Mathematical models of multiservice networks

F.P. Kelly

University of Cambridge

This paper describes several simple models that have helped our understanding of communication networks, and describes some of the new problems that arise in connection with the multiservice networks planned for the future.

1 Introduction

Throughout this century problems that have arisen in the design and analysis of communication networks have provided motivation for the development of stochastic modelling techniques. In this paper we describe some classical models that have helped our understanding of communication networks, and discuss some of the new problems that arise in connection with the multiservice networks planned for the future.

We begin, in Section 2, by describing Erlang's formula, and its generalization to networks with fixed routing. The classical example of this model is a telephone network, but the model arises naturally in the study of local area networks, multiprocessor interconnection architectures, database structures, mobile radio and broadband packet networks ([13], [16], [18]). In computer communication networks, and increasingly in telephone networks, the circuits are virtual rather than physical: for example, a fixed proportion of the transmission capacity of a communication channel. The term 'circuit-switched' is common in some application areas, where it is used to describe systems in which before a request (which may be a call, a task or a customer) is accepted it is first checked that sufficient resources are available to deal with each stage of the request.

The trend of current developments in telecommunication networks is towards systems which will allow a number of widely disparate traffic streams to share the same broadband channel ([7], [23], [25]). A call, which might be a mixture of voice, video and data, would appear to the network as a stream of cells, and the hope is that calls with a broad range of burstiness characteristics can be efficiently integrated, through statistical multiplexing, to share a common resource. In Section 3 we describe how it is possible to associate an effective bandwidth with a source type such that, provided the sum of the effective bandwidths of the sources using a resource is less than a certain level, then the resource can deliver a performance guarantee, expressed in terms of the probability that delay exceeds a threshold or that a cell is lost.

The effective bandwidth depends on characteristics of the source such as its peak rate. Current plans [5] are that at call admission a contract would be made between user and network specifying in more or less detail the statistical properties of the call, and that policing mechanisms would enforce the contract. In Section 4 we discuss some of the issues and models that arise in connection with the policing of peak rates.

Several time-scales are involved in discussions of traffic in multiservice networks. Thus a call continuing for several minutes may be composed of bursts that last for less than a second, while bursts themselves are formed from sequences of cells each lasting only microseconds. It is possible to view the models of Sections 2, 3 and 4 as addressing respectively the call, burst and cell time-scales. A theme we begin to develop in this paper (see also [20]) concerns the integration of these different levels: we shall see how derived characteristics at one level provide the parameters for models at a higher level.

2 Loss Networks

In 1917 the Danish mathematician A.K. Erlang published his famous formula

$$E(\nu, C) = \frac{\nu^C}{C!}\left[\sum_{n=0}^{C} \frac{\nu^n}{n!}\right]^{-1} \tag{2.1}$$

for the loss probability of a telephone system ([4], p. 139). The problem considered by Erlang can be phrased as follows. Calls arrive at a link as a Poisson process of rate ν. The link comprises C circuits, and a call is blocked and lost if all C circuits are occupied. Otherwise the call is accepted and occupies a single circuit for the holding period of the call. Call holding periods are independent of each other and of arrival times, and are identically distributed with unit mean. Then *Erlang's formula* (2.1) gives the proportion of calls that are lost.

But what happens if the system consists of many links, and if calls of different types (perhaps voice, video or conference calls) require different resources? We now describe a generalization of Erlang's model which allows a network of links, and which allows the number of circuits required to depend upon the call. Consider then a network with K links, labelled $1, 2, \ldots, K$, and suppose that link k comprises C_k circuits. A call on route r uses A_{kr} circuits from link k, where $A_{kr} \in \mathbb{Z}_+$. Let $\mathcal{R}$ be the set of possible routes. 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 k, $k = 1, 2, \ldots K$, there are less than A_{kr} circuits free. Otherwise the call is connected and simultaneously holds

A_{kr} circuits from link k, $k = 1, 2, \ldots, K$, for the holding period of the call. The call holding period is independent of earlier arrival times and holding periods; holding periods of calls on route r are identically distributed with unit mean.

Let $n_r(t)$ be the number of calls in progress at time t on route r, and define the vectors $n(t) = (n_r(t), r \in \mathcal{R})$ and $C = (C_1, C_2, \ldots, C_K)$. Then the stochastic process $(n(t), t \geq 0)$ has a unique stationary distribution and under this distribution $\pi(n) = P\{n(t) = n\}$ is given by

$$\pi(n) = G(C)^{-1} \prod_{r \in \mathcal{R}} \frac{\nu_r^{n_r}}{n_r!} \qquad n \in \mathcal{S}(C) \tag{2.2}$$

where

$$\mathcal{S}(C) = \{n \in \mathbb{Z}_+^{\mathcal{R}} : An \leq C\} \tag{2.3}$$

and $G(C)$ is the normalizing constant (or partition function)

$$G(C) = \left(\sum_{n \in \mathcal{S}(C)} \prod_{r \in \mathcal{R}} \frac{\nu_r^{n_r}}{n_r!} \right). \tag{2.4}$$

This result is easy to check in the case where holding times are B exponentially distributed: then $(n(t), t \geq 0)$ is a Markov process and the distribution (2.2) satisfies the detailed balance conditions

$$\pi(n).\nu_r = \pi(n + e_r).(n_r + 1) \qquad n, n + e_r \in \mathcal{S}(C) \tag{2.5}$$

where $e_r = (\mathrm{I}[r' = r], r' \in \mathcal{R})$ is the unit vector describing just one call in progress on route r.

Most quantities of interest can be written in terms of the distribution (2.2) or the partition function (2.4). For example let L_r be the stationary probability that a call requesting route r is lost. Since the arrival stream of calls requesting route r is Poisson, B

$$1 - L_r = \sum_{n \in \mathcal{S}(C - Ae_r)} \pi(n) = G(C)^{-1} G(C - Ae_r). \tag{2.6}$$

Observe that the distribution (2.2) is simply that of independent Poisson random variables truncated to a linearly constrained region (2.3): thus from expression (2.6) we obtain Erlang's formula (2.1) in the case of a single truncated Poisson random variable. For more complex networks the explicit form (2.6) may be hard to compute [22], but there now exist many methods of approximation and analysis: these extend to generalizations of the model which allow more complex routing choices, and permit consideration of issues such as dynamic routing and network planning. For recent reviews see [18], [19].

3 Effective Bandwidths

What happens if a call's resource requirement can vary randomly over the lifetime of a call? Hui ([12], [13]) has shown that, for a simple model of an unbuffered resource, the probability of resource overload can be held below a desired level by requiring that B the number of calls n_j accepted from sources of class j, $j = 1, 2, \ldots, J$, satisfies

$$\sum_j \alpha_j n_j \le C, \tag{3.1}$$

where C is interpreted as the capacity of the resource, and α_j is the *effective bandwidth* at the resource of each source of class j. The effective bandwidth α_j lies between the mean and the peak resource requirement of a source of class j: it depends on characteristics of the source such as its burstiness, and on the degree of statistical multiplexing possible at the resource. The bandwidth of a source may vary over the different resources in a network, just as in the classical model of the last Section the requirements A_{kr} of a call may vary over the links k along its route. The linearity of the constraint (3.1) encourages the prospect that the insights available from the classical model may readily transfer to the case where a call's resource requirement may vary over the lifetime of a call.

We now review Hui's model of an unbuffered resource. We begin by recalling Chernoff's bound on the tail behaviour of sums of random variables. Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with common logarithmic moment generating function

$$M(s) = \log \mathbb{E}[e^{sX_1}]. \tag{3.2}$$

Now for any random variable Y

$$\mathbb{P}\{Y \ge 0\} = \mathbb{P}\{e^{sY} \ge 1\} \le \mathbb{E}[e^{sY}], \tag{3.3}$$

where here and throughout $s \ge 0$. Hence

$$\frac{1}{n} \log \mathbb{P}\{X_1 + X_2 + \cdots X_n \ge 0\} \le \inf_s M(s). \tag{3.4}$$

This bound is often used as an approximation, the *large deviations* approximation, and is asymptotically exact: Chernoff's theorem ([1], pp. 147–149) establishes that if $\mathbb{E}[X_n] < 0$ and $\mathbb{P}\{X_n > 0\} > 0$ then

$$\lim_{n \to \infty} \frac{1}{n} \log \mathbb{P}\{X_1 + X_2 + \cdots + X_n \ge 0\} = \inf_s M(s). \tag{3.5}$$

Let

$$S = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji} \tag{3.6}$$

where X_{ji} are independent random variables, with logarithmic moment generating functions

$$M_j(s) = \log \mathbb{E}\left[e^{sX_{ji}}\right]. \tag{3.7}$$

Interpret X_{ji} as the load placed on an unbuffered resource by a source of class j, and n_j as the number of sources of class j. Let C be the capacity of the resource and suppose $\mathbb{E}[S] < C$ and $\mathbb{P}\{S > C\} > 0$. Then Chernoff's bound gives

$$\begin{aligned} \log \mathbb{P}\{S \geq C\} \quad &\leq \log \mathbb{E}\left[e^{s(S-C)}\right] \\ &= \textstyle\sum_{j=1}^{J} n_j M_j(s) - sC, \end{aligned}$$

and the large deviations approximation is

$$\log \mathbb{P}\{S \geq C\} \approx \inf_s \left[\sum_{j=1}^{J} n_j M_j(s) - sC\right]. \tag{3.8}$$

The constraint on tail behaviour $\log \mathbb{P}\{S \geq C\} \leq -\gamma$ will certainly be satisfied if

$$\inf_s \left[\sum_{j=1}^{J} n_j M_j(s) - sC\right] \leq -\gamma. \tag{3.9}$$

Note that the term in square brackets is linear in $n = (n_1, n_2, \ldots, n_J)$. Hence the acceptance region A, consisting of values $n \in \mathbb{R}_+^J$ satisfying condition (3.9), has a convex complement in $\mathbb{R}_+^J$, since this complement is defined as the intersection of $\mathbb{R}_+^J$ with a family of half spaces. The tangent plane at a point n^* on the boundary of the region A is

$$\sum_{j=1}^{J} n_j M_j(s^*) - s^* C = -\gamma \tag{3.10}$$

where s^* attains the infimum in (3.9) with n replaced by n^*. Thus the acceptance region

$$A(n^*) = \left\{ n : \sum_{j=1}^{J} \alpha_j^* n_j + \frac{\gamma}{s^*} \leq C \right\} \tag{3.11}$$

B where

$$\alpha_j^* = \frac{M_j(s^*)}{s^*} \tag{3.12}$$

will assure satisfaction of the constraint $\log \mathbb{P}\{S > C\} \leq -\gamma$, and this linearly constrained region touches the boundary of the acceptance region A defined by (3.9) at the point n^* defining s^*. One could, for example, define n^* in terms of the expected mix of source classes. The acceptance region $A(n^*)$ assures satisfaction of the tail probability constraint whatever the mix of source classes, and is the best possible linearly constrained region for the expected mix. For many realistic examples of source classes the region $A(n^*)$ is not that sensitive to the precise choice of n^* – the boundary of A is approximately a hyperplane – see for example [10].

Next we show, following [17], that the notion of an effective bandwidth, additive over sources of different classes, generalizes to certain models of a buffered resource. Our first model of a buffered resource is as follows. Suppose that bursts from a source of class j arrive in a Poisson stream of rate ν_j and have lengths with distribution G_j. Burst lengths and the Poisson streams associated with different sources are assumed independent. The time taken to serve a burst is equal to its length, and thus the resource operates as an M/G/1 queue with arrival rate ν and service time distribution G, where

$$G(x) = \sum_{j=1}^{J} p_j G_j(x) \tag{3.13}$$

$$\nu = \sum_{j=1}^{J} \nu_j n_j, \qquad p_j = \nu_j n_j / \nu. \tag{3.14}$$

Here n_j is the number of sources of class j, and G_j is the distribution of burst length from sources of class j. The Pollaczek–Khintchine formula gives the stationary distribution of B, the buffer space required by the server, as

$$\mathbb{P}\{B \leq b\} = (1 - \nu\mu) \sum_{r=0}^{\infty} (\nu\mu)^r G_e^{(r)}(b) \tag{3.15}$$

where $\mu(< \nu^{-1})$ is the mean of the distribution G, and $G_e^{(r)}(b)$ is the distribution function of the sum of r independent random variables each with distribution function

$$G_e(b) = \frac{1}{\mu} \int_0^b (1 - G(x)) dx. \tag{3.16}$$

¿From (3.16), or directly,

$$\mathbb{P}\{B = 0\} = 1 - \nu\mu = 1 - \sum_{j=1}^{J} \nu_j n_j \mu_j. \tag{3.17}$$

The *utilization* of the resource, U, is thus $\sum_j \nu_j n_j \mu_j$. Hence a condition of the form $U \leq K$ becomes a linear constraint

$$\sum_{j=1}^{J} \alpha_j n_j \leq K \tag{3.18}$$

where

$$\alpha_j = \nu_j \mu_j \tag{3.19}$$

is the effective bandwidth of each source of class j. Of course α_j is just the traffic intensity due to a source of class j. This (near trivial) result clearly extends far beyond the M/G/1 setting: we include it since it will emerge as a limiting form from later constraints on queue behaviour. Next we turn to a less obvious case of exact linearity.

A consequence of the distributional form (3.15) is that

$$\mathbb{E}(B) = \frac{\nu(\mu^2 + \sigma^2)}{2(1 - \nu\mu)} \tag{3.20}$$

where μ and σ^2 are the mean and variance respectively of the distribution G (see, for example, [14], p. 81). Let μ_j and σ_j^2 be the mean and variance respectively of G_j, the burst size distribution for sources of class j. Then

$$\mu = \sum_{j=1}^{J} p_j \mu_j, \quad \mu^2 + \sigma^2 = \sum_{j=1}^{J} p_j (\mu_j^2 + \sigma_j^2), \tag{3.21}$$

and so

$$\nu\mu = \sum_{j=1}^{J} \nu_j n_j \mu_j, \quad \nu(\mu^2 + \sigma^2) = \sum_{j=1}^{J} \nu_j n_j (\mu_j^2 + \sigma_j^2). \tag{3.22}$$

Thus a condition $\mathbb{E}(B) \leq L$ is, from (3.20), exactly the condition

$$\sum_{j=1}^{J} \nu_j n_j (\mu_j^2 + \sigma_j^2) \leq 2 \left(1 - \sum_{j=1}^{J} \nu_j n_j \mu_j \right) L. \tag{3.23}$$

Rearranging terms, this is equivalent to

$$\sum_{j=1}^{J} n_j [\nu_j (\mu_j^2 + \sigma_j^2) + 2\nu_j \mu_j L] \leq 2L. \tag{3.24}$$

Thus the effective bandwidth of a source of type j can be defined to be

$$\alpha_j = \nu_j \left[\mu_j + \frac{1}{2L}(\mu_j^2 + \sigma_j^2)\right] \tag{3.25}$$

since under this identification the constraint $\mathbb{E}(B) \leq L$ *becomes* the linear constraint

$$\sum_{j=1}^{J} \alpha_j n_j \leq 1. \tag{3.26}$$

The analytical expression (3.25) for bandwidth α_j is illuminating. Observe, for example, the dependence of bandwidth on L, the constraint on mean workload. If L is large enough α_j reduces to (3.19), the effective bandwidth in the utilization constrained formulation. If L is small the burst size distribution, as well as its mean, is important. For example if the distribution G_j is exponential, then $\sigma_j^2 = \mu_j^2$, and so the bandwidth α_j has a quadratic dependence, proportional to $\mu_j + L^{-1}\mu_j^2$, on the mean burst size. If burst sizes are constant, so that $\sigma_j^2 = 0$, then bandwidth again has a quadratic dependence on burst size, but now proportional to $\mu_j + (2L)^{-1}\mu_j^2$.

Often constraints on the probability that buffer space or delay exceeds a threshold are more important than constraints on mean values. Fortunately there exist manageable estimates and bounds for tail behaviour, provided by Kingman [21] and Ross [26] for the more general GI/G/1 queue. If A is a random variable with the interarrival time distribution, X a random variable with the service time distribution G, and κ a positive constant such that

$$\mathbb{E}(e^{\kappa X})\mathbb{E}(e^{-\kappa A}) = 1 \tag{3.27}$$

then the stationary distribution of B, the unfinished work found by an arriving customer, satisfies

$$a_1 e^{-\kappa b} \leq \mathbb{P}\{B > b\} \leq a_2 e^{-\kappa b} \qquad b \geq 0 \tag{3.28}$$

for constants $a_1, a_2 \leq 1$.

Consider the M/G/1 queue. The constraint on tail behaviour B $\log \mathbb{P}\{B > b\} \leq -\gamma$ will certainly be satisfied if κ, the solution to equation (3.27), satisfies $\kappa b \geq \gamma$, or equivalently

$$\nu \int_0^\infty e^{\gamma x/b}(1 - G(x))dx \leq 1. \tag{3.29}$$

Suppose again that G is defined by (3.13) and (3.14). Then (3.29) becomes

$$\sum_{j=1}^{J} \nu_j n_j \int_0^\infty e^{\gamma x/b}(1 - G_j(x))dx \leq 1, \tag{3.30}$$

or equivalently

$$\sum_{j=1}^{J} \alpha_j n_j \leq 1 \tag{3.31}$$

where

$$\alpha_j = \nu_j \int_0^\infty e^{\gamma x/b}(1 - G_j(x))dx. \tag{3.32}$$

Again we obtain a linearly constrained acceptance region, and again the analytical form (3.23) for the bandwidth is illuminating. Observe that as γ shrinks to zero, α_j reduces to (3.19), the effective bandwidth in the utilization constrained formulation. As γ increases, the tail of the distribution G_j becomes more and more important.

Our model assumes that arriving bursts are not lost when the buffer level exceeds b: they may for example be held at resources leading to the particular resource under consideration, and forwarded later. The provision of a buffer area is intended to prevent this happening too often: if such blocking is indeed an infrequent occurrence and if our assumption concerning arrival streams is valid, perhaps in a network with sufficiently diverse routing, then it should be possible to analyse different resources as independent systems and to use the loss network results from Section 2. Of course buffers arranged strictly in series exhibit a quite different behaviour, owing to the strong dependence between the service mechanism at one buffer and the arrival stream at the next ([3], [15]).

For other and more general models of buffered resources that allow a similar analysis, see [8], [9] and [27].

4 Policing Peak Rates

One of the interesting issues that arises in connection with multiservice networks concerns the policing of peak rates. Suppose that a user declares that the greatest load it will put on the network will be a regular periodic stream of cells with period d. By the time this stream enters the network it may well have been subjected to some perturbation, perhaps by a multiplexing stage. How can the network effectively police the perturbed stream?

A natural model is illustrated in Figure 1. We consider a FIFO queue handling the superposition of a periodic stream of cells of period d, and a Poisson stream of rate λ, where the unit of time is the cell transmission time (Figure 1). The load (traffic intensity) is thus $\rho = \lambda + 1/d$. The Poisson stream represents the perturbing traffic, and after passage through the queue the nominally periodic stream passes through a leaky bucket policer [24], that is a queue with constant service time a and a finite buffer of depth b where cells that overflow the buffer are discarded. How should a and b be chosen so that test streams which initially conform to the user's

Figure 1. An M+D/D/1 multiplexer

declaration of period d, but are perturbed by the FIFO queue, do not suffer?

The squared coefficient of variation of the inter-exit time distribution for cells from the periodic stream has been calculated by Guillemin and Roberts [11] for the cases $\rho = 0.7, 0.85$ and is shown in Table 1. As ρ approaches 1 the exit times of the periodic stream approach a renewal process, with inter-exit time distribution $1 + P(d-1)$, where $P(\lambda)$ represents a Poisson distribution with mean λ. This inter-exit time distribution has squared coefficient of variation $(d-1)/d^2$, and this simple expression provides the case $\rho = 1$ in Table 1.

For $\rho < 1$ the exit times of the periodic stream do not form a renewal process: as discussed by Blaabjerg [2] the exit stream appears more regular over large intervals, and thus approximating the stream by a renewal process with the calculated inter-exit time distribution will be conservative. As ρ increases to 1, the variability of the stream increases, and the degree of conservatism decreases. For both reasons we should expect the renewal stream obtained when $\rho = 1$ to be a worst case bound.

We have seen that a bound on the waiting time for the GI/G/1 queue is

$$\mathbb{P}\{W > w\} \leq e^{-\kappa w} \tag{4.1}$$

where κ is the positive constant solving equation (3.27). In the case where $\rho = 1$ and the input stream is a renewal process with inter-arrival distribution $1 + P(d-1)$,

$$\mathbb{E}(e^{-\kappa A}) = e^{-\kappa} g_{d-1}(e^{-\kappa}) \tag{4.2}$$

where

$$g_\alpha(z) = e^{-\alpha(1-z)} \tag{4.3}$$

is the probability generating function of a Poisson random variable with mean α. Thus equation (3.27) becomes

$$\kappa(a-1) = (d-1)(1-e^{-\kappa}). \tag{4.4}$$

Let $\gamma = \kappa w$, so that the bound (4.1) is, as before, $\exp(-\gamma)$. Then equation (4.4) becomes

$$\frac{a-1}{d-1} = f\left(\frac{\gamma}{ba}\right) B \tag{4.5}$$

where

$$f(\kappa) = (1-e^{-\kappa})/\kappa\,, \tag{4.6}$$

a function decreasing from 1 to 0 as κ increases from 0 to infinity. Equation (4.5) is thus a relationship between a, b, d and γ that ensures the leaky bucket policer loses less than 1 in e^γ cells from an initially periodic stream of period d that has been perturbed by a critically loaded M+D/D/1 queue.

If we fix $w = ba$, the time taken to empty the leaky bucket, then we can use equation (4.5) to give the required leak rate:

$$a - 1 = (d-1) f\left(\frac{\gamma}{w}\right)\ . \tag{4.7}$$

For example, if $\gamma = 20$, $w = 20$,

$$a - 1 = (1-e^{-1})(d-1), \tag{4.8}$$

or if $\gamma = 23$ (a loss probability of 10^{-10}) and $w = 200$,

$$a - 1 = 0.94(d-1). \tag{4.9}$$

This suggests a natural policing strategy. Fix the time taken to empty a leaky bucket, $w = ba = 200$, say, with leak rate a^{-1} given by equation (4.9), and use this leaky bucket to police a source of declared rate d^{-1}. Thus dimensioned, the leaky bucket loses less than one cell in 10^{10} if input comes from a deterministic stream of rate d^{-1} perturbed by a critically loaded FIFO M+D/D/1 queue. A stream of rate up to a^{-1} could pass through the leaky bucket, but this will only increase the nominal rate d^{-1} of the stream by a factor $d/(1+0.94(d-1))$, a percentage increase of at most 6%.

Table 1. Squared coefficient of variation of inter-exit time

load (ρ)	$d=2$	$d=5$	$d=10$	$d=15$
0.70	0.100	0.064	0.027	0.014
0.85	0.175	0.104	0.050	0.030
1.00	0.250	0.160	0.090	0.062

In practice [5] leaky bucket policers may be virtual rather than real: a virtual leaky bucket discards the same cells as a real leaky bucket, but does not delay the other cells. It is interesting to consider the next multiplexing stage in a network, where the output from several leaky buckets (real or virtual) may form the arrival process at a queue which we may represent as a (real) leaky bucket. Consider a leaky bucket with parameters (a_i, b_i). The number of cells output in the period $(0,t)$ is bounded above by

$$b_i + \lfloor t a_i^{-1} \rfloor \,. \tag{4.10}$$

Thus the number of cells output in the period $(0,t)$ by a collection of leaky buckets with parameters (a_i, b_i), for $i = 1, 2, \ldots, I$, will be bounded above by

$$\sum_{i=1}^{I} \left(b_i + \lfloor t a_i^{-1} \rfloor \right) \le \sum_{i=1}^{I} b_i + \lfloor t \sum_{i=1}^{I} a_i^{-1} \rfloor \,. \tag{4.11}$$

But the latter bound is enough for those cells to pass without loss through a leaky bucket with parameters

$$\left(\left(\sum_{i=1}^{I} a_i^{-1} \right)^{-1} , \sum_{i=1}^{I} b_i \right) \,. \tag{4.12}$$

(For a very full analysis of this form of deterministic bounding see Cruz [6].) Thus if

$$B \sum_{i=1}^{I} a_i^{-1} \le \rho \,, \qquad \sum_{i=1}^{I} b_i \le bB \tag{4.13}$$

then the next multiplexer stage, modelled as a leaky bucky with parameters (ρ^{-1}, b) will lose *no* cells.

If $a_i b_i = w$ for $i = 1, 2, \ldots, I$, and if $b = \rho w$ then constraints (4.13) become the single constraint

$$\sum_{i=1}^{I} a_i^{-1} \le \rho \,. \tag{4.14}$$

In either case the linearity of constraints (4.13) or (4.14) allows use, at higher levels, of the classical model of Section 2. B

Acknowledgement. I am indebted to colleagues at Cambridge and at BT Laboratories, Martlesham, for discussions over several years on the models outlined in this paper: in particular, the ideas described in Section 4 have emerged during a recent joint project involving Richard Gibbens, Rhodri Griffiths, Peter Key and Andreas Skliros, to all of whom I am grateful.

Bibliography

1. B Billingsley P. (1986) *Probability and Measure*, 2nd ed. Wiley, New York.
2. Blaabjerg S. (1992) Estimating the effect of cell delay variation by an application of the heavy traffic limit approximation. COST 242 document.
3. Boxma O J and Konheim A G. (1981) Approximate analysis of exponential queueing systems with blocking. *Acta Informatica*, **15**, 19–66.
4. Brockmeyer E. (1948) *The Life and Works of A.K. Erlang.* The Copenhagen Telephone Company.
5. CCITT (1992) Recommendation I.371: Traffic Control and Congestion Control in B-ISDN. CCITT, Geneva.
6. Cruz R L. (1991) A calculus for network delay. *IEEE Trans. Information Theory*, **37**, 114–141.
7. Decina M and Trecordi V (eds). (1992) Traffic management and congestion control for ATM networks. *IEEE Network* **6**, No. 5, September.
8. Elwalid A I and Mitra D. (1993) Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE–ACM Trans. Networking* **1**, 329–343.
9. Gibbens R J and Hunt P J. (1991) Effective bandwidths for the multi-type UAS channel. *Queueing Systems*, **9**, 17–28.
10. Griffiths T R. (1990) Analysis of connection acceptance strategies in asynchronous transfer mode networks. *7th UK Teletraffic Symposium*, Durham.
11. Guillemin E and Roberts J W. (1991) Jitter and bandwidth enforcement. Globecom '91.
12. Hui J Y. (1988) Resource Allocation for Broadband Networks. *IEEE Journal on Selected areas in Communications*, SAC-6(9):1598–1608.
13. Hui J Y. (1990) *Switching and Traffic Theory for Integrated Broadband Networks*. Kluwer, Boston.
14. Kelly F P. (1979) *Reversibility and Stochastic Networks.* Wiley, Chichester.
15. Kelly F P. (1982) The throughput of a series of buffers. *Advan. Appl. Prob* **14**, 633–653.
16. Kelly F P. (1985) Stochastic models of computer communication systems. *J. Roy. Statist. Soc.* Series B **47**, 379–395.

17. Kelly F P. (1991) Effective bandwidths at multi-class queues. *Queueing Systems*, **9**, 5–16.
18. Kelly F P. (1991) Loss networks. *Ann. Appl. Prob.* **1**, 319–378.
19. Key P. (1990) Optimal control and trunk reservation in loss networks. *Prob. Eng. Inf. Sci.* **4**, 203–242.
20. Key P. (1993) Worths, controls and trunk reservation in a multi-service network. COST242 document, British Telecommunications.
21. Kingman J F C. (1970) Inequalities in the theory of queues. *J. Roy. Stat. Soc.* Series B **32**, 102–110.
22. Louth G M, Mitzenmacher M and Kelly F P. (1994) Computational complexity of loss networks. *Theoretical Computer Science*.
23. Mitra D and Mitrani I. (1991) Editorial introduction to communication systems. Special issue of *Queueing Systems* **9**, 1–4.
24. Niestegge G. (1990) The ‘leaky bucket’ policing method in the ATM (asynchronous transfer mode) network. *Int. J. Digital and Analog Comm. Syst.*, **3**, 187–197.
25. Roberts J W (ed). (1992) Performance evaluation and design of multiservice networks. Commission of the European Communities.
26. Ross S M. (1974) Bounds on the delay distribution in GI/G/1 queues. *J. Appl. Prob.* **11**, 417–421.
27. Whitt W. (1992) Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. A.T.&T. Bell Laboratories.