DYNAMIC ROUTING IN STOCHASTIC NETWORKS

F.P. KELLY*

Abstract. This paper reviews some current work on routing in loss and queueing networks. We describe two classes of bound on the performance of any dynamic routing scheme, together with some open questions concerning whether the bounds can be approached under certain limiting regimes. The first class of bound is particularly appropriate for networks in heavy traffic, where the key feature of a good routing scheme is its effective utilization of various pooled resources identified by a fluid version of the routing problem. The second class of bound is particularly appropriate for highly connected networks, with many alternative paths. Again, a network flow representation is central, but this time involving a collection of Markov decision processes, one for each resource of the network. Despite their simplicity, the bounds are able to identify the great variety of qualitatively distinct behaviour expected of a good dynamic routing scheme, depending on a network's size, connectivity, asymmetry and degree of overload.

1. Introduction. Modern telecommunications networks provide a variety of fascinating challenges for the mathematician interested in applications of probability and optimization. Dynamic routing schemes, for example, allow a network to respond to fluctuating demands and failures by rerouting traffic and reallocating resources. While the detailed design of a dynamic routing scheme is much influenced by the rapidly evolving technological environment, some important principles are apparent from an analysis of loss probabilities and queueing delays in simple network models.

In this paper we describe some current work on routing in loss and queueing networks. One of the promising approaches of the last few years has been the development of upper bounds on the performance of *any* dynamic routing scheme for a stochastic network ([1],[2],[19],[25]). We describe two classes of bound, each with an associated limiting regime under which it is conjectured that the bound is tight. Further open questions concern whether the bounds can be approached by dynamic routing schemes that are simple in some sense, perhaps depending only on local rules and thresholds. Despite their simplicity and potential coarseness, the bounds are able to identify the great variety of qualitatively distinct behaviour expected of a good dynamic routing scheme, depending on a network's size, connectivity, asymmetry and degree of overload.

In a queueing network an arriving demand makes use of different resources sequentially in time, and congestion causes delay. In a loss network an arriving demand requests simultaneous use of different resources, and congestion causes blocking. Despite these differences there is a remarkable similarity between the methods used to analyse loss and queueing networks:

* Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England. Part of this paper was written whilst the author was visiting the Institute for Mathematics and its Applications, whose financial support is gratefully acknowledged.

see, for example, the discussion in [9] of product-form solutions, approximation procedures and optimization issues, and the study in [7] of least busy alternative routing. A subsidiary aim of this paper is to further indicate the potential for cross-fertilization of ideas between studies of the two forms of network.

In Section 2 we describe the multicommodity network flow results that apply to the deterministic, or fluid, version of the network routing problem. The flows that are achievable in the fluid version are defined by a set of generalized cut constraints, written in terms of the load upon, and the capacity of, various pooled resources. These constraints provide the starting point for the heavy traffic analysis of stochastic networks. In a queueing network heavy traffic is defined by a limiting regime where the load upon pooled resources approaches their capacity. In a loss network heavy traffic is defined within a limiting regime where the load upon, and capacity of, pooled resources increase together, with ratios held fixed. It is possible to bound the performance of a stochastic network in terms of the performance of its pooled resources, but there remain several open questions concerning whether these bounds can be approached and, if so, by what form of dynamic routing scheme. The analysis does, however, make clear that in heavy traffic the key determinant of network performance is the behaviour of the pooled resources identified by the fluid version of the routing problem.

It is also clear that, under a good routing scheme, the proportion of alternatively routed traffic within the network, routed along other than the first choice routes identified by the fluid version, should vanish in the heavy traffic limit. Rather different behaviour should be expected in highly connected networks with relatively small traffic and capacity at each resource. In Section 3 we outline a class of bound particularly appropriate for such networks, and discuss a limiting regime under which the proportion of alternatively routed traffic within the network does not vanish. Again a network flow representation is central, but this time involving a collection of Markov decision processes, one for each resource of the network. Once again, there remain several open questions concerning whether the bounds can be approached and, if so, by what form of dynamic routing scheme.

Finally in Section 4 we describe how, for any given finite network, the two classes of bound may be combined. In future numerical work with combined bounds it is hoped to get a clearer view of how the structural features of a given network affect the relative importance and accuracy of the various forms of bound.

2. Deterministic bounds, and heavy traffic. We begin by outlining the basic notation we use to describe loss and queueing networks.

2.1. Loss and queueing networks. The classical example of a loss network is a telephone network, and we shall phrase our description in terms of calls, circuits and routes. However the reader will observe that our model applies more widely to systems in which, before a request (which may be a call, or a task, or a customer) is accepted, it is first checked that sufficient resources are available to deal with the request.

Consider then a network with links labelled $i \in I$, for some finite set I, and suppose that link i comprises C_i circuits. A call on route $r \in R$ uses A_{ir} circuits from link i, where $A_{ir} \geq 0$. Thus if n_r calls are in progress on route r then

$$An \leq C \tag{2.1}$$

where $A = (A_{ir}, i \in I, r \in R)$, $n = (n_r, r \in R)$ and $C = (C_i, i \in I)$.

Suppose that calls of type $j \in J$ arrive at the network at rate ν_j, where J is the set of call types. Set $D_{jr} = 1$ if a call of type j can be carried on route r, and set $D_{jr} = 0$ otherwise. This defines a $0-1$ matrix $D = (D_{jr}, j \in J, r \in R)$. When a call of type j arrives at the network, a dynamic routing scheme must decide which route in the set $R_j = \{r : D_{jr} = 1\}$ should be used to carry the call, or whether the call should be rejected. The call *must* be rejected if acceptance would lead to vector n violating the constraint (2.1). Once a call is accepted it has a holding period which, for definiteness, we shall assume has unit mean and is independent of earlier arrival times and holding periods.

We describe a queueing network in very similar terms. Suppose the network has stations labelled $i \in I$, for some finite set I, and that a customer on route $r = (r(1), r(2), \ldots, r(M_r))$ passes through the sequence of queues

$$r(1), r(2), \ldots, r(M_r)$$

before leaving the system. Thus the queue which a customer on route r visits at stage m $(= 1, 2, \ldots, M_r)$ of its route is queue $r(m)$, and M_r is the number of stages on route r. Let station i have a single server, capable of dispensing service effort at rate C_i, and, for definiteness, assume that all service requirements, even of the same customer at different stages of its route, are independent of each other and of earlier arrival times, and have unit mean. Let

$$A_{ir} = |\{m : r(m) = i\}|,$$

the number of times a customer on route r visits station i. Suppose that customers of type $j \in J$ arrive at the network at rate ν_j. As before set $D_{jr} = 1$ if a customer of type j can be carried on route r, and set $D_{jr} = 0$ otherwise. Since resources are used sequentially in time it may not be necessary for a dynamic routing scheme to choose the entire route r of a customer when the customer enters the network – it may, for example, be possible to select a first stage for the route, but leave open the choice of second stage until the first stage is completed. In a queueing network it is also necessary to make sequencing decisions at stations: a station must decide the order in which to serve the various customers on different routes that may be queueing there. We view sequencing choices as part of the overall dynamic routing scheme.

2.2. Deterministic bounds. Before developing bounds for a stochastic network it is useful to consider a deterministic or fluid version of the routing problem, using standard multicommodity network flow results.

Let

$$S = \{\nu \geq 0 : \exists\, x \geq 0 \text{ with } Dx = \nu, Ax \leq C\}.$$

We can view S as the achievable region for the fluid analogue: for any $\nu \in S$ there exists a deterministic flow pattern x which will satisfy the demand pattern ν (since $Dx = \nu$), but will not overload any resource (since $Ax \leq C$). Let

$$T = \{x \geq 0 : Ax \leq C\}.$$

Thus T is the intersection of the half-spaces $(x : (Ax)_i \leq C_i), i \in I$, and the non-negative orthant. With no loss of generality assume no column of A is null: hence the set T is bounded, and is the convex hull of a finite number of extreme points. Hence $S = DT$ is the convex hull of a finite number of extreme points or, equivalently, the bounded intersection of a finite set of half-spaces. Thus there exists a representation

$$S = \{\nu \geq 0 : \overline{A}\nu \leq \overline{C}\} \tag{2.2}$$

for some choice of $\overline{A}, \overline{C}$. Moreover $\overline{A}, \overline{C}$ can be chosen to have non-negative element (since $0 \in S$; and $0 \leq \nu' \leq \nu \in S$ implies $\nu' \in S$).

This result, that the achievable region (2.2) for the fluid analogue is subject to the same form of linear constraint (2.1) as the original loss network, is a well known result [5] . The usual derivation is less direct, but achieves a valuable insight from its use of dual variables. In [5] it is shown that $\nu \in S$ if and only if for all $\beta \geq 0$

$$\sum_j \alpha_j \nu_j \leq \sum_i \beta_i C_i \tag{2.3}$$

where

$$\alpha_j = \min_{r : D_{jr} = 1} \left\{ \sum_i \beta_i A_{ir} \right\} \quad j \in J. \tag{2.4}$$

Here the dual variable β_i may be viewed as a price charged for the use of one unit of capacity at resource i: if there exists a vector $\beta \geq 0$ such that (2.3) is violated then clearly the demand pattern ν cannot be achieved, since with these prices the maximum revenue from resources is exceeded by the amount demands must pay to cross the network. Laws [17] notes that (2.3) and (2.4) will follow for all $\beta \geq 0$ if they hold for a certain finite collection of vectors β: thus $\nu \in S$ if and only if

$$\sum_j \alpha_{kj}\nu_j \leq \sum_i \beta_{ki} C_i \qquad k \in K \tag{2.5}$$

for some collection $\{\{\alpha_{kj}, \beta_{ki}, i \in I, j \in J\}, k \in K\}$, where K is finite. The linear constraints (2.5) defining S are termed *generalized cut constraints*. Each such constraint describes the load on, and the capacity of, a single *pooled resource*.

2.3. Heavy traffic. What can we deduce about the behaviour of a stochastic network from the above fluid version? Consider first a loss network, and suppose that the arrival rates ν are such that at least one of the constraints (2.5) is violated: let us say that the constraint labelled $k = 0$ is violated. Then the performance of the loss network is clearly dominated by the performance of a single link, of capacity $\sum_i \beta_{oi} C_i$ circuits, at which calls of type j arrive, as in the network, at rate ν_j, each such call requesting α_{oj} circuits. More generally, if $(n_j, j \in J)$ describes the number of calls of type j in progress then necessarily

$$\sum_j \alpha_{kj} n_j \leq \sum_i \beta_{ki} C_i \qquad k \in K.$$

Thus the performance of a loss network under *any* dynamic routing scheme is dominated by the performance of a loss network, with no routing choices, where calls of type j arrive at rate ν_j, and require α_{kj} circuits from link k, a link of capacity $\sum_i \beta_{ki} C_i$, for $k \in K$.

When might the bounds[1] implied by the above comparison be good bounds, that can be approximately achieved? Consider a sequence of networks where $(\nu_j(N), j \in J)$, $(C_i(N), i \in I)$ label the arrival rates and capacities, respectively, in the N^{th} network, and suppose that as $N \to \infty$

$$\frac{\nu_j(N)}{N} \to \nu_j, \ \frac{C_i(N)}{N} \to C_i. \tag{2.6}$$

Then it is possible to show (under rather weak conditions on arrival processes: see [13],[14] for the case of Poisson arrivals) that the set of loss probabilities $(L_j(N), j \in J)$ achievable in network N approaches the set of loss probabilities achievable in the fluid version. Interest then focuses on whether simple decentralized schemes, using threshold rules like trunk reservation, are able to perform close to optimally (see [14],[18]: for networks with a small number of links it is possible to compute the optimal control [13], [16], but this seems impractical for a large network such as that

[1] When calls may be repacked the bound becomes a characterization: see [10] for a discussion of this interesting case, which here we consider no further.

studied in [3]). To give an example, suppose that arrivals of different types of call form independent Poisson processes, that the limit vectors ν, C from (2.6) satisfy with strict inequality all but one ($k = 0$) of the generalized cut constraints (2.5), and that

$$\sum_j \alpha_{oj}\nu_j(N) - \sum_i \beta_{oi}C_i(N) = \gamma N^{\frac{1}{2}} \tag{2.7}$$

for some constant γ. Then for the bounding single link it seems plausible (cf. [18],[27]) that, by rejecting traffic of type $j = 1$ if the pooled resource is closer to saturation than, say, $\log N$ circuits, it is possible to achieve loss probabilities satisfying

$$L_j(N) = o(N^{\frac{1}{2}}) \qquad j \neq 1,$$

$$N^{\frac{1}{2}}L_1(N) \to \frac{\left(\sum_j \alpha_{oj}^2\nu_j\right)^{\frac{1}{2}}}{\alpha_{01}\nu_1}\,\frac{\phi(\delta)}{1-\Phi(\delta)}$$

where ϕ, Φ are respectively the normal density and distribution function, and

$$\delta = \gamma\left(\sum_j \alpha_{oj}^2\nu_j\right)^{-\frac{1}{2}}.$$

(Note that if $\alpha_{oj} = 1, j \in J$, $\beta_{oi} = 1, i \in I$, and $\sum_j \nu_j = 1$, then $\delta = \gamma$ and these expressions become those established for a single link by Reiman [27]. For more general parameter choices the constant γ must be renormalized, and some insight into the expression for δ is obtained by observing that if n_j is a Poisson random variable with mean ν_j, then the variance of $\sum_j \alpha_{oj}n_j$ is $\sum_j \alpha_{oj}^2\nu_j$.) It is a challenge to determine if a dynamic routing scheme for the loss network could approach the above putative bound, under the limiting regime (2.6) and (2.7).

Consider next a queueing network, and suppose the arrival rates ν are such that all of the constraints (2.5) are satisfied. Nevertheless choose one, say the constraint labelled $k = 0$. Then the performance of the queueing network is dominated by the performance of a single station constructed as follows. Let the single station have multiple servers, a server i replicating the server at station i for each station i with $\beta_{oi} > 0$. Let customers of type j arrive at this station in a Poisson stream of rate ν_j. Any customer may be served by any server, but a single service completion at server i corresponds to the accrual of β_{oi} credits by the customer, and a customer of type j cannot leave until it has accrued at least α_{oj} credits. (Note that, by relation (2.4), it is certainly possible for a customer of type j to acquire exactly α_{oj} credits.)

Again it is reasonable to consider when such bounds may be approximately achieved. Consider a sequence of networks, where $(\nu_j(N), j \in J)$ label the arrival rates in the N^{th} network, and where as $N \to \infty$

$$\nu_j(N) \to \nu_j. \tag{2.8}$$

Suppose that all of the constraints (2.5) are satisfied with strict inequality, save one, labelled $k = 0$, for which

$$\sum_j \alpha_{oj}\nu_j(N) = \sum_i \beta_{oi}C_i - \theta N^{-\frac{1}{2}} \tag{2.9}$$

where $\theta > 0$. Further suppose that arrivals of type j customers form a renewal process, where interarrival times have mean $\nu_j(N)^{-1}$ and variance a_j, and service times at station i have mean C_i^{-1} and variance s_i. Then for the bounding single station it seems plausible ([26],[28],[18]) that, by giving priority to all customers except those for which α_{oj} is maximal ($= \alpha$, say), it is possible to achieve a mean number of customers in the queue of order

$$\frac{N^{\frac{1}{2}}}{2\alpha\theta}\left(\sum_j \alpha_{oj}^2\nu_j^3 a_j + \sum_i \beta_{oi}^2 C_i^3 s_i\right)$$

Again it is a challenge to determine if a dynamic routing scheme for the queueing network could approach this putative bound, under the limiting regime (2.8) and (2.9).

Laws and Louth [19] analyse a queueing network where *two* of the generalized cut constraints (2.5) are close to saturation: they show that a bound can be constructed using the pathwise minimum of two queues, each representing a pooled resource, and provide numerical evidence that this bound is approached under the limiting regime. See [12] for further discussion of a variety of examples, emphasizing the use of the Brownian network models of Harrison [6].

3. Diverse routing. In this Section we describe a further class of bound, more appropriate to the case of a highly connected network where the network contains many alternative routes and a typical resource lies, potentially, on a large number of alternative routes.

3.1. Loss networks. Let I label the set of possible demands on a network, and also the set of network resources. Assume that demands of type i arrive at the network in a Poisson stream of rate ν_i, for $i \in I$, and that as i varies it indexes independent Poisson streams. Let C_i be the *capacity* of resource i. Again, we shall refer to a resource as a *link*, interpret C_i as the number of *circuits* on link i, and refer to a demand labelled i as a *call* of type i. Suppose that an arriving call of type i may potentially be routed directly, on link i, or alternatively, on a route $r \in R(i)$. Here a route r identifies a subset of I, and $R(i)$ is the set of alternative routes potentially available to a call of type i. Label routes so that the sets $R(i), i \in I$, are disjoint: let $R = \bigcup_{i \in I} R(i)$, and let $i(r)$ be the unique element in I such that $r \in R(i(r))$. If an arriving call of type i is sent to route $r \in \{\{i\}\} \cup R(i)$ then it uses one circuit from each link $j \in r$ for the holding period of the call. A call may only be sent to a route r with at least one free circuit on each link $j \in r$. The call may also be discarded, and it *must* be discarded if there are no routes with spare capacity available. The holding period of a call is arbitrarily distributed with unit mean, is independent of earlier arrival times and holding periods, and is unaffected by the route used to carry the call.

For example, we could model a fully connected telephone network on n nodes by letting I be the set of $\frac{1}{2}n(n-1)$ unordered pairs of nodes. A call labelled $i \in I$ desires to connect the two nodes identified by i, and C_i is the number of circuits on the direct link between these two nodes. Note that we can interpret I as the set of edges of the complete graph on n nodes, and then a route connecting the node pair i is just a path in the graph between these two nodes. We might let $R(i)$ be the set of $(n-2)$ two-link paths connecting the node pair i. Or $R(i)$ might be the entire set of paths of length two or more connecting the node pair i, or any given subset of this. If the network is not fully connected, then we just set $C_i = 0$ for the missing links.

A dynamic routing scheme is a control policy which decides for each arriving call whether to discard the call, or, if not, by which route to carry the call. We allow the control policy to have full information on the history of call arrivals, routing decisions and expired holding times, but do not allow it to anticipate future events, and in particular do not allow it to have any information on the future holding time of an arriving call. Suppose that an accepted call of type i is worth w_i, and that our objective is to maximize the expected average reward per unit time. If holding times are exponentially distributed then we may formulate the control problem as a finite state Markov decision process, where the state is a complete list of the routes used by all calls currently in progress (although it may be difficult to make computational progress with this formulation: see [22]). With arbitrarily distributed holding times a Markov description requires additional information on the elapsed holding times of the calls currently in progress. Throughout we shall restrict attention to policies under which the number of calls of type i accepted over the period $[0, t]$ converges, in expectation and almost surely, as $t \to \infty$. We lose no essential generality by this restriction, since the event that the system is empty has bounded mean recurrence time under any policy, by the assumption of Poisson arrivals.

Our aim is to obtain an upper bound on the expected reward per unit time. Before describing the bound we shall find it helpful to analyze a much simpler dynamic optimization problem. Consider a single link, of capacity C circuits, offered two streams of traffic. Suppose that acceptance of a type 1 call generates a reward w_1, and acceptance of a type 2 call generates a reward w_2, where $w_1, w_2 \geq 0$. Suppose that the two streams of arriving traffic form independent Poisson processes of rates ν_1 and ν_2 respectively, and that accepted calls have holding times which are arbitrarily distributed with unit mean, independently of earlier arrival and holding times. When a call arrives a decision is made to accept or reject the call, and the decision can depend on the type of the call and the entire history of the link.

Let $W(\nu_1, \nu_2, C; w_1, w_2)$ be the maximal expected reward per unit time over all policies. Observe that W is a convex function of (w_1, w_2), since it can be expressed as a supremum of linear functions of (w_1, w_2),

$$W(\nu_1, \nu_2, C; w_1, w_2) = \sup_{\pi}\{w_1 x(\pi) + w_2 y(\pi)\},$$

where $x(\pi)$ and $y(\pi)$ are the mean acceptance rate of calls of types 1 and 2 respectively under a policy π. Let

$$M(\nu_1, \nu_2, C; x) = \sup_{\pi}\{y(\pi) : x(\pi) \geq x\},$$

the maximal mean acceptance rate of calls of type 2, subject to the requirement that the mean acceptance rate of calls of type 1 must be at least x. M is a concave function of x – indeed W and M are conjugates:

$$\begin{aligned} W(\nu_1, \nu_2, C; w_1, w_2) &= \sup_{x \geq 0}\{w_1 x + w_2 M(\nu_1, \nu_2, C; x)\}, \\ M(\nu_1, \nu_2, C; x) &= \inf_{w \geq 0}\{W(\nu_1, \nu_2, C; w, 1) - wx\}. \end{aligned}$$

If holding times are exponentially distributed with unit mean, then it is sufficient to consider policies which summarize the entire history of the link by a single integer, the number of calls currently in progress through the link. By a monotonicity argument [21] the optimal policy for the resulting Markov decision process is (if $w_1 \geq w_2 \geq 0$) to accept type 1 calls provided the link is not full, and to accept type 2 calls provided the number of spare circuits is above a certain integer threshold. Policies of this form are termed *trunk reservation* policies. If $w_2 > w_1 \geq 0$ then the optimal policy is to use trunk reservation against type 1 calls. To achieve the maximal flow M at arbitrary values of x we must extend the definition of a trunk reservation policy to allow randomized acceptance when the number of spare circuits is exactly equal to the threshold. If holding times are not exponentially distributed then the optimal policy may not be a trunk reservation policy, but such policies will generally be near optimal, especially for larger values of the capacity C.

Return now to the network model, and let

$$M_i(x) = M(\nu_i, \sum_{j : i \in R(j)} \nu_j, C_i; x),$$

$$W_i(w_1, w_2) = W(\nu_i, \sum_{j : i \in R(j)} \nu_j, C_i; w_1, w_2).$$

Then in [11] (see also [8],[15],[24] for earlier, related, work) it is shown that under any policy the expected reward per unit time is bounded above by the value attained in the maximum flow problem

$$\text{(3.1a)} \qquad \text{maximize} \qquad \sum_i w_i (x_i + \sum_{r \in R(i)} y_r)$$

$$\text{(3.1b)} \qquad \text{subject to} \qquad x_i + \sum_{r \in R(i)} y_r \leq \nu_i \qquad i \in I$$

$$\text{(3.1c)} \qquad \text{and} \qquad \sum_{r: i \in r} y_r \leq M_i(x_i) \qquad i \in I$$

$$\text{(3.1d)} \qquad \text{over} \qquad x_i \geq 0, i \in I, \quad \text{and} \quad y_r \geq 0, r \in R\ .$$

or, equivalently, the value attained in the dual minimum cost problem

$$\text{(3.2a)} \qquad \text{minimize} \qquad \sum_i [W_i(w_i - s_i, c_i) + s_i \nu_i]$$

$$\text{(3.2b)} \qquad \text{subject to} \qquad w_{i(r)} - \sum_{j \in r} c_j \leq s_{i(r)} \qquad r \in R$$

$$\text{(3.2c)} \qquad \text{over} \qquad c_i, s_i \geq 0 \qquad i \in I.$$

The primal problem is perhaps the easier to interpret, with x_i, y_r representing mean flows along routes $\{i\}, r$, respectively: thus inequality (3.1c) captures the insight that link i cannot achieve higher acceptance rates as part of a network than it could if the rest of the network were transparent. A sketch of the proof via the dual problem is instructive. Consider a feasible choice of c, s; that is, a collection $c_i, s_i, i \in I$, satisfying (3.2b) and (3.2c). Suppose that a call of type i is charged an amount s_i by the network when it is offered to the network, and in addition is charged an amount $w_i - s_i$ by link i if it is carried directly on link i, or an amount c_j at each link $j \in r$ if it is carried on alternative route r. The maximum revenue that can possibly be collected thus cannot exceed the objective function (3.2a), for any feasible choice of s, c. Since s, c satisfy (3.2b) each routed call of type i pays at least w_i, whether the call is routed directly or alternatively, and the result follows.

When holding times are exponentially distributed, the function $M_i(x_i)$ in piecewise linear in x_i, for $i \in I$. In this case both the primal and dual problems (3.1) and (3.2) may be written as linear programming problems, a fact exploited in [4].

It is interesting to interpret the complementary slackness conditions that interlink the solutions to problems (3.1) and (3.2) in terms of the charges used in the proof sketch. The main condition shows that if the charges along route r are too high (i.e. $\sum_{j \in r} c_j > w_{i(r)} - s_{i(r)}$) then the route is not used (i.e. $y_r = 0$). The interpretation of the remaining complementary slackness conditions is more familiar: the charge c_i at link i is zero if there is spare capacity at this link, as indicated by slackness in the constraint (3.1c); while the charge s_i on an offered call of type i is zero if calls of type i are being lost, as indicated by slackness in the constraint

(3.1b). We might term s_i the *surplus value* of an additional call of type i, and c_j the *implied cost* of using a circuit from link j: if a call of type i is accepted on route r it will earn w_i directly but at an implied cost of c_j for each circuit used from link j, leaving a surplus value of $s_i = w_i - \sum_{j \in r} c_j$ (cf. [9]).

The above discussion is very suggestive of routing schemes where each link operates a trunk reservation policy and calls are routed by the cheapest route provided that route costs less than the worth of the call. For certain forms of symmetric network, for example a fully connected network on n nodes, it is known that such schemes achieve a reward per link that approaches the bound as the number of nodes increases ([23], [8]). This phenomenon has also been observed in simulations of asymmetric networks [3]. It seems reasonable to expect the bounds to be approached by trunk reservation schemes whenever calls have many alternative routes to choose from, and each link lies, potentially, on a large number of alternative routes. It seems difficult, however, to prove this under appropriately general conditions (some general results are available when calls may be repacked: see [11], Section 3).

Next we show that a parallel development is able to provide bounds on the performance of arbitrary dynamic routing schemes for queueing networks.

3.2. Queueing networks. Again let I label the set of possible demands on the network as well as the collection of network resources. Assume that arrivals at the network of demands labelled i form a stationary stochastic process of rate ν_i. Assume now that resources are queues, and that demands are customers. Suppose that a customer of type i may be routed directly, via queue i, or alternatively on a route $r \in R(i)$. A route r now identifies an ordered subset of I, and $R(i)$ is the set of alternative routes potentially available to a call of type i. If a customer is sent via route r then it must be served in succession at each queue $j \in r$. Assume that the service requirement of a customer arriving at the network is arbitrarily distributed with unit mean, and that a customer's service requirement is the same at each queue that it visits. Let queue i serve at rate C_i, so that the service time of a customer at queue i is that customer's service requirement divided by C_i.

A dynamic routing scheme is a control policy which decides the route of each arriving customer, and decides for each queue who the server should deal with next (pre-emption may be either allowed or disallowed). We suppose that it costs w_i per unit of time that a customer of type i remains in the system, and our objective is to minimize the expected cost per unit time. Again our aim is to bound the performance of any policy, and again we begin by considering a much simpler dynamic optimization problem.

Consider first a single queue at which customer arrivals form a stationary stochastic process of rate ν. On arrival a customer may be accepted or rejected. Accepted customers queue, and have service times that are arbitrarily distributed with mean C^{-1}. (For notational convenience, and with no loss of essential generality, we shall assume that ν fixes the distribution of the arrival process, as well as its rate, and that C fixes the distribution of service times, as well as their mean.) Suppose that an accepted customer costs w per unit of time that it remains in the queue, and generates a reward s on service completion. Let $W(\nu, C; w, s)$ be the maximal net rate of return over all admission and scheduling policies. Observe that W is a convex function of (w, s), since it can be expressed as a supremum of linear functions of (w, s),

$$W(\nu, C; w, s) = \sup_{\pi}\{sx(\pi) - wn(\pi)\},$$

where $x(\pi)$ is the mean acceptance rate of customers and $n(\pi)$ is the mean number of customers in the queue under a policy π. Let

$$N(\nu, C; x) = \inf_{\pi}\{n(\pi) : x(\pi) \geq x\},$$

the minimal mean number of customers in the queue, subject to the requirement that the mean acceptance rate of customers must be at least x. N is a convex function of x – indeed W and N are conjugates:

$$\begin{aligned} W(\nu, C; w, s) &= \sup_{x}\{sx - wN(\nu, C; x)\}. \\ N(\nu, C; x) &= \sup_{s}\{sx - W(\nu, C; 1, s)\}. \end{aligned}$$

Under simplified assumptions concerning arrival and service times it is possible to determine explicitly the functions W and N. For example, suppose that arrivals of customers form a Poisson process of rate ν, and that service times are exponentially distributed with mean C^{-1} independently of earlier arrival and service times. Then the optimal policy for the resulting Markov decision process is to accept an arriving customer if the number of customers already in the queue is less than a certain integer, and to reject the customer otherwise [21].

Return now to the network model, and let

$$W_i(w, s) = W(\nu_i, C_i; w, s)$$

$$N_i(x) = N(\nu_i, C_i; x).$$

Then in [11] it is shown that under any policy the expected cost per unit time is bounded below by the value attained in the problem

(3.3a) minimize $\sum_i w_i N_i(x_i)$

(3.3b) subject to $x_i + \sum_{r \in R(i)} y_r = \nu_i \qquad i \in I$

(3.3c) and $$x_i + \sum_{r:i\in r} y_r \le C_i \qquad i \in I$$

(3.3d) over $$x_i \ge 0, i \in I, \quad \text{and} \quad y_r \ge 0,\, r \in R,$$

or, equivalently, the value attained in the dual problem

(3.4a) maximize $$\sum_i [s_i\nu_i - c_iC_i - W_i(w_i, s_i - c_i)]$$

(3.4b) subject to $$s_{i(r)} \le \sum_{j\in r} c_j \qquad r \in R$$

(3.4c) over $$c_i \ge 0,\, i \in I, \quad \text{and} \quad s_i, i \in I.$$

A sketch of the proof is again instructive. Consider values s, c that satisfy (3.4b) and (3.4c). Suppose a network operator gives to each arriving customer of type i an amount s_i, but let queue i charge an amount s_i to directly routed customers and an amount c_i to alternatively routed customers. Note that $\sum_i s_i\nu_i$ is the rate of disbursement to arriving customers and that each customer pays to the queues at least the amount received from the network operator, by condition (3.4b). Suppose that queue i pays to the network operator an amount w_i per unit of time it holds a directly routed customer. Then

$$W_i(w_i, s_i - c_i) + c_iC_i$$

is an upper bound on the net rate of return that queue i can attain, and, since at least $\sum_i s_i\nu_i$ accrues to the queues from charges to customers,

$$\sum_i s_i\nu_i - \sum_i [W_i(w_i, s_i - c_i) + c_iC_i]$$

is a lower bound on the delay costs that are incurred in the system.

The complementary slackness conditions again have an interesting interpretation. The main condition shows that if the charges along route r are too high (i.e. $\sum_{j\in r} c_j > s_{i(r)}$) then the route is not used ($y_r = 0$). The other condition is more familiar: the charge c_i to alternatively routed customers at queue i is zero if there is spare capacity at this queue, as indicated by slackness in the constraint (3.3c).

Under exponential assumptions concerning arrival and service times, the function $N_i(x_i)$ is a piecewise linear convex function of x_i, for $i \in I$. In this case both the primal and dual problems (3.3) and (3.4) can be rewritten as linear programming problems.

When might we expect the above bound to be tight? From its derivation the bound will be tight if alternatively routed flows incur zero delay. This seems a condition unlikely to be satisfied, but we note a circumstance when the condition may be approached. Suppose that for each demand of type i there is a large number of alternative routes, that through each resource j there pass a large number of alternative routes, and that it is possible to break a demand into small sub-demands. The sub-demands have service requirements then add to the service requirement of the original demand, but they can be routed separately: the delay incurred by the original demand is then the maximum delay across the network incurred by any sub-demand. Under these circumstances we expect the bound to be approached (see [7],[8] for methods which are able to establish this for sufficiently symmetric networks, under Poisson arrival and exponential requirement assumptions).

If it is not possible to break a demand into smaller sub-demands, then the above bound will not be tight, since an alternatively routed demand must incur delay, and may cause additional delay to directly routed demands. In [11] there is developed a bound that takes these delays into account, but simplifies other aspects of structure.

3.3. Pipelined networks. We finish the Section with another variation which combines certain interesting aspects of queueing and loss networks. Suppose that an alternatively routed demand must be *pipelined*: that is it must be served simultaneously and without pre-emption at each queue on the alternative route, necessarily having the same service time at each of those queues. Suppose further that all service times are independent and exponentially distributed with unit mean, that it costs w_i per unit of time that a demand of type i remains in the system, and that arrival streams at the network form independent Poisson processes. (The exponential and Poisson assumptions are made for simplicity of exposition; the model and bounds could be developed under more general assumptions – see [11].)

Again consider a single queue with two types of customer. Customers of type 1 arrive in a Poisson stream of rate ν, and each customer may be accepted or rejected. There is an infinite reservoir of type 2 customers. The arrival of a type 2 customer may be triggered at any instant the server chooses, but after arrival type 2 customers have priority at the server. All service times are exponentially distributed with unit mean. Customers of type 1 cost an amount w per unit of time they are in the queue, and a customer of type i generates a reward s_i on service completion. Let $S(\nu; w, s_1, s_2)$ be the maximal net rate of return over all admission policies. Again the form of the optimal policy is clear: type 2 customers will be accepted, if at all, only when the queue is empty, and type 1 customers will be accepted if the queue size is below a threshold. Let $N(\nu; x, y)$ be the minimal mean number of type 1 customers in the queue for a given flow x of type 1 customers and y of type 2 customers. Then

$$\begin{aligned} S(\nu; w, s_1, s_2) &= \sup_{x,y \geq 0} \{s_1 x + s_2 y - wN(\nu; x, y)\}, \\ N(\nu; x, y) &= \sup_{s_1, s_2 \geq 0} \{s_1 x + s_2 y - S(\nu; 1, s_1, s_2)\}. \end{aligned}$$

Further, the expected cost per unit time for the network is bounded below by the value attained in the problem

$$\begin{array}{lll} \text{minimize} & \sum_i w_i[N(\nu_i; x_i, v_i) + \sum_{r\in R(i)} y_r] & \\ \text{subject to} & x_i + \sum_{r\in R(i)} y_r = \nu_i & i \in I \\ \text{and} & \sum_{r:i\in r} y_r = v_i & i \in I \\ \text{over} & x_i, v_i \geq 0, i \in I, \quad \text{and} \quad y_r \geq 0, r \in R, & \end{array}$$

or, equivalently, the value attained in the dual problem

$$\begin{array}{lll} \text{maximize} & \sum_i [s_i\nu_i - S(\nu_i; w_i, s_i, c_i)] & \\ \text{subject to} & s_{i(r)} \leq w_{i(r)} + \sum_{j\in r} c_j & r \in R \\ \text{over} & c_i, s_i, \quad i \in I. & \end{array}$$

We expect this bound to be approached in large highly connected networks under a form of threshold routing, where a demand is sent to an idle alternative route whenever the queue for its direct route exceeds a threshold.

4. Combined bounds. The bounds of Sections 2 and 3 are expected to be tight under the different limiting regimes of heavy traffic and diverse routing respectively. For a given finite network it is often possible to combine the bounds in a manner that gives improved accuracy and further insight into the qualitative behaviour of the network. We illustrate this point with two loss network examples.

4.1. Vertex constraints. Consider a loss network, as described in Section 3.1. Make the additional and natural assumption that the set of resources I are the edges of a graph with vertex set V, and that a route $r \in R(i)$ is a path through this graph connecting the vertices at either end of the edge i. Write $v \in i$ if vertex v is at an end of the edge i, and write $v \in T(r)$ if the route r passes through vertex v, but v is not an end vertex of the path r. Thus $T(r)$ is the set of *tandem* vertices on route r.

Let

$$\begin{aligned} M_v(x) &= M(\sum_{j:v\in j} \nu_j, \infty, \sum_{j:v\in j} C_j; x) \\ W_v(w_1, w_2) &= W(\sum_{j:v\in j} \nu_j, \infty, \sum_{j:v\in j} C_j; w_1, w_2). \end{aligned}$$

Consider the flow of traffic into and out of vertex v. The amount of tandem traffic is

$$2 \sum_{r:v\in T(r)} y_r,$$

since each call that uses vertex v as a tandem vertex occupies a circuit into, *and* a circuit out of, vertex v. Hence under any dynamic routing scheme

$$(4.1)\qquad 2\sum_{r:v\in T(r)} y_r \leq M_v\Big(\sum_{j:v\in j}(x_j+\sum_{r\in R(j)} y_r)\Big),$$

providing a single additional constraint for each vertex $v \in V$. We can improve the bound provided by problem (3.1) by appending to that problem the additional constraints (4.1). The corresponding dual problem takes the form

$$\begin{array}{lll}
\text{minimize} & \sum_i [W_i(e_i,c_i)+s_i\nu_i] + \sum_v W_v(g_v,d_v) & \\
\text{subject to} & w_i \leq s_i + e_i + \sum_{v\in i} g_v & i \in I \\
\text{and} & e_{i(r)} \leq \sum_{j\in r} c_j + 2\sum_{v\in T(r)} d_v & r \in R \\
\text{over} & c_i, s_i, e_i \geq 0, i \in I, \quad \text{and} & g_v, d_v \geq 0, v \in V.
\end{array}$$

Comparing this formulation with problem (3.4), we observe the introduction of additional implied costs g_v and d_v: the vertex v charges an amount g_v for the use of a single circuit out of vertex v by a call terminating at vertex v, and an amount $2d_v$ for the use of two circuits out of vertex v by a call using vertex v as a tandem vertex. The conceptual simplicity is notable: there is a charge for each vertex, as well as each link, used by a call.

4.2. Other cut constraints. Each of the constraints (4.1) corresponds to the cut formed between a single vertex and the other vertices. A network's size, connectivity and asymmetry may make some other key subset of resources important: we next describe an example illustrating how an additional cut constraint may be included.

Consider again the loss network described in Section 3.1. Let $J \subset I$ be such that

$$j \in J, r \in R(j) \Rightarrow |r \cap J| = 1,$$

where $|r \cap J|$ is the number of resources from J on route r. Thus J forms a cut, in that every demand from J requires to use a single resource from J however it is routed. Let

$$\begin{aligned}
M_J(x) &= M(\sum_{j\in J}\nu_j, \infty, \sum_{j\in J} C_j; x) \\
W_J(w_1,w_2) &= W(\sum_{j\in J}\nu_j, \infty, \sum_{j\in J} C_j; w_1, w_2).
\end{aligned}$$

Consider the flow of traffic across the cut J. Under any dynamic routing scheme

$$\sum_{k\notin J}\sum_{r\in R(k)} |r\cap J| y_r \leq M_J\Big(\sum_{j\in J}(x_j + \sum_{r\in R(j)} y_r)\Big). \tag{4.2}$$

We can improve the bound provided by problem (3.1) by appending to that problem that additional constraint (4.2). The corresponding dual problem takes the form

$$\begin{array}{lll} \text{minimize} & \sum_i \big[W_i(e_i,c_i) + s_i\nu_i\big] + W_J(g_J,d_J) & \\ \text{subject to} & w_i \leq s_i + e_i + I[i\in J]g_J & i\in I \\ \text{and} & e_{i(r)} \leq \sum_{j\in r} c_j + |r\cap J|d_J & r\in R \\ \text{over} & c_i, s_i, e_i \geq 0, i\in I, \quad \text{and} \quad g_J, d_J\geq 0. & \end{array}$$

Again the conceptual simplicity is notable: each call serving a demand $i\in J$ must pay a charge g_J, while other calls must pay a charge d_J per circuit used from the cut J.

Work in progress is concerned with illustrating these various bounds and assessing their accuracy for a variety of realistic network architectures: for early results the reader is referred to [3] and [4].

REFERENCES

[1] D. Bertsimas, I. Ch. Paschalidis and J.N. Tsitsiklis, *Scheduling of multiclass queueing networks: bounds on achievable performance*, Ann. Appl. Prob.

[2] R.J. Gibbens and F.P. Kelly, *Dynamic routing in fully connected networks*, IMA J. Math. Cont. Inf. **7** (1990), pp. 77–111.

[3] R.J. Gibbens, F.P. Kelly, and S.R.E. Turner, *Dynamic routing in multiparented networks*, IEEE Trans. Networking **1** (1993), pp. 261–270.

[4] R.J. Gibbens and P. Reichl, *Performance bounds applied to loss networks*, in: Complex Stochastic Systems and Engineering, editor D.M. Titterington (Oxford University Press, 1994).

[5] M. Gondran and M. Minoux, Graphs and Algorithms (Wiley, New York, 1984).

[6] J.M. Harrison, *Brownian models of queueing networks with heterogeneous customer populations*, in: Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol 10, editors W. Fleming and P.-L. Lions (Springer, New York, 1988), pp. 147–186.

[7] P.J. Hunt and C.N. Laws, *Least busy alternative routing in queueing and loss networks*, Probab. Eng. Inf. Sci. **6** (1992), pp. 439–456.

[8] P.J. Hunt and C.N. Laws, *Asymptotically optimal loss network control*, Math. Oper. Res. **18** (1993), pp. 880–900.

[9] F.P. Kelly, *The optimization of queueing and loss networks*, in: Queueing Theory and its Applications, editors O.J. Boxma and R. Syski (Elsevier, Amsterdam, 1988), pp. 375–392.

[10] F.P. Kelly, *Loss networks*, Ann. Appl. Prob. **1** (1991), pp. 317–378.

[11] F.P. Kelly, *Bounds on the performance of dynamic routing schemes for highly connected networks*, Math. Oper. Res. **19** (1994), pp. 1–20.

[12] F.P. Kelly and C.N. Laws, *Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling*, Queueing Systems **13**, (1993), pp. 47–86.

[13] P.B. Key, *Optimal control and trunk reservation in loss networks*, Prob. Engineering and Information Sciences **4** (1990), pp. 203–242.

[14] P.B. Key, *Some control issues in telecommunication networks*, in: Probability, Statistics and Optimization, editor F.P. Kelly (Wiley, Chichester, 1994), pp. 383–395.

[15] K.R. Krishnan and T.J. Ott, *State-dependent routing for telephone traffic: theory and results*, Proc. 25th Conf. Decision and Control, Athens (1986).

[16] H.J. Kushner, *Control of trunk line systems in heavy traffic*, Division of Applied Mathematics, Brown University (1992).

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

[18] C.N. Laws, *Paper in this volume.*

[19] C.N. Laws and G.M. Louth, *Dynamic scheduling of a four-station queueing network*, Prob. Eng. Inf. Sci. **4** (1990), pp. 131–156.

[20] V.G. Lazarev and S.M. Starobinets, *The use of dynamic programming for optimization of control in networks of commutation of channels*, Engineering Cybernet. **15** (1977), pp. 107–117.

[21] S.A. Lippman, *Applying a new device in the optimization of exponential queueing systems*, Oper. Res. **23** (1975), pp. 687–710.

[22] G.M. Louth, M. Mitzenmacher and F.P. Kelly, *Computational complexity of loss networks*, Theor. Comp. Sci. **125** (1994), pp. 45–59.

[23] V.V. Marbukh *Asymptotic investigation of a complete communications network with a large number of points and bypass routes*, Problemy Peradači Informacii **16** (1981), pp. 89–95.

[24] D. Mitra and R.J. Gibbens, *State-dependent routing on symmetric loss networks with trunk reservations, II: asymptotics, optimal design*, Ann. Oper. Res. **35** (1992), pp. 3–30.

[25] J. Ou and L.M. Wein, (1992) *Performance bounds for scheduling queueing networks*, Ann. Appl. Probab. **2** (1992), pp. 460–480.

[26] M.I. Reiman, *A multiclass feedback queue in heavy traffic*, Adv. Appl. Prob. **20** (1988), pp. 179–207.

[27] M.I. Reiman, *Optimal trunk reservation for a critically loaded network*, Proc. 13th ITC. In Teletraffic and Datatraffic, editors A. Jensen and V.B. Iverson (Elsevier, Amsterdam, 1991).

[28] M.I. Reiman and B. Simon, *A network of priority queues in heavy traffic: one bottleneck station*, Queueing systems **6** (1990), pp. 33–58.