J. Austral. Math. Soc. Ser. B 31(1989), 150–175

A MATHEMATICAL MODEL AND RELATED PROBLEMS OF OPTIMAL MANAGEMENT AND DESIGN IN A BROADBAND INTEGRATED SERVICES NETWORK

SUZANNE P. EVANS[1]

(Received December 1988)

Abstract

This paper describes a mathematical model for a broadband integrated services network offered traffic of many different types. Performance measures are introduced related to revenue generation and overall grade-of-service, providing criteria for the optimal management of resources. Simple asymptotic expressions are derived for quantities termed the "implied costs", which measure the effect on performance of changes in parameters that are controllable by network management, or that are subject to variation. These implied costs may be used, both to implement optimal bandwidth allocation polices, and also to indicate which services may share a single facility without adversely affecting performance, and which might require a dedicated facility. Asymptotic results are also used to examine how to make efficient use of capacity that is shared between calls with fluctuating bit-rate requirements.

1. Introduction

Broadband integrated service networks (B-ISDN) are currently a subject of great interest in the field of telecommunications. In Australia, experts at Telecom's Research Laboratories have been concerned with information transfer protocols for B-ISDN. (See, for example, [2] and [3]). In particular these papers propose a technique known as virtual direct routing. A similar technique of virtual paths has been developed independently at NTT in Japan and, together, these ideas could form the basis of a new international standard. This paper looks at the problems of assigning and controlling the resources in a high capacity integrated services network employing virtual path techniques. Section 2 looks briefly at current ideas on B-ISDN that are relevant to the

[1]Teletraffic Research Centre, Dept. of Applied Mathematics, University of Adelaide, S.A. 5001.

work reported here, and outlines the problems to be addressed. The long Section 3 describes a mathematical model and related theoretical results for the case of a single virtual path (VP) offered traffic of several different types. This provides the theoretical background necessary to address the problems outlined in Section 2. In Section 4 two problems of optimal resource sharing are discussed in terms of the model of Section 3.

2. Virtual paths and bandwidth switching in B-ISDN networks

One of the main factors behind the current move towards B-ISDN is the advent of optical fibres, which provide very high capacities at low cost per unit capacity. Because the low unit cost of capacity is only achievable if very high capacity systems are provided, there is a need to allow a range of communication services to share the resources of a fibre-optic network and so make use of the greater available bandwidth. It is now generally accepted that fast packet switched (FPS) networks would allow the available bandwidth to be used flexibly and efficiently by a mix of services, provided that call establishment could be carried out with minimal delay. The technique of virtual paths was designed primarily to achieve this.

A virtual path (VP) refers to a pre-calculated route between a pair of network exchanges, together with an associated capacity; the route may include any number of exchanges. A VP may be dedicated to a particular service type or may carry a mix of services. As far as the call sub-layer is concerned, the VP appears as a direct link between its origin and destination, with a fixed capacity, and if an arriving call finds sufficient capacity available it may be connected immediately. Only if insufficient capacity is available is it necessary to go outside the call sub-layer and discover whether the network VP manager is able to allocate more capacity to that VP. If more capacity is not assigned, the call is lost. As well as reducing the complexity of call control, the technique of virtual paths has an added bonus; it allows easy reconfiguration of the network in case of failure of network equipment. This feature is of great importance in improving network availability and reliability, and is considered further in [1].

This paper is concerned with the central core network of a B-ISDN system, where the use of fibre-optic transmission results in very large link capacities, and the call concentration function performed by the access networks produces very high arrival rates to each VP. In these circumstances asymptotic results are applicable.

For a given network with given resources, the problems that we wish to analyse concern the optimal management of those resources by control of VP's and their associated capacities. The decisions available include the

setting up of VP's, to carry a mix of services or just a single service, the reservation of capacity for new VP's, switching of capacity (or bandwidth) between existing VP's, and clearing down of VP's where there is insufficient demand. At the network design stage on the other hand, we need to look at the problems of designing for optimal performance subject to budget constraints, or designing to meet given performance criteria at minimum cost.

Performance criteria need to operate at two different levels, at the call level and at the packet level. In accepting or denying access to calls, both grade-of-service and economic considerations are relevant. However a call should not be accepted unless a minimum quality of service standard can be met over the duration of the call, that is unless the probability of packet loss can be kept acceptably low. Calls will generally have randomly varying bit-rate requirements over the duration of a single holding time. We need to know the properties of these random fluctuations in order to develop call acceptance policies that will maintain an optimal balance between minimising call access denial probabilities, maximising revenue, and keeping the probability of packet loss to an acceptable minimum.

The different services using a B-ISDN network are likely to have very different characteristics. For example bit-rate requirements are likely to range over several orders of magnitude (from a few hundred bits/sec up to possibly 135 Mbits/sec), and call arrival rates and holding times will also be very different. This raises the question of which services can share a VP without adversely affecting performance, and which services might require a dedicated facility.

In the next section we develop a model framework that takes most of the above factors into account, and will enable us to formulate and begin to analyse the problems of optimal VP management and network design. There already exists a large body of work on the analysis of mixtures of different traffic types, both with respect to packet delay and call access denial. The approach in this paper is chiefly indebted to the work of Kelly and Hunt whose results on asymptotics, optimality criteria and implied costs is reported in [6], [7], [9] and [10]. Their application of these results is mainly in the context of optimal routing policies; the application to decisions on capacity increases is also discussed but not in the context of the flexible allocation and rearrangement of capacity in a bandwidth switching network.

3. Analysis of a single VP

3.1 The call acceptance model and its exact solution

We begin this section with a description of the model used to determine whether an arriving call is accepted or not. Its essential feature is that an

arriving call's peak bit-rate is the only information that is used at this stage. To take account of the random fluctuations in each call's bit-rate requirement, capacity is assumed to have been overallocated, thus increasing call acceptance probabilities so that efficient use can be made of the available capacity. Discussion on how to do this, while keeping the probability of packet loss acceptably low, is deferred until Section 3.6.

Consider first a single VP (or, equivalently, a single link) which carries a mix of S different services. For the present we take the capacity of the VP to be a given and fixed integer M, where M will generally represent a notional capacity rather than the true capacity. The units in which capacity is measured require some explanation. We shall follow Zukerman and Kirton [19] in their use of the term Fundamental Capacity Unit (FCU) and define an FCU as the largest amount of capacity, say ξ kbits/sec., such that the peak bit-rates of calls of type $s (s = 1, \cdots, S)$ are all integral multiplies of ξ.

For each service s we make the following assumptions:

(i) Calls arrive according to a Poisson process with a constant arrival rate λ_s.

(ii) Call holding times are independent and identically distributed with mean μ_s^{-1}. (Note that we do not need to assume exponential holding times).

(iii) Each call has a peak bit-rate or capacity requirement of a_s FCU's where a_s is an integer.

(iv) Throughout holding times, call bit-rate or capacity requirements at any instant of time are independent and identically distributed with mean m_s and variance v_s. (This assumption is used only in determining the extent to which capacity can be overallocated. See Section 3.6).

Let $\mathbf{a}$ and $\mathbf{n}$ be the column vectors $\mathbf{a} = (a_1, a_2, \ldots, a_S)'$ and $(n_1, n_2, \ldots, n_S)'$ where n_s is the number of type s calls in progress. Then calls are accepted or rejected according to the following rule. An arriving call of type s is accepted provided that

$$\mathbf{a}'\mathbf{n} \leq M - a_s.$$

Otherwise it is rejected and lost. [Note that in a network context it may be possible to increase the capacity of the VP and thus accept the call. For the present it is assumed fixed. Note also that other characteristics of the arriving call, m_s and v_s, are ignored for the purpose of deciding whether the call is to be accepted or not.]

Under all these assumptions the unique, invariant distribution for $\mathbf{n}$ is given by

$$\pi(\mathbf{n}) = G(M)^{-1} \prod_{s=1}^{S} \frac{\nu_s^{n_s}}{n_s!} \qquad \mathbf{n} \in S(M) \tag{1}$$

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

where

$$\nu_s = \lambda_s \mu_s^{-1} \tag{2}$$

$$S(M) = \{\mathbf{n} \in Z_+^S : \mathbf{a}'\mathbf{n} \leq M\} \tag{3}$$

and

$$G(M) = \sum_{\mathbf{n} \in S(M)} \prod_{s=1}^{S} \frac{\nu^{n_s}}{n_s!}. \tag{4}$$

(See, for example, [4].) If we define $L_s(M)$ to be the stationary probability that a call of type s is rejected then

$$(1 - L_s(M)) = G(M - a_s)G(M)^{-1} \tag{5}$$

and from the probability distribution for $\mathbf{n}$ given by (1)–(4) it is easily shown that

$$\begin{aligned} E(n_s) &= \nu_s(1 - L_s(M)) \\ &= \alpha_s(M) \quad \text{(definition)} \end{aligned} \tag{6}$$

$$\mathrm{Var}(n_s) = \alpha_s(M)\{1 - (\alpha_s(M) - \alpha_s(M - a_s))\} \tag{7}$$

$$\begin{aligned} \mathrm{Cov}(n_s, n_t) &= -\alpha_s(M)\{\alpha_t(M) - \alpha_t(M - a_s)\} \\ &= -\alpha_t(M)\{\alpha_s(M) - \alpha_s(M - a_t)\}. \end{aligned} \tag{8}$$

Roberts [16] and Kaufman [8] have independently shown that the $L_s(M)$ may be obtained using a simple recurrence relation, and hence we can obtain $E(n_s)$, $\mathrm{Var}(n_s)$ and $\mathrm{Cov}(n_s, n_t)$ exactly.

3.2 Performance measures W_1 and W_2

Consider first, performance functions that may be written in the form

$$\begin{aligned} W_1(\lambda, \boldsymbol{\mu}; M) &= E\left(\sum_s w_s \mu_s n_s\right) \\ &= \sum_s w_s \mu_s \alpha_s(M), \end{aligned} \tag{9}$$

where w_s may be regarded as the expected reward earned by accepting a call of type s, so that $w_s\mu_s$ is the expected reward earned per unit time. By setting w_s appropriately, such performance functions are able to take into account both revenue generation and overall grade-of-service at the call level. (It is assumed that the overallocation of capacity will be adjusted to take care of the grade-of-service at the packet level.) The implications of setting w_s equal to 1, μ_s^{-1}, and $a_s\mu_s^{-1}$ respectively, are discussed in [5]. The setting $w_s = r_s + g_s$ is also discussed, where the expected reward, w_s, is explicitly separated into the expected revenue, r_s, and the expected goodwill, g_s, valued in the same units as revenue. The main limitation of the linearity of W_1 with respect to the $\alpha_s(M)$ is that accepting a new call of type s has the same value w_s, regardless

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

of the current grade-of-service provided. Performance functions of the form given in (9) have been considered by Kelly [10] and Hunt [7], and lead to tractable expressions for the derivatives of the function with respect to the parameters λ_s and M. In [10], approximations to such derivatives form the basis of a suggested procedure for adaptive routing within a network where several routes are possible between any pair of nodes. The aim is gradually to vary the routing patterns in response to changes in the demands placed on the network, in such a way as to improve the value of the performance function. In the context of optimal resource management we might wish gradually to vary the number, types and capacities of VP's with a similar aim. In the context of the adaptive routing scheme Kelly remarks that the quantities w_s need not be regarded as fixed but could be adjusted when necessary to force routing changes that would reduce unacceptably high loss probabilities. This would require intervention by a human operator. We shall introduce a grade-of-service measure that is nonlinear in the $\alpha_s(M)$, and values call acceptances more highly as access denial probabilities increase.

For each s, suppose that f_s is a function, defined on $[0, \infty)$, with the following properties.

(i) $f_s(0) = 1$.

(ii) $f_s(x) = 0$ if $x \geq 1$.

(iii) Over $[0, 1)$ f_s is continuously differentiable and strictly monotonically decreasing, with an "elbow" at $x = p_s (0 < p_s < 1)$, so that f_s' is negative throughout $[0, 1)$, but small in magnitude between 0 and p_s and large negative thereafter as x increases to 1. [The value of p_s is assumed to represent a desirable call acceptance probability for calls of type s].

Suppose now, that for each service s, we can define a number h_s representing an upper limit to the call arrival rate for which we would aim to provide a grade-of-service p_s. Suppose, further, that defining

$$g_s(x) = f_s(x/h_s) \qquad 0 \leq x < \infty$$

we can choose a number q_s, so that accepting a call of type s, when the average number being accepted per unit time is x_s, is assigned a value of $q_s g_s(x_s)$ dollars per unit time over time μ_s^{-1}. Now let

$$G_s(y) = \int_0^y g_s(x)\, dx$$

and define the grade-of-service performance function W_2 as

$$\begin{aligned} W_2(\lambda, \boldsymbol{\mu}; M) &= \sum_s q_s G_s(\lambda_s(1 - L_s(M))) \\ &= \sum_s q_s G_s(\mu_s \alpha_s(M)). \end{aligned} \tag{10}$$

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

Here $q_s G_s(\mu_s \alpha_s(M))$ is assumed to represent the total value, in dollars per unit time, of all calls of type s being carried on the VP.

NOTE 1. It might be thought appropriate to use $E(G_s(\mu_s n_s))$ in (10) rather than $G_s(\mu_s E(n_s))$. However, using the asymptotic results in Section 3.4, it is possible to show that the difference becomes negligible for large λ_s and M.

NOTE 2. The priority of the different services can be reflected by adjusting the design grade-of-service parameters p_s in the functions f_s, or by adjusting the quantities q_s, and it may be sufficient to keep one of these constant over all services and vary the other.

NOTE 3. If the functions f_s are identical ($\equiv f$), then the terms $G_s(\mu_s \alpha_s(M))$ in W_2 may be thought of as design grade-of-service measures, weighted by the reference arrival rates h_s. This can be seen by showing, as is easily done, that if two services, s and t, have the same call acceptance probabilities, and $\lambda_s = h_s$, $\lambda_t = h_t$, then

$$G_s(\mu_s \alpha_s(M))/G_t(\mu_t \alpha_t(M)) = h_s/h_t.$$

A network performance function reflecting both grade-of-service, as measured by W_2, and generated revenue in the form W_1, can be defined by

$$W(\lambda, \boldsymbol{\mu}; M) = \sum_s (r_s \mu_s \alpha_s(M) + q_s G_s(\mu_s \alpha_s(M))) \tag{11}$$

where r_s is the expected revenue generated by a call of type s. The analogy with setting $w_s = r_s + g_s$ in W_1 is obvious.

In the next section we look at derivatives of W_1 and W_2 with respect to the model parameters λ_s and M. These derivatives are sometimes called "implied costs", as in [7], or "shadow prices" as in [10]. Clearly derivatives with respect to M are relevant when considering capacity alterations. Derivatives with respect to the λ_s have a number of possible uses. First, they turn out to be very simply related to the derivatives with respect to capacity, and give a simple and intuitive understanding of the effect of accepting a single call of type s. Secondly they indicate where accepting a call of type s results in a nett loss in terms of the given performance function; the possible implications in this case are discussed in Section 4.1.

3.3 W_1, W_2: Exact derivatives and implied costs

We begin by considering the performance measure W_1 given by (9). We define

$$d_s = \frac{dW_1}{d\lambda_s}(\lambda, \boldsymbol{\mu}; M) \qquad 1 \le s \le S.$$

Defining the derivative of W_1 with respect to M is less simple. Since all the capacity requirements a_s are measured in integral numbers of FCU's, a capacity alteration of less than one FCU would have no effect on the probability

distribution $\pi(\mathbf{n})$ and hence no effect on W_1. Let us define

$$e^+ = \frac{dW_1}{dM_+}(\lambda, \boldsymbol{\mu}; M)$$
$$= W_1(\lambda, \boldsymbol{\mu}; M+1) - W_1(\lambda, \boldsymbol{\mu}; M)$$

and

$$e^- = \frac{dW_1}{dM_-}(\lambda, \mu; M)$$
$$= W_1(\lambda, \boldsymbol{\mu}; M) - W_1(\lambda, \boldsymbol{\mu}; M-1).$$

Thus e^+ and e^- are the changes in reward earned per unit time caused by respectively adding or subtracting one unit of capacity. These quantities are not necessarily useful, however, in the context of capacity alteration decisions. Capacity alterations will usually involve more than one unit of capacity, and the constraint, $\mathbf{a}'\mathbf{n} \leq M$ that determines the feasible state space $S(M)$, could operate in such a way that larger changes in capacity result in changes in W_1 that are not locally linear functions of e^+ or e^-. It is more convenient at this stage to consider the implied costs c_s defined by

$$c_s = \mu_s^{-1}[W_1(\lambda, \boldsymbol{\mu}; M) - W_1(\lambda, \boldsymbol{\mu}; M - a_s)] \qquad 1 \leq s \leq S. \tag{12}$$

Let $\mathbf{c}$ and $\mathbf{d}$ be the $S \times 1$ column vectors $(c_1, \ldots, c_S)'$ and $(d_1, \ldots, d_S)'$ respectively. The vector $\mathbf{c}$ is closely related to $\mathbf{d}$ as we shall show. Moreover, as pointed out in [7], it is of interest in its own right since c_s is the expected cost (when in equilibrium) of accepting a call of type s, reflecting the fact that the VP must operate at a capacity $M - a_s$ rather than M for an average holding time μ_s^{-1}. In Section 3.5 we shall show that, asymptotically, $e^+ = e^- = e$, and c_s is linearly related to e. In this section we give exact expressions for the elements of $\mathbf{d}$ and $\mathbf{c}$ and show the relationship between them.

THEOREM 1.

(i) $d_s \equiv \frac{dW}{d\lambda_s}(\lambda, \boldsymbol{\mu}; M) = (1 - L_s(M))(w_s - c_s)$ *where* c_s *is as defined in* (12).

(ii) $c_s = \mu_s^{-1} \sum_t w_t \mu_t \left[\delta_{ts} - \frac{\mathrm{Cov}(n_t, n_s)}{E(n_s)}\right]$ *where* $\delta_{ts} = \begin{cases} 1 & t = s, \\ 0 & \text{otherwise.} \end{cases}$

[*Recall that* $\mathrm{Cov}(n_s, n_s) = \mathrm{Var}(n_s)$].

(iii) $d_s = \lambda_s^{-1} \sum_t w_t \mu_t \mathrm{Cov}(n_t, n_s)$.

PROOF. Part (i) may be proved by simple differentiation and applying the definition of c_s given in (12). Part (ii) follows from (12) and the expressions for $\mathrm{Var}(n_s)$ and $\mathrm{Cov}(n_s, n_t)$ given in (7) and (8), and part (iii) is an immediate consequence of parts (i) and (ii). (See [5], or [10] and [7] for a proof that covers the general network case).

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

NOTE 1. Theorem 1 expresses c_s and d_s in terms of moments of the probability distribution $\pi(\mathbf{n})$. Using Kelly's central limit theorem for the distribution $\pi(\mathbf{n})$ ([9] and [6]), asymptotic expressions for c_s and d_s follow immediately [7]. Note, however, that by using the recurrence relations of Roberts [16] and Kaufman [8] exact values for the c_s, and hence the d_s are easily calculable.

NOTE 2. The relationship between d_s and c_s given by Theorem 1(i), can be interpreted as follows. An additional call of type s offered to the VP will be accepted with probability $(1 - L_s(M))$; if accepted it will earn w_s directly but at a cost c_s, where c_s measures the effect of accepting a call of type s in terms of the reward lost from other calls during its holding time. We shall follow [7] and use the term "implied costs" when referring to the c_s and d_s.

Results similar to those given in Theorem 1 can be derived for the performance function W_2.

THEOREM 1′.

(i) $$\hat{d}_s \equiv \frac{dW_2}{d\lambda_s}(\lambda, \boldsymbol{\mu}; M) = (1 - L_s(M))(\hat{w}_s - \hat{c}_s)$$

where we define

$$\hat{w}_s = q_s g_s(\mu_s \alpha_s(M)) = q_s g_s(\mu_s E(n_s))$$

and

$$\hat{c}_s = \mu_s^{-1} \sum_t \hat{w}_t \mu_t (\alpha_t(M) - \alpha_t(M - a_s)).$$

(ii) $$\hat{c}_s = \mu_s^{-1} \sum_t \hat{w}_t \mu_t [\delta_{ts} - (\mathrm{Cov}(n_t, n_s)/E(n_s))].$$

(iii) $$\hat{d}_s = \lambda_s^{-1} \sum_t \hat{w}_t \,\mathrm{Cov}(n_s, n_t).$$

PROOF. See [5].

Note that with the above definitions of $\hat{d}_s$, $\hat{w}_s$ and $\hat{c}_s$, Theorem 1′ is directly analogous to Theorem 1. Note also, however, that although $\hat{d}_s$ is directly analogous to d_s, $\hat{c}_s$ is a linearised version of c_s and $\hat{w}_s$, the value of accepting a call of type s as explained in Section 3.2, is no longer a constant but a function of $\mu_s E(n_s)$, the expected number of calls accepted per unit time; it is therefore a function of λ and M.

In the next section we look at a limiting regime in which both capacity M, and the arrival rates λ_s, are increased together. Asymptotically, the pattern of traffic carried on the VP has a very simple description enabling asymptotic values for d_s, c_s, $\hat{d}_s$ and $\hat{c}_s$ to be calculated very simply.

3.4 The limiting regime and related asymptotics

The analysis starts with the probability distribution $\pi(\mathbf{n})$, given by (1)–(4) and looks at the problem of finding the most likely state $\mathbf{n}$. Following the approach of Kelly [9], with some additions to the theory to handle the difficulties that arise if $n_s = 0$ for some s, we are able to prove the following theorem.

THEOREM 2. *Provided that $M > 0$, there exists a unique vector, $\mathbf{x}^*$, and a unique scalar, B^*, such that*

$$x_s^* = \nu_s(1 - B^*)^{a_s} \qquad 1 \le s \le S \tag{13}$$

$$B^* = 0 \quad \text{if } \mathbf{a}'\mathbf{x}^* < M \tag{14}$$

$$B^* \ge 0 \quad \text{if } \mathbf{a}'\mathbf{x}^* = M \tag{15}$$

and

$$B^* \in [0, 1). \tag{16}$$

The vector $\mathbf{x}^*$ is the unique solution to the problem of finding the most likely state vector, where the integer vector $\mathbf{n}$ has been replaced by a real vector $\mathbf{x}$. The scalar y^*, defined by

$$(1 - B^*) = \exp(-y^*)$$

is the unique solution to the corresponding dual problem.

PROOF. See [5].

NOTE 1. Theorem 2 is the single link equivalent of the theorem for a more general network given by Kelly in [9]. His proof appeals to the strong Lagrangian Principle and requires the primal objective function, which is related to the probability of state $\mathbf{x}$, and is given by

$$P(\mathbf{x}) = \sum_s (x_s \log \nu_s - x_s \log x_s + x_s)$$

to be differentiable over the cone $\{\mathbf{x}: x_s \ge 0, 1 \le s \le S\}$. However more extensive analysis is needed to handle the lack of differentiability of $P(\mathbf{x})$ whenever $x_s = 0$ for some s. Reference [5] uses the theorems proved by Rockafellar in [17] to establish the results of Theorem 2. In the more general network case considered by Kelly, the scalar B^* is replaced by a vector $\mathbf{B}^* = (B_1^*, \ldots, B_L^*)$ where L is the number of links in the network. Unlike the scalar B^* of Theorem 2, the vector $\mathbf{B}^*$ is not in general unique and some care is required when dealing with networks that exhibit such non-uniqueness. For further details see [9].

NOTE 2. As pointed out in [9], relations (13)–(16) have a simple fluid interpretation. Suppose that a service $s (1 \le s \le S)$ offers a flow of size ν_s to

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

the link. This flow is then thinned by a factor $(1 - B^*)^{a_s}$ so that a flow of $\nu_s(1 - B^*)^{a_s}$ remains. Suppose now that a unit of flow of service s uses a_s units of capacity. The relations (13)–(16) state that, if the capacity of the link is not fully utilised by the superposition of the flows $\nu_s(1 - B^*)^{a_s}$, then no thinning of the offered flows occurs. That is $B^* = 0$. On the other hand, if the capacity is fully utilised then two possible cases can arise. Either the capacity requirement of the superposition of offered flows ν_s exactly matches the link capacity, M, in which case $B^* = 0$, or we have $B^* > 0$, and the offered flows ν_s are thinned until the capacity requirement of the superposition of flows $\nu_s(1 - B^*)^{a_s}$ exactly matches the link capacity. We need to distinguish these three cases since the asymptotic behaviour of the system we shall consider is different in each case. In particular the case where offered flow exactly matches link capacity requires careful handling; it is discussed in detail in [6] and [7]. We label the cases as follows:

CASE 1. $B^* = 0$, $x_s^* = \nu_s$ $(1 \le s \le S)$, $\mathbf{a}'\mathbf{x}^* < M$.

CASE 2. $B^* = 0$, $x_s^* = \nu_s$ $(1 \le s \le S)$, $\mathbf{a}'\mathbf{x}^* = M$.

CASE 3. $B^* > 0$, $x_s^* = \nu_s(1 - B^*)^{a_s}$ $(1 \le s \le S)$, $\mathbf{a}'\mathbf{x}^* = M$.

NOTE 3. It is easy to show that given ν_s, $a_s(1 \le s \le S)$ and M, it is simple, computationally, to find the corresponding unique B^* and $\mathbf{x}^*$ of Theorem 2. To see this consider the function

$$H(B) = \sum_s a_s \nu_s (1 - B)^{a_s} \qquad B \in [0, 1]$$

and observe that $H(0) = \sum_s a_s \nu_s$ and $H(1) = 0$. If $H(0) \le M$ then clearly $B^* = 0$ and $\mathbf{x}^* = (\nu_1, \ldots, \nu_S)'$ are the required unique B^* and $\mathbf{x}^*$. If $H(0) > M$ then, since $H(B)$ is a continuous and strictly decreasing function of B on $[0, 1]$, it is easy to find B^* such that $H(B^*) = M$ (for example, by bisection or using a golden section algorithm); the x_s^* follow immediately from (13). The entities $\mathbf{x}^*$ and B^* are the basic parameters for the asymptotic results that follow.

Consider now a sequence of distributions of the form (1)–(4), where the parameters $\lambda_s(1 \le s \le S)$ and M are replaced by $\lambda_s(N)(1 \le s \le S)$ and $M(N)$, and the state vector $\mathbf{n}$ is replaced by $\mathbf{n}(N) = (n_s(N))$. We make two assumptions about these sequences. The first is that

$$\begin{aligned} \lambda_s(N)/N \to \lambda_s \quad &\text{as } N \to \infty \quad 1 \le s \le S \\ M(N)/N \to M \quad &\text{as } N \to \infty. \end{aligned} \tag{17}$$

Thus the arrival rates and the link capacity are increased in line with one another. To examine the limiting behaviour of the stationary probability distributions indexed by N, we examine first the limiting behaviour of the most probable state and the corresponding blocking probability. Here we

follow [6] which corrects the corresponding analysis in [9]. For each N we define $\mathbf{x}^*(N) = (x_s^*(N))$ and $B^*(N)$ to be the unique quantities of Theorem 2 with the $\nu_s = \mu_s^{-1}\lambda_s$ replaced by $\nu_s(N) = \mu_s^{-1}\lambda_s(N)$, and M replaced, either by $M(N)$ (if Case 3 holds for M and $\boldsymbol{\nu}$ and the corresponding $\mathbf{x}^*$ and B^*), or infinity (if Case 1 or Case 2 holds for $M, \boldsymbol{\nu}, \mathbf{x}^*$ and B^*). Then it is easy to show ([6] and [9]) that

$$\mathbf{x}^*(N)/N \to \mathbf{x}^* \quad \text{as } N \to \infty \tag{18}$$

and

$$B^*(N) \to B^* \quad \text{as } N \to \infty. \tag{19}$$

It follows that

$$(M(N) - \mathbf{a}'\mathbf{x}^*(N))/N \to M - \mathbf{a}'\mathbf{x}^* \quad \text{as } N \to \infty \tag{20}$$

which is either strictly greater than 0 (Case 1) or equal to 0 (Cases 2 and 3). Our second assumption concerns the rate of convergence to 0 in (20). We assume that

$$M(N) - \mathbf{a}'\mathbf{x}^*(N) = o(N^{1/2}) \quad \text{for Case 3} \tag{21}$$

where this assumption is without loss of generality (see [9]). We further assume that

$$M(N) - \mathbf{a}'\mathbf{x}^*(N) = \gamma N^{1/2} + o(N^{1/2}) \quad \text{for Case 2} \tag{22}$$

where γ is some constant. That is to say, for Case 2 we consider a sequence in which offered load matches link capacity to order $N^{1/2}$ since $\mathbf{x}^*(N) = \boldsymbol{\nu}(N)$. For each N we now define the vector $\mathbf{u}(N) = (u_1(N), u_2(N), \ldots, u_S(N))'$ by setting

$$u_s(N) = N^{-1/2}[n_s(N) - x_s^*(N)] \qquad 1 \le s \le S. \tag{23}$$

Thus the vector $\mathbf{u}(N)$ is obtained by centering and normalising $\mathbf{n}(N)$. The next theorem follows from the results of [9] and [6].

THEOREM 3.

Case 1. $B^* = 0$, $\mathbf{a}'\mathbf{x}^* < M$

The distribution of $\mathbf{u}(N)$ *converges weakly to the distribution of the vector* $\mathbf{u}$ *whose components are independent normal random variables,* $u_s \sim N(0, x_s^*)$. *If we define the* $S \times S$ *diagonal matrix* $\Sigma = \operatorname{diag}(x_s^*)$, *then* $\mathbf{u}$ *has the multivariate normal distribution* $MVN(\mathbf{0}, \Sigma)$ *and the distribution of* $\mathbf{u}(N)$ *converges weakly to* $MVN(\mathbf{0}, \Sigma)$.

Case 2. $B^* = 0$, $\mathbf{a}'\mathbf{x}^* = M$

The distribution of $\mathbf{u}(N)$ *converges weakly to the distribution of vector* $\mathbf{u} = (u_1, \ldots, u_S)'$ *formed by conditioning independent normal random variables*

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

$u_s \sim N(0, x_s^*)$ on $\mathbf{a}'\mathbf{u} \leq \gamma$, where γ is the constant appearing in (22). *Thus the distribution of* $\mathbf{u}$ *has probability density function*

$$f(\mathbf{u}) = \prod_{s=1}^{S} (x_s^*)^{-1/2} \phi(u_s (x_s^*)^{-1/2}) / \Phi(\gamma \theta^{-1/2}), \qquad \mathbf{a}'\mathbf{u} \leq \gamma \tag{24}$$

where

$$\theta = \sum_s a_s^2 x_s^* \tag{25}$$

and ϕ and Φ are the standard normal density and distribution functions respectively.

Case 3. $B^* > 0$, $\mathbf{a}'\mathbf{x}^* = M$

The distribution of $\mathbf{u}(N)$ *converges weakly to the distribution of a vector* $\mathbf{u} = (u_1, \ldots, u_S)'$ *formed by conditioning independent normal random variables $u_s \sim N(0, x_s^*)$ on $\mathbf{a}'\mathbf{u} = 0$. That is the distribution of* $\mathbf{u}$ *is multivariate normal $MVN(\mathbf{0}, \Sigma^*)$, where the $S \times S$ matrix $\Sigma^* = (\sigma_{st}^*)$ has elements given by*

$$\sigma_{st}^* = \delta_{st} x_s^* - a_s a_t x_s^* x_t^* / \sum_s a_s^2 x_s^*. \tag{26}$$

(*see* [15])

In each of Cases 1, 2 *and* 3, *the moments of* $\mathbf{u}(N)$ *converge to the moments of the corresponding distribution of* $\mathbf{u}$ *and*

$$E(n_s(N))/N \to x_s^* \quad \text{as } N \to \infty \quad 1 \leq s \leq S. \tag{27}$$

Also in each case

$$L_s(N) \to 1 - (1 - B^*)^{a_s} \quad \text{as } N \to \infty \tag{28}$$

where $L_s(N)$ is the probability that a call of type s is lost under the N'th probability distribution $\pi(\mathbf{n}(N))$. This completes the statement of Theorem 3.

From (27) it is clear that

$$W_1(\lambda(N), \boldsymbol{\mu}; M(N))/N \to W_1^*(\lambda, \boldsymbol{\mu}; M) \quad \text{as } N \to \infty \tag{29}$$

where

$$W_1^*(\lambda, \boldsymbol{\mu}; M) = \sum_s w_s \mu_s x_s^*. \tag{30}$$

To examine the asymptotic behaviour of W_2 we need to introduce a sequence $h_s(N)$ such that

$$h_s(N)/N \to h_s \quad \text{as } N \to \infty. \tag{31}$$

We define

$$W_2(\lambda(N), \boldsymbol{\mu}; M(N)) = \sum_s q_s G_s^{(N)}(\mu_s E(n_s(N))) \tag{32}$$

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

where

$$G_s^{(N)}(x) = \int_0^x f_s\left(\frac{y}{h_s(N)}\right) dy \tag{33}$$

and then it is not difficult to show that

$$W_2(\lambda(N), \boldsymbol{\mu}; M(N))/N \to W_2^*(\lambda, \boldsymbol{\mu}; M) \quad \text{as } N \to \infty \tag{34}$$

where

$$W_2^*(\lambda, \boldsymbol{\mu}; M) = \sum_s q_s G_s(\mu_s x_s^*). \tag{35}$$

This completes our description of the limiting regime and the associated asymptotic results. In the next section we use these results to look at the asymptotic behaviour of the implied costs defined in Section 3.3.

3.5 W_1, W_2: asymptotic behaviour of implied costs

The following lemma establishes the basic results that will allow us to describe the asymptotic behaviour of implied costs.

LEMMA 1.

(i) $$\frac{\mathrm{Cov}(n_t(N), n_s(N))}{E(n_s(N))} \to \frac{\mathrm{Cov}(u_t, u_s)}{x_s^*} \quad \text{as } N \to \infty$$

(ii) $$\frac{\mathrm{Cov}(n_t(N), n_s(N))}{\lambda_s(N)} \to \frac{\mathrm{Cov}(u_t, u_s)}{\lambda_s} \quad \text{as } N \to \infty$$

(iii) $$\alpha_s(M(N)) - \alpha_s(M(N) - a_t) \to \delta_{ts} - \mathrm{Cov}(u_s, u_t)/x_t^* \quad \text{as } N \to \infty.$$

PROOF. Results (i) and (ii) are proved in [7]. They are an easy consequence of the definition of $\mathbf{u}(N)$, the fact that moments of $\mathbf{u}(N)$ converge to the corresponding moments of $\mathbf{u}$ (Theorem 3), and (27). Result (iii) follows from (7), (8) and Result (i) above.

Theorems 4 and 4′, which follow, demonstrate the convergence of the implied costs associated, respectively, with W_1 and W_2.

THEOREM 4. *The implied costs $d_s(N)$ and $c_s(N)$, where these are defined in the obvious way, converge to limiting costs as $N \to \infty$. The limiting costs are given by*

(i) $$c_s^* = \mu_s^{-1} \sum_t w_t \mu_t \{\delta_{ts} - \mathrm{Cov}(u_t, u_s)/x_s^*\} \qquad 1 \le s \le S$$

(ii) $$d_s^* = \lambda_s^{-1} \sum_t w_t \mu_t \,\mathrm{Cov}(u_t, u_s) \qquad 1 \le s \le S$$

and

(iii) $$d_s^* = (1 - B^*)^{a_s}(w_s - c_s^*).$$

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

PROOF. Results (i) and (ii) follow immediately from Lemma 1(i) and (ii). Result (iii) follows from (i) and (ii).

Consider now the implied costs associated with W_2. By analogy with Theorem 1′ we define $\hat{d}_s(N)$ and $\hat{c}_s(N)$ in the obvious way as

$$\hat{d}_s(N) = \lambda_s(N)^{-1} \sum_t \hat{w}_t(N)\mu_t \operatorname{Cov}(n_s(N), n_t(N)) \qquad 1 \le s \le S \tag{36}$$

and

$$\hat{c}_s(N) = \mu_s^{-1} \sum_t \hat{w}_t(N)\mu_t[\delta_{ts} - \operatorname{Cov}(n_t(N), n_s(N))/E(n_s(N))] \qquad 1 \le s \le S \tag{37}$$

where

$$\hat{w}_t(N) = q_t f_t(\mu_t E(n_t(N))/h_t(N)) \qquad 1 \le t \le S. \tag{38}$$

THEOREM 4′. *The implied costs $\hat{d}_s(N)$ and $\hat{c}_s(N)$ converge to limiting costs as $N \to \infty$ where these are given by*

(i) $$\hat{c}_s^* = \mu_s^{-1} \sum_t \hat{w}_t^* \mu_t \{\delta_{ts} - \operatorname{Cov}(u_t, u_s)/x_s^*\}$$

and

(ii) $$\hat{d}_s^* = \lambda_s^{-1} \sum_t \hat{w}_t^* \mu_t \operatorname{Cov}(u_t, u_s)$$

where

$$\hat{w}_t^* = q_t f_t(\mu_t x_t^*/h_t) \tag{39}$$

and

(iii) $$\hat{d}_s^* = (1 - B^*)^{a_s}(\hat{w}_s^* - \hat{c}_s^*).$$

PROOF. Since f_t is continuous it is clear that

$$\hat{w}_t(N) \to \hat{w}_t^* \quad \text{as } N \to \infty$$

from (27) and (31). The results are then immediate from Lemma 1(i) and (ii).

An immediate Corollary to Theorems 4 and 4′ gives closed form expressions for the c_s^* and $\hat{c}_s^*$, and hence for the d_s^* and $\hat{d}_s^*$, in Cases 1 and 3.

COROLLARY 4.1.
Case 1. $B^* = 0$, $a'\mathbf{x}^* < M$
(i) $c_s^* = 0$ (*and hence* $d_s = w_s$), $\quad 1 \le s \le S$
(ii) $\hat{c}_s^* = 0$ (*and hence* $\hat{d}_s^* = \hat{w}_s^*$), $\quad 1 \le s \le S$.

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

Case 3. $B^* > 0$, $\mathbf{a}'\mathbf{x}^* = M$

(i) $c_s^* = a_s\mu_s^{-1}(\sum_t w_t\mu_t a_t x_t^* / \sum_t a_t^2 x_t^*), \quad 1 \le s \le S$

(ii) $\hat{c}_s^* = a_s\mu_s^{-1}(\sum_t \hat{w}_t^*\mu_t a_t x_t^* / \sum_t a_t^2 x_t^*), \quad 1 \le s \le S.$

PROOF. The results follow from Theorems 4 and 4′, and the expressions for the moments of $\mathbf{u}$ given in Theorem 3 for Cases 1 and 3. Expressions for d_s^* and $\hat{d}_s^*$ for Case 3 are immediate from Theorems 4 and 4′, result (iii).

Note that in Case 2 it is not possible to find simpler expressions for c_s^* and $\hat{c}_s^*$ than are given by Theorem 3 and Theorems 4 and 4′. The covariance matrix of the distribution specified by (24) is required, and its elements are not expressible in closed form.

The following Theorem 5 contains a result that, asymptotically, will enable us to attach meanings to the derivatives of W_1 and W_2 with respect to VP capacity.

THEOREM 5.

(i) *There exists a quantity* e^* *such that*

$$\mu_s c_s^* = a_s e^* \qquad 1 \le s \le S. \tag{40}$$

(ii) *There exists a quantity* $\hat{e}^*$ *such that*

$$\mu_s \hat{c}_s^* = a_s \hat{e}^* \qquad 1 \le s \le S. \tag{41}$$

We may interpret (40) (and similarly (41)) as follows. By definition, $\mu_s c_s^*$ measures the worth per unit time of a_s FCU's, as measured by W_1. Equation (40) says that we may assign a worth of e^* to a single FCU such that the worth of a_s FCU's for any $s = 1, \ldots, S$ is just $a_s e^*$, the appropriate multiple of e^*.

PROOF. Result (i) is proved in [7]. Result (ii) has an exactly analogous proof provided we can show that, for all $\mathbf{k} \in Z^S$

$$\sum_s \hat{w}_s(N)\mu_s\{\alpha_s(M(N)) - \alpha_s(M(N) - \mathbf{k}'\mathbf{a})\} \to \sum_s \mu_s k_s \hat{c}_s^* \quad \text{as } N \to \infty.$$

We can write this as a finite sum of $K = \sum_s |k_s|$ terms of the form

$$\pm\left\{\sum_s \hat{w}_s(N)\mu_s\{\alpha_s(M(N) - \gamma'\mathbf{a}) - \alpha_s(M(N) - \gamma'\mathbf{a} - a_t)\}\right\} \tag{42}$$

for some $\gamma \in Z^S$. Now $\hat{w}_s(N) \to \hat{w}_s^*$ as $N \to \infty$. (See proof of Theorem 4′). Moreover the same asymptotics hold for a sequence of link capacities $M(N) - \gamma'\mathbf{a}$ as they do for the sequence $M(N)$, and so by Lemma 1(iii), (42) converges to $\pm\mu_t\hat{c}_t^*$ as $N \to \infty$ and the result follows. This completes the proof of Theorem 5.

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

NOTE. The proof of Theorem 5 also demonstrates that, asymptotically,

(i) c_s^* (or $\hat{c}_s^*$) is the amount we would gain by adding a_s FCU's to the link, as well as the amount we would lose by removing a_s FCU's and

(ii) K units of capacity are worth Ke^* (or $K\hat{e}^*$) if K can be expressed as $K = \mathbf{a}'\mathbf{k}$ for some $\mathbf{k} \in Z^S$ where the elements of $\mathbf{k}$ are $o(N)$. Thus Theorem 5 tells us that, asymptotically, we may define

$$\frac{dW_1}{dM(N)}(\lambda(N), \boldsymbol{\mu}; M(N)) = e^*$$

and

$$\frac{dW_2}{dM(N)}(\lambda(N), \boldsymbol{\mu}; M(N)) = \hat{e}^*.$$

In Cases 1 and 3, we can find explicit expressions for e^* and $\hat{e}^*$. By inspecting the expressions for c_s^* and $\hat{c}_s^*$ in Corollary 4.1, we see that

$$e^* = \hat{e}^* = 0 \quad \text{Case 1} \tag{43}$$

$$e^* = \sum_s w_s \mu_s a_s x_s^* / \sum_s a_s^2 x_s^* \quad \text{Case 3} \tag{44}$$

and

$$\hat{e}^* = \sum_s \hat{w}_s^* \mu_s a_s x_s^* / \sum_s a_s^2 x_s^* \quad \text{Case 3.} \tag{45}$$

For large N, (44) and (45) are clearly well approximated by the corresponding expressions with the x_s^* replaced by $x_s^*(N)$, and the $\hat{w}_s^*$ replaced by $\hat{w}_s(N)$.

In Case 3 ($B^* > 0, \mathbf{a}'\mathbf{x}^* = M$) we are able to get a better idea of the behaviour of W_1 by looking at the behaviour of the function W_1^*. We use the relations

$$\mathbf{a}'\mathbf{x}^* = M \tag{46}$$

and

$$\mathbf{a}'\mathbf{x}^*(N) = M(N) \quad \text{for all } N > N' \tag{47}$$

where N' is chosen so that $B^*(N) > 0$ for all $N > N'$. Differentiating (46) and (47) implicitly we obtain expressions for the derivatives of B^*, and hence x_t^*, with respect to M and λ_s, and analagous expressions for the derivatives of $B^*(N)$, and hence $x_t^*(N)$, with respect to $M(N)$ and $\lambda_s(N)$. It is easy to see then that $dx_t^*(N)/dM(N)$ and $dx_t^*(N)/d\lambda_s(N)$ are $O(1)$ and converge to dx_t^*/dM and $dx_t^*/d\lambda_s$ respectively, while $dB^*(N)/dM(N)$ and $dB^*(N)/d\lambda_s(N)$ are $O(N^{-1})$, and $NdB^*(N)/dM(N)$ and $NdB^*(N)/d\lambda_s(N)$ converge to dB^*/dM and $dB^*/d\lambda_s$ respectively. Consider now the function W_1^* defined in (30). It follows from (18) that

$$W_1^*(\boldsymbol{\lambda}(N), \boldsymbol{\mu}; M(N))/N \to W_1^*(\boldsymbol{\lambda}, \boldsymbol{\mu}; M) \quad \text{as } N \to \infty.$$

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

By using the expressions for the derivatives of $x_s^*(N)$, $x_s^*, B^*(N)$ and B^*, we can also show that

$$\frac{dW_1^*}{dM}(\lambda, \mu; M) = e^*$$

$$\frac{dW_1^*}{dM(N)}(\lambda(N), \mu; M(N)) \to e^* \quad \text{as } N \to \infty$$

and

$$N\frac{d^2W_1^*}{dM(N)^2}(\lambda(N), \mu; M(N)) \to \frac{d^2W_1^*}{dM^2}(\lambda, \mu; M).$$

Thus, while $W_1^*(\lambda(N), \mu; M(N))$ is $O(N)$, the derivative of $W_1^*(\lambda(N), \mu; M(N))$ with respect to $M(N)$ is $O(1)$ (and $\equiv e^*$), and the second derivative with respect to $M(N)$ is $O(N^{-1})$ so that, over capacity ranges of $o(N)$, $W_1^*(\lambda(N), \mu; M(N))$ is nearly a linear function of $M(N)$ while Case 3 holds.

The behaviour of $W_1^*(\lambda(N), \mu; M(N))$ reflects the behaviour of $W_1(\lambda(N), \mu; M(N))$ as we can easily see. From (29)

$$W_1(\lambda(N), \mu; M(N))/N \to W_1^*(\lambda, \mu; M) \quad \text{as } N \to \infty.$$

From Theorem 5 (i) the first differences

$$\Delta_{a_s} W_1(M(N)) = W_1(\lambda(N), \mu; M(N)) - W_1(\lambda(N), \mu; M(N) - a_s)$$

converge to $a_s e^*$ as $N \to \infty$ for all $1 \le s \le S$. Finally, it is easily shown, by an argument similar to that of Theorem 5(ii), that the second differences

$$\Delta^2_{a_s, a_t} W_1(M(N)) = \Delta_{a_s} W_1(M(N)) - \Delta_{a_s} W_1(M(N)) - a_t)$$

converge to 0 as $N \to \infty$ for all $1 \le s, t \le S$. Thus, over capacity ranges of $o(N)$, $W_1(\lambda(N), \mu; M(N))$ is nearly a linear function of $M(N)$.

The behaviour of the function W_2 for Case 3 can be examined in a similar way.

This completes Section 3.5. In the next section we shall look at the actual capacity use of a VP, as compared with the nominal use based on the assumption that calls of type s ($1 \le s \le S$) occupy a_s units of capacity throughout their holding times.

3.6 Capacity overallocation

For each service $s (1 \le s \le S)$, we recall that call capacity requirements, at any instant of time throughout a call's holding time, are independent and identically distributed random variables with mean m_s and variance v_s. We shall begin by assuming that the link (or VP) has infinite capacity, and that the λ_s are large so that, asymptotically, the $n_s \sim NID(\nu_s, \nu_s)$. Let us consider the distribution of y, the actual capacity used at any instant of time.

Suppose, first, that we fix a value for the state vector $\mathbf{n} = (n_1, \ldots, n_S)'$, where n_s is the number of calls of type s in progress ($1 \le s \le S$). Since

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

the λ_s are assumed large, the ν_s and n_s may also be assumed large. While in state $\mathbf{n}$, the capacity requirement of type s calls at any instant of time is a random variable y_s, which is the sum of n_s independent and identically distributed random variables with mean m_s and variance v_s. Thus, by the central limit theorem, y_s, conditioned on the state vector $\mathbf{n}$, is approximately normally distributed with mean $m_s n_s$ and variance $v_s n_s$. It follows that if we define $\mathbf{m} = (m_1, \ldots, m_S)'$ and $\mathbf{v} = (v_1, \ldots, v_S)'$ then, conditional on the value of $\mathbf{n}$, the instantaneous total capacity requirement of all calls, $y = \sum_s y_s$, is a random variable which is approximately normally distributed with mean $\mathbf{m}'\mathbf{n}$ and variance $\mathbf{v}'\mathbf{n}$. That is

$$y/\mathbf{n} \sim N(\mathbf{m}'\mathbf{n}, \mathbf{v}'\mathbf{n}) \tag{48}$$

and

$$\mathbf{n} \sim MVN(\boldsymbol{\nu}, \Sigma) \tag{49}$$

where

$$\Sigma = \operatorname{diag}(\nu_s).$$

Now if we define the random vector $\mathbf{z} = (z_1, z_2)'$ by setting

$$z_1 = \mathbf{m}'\mathbf{n} \quad \text{and} \quad z_2 = \mathbf{v}'\mathbf{n}$$

then, since $\mathbf{n} \sim MVN(\boldsymbol{\nu}, \Sigma)$, $\mathbf{z}$ has the bivariate normal distribution with mean $\boldsymbol{\gamma} = (\mathbf{m}'\boldsymbol{\nu}, \mathbf{v}'\boldsymbol{\nu})'$ and covariance matrix $\Psi = (\psi_{ij})$ where

$$\psi_{11} = \sum_s m_s^2 \nu_s, \qquad \psi_{22} = \sum_s v_s^2 \nu_s, \qquad \psi_{12} = \psi_{21} = \sum_s m_s v_s \nu_s.$$

See [14].

It is not difficult to see that the distribution of y, defined by (48) and (49), can be equivalently defined by

$$y/\mathbf{z} \sim N(z_1, z_2) \tag{50}$$

and

$$\mathbf{z} \sim BVN(\boldsymbol{\gamma}, \Psi). \tag{51}$$

Let

$$p(z_1, z_2) = \frac{1}{2\pi|\Psi|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\gamma})'\Psi^{-1}(\mathbf{z} - \boldsymbol{\gamma})\right]$$

be the bivariate normal density function with mean vector $\boldsymbol{\gamma}$ and covariance matrix Ψ, and

$$q(y; z_1, z_2) = \frac{1}{\sqrt{2\pi z_2}} \exp\left[-\frac{1}{2}\frac{(y - z_1)^2}{z_2}\right]$$

be the normal density function with mean z_1 and variance z_2. Then the probability density function of y is approximately given by

$$g(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q(y; z_1, z_2) p(z_1, z_2) dz_1\, dz_2. \tag{52}$$

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

[Note that the normal approximations which lead to this formula are a consequence of the assumption that the components of $\mathbf{n}$ and $\boldsymbol{\nu}$ are assumed large, of $O(N)$ say, and hence the components of $\mathbf{z}$ and γ, the elements of Ψ, and the scalar y are also $O(N)$. This means that, practically speaking, the limits of integration in (52) may be severely truncated, and, in particular, that the lower limits of integration may be replaced by positive values; the problem of $z_1 \leq 0$ in $q(y; z_1, z_2)$ need not arise. The expression given in (52) is, therefore, quite tractable computationally. Notice, also, that the distribution of y given by (52) may be thought of as a generalisation of the familiar concept of offered traffic.]

Using (52) it is easy to show that

$$E(y) = \gamma_1 = \mathbf{m}'\boldsymbol{\nu}$$

and

$$\mathrm{Var}(y) = \gamma_2 + \psi_{11} = \sum_s (v_s + m_s^2)\nu_s.$$

We may compare the distribution of y, the actual instantaneous total capacity requirement, with the distribution of $\mathbf{a}'\mathbf{n}$ which is approximately normally distributed with mean $\sum_s a_s\nu_s$ and variance $\sum_s a_s^2\nu_s$. Comparing moments it is easy to show that $E(y) \leq E(\mathbf{a}'\mathbf{n})$, $\mathrm{Var}(y) \leq \mathrm{Var}(\mathbf{a}'\mathbf{n})$.

One approach to the problem of making the best use of the capacity of a VP might be to use the distributions of y and $\mathbf{a}'\mathbf{n}$ just derived for the case of a link of infinite capacity, to determine an overallocation factor, f, for the associated mix of traffic. This would need to be done in such a way that the resulting probability of packet loss was small. For example we might set

$$f = \frac{E(\mathbf{a}'\mathbf{n}) + k\sqrt{\mathrm{Var}(\mathbf{a}'\mathbf{n})}}{E(y) + k\sqrt{\mathrm{Var}(y)}} \quad \text{for some } k$$

or

$$f = M_1(\varepsilon)/M_2(\varepsilon) \quad \text{for some } \varepsilon$$

where $P(\mathbf{a}'\mathbf{n} \geq M_1(\varepsilon)) = \varepsilon$ and $P(y \geq M_2(\varepsilon)) = \varepsilon$. Both these suggested formulae for f depend on the assumed arrival rates λ_s. It may be that these f are very sensitive to changes in the λ_s and we discuss this possibility and its consequences in Section 4.1.

Another approach to the overallocation of capacity is the following. If the actual capacity available on a VP may be assumed known, and is denoted, say, by M_0, then the problem is to find a nominal capacity M, and hence a call acceptance policy, such that the resulting random variable y, representing the instantaneous actual capacity requirement, has an acceptably low probability of exceeding M_0. Assume first that we choose a nominal capacity M so big that all arriving calls are accepted. Then the distribution of the instantaneous

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

capacity requirement, y, is given by (52), and we may use this to check whether $P(y > M_0)$ is acceptably small. If it is not then let us choose an M such that $\mathbf{a}'\mathbf{n} > M$. Then, in our limiting regime, $B^* > 0$ and Case 3 applies. Theorem 3 now tells us that, asymptotically, $\mathbf{n} \sim MVN(\mathbf{x}^*, \Sigma^*)$ where

$$x_s^* = \nu_s(1 - B^*)^{a_s}, \quad 1 \leq s \leq S \quad \text{and} \quad \Sigma^* = (\sigma_{st}^*)$$

where $\sigma_{st}^* = \delta_{st}x_s^* - (a_s a_t x_s^* x_t^*)/(\sum_s a_s^2 x_s^*)$, $1 \leq s, t \leq S$.

It is then easily shown that the distribution of y is given by

$$y/\mathbf{z} \sim N(z, z_2)$$

and

$$\mathbf{z} \sim BVN(\gamma^*, \Psi^*)$$

where $\gamma^* = (\mathbf{m}'\mathbf{x}^*, \mathbf{v}'\mathbf{x}^*)'$ and $\Psi^* = (\mathbf{m}, \mathbf{v})'\Sigma^*(\mathbf{m}, \mathbf{v})$. (See [14].) Using these results it is a simple matter computationally to adjust M so that some chosen criterion is satisfied. For example, $P(y > M_0) \leq \varepsilon$. The above procedure finds a nominal capacity $M \geq M_0$, depending on both the call arrival rates λ_s, and M_0, such that the chosen criterion is satisfied.

This completes the model description for the case of a single VP or link. Its purpose has been to provide tools for analysing problems concerned with the optimal sharing of capacity between services and VP's in a high capacity core network, and also problems concerned with design and provision of capacity in such a network. In Section 4 we discuss briefly two problems of optimal resource management. Section 4.1 looks at the case of different services sharing the resources of a single VP and Section 4.2 considers the case of different VP's sharing the resources of a given network. Problems of optimal network design can also be formulated in terms of this model but are not considered in this paper. Before moving on to Section 4 let us examine for a moment some of the assumptions on which the model is based.

3.7 Discussion

Two assumptions have been made in this section that bear closer examination in cases where services differ markedly, either with respect to their capacity requirements, or with respect to the time scales on which they operate, or both. The first relates to the limiting regime, in which we have assumed that arrival rates and link capacity are increased in line with one another, and that asymptotic results may be used when these quantities become sufficiently large. For this purpose we require that the total link capacity becomes very large compared with each individual a_s; for a service such as video, this may not be the case. The second assumption concerns our use of stationary probability distributions. When considering together services that operate on very different time scales, for example voice and video, it is possible that stationary probabilities may not tell us enough about the behaviour

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

of the system. The problem of sharing resources among services with such markedly different characteristics is currently the subject of a separate study.

4. Optimal resource management

4.1 Sharing resources between services

A problem, already mentioned in this paper, is how to discover which services can successfully share the resources of a single VP, and which services would require a dedicated VP tailored to their particular requirements. Some of the quantities introduced in Section 3 may indicate answers to such questions. For example, the elements of the implied cost vector $\mathbf{d} = (dW/d\lambda_s)$ may be informative, where W may be of the form W_1 or W_2, or a linear combination of the two. If $dW/d\lambda_s < 0$ for some s, then the gain of accepting a call of type s on the VP is not sufficient to offset the cost incurred by reducing the capacity available to other services. This could indicate that the particular service for which $dW/d\lambda_s < 0$ should have a dedicated VP so that its effect on other services can be controlled by altering its own VP capacity. Alternatively it could indicate that the call acceptance model should be adjusted to take account of factors other than available capacity. (It might also, of course, indicate that a call of this type should be charged more so as to compensate for the effect on other calls.) Another quantity of interest is the capacity overallocation factor discussed in Section 3.6. If overallocation factors, calculated as we have described in Section 3.6, turn out to be unduly sensitive to changes in the call arrival rates, λ_s, this could indicate an instability in actual capacity requirement caused by an inappropriate mixture of services. It could mean, however, that what is inappropriate is the system of accepting or rejecting calls of different services on the basis of each call's peak capacity requirements, and a nominal VP capacity calculated using a single overallocation factor. Another possibility is to work with the actual VP capacity, to introduce new quantities, b_s, representing "typical" call capacity requirements, and to accept a call of type s if and only if at least b_s units remain unallocated. The model, its solutions, and all the asymptotic results, remain essentially unaltered, but it is possible that the b_s will be less sensitive to changes in the λ_s than the single overallocation factor f. A possible way of choosing the b_s would be to set $b_s = m_s + k_s\sqrt{v_s}$ for some k_s. It is not clear yet which is the best way to handle this problem of the efficient use of capacity being shared between calls with fluctuating bit-rate requirements, but as a start we would wish to examine the feasibility of the methods suggested in Section 3.6 with respect to sensitivity to changes in arrival patterns.

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

4.2 Sharing resources between VP's

As we have noted, the models introduced in this paper are designed to give a macroscopic description of network use when the system is in statistical equilibrium. This makes them very suitable for analysis at the network design stage, for example for dimensioning purposes. However their use also extends to problems of network VP management when the aim is to alter VP structure and capacities in response to changes in demand patterns and network form, rather than to optimise capacity ultilisation on the timescale of very short term statistical fluctuations, or to control capacity use at the packet or burst level. An adaptive scheme based on the models of this section would make adjustments to the VP structure and capacities as indicated by changes in the implied cost vectors, and these adjustments would be made gradually to ensure stability in the resulting traffic patterns. For changes on such a time scale, equilibrium models would not be inappropriate. Optimising network performance at the network design stage leads to problems of global optimisation, while an adaptive VP management scheme would be more concerned with local optimisation. An interesting question is whether a local approach to optimisation via an adaptive capacity allocation scheme would tend to move the system in the direction of the global maximum.

In this section we shall briefly indicate how the results of Section 3 could be made the basis of an algorithm for switching capacity or bandwidth between VP's so as to optimise network performance. The study of these and other related optimisation problems is proceeding.

Suppose that there are C different VP's sharing the resources of a core network. For each VP, c, $(1 \leq c \leq C)$ we make the following assumptions.

(i) There are S_c services that share VP c, and for each service s $(1 \leq s \leq S_c)$ we make the same assumptions that are made in the model of Section 3.1, with $\lambda_s, \mu_s, a_s, m_s$ and v_s replaced by $\lambda_{sc}, \mu_{sc}, a_{sc}, m_{sc}$ and v_{sc} respectively.

(ii) There is a nominal capacity M_c assigned to VP c, and an arriving call of type s is accepted provided that $\mathbf{a}_c'\mathbf{n}_c \leq M_c - a_{sc}$, where $\mathbf{a}_c = (a_{1c}, a_{2c}, \ldots, a_{S_c c})'$ and $\mathbf{n}_c = (n_{1c}, n_{2c}, \ldots, n_{S_c c})'$.

(iii) The VP performance function W_c is a function of λ_c and M_c and for definiteness we shall assume it is of type W_1; that is, it can be written as

$$W_c(\lambda_c, \mu_c; M) = \sum_{s=1}^{S_c} w_{sc}\mu_{sc}E(n_{sc}).$$

(iv) An overallocation factor f_c can be calculated, where f_c may be a function of λ_c.

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

We define the overall performance function W by

$$W(\lambda_1,\ldots,\lambda_C,\mu_1,\ldots,\mu_C;M_1,\ldots,M_C)=\sum_{c=1}^{C}W_c(\lambda_c,\mu_c;M_c). \tag{53}$$

Suppose, now, that the λ_c are fixed and known, the corresponding f_c are known, and the current capacity allocations M_c are given. We may calculate the gradient or implied cost vector $\mathbf{e}^*$ given by $\mathbf{e}^* = (e_1^*,\ldots,e_C^*)'$, where $e_c^* = dW_c/dM_c$, and consider altering the vector $\mathbf{M} = (M_1,\ldots,M_C)'$ so as to improve the overall performance function W subject to constraints imposed by the actual resources available on the links of the network. In the simplest situation, in which a single link of actual capacity D is shared between the C VP's, these reduce to a single constraint given by

$$\sum_{c=1}^{C}f_c^{-1}M_c \le D. \tag{54}$$

If the network consists of L links, each of actual capacity $D_l (1 \le l \le L)$, then the constraints are of the form

$$BF\mathbf{M} \le \mathbf{D} \tag{55}$$

where $\mathbf{D} = (D_1,\ldots,D_L)'$, $F = \mathrm{diag}(f_c^{-1})$ and $B = (b_{lc})$ is an $L \times C$ matrix whose elements are given by

$$b_{lc} = \begin{cases} 0 & \text{if link } l \text{ is not part of VP } c, \\ 1 & \text{if link } l \text{ is part of VP } c. \end{cases} \tag{56}$$

An obvious approach to improving the value of W is to use one of the methods used at each iteration of a global optimisation algorithm to choose a new direction in which to move. A number of these are discussed in [11] and [18], and include a variety of projected-gradient or reduced-gradient techniques. For the single constraint (54) all these methods give easily computable new directions. For the constraints (55), determining a new direction is computationally more complex, commonly involving a matrix inversion.

4.3 Discussion

In Section 4.2 we have taken a preliminary look at the local optimisation of the performance function W. A deeper study of the properties of W and the nature of these optimisation problems is proceeding. The function W is not, in general, concave so that no easy results are likely. However the linearity of the constraints, the decomposability of the objective function, and its near linearity with respect to $\mathbf{M}$ in Case 3, point to a number of possible approaches to both the local and global optimisation problems via duality, and decomposition. An interesting discussion in [10], referring to the behaviour

https://doi.org/10.1017/S0334270000006561 Published online by Cambridge University Press

of an adaptive routing scheme rather than an adaptive capacity allocation scheme, considers the behaviour of the scheme in the event that W has local maxima other than the global maximum. By analogy with the convergence of simulated annealing algorithms (see [12] and [13]), it is suggested that an adaptive scheme that relies, at each stage, on estimates of current traffic parameters to indicate optimal local changes, has inherent stochastic fluctuations that might allow it to escape from the region around a nonoptimal local maximum. It is further suggested that if the global maximum is sufficiently greater than other maxima, the equilibrium distribution of the scheme might assign a relatively high probability to the region around the global maximum.

Optimisation problems that arise in the context of the design and dimensioning of a core B-ISDN network may involve nonlinearities in the constraints as well as in the objective function and the feasible region need not be convex. Further study of such problems is required.

5. Conclusions

The model formulated in this paper gives a macroscopic description of a B-ISDN network employing virtual path techniques. Using performance measures that reflect both revenue and grade-of-service at the call level, implied cost vectors can be calculated that quantify the effect of small changes in network parameters that are either controllable by network management, or liable to vary. Within this framework a number of problems can be addressed. For example the problem of optimal management of capacity can be formulated mathematically and the implied cost vectors used to suggest locally optimal changes. Implied cost vectors, as well as suggested capacity overallocation factors, can also be used to indicate answers to questions concerning optimal mixes of services sharing the same VP.

Acknowledgement

The author is pleased to acknowledge the funding of Telecom Australia.

References

[1] R. G. Addie, "B-ISDN protocol architecture and network reliability", *Proc. 3rd ATERB FPS Workshop* (1988).

[2] R. G. Addie, J. L. Burgin and S. L. Sutherland, "Information transfer protocols for the broadband ISDN", *Proc. GLOBECOM'88* (1988) 716–720.

[3] R. G. Addie and R. E. Warfield, "Teletraffic engineering of new network structures", *Proc. 2nd Australian Teletraffic Research Seminar* (1987).
[4] D. Y. Burman, J. P. Lehoczky and Y. Lim, "Insensitivity of blocking probabilities in a circuit-switched network", *J. Appl. Prob.* **21** (1984) 850–859.
[5] S. P. Evans, "Integration of services and bandwidth allocation in a B-ISDN network", *Teletraffic Research Centre, University of Adelaide Rep.* TRC 22/88 (1988).
[6] P. J. Hunt and F. P. Kelly, "On critically loaded loss networks", *Adv. Appl. Prob.* **21** (1989).
[7] P. J. Hunt, "Implied costs in loss networks", *Adv. Appl. Prob.* **21** (1989).
[8] J. S. Kaufman, "Blocking in a shared resource environment", *IEEE Trans. Commun. Vol. COM-29* (10) (1981) 1474–1481.
[9] F. P. Kelly, "Blocking probabilities in large circuit-switched networks", *Adv. Appl. Prob.* **18** (1986) 473–505.
[10] F. P. Kelly, "Routing in circuit-switched networks: shadow prices and decentralization", *Adv. Appl. Prob.* **20** (1988) 112–144.
[11] L. S. Lasdon, *Optimization theory for large systems* (Macmillan, New York, 1970)
[12] M. Lundry and A. Mees, "Convergence of the annealing algorithm", *Math. Programming* **34** (1986) 111–124.
[13] D. Mitra, F. Romeo and A. Sangiovanni-Vincentelli, "Convergence and finite-time behaviour and simulated annealing", *Adv. Appl. Prob.* **18** (1986) 747–771.
[14] D. F. Morrison, *Multivariate statistical methods* (McGraw-Hill, 1967).
[15] C. R. Rao, *Linear statistical inference and its applications* (Wiley, New York, 1965).
[16] J. W. Roberts, "A service system with heterogeneous user requirements—application to multiservice telecommunications systems", in *Performance of data communications systems and their applications* (ed. G. Pujolle), (North-Holland, 1981) 423–431.
[17] R. T. Rockafellar, "Convex programming and systems of elementary monotonic relations", *J. Math. Anal. Appl.* **19** (3) (1967) 543–564.
[18] P. Wolfe, "Methods of nonlinear programming", in *Recent advances in mathematical programming* (eds. R. L. Graves and P. Wolfe), (McGraw-Hill, New York, 1963) 67–86.
[19] M. Zuckerman and P. Kirton, "Queueing analysis of a B-ISDN switching system", *Proc. 3rd ATERB FPS Workshop* (1988).