J. Austral. Math. Soc. Ser. B 31(1989), 204–218

FIXED POINT MODELS OF LOSS NETWORKS

F. P. KELLY[1]

(Received December 1988; revised February 1989)

Abstract

In this paper we review a simple class of fixed point models for loss networks. We illustrate how these models can readily deal with heterogeneous call types and with simple dynamic routing strategies, and we outline some of the recent mathematical advances in the study of such models. We describe how fixed point models lead to a natural and tractable definition of the *implied cost* of carrying a call, and how this concept is related to issues of routing and capacity expansion in loss networks.

1. Introduction

Fixed point models of loss networks have a long history in the telecommunications literature (see, for example, [3], [6], [19], [20] and [26]) and continue to pose interesting and difficult challenges to mathematicians. In this paper we review a very simple class of fixed point models, aiming to illustrate the interplay between the practical and theoretical issues raised. We begin, in Section 2, by describing the basic model in the case of a loss network operating under fixed routing. In Section 3 we describe how the model leads to a natural and tractable definition of the *implied cost* of carrying a call, and how this concept is related to issues of routing and capacity allocation in loss networks. In Sections 4 and 5 we outline how the basic model can be extended to loss networks with alternative routing and trunk reservation, and illustrate the form taken by implied costs in simple examples.[1]

There is currently considerable interest in schemes which can dynamically control the routing of calls within a network. The purpose of such dynamic routing schemes is to adjust routing patterns with a network in accordance with varying or uncertain offered traffics, to make better use of spare capacity

[1]Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England.
© Copyright Australian Mathematical Society 1989, Serial-fee code 0334-2700/89

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

in the network resulting from dimensioning upgrades or forecasting errors, and to provide extra flexibility and robustness to respond to failures or overloads. In Section 6 we describe a scheme, Dynamic Alternative Routing, which is now being implemented in the UK trunk network [23]. An important contributing factor in the development of this scheme was the ease with which fixed point methods could be adapted to provide accurate estimates of network performance under a wide range of failure and overload conditions ([6], [8]).

Finally, in Section 7, we briefly outline an extension of our basic model to the case of heterogeneous call types, and consider especially the interaction of trunk reservation and variable holding period distributions. This topic is becoming increasingly important with the advent of integrated services digital networks, where a single network may have to handle demands with widely differing service requirements.

2. Fixed routing

We begin by describing our basic model of a loss network operating under fixed routing. Consider a network with J links, labelled $1, 2, \ldots, J$, and suppose that link j comprises C_j circuits. A call on route r uses A_{jr} circuits from link j, where $A_{jr} \in \mathbf{Z}_+$. Let $\mathscr{R}$ be the set of possible routes. In the important special case where each element of the matrix $A = (A_{jr}, j = 1, 2, \ldots, J; r \in \mathscr{R})$ is either 0 or 1, a route r can be identified with a subset of the set of links $\{1, 2, \ldots, J\}$: just set $r = \{j: A_{jr} = 1\}$. Calls requesting route r arrive as a Poisson stream of rate ν_r, and as r varies it indexes independent Poisson streams. A call requesting route r is blocked and lost if on any link j, $j = 1, 2, \ldots, J$, there are less than A_{jr} circuits free. Otherwise the call is connected and simultaneously holds A_{jr} circuits from link j, $j = 1, 2, \ldots, J$, for the holding period of the call. The call holding period is independent of earlier arrival times and holding periods; holding periods of calls on route r are identically distributed with unit mean.

It is possible to provide an analytical formula for the stationary distribution of the stochastic process described above and hence for loss probabilities (see, for example, [2], [12]), but these explicit forms are computationally intractable for networks in which the number of routes $|\mathscr{R}|$ or the link capacities $C_1, C_2, \ldots, C_J$ are large. Fortunately there is an approximation to hand. Let $B_1, B_2, \ldots, B_J$ be a solution to the equations

$$B_j = E(\rho_j, C_j) \qquad j = 1, 2, \ldots, J \tag{2.1}$$

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

where

$$\rho_j = (1 - B_j)^{-1} \sum_r A_{jr} \nu_r \prod_i (1 - B_i)^{A_{ir}} \tag{2.2}$$

and

$$E(\nu, C) = \frac{\nu^C}{C!} \left(\sum_{n=0}^{C} \frac{\nu^n}{n!} \right)^{-1}. \tag{2.3}$$

Here $E(\nu, C)$ is just Erlang's formula for the loss probability of a single link of capacity C offered Poisson traffic at rate ν. Then an approximation for the loss of probability on route r is

$$L_r = 1 - \prod_j (1 - B_j)^{A_{jr}}. \tag{2.4}$$

The right hand side of equations (2.1), regarded as a function of $(B_1, B_2, \ldots, B_J)$, defines a continuous mapping from the compact convex set $[0, 1]^J$ into itself, and so, by the Brouwer fixed point theorem, there exists a solution $(B_1, B_2, \ldots, B_J)$ to equations (2.1). In [13] it is proved that the solution is unique, by showing that it is a stationary point of a strictly convex potential function. The solution has been termed the Erlang fixed point [13] or, when A is a $0 - 1$ matrix, the reduced load approximation [24]. The approximation (2.1)–(2.4) has a long history, at least in the case where A is a $0 - 1$ matrix: for early examples of its use see [3], [26]. The underlying idea is simple to explain. If a request for a circuit from link i is denied with probability B_i, and if we make the approximation that all such requests are granted or denied independently, then the traffic offered to link j will comprise independent Poisson streams, and the level of carried traffic on link j will be $\sum_r A_{jr} \nu_r \prod_i (1 - B_i)^{A_{ir}}$. Equations (2.1) and (2.2) simply state that the blocking probability on link j should be consistent with this level of carried traffic, under the Erlang model of a single link offered Poisson single-circuit traffic. Call ρ_j, given by expression (2.2), the *reduced load* on link j.

For small networks the approximation (2.4) can be checked against the exact loss probability and it is known that the approximation can be fairly accurate. In [13] it is further shown that if capacities C_j, $j = 1, 2, \ldots, J$, and offered traffics ν_r, $r \in \mathscr{R}$, are increased together (with ratios C_j/ν_r held fixed) then the approximation (2.4) converges to the correct value. This result indicates that the larger the capacities in a network the more accurate the approximation will be, and complements the work of Whitt [24] and Ziedins and Kelly [28]: by considering certain networks in which J and $\mathscr{R}$ become large they obtained results which indicate that the more diverse the routing within a network the more accurate the approximation procedure will be.

Of course there are circumstances where the approximation procedure should not be expected to perform well. For example, if a number of small

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

capacity links are arranged one after another in a line then we would expect considerable dependence between the number of free circuits on adjacent links. This example, more typical of local area networks than large scale telecommunications networks, is considered in detail in [14] and [27]. Also, note that our model assumes that routes are fixed. For models involving alternative routing the equivalent approximation procedure may not lead to a unique fixed point ([13], [19], [21]). This is not necessarily a fundamental flaw in the procedure: nonuniqueness may indicate instabilities in the network, with a number of distinct modes of behaviour possible ([1], [7]).

3. Implied costs

Important issues concern how routes should be chosen or capacity allowed in loss networks. These issues are complicated by the fact that small changes in one part of the network may have repercussions over a large area, and these knock-on effects must be taken into account. A related issue concerns the extent to which control can be decentralised. Over a period of time the form of the network or the demands placed on it may change, and routings may need to adapt accordingly. A single node could perhaps control this, receiving information from everywhere in the network and making all decisions about routing. But this approach has drawbacks, particularly if links or nodes may fail. Could control be distributed over the nodes of the network, with computations and decisions made locally? A distributed control scheme should be able to react rapidly to a local disturbance at the point of the disturbance, with slower adjustments in the rest of the network as effects propagate outwards.

Some insight into these issues can be gained from further analysis of the fixed point model described in Section 2. Suppose that each call carried on route r is worth w_r. Then, under the approximation (2.4), the rate of return from the network will be $W(\nu; C) = \sum_r w_r \lambda_r$ where $\lambda_r = \nu_r(1 - L_r)$ corresponds to the traffic carried on route r. We emphasise the dependence of W on the vectors of offered traffics $\nu = (\nu_r, r \in \mathscr{R})$ and capacities $C = (C_1, C_2, \ldots, C_J)$. Let $\delta_j = \rho_j(E(\rho_j, C_j - 1) - E(\rho_j, C_j))$, extend the definition (2.3) to non-integral values of the scalar C by linear interpolation, and at integer values of C_j define the derivative of $W(\nu; C)$ with respect to C_j to be the left derivative. Then it is possible to prove that

$$\frac{d}{d\nu_r} W(\nu; C) = (1 - L_r)s_r \tag{3.1}$$

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

and

$$\frac{d}{dC_j}W(\nu;C) = c_j \tag{3.2}$$

where $s = (s_r, r \in \mathscr{R})$ and $c = (c_1, c_2, \ldots, c_J)$ are the unique solution to the linear equations

$$s_r = w_r - \sum_j c_j A_{jr} \tag{3.3}$$

$$c_j = \delta_j \sum_r A_{jr}\lambda_r(s_r + c_j) / \sum_r A_{jr}\lambda_r. \tag{3.4}$$

We can interpret s_r as the *surplus value* of a call on route r: if such a call is accepted it will earn w_r directly but at an *implied cost* of c_j for each circuit used from link j. The implied costs c measure the expected knock-on effects of accepting a call upon later arrivals at the network. From (3.2) it follows that c_j is also a *shadow price*, measuring the sensitivity of the rate of return to the capacity C_j of link j. The local character of equations (3.3) and (3.4) is striking. The right hand side of (3.3) involves costs c_j only for links j on the route r, while (3.4) exhibits c_j in terms of an average, weighted over just those routes through link j, of $s_r + c_j$.

The formal mathematical derivation of the relationships (3.1)–(3.4) is, in a certain sense, elementary. These are, after all, simply relationships between the derivatives of an implicitly defined function. The elementary approach is, however, tedious. It is illustrated in [15] where a frontal assault is made on (2.1)–(2.3), involving calculation of partial and total derivatives of B_1, $B_2, \ldots, B_J$ with respect to ν and C, and subsequent reduction of the equations obtained. An elegant alternative approach is suggested by the work of Whittle [25]. The fixed point $B_1, B_2, \ldots, B_J$ locates a stationary point of a potential function, and so derivatives of W can be deduced from derivatives of the potential function (note that Whittle [25] focuses on an alternative saddle-point approximation, but his approach applies in the present context also). Unfortunately this approach does not appear capable of extension to the more complex models, involving trunk reservation, to be considered in later sections: these models lack the required characterisation of fixed points as stationary points of a potential function. A third approach [16] is based on the differentiation of W on various carefully constructed manifolds around the point $(\nu, C) \in \mathbf{R}^{\mathscr{R}} \times \mathbf{R}^J$. Currently this approach seems to be the most widely applicable; it also seems to be the most direct, in that equations possessing the local character of (3.3) and (3.4) emerge naturally.

We illustrate the third approach by deriving the relations (3.2)–(3.4) in the case where A is a $0-1$ matrix. Suppose, without loss of generality, that there exist marker routes $\{j\} \in \mathscr{R}$ for $j = 1, 2, \ldots, J$, with $\nu_{\{j\}} = w_{\{j\}} = 0$.

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

For notational simplicity write ν_j for $\nu_{\{j\}}$. We aim to differentiate W on a manifold around $(\nu_k,\ k = 1,2,\ldots,J)$ constructed so that on this manifold B_k is constant for $k \neq j$. An informal description is as follows. Alter the offered traffic ν_j. This will affect directly the blocking probability B_j at link j, and hence the carried traffics λ_r for routes r through link j. This in turn will have indirect effects upon other links through which these routes pass. We can, however, cancel out these indirect effects by judicious alterations to ν_k for $k \neq j$. The alterations to ν_k have to be such as to leave the reduced load ρ_k constant for $k \neq j$, since then, from (2.1), the blocking probability B_k will be left constant for $k \neq j$. Let us begin by calculating the direct effect of the change in ν_j on the carried traffic λ_r along a route through link j. From the relation

$$\lambda_r = \nu_r \prod_k (1 - B_k)^{A_{kr}}$$

the direct effect is

$$d\lambda_r = -A_{jr}(1 - B_j)^{-1}\lambda_r \cdot \frac{\partial B_j}{\partial \nu_j} \cdot d\nu_j. \tag{3.5}$$

Next we calculate the necessary alterations to ν_k for $k \neq j$. In order that ρ_k be left constant the change $d\lambda_r$ must be balanced by a change

$$d\nu_k = -A_{kr}(1 - B_k)^{-1} d\lambda_r,$$

from (2.2); observe that $A_{kr}(1 - B_k)^{-1}\lambda_r$ is the contribution to the reduced load ρ_k from route r. Observe also that, apart from marker routes, the only routes for which λ_r changes are routes through link j; the effect on $W(\nu; C)$ can thus be calculated from (3.5). Formally, we have an evaluation for the differential form

$$\left[\frac{d}{d\nu_j} + \sum_r A_{jr}(1 - B_j)^{-1}\lambda_r \cdot \frac{\partial B_j}{\partial \nu_j} \cdot \sum_{k \neq j} A_{kr}(1 - B_k)^{-1} \frac{d}{d\nu_k}\right] W(\nu; C)$$
$$= -\sum_r A_{jr}(1 - B_j)^{-1}\lambda_r \cdot \frac{\partial B_j}{\partial \nu_j} \cdot w_r. \tag{3.6}$$

Now an elementary partial derivative calculation from Erlang's formula shows that

$$\frac{\partial B_j}{\partial \nu_j} = (1 - B_j)\delta_j \rho_j^{-1}. \tag{3.7}$$

A partial differentiation with respect to the second argument of Erlang's formula further establishes that c_j, defined by (3.2), satisfies

$$c_j = -(1 - B_j)^{-1} \frac{d}{d\nu_j} W(\nu; C). \tag{3.8}$$

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

Using the relations (3.7) and (3.8), (3.6) can be rewritten in the desired form (3.3) and (3.4). We have thus established relations (3.2)–(3.4). Relation (3.1) can be obtained similarly, by differentiating W on a manifold around $\nu \in \mathbf{R}^{\mathscr{R}}$ constructed so that ν_r and ν_j, $j \in r$, are allowed to alter, but $B_1, B_2, \ldots, B_J$ are held fixed.

The derivatives (3.1) and (3.2) are exact, but are calculated from a fixed point model which is itself only approximate. How accurate are these derivatives? Whittle [25] and Hunt [10] have shown that if capacities and offered traffics are increased to infinity, with ratios held fixed, then derivatives calculated from the exact stationary distribution and derivatives calculated from the Erlang fixed point converge to the same values provided the network has no *critically loaded* links. A critically loaded link is one at which the reduced load, ρ_j, very nearly matches the capacity, C_j, in that $\rho_j - C_j = o(C_j^{1/2})$ under the limiting regime. At a critically loaded link the blocking probability approaches zero: the Erlang fixed point is asymptotically correct at this level of detail, but it is incorrect in its estimation of the rate of convergence to zero under the limiting regime [11]. This finer level of detail matters for derivatives: Hunt [10] shows that in a network containing critically loaded links, derivatives calculated from the exact stationary distribution and from the Erlang fixed point may converge to different values. The extent of the discrepancy depends on the diversity of routing within the network. Hunt [10] also provides an example of a network in which all links have just unit capacity, but in which the discrepancy between the derivatives disappears as the number of links, and the number of routes through each link, increases.

4. Alternative routing

In this section we indicate how the fixed point model can deal with alternative routing, where a call which is blocked on a route may be allowed to try again on another route. We begin by describing a very general form of alternative routing, where the arrival rate for a route is allowed to depend arbitrarily upon which links are full. Henceforth assume A is a $0-1$ matrix, so that a call requires at most one circuit from a link. Let $b = (b_1, b_2, \ldots, b_J)$ denote the blocking configuration of the links: $b_j = 1$ or 0 according as link j has free circuits or not. Write $p(b, B) = \prod_j B_j^{1-b_j}(1 - B_j)^{b_j}$. Thus $p(b, B)$ is the probability of blocking configuration b under the assumption that links $1, 2, \ldots, J$ block independently, link j blocking with probability B_j. Write $\lambda_r(b)$ for the traffic offered to route r when the blocking configuration is b. Insist that $\lambda_r(b) = 0$ if $\prod_j b_j^{A_{jr}} = 0$ where here and throughout $0^0 = 1$. Thus

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

no traffic is offered to route r if it cannot be accepted there, and so $\lambda_r(b)$ is also the rate of accepted traffic on route r when the blocking condition is b. Then the generalisation of the fixed point equations (2.1) and (2.2) is

$$B_j = E(\rho_j, C_j) \qquad j = 1, 2, \ldots, J \tag{4.1}$$

where

$$\rho_j = (1 - B_j)^{-1} \sum_r A_{jr} \sum_b p(b, B) \lambda_r(b). \tag{4.2}$$

Observe that we recover (2.1) and (2.2) if we set $\lambda_r(b) = \nu_r \prod_j b_j^{A_{jr}}$, corresponding to the case of fixed routing.

We devote the rest of this section to a more complex example, in which we suppose that the label r fixes a pair $(\phi(r), \psi(r))$, where $\phi(r)$ and $\psi(r)$ are both sets of links. Interpret $\psi(r)$ as the path used by a call on route r, and $\phi(r)$ as a set of links each of which must be blocked in order for this path to be attempted. For instance, suppose a call tries first a path through links $\{1, 2\}$; if link 1 is blocked it then tries a path through links $\{3, 4, 5\}$; while if link 1 is free and link 2 is blocked it tries a path through links $\{1, 6, 7\}$. This pattern of choices can be represented by the following three pairs $(\phi(r), \psi(r))$:

$$(\varnothing, \{1, 2\}), \qquad (\{1\}, \{3, 4, 5\}), \qquad (\{2\}, \{1, 6, 7\}).$$

Any routing scheme not involving crankback can be thus represented by pairs $(\phi(r), \psi(r))$, and with such a representation

$$\lambda_r(b) = \nu_r \prod_{j \in \phi(r)} (1 - b_j) \prod_{k \in \psi(r)} b_k \tag{4.3}$$

where ν_r is the arrival rate at the network of calls potentially served by route r.

Next we consider whether implied costs and surplus values can be defined and calculated as in the case of direct routing. Let $\lambda_r = \sum_b p(b, B) \lambda_r(b)$, where $\lambda_r(b)$ is given by (4.3); thus λ_r corresponds to the total carried traffic on route r. Again define the rate of return from the network by $W(\nu; C) = \sum_r w_r \lambda_r$, and let

$$\frac{d}{dC_j} W(\nu; C) = c_j. \tag{4.4}$$

Then an elementary (but lengthy) analysis of the implicit equations (4.1) and (4.2) defining $B_1, B_2, \ldots, B_J$, and hence λ_r and $W(\nu; C)$, leads to the following equations for $c_1, c_2, \ldots, c_J$ [15]:

$$s_r = w_r - \sum_{j \in \psi(r)} c_j \tag{4.5}$$

$$c_j = \eta_j \left[(1 - B_j)^{-1} \sum_{r:\ j \in \psi(r)} \lambda_r(s_r + c_j) - B_j^{-1} \sum_{r:\ j \in \phi(r)} \lambda_r s_r \right] \tag{4.6}$$

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

where $\eta_j = E(\rho_j, C_j) - E(\rho_j, C_j - 1)$. Further

$$\frac{d}{d\nu_r} W(\nu; C) = s_r \prod_{j \in \phi(r)} B_j \prod_{k \in \psi(r)} (1 - B_k).$$

Again we can interpret s_r as the surplus value of a call on route r, and c_j as the implied cost of using a circuit from link j. Equations (4.5) and (4.6) retain a local character, but it is weaker than that apparent in (3.3) and (3.4). The right hand side of (4.6) involves the surplus values not just of those routes which pass through link j, but also of those routes which require link j to be blocked before they are attempted. The relation (4.4) shows that the implied cost c_j retains its dual role as a shadow price.

5. Trunk reservation

Trunk reservation is an easily implemented control mechanism which allows priority to be given to chosen traffic streams. It is especially helpful in networks which allow alternative routing, where without its use performance may degrade significantly ([19], [22]). Consider a single link with a capacity of C circuits that is offered a stream of priority traffic at rate ρ_1 and a stream of non-priority traffic at rate ρ_2. A priority call is accepted if there is a circuit free on the link, while a non-priority call is accepted only if there are more than t circuits free on the link. Here t is called the *trunk reservation parameter.* If the arrival streams are independent Poisson processes and if call holding periods are independent of earlier arrival times and holding periods and exponentially distributed with unit mean, then the number of circuits busy is a birth and death process whose equilibrium distribution is readily calculated. It follows from this distribution that

$$E_1(\rho_1, \rho_2, C, t) = G(\rho_1, \rho_2, C, t)(\rho_1 + \rho_2)^{C-t} \rho_1^t / C! \tag{5.1}$$

is the proportion of priority calls blocked and that

$$E_2(\rho_1, \rho_2, C, t) = G(\rho_1, \rho_2, C, t)(\rho_1 + \rho_2)^{C-t} \sum_{n=C-t}^{C} \rho_1^{n-C+t} / n! \tag{5.2}$$

is the proportion of non-priority calls blocked, where

$$G(\rho_1, \rho_2, C, t) = \left[\sum_{n=0}^{C-t-1} (\rho_1 + \rho_2)^n / n! + (\rho_1 + \rho_2)^{C-t} \sum_{n=C-t}^{C} \rho_1^{n-C+t} / n! \right]^{-1}.$$

Observe that $E_1(\rho_1, 0, C, t)$ reduces to Erlang's formula $E(\rho_1, C)$.

Using the functions E_1, E_2 it is straightforward to extend (4.1) and (4.2) to deal with the possible use of trunk reservation in a network operating under

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

alternative routing. Suppose that link j has a trunk reservation parameter t_j; if n_j is the number of circuits occupied on link j let

$$\begin{aligned} b_j &= 0 \quad \text{if } n_j = C_j, \\ b_j &= 1 \quad \text{if } C_j - t_j \leq n_j < C_j, \\ b_j &= 2 \quad \text{if } n_j < C_j - t_j. \end{aligned}$$

Let $b = (b_1, b_2, \ldots, b_J)$ denote the blocking configuration of the network, and write $\lambda_r(b)$ for the traffic offered to route r when the blocking configuration is b. In this way the arrival rate for a route is allowed to depend arbitrarily upon which links are full, which links are occupied beyond their trunk reservation parameter and which links are neither. In [16] the resulting fixed point equations are presented, and it is shown that they lead to a natural definition of implied costs, surplus values and shadow prices. Implied costs and shadow prices are no longer identical, as they are in networks without trunk reservation, but they can both be readily calculated.

We end this section with a simple example which illustrates how implied costs can be used to provide insight into the fascinating phenomena that arise in networks involving alternative routing. Consider a fully connected network with K nodes. Calls between each pair of nodes arrive at rate ν, and each pair of nodes is connected by a link of capacity C. A call between two nodes is carried on the link joining the nodes if possible. If this link is full, the call is offered to one randomly selected two-link path between the nodes. Trunk reservation operates against alternatively routed calls: such a call is carried if both links have more than t free circuits and is lost otherwise. Suppose now that an additional call is offered to a link of the network, to be routed directly on that link provided the link is not full and to be lost otherwise. If this additional call is accepted there will be an expected net effect upon the network, measured in terms of calls lost that would otherwise have been carried. Call this expected net effect the implied cost. Under the fixed point model this implied cost is readily calculated, and is shown in Figure 1 for a network with links of capacity 120. The dashed curve corresponds to using the optimal trunk reservation parameter at each level of offered traffic. The dotted curve corresponds to using a trunk reservation parameter of zero or, equivalently, not using trunk reservation. Observe the steep rise of the dotted curve, up to a value of almost 2, at an offered traffic of about 107 Erlangs. At this point carrying an additional directly routed call causes the subsequent loss of nearly *two* calls. The additional call can cause this much damage, on average, because it can force a later call to be rerouted: this uses more network resource and may in turn cause further calls to be rerouted. Graphs of shadow prices (the marginal benefit of increased capacity) display very similar features.

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

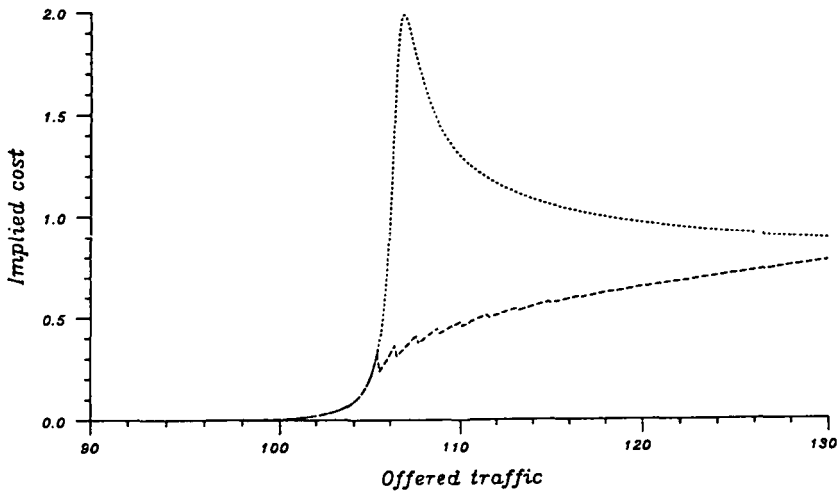

FIGURE 1. Implied costs in a symmetric fully connected network

The implied costs illustrated in Figure 1 are exact deductions from a fixed point model which itself gives only approximations for loss probabilities. Gibbens and Whiting [9] study the validity of the procedure, by comparing implied costs calculated analytically from the model with implied costs estimated by simulation. Preliminary conclusions are that the peak shown in Figure 1 is a real effect, although its height and narrowness are exaggerated by comparison with simulations. There appears to be good general agreement between analysis and simulation when trunk reservation parameters are positive. It seems reasonable to expect that under certain limiting regimes the implied costs calculated from the fixed point approximation will be asymptotically exact: see Hunt [10] for a detailed study in the case of fixed routing without trunk reservation. In practice implied costs have been of use in large asymmetric networks with complex routing patterns ([4], [5], [17], [18]): derivative information is especially valuable for optimising functions defined on a high-dimensional space; and the structure of such networks, with traffic through a link coming from a diverse range of routes, lends credence to the fixed point approximation.

6. Dynamic Alternative Routing

DAR is a simple but effective dynamic routing strategy, which is decentralised and uses only local information ([6], [8], [23]). Its definition for a fully connected network is as follows. Suppose there are K nodes in the network, with link $\{i, j\}$ having capacity C_{ij}. Each link is assigned a trunk

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

reservation parameter r_{ij}, and each source-destination pair stores the identity of its current tandem k for use in two-link alternative routes. A call between nodes i and j is first offered to the direct link and a call is always routed along that link if there is a free circuit. Otherwise, the call attempts the two-link alternative route via tandem node k with trunk reservation applied to both links. If the call fails to be routed via k, this call is lost and, further, the identity of the tandem node is reselected (at random perhaps) from the set $\{1, 2, \ldots, K\} - \{i, j\}$. Note especially that the tandem node is not reselected if the call is successfully routed on either the direct link or the two-link alternative route. In practice it may be simpler to reselect a tandem node by cycling around a fixed tandem permutation; the point is that reselection is not based on any collected data, only the important information that a call has just failed.

Let $p_k(i, j)$ denote the long-run proportion of calls between i and j which are offered to tandem node k, and let $q_k(i, j)$ be the long-run proportion of those calls between i and j and offered to tandem node k which are blocked. Then, under uniform reselection,

$$p_a(i, j)q_a(i, j) = p_b(i, j)q_b(i, j) \qquad a, b \neq i, j.$$

Observe that this simple ergodic result is exact for either random reselection or reselection using a fixed permutation. More generally, suppose the DAR mechanism for reselection of the tandem node between i and j chooses node k with long-run frequency f_k where $\sum_{k \neq i,j} f_k = 1$. Then each selection of node k is paired with a failed call via node k, and so

$$p_a(i, j)q_a(i, j) : p_b(i, j)q_b(i, j) = f_a : f_b \qquad a, b \neq i, j.$$

We approximately estimate $q_k(i, j)$ by $L_k(i, j)$, the loss probability on the two-link path $i - k - j$ given by a fixed point model. The natural fixed point model of DAR is that for a network with alternative routing and trunk reservation, but with the overflow stream from node i to node j divided over tandem nodes $k \in \{1, 2, \ldots, K\} - \{i, j\}$ in proportions

$$p_k(i, j) = \frac{f_k}{L_k(i, j)} \left[\sum_{a \neq i,j} \frac{f_a}{L_a(i, j)} \right]^{-1} \tag{6.1}$$

Observe that if the offered traffic to a link increases, or the capacity of the link decreases, then the blocking probabilities on that link will increase, and this will affect the proportions (6.1). This is the means by which the fixed point model mimics the flexible handling by the routing scheme of overloads and failures.

Simulation experiments to assess the accuracy of the model are reported in [8]. In general there is found to be excellent agreement between simulations

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

and the model, both for overall and for stream grades of service. Key [17] describes an approach to the calculation of implied costs and shadow prices for a fully connected network using DAR, and shows how they can be used interactively to dimension a network.

7. Heterogeneous call types

Our basic model of a loss network, described in Section 2, assumed that all call holding periods had unit mean. This involved no real loss of generality, since if the holding periods of calls on route r have mean μ_r, then the exact analytical formula for loss probabilities in the network of Section 2 depends only on the products $(\nu_r\mu_r, r \in \mathscr{R})$: see, for example, [2], [12]. In a network using trunk reservation, however, the consequences of differing mean holding periods between priority and non-priority traffic are less predictable. In this section we outline a generalisation of our basic model that allows differing mean holding periods in a network with trunk reservation. We describe only the case of fixed routing; the extension to alternative routing is natural through the approach of Section 4.

Amend the model of Section 2 as follows. Let calls on route r have mean holding period μ_r, and suppose that a call on route r requires A_{ljr} circuits from link j at priority level l, where $l = 1, 2$, and A_{1jr}, A_{2jr} and $A_{1jr} + A_{2jr} \in \{0, 1\}$. Let C_j, t_j be the capacity and the trunk reservation parameter respectively for link j. Let $(B_{1j}, B_{2j}, j = 1, 2, \ldots, J)$ solve the equations

$$B_{lj} = E_l(\rho_{1j}, \rho_{2j}, C_j, t_j) \qquad l = 1, 2; \quad j = 1, 2, \ldots, J$$

where

$$\rho_{lj} = \mu_j(1 - B_{lj})^{-1} \sum_r A_{ljr}\nu_r(1 - L_r),$$

$$1 - L_r = \prod_i (1 - B_{1i})^{A_{1ir}}(1 - B_{2i})^{A_{2ir}},$$

$$\mu_j = \sum_r \sum_l A_{ljr}\nu_r(1 - L_r)\mu_r \Big/ \sum_r \sum_l A_{ljr}\nu_r(1 - L_r)$$

and the functions E_1, E_2 are defined by (5.1) and (5.2). Here L_r is the approximation for the loss probability on route r, and μ_j is the approximate mean holding period for a typical circuit on link j. Preliminary theoretical and numerical investigations suggest that this model becomes increasingly accurate in networks with diverse routing as capacities and offered traffics are increased together, with ratios C_j/ν_r held fixed.

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

References

[1] R. G. Ackerley, "Hysteresis-type behaviour in networks with extensive overflow", *Br. Telecom Technol. J.* **5** (1987).

[2] D. Y. Burman, J. P. Lehoczky and Y. Lim, "Insensitivity of blocking probabilities in a circuit switching network", *J. Appl. Prob.* **21** (1984) 850–859.

[3] R. B. Cooper and S. S. Katz, "Analysis of alternate routing networks account taken of the nonrandomness of overflow traffic", Bell Laboratories (1964).

[4] G. A. Cope, "Data structures for descriptions of routing strategies in circuit-switched networks and efficient evaluation of implied costs", Performance Engineering Division, British Telecom Research Laboratories (1988).

[5] G. A. Cope and F. P. Kelly, "The use of implied costs for dimensioning and routing", report prepared by Stochastic Networks Group, Cambridge, for British Telecom Research Laboratories (1986).

[6] R. J. Gibbens, *Dynamic Routing in Circuit-switched Networks: the Dynamic Alternative Routing Strategy*, Ph. D. thesis, University of Cambridge (1988).

[7] R. J. Gibbens, P. J. Hunt and F. P. Kelly, "Bistability in communication networks", *Festschrift for J. M. Hammersley* (to appear).

[8] R. J. Gibbens, F. P. Kelly and P. B. Key, "Dynamic Alternative Routing—modelling and behaviour", *Proc. 12th Int. Teletraffic Cong., Turin* (1988), Ed. M. Bonatti (Elsevier, Amsterdam).

[9] R. J. Gibbens and P. A. Whiting, "An investigation of the accuracy of implied cost methods for circuit-switched network optimization", 5th UK Teletraffic Symposium, Aston (1988).

[10] P. J. Hunt, "Implied costs in loss networks", *Adv. Appl. Prob.* **21** (1989) (to appear).

[11] P. J. Hunt and F. P. Kelly, "On critically loaded loss networks", *Adv. Appl. Prob.* **21** (1989) (to appear).

[12] F. P. Kelly, *Reversibility and Stochastic Networks*, (Wiley, Chichester, 1979).

[13] F. P. Kelly, "Blocking probabilities in large circuit-switched networks", *Adv. Appl. Probab.* **18** (1986) 473–505.

[14] F. P. Kelly, "One-dimensional circuit-switched networks", *Ann. Prob.* **15** (1987) 1166–1179.

[15] F. P. Kelly, "Routing in circuit-switched networks: optimization, shadow prices and decentralization", *Adv. Appl. Prob.* **20** (1988) 112–144.

[16] F. P. Kelly, "Routing and capacity allocation in networks with trunk reservation" (submitted).

[17] P. B. Key, "Implied cost methodology and software tools for a fully connected network with DAR and trunk reservation", *Br. Telecom Technol. J.* **6** (1988) 52–65.

[18] P. B. Key and M. J. Whitehead, "Cost-effective use of networks employing Dynamic Alternative Routing", *Proc. 12th Int. Teletraffic Cong., Turin* (1988), Ed. M. Bonatti (Elsevier, Amsterdam).

[19] R. S. Krupp, "Stabilization of alternate routing networks", IEEE International Communications Conference, Philadelphia (1982).

[20] P. M. Lin, B. J. Leon and C. R. Stewart, "Analysis of circuit-switched networks employing originating-office control with spill-forward", *IEEE Trans. Comm.* **26** (1978) 754–765.

[21] Y. Nakagomi and H. Mori, "Flexible routing in the global communication network", 7th International Teletraffic Congress (1973).

[22] D. J. Songhurst, "Protection against traffic overload in hierarchical networks employing alternative routing", Telecommunications Networks Planning Symposium, Paris (1980).

[23] R. R. Stacey and D. J. Songhurst, "Dynamic Alternative Routing in the British Telecom trunk network", International Switching Symposium, Phoenix (1987).

[24] W. Whitt, "Blocking when service is required from several facilities simultaneously", *A. T. & T. Tech. J.* **64** (1985) 1807–1856.

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press

[25] P. Whittle, "Approximation in large-scale circuit-switched networks", *Prob. Eng. Inf. Sci.* **2** (1988) 279–291.
[26] R. I. Wilkinson, "Theory for toll traffic engineering in the USA", *Bell Syst. Tech. J.* **35** (1956) 421–513.
[27] I. B. Ziedins, "Quasi-stationary distributions and one-dimensional circuit-switched networks", *J. Appl. Prob.* **24** (1987) 965–977.
[28] I. B. Ziedins and F. P. Kelly, "Limit theorems for loss networks with diverse routing", *Adv. Appl. Prob.* **21** (1989) (to appear).

https://doi.org/10.1017/S0334270000006597 Published online by Cambridge University Press