, Q_2 > r\}$ by hitting the line $\{Q_1 = r\}$. The probability this sample path hits $\{Q_1 = 0\}$ before leaving $\{Q_1 \leqslant r, Q_2 > r\}$ is bounded above by $[\mu/(\lambda_1 + \lambda_3)]^r$, by a simple gambler's ruin calculation. A coupling argument then shows that

$$P \leqslant \frac{2\rho + 1}{1 - \rho} \left(\frac{\mu}{\lambda_1 + \lambda_3} \right)^r .$$

Under the heavy traffic regime defined earlier, P will approach zero provided $r(n)$, the threshold associated with the nth system, satisfies

$$r(n) \geqslant \frac{1 + \epsilon}{2 \log \left(\dfrac{\lambda_1 + \lambda_3}{\mu} \right)} \log n \tag{2.4}$$

for some $\epsilon > 0$. The coupling argument can be extended to show that provided (2.4) holds, the normalized total queue length process, $Z_1^{(n)}(t) + Z_2^{(n)}(t)$, satisfies (2.3) where $Z(t)$ is again the reflected Brownian motion describing the two server queue. The limit process $(Z_1(t), Z_2(t))$ is now two-dimensional: it is a Brownian motion in the positive orthant reflected on the axes at angles of $\pi/4$ towards the origin and out from the origin along the diagonal. (This process is of some interest in its own right: note that the angles of reflection on the axes are critical, just failing the condition of Harrison and Reiman [20]. See Varadhan and Williams [42] for the existence and uniqueness of the process, and Williams [50,51] for further properties.)

Thus the halving of mean queue lengths, achieved in heavy traffic by the shorter queue policy, can also be achieved by a threshold strategy. Further, the relation (2.4) establishes that the threshold need only grow rather slowly with n: recall that the mean number in the system is of order $n^{1/2}$. Of course for any given value of n the shorter queue policy will improve on the performance of threshold routing. However, the relative improvement is generally slight, and disappears in heavy traffic. When later we discuss more general networks, we shall see that threshold strategies are able to achieve additional effects which improve relative performance and come to dominate in heavy traffic.

Foschini [7] has considered the heavy traffic advantage of an $M/M/K$ system with service rates $(\mu_1, \mu_2, \ldots, \mu_K)$ over a system of K independent $M/M/1$ queues

with service rates $\mu_1, \mu_2, \ldots, \mu_K$ respectively. Suppose that arrivals at the latter system are distributed over the K queues in accordance with the roll of an optimally biased die, where the best bias can be found by a simple Lagrangian analysis. Then the mean number in the $M/M/K$ system is smaller by a factor of

$$\frac{\left(\sum_{k=1}^{K} \mu_k^{1/2}\right)^2}{\sum_{k=1}^{K} \mu_k} \tag{2.5}$$

in heavy traffic. Again the improvement can be obtained by any strategy which keeps all servers busy when there is substantial work to be done in the system.

3. Dynamic sequencing

In this section we consider the two station network in fig. 2. This is the system studied by Harrison and Wein [21], whose analysis we follow. Customers of types A and B arrive at station 1 according to independent renewal processes. Assume interarrival times from each of these processes have mean λ^{-1} and variance a. Type A customers require service at station 1 and then at station 2; type B customers require service at station 1 only. Servers 1 and 2 process independently customers according to general service time distributions with means of $(2\mu)^{-1}$ and μ^{-1}, and variances of s_1 and s_2.

In contrast with the system of the previous section, there are no routing decisions in this network. Rather, dynamic sequencing decisions specify, at each point in time, which type of customer to serve at station 1. The traffic intensity at each station is $\rho = \lambda/\mu$, and the Brownian approximation we shall use requires ρ to be near 1 so that the system is close to heavy traffic. More formally, the heavy traffic limit concerns a sequence of systems, indexed by n, such that $\lambda(n) \to \lambda$, $\mu(n) \to \mu$ and

$$n^{1/2}(\lambda(n) - \mu(n)) \to \theta \quad \text{as } n \to \infty,$$

where $\theta < \infty$.

Type A customers wait in queue 1 for server 1 and then in queue 2 for server 2, and type B customers wait in queue 3 for server 1. Define the queue length process $\mathbf{Q} = (Q_1, Q_2, Q_3)$ where $Q_k(t)$ is the number of customers in queue k (including the one in service, if any) at time t. Also define the idleness process $\mathbf{I} = (I_1, I_2)$ where $I_i(t)$ is the cumulative amount of time that server i is idle in the interval $[0, t]$. With the parameter n fixed, now define the scaled queue length and idleness processes

$$\mathbf{Z}(t) = \frac{\mathbf{Q}(nt)}{n^{1/2}}, \quad \mathbf{U}(t) = \frac{\mathbf{I}(nt)}{n^{1/2}}.$$

Note that the process $\mathbf{U}$ is necessarily non-decreasing, since it represents *cumulative* idleness.

We shall find it helpful to define sequencing controls relative to a simple nominal policy, under which the behaviour of the network is well understood. Under the nominal policy, suppose that server 1 devotes precisely half its service effort to queue 1 and half to queue 3. Under this nominal policy, queues 1 and 3 will behave as independent queues, each heavily loaded. Under a more general sequencing control, let $T_k(t)$ be the cumulative service time so far received at queue k in $[0, t]$. Now let

$$Y_k(t) = \begin{cases} \dfrac{\frac{1}{2}nt - T_k(nt)}{n^{1/2}} & k = 1, 3, \\ \dfrac{nt - T_k(nt)}{n^{1/2}} & k = 2. \end{cases}$$

Thus $Y_k(t)$ is a centred and scaled measure of how much service has been received by customers in queue k in $[0, t]$, centred by the maximum amount that could have been received under the nominal policy. The process $\mathbf{Y} = (Y_1, Y_2, Y_3)$ thus represents the sequencing policy. From their definitions we have the following link between the (scaled) cumulative idle time of servers $\mathbf{U}$ and the (centred and scaled) sequencing policy $\mathbf{Y}$:

$$\mathbf{AY}(t) = \mathbf{U}(t), \tag{3.1}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

The matrix $\mathbf{A}$ simply records which queues are attached to which stations: A_{ik} is 1 or 0 according to whether queue k is at station i. Thus the first component of the vector equation (3.1) records the fact that server 1 is either serving queue 1 or queue 3 or is idle: the second component of (3.1) records that server 2 is either serving queue 2 or idle.

We can write the relationship between the (scaled) queue length process $\mathbf{Z}$ and the (centred and scaled) sequencing policy $\mathbf{Y}$ as

$$\mathbf{Z}(t) = \mathbf{X}(t) + \mathbf{RY}(t), \tag{3.2}$$

where

$$\mathbf{R} = \mu \begin{pmatrix} 2 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Here $\mathbf{X}$ is a three-dimensional process which we shall approximate by a Brownian motion with drift vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Gamma}$ where

$$\boldsymbol{\theta} = n^{1/2}(\lambda - \mu)\,(1,0,1)^{\mathrm{T}}$$

and

$$\boldsymbol{\Gamma} = \begin{pmatrix} \lambda^3 a + 4\mu^3 s_1 & -4\mu^3 s_1 & 0 \\ -4\mu^3 s_1 & \mu^3(4s_1 + s_2) & 0 \\ 0 & 0 & \lambda^3 a + 4\mu^3 s_1 \end{pmatrix}.$$

The process $\mathbf{X}$ can be interpreted as follows. Suppose for the moment that $Z_1(0), Z_2(0), Z_3(0) > 0$, and that station 1 devotes precisely half its service effort to queue 1 and half to queue 2, so that $Y_k(t)$ only increases when $Z_k(t) = 0$: this is the nominal policy described earlier. Then $X_k(t)$ represents the (scaled) queue lengths up until the time that one of the queue lengths hits zero. Further, in the heavy traffic limit the queue lengths $\mathbf{Z}$ are indeed given by a reflected Brownian motion (3.2), with reflection matrix $\mathbf{R}$ (Harrison and Reiman [20], Reiman [39]). Note that R_{kl} is just the rate at which service of queue l depletes queue k. Under other sequencing policies $\mathbf{Y}$ the process $\mathbf{Z}$ can take other forms: for example if server 1 serves the longer queue then $Z_1(t)$ and $Z_2(t)$ will be held approximately equal. Harrison's [16] Brownian approximation is to take $\mathbf{X}$ to be a $(\boldsymbol{\theta}, \boldsymbol{\Gamma})$-Brownian motion in (3.2), whatever the sequencing policy $\mathbf{Y}$.

Define the matrix $\mathbf{M} = (M_{ik})$ by

$$\mathbf{M} = \mathbf{A}\mathbf{R}^{-1} = \mu^{-1}\begin{pmatrix} 1/2 & 0 & 1/2 \\ 1 & 1 & 0 \end{pmatrix}.$$

Then M_{ik} is the expected amount of time that server i must devote to a customer in queue k before the customer leaves the system. Define $\mathbf{W} = (W_1, W_2)$ by

$$\mathbf{W}(t) = \mathbf{M}\mathbf{Z}(t)\,. \tag{3.3}$$

Thus $W_i(t)$ is the (scaled) workload for server i in the system at time t. We can find an equation for the workload process $\mathbf{W}$ by multiplying (3.2) by $\mathbf{M}$ and using (3.1) to obtain

$$\mathbf{W}(t) = \mathbf{B}(t) + \mathbf{U}(t)\,, \tag{3.4}$$

where $\mathbf{B}(t) = \mathbf{M}\mathbf{X}(t)$ and is hence a Brownian motion with drift vector $\mathbf{M}\boldsymbol{\theta}$ and covariance matrix $\mathbf{M}\boldsymbol{\Gamma}\mathbf{M}^{\mathrm{T}}$. Thus given a sequencing policy $\mathbf{Y}$ we can construct $\mathbf{U}$ and $\mathbf{Z}$ satisfying (3.3) and (3.4). Conversely, given $\mathbf{U}$ and $\mathbf{Z}$ satisfying (3.3) and (3.4),

$$\mathbf{Y}(t) = \mathbf{R}^{-1}(\mathbf{Z}(t) - \mathbf{X}(t))$$

satisfies (3.1) and (3.2). The problem of choosing a pair $(\mathbf{U}, \mathbf{Z})$, with $\mathbf{U}$ non-decreasing and $\mathbf{Z}$ non-negative, to satisfy (3.3) and (3.4) and to minimize a performance measure is termed the *workload formation* of the Brownian network model.

Suppose now that the system starts from empty at time $t = 0$: take $\mathbf{Z}(0)=\mathbf{X}(0)=\mathbf{0}$. Consider the equation (3.4). The Brownian motion $\mathbf{B}$ is given, and so we can minimize $W_i(t)$ for $i = 1, 2$ over non-decreasing U_i by taking

$$U_i(t) = -\inf_{0 \leqslant s \leqslant t} B_i(s). \tag{3.5}$$

Then, from (3.3) and (3.4),

$$\begin{aligned} \tfrac{1}{2}(Z_1(t) + Z_3(t)) &= b_1(t), \\ Z_1(t) + Z_2(t) &= b_2(t), \end{aligned}$$

where

$$b_i(t) = \mu\left(B_i(t) - \inf_{0 \leqslant s \leqslant t} B_i(s)\right),$$

and hence

$$\sum_{k=1}^{3} Z_k(t) \geqslant 2b_1(t) \vee b_2(t). \tag{3.6}$$

However, observe that

$$\begin{aligned} Z_1(t) &= 2b_1(t) \wedge b_2(t), \\ Z_2(t) &= [b_2(t) - 2b_1(t)]^+, \\ Z_3(t) &= [2b_1(t) - b_2(t)]^+ \end{aligned} \tag{3.7}$$

attain the bound (3.6). Hence we can find a *pathwise solution* (Harrison [16]): that is a pair $(\mathbf{U}, \mathbf{Z})$, given by (3.5) and (3.7), which minimizes $\sum_k Z_k(t)$ at all times $t \geqslant 0$ simultaneously.

Under the pathwise solution,

$$U_i \text{ only increases when } W_i = 0 \quad i = 1, 2 \tag{3.8}$$

and

$$\text{either } Z_2(t) \text{ or } Z_3(t) = 0. \tag{3.9}$$

Condition (3.8) will be satisfied for $i = 1$ provided that server 1 is never idle when there are customers waiting at station 1. An extreme way to attempt to satisfy (3.8) for $i = 2$ is to give priority to type A customers at station 1. Although even under this policy server 2 may be idle when its workload is non-zero (the first arrival after $t = 0$ is a type A customer, for example) the policy does satisfy (3.8) in heavy traffic (Peterson [36]). However, this static priority discipline violates condition (3.9): it retains more customers in the system than necessary in the short term since type B customers leave the system immediately after service at 1 whereas type A do not. To attempt to satisfy (3.8) and (3.9), type A customers should be given priority at

station 1 only when station 2 is threatened with idleness. Perhaps the simplest way to do this is to give priority to type B customers at station 1 unless the number of customers at station 2 is less than some threshold value, in which case priority is given to type A customers.

The above example illustrates an important feature in the control of open queueing networks: there is a trade-off between short term minimization of the system population and long term minimization of server idleness. The first of these objectives suggests giving priority to customers of type B over type A at station 1 while the second suggests reversing these priorities. The above dynamic priority policy attempts to handle this trade-off. Simulation results (Harrison and Wein [21]) suggest that the threshold policy will achieve its aim, and that the improvements of the policy over first come first served or either of the two possible static priority disciplines at station 1 are significant.

4. Routing and sequencing

We now extend the two station network of section 3 to that shown in fig. 3; this example is from Wein [46]. Type B customers now require service at either station 1 or station 2: at the time of arrival, a type B customer is routed to either queue 3 or queue 4.

As in section 3 the two types of customers arrive according to independent renewal processes, interarrival times having mean λ^{-1} and variance a. Assume that servers 1 and 2 process customers independently and according to general service time distributions, each with mean μ^{-1} and variance s. Also, assume that $3\lambda - 2\mu$ is of order $n^{-1/2}$ so that both stations are heavily loaded. With n fixed, define scaled queue length and idleness processes $\mathbf{Z} = (Z_k, k = 1,2,3,4)$ and $\mathbf{U} = (U_i, i = 1,2)$ as in section 3. Again we shall find it helpful to describe a simple nominal policy, under which the behaviour of the network is well understood: both routing and sequencing controls will be defined relative to the nominal policy. Under the nominal policy suppose customers of type B are routed randomly to queues 3 and 4, each with probability 1/2 and independently of previous routing decisions. Further, suppose that server 1 devotes precisely 2/3 of its effort to queue 1, and 1/3 to queue 3. Thus queues 1 and 3 will behave as independent queues, and the (2/3) : (1/3) split ensures both are heavily loaded. Similarly suppose that server 2 devotes 2/3 of its effort to queue 2, and 1/3 to queue 4. Let

$$Y_k(t) = \begin{cases} \dfrac{\frac{2}{3}nt - T_k(nt)}{n^{1/2}} & k = 1,2, \\ \dfrac{\frac{1}{3}nt - T_k(nt)}{n^{1/2}} & k = 3,4, \end{cases}$$

where $T_k(t)$ is the amount of service time received at queue k in $[0, t]$.

Define the matrix $\mathbf{P} = (P_{kl}, k, l = 1,2,3,4)$ where P_{kl} is the probability that a customer joins queue l after queue k; so $P_{12} = 1$ and $P_{kl} = 0$ otherwise. Let $\mathbf{I}_4$ be the four-dimensional identity matrix and let

$$\mathbf{R} = \mu(\mathbf{I}_4 - \mathbf{P}^{\mathrm{T}}),$$

so that, as in section 3, R_{kl} is the (possibly negative) rate at which service of queue l depletes queue k. Then the relationship between $\mathbf{Z}$ and the sequencing policy $\mathbf{Y} = (Y_k, k = 1,2,3,4)$ is given by

$$\mathbf{Z}(t) = \mathbf{X}(t) + \mathbf{R}\mathbf{Y}(t).$$

We make the approximation, which is again that of Harrison [16] and justified for the nominal policy by Reiman [39], that $\mathbf{X}$ is a Brownian motion with drift vector $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Gamma}$ where

$$\boldsymbol{\theta} = \frac{n^{1/2}(3\lambda - 2\mu)}{6}(2,0,1,1)^{\mathrm{T}}$$

and

$$\boldsymbol{\Gamma} = \begin{pmatrix} \lambda^3 + \frac{2}{3}\mu^3 s & -\frac{2}{3}\mu^3 s & 0 & 0 \\ -\frac{2}{3}\mu^3 s & \frac{4}{3}\mu^3 s & 0 & 0 \\ 0 & 0 & \frac{1}{4}(\lambda^3 a + \lambda) + \frac{1}{3}\mu^3 s & \frac{1}{4}(\lambda^3 a - \lambda) \\ 0 & 0 & \frac{1}{4}(\lambda^3 a - \lambda) & \frac{1}{4}(\lambda^3 a + \lambda) + \frac{1}{3}\mu^3 s \end{pmatrix}.$$

To include the effect of routing decisions on $\mathbf{Z}$ let $\hat{V}_l(t)$ be the number of arrivals actually routed to queue l minus the number routed to l under the nominal (random routing) policy in $[0, t]$, for $l = 3,4$, and let

$$V_l(t) = \frac{\hat{V}_l(nt)}{n^{1/2}} \quad l = 3,4.$$

Thus $V_3(t) + V_4(t) = 0$. Let $\mathbf{V}(t) = (V_3(t), V_4(t))$. Then the approximate relationship between the queue length process $\mathbf{Z}$, the control policy $(\mathbf{Y}, \mathbf{V})$ and the $(\boldsymbol{\theta}, \boldsymbol{\Gamma})$-Brownian motion $\mathbf{X}$ is given by

$$\mathbf{Z}(t) = \mathbf{X}(t) + \mathbf{R}\mathbf{Y}(t) + \mathbf{G}\mathbf{V}(t), \tag{4.1}$$

where the matrix $\mathbf{G} = (\mathbf{0}_2, \mathbf{I}_2)^{\mathrm{T}}$ links the routing controls (V_3, V_4) directly to the queue lengths (Z_3, Z_4). Here $\mathbf{0}_2$ and $\mathbf{I}_2$ are the 2×2 zero and identity matrices respectively. The control policy $(\mathbf{Y}, \mathbf{V})$ and idleness process $\mathbf{U}$ satisfy

$$\mathbf{A}\mathbf{Y}(t) = \mathbf{U}(t), \quad \mathbf{H}\mathbf{V}(t) = 0, \tag{4.2}$$

where

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \quad \mathbf{H} = (1,1).$$

So under the Brownian network model we choose a sequencing policy **Y** and a routing policy **V** satisfying eqs. (4.1) and (4.2), with **U** non-decreasing and **Z** non-negative. As in section 3, we can find a reduced form of this model in terms of a workload process. Again the (i,k) element of the matrix

$$\mathbf{AR}^{-1} = \mu^{-1}\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$$

gives the expected amount of time that server i must devote to a customer in queue k before the customer leaves the system. But now define

$$\mathbf{M} = (1,1)\mathbf{AR}^{-1} = \mu^{-1}(2,1,1,1);$$

M_k is thus the expected amount of time that must be devoted to a customer in queue k by either or both servers before the customer leaves the system. Again let

$$W(t) = \mathbf{MZ}(t). \tag{4.3}$$

Then W is the system workload process: $W(t)$ is the (scaled) expected amount of service time required at stations 1 and 2 by the customers in the system at time t. Using equations (4.1) and (4.2) we find

$$W(t) = B(t) + U_1(t) + U_2(t), \tag{4.4}$$

where $B(t) = \mathbf{MX}(t)$. So given a control policy (**Y**, **V**) we can construct a pair (**U**, **Z**) which satisfy eqs. (4.3) and (4.4). Similarly given any pair (**U**, **Z**) which satisfy the equations for the reduced system model, (4.3) and (4.4), it can be shown that there is a (**Y**, **V**) satisfying eqs. (4.1) and (4.2). Hence the workload formulation of the Brownian network model, that of choosing a pair (**U**, **Z**) with **U** non-decreasing and **Z** non-negative to satisfy eqs. (4.3) and (4.4), is equivalent to the original formulation of the model in terms of the control policy (**Y**, **V**).

Since U_1, U_2 represent cumulative idleness, we minimize $W(t)$ at all times simultaneously by choosing U_1, U_2 non-decreasing such that

$$U_1(t) + U_2(t) = -\inf_{0 \leqslant s \leqslant t} B(s). \tag{4.5}$$

Then eqs. (4.3) and (4.4) imply

$$\sum_{k=1}^{4} Z_k(t) \geqslant \tfrac{1}{2} b(t), \tag{4.6}$$

where

$$b(t) = \mu\left(B(t) - \inf_{0 \leqslant s \leqslant t} B(s)\right).$$

Further, the bound (4.6) is attained for all t by

$$Z_1(t) = \tfrac{1}{2}b(t)\,,$$
$$Z_k(t) = 0 \qquad k = 2,3,4\,. \tag{4.7}$$

As in the previous section we can minimize $\sum_k Z_k(t)$ for all $t \geqslant 0$, through the pair (**U**, **Z**) given by (4.5), (4.7), to obtain a pathwise solution. Again the existence of a pathwise solution implies that the system population over the short term and server idleness over the long term can be minimized simultaneously. However, here the equations for the reduced system model, (4.3) and (4.4), involve the combined workload for stations 1 and 2 and not the workloads for these stations individually. It is the inclusion of routing decisions which causes these two workloads to merge. As a result, the system exhibits resource pooling under the pathwise solution: when one of servers 1 and 2 is busy so is the other, and both are idle only when there is no work anywhere in the system. It is as though servers 1 and 2 are combined to form a single pooled resource. Further, a customer in queue 1 comprises twice the system workload of a customer in another queue and, since W is constrained by (4.4), the system population is minimized by taking $Z_k > 0$ if and only if $k = 1$. Various forms of control can achieve these desired effects in heavy traffic: for example, route a customer of type B to queue 3 if there are fewer customers there than in queues 2 and 4 together, and give priority to queue 3 at station 1 except when the total number of customers at station 2 is below a threshold.

For simplicity of exposition we have assumed that servers 1 and 2 are identical, and that the two arrival streams have the same characteristics. The analysis can be generalized to the case where service times at station i have a general distribution with mean μ_i^{-1} and variance s_i, for $i = 1, 2$, and where interarrival times of type A (respectively B) customers have mean λ_A^{-1} (λ_B^{-1}) and variance a_A (a_B). Assume $2\lambda_A + \lambda_B$ approaches $\mu_1 + \mu_2$ so that the system is in heavy traffic, and that $\lambda_A < \mu_1, \mu_2$ so that no server is individually overloaded. Under the optimal control the total (scaled) number of service completions necessary to empty the system is approximated by a reflected Brownian motion with drift $n^{1/2}(2\lambda_A + \lambda_B - \mu_1 - \mu_2)$ and variance $(4\lambda_A^3 a_A + \lambda_B^3 a_B + \mu_1^3 s_1 + \mu_2^3 s_2)$. This represents twice the (scaled) number of customers in queue 1: hence in heavy traffic the total (scaled) number of customers in the system (equivalently, in queue 1) will be a reflected Brownian motion with drift $\theta = n^{1/2}(2\lambda_A + \lambda_B - \mu_1 - \mu_2)/2$ and variance $(4\lambda_A^3 a_A + \lambda_B^3 a_B + \mu_1^3 s_1 + \mu_2^3 s_2)/4$. Assuming $\theta < 0$, the mean of the stationary distribution of this reflected Brownian motion is

$$\tfrac{1}{8}(4\lambda_A^3 a_A + \lambda_B^3 a_B + \mu_1^3 s_1 + \mu_2^3 s_2)|\theta|^{-1}\,. \tag{4.8}$$

To illustrate the magnitude of the improvement possible, consider the special case where the arrival processes are Poisson, service times are exponential, type B customers are routed randomly in accordance with the toss of a fair coin, queues operate a first come first served discipline and $\lambda_1 = \lambda_2 = \lambda, \mu_1 = \mu_2 = \mu$. Then the mean number of customers in the system is readily calculated to be $6\lambda(2\mu - 3\lambda)^{-1}$. If the system of two servers is replaced by a single server operating at twice the

rate, with customers of type A recycled to the end of the queue after their first service, then the mean number of customers becomes $3\lambda(2\mu - 3\lambda)^{-1}$. Thus resource pooling can be viewed as responsible for a factor of 2 improvement. By holding the workload in queue 1 the optimal policy improves on this by a further factor of 3/2.

5. A four station network

We next consider control of the network with four single-server stations shown in fig. 4. The system is used by two different types of customers, horizontal and vertical, and there are two queues at each station to distinguish between these different customer types. Horizontal customers pass through the system by using either the top row, queueing at station 1 and then at station 2 before leaving, or the bottom row, queueing at station 3 then station 4. Similarly vertical customers use either the left or right hand column. A network control policy determines both routing and sequencing decisions. A routing policy specifies, at the time of a customer's arrival, which route it takes through the network. A sequencing policy specifies which type of customer to serve at each station at each point in time.

Horizontal and vertical customers arrive according to independent renewal processes and each server processes customers according to a general service time distribution, the service times of all customers being independent. Suppose horizontal and vertical customers arrive at rates λ_H and λ_V respectively, and that all service times have mean μ^{-1}.

We again use the general framework of Harrison [16] to obtain a Brownian network model which approximates the system described above. Assume that $\lambda_H + \lambda_V - 2\mu$ is of order $n^{-1/2}$ so that the system is close to heavy traffic. With n fixed, define the scaled idleness and queue length processes $\mathbf{U}$ and $\mathbf{Z}$ (four- and eight-dimensional, respectively) as in section 3. The obvious nominal policy is to route each arrival to one of the two available routes with probability 1/2, and to have each server devote proportions p_H and p_V of its service effort to horizontal and vertical customers respectively, where

$$p_H = \frac{\lambda_H}{\lambda_H + \lambda_V}, \quad p_V = \frac{\lambda_V}{\lambda_H + \lambda_V}.$$

Let

$$Y_k(t) = \begin{cases} \dfrac{p_H nt - T_k(nt)}{n^{1/2}} & k = 1,3,5,7, \\ \dfrac{p_V nt - T_k(nt)}{n^{1/2}} & k = 2,4,6,8, \end{cases}$$

where $T_k(t)$ is the amount of service time received at queue k in $[0,t]$. Let $\mathbf{P} = (P_{kl}, k, l = 1,2,\ldots,8)$ where P_{kl} is the probability that a customer joins queue

l after queue k; so $P_{k,k+4} = 1$ for $k = 1, \ldots, 4$ and $P_{kl} = 0$ otherwise. Also, let $\mathbf{I}_8$ be the eight-dimensional identity matrix and let

$$\mathbf{R} = \mu(\mathbf{I}_8 - \mathbf{P}^{\mathrm{T}}) .$$

Then the relationship between $\mathbf{Z}$ and the (centred and scaled) sequencing policy $\mathbf{Y} = (Y_k, k = 1, 2, \ldots, 8)$ is given by

$$\mathbf{Z}(t) = \mathbf{X}(t) + \mathbf{R}\mathbf{Y}(t) ,$$

where we approximate the eight-dimensional process $\mathbf{X}$ by a Brownian motion with a certain drift vector and covariance matrix. That $\mathbf{X}$ is indeed an approximate Brownian motion follows from a heavy traffic limit theorem of Reiman [39], at least for the nominal policy.

To include the effect of routing controls on $\mathbf{Z}$, let $\hat{V}_l(t)$ be the number of arrivals actually routed to queue l minus the number routed to l under the nominal (random routing) policy in $[0, t]$, and let

$$V_l(t) = \frac{\hat{V}_l(nt)}{n^{1/2}} \quad l = 1, 2, 3, 4 .$$

Then the approximate relationship between $\mathbf{Z}$, the control policy $\mathbf{Y} = (Y_k, k = 1, 2, \ldots, 8)$, $\mathbf{V} = (V_l, l = 1, 2, 3, 4)$, and the Brownian motion $\mathbf{X}$ is given by

$$\mathbf{Z}(t) = \mathbf{X}(t) + \mathbf{R}\mathbf{Y}(t) + \mathbf{G}\mathbf{V}(t) , \tag{5.1}$$

where $\mathbf{G} = (\mathbf{I}_4, \mathbf{0}_4)^{\mathrm{T}}$, and $\mathbf{I}_4$ and $\mathbf{0}_4$ are the 4×4 identity and zero matrices. The control policy $(\mathbf{Y}, \mathbf{V})$ must also satisfy

$$\mathbf{A}\mathbf{Y}(t) = \mathbf{U}(t), \quad \mathbf{H}\mathbf{V}(t) = \mathbf{0} , \tag{5.2}$$

where $\mathbf{U}$ is the (scaled) idleness process. Here $\mathbf{A} = (A_{ik}, i = 1, 2, 3, 4;\ k = 1, 2, \ldots, 8)$ and A_{ik} is 1 or 0 according to whether queue k is at station i, and

$$\mathbf{H} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} .$$

Thus again $\mathbf{A}$ and $\mathbf{H}$ simply record which queues are at which stations, and which queues are accessed by which arrival streams.

As in sections 3 and 4 we can reformulate our Brownian network control problem in terms of workloads. Again $(\mathbf{A}\mathbf{R}^{-1})_{ik}$ gives the expected amount of time that server i must devote to a customer in queue k before the customer leaves the system. But now let

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \mathbf{A}\mathbf{R}^{-1} = \mu^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \end{pmatrix} .$$

Define the workload process $\mathbf{W} = (W_{14}, W_{23})$ by

$$\mathbf{W}(t) = \mathbf{M}\mathbf{Z}(t)\,. \tag{5.3}$$

Then W_{14} (respectively W_{23}) is the expected amount of (scaled) service time required at stations 1 and 4 (stations 2 and 3) before all customers currently in the system leave. Multiplying (5.1) by $\mathbf{M}$ and using (5.2) gives the pair of equations

$$\begin{aligned} W_{14}(t) &= B_1(t) + U_1(t) + U_4(t)\,, \\ W_{23}(t) &= B_2(t) + U_2(t) + U_3(t)\,, \end{aligned} \tag{5.4}$$

where $\mathbf{B}(t) = \mathbf{M}\mathbf{X}(t)$. The original formulation of the Brownian model is in terms of the control policy $(\mathbf{Y}, \mathbf{V})$, while the workload formulation requires choosing a pair $(\mathbf{U}, \mathbf{Z})$, with $\mathbf{U}$ non-decreasing and $\mathbf{Z}$ non-negative, subject to (5.3) and (5.4). These two formulations are equivalent (see Laws and Louth [34] for a proof) and we now concentrate on that based on workloads.

To minimize both $W_{14}(t)$ and $W_{23}(t)$ for all $t \geqslant 0$ choose $\mathbf{U}$ such that

$$\begin{aligned} U_1(t) + U_4(t) &= -\inf_{0 \leqslant s \leqslant t} B_1(s)\,, \\ U_2(t) + U_3(t) &= -\inf_{0 \leqslant s \leqslant t} B_2(s)\,. \end{aligned} \tag{5.5}$$

Now

$$\sum_{k=1}^{8} Z_k(t) \geqslant \mu(W_{14}(t) \vee W_{23}(t))\,,$$

and hence

$$\sum_{k=1}^{8} Z_k(t) \geqslant b_1(t) \vee b_2(t)\,, \tag{5.6}$$

where

$$b_r(t) = \mu\left(B_r(t) - \inf_{0 \leqslant s \leqslant t} B_r(s)\right).$$

In fact we can attain the bound (5.6) at all times $t \geqslant 0$ by choosing $\mathbf{Z}$ such that

$$\begin{aligned} \sum_{k=1}^{4} Z_k(t) &= b_1(t) \wedge b_2(t)\,, \\ Z_5(t) + Z_6(t) &= [b_2(t) - b_1(t)]^+\,, \\ Z_7(t) + Z_8(t) &= [b_1(t) - b_2(t)]^+\,. \end{aligned} \tag{5.7}$$

So as in previous sections we can minimize $\sum_k Z_k(t)$ for all $t \geqslant 0$ simultaneously, by setting $(\mathbf{U}, \mathbf{Z})$ as given by (5.5) and (5.7).

The behaviour of the network under this pathwise solution can be interpreted in terms of the reduced fork–join queueing system illustrated in fig. 5. Consider two stations each with a single first come first served (FCFS) queue and a single server of rate 2μ. The first server represents the combination of servers 1 and 4, and the second the combination of servers 2 and 3. Customers arrive as for the original system, at a total rate of $\lambda_H + \lambda_V$. Each arriving customer sends a token to both queues simultaneously and these progress independently until service completion when they return to the customer. A customer leaves the system once both tokens have returned. Let the number of tokens waiting for the (1, 4)-queue and the (2, 3)-queue be w_{14} and w_{23} respectively. Since both queues are FCFS, the number of customers in the reduced system is $w_{14} \vee w_{23}$.

Under the pathwise solution, servers 1 and 4 (respectively 2 and 3) are only idle when $W_{14} = 0$ ($W_{23} = 0$), the workload for servers 1 and 4 (2 and 3) is b_1 (b_2) and the system population is $b_1 \vee b_2$. Hence, after suitable scaling, the behaviour of the fork–join system and the original network under the pathwise solution are identical. So, in heavy traffic and under the optimal routing and sequencing policies, it is as though servers 1 and 4, and servers 2 and 3, are combined to form two pooled resources. Also, the network behaves as if customers queue for their two services in parallel rather than in series.

The results above do not require symmetric service rates: they hold for arrival rates λ_H, λ_V and service rates $\mu_i, i = 1, 2, 3, 4$ provided that

$$\lambda_H + \lambda_V \approx \mu_1 + \mu_4 \approx \mu_2 + \mu_3\,, \tag{5.8}$$

$$\lambda_H < (\mu_1 + \mu_3) \wedge (\mu_2 + \mu_4)\,, \tag{5.9}$$

$$\lambda_V < (\mu_1 + \mu_2) \wedge (\mu_3 + \mu_4)\,. \tag{5.10}$$

More formally, we would consider a sequence of systems, indexed by n, whose limiting rates satisfied (5.8) with exact equality, (5.9) and (5.10). Further, as in all of our examples, the processes generating arrival and service events need not be independent renewal processes. It is sufficient that arrival and service processes jointly satisfy a functional central limit theorem: that is, in the limit $n \to \infty$, the arrival and service processes, suitably centred and scaled, converge to a multidimensional Brownian motion. Subject to this requirement arrivals and services can be arbitrary and, in particular, they need not be independent. In this four station example, the Brownian model would only be affected via the covariance matrix of the Brownian motion **B**: the pathwise solution and its fork–join queueing interpretation would still hold.

In Laws and Louth [34] a simple control policy is discussed, which attempts to achieve the heavy traffic performance of the pathwise solution and to perform well in moderate traffic. One feature of the policy is a threshold at station 4, which indicates when this station is threatened with idleness. Stations 2 and 3 give priority to the exiting customers in queues 5 and 6, except when the total number of customers in queues 7 and 8 drops below threshold. By this means the policy attempts to

keep *either* queues 5 and 6 near empty *or* queues 7 and 8 near empty, as in the pathwise solution (5.7). Customers are routed so as to equalize the workloads for stations 1 and 4, and to equalize the workloads for stations 2 and 3. By this means the policy attempts to keep stations 1 and 4 busy together, and stations 2 and 3 busy together. Simulation results (Laws and Louth [34]) show the benefits of dynamic rather than static control policies, and of dynamic routing in particular, to be substantial.

6. Further examples

In this section we show that the analysis of the four station network extends straightforwardly to certain other forms of network. However, we also find that the existence of a pathwise solution is not guaranteed, and that it is necessary to generalize our definition of a pooled resource. Our examples are taken from Laws [32].

THE $2 \times 2 \times 2$ CUBE

We consider the extension of the four station, 2×2 square of section 5 to the $2 \times 2 \times 2$ cube shown in fig. 9. The eight stations, situated at the vertices of the cube, are labelled as shown and there are three arrival streams. Customers arrive at the top, front and left hand faces of the cube and they leave the system at the opposite face. The four possible routes for a given customer type correspond to the four edges connecting the face at which these customers arrive and the opposite

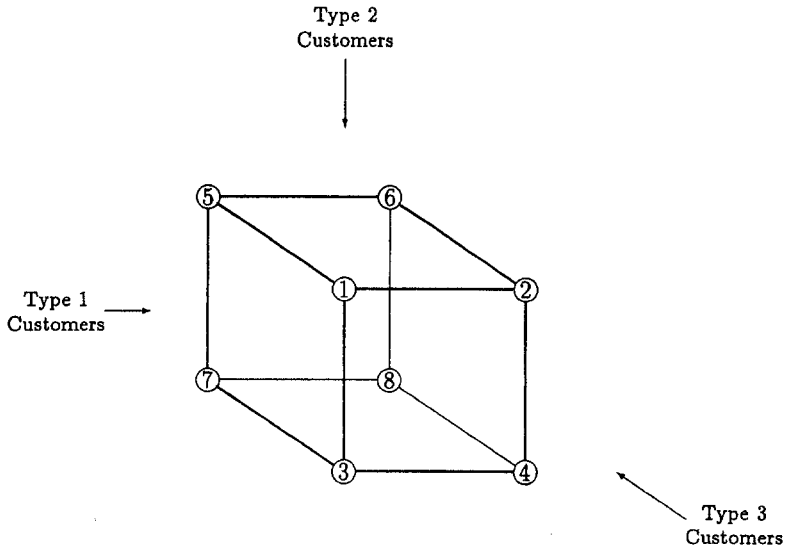

Fig. 9. The $2 \times 2 \times 2$ cube.

face. Customers of type 1 use one of routes (1, 2), (3, 4), (5, 6) and (7, 8); type 2 use one of routes (1, 3), (2, 4), (5, 7) and (6, 8); and type 3 use one of routes (1, 5), (2, 6), (3, 7) and (4, 8). (In general, a customer using route $r = (i_1, i_2, \ldots, i_n)$ queues for service at each of stations $i_1, i_2, \ldots, i_n$ in turn before leaving the system.)

Suppose that the three arrival rates are λ, that all eight service rates are μ, and that $3\lambda \approx 4\mu$ so that the system is close to heavy traffic. Define the sets of servers

$$\mathcal{S}_1 = \{1, 4, 6, 7\}, \quad \mathcal{S}_2 = \{2, 3, 5, 8\}.$$

Observe that every arrival requires exactly one service from each set of servers $\mathcal{S}_1$ and $\mathcal{S}_2$. Approximating the system behaviour by a Brownian network model, and proceeding as in the previous sections, we can obtain the workload formulation of the Brownian model. The important workload process is two-dimensional, the two components corresponding to the system workload for the sets of servers $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. Thus the reduced system model of the $2 \times 2 \times 2$ cube is similar to that of the four station network, except that the sets of servers $\mathcal{S}_1$ and $\mathcal{S}_2$ replace $\{1, 4\}$ and $\{2, 3\}$. As before, under the Brownian network model we can find a control policy which minimizes the total system population at all times $t \geqslant 0$ simultaneously, a pathwise solution. The interpretation of this solution in terms of a fork–join queueing system is the same as that in section 5, except that the two pooled resources are now the combinations of the sets of stations $\mathcal{S}_1$ and $\mathcal{S}_2$. In fact, the results described here extend to an N-dimensional hypercube of 2^N stations and N arrival streams with 2^{N-1} routes for each arrival stream. The two pooled resources are the set of *odd* stations and the set of *even* stations, a station being odd or even according to whether the number of steps (via edges of the cube) to reach it from station 1 is odd or even.

A SIX STATION NETWORK

Consider the six station network with three arrival streams shown in fig. 10. Customers of type 1 use either of routes (1, 6, 5) and (2, 3, 4); type 2 use either of routes (2, 1, 6) and (3, 4, 5); and type 3 use either of routes (3, 2, 1) and (4, 5, 6).

Consider the symmetric case where the three arrival rates are λ and the six service rates are μ, and suppose $3\lambda \approx 2\mu$, so that the system is close to heavy traffic. The system performance under the Brownian model is constrained by the total workload for the pairs of servers $\{1, 4\}$, $\{2, 5\}$ and $\{3, 6\}$. For $r = 1, 2, 3$ let $\hat{W}_r(t)$ be the total (scaled) number of customers in the system at time t requiring service from either of stations r and $r + 3$, let $W_r(t) = \mu^{-1} \hat{W}_r(t)$, and let $U_i(t)$ denote the cumulative (scaled) idleness of server i over $[0, t]$. Then the constraints on the workloads W_r are

$$W_r(t) = B_r(t) + U_r(t) + U_{r+3}(t) \quad r = 1, 2, 3,$$

where $\mathbf{B} = (B_1, B_2, B_3)$ is a Brownian motion with a certain drift vector and covariance matrix.

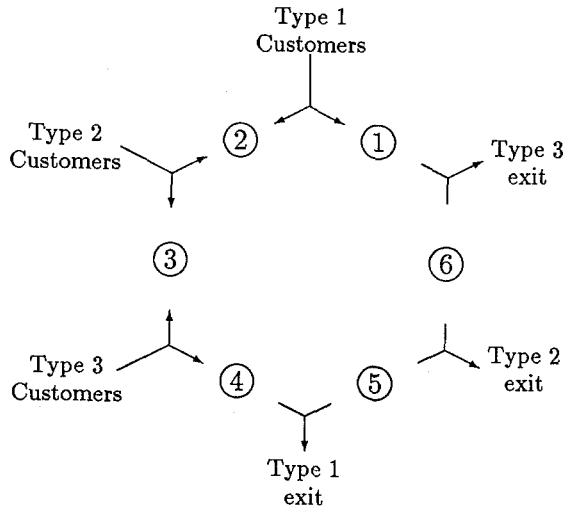

Fig. 10. The six station network.

Choosing server idleness processes such that

$$U_r(t) + U_{r+3}(t) = -\inf_{0 \leqslant s \leqslant t} B_r(s) \quad r = 1, 2, 3$$

minimizes $W_r(t)$ at all times $t \geqslant 0$ simultaneously. Further, there is a choice of queue lengths under which the total system population, $\sum_k Z_k(t)$, satisfies

$$\sum_k Z_k(t) = b_1(t) \vee b_2(t) \vee b_3(t)\,,$$

where

$$b_r(t) = \mu\left(B_r(t) - \inf_{0 \leqslant s \leqslant t} B_r(s)\right).$$

This choice of queue lengths is a pathwise solution in that it minimizes the system population at all times $t \geqslant 0$. Under this solution the network behaves like the reduced fork–join queueing system of fig. 11, which operates in a similar manner to that of fig. 5. Servers r and $r+3$ serve a single queue according to a FCFS discipline, and each arriving customer sends a token to all three queues and leaves once all of its tokens have returned. Again there is a pooling of resources, the three pooled resources being $\{1, 4\}$, $\{2, 5\}$ and $\{3, 6\}$, and in the reduced system customers queue at these resources in parallel rather than in series. As in the previous examples of this section and of section 5, the pooled resources are identified by the fact that they represent cut sets of servers, from which each arrival requires exactly one service.

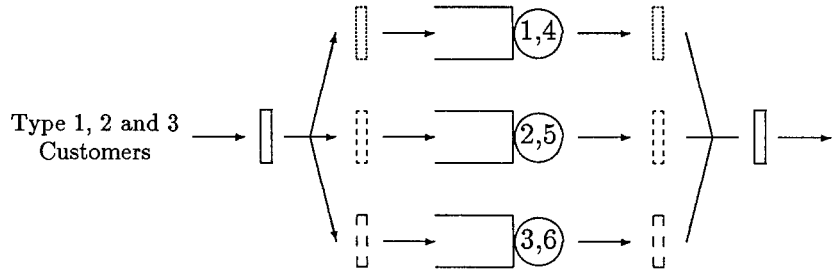

Fig. 11. The reduced fork–join system.

We can extend the example of this section to a network with eight stations in a ring and four arrival streams, where each customer uses either the four stations to the left or the four stations to the right of its arrival point. The workload formulation of the Brownian network model is of the same form as that above, there being constraints on the workloads, W_r, for the pairs of servers $\{r, r+4\}$, $r = 1, \ldots, 4$. However there is no control policy which minimizes the system population at all times $t \geqslant 0$ (see Laws [32]). In contrast with our previous examples, we cannot minimize both the short term system population and long term server idleness simultaneously: the trade-off between these two objectives has an important effect. Although we can minimize server idleness over $[0, t]$ for all $t \geqslant 0$, and hence minimize $W_r(t)$ for all $t \geqslant 0$, this minimization requires a higher system population than necessary on some time intervals. For work on necessary and sufficient conditions for pathwise minimization of the objective function $\sum_k h_k Z_k$, for constant holding costs $h_k \geqslant 0$, see Yang [53] for networks with sequencing controls and Laws [32] for networks with routing and sequencing controls.

THE 2 × 3 NETWORK

We now consider extending the network of section 5 to the 2 × 3 network of fig. 6. Consideration of the symmetric version of this network leads on to the general case discussed in the following section.

Type 1 customers use either route (1, 2, 3) or route (4, 5, 6); type 2 customers use one of routes (1, 4), (2, 5) and (3, 6). Suppose both arrival streams have rate λ, all six servers have rate μ and that $5\lambda \approx 6\mu$. Then the network is close to heavy traffic and, as before, we use a Brownian network model to approximate the system's behaviour. Previously the important workloads could be identified by finding cut sets: that is, sets of servers such that every arriving customer required exactly one service from each set. This is not possible here, but we can find sets of servers with similar properties provided we attach different weights to the servers. Let $\mathbf{A} = (A_{ik})$ where A_{ik} is 1 or 0 according to whether queue k is at station i, let $\mathbf{P} = (P_{kl}, k, l = 1, 2, \ldots, 12)$ where P_{kl} is 1 or 0 according to whether a customer leaving queue k joins queue l, and let

$$\mathbf{R} = \mu(\mathbf{I}_{12} - \mathbf{P}^{\mathrm{T}}),$$

where $\mathbf{I}_{12}$ is the 12×12 identity matrix. Then define

$$\mathbf{M} = \mathbf{DAR}^{-1} = \mu^{-1} \begin{pmatrix} 3 & 3 & 2 & 2 & 2 & 1 & 3 & 0 & 2 & 1 & 1 & 1 \\ 3 & 3 & 2 & 2 & 2 & 1 & 3 & 0 & 1 & 2 & 0 & 2 \\ 3 & 3 & 2 & 2 & 2 & 3 & 1 & 2 & 0 & 1 & 1 & 1 \end{pmatrix},$$

where

$$\mathbf{D} = \begin{pmatrix} 2 & 0 & 1 & 0 & 2 & 1 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 0 & 2 & 1 & 2 & 0 & 1 \end{pmatrix}.$$

Also define the workload process $\mathbf{W} = (W_r, r = 1,2,3)$ in terms of the queue length process $\mathbf{Z} = (Z_k, k = 1,2,\ldots,12)$ by

$$\mathbf{W}(t) = \mathbf{MZ}(t). \tag{6.1}$$

The Brownian model relates this workload process to the system idleness process $\mathbf{U} = (U_i, i = 1,2,\ldots,6)$ via

$$\mathbf{W}(t) = \mathbf{B}(t) + \mathbf{DU}(t), \tag{6.2}$$

where $\mathbf{B}$ is a three-dimensional Brownian motion with a certain drift vector and covariance matrix.

Observe that, in contrast with our previous examples, different queue lengths Z_k no longer contribute equally to the workloads W_r. Indeed the process W_1 can only be interpreted as the workload for stations 1, 3, 5 and 6 if we regard a service at stations 1 or 5 as being 2 units of work, and a service at stations 3 or 6 as being 1 unit of work. Both W_2 and W_3 have similar interpretations, the corresponding number of units of work being given by the second and third rows of $\mathbf{D}$. So the components of $\mathbf{W}$ can be interpreted as workloads for the combined resources $\{1,3,5,6\}$, $\{1,2,5,6\}$ and $\{2,3,4,6\}$. In fact, by symmetry, we can find three further workloads with similar properties to the W_r. However, the constraints on these additional workloads are linearly dependent on those given by equation (6.2) and hence we can ignore them.

Another important difference from our previous examples is that the process $\mathbf{W}$ can no longer lie anywhere in the positive orthant: it is constrained to lie in the cone $\mathcal{W}$ given by

$$\mathcal{W} = \{\mathbf{W} \in \mathbb{R}^3 : \mathbf{W} = \mathbf{MZ}, \mathbf{Z} \in \mathbb{R}_+^{12}\},$$
$$= \{\mathbf{W} : \mathbf{CW} \geqslant 0\}$$

where

$$\mathbf{C} = \begin{pmatrix} 2 & -1 & 0 \\ 1 & -2 & 3 \\ -1 & 2 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

Suppose that $\mathbf{W}(t)$ lies on the interior of the boundary face $W_2 = 0$ of $\mathcal{W}$. From the second row of $\mathbf{D}$, we will need to incur idleness at one (or more) of stations 1, 2, 5 or 6 just after time t to prevent $\mathbf{W}$ leaving $\mathcal{W}$ via the face $W_2 = 0$. However, increasing only U_2 minimizes the increase of $(\mathbf{DU})_1$ and W_1, whereas increasing only U_1 or U_5 (or both) minimizes the increase of $(\mathbf{DU})_3$ and W_3. After leaving the face $W_2 = 0$ there is positive probability of moving to a region where it would be optimal to increase $(\mathbf{DU})_1$ at time t, and there is also positive probability of moving to a region where it would be optimal to increase $(\mathbf{DU})_3$ at time t. With a non-anticipating control policy, there is positive probability of choosing to increase the wrong component of $\mathbf{DU}$ just after time t. So, in contrast with earlier results of this and previous sections, there is no choice of idleness which minimizes all components of $\mathbf{W}(t)$ for all $t \geqslant 0$ and, moreover, this implies (see Laws [32]) that it is not possible to minimize $\sum_k Z_k(t)$ for all $t \geqslant 0$ simultaneously.

7. General networks

In this section we consider dynamic routing and sequencing in general networks with asymmetric arrival and service rates, as in Laws [33]. Rather than starting with the most general case immediately, we consider an illustrative example.

THE 2 × 3 NETWORK

Consider the network of fig. 6. Suppose the arrival and service rate vectors of the (mutually independent) renewal arrival and general service processes are $\boldsymbol{\lambda} = (\lambda_j, j = 1, 2)$ and $\boldsymbol{\mu} = (\mu_i, i = 1, 2, \ldots, 6)$. Suppose also that the variances of type j interarrival times and station i service times are a_j and s_i respectively. Before using the Brownian approximation, consider the heavy traffic conditions under which we expect it to be accurate. There are a set of constraints, arising from deterministic network flow theory (Gondran and Minoux [12]), which determine the *capacity region* within which the system can cope with the arriving demands. The 29 constraints for this example include the following six constraints:

$$\lambda_1 \leqslant \mu_1 + \mu_4 , \tag{7.1}$$

$$\lambda_1 \leqslant \mu_1 + \mu_5 , \tag{7.2}$$

$$\lambda_2 \leqslant \mu_1 + \mu_2 + \mu_3 , \tag{7.3}$$

$$\lambda_1 + \lambda_2 \leqslant \mu_1 + \mu_2 + \mu_6 , \tag{7.4}$$

$$2\lambda_1 + \lambda_2 \leqslant \mu_1 + \mu_2 + 2\mu_6 , \tag{7.5}$$

$$3\lambda_1 + 2\lambda_2 \leqslant 2\mu_1 + 2\mu_5 + \mu_3 + \mu_6 . \tag{7.6}$$

The remaining 23 constraints can be obtained from those above by symmetry. The system is close to heavy traffic when all of the constraints are satisfied with strict inequality and one (or more) of them is close to equality.

Constraints (7.1)–(7.4) are *cut constraints*. Consider constraint (7.4), for example. Stations 1, 2 and 6 form a cut for type 1 and type 2 traffic: their removal from the system prevents the flow of customers of types 1 and 2. Hence the minimum arrival rate at this cut, $\lambda_1 + \lambda_2$, must not exceed the service rate of the cut, $\mu_1 + \mu_2 + \mu_6$, if the system is to be stable. Constraint (7.6) can be interpreted as follows. Customers are charged 2, 2, 1 and 1 units at stations 1, 5, 3 and 6 respectively; stations 2 and 4 have zero cost. To pay to get through the network, type 1 and 2 customers enter the system with 3 and 2 units respectively. If there is a feasible flow then the total arrival rate of revenue, $3\lambda_1 + 2\lambda_2$, cannot exceed the maximal charging rate, $2\mu_1 + 2\mu_5 + \mu_3 + \mu_6$. Constraint (7.5) has a similar interpretation with costs of 1, 1 and 2 units at stations 1, 2 and 6, and where type 1 and 2 customers arrive with 2 and 1 units respectively (route (3, 6) is available to type 2 customers, but is too expensive). We call the constraints (7.5) and (7.6), and the others obtained from them by symmetry, *generalized cut constraints*.

In the symmetric version of this example considered in section 6, all six constraints on $3\lambda_1 + 2\lambda_2$ are close to being tight. However, in many respects, this is not the typical situation. Consider a regime in which $\boldsymbol{\lambda}$ is scaled by a factor ρ, where ρ is increased from zero until one or more of the constraints first hit equality, at $\rho = \rho^*$ say. In many contexts it will be natural for exactly one of the constraints to become tight in heavy traffic, that is at $\rho = \rho^*$. If $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are chosen from some continuous probability distribution over $\mathbb{R}^8_+$, for example, then there will be a single tight constraint with probability 1. In a communication network where arrival rates vary over time we might expect to find at most a single constraint close to equality. However, if arrival and service rates are carefully matched (perhaps arrival rates themselves can be controlled) as may be the case if the network represented a manufacturing system, for example, then it is possible that several constraints will be close to equality (cf. Wein [44,45]).

Consider using a Brownian network model to approximate the system behaviour when only the generalized cut constraint (7.6) is close to equality. Suppose that $3\lambda_1 + 2\lambda_2 - (2\mu_1 + 2\mu_5 + \mu_3 + \mu_6)$ is of order $n^{-1/2}$: for example, $\boldsymbol{\lambda} = (2\rho, 4\rho)$, $\boldsymbol{\mu} = (2, 6, 3, 6, 2, 3)$ where $(1 - \rho)$ is of order $n^{-1/2}$. Define the scaled queue length and idleness processes $\mathbf{Z}$ and $\mathbf{U}$ as before. Since (7.6) is the constraint determining heavy traffic, it is impossible for stations 2 and 4 to be heavily loaded (this is clear for the $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ given above; in general see Laws [33]). Only heavily loaded stations appear in the Brownian model: in comparison with those at heavily loaded sta-

tions, queue lengths and delays at lightly loaded stations are small (cf. Harrison [16], Harrison and Williams [23]), and we take $Z_k = 0$ for $k = 2, 4, 6, 8$.

Let

$$M_k = \begin{cases} 3 & k = 1, 7, \\ 2 & k = 3, 5, 9, \\ 1 & k = 10, 11, 12, \\ 0 & k = 2, 4, 6, 8, \end{cases}$$

and let

$$W(t) = \sum_k M_k Z_k(t).$$

The interpretation of W as the system workload process is similar to that in section 6: we regard a service at stations 1 or 5 as being 2 units of work, and a service at stations 3 or 6 as 1 unit of work. (A minor difference is the factor μ^{-1} implicit, through $\mathbf{M}$, in the definition (6.1) of section 6. Now, in the asymmetric case, it is important for W to represent a weighted count of customers, rather than a weighted count of their expected service times.) Under the workload formulation of the Brownian network model, the choice of dynamic routing and sequencing policies for the system is equivalent to choosing the pair $(\mathbf{U}, \mathbf{Z})$ subject to the constraint

$$W(t) = B(t) + U(t), \tag{7.7}$$

where the scalar function U is the weighted idleness process

$$U(t) = 2\mu_1 U_1(t) + 2\mu_5 U_5(t) + \mu_3 U_3(t) + \mu_6 U_6(t)$$

and B is the Brownian motion with drift

$$n^{1/2}[3\lambda_1 + 2\lambda_2 - (2\mu_1 + 2\mu_5 + \mu_3 + \mu_6)]$$

and variance

$$9\lambda_1^3 a_1 + 4\lambda_2^3 a_2 + 4\mu_1^3 s_1 + 4\mu_5^3 s_5 + \mu_3^3 s_3 + \mu_6^3 s_6.$$

Notice the dependence of the Brownian model on the heavily loaded generalized cut constraint (7.6): the workload W, eq. (7.7) and the parameters of B are all obtained from the coefficients of the λ_j and μ_i in constraint (7.6). In this asymmetric case, the service rates μ_i are no longer used to define the workload W: instead they appear in eq. (7.7) via the weighted idleness process U. The reader will observe the very simple interpretation of $W(t)$ and $U(t)$ in terms of the charging argument behind constraint (7.6): the workload $W(t)$ is just the number of units held by customers in the system at time t, and $U(t)$ is just the cumulative revenue lost, through server idleness, over the period $[0, t]$.

In order to minimize both the weighted idleness and the system workload for all $t \geqslant 0$, choose $\mathbf{U}$ such that

$$U(t) = -\inf_{0 \leqslant s \leqslant t} B(s). \tag{7.8}$$

Further, we then minimize $\sum_k Z_k(t)$ for all $t \geqslant 0$ by choosing $\mathbf{Z}$ such that

$$\begin{aligned} Z_1(t) + Z_7(t) &= \tfrac{1}{3}b(t), \\ Z_k(t) &= 0 \qquad k \neq 1, 7, \end{aligned} \tag{7.9}$$

where

$$b(t) = B(t) - \inf_{0 \leqslant s \leqslant t} B(s).$$

The interpretation of the pathwise solution, given by (7.8) and (7.9), is as follows. Since the workload for the system, W, is constrained by (7.7) we minimize the number of customers in the system by maximizing the workload per customer: that is, $Z_k > 0$ only if $k = 1, 7$. Also, the solution shows that $U_i, i = 1, 3, 5, 6$, only increase when the system workload is zero. So, whenever one of the heavily loaded servers is working, all of them are working, and they are all idle only when there are no customers anywhere in the system. If the network is operated without dynamic routing then the Brownian model would have a separate constraint on the workload of each of servers 1, 3, 5 and 6 (Harrison [16]). So we again observe the important resource pooling effect of dynamic routing, which merges the individual workloads for servers 1, 3, 5 and 6 to form a single system workload process. While servers 2 and 4 have no effect and can be regarded as having infinite capacity, servers 1, 3, 5 and 6 behave as if they form a single pooled resource.

The benefit of this resource pooling can be calculated by considering the equilibrium behaviour of the system. Suppose arrival processes are Poisson and service times are independent and exponential so that $a_j = \lambda_j^{-2}, s_i = \mu_i^{-2}$. Consider a sequence of networks, indexed by n, in which $\boldsymbol{\lambda} = (2\rho(n), 4\rho(n)), \boldsymbol{\mu} = (2, 6, 3, 6, 2, 3)$ where

$$n^{1/2}|1 - \rho(n)| \to \theta < 0 \quad \text{as } n \to \infty. \tag{7.10}$$

Then in the heavy traffic limit under the pathwise solution, the normalized number of customers in the system is a reflected Brownian motion with drift $(14/3)\theta$ and variance $56/9$. From the stationary distribution of this reflected Brownian motion, the mean number in the system is $m^* = (2/3)|\theta|^{-1}$.

We compare the system population m^* with that for an optimal random routing policy. Since we are interested in heavy traffic we ignore stations 2 and 4, as in the Brownian model. Suppose all sequencing is FCFS and that arrivals choose their route by rolling a (biased) die. Then, subject to stability, the network has a product-form stationary distribution and under the optimal bias, which can be found by a Lagrangian analysis, the mean system population is

$$\frac{14 + 8\sqrt{3}}{7(1 - \rho(n))} - 4.$$

Using the same scaling as in the Brownian model, the mean system population in the heavy traffic limit is

$$\hat{m} = \lim_{n\to\infty} n^{-1/2}\left(\frac{14+8\sqrt{3}}{7(1-\rho(n))} - 4\right)$$
$$= \frac{2}{7}(7+4\sqrt{3})|\theta|^{-1}.$$

Hence the improvement of the pathwise solution over the optimal random routing policy is by a factor of a where

$$a = \frac{\hat{m}}{m^*} = \frac{3}{7}(7+4\sqrt{3}) \approx 5.97.$$

We now consider the behaviour of a particular dynamic routing policy. Let $n_i(t)$ be the total number of customers at station i at time t and suppose all sequencing is FCFS. Suppose a type 1 customer arriving at time t uses the route $r \in \{(1,2,3),(4,5,6)\}$ which minimizes $d_r(t)$ where

$$d_r(t) = \sum_{i\in r} \frac{n_i(t)}{\mu_i}.$$

Similarly suppose a type 2 customer arriving at time t uses the route $r \in \{(1,4),(2,5),(3,6)\}$ which minimizes $d_r(t)$. We regard this policy as a form of shortest expected delay routing (SDR) policy. Although, for a customer arriving at time t, the number of customers at station i may not be $n_i(t)$ when the customer reaches i, the *snapshot relation* (Foschini [8], Reiman [37,39]) suggests that $d_r(t)$ will be an accurate estimate in heavy traffic. The relation is in the form of a limit theorem for networks without dynamic routing and can be informally described as follows: when close to heavy traffic, the system queue length process is approximately unchanged during a customer's sojourn in the network. Hence it seems reasonable to estimate the delay on route r by $d_r(t)$ for a customer arriving at time t.

In heavy traffic queue lengths at stations 2 and 4 will be small compared with those at other stations, since these stations are not heavily loaded, so make the approximation

$$n_2(t) = n_4(t) = 0. \tag{7.11}$$

As a heavy traffic approximation, the SDR policy equalizes $d_r(t)$ over the routes available to type 1 customers, and over the routes available to type 2. That is, under SDR, station queue lengths satisfy

$$\frac{n_1(t)}{\mu_1} + \frac{n_3(t)}{\mu_3} = \frac{n_5(t)}{\mu_5} + \frac{n_6(t)}{\mu_6},$$
$$\frac{n_1(t)}{\mu_1} = \frac{n_5(t)}{\mu_5} = \frac{n_3(t)}{\mu_3} + \frac{n_6(t)}{\mu_6}. \tag{7.12}$$

Notice that eqs. (7.11) and (7.12) are equivalent to

$$(n_i(t)) \propto (2\mu_1, 0, \mu_3, 0, 2\mu_5, \mu_6)\,. \tag{7.13}$$

So, from relation (7.13), under SDR and FCFS sequencing, servers 1, 3, 5 and 6 are all busy whenever there is work anywhere in the system. Transferring these results to the Brownian model, eq. (7.8) is satisfied,

$$W(t) = b(t)\,, \tag{7.14}$$

and hence SDR minimizes $W(t)$ at all times $t \geqslant 0$. So SDR has the important property that it achieves the resource pooling effect of the pathwise solution. However, the policy does not succeed in minimizing the system population: relation (7.9) requires zero queue lengths at stations 3 and 6, and from eq. (7.13) this will not hold under SDR.

Consider again the case of Poisson arrivals and exponential service times with $\lambda = (2\rho(n), 4\rho(n))$, $\mu = (2, 6, 3, 6, 2, 3)$ where $\rho(n)$ satisfies (7.10). Let $N_i(t)$ be the (scaled) number of customers at station i in the Brownian model of SDR. Transferring eq. (7.13) to the Brownian model, we obtain

$$(N_i(t)) = \frac{\sum_i N_i(t)}{14}(4, 0, 3, 0, 4, 3)\,. \tag{7.15}$$

Heavy traffic analysis of systems without dynamic routing (Peterson [36], Reiman [40]) shows that, under FCFS sequencing, the length of queue k at station i is a fixed fraction of the total queue length at i, the fraction being the proportion of its busy time that server i devotes to queue k. Assume the same result holds for our network with dynamic routing. Then, in heavy traffic,

$$Z_1(t) = Z_3(t) = \tfrac{1}{2}N_1(t), \quad Z_5(t) = 2Z_{11}(t) = \tfrac{2}{3}N_3(t)\,,$$
$$Z_7(t) = Z_9(t) = \tfrac{1}{2}N_5(t), \quad Z_{10}(t) = 2Z_{12}(t) = \tfrac{2}{3}N_6(t)\,.$$

Hence under SDR and FCFS sequencing, using the relationships (7.14) and (7.15),

$$\sum_i N_i(t) = \tfrac{1}{2}b(t)\,.$$

So, pushing to a conclusion our various heavy traffic approximations, the scaled number in the system is a reflected Brownian motion with drift 7θ and variance 14. Hence the mean system population is $m' = |\theta|^{-1}$. Thus in heavy traffic, under SDR and FCFS sequencing, the anticipated improvement over the optimal random routing strategy is by a factor of

$$a' = \frac{\hat{m}}{m'} \approx 3.98\,.$$

While this improvement is smaller than that obtained under the pathwise solution, it is still substantial and, moreover, it is attained by a simple routing policy.

The use of dynamic sequencing with SDR allows further improvement, to an overall factor of 4.69: see Laws [33] for further discussion and numerical evidence.

THE GENERAL CASE

In general consider a network of I single-server stations. Customers of type j, $j = 1, 2, \ldots, J$, arrive according to a renewal process, interarrival times having mean λ_j^{-1} and variance a_j. Service times at station i, $i = 1, 2, \ldots, I$, form an independent identically distributed sequence of random variables with mean μ_i^{-1} and variance s_i. All arrival and service processes are independent. A route r is an ordered series of stations $(i_1, i_2, \ldots, i_n)$ and a customer using route r queues for service at each of these stations in turn before leaving the system. Let $\mathcal{R}_j$ be the set of routes available to type j customers and let $a_{ir} \in \mathbb{Z}_+$ be the number of times that route r visits station i; a_{ir} is 0 or 1 in all of our previous examples. As well as there being customers on different routes that queue at station i, there may also be customers queueing at i who are at different stages of the same route (since routes may visit i more than once). At station i distinguish between these different kinds of customers by having a different queue for each stage of each route that passes through i. The network is controlled via dynamic routing and sequencing. A routing policy specifies, at its arrival time, the route $r \in \mathcal{R}_j$ along which a type j customer is routed. A sequencing policy specifies which queue to serve at each station at each point in time.

As in the 2×3 example there are a set of generalized cut constraints determining the capacity region of the network. Each constraint is of the form

$$\sum_j \alpha_j \lambda_j \leqslant \sum_i \beta_i \mu_i \,, \tag{7.16}$$

where $\boldsymbol{\alpha} = (\alpha_j)$ and $\boldsymbol{\beta} = (\beta_i)$ are known vectors satisfying

$$\alpha_j = \min_{r \in \mathcal{R}_j} \left(\sum_i a_{ir} \beta_i \right).$$

As above, consider the simplest heavy traffic scenario: that is, assume that all constraints of the form (7.16) are satisfied and that exactly one is close to being a tight constraint. Suppose this constraint is given by the vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$ where $\alpha_j > 0$ for all j, $\beta_i > 0$ for all i and suppose further that, in heavy traffic, it is possible for all routes $r \in \mathcal{R}_j$ to be in use. (If these conditions on $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and $\mathcal{R}_j$ do not hold then we should reduce the original network to its bottleneck subnetwork, see Harrison [16], Laws [33], before proceeding with heavy traffic analysis.) With these assumptions

$$\alpha_j = \sum_i a_{ir} \beta_i \quad \text{for all } r \in \mathcal{R}_j \,.$$

We now describe the workload formulation of the Brownian model of this general network. Assume that $(\sum_j \alpha_j \lambda_j - \sum_i \beta_i \mu_i)$ is of order $n^{-1/2}$, and let $\mathbf{Z} = (Z_k)$

and $\mathbf{U} = (U_i)$ be the scaled queue length and idleness processes, as before. Let M_{ik} be the number of services (including the present one) that a customer currently in queue k requires from station i before leaving the network. Let

$$M_k = \sum_i M_{ik}\beta_i$$

and define the process W by

$$W(t) = \sum_k M_k Z_k(t)\,.$$

Since M_k is a weighted average of the remaining number of services of a customer in queue k, we can regard $M_k Z_k(t)$ as the total workload for the system in queue k at time t: a service at station i is regarded as being β_i units of work. Then $W(t)$ is the total work in the system at time t. Under the Brownian network model the choice of dynamic routing and sequencing policies is equivalent to choosing $(\mathbf{U}, \mathbf{Z})$ subject to

$$W(t) = B(t) + \sum_i \beta_i \mu_i U_i(t)\,, \tag{7.17}$$

where B is a Brownian motion with drift $\theta = n^{1/2}(\sum_j \alpha_j \lambda_j - \sum_i \beta_i \mu_i)$ and variance $(\sum_j \alpha_j^2 \lambda_j^3 a_j + \sum_i \beta_i^2 \mu_i^3 s_i)$. As in the 2×3 example above, observe how the Brownian model depends on the vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$ which define heavy traffic. Again the system workload and the total lost service effort are linked, by eq. (7.17), lost service effort at station i being weighted by a factor β_i.

Since the workload process is one-dimensional we can again find a pathwise solution. Let

$$\bar{M} = \max_k M_k$$

and define the set of customer classes $\mathcal{K} = \{k : M_k = \bar{M}\}$. To minimize cumulative idleness over $[0, t]$ and hence minimize $W(t)$, for all $t \geqslant 0$, choose $\mathbf{U}$ such that

$$\sum_i \beta_i \mu_i U_i(t) = -\inf_{0 \leqslant s \leqslant t} B(s)\,.$$

Further, we minimize the system population, $\sum_k Z_k(t)$, for all t by choosing $\mathbf{Z}$ such that

$$\sum_{k \in \mathcal{K}} Z_k(t) = \frac{1}{\bar{M}} b(t)\,,$$

$$Z_k(t) = 0 \quad k \notin \mathcal{K}\,,$$

where

$$b(t) = B(t) - \inf_{0 \leqslant s \leqslant t} B(s)\,.$$

The interpretation of this pathwise solution is again in terms of a single pooled resource. Under the solution the servers of the network are only idle when there are no customers anywhere in the system, and even when there is work in the system, the system population is minimized by allowing only queues $k \in \mathcal{K}$ to have positive queue lengths. Hence the system behaves as if servers $1, 2, \ldots, I$ are combined to form a single pooled resource.

Under the pathwise solution the system population is a reflected Brownian motion and, assuming $\theta < 0$, the mean scaled number in the system is

$$m^* = \frac{1}{2\bar{M}} \left(\sum_j \alpha_j^2 \lambda_j^3 a_j + \sum_i \beta_i^2 \mu_i^3 s_i \right) |\theta|^{-1} . \tag{7.18}$$

To make a comparison with the optimal random routing strategy, suppose arrivals are Poisson and services are independent and exponential. Suppose all sequencing is FCFS and that arrivals are routed according to the roll of an optimally biased die, and let $\hat{m}$ be the mean scaled system population. Then a Lagrangian analysis shows that in the heavy traffic limit the improvement of the pathwise solution over optimal random routing is by a factor of a where

$$a = \frac{\hat{m}}{m^*} \geqslant \frac{2\bar{M}(\sum_i \sqrt{\beta_i \mu_i})^2}{(\sum_j \alpha_j^2 \lambda_j + \sum_i \beta_i^2 \mu_i)} . \tag{7.19}$$

The bound (7.19) is certainly attained in the simple case considered in section 2: see expression (2.5).

Finally we consider the behaviour of shortest expected delay routing (SDR). Suppose sequencing is FCFS and let $n_i(t)$ be the total number of customers at station i at time t. Under SDR a customer of type j arriving at time t is routed via the route $r \in \mathcal{R}_j$ which minimizes

$$d_r(t) = \sum_i a_{ir} \frac{n_i(t)}{\mu_i} , \tag{7.20}$$

where $d_r(t)$ is an estimate of the delay via route r (supported by the snapshot relation as before). Now make the heavy traffic approximation that under SDR

$$d_r(t) = d_{r'}(t) \quad \text{for all } r, r' \in \mathcal{R}_j . \tag{7.21}$$

Equations (7.21) and (7.20) imply that

$$n_i(t) \propto \beta_i \mu_i ,$$

and so, that under SDR, all servers are kept busy whenever there are customers anywhere in the system. Hence the general network behaves as if all of its servers are combined to form a single pooled resource.

In the above discussion we have made several approximations which, while they appear plausible in heavy traffic, are not supported by existing convergence

results. A major challenge remains the identification of a framework which will allow a rigorous treatment of pathwise solutions for general networks.

Acknowledgements

We are grateful to Gerry Foschini, Mike Harrison, P.R. Kumar and Ruth Williams for their helpful comments on an earlier draft of this paper.

References

[1] I.J.B.F. Adan, J. Wessels and W.H.M. Zijm, Analysis of the symmetric shortest queue problem, Stochastic Models 6 (1990) 691–713.
[2] D.P. Bertsekas and R.G. Gallager, *Data Networks* (Prentice–Hall, Englewood Cliffs, 1987).
[3] J.G. Dai and J.M. Harrison, Reflected Brownian motion in an orthant: numerical methods for steady-state analysis, Ann. Appl. Prob. 2 (1992) 65–86.
[4] J.G. Dai and Y. Wang, Nonexistence of Brownian models for certain multiclass queueing networks, Queueing Systems 13 (1993), this issue.
[5] A. Ephremides, P. Varaiya and J. Walrand, A simple dynamic routing problem, IEEE Trans. Autom. Control AC-25 (1980) 690–693.
[6] L. Flatto and H.P. McKean, Two queues in parallel, Commun. Pure Appl. Math. 30 (1977) 255–263.
[7] G.J. Foschini, On heavy traffic diffusion analysis and dynamic routing in packet switched networks, in: *Computer Performance*, eds. K.M. Chandy and M. Reiser (North-Holland, Amsterdam, 1977) pp. 499–513.
[8] G.J. Foschini, Equilibria for diffusion models of pairs of communicating computers – symmetric case, IEEE Trans. Inf. Theory IT-28 (1982) 273–284.
[9] G.J. Foschini and J. Salz, A basic dynamic routing problem and diffusion, IEEE Trans. Commun. COM-26 (1978) 320–327.
[10] R.G. Gallager, A minimum delay routing algorithm using distributed computation, IEEE Trans. Commun. COM-25 (1977) 73–85.
[11] E. Gelenbe and G. Pujolle, *Introduction to Queueing Networks* (Wiley, Chichester, 1987).
[12] M. Gondran and M. Minoux, *Graphs and Algorithms* (Wiley–Interscience, New York, 1984).
[13] F.A. Haight, Two queues in parallel, Biometrika 45 (1958) 401–410.
[14] S. Halfin, The shortest queue problem, J. Appl. Prob. 22 (1985) 865–878.
[15] J.M. Harrison, *Brownian Motion and Stochastic Flow Systems* (Wiley, New York, 1985).
[16] J.M. Harrison, Brownian models of queueing networks with heterogeneous customer populations, in: *Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Vol. 10, eds. W. Fleming and P.-L. Lions (Springer, New York, 1988) pp. 147–186.
[17] J.M. Harrison and V. Nguyen, The QNET method for two-moment analysis of open queueing networks, Queueing Systems 6 (1990) 1–32.
[18] J.M. Harrison and V. Nguyen, Brownian models of multiclass queueing networks: current states and open problems, Queueing Systems, 13 (1993), this issue.
[19] J.M. Harrison and M.I. Reiman, On the distribution of multidimensional reflected Brownian motion, SIAM J. Appl. Math. 41 (1981) 345–361.
[20] J.M. Harrison and M.I. Reiman, Reflected Brownian motion on an orthant, Ann. Prob. 9 (1981) 302–308.

[21] J.M. Harrison and L.M. Wein, Scheduling networks of queues: heavy traffic analysis of a simple open network, Queueing Systems 5 (1989) 265–280.
[22] J.M. Harrison and L.M. Wein, Scheduling networks of queues: heavy traffic analysis of two-station closed network, Oper. Res. 38 (1990) 1052–1064.
[23] J.M. Harrison and R.J. Williams, Brownian models of open queueing networks with homogeneous customer populations, Stochastics 22 (1987) 77–115.
[24] D.J. Houck, Comparison of policies for routing customers to parallel queueing systems, Oper. Res. 35 (1987) 306–310.
[25] F.P. Kelly, The optimization of queueing and loss networks, in: *Queueing Theory and its Applications*, eds. O.J. Boxma and R. Syski (North- Holland, Amsterdam, 1988) pp. 375–392.
[26] F.P. Kelly, On a class of approximations for closed queueing networks, Queueing Systems 4 (1989) 69–76.
[27] J.F.C. Kingman, Two similar queues in parallel, Ann. Math. Statist. 32 (1961) 1314–1323.
[28] L. Kleinrock, *Queueing Systems*, Vol. 2: *Computer Applications* (Wiley, New York, 1976).
[29] H.J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations* (Academic Press, New York, 1977).
[30] H.J. Kushner and L.F. Martins, Numerical methods for stochastic singular control problems, SIAM J. Control Optim. 29 (1991) 1443–1475.
[31] H.J. Kushner and K.M. Ramachandran, Optimal and approximately optimal control policies for queues in heavy traffic, SIAM J. Control Optim. 27 (1989) 1293–1318.
[32] C.N. Laws, Dynamic routing in queueing networks, PhD thesis, Statistical Laboratory, University of Cambridge (1990).
[33] C.N. Laws, Resource pooling in queueing networks with dynamic routing, Adv. Appl. Prob. 24 (1992) 699–726.
[34] C.N. Laws and G.M. Louth, Dynamic scheduling of a four-station queueing network, Prob. Eng. Inf. Sci. 4 (1990) 131–156.
[35] L.F. Martins and H.J. Kushner, Routing and singular control for queueing networks in heavy traffic, SIAM J. Control Optim. 28 (1990) 1209–1233.
[36] W.P. Peterson, A heavy traffic limit theorem for networks of queues with multiple customer types, Math. Oper. Res. 16 (1991) 90–118.
[37] M.I. Reiman, The heavy traffic diffusion approximation for sojourn times in Jackson networks, in: *Applied Probability – Computer Science: The Interface*, Vol. 2. eds. R.L. Disney and T.J. Ott (Birkhäuser, Boston, 1982) pp. 409–422.
[38] M.I. Reiman, Some diffusion approximations with state space collapse, in: *Proc. Int. Seminar on Modelling and Performance Evaluation Methodology*, Lecture Notes in Control and Informational Sciences 60, eds. F. Baccelli and G. Fayolle (Springer, Berlin, 1983) pp. 209–240.
[39] M.I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (1984) 441–458.
[40] M.I. Reiman, A multiclass feedback queue in heavy traffic, Adv. Appl. Prob. 20 (1988) 179–207.
[41] M. Schwartz, *Telecommunication Networks* (Addison–Wesley, Reading, MA, 1987).
[42] S.R.S. Varadhan and R.J. Williams, Brownian motion in a wedge with oblique reflection, Commun. Pure Appl. Math. 38 (1985) 405–443.
[43] R.R. Weber, On the optimal assignment of customers to parallel servers, J. Appl. Prob. 15 (1978) 406–413.
[44] L.M. Wein, Optimal control of a two-station Brownian network, Math. Oper. Res. 15 (1990) 215–242.
[45] L.M. Wein, Scheduling networks of queues: heavy traffic analysis of a two-station network with controllable inputs, Oper. Res. 38 (1990) 1065–1078.
[46] L.M. Wein, Brownian networks with discretionary routing, Oper. Res. 39 (1991) 322–340.
[47] L.M. Wein, Scheduling networks of queues: heavy traffic analysis of a multistation network with controllable inputs, Sloan School of Management, MIT, Boston, MA.

[48] W. Whitt, Open and closed models for networks of queues, AT&T Bell Lab. Techn. J. 63 (1984) 1911–1979.
[49] W. Whitt, Deciding which queue to join: some counterexamples, Oper. Res. 34 (1986) 55–62.
[50] R.J. Williams, Recurrence classification and invariant measure for reflected Brownian motion in a wedge, Ann. Prob. 13 (1985) 758–778.
[51] R.J. Williams, Reflected Brownian motion in a wedge: semimartingale property, Z. Wahrscheinlichkeitstheorie verw. Gebiete 69 (1985) 161–176.
[52] W. Winston, Optimality of the shortest line discipline, J. Appl. Prob. 14 (1977) 181–189.
[53] P. Yang, Pathwise solutions for a class of linear stochastic systems, PhD thesis, Dept. of Operations Research, Stanford University (1988).