PUBLISHED VERSION

Bean, Nigel Geoffrey ; Taylor, P. G.
Maximal profit dimensioning and tariffing of loss networks, *Probability in the Engineering and Informational Sciences*, 1995; 9(3):323-340.

Copyright © 1995 Cambridge University Press

PERMISSIONS

http://journals.cambridge.org/action/stream?pageId=4088&level=2#4408

The right to post the definitive version of the contribution as published at Cambridge Journals Online (in PDF or HTML form) in the Institutional Repository of the institution in which they worked at the time the paper was first submitted, or (for appropriate journals) in PubMed Central or UK PubMed Central, no sooner than one year after first publication of the paper in the journal, subject to file availability and provided the posting includes a prominent statement of the full bibliographical details, a copyright notice in the name of the copyright holder (Cambridge University Press or the sponsoring Society, as appropriate), and a link to the online edition of the journal at Cambridge Journals Online. Inclusion of this definitive version after one year in Institutional Repositories outside of the institution in which the contributor worked at the time the paper was first submitted will be subject to the additional permission of Cambridge University Press (not to be unreasonably withheld).

10th December 2010

http://hdl.handle.net/2440/613

Probability in the Engineering and Informational Sciences, 9, 1995, 323–340. Printed in the U.S.A.

MAXIMAL PROFIT DIMENSIONING AND TARIFFING OF LOSS NETWORKS

N. G. BEAN AND P. G. TAYLOR
Department of Applied Mathematics
University of Adelaide
S.A. 5005, Australia

In this paper we present a unified approach to the optimal dimensioning and tariffing of loss networks. In our formulation the optimum is chosen to maximize the profit for the company operating the loss network. We assume that the operating company has the flexibility to determine tariffs and grade of service—although both of these can possibly be subject to regulatory constraints. The fact that the tariffing may affect demand and, hence, the dimensioning makes it essential that the operating company include the tariff/demand trade-off in determining the optimal way to dimension the loss network. A consequence of our formulation is that the optimal tariff structure has a particularly simple form, with the optimal tariff on a particular route separating into a term related to the tariff/demand trade-off on that route and a term that reflects the cost of the circuits used by the route.

1. INTRODUCTION

Loss networks have been used to model many systems in which users of different types arrive at sets of resources and try to access one or more of the resources from some of the sets. If the required resources are not available, one or more alternative sets may be tried, but ultimately a user whose request cannot be satisfied is lost from the system. Among systems well modelled by loss

Research supported by Australian Research Council grant A69132151.

© 1995 Cambridge University Press 0269-9648/95 $11.00 + .10

networks are local area networks, multiprocessor architectures, database structures, cellular mobile phone networks, computer communication networks, and, of course, circuit-switched telephone networks, from which the terminology used to describe loss networks is taken (see Kelly [12] and references therein).

The issue of how a loss network should be dimensioned has long been of interest and much has been written about it. A good summary of the field was given by Girard [5]. Early work was due to Truitt [16] and Pratt [15], who discussed the dimensioning of loss networks under hierarchical routing. Katz [9], Berry [2], Ash, Cardwell and Murray [1], and Girard and Liau [6] have presented methods for dimensioning networks under various types of nonhierarchical alternative routing.

Each of these models dimensions a network by attempting to minimize its cost, provided certain grade of service constraints are satisfied. Dimensioning can be considered to proceed in a manner appropriate to a company operating in a "public utility" framework, that is, a company with a charter to provide a given service to the public in as efficient a way as possible. However, in the increasingly competitive environment that is faced by modern operating companies, it seems more appropriate that networks should be dimensioned according to a principle first put forward by Moe in the late 1920s (see Girard [5] or Jensen [8])—that networks should be dimensioned such that the difference between the revenue generated by servicing users and the cost of providing resources (i.e., the profit generated by the network) is maximized.

This approach was used for dimensioning loss networks by Kelly [10–12], and a similar concept was used by Low and Varaiya [14] with respect to service provision in ATM networks. Kelly's work focussed on perturbations to a given network. He studied ways to calculate the incremental cost to the network of accepting an extra call on a particular link and the surplus value gained from accepting a new call on a given route. This approach is admirably suited to the study of optimal routing of calls and optimal upgrades for a preexisting network but is less useful for dimensioning a new network, to be built from scratch. This is the problem that we examine in this paper: to formulate the problem of dimensioning a "greenfield" loss network in such a way as to maximize the profit generated for the operating company.

We incorporate a further feature that, to our knowledge, has not appeared in the literature on loss networks. We take into account the fact that there is a relationship between the tariff charged to users and the traffic offered to the network. This relationship is encapsulated in a traffic elasticity function, which, in general, assumes that offered traffic is a decreasing function of tariff and whose specific form could reflect market factors such as a competitor's tariff structure and, indeed, the grade of service offered by the network.

The paper is organized as follows. In Section 2 we define our model and discuss some basic results from the literature. In Section 3 we discuss how a network with fixed routing can be optimally dimensioned and finish with a discussion of how our method can be extended to cover networks with alternate routing. Section 4 contains a numerical example. We present a further investigation of our model in Section 5. In that section we show that, under reasonable assumptions, the optimal tariff for a particular route splits into a term related to the cost of carrying a call on that route plus a term related to the traffic elasticity function. Our conclusions are stated in Section 6.

2. THE MODEL

In this section we introduce the classical loss network operating under fixed routing and present both exact and approximate approaches for analysis. We follow the notation of Kelly [12]. The standard example of this model is a circuit-switched telephone network. Accordingly, we use the terms calls, links, and circuits. The model also arises naturally in many other types of communication networks. Generalizations to dynamic and alternative routing are possible.

Consider a loss network where the set of links is labelled $\mathcal{J}$ and the set of routes is labelled $\mathcal{R}$. A route is considered to be a collection of links (not necessarily connected) and so $r \subset \mathcal{J}$, $r \in \mathcal{R}$. Suppose that link j comprises C_j circuits, for all $j \in \mathcal{J}$. A call on route $r \in \mathcal{R}$ uses A_{jr} circuits from link j, where $A_{jr} \in \mathbb{Z}_+$. 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 $j \in \mathcal{J}$ there are fewer than A_{jr} free circuits. Otherwise, the call is connected and simultaneously holds A_{jr} circuits from link $j \in \mathcal{J}$ for the holding time of the call. If the call is connected, it is charged α_r units. The holding times of calls on route $r \in \mathcal{R}$ are identically distributed with unit mean, and holding times are independent of all earlier arrival times and holding times. The focus in this paper is the optimal choice of $C_j, j \in \mathcal{J}$ and $\alpha_r, r \in \mathcal{R}$, the variables over which the operating company will usually have some control.

Let $n_r(t)$ be the number of calls in progress at time t on route r and define the vectors $\mathbf{n}(t) = (n_r(t), r \in \mathcal{R})$ and $\mathbf{C} = (C_j, j \in \mathcal{J})$. Then the stochastic process $(\mathbf{n}(t), t \geq 0)$ has a unique stationary distribution $\pi(\cdot)$ given by

$$\pi(\mathbf{n}) = G(\mathbf{C})^{-1} \prod_{r \in \mathcal{R}} \frac{\nu_r^{n_r}}{n_r!}, \qquad \mathbf{n} \in \mathcal{S}(\mathbf{C}), \tag{2.1}$$

where

$$\mathcal{S}(\mathbf{C}) = \{\mathbf{n} \in \mathbb{Z}_+^{\mathcal{R}} : A\mathbf{n} \leq \mathbf{C}\} \tag{2.2}$$

and $G(\mathbf{C})$ is the normalizing constant

$$G(\mathbf{C}) = \left(\sum_{n \in \mathcal{S}(\mathbf{C})} \prod_{r \in \mathcal{R}} \frac{\nu_r^{n_r}}{n_r!} \right). \tag{2.3}$$

This result is easy to check when call holding times are exponentially distributed because the stochastic process is a reversible Markov process and for general distributions the result holds from the theory of insensitivity (see, e.g., Burman, Lehoczky, and Lim [3]).

The performance measure of interest in such networks is the probability that a call on route $r \in \mathcal{R}$ will be blocked, known as the blocking probability for route r and denoted $B_r, r \in \mathcal{R}$. Let $\mathcal{B}_r(\mathbf{C})$ represent the set of states for which a call of type r is blocked. Then, for each $r \in \mathcal{R}$,

$$B_r = \sum_{\mathbf{n} \in \mathcal{B}_r(\mathbf{C})} \pi(\mathbf{n}). \tag{2.4}$$

The equilibrium distribution $\pi(\cdot)$ of Eq. (2.1) is of a very simple form. However, the determination of the normalizing constant $G(\mathbf{C})$, defined by Eq. (2.3), is difficult and has recently been shown by Louth, Mitzenmacher, and Kelly [13] to be #P-complete in the number of distinct routes. Accordingly, some method of approximation must be used, and the usual method for approximating the blocking probabilities is the Erlang fixed-point technique [12].

Let $E_j, j \in \mathcal{J}$ be the unique solution to the equations

$$E_j = E(\rho_j, C_j), \tag{2.5}$$

where

$$\rho_j = \sum_{r:j \in r} \nu_r (1 - E_j)^{-1} \prod_{i \in r} (1 - E_i)^{A_{ir}} \tag{2.6}$$

and the function E is Erlang's formula

$$E(\rho, C) = \frac{\rho^C}{C!} \left[\sum_{l=0}^{C} \frac{\rho^l}{l!} \right]^{-1}, \qquad j \in \mathcal{J}. \tag{2.7}$$

Then the vector $(E_j, j \in \mathcal{J})$ is called the *Erlang fixed point*, and an approximation for the loss probability on route r is given by

$$B_r \approx 1 - \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}}, \tag{2.8}$$

where the link blocking probabilities are now sufficient to describe the grades of service.

The idea behind this approximation is very simple. Suppose that a Poisson stream of rate ν_r is thinned randomly by a factor of $(1 - E_i)^{A_{ir}}$ at each link on route r other than link j and by $(1 - E_j)^{A_{jr}-1}$ at link j. If these thinnings were independent over both the links and the routes (clearly this is not true), then the traffic offered to link j would be Poisson at rate (2.6), the link blocking probability would be given by Eq. (2.5), and the loss probability on route r would satisfy Eq. (2.8) exactly. We call Eq. (2.6) the *reduced load* on link j.

At this stage we have assumed that calls on route r arrive as a Poisson process of rate ν_r and pay a tariff of α_r if accepted. However, in reality the traffic offered to route r is likely to depend on both the tariff α_r and the route blocking probability B_r. Thus, henceforth we treat the arrival rate ν_r as a function of both the tariff and the grade of service on route r; that is,

$$\nu_r = \nu_r(\alpha_r, B_r), \qquad r \in \mathcal{R}. \tag{2.9}$$

3. OPTIMIZING PROFIT

If the network provider is to install capacity on a link, then there will be an associated cost to the provider. This may be in the form of a charge per unit time if the capacity is leased or it may involve a capital outlay. In the latter case, the installation will have an economic lifetime and the capital outlay must be written off over this period of time. In such a way the provider can again describe the cost as a charge per unit time. Let the charge per unit time of installing capacity C_j on link $j \in \mathcal{J}$ be given by $\beta_j(C_j)$ depending both on the nature of link j and on the amount of capacity required. In a telecommunications environment, for example, it will normally be the case that the charge $\beta_j(C)$ will be made up of three components: one that is independent of j and C, representing the provision of space costs at either end of the link; another that is independent of C, representing the cost of accessing the corridor for link j depending on the length of the link and the terrain; and finally a component that does not depend on the link but depends only on the capacity, representing the terminal equipment. Thus, $\beta_j(C)$ will have the form

$$\beta_j(C) = \gamma + \delta_j + \epsilon C. \tag{3.1}$$

However, different functions $\beta_j(C)$ are possible.

The traditional method of dimensioning loss networks has involved the provider attempting to minimize the cost of provision, that is, minimize

$$\sum_{j \in \mathcal{J}} \beta_j(C_j), \tag{3.2}$$

subject to certain regulatory constraints. This approach is probably due to the fact that in the recent past most operating companies were public utilities. In the new era of private ownership and competition, the providers must dimension to maximize profit, that is, maximize the difference between revenue and cost. Also, the companies can now alter their tariffs and perhaps offer different grades of service. All this points to a new optimization problem with new constraints and a new objective function.

Recall that the probability that a call on route $r \in \mathcal{R}$ is blocked is given by B_r. Therefore, the rate that calls are accepted into the network, when the tariff charged is α_r and the blocking probability is B_r, is given by $\nu_r(\alpha_r, B_r)(1 - B_r)$. Accordingly, the network can expect to receive revenue on route $r \in \mathcal{R}$ at a rate of $\alpha_r \nu_r(\alpha_r, B_r)(1 - B_r)$ per unit time. Thus, the total expected revenue per unit time is given by

$$\sum_{r \in \mathcal{R}} \alpha_r \nu_r(\alpha_r, B_r)(1 - B_r). \tag{3.3}$$

The network provider wishes to maximize its profit and so wishes to maximize its revenue minus its cost, that is,

$$\sum_{r\in\mathcal{R}} \alpha_r \nu_r(\alpha_r, B_r)(1 - B_r) - \sum_{j\in\mathcal{J}} \beta_j(C_j), \tag{3.4}$$

where the route blocking probabilities are related to the link capacities through Eq. (2.4) and the equilibrium distribution $\pi(\cdot)$ is given in Eq. (2.1). There may also be regulatory constraints that the tariff $\alpha_r, r \in \mathcal{R}$ charged on route r has to lie in the interval $[\underline{\alpha_r}, \overline{\alpha_r}]$ and the blocking probability $B_r, r \in \mathcal{R}$ on route r also has to lie in some interval $[\underline{B_r}, \overline{B_r}]$.

Written formally, the optimization problem is as follows.

3.1. Formulation 1

Variables: $\alpha_r, r \in \mathcal{R},$ (3.5)

$C_j, j \in \mathcal{J}.$ (3.6)

Objective: $\max \sum_{r\in\mathcal{R}} \alpha_r \nu_r(\alpha_r, B_r)(1 - B_r) - \sum_{j\in\mathcal{J}} \beta_j(C_j).$ (3.7)

Constraints: $\underline{\alpha_r} \le \alpha_r \le \overline{\alpha_r},\ r \in \mathcal{R},$ (3.8)

$0 \le C_j < \infty,\ j \in \mathcal{J},$ (3.9)

$\underline{B_r} \le B_r \le \overline{B_r},\ r \in \mathcal{R},$ (3.10)

where $B_r = \sum_{\mathbf{n}\in\mathcal{B}_r(\mathbf{C})} \pi(\mathbf{n}),\ r \in \mathcal{R}$ (3.11)

and $\pi(\mathbf{n}) = G(\mathbf{C})^{-1} \prod_{r\in\mathcal{R}} \frac{\nu_r(\alpha_r, B_r)^{n_r}}{n_r!},\ \mathbf{n} \in \mathcal{S}(\mathbf{C}).$ (3.12)

Finding the solution of this optimization problem is an extremely difficult task. As an illustration, consider the problem of finding an initial feasible point, that is, a set of route tariffs and link capacities that obeys all the constraints. First, the solution to the set of Eqs. (3.12) and (3.11) needs to be found. Due to the fact that the arrival rate, $\nu_r(\alpha_r, B_r)$, may depend on the grade of service, B_r, this could become a complex process involving iteration to determine $B_r, r \in \mathcal{R}$ and, hence, $\nu_r, r \in \mathcal{R}$. However, each iteration requires that the exact equilibrium distribution be found. This is itself a computationally intensive task, and, as we mentioned before, the determination of the normalizing constant, $G(\mathbf{C})$, has been shown to be #P-complete in the number of distinct routes. Once this iterative procedure has converged, constraint (3.10) must be checked to see whether or not it is obeyed. Clearly, constraint (3.8) will be obeyed if the tariffs have been chosen sensibly and constraint (3.9) is a natural bound presented only for completeness.

It would be a great advantage if the #P-complete procedure of solving for the equilibrium distribution and, hence, the blocking probabilities could be avoided. Next, we take advantage of the Erlang fixed point and so no longer need to determine the entire equilibrium distribution. The penalty for this added convenience is that the results are no longer exact.

Because B_r is (approximately) determined by $\mathbf{E} = (E_j, j \in \mathcal{J})$ through Eq. (2.8), we can rewrite the expected revenue as

$$\sum_{r \in \mathcal{R}} \alpha_r \nu_r(\alpha_r, \mathbf{E}) \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}}. \tag{3.13}$$

The network provider wishes to maximize its profit and so wishes to maximize its revenue minus its cost, that is,

$$\sum_{r \in \mathcal{R}} \alpha_r \nu_r(\alpha_r, E) \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}} - \sum_{j \in \mathcal{J}} \beta_j(C_j), \tag{3.14}$$

where the link blocking probabilities are a function of the link capacities through Eq. (2.5). The constraints are the same as above except that the blocking probability $B_r, r \in \mathcal{R}$ is now written as $B_r = 1 - \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}}$. The Eqs. (3.12) and (3.11) are also replaced by the Eqs. (2.5) and (2.6) which determine E_j as a function of C_k, $k \in \mathcal{J}$.

Written formally, the optimisation problem is as follows:

3.2. Formulation 2

Variables: $\alpha_r, r \in \mathcal{R},$ (3.15)

$C_j, j \in \mathcal{J}.$ (3.16)

Objective: $\max \sum_{r \in \mathcal{R}} \alpha_r \nu_r(\alpha_r, \mathbf{E}) \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}} - \sum_{j \in \mathcal{J}} \beta_j(C_j).$ (3.17)

Constraints: $\underline{\alpha_r} \leq \alpha_r \leq \overline{\alpha_r}, r \in \mathcal{R},$ (3.18)

$0 \leq C_j < \infty, j \in \mathcal{J},$ (3.19)

$\underline{B_r} \leq 1 - \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}} \leq \overline{B_r}, r \in \mathcal{R},$ (3.20)

where $E_j = E(\rho_j, C_j), j \in \mathcal{J},$ (3.21)

and $\rho_j = \sum_{r: j \in r} \nu_r(\alpha_r, \mathbf{E})(1 - E_j)^{-1} \prod_{i \in r} (1 - E_i)^{A_{ir}}, j \in \mathcal{J}.$ (3.22)

Observe that it is not that much easier to solve Formulation 2 than Formulation 1. First, the solution to constraints (3.21) and (3.22) needs to be found. This is a complex procedure involving a network-wide iteration. Then, constraint (3.20) must be checked to see whether or not it is obeyed. Clearly, constraint (3.18) will be obeyed if the tariffs at this point have been chosen sensibly and constraint (3.19) is a natural bound presented purely for completeness. Finally, the objective function can be evaluated.

It would be a great advantage if a procedure could be found that did not involve the solution of a complex set of equations (such as Eqs. (3.11) and (3.12) or Eqs. (3.21) and 3.22)) just to determine feasibility. Formulation 3, which uses link blocking probabilities rather than link capacities as variables, contains such a procedure. In a slightly different context, Girard [5, p. 382] presented an approach that also used link blocking probabilities as the variables.

For a fixed value of $\rho_j \geq 0$, the function $E: \mathbb{Z}_+ \to [0,1]$, defined in Eq. (2.7), is a one-to-one and strictly decreasing function. If we extend the definition so that the domain of E is $\mathbb{R}_+$ by linearly interpolating between integer points, then we can define an inverse function $C: [0,1] \to \mathbb{R}_+$ such that

$$C_j = C(\rho_j, E_j), \qquad j \in \mathcal{J}. \tag{3.23}$$

C_j is then the capacity required in an Erlang loss system to accommodate an offered traffic of ρ_j with a blocking probability of E_j. Clearly, the function $C(\cdot,\cdot)$ no longer has a nice analytical form, but the advantages of its use in dimensioning are compelling.

Using the function C, the optimization problem is as follows.

3.3. Formulation 3

Variables: $\alpha_r, r \in \mathcal{R}$, (3.24)

$E_j, j \in \mathcal{J}$. (3.25)

Objective: $$\max \sum_{r \in \mathcal{R}} \alpha_r \nu_r(\alpha_r, \mathbf{E}) \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}} - \sum_{j \in \mathcal{J}} \beta_j(C_j). \tag{3.26}$$

Constraints: $\underline{\alpha_r} \leq \alpha_r \leq \overline{\alpha_r},\ r \in \mathcal{R}$, (3.27)

$0 \leq E_j \leq 1,\ j \in \mathcal{J}$, (3.28)

$$\underline{B_r} \leq 1 - \prod_{j \in \mathcal{J}} (1 - E_j)^{A_{jr}} \leq \overline{B_r},\ r \in \mathcal{R}, \tag{3.29}$$

where $C_j = C(\rho_j, E_j),\ j \in \mathcal{J}$, (3.30)

$$\text{and } \rho_j = \sum_{r: j \in r} \nu_r(\alpha_r, \mathbf{E})(1 - E_j)^{-1} \prod_{i \in r} (1 - E_i)^{A_{ir}},\ j \in \mathcal{J}. \tag{3.31}$$

The importance of this formulation is that Eqs. (3.30) and (3.31) are not required to determine feasibility but merely allow a neat statement of the objective function.

The tariffs, $(\alpha_r, r \in \mathcal{R})$, and link blocking probabilities, $(E_j, j \in \mathcal{J})$, are exactly the variables in the optimization problem and can be chosen so that constraints (3.27)–(3.29) are obeyed. Hence, $\rho_j, j \in \mathcal{J}$ is available immediately. This, in turn, simply provides the required link capacities by using the functions $C(\rho_j, E_j), j \in \mathcal{J}$, and so the objective function can be simply evaluated.

It may at first seem that the complexity has been hidden from view in the definition of the function $C(\cdot,\cdot)$ defined implicitly as the inverse of Erlang's function; however, this is not the case. Recall that Erlang's function is defined in Eq. (2.7) to be

$$E(\rho,C) = \frac{\rho^C}{C!}\left[\sum_{l=0}^{C} \frac{\rho^l}{l!}\right]^{-1}. \tag{3.32}$$

This is usually evaluated using the iterative form

$$E(\rho,C) = \frac{\rho E(\rho,C-1)}{C + \rho E(\rho,C-1)}, \tag{3.33}$$

with the boundary condition $E(\rho,0) = 1$. Accordingly, exactly C steps are required to evaluate $E(\rho,C)$. Recall also that $E(\rho,C)$ is strictly decreasing in C, and so the inverse function $C(\rho,E)$ could be found using any of the normal iterative techniques, involving evaluations of $E(\rho,C)$ for different values of C. However, a more efficient procedure is simply to use recursion (3.33) until $E(\rho,C)$ is less than the desired blocking probability, followed by linear interpolation. Therefore, the complexity of evaluation of $C_j = C(\rho,E_j)$ is equivalent to the complexity of evaluation of $E_j = E(\rho,C_j)$.

A further advantage of the approach of using the link blocking probabilities as the variables of the optimization procedure (along with the tariffs) is that the constraints under this formulation can all be made to be linear via a simple transformation (as in Kelly [12, Sect. 3]). The theoretical and computational advantages of having linear rather than nonlinear constraints are well known.

This can be achieved by setting

$$y_j = -\log(1 - E_j), j \in \mathcal{J} \tag{3.34}$$

so that

$$E_j = 1 - \exp(-y_j), j \in \mathcal{J}. \tag{3.35}$$

Finally, write $\mathbf{y} = (y_j, j \in \mathcal{J})$ and rewrite $\nu_r(\alpha_r,\mathbf{E})$ as $\nu_r(\alpha_r,\mathbf{y})$.

Then Formulation 3 can be written in the following way, where constraint (3.41) is now a linear constraint.

3.4. Formulation 4

Variables: $\alpha_r, r \in \mathcal{R},$ (3.36)

$y_j, j \in \mathcal{J}.$ (3.37)

Objective: $$\max \sum_{r\in\mathcal{R}} \alpha_r \nu_r(\alpha_r,\mathbf{y}) \exp\left(-\sum_{j\in\mathcal{J}} y_j A_{jr}\right) - \sum_{j\in\mathcal{J}} \beta_j(C_j). \tag{3.38}$$

Constraints: $\underline{\alpha_r} \le \alpha_r \le \overline{\alpha_r},\ r \in \mathcal{R},$ (3.39)

$$0 \le y_j \le \infty,\ j \in \mathcal{J}, \tag{3.40}$$

$$-\log(1 - \underline{B_r}) \le \sum_{j \in \mathcal{J}} A_{jr} y_j \le -\log(1 - \overline{B_r}),\ r \in \mathcal{R}, \tag{3.41}$$

$$\text{where } C_j = C(\rho_j, E_j),\ j \in \mathcal{J}, \tag{3.42}$$

$$\text{and } \rho_j = \sum_{r:j \in r} \nu_r(\alpha_r, y)\exp(y_j)\exp\left(-\sum_{j \in \mathcal{J}} y_j A_{jr}\right),\ j \in \mathcal{J}, \tag{3.43}$$

Up to this point we have only discussed networks with fixed routing—that is, networks where each route has exactly one path through the network associated with it. In the case of alternative routing networks, more than one path is associated with each route. Usually there is one path that is regarded as a *first choice path*, or *direct path*, and any number of alternative or *indirect paths*. The methods of choice of alternative path discussed in the literature vary from very complex decisions using information gathered from all over the network to simple decentralized decisions. They also vary from the fully *dynamic* decision relying on exact and up-to-date information on the state of the network through to *static* decisions that are independent of the state of the network.

Using a reduced load of the preceding form, it is clear that an approach similar to that above can be used with the link capacities as the variables in the dimensioning part of the optimization procedure. This, however, will suffer from the same inefficiencies as did Formulation 2. For certain forms of alternative routing, it is possible to use the link blocking probabilities as variables in the optimization procedure, as in Formulations 3 and 4. This will provide similar benefits to those described there. Usually, however, extra information will be required, such as the probability that alternatively routed calls will be blocked on each link (due, perhaps, to a trunk reservation scheme). These also become variables in the optimization procedure. This will have the added advantage of optimally choosing the trunk reservation parameters to maximize profit for the operating company. For a detailed discussion of reduced load approximations as applied to state-dependent alternative routing schemes, see Chung, Kashper, and Ross [4].

4. A FIXED ROUTING EXAMPLE

Consider the network shown in Figure 1 operating under fixed routing. The network could represent a possible backbone network in Australia. Let each origin–destination pair correspond to a route and so the network has 9 nodes, 10 links, and 36 routes. For each route, the path that a call will take is given by the shortest possible path (in the sense of the number of links) and the matrix A_{jr} is so defined.

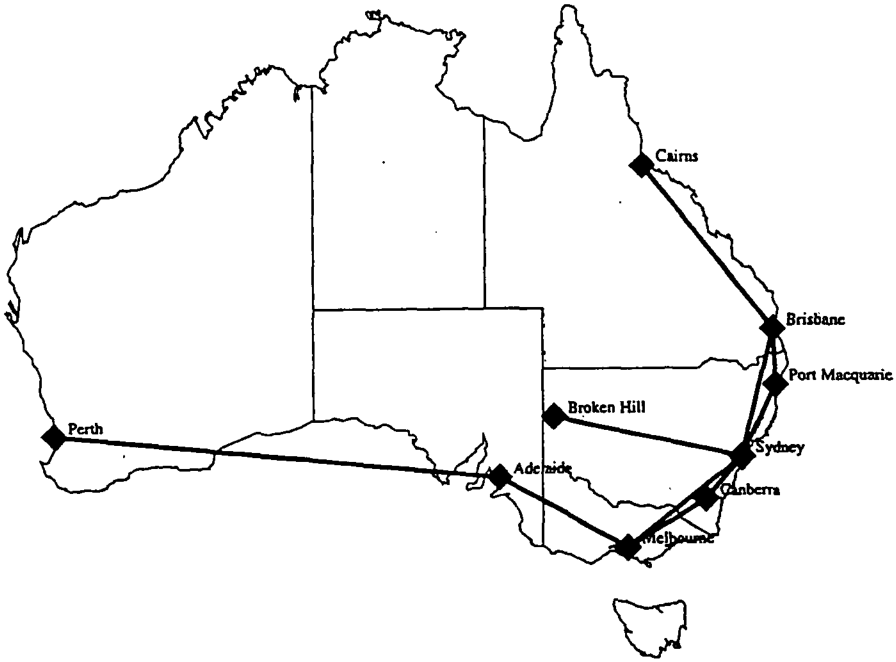

FIGURE 1. An example loss network.

Assume that the regulator has imposed the constraint that the route blocking probability on all routes must be bounded above by 0.01. That is, $\overline{B_r} = 0.01$, $r \in \mathcal{R}$. No constraint has been imposed on the lower bound so $\underline{B_r} = 0, r \in \mathcal{R}$. Assume, further, that the regulator has imposed an upper bound of 6 on the tariff for all routes. That is, $\overline{\alpha_r} = 6$, $\underline{\alpha_r} = 0$, $r \in \mathcal{R}$. Note that for all ρ, $C > 0$, it follows that $1 > E(\rho, C) > 0$. This means that in Formulation 3 we usually assume that $0 < E_j < 1, j \in \mathcal{J}$. This has the effect in Formulation 4 of making $0 < y_j < \infty, j \in \mathcal{J}$. In practice, it usually pays to impose explicit bounds, and as $\overline{B_r} = 0.01$, $r \in \mathcal{R}$, the upper bound is imposed at $y_j \leq -\log(0.99) = 0.01005$, $j \in \mathcal{J}$. However, there is no such natural lower bound, and so we impose an artificial constraint of $y_j \geq 10^{-5}, j \in \mathcal{J}$, and so $E_j \geq 10^{-5}, j \in \mathcal{J}$, as well. The solution (presented later) indicates that this artificial constraint does not bite and so has no effect on the solution.

We must now define the form of the function $\nu_r(\alpha_r, B_r)$ that defines the traffic intensity on route $r, r \in \mathcal{R}$ given that the tariff is α_r and the route blocking probability is B_r. It is clear that $\alpha_r \nu_r(\alpha_r, B_r)$ must decay quite rapidly as $\alpha_r \to \infty$, for all $r \in \mathcal{R}$, as otherwise the reward for accepting one call will be so large that the network will be highly profitable if it provides only one circuit along route r and so accepts a call on route r very occasionally. Also, $\nu_r(\alpha_r, B_r)$

must be bounded above as $\alpha_r \to 0$, for all $r \in \mathcal{R}$, as otherwise the optimization procedure would have to evaluate the profit of a network offering infinite capacity on some routes. Economic models for such a function use a sigmoid shape, representing the fact that under competition there will usually be a marked reduction (increase) in demand if your tariff is more (less) expensive than your competitor's.

The following is artificial, and practical implementations would require research to be carried out to determine the form of the functions involved and the parameter values.

We propose that $\nu_r(\alpha_r, B_r)$ be the following function, for all $r \in \mathcal{R}$:

$$\nu_r(\alpha_r, B_r) = \begin{cases} \nu_r^* \exp(\alpha_r^* - \alpha_r), & \text{if } \alpha_r \geq \alpha_r^*, \\ \nu_r^*(2 - \exp(\alpha_r - \alpha_r^*)), & \text{otherwise.} \end{cases} \tag{4.1}$$

This function is clearly continuous and differentiable with respect to $\alpha_r \in (0, \infty)$. Note also, that $\nu_r(\alpha_r, B_r)$ is bounded above by $\nu_r^*(2 - \exp(-\alpha_r^*))$, and this bound is attained when $\alpha_r = 0$.

Therefore, for each route $r \in \mathcal{R}$, we need to define α_r^* (representing the tariff of the competitor and, hence, the point of inflection) and ν_r^* (the traffic demand if the tariff is α_r^*). For simplicity, let $\alpha_r^* = 1$, $r \in \mathcal{R}$. We still need to provide the values $\nu_r^*, r \in \mathcal{R}$ representing a base traffic demand. Using approximate subscriber figures for the appropriate regions, we have created a possible base traffic demand, as given in Table 1.

It now remains to define the cost per unit time of providing C_j circuits on link $j \in \mathcal{J}$, labelled $\beta_j(C_j)$. As described in Section 3, we assume that $\beta_j(\cdot)$ has a form given by

$$\beta_j(C_j) = \gamma + \delta_j + \epsilon C_j, \qquad \text{for all } j \in \mathcal{J}. \tag{4.2}$$

For simplicity, let $\epsilon = 1$, $\gamma = 100$, and $\delta_j = 100$, $j \in \mathcal{J}$.

To solve the optimization problem, the Erlang fixed point approximation must be used, as the network is far too large for an exact analysis. We used Formulation 4, where the blocking probabilities on each link, rather then the capacities on each link, are treated as variables in the optimization procedure and the constraints are all linear. A standard routine available in the NAG Library of routines found the optimal operating point, that is, the optimal tariffs to charge and the optimal dimensions to install. These are presented in Tables 1 and 2.

Note that with the form of traffic elasticity function and parameter values that we have used (remember they are artificial) it has turned out that capacity is expensive with respect to tariff. Thus, the optimal way to operate the network is to charge high tariffs on all routes (i.e., $\alpha_r > \alpha_r^*$, $r \in \mathcal{R}$), to reduce the traffic and, hence, the cost of provision of capacity. In general, tariffs are higher on longer routes, which makes sense, because longer routes use more resources.

TABLE 1. Table of Base Traffic Demands, Optimal Traffic Demands, and Optimal Tariffs

Route Label	Base Traffic ν_r^*	Optimal Traffic $\nu_r(\alpha_r, B_r)$	Tariff α_r
Adelaide-Brisbane	12,556.80	621.66	4.01
Adelaide-Broken Hill	8603.40	426.12	4.01
Adelaide-Cairns	7040.30	127.49	5.01
Adelaide-Canberra	677.70	90.03	3.02
Adelaide-Melbourne	38,632.80	14,182.20	2.00
Adelaide-Perth	9776.60	3586.71	2.00
Adelaide-Port Macquarie	8327.50	412.50	4.01
Adelaide-Sydney	25,750.90	3473.98	3.00
Brisbane-Broken Hill	10,764.60	1450.83	3.00
Brisbane-Cairns	8967.00	3281.41	2.01
Brisbane-Canberra	848.20	112.57	3.02
Brisbane-Melbourne	48,263.70	6507.85	3.00
Brisbane-Perth	12,232.00	222.80	5.01
Brisbane-Port Macquarie	10,419.70	3812.11	2.01
Brisbane-Sydney	32,191.00	11,813.95	2.00
Broken Hill-Cairns	6034.60	297.23	4.01
Broken Hill-Canberra	580.70	77.13	3.02
Broken Hill-Melbourne	33,143.90	4472.85	3.00
Broken Hill-Perth	8380.60	152.09	5.01
Broken Hill-Port Macquarie	7138.00	962.26	3.00
Broken Hill-Sydney	22,084.10	8111.23	2.00
Cairns-Canberra	475.00	23.08	4.02
Cairns-Melbourne	27,147.10	1339.55	4.01
Cairns-Perth	6858.00	46.21	6.00
Cairns-Port Macquarie	5840.90	781.72	3.01
Cairns-Sydney	18,081.20	2427.38	3.01
Canberra-Melbourne	2622.40	947.30	2.02
Canberra-Perth	660.00	32.07	4.02
Canberra-Port Macquarie	562.00	74.48	3.02
Canberra-Sydney	1744.00	633.11	2.01
Melbourne-Perth	37,638.50	5069.48	3.00
Melbourne-Port Macquarie	32,086.40	4326.67	3.00
Melbourne-Sydney	98,224.40	36,088.93	2.00
Perth-Port Macquarie	8111.90	147.45	5.01
Perth-Sydney	25,086.70	1241.96	4.01
Port Macquarie-Sydney	21,378.10	7845.37	2.00

TABLE 2. Table of Optimal Link Capacities and Link Blocking Probabilities

Link Label	Capacity C_j	Blocking Probability E_j
Adelaide-Melbourne	26,355.36	0.0020423
Adelaide-Perth	10,601.99	0.0021478
Brisbane-Cairns	8422.97	0.0021037
Brisbane-Port Macquarie	4629.00	0.0071315
Brisbane-Sydney	25,098.51	0.0020586
Broken Hill-Sydney	15,702.77	0.0038879
Canberra-Melbourne	1111.22	0.0058387
Canberra-Sydney	959.95	0.0058664
Melbourne-Sydney	56,699.47	0.0016877
Port Macquarie-Sydney	14,114.85	0.0041053

Consider the structure of the network shown in Figure 1. There are two clear classes of link, those that form the backbone of the network (Adelaide-Melbourne, Adelaide-Perth, Brisbane-Cairns, Brisbane-Sydney, Melbourne-Sydney) and those that represent side branches (Broken Hill-Sydney, Brisbane-Port Macquarie, Canberra-Melbourne, Canberra-Sydney, Port Macquarie-Sydney). Table 2 shows that this classification is repeated in the optimal allocation of link blocking probabilities; that is, the link blocking probabilities for the backbone links are uniformly less than the link blocking probabilities for the side branch links. Note that this is not a function of these links also being low-capacity links because the link Brisbane-Cairns has less capacity than the link Port Macquarie-Sydney.

5. OBSERVATIONS ABOUT OPTIMAL TARIFFING

The optimal tariffs for the example presented in Section 4 are very striking. It is clear that the tariff for each route is the number of links on the route plus one. This description depends, of course, on the functional forms of $\nu_r(\cdot,\cdot)$, $r \in \mathcal{R}$ and $\beta_j(\cdot), j \in \mathcal{J}$ and their parameters. However, by using Formulation 3, we can prove that a relationship of a similar form always exists. In fact, the optimal tariff always consists of a term representing the cost to the network of carrying the call plus a constant depending only on the elasticity function of the traffic, $\nu_r(\cdot,\cdot)$. First we require a few preliminary lemmas about the partial derivatives of Erlang's function, $E(\rho,C)$, which is defined in Eq. (2.7).

Many papers (e.g., Jagerman [7]) present an extended form of Erlang's function that is defined for complex-valued ρ and C and passes through the correct values at the positive integer lattice points. The proof of the following is given in Jagerman [7].

LEMMA 1:

$$\frac{\partial E(\rho,C)}{\partial \rho} = E(\rho,C)\left(\frac{C}{\rho} - 1 + E(\rho,C)\right). \tag{5.1}$$

The function $E(\rho,C)$ defined in Jagerman [7] is extremely hard to evaluate for noninteger C, and its partial derivative with respect to C is given only as an approximation. In Section 3 we extended the definition of $E(\rho,C)$ to noninteger C by using linear interpolation between its value at integer points. We again use this extension and therefore define the derivative to be the left derivative, in the knowledge that the derivative will not be continuous.

LEMMA 2: *For all $x \in (C^* - 1, C^*]$, $C^* \in \mathbb{Z}_+$, the partial derivative with respect to C is given by*

$$\left.\frac{\partial E(\rho,C)}{\partial C}\right|_{C=x} = E(\rho,C^*) - E(\rho,C^* - 1), \tag{5.2}$$

$$= E(\rho,C^*) - \frac{C^* E(\rho,C^*)}{\rho(1 - E(\rho,C^*))}. \tag{5.3}$$

PROOF: The first equality is a simple consequence of the definition of Erlang's function for real-valued C. The second equality can easily be proved by using the well-known recursion for Erlang's function given in Eq. (3.33). ■

Recall that in Section 3 we defined the function $C(\rho,E)$ as the inverse of Erlang's function for fixed ρ. That is, it is defined implicitly by

$$E(\rho,C(\rho,E)) = E. \tag{5.4}$$

We can now use the preceding two lemmas and this equation to determine the partial derivative of $C(\rho,E)$ with respect to ρ. This partial derivative is again a left derivative and will not be continuous with respect to ρ.

LEMMA 3:

$$\frac{\partial C(\rho,E)}{\partial \rho} = \frac{E(1 - E) - E\dfrac{C(\rho,E)}{\rho}}{E(\rho,C^*) - \dfrac{C^* E(\rho,C^*)}{\rho(1 - E(\rho,C^*))}}, \tag{5.5}$$

$$= 1 - E, \qquad \text{as } E(\rho,C^*) \to E, \tag{5.6}$$

where $C^* = \lceil C(\rho,E) \rceil$, the smallest integer greater than or equal to $C(\rho,E)$.

PROOF: The first equality is a simple consequence of differentiating Eq. (5.4) with respect to ρ and using the results of the previous two lemmas. The limit result is again a result of simple manipulation. ■

This limit result is a good approximation whenever E is small. This is because, in this situation, $E(\rho, C^*) \approx E$. The limit result is now intuitively obvious. Consider two Erlang loss systems offered traffic at rate ρ and $\rho + 1$, respectively, where both systems must achieve a loss probability of E. The second system is offered an extra Erlang of traffic but needs to carry only $1 - E$ extra Erlangs of traffic. Because the systems are very nearly deterministic (ρ, C are usually large and E is small), the second system needs an extra $1 - E$ capacity to carry the extra $1 - E$ Erlangs of traffic.

We are now able to prove the result about the optimal choice of tariff.

THEOREM 4: *For fixed link blocking probabilities* $\mathbf{E}$, *the optimal tariff*, $\hat{\alpha}_r$, *so that network profit is maximized, when the tariff is unconstrained, satisfies the following equation in* α_r:

$$\alpha_r = \sum_{j \in r} \frac{d\beta_j(C(\rho_j, E_j))}{dC(\rho_j, E_j)} \frac{\partial C(\rho_j, E_j)}{\partial \rho_j} \frac{1}{1 - E_j} - \nu_r(\alpha_r, \mathbf{E}) \left[\frac{\partial \nu_r(\alpha_r, \mathbf{E})}{\partial \alpha_r} \right]^{-1}. \quad (5.7)$$

PROOF: Simple differential calculus applied to Eq. (3.26) is sufficient to find the location of this extremal point. ∎

The summand on the right-hand side of Eq. (5.7) is the rate of change of the cost of capacity on link j with respect to the traffic of route r and the summation is over all the links used by route r. The theorem thus explains the optimal tariff as the marginal cost to the network of carrying a call plus a term depending on the elasticity function. This result could also be used to simplify the optimization procedure so that the variables are just the $E_j, j \in \mathcal{J}$, as the optimal tariff can now be determined for any set of link blocking probabilities $\mathbf{E}$. However, this is not a simple task, as each ρ_j depends on possibly all the tariffs $\alpha_r, r \in \mathcal{R}$, and so all the equations are dependent on each other.

For the case where $E_j, j \in \mathcal{J}$ are small, an approximate result can be used that removes this dependence and so could greatly assist with the optimization procedure. We present this in the following corollary.

COROLLARY 5: *If* $E_j, j \in \mathcal{J}$ *are small, then* $\hat{\alpha}_r$ *is approximately the unique solution to the following equation in* α_r,

$$\alpha_r = \sum_{j \in r} \frac{d\beta_j(C_j(\rho_j, E_j))}{dC(\rho_j, E_j)} - \nu_r(\alpha_r, \mathbf{E}) \left[\frac{\partial \nu_r(\alpha_r, \mathbf{E})}{\partial \alpha_r} \right]^{-1}. \quad (5.8)$$

PROOF: Lemma 3 shows that $\partial C(\rho_j, E_j)/\partial \rho_j \approx 1 - E_j$ when E_j is small. Therefore, the fact that α_r satisfies Eq. (5.8) is a simple consequence of Lemma 3 and the previous theorem. Uniqueness follows from an argument using the sigmoid nature of $\upsilon_r(\alpha_r, \mathbf{E})$ as a function of α_r. ∎

We finish this section by considering the example presented in Section 4. There we defined the function $\beta_j(C)$ to be

$$\beta_j(C) = \gamma + \delta_j + \epsilon C, \quad (5.9)$$

where $\epsilon = 1$, $\gamma = 100$, and $\delta_j = 100, j \in \mathcal{J}$. Therefore, $d\beta_j(C)/dC = 1$. Also, the function $\nu_r(\alpha_r, \mathrm{E})$ was defined for $\alpha_r \geq \alpha_r^* = 1$ to be

$$\nu_r(\alpha_r, \mathrm{E}) = \nu_r^* \exp(1 - \alpha_r). \tag{5.10}$$

Therefore, $\hat{\alpha}_r$ is given by

$$\hat{\alpha}_r = \sum_{j \in r} 1 + 1, \tag{5.11}$$

which is the number of links on route r plus one.

6. CONCLUSIONS

In this paper we have presented a unified formulation for the determination of the optimal dimensions and tariffs for a loss network in order to maximize the operating company's profit. We allow the traffic arrival intensity on a route to depend on the tariff charged on that route and the grade of service provided on that route (in the form of the route blocking probability). By exploiting the Erlang fixed point approximation, we have presented two approximate approaches to this problem. The first is the intuitively obvious approach of replacing the exact analysis by the fixed point approximation and so uses the link capacities as the variables. However, by considering the link blocking probabilities to be the variables, a second approach can be developed. This approach enables enormous computational savings to be made as the determination of feasibility is immediate. Further, the constraints can be treated as linear constraints by the use of a simple transformation.

Finally, we used this formulation to prove a result that explains the optimal tariff as the marginal cost to the network of carrying the call plus a term depending only on the elasticity function. This result could be used in certain circumstances, such as where the traffic arrival rates are high and the link blocking probabilities are low, to simplify the optimization procedure.

Acknowledgments

The authors thank Richard Gibbens for useful discussions on the work presented, the Australian Research Council for supporting the research, and an anonymous referee for some useful comments.

References

1. Ash, G.R., Cardwell, R.H., & Murray, R.P. (1981). Design and optimization of networks with dynamic routing. *Bell System Technical Journal* 60: 1787–1820.
2. Berry, L.T.M. (1971). A mathematical model for optimizing telephone networks. Ph.D. thesis, University of Adelaide.
3. Burman, D.Y., Lehoczky, J.P., & Lim, Y. (1984). Insensitivity of blocking probabilities in a circuit-switched network. *Journal of Applied Probability* 21: 850–859.
4. Chung, S.-P., Kashper, A., & Ross, K.W. (1993). Computing approximate blocking probabilities for large loss networks with state-dependent routing. *IEEE/ACM Transaction on Networking* 1(1): 105–115.

5. Girard, A. (1990). *Routing and dimensioning in circuit switched networks.* Addison-Wesley, 1990.
6. Girard, A. & Liau, B. (1993). Dimensioning of adaptively routed networks. *IEEE Transactions on Networking* 1: 460–468.
7. Jagerman, D.L. (1974). Some properties of the Erlang loss function. *Bell System Technical Journal* 53: 525-551.
8. Jensen, A. (1950). Moe's principle: An econometric investigation intended as an aid in dimensioning and managing telephone plants. Copenhagen Telephone Company.
9. Katz, S.S. (1971). Trunk engineering of non-hierarchical networks. In *Proceedings of the International Teletraffic Congress, ITC 6, Munich*, pages 142.1-142.8.
10. Kelly, F.P. (1988). Routing in circuit-switched networks: optimization, shadow prices and decentralization. *Advances in Applied Probability* 20: 112-144.
11. Kelly, F.P. (1990). Routing and capacity allocation in networks with trunk reservation. *Mathematics of Operations Research* 15.
12. Kelly, F.P. (1991). Loss networks. *The Annals of Applied Probability* 1: 319-378.
13. Louth, G., Mitzenmacher, M., & Kelly, F.P. (1993). Computational complexity of loss networks. Research Report No. 93-7, The Statistical Laboratory, University of Cambridge, 16 Mill Lane, Cambridge, CB2 1SB, U.K.
14. Low, S.H. & Varaiya, P.P. (1993). A new approach to service provisioning in ATM networks. *IEEE/ACM Transaction on Networking* 1(5): 547-553.
15. Pratt, C.W. (1967). The concept of marginal overflow in alternate routing. *Australian Telecommunications Research* 1(2): 76–82.
16. Truitt, C.J. (1954). Traffic engineering techniques for determining trunk requirements in alternate routing trunk networks. *Bell System Technical Journal* 33:277-302.