EFFECTIVE BANDWIDTHS AT MULTI-CLASS QUEUES

F.P. KELLY

Statistical Laboratory, University of Cambridge, 16 Mill Lane,
Cambridge CB2 1SB, England

Abstract

Consider a queue which serves traffic from a number of distinct sources and which is required to deliver a performance guarantee, expressed in terms of the mean delay or the probability the delay exceeds a threshold. For various simple models we show that an effective bandwidth can be associated with each source, and that the queue can deliver its performance guarantee by limiting the sources served so that their effective bandwidths sum to less than the capacity of the queue.

Keywords: large deviations, M/G/1 queue, circuit-switched network, connection acceptance control.

1. Introduction

The traditional model of a circuit-switched network assumes that each link k of the network has a capacity C_k and that each call carried of class j requires a known amount of capacity, α_{jk} say, at link k. The network is able to carry n_j calls of class j, $j = 1, 2, \ldots, J$, if

$$\sum_j n_j \alpha_{jk} \leq C_k \tag{1.1}$$

for each link k of the network. There is a rich theory of such networks (see, for example, [1], [7], [13], [19]), able to provide insight into such topics as trunk reservation, dynamic routing and network planning.

What happens if a call's resource requirements vary randomly over the lifetime of the call? Hui ([9], [10]; see also [4]) 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 the number of calls n_j accepted from sources of class j, $j = 1, 2, \ldots, J$, satisfies

$$\sum_j \alpha_j n_j \leq C, \tag{1.2}$$

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 depends on characteristics of a source of class j 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 traditional model the requirements α_{jk} of a call may vary over the links k along its route.

Our aim in this paper is to show that the notion of an effective bandwidth, additive over sources of different classes, generalizes to certain models of a buffered resource. This further encourages the prospect (noted already in [4], [8] and [9]) that the insights available from the traditional model of a circuit-switched network will transfer to a wide range of newly emerging communication networks.

We now define our model of a buffered resource. 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. Suppose the buffer space required by the server at any instant is simply the time it would take for the server to clear its backlog if no more bursts were to arrive. The stationary distribution of the buffer space required by the server is then given by the known distribution for the unfinished work in an M/G/1 queue. The mean buffer space required, or equivalently the mean queueing delay under a first-come-first-served discipline, is given by the Pollaczek-Khintchine formula. In Section 3 we show that this measure of performance is held below any given value if and only if the number of sources of each class satisfies a linear inequality of the form (1.2).

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. In Section 4 we use Cramér's asymptotic estimates [5] and the bounds of Kingman [16] and Ross [17] to show that the probability that required buffer space exceeds a threshold can be held below any given value by requiring that the numbers of sources of each class satisfy a linear inequality of the form (1.2). It is interesting to note that the effective bandwidth α_j of a source of class j has a very similar analytical form to that obtained from the earlier model of an unbuffered resource.

The bounds of Kingman [16] and Ross [17] apply more generally to GI/G/1 queues, and our results on effective bandwidths generalize to renewal arrival streams. A particular example, the slotted/batch model, is considered in Section 5. The model is a form of discrete time queue, and has a close formal relationship with the unbuffered model of Section 2.

We conclude in Section 6 with a simple example indicating how the traditional model of a circuit-switched link can shed light on the problems of connection acceptance control.

2. Unbuffered resources

In this section we review Hui's model of an unbuffered resource ([9],[10]). 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

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

Now for any random variable Y

$$P\{Y \geq 0\} = P\{e^{sY} \geq 1\} \leq E[e^{sY}],$$

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

$$\frac{1}{n} \log P\{X_1 + X_2 + \cdots X_n \geq 0\} \leq \inf_s M(s).$$

This bound is often used as an approximation, the *large deviations* approximation, and is asymptotically exact: Chernoff's theorem ([2], pp.147–149) establishes that if $E[X_n] < 0$ and $P\{X_n > 0\} > 0$ then

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

Let

$$S = \sum_{j=1}^{J} \sum_{i=1}^{n_j} X_{ji} \tag{2.1}$$

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

$$M_j(s) = \log E\left[e^{sX_{ji}}\right]. \tag{2.2}$$

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 $E[S] < C$ and $P\{S > C\} > 0$. Then Chernoff's bound gives

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

and the large deviations approximation is

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

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

$$\inf_s \left[\sum_{j=1}^{J} n_j M_j(s) - sC\right] \leq -\gamma. \tag{2.3}$$

Note that the term in square brackets is linear in $n = (n_1, n_2, \ldots, n_J)$. Here the acceptance region A, consisting of values $n \in R_+^J$ satisfying condition (2.3), has a convex complement in R_+^J, since this complement is defined as the intersection of R_+^J with a family of half spaces. The tangent plane at a point n^* on the boundary of the region A is

$$\sum_{j=1}^{J} n_j M_j(s^*) - s^* C = -\gamma \tag{2.4}$$

where s^* attains the infimum in (2.3) with n replaced by n^*. Thus the acceptance region

$$A(n^*) = \left\{ n : \sum_{j=1}^{J} \alpha_j^* n_j + \frac{\gamma}{s^*} \leq C \right\} \tag{2.5}$$

where

$$\alpha_j^* = \frac{M_j(s^*)}{s^*} \tag{2.6}$$

will assure satisfaction of the constraint $\log P\{S \leq C\} \leq -\gamma$, and this linearly constrained region touches the boundary of the acceptance region A defined by (2.3) 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 [4], [8], [9].

3. Constraints on the mean workload

Consider the M/G/1 queue described in the Introduction, with arrival rate ν and service time distribution G, where

$$G(x) = \sum_{j=1}^{J} p_j G_j(x) \tag{3.1}$$

$$\nu = \sum_{j=1}^{J} \nu_j n_j, \qquad p_j = \nu_j n_j / \nu. \tag{3.2}$$

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

$$P\{B \leq b\} = (1 - \nu\mu) \sum_{r=0}^{\infty} (\nu\mu)^r G_e^{(r)}(b) \tag{3.3}$$

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

$$G_e(b) = \frac{1}{\mu} \int_0^b (1 - G(x))dx. \tag{3.4}$$

¿From (3.3), or directly,

$$\begin{aligned} P\{B = 0\} &= 1 - \nu\mu \\ &= 1 - \sum_{j=1}^{J} \nu_j n_j \mu_j. \end{aligned}$$

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

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

where

$$\alpha_j = \nu_j \mu_j \tag{3.5}$$

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 (3.3) is that

$$E(B) = \frac{\nu(\mu^2 + \sigma^2)}{2(1 - \nu\mu)} \tag{3.6}$$

where μ and σ^2 are the mean and variance respectively of the distribution G (see, for example, [11], 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

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

and so

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

Thus a condition $E(B) \le L$ is, from (3.6), exactly the condition

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

Rearranging terms, this is equivalent to

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

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

$$\alpha_j = \nu_j \left[\mu_j + \frac{1}{2L} (\mu_j^2 + \sigma_j^2) \right] \tag{3.7}$$

since under this identification the constraint $E(B) \le L$ *becomes* the linear constraint

$$\sum_{j=1}^{J} \alpha_j n_j \le 1.$$

The analytical expression (3.7) 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 (3.5), 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 $\mu_j + (2L)^{-1}\mu_j^2$.

4. Constraints on tail probabilities

Next we consider constraints on the tail behaviour of the distribution (3.3): in many circumstances such constraints are more appropriate than constraints on utilization or mean workload.

Cramér's estimate [5], originally derived for a related ruin problem, describes the tail behaviour of the distribution (3.3). Suppose there exists a finite constant κ such that

$$\nu \int_0^\infty e^{\kappa x}(1 - G(x))dx = 1 \tag{4.1}$$

and suppose that

$$\eta = \nu \int_0^\infty e^{\kappa x}(1 - G(x))x dx \tag{4.2}$$

is finite. Then Cramér's estimate is

$$P\{B > b\} \sim \frac{1 - \nu\mu}{\kappa\eta} e^{-\kappa b} \qquad \text{as} \quad b \to \infty. \tag{4.3}$$

Kingman [16] and Ross [17] discuss closely related bounds 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

$$E(e^{\kappa X})E(e^{-\kappa A}) = 1 \tag{4.4}$$

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

$$a_1 e^{-\kappa b} \leq P\{B > b\} \leq a_2 e^{-\kappa b} \qquad b \geq 0 \tag{4.5}$$

for constants $a_1, a_2 \leq 1$.

If traffic intensity $\nu\mu$ is close to 1 the bound a_2 is close to 1 and the constant κ is approximately

$$\frac{2(EA - EX)}{\mathrm{Var}(A) + \mathrm{Var}(X)}. \tag{4.6}$$

This is consistent with heavy traffic results for the GI/G/1 queue, which show that the stationary distribution of B is approximately exponential with parameter (4.6), although the limiting regime is different ([15],[18]). Simple bounds on the constant a_1 are given by Kingman [16]. For example, if $P\{X \leq M\} = 1$ then

$$e^{-\kappa(b+M)} \leq P\{B > b\} \leq e^{-\kappa b} \qquad b \geq 0. \tag{4.7}$$

Equation (4.4) reduces to equation (4.1) when A has an exponential distribution with parameter ν. Consider further this case, the M/G/1 queue. The constraint on tail behaviour $\log P\{B > b\} \leq -\gamma$ will certainly be satisfied if κ, the solution to equation (4.1), satisfies $\kappa b \geq \gamma$, or equivalently

$$\nu \int_0^\infty e^{\gamma x/b}(1 - G(x))dx \leq 1. \tag{4.8}$$

Suppose again that G is defined by (3.1) and (3.2). Then (4.8) becomes

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

or equivalently

$$\sum_{j=1}^{J} \alpha_j n_j \leq 1 \tag{4.9}$$

where

$$\alpha_j = \nu_j \int_0^\infty e^{\gamma x/b}(1 - G_j(x))dx. \tag{4.10}$$

Again we obtain a linearly constrained acceptance region, and again the analytical form (4.10) for the bandwidth is illuminating. Observe that as γ shrinks to zero, α_j reduces to (3.5), the effective bandwidth in the utilization constrained formulation. As γ increases, the tail of the distribution G_j becomes more and more important. Note that the expression (4.10) can also be written as

$$\begin{aligned}\alpha_j &= \frac{\nu_j b}{\gamma} \int_0^\infty (e^{\gamma x/b} - 1)dG_j(x) \\ &= \frac{\nu_j b}{\gamma}\left[\exp\left(M_j\left(\frac{\gamma}{b}\right)\right) - 1\right],\end{aligned} \tag{4.11}$$

using expression (2.2).

The more refined approximation (4.3) will not in general produce a linearly constrained acceptance region, but the region it does produce is usually only slightly larger than that defined by (4.9) and (4.10).

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. 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], [12]).

5. A slotted/batch model

Suppose that time is divided into slots of unit length, and that independent batches of bursts arrive at the start of each slot. The model we consider is thus the special case of the GI/G/1 queue where the renewal process describing the arrival stream is deterministic. An attraction of the model, introduced and discussed further in [6], is its close formal relationship with the unbuffered model of Section 2. In the unbuffered model each slot is treated independently: excessive arrivals are lost. In the buffered model excessive arrivals are held over to be dealt with in the next, or following, slots.

Equation (4.4) becomes

$$E\left(e^{\kappa X}\right) = e^{\kappa}, \tag{5.1}$$

since slots are of unit length. The random variable X now describes the *batch* of arriving bursts. Thus if there are n_j sources of class j, we can set

$$X = \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ji}$$

where, parallelling equations (2.1) and (2.2), the random variables Y_{ji} are independent and Y_{ji} has logarithmic moment generating function $M_j(s)$. Equation (5.1) becomes

$$\sum_{j=1}^{J} n_j M_j(\kappa) = \kappa.$$

Thus, using the bound (4.5), the constraint $\log P\{B > b\} \leq -\gamma$ on the workload found at the end of a slot will certainly be satisfied if $\kappa b \geq \gamma$, or equivalently

$$\sum_{j=1}^{J} n_j M_j\left(\frac{\gamma}{b}\right) \leq \frac{\gamma}{b}\,.$$

But this is just the inequality (4.9) with

$$\begin{aligned} \alpha_j &= \frac{b}{\gamma} \log \int_0^\infty e^{\gamma x/b} dG(x) \\ &= \frac{b}{\gamma} M_j\left(\frac{\gamma}{b}\right)\,. \end{aligned} \tag{5.2}$$

It is interesting to compare the form (5.2) with our earlier results. First, note that in an M/G/1 queue the stationary distribution of unfinished work found by an arriving customer is the same as the stationary distribution of work at the server at integer time points. But in an M/G/1 queue the arriving workload from a single source of class j over a unit time interval, Y_{ji} say, has the compound Poisson distribution

$$\log E[e^{sY_{ji}}] = \nu_j[\exp(M_j(s)) - 1]\,:$$

here $M_j(s)$ is, as in Section 4, the logarithmic moment generating function of a single burst from a source of class j. This confirms the necessary correspondence between the expressions (4.11) and (5.2). Secondly, the form (5.2) parallels the earlier form (2.6) arising from the model of an unbuffered resource. The similarity is perhaps to be expected in view of the close formal relationship between the models. Note also the similarity between the asymptotic regimes involved: the large deviations approximation of Section

2 becomes more accurate as the number of sources increases and the tail probability decreases; the bounds (4.5) and (4.7) determine κ, and hence effective bandwidths, more precisely as the buffer size increases and the tail probability decreases.

6. Concluding remarks

For various simple models of a multi-class queue we have seen that there exists a notion of effective bandwidth, such that the queue can deliver its performance guarantee by limiting the sources served so that the inequality (1.2) is satisfied. To illustrate the utility of this concept, we briefly mention two issues of current interest. First, how finely should sources be classified? For example, if the sources of a particular class could be identified as belonging to distinct subclasses, then this is likely to be worthwhile only if the subclasses have substantially differing effective bandwidths. The analysis of earlier sections helps quantify and explore this issue. Secondly, consider the problem of connection acceptance control. Suppose that new sources of class j request connection in a Poisson stream of rate g_j and that, if accepted, a source of class j remains connected for a holding period with mean h_j. One possible connection acceptance control is to accept a new source provided the vector $n = (n_1, n_2, \ldots, n_J)$ remains such that equation (1.2) is satisfied. This connection acceptance control will certainly ensure that the queue delivers its performance guarantee. However it can have a serious drawback in overload: if the capacity C is large, but the effective offered traffic $\sum_j g_j h_j \alpha_j$ is larger still, then the sources accepted will be biased towards those with low effective bandwidths. Indeed the probability of connection for a new source of class j will be approximately $\exp(-y\alpha_j)$, for some positive constant y, and thus will decay rapidly with the effective bandwidth α_j (see [14]). This difficulty is well understood in circuit-switched networks, and is dealt with by the technique known as *trunk reservation*: accept a call of class j if and only if the vector n remains such that

$$\sum_{i=1}^{J} \alpha_i n_i \leq C - r_j,$$

where $r_j \geq 0$. By choosing higher values of r_j to accompany lower values of α_j, for example, high bandwidth connections can be protected in overload. Work in progress concerns how such connection acceptance controls can be generalized to deal with sources and queues whose parameters are dynamically estimated.

Acknowledgements

This paper emerged from work sponsored by British Telecom, and from discussions with James Appleton, Richard Gibbens, Rhodri Griffiths, Steve Johnson and, indirectly, Ward Whitt, to all of whom I am grateful.

References

[1] E. Brockmeyer, H.L. Halstrom and A. Jensen, *The Life and Works of A.K. Erlang.* (Academy of Technical Sciences, Copenhagen, 1948).

[2] P. Billingsley, *Probability and Measure*, 2nd edition. (Wiley, New York, 1986).

[3] O.J. Boxma and A.G. Konheim, Approximate analysis of exponential queueing systems with blocking. *Acta Informatica* **15**, (1981), 19–66.

[4] Z. Dziong, J. Choquette, K.-Q. Liao and L. Mason. Admission control and routing in ATM networks. ITC Specialist Seminar, Adelaide (1989).

[5] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol.II, 2nd edition. (Wiley, New York, 1971).

[6] R.J. Gibbens, P.J. Hunt and F.P. Kelly, On models of buffering. In preparation.

[7] R.J. Gibbens and F.P. Kelly, Dynamic routing in fully connected networks. *IMA Journal of Mathematical Control and Information* **7**, (1990), 77–111.

[8] T.R. Griffiths, Analysis of connection acceptance strategies in asynchronous transfer mode networks. 7th UK Teletraffic Symposium, Durham, (1990).

[9] J.Y. Hui, Resource allocation for broadband networks. *IEEE Journal on Selected Areas in Communications* **6**, (1988), 1598–1608.

[10] J.Y. Hui, *Switching and Traffic Theory for Integrated Broadband Networks.* (Kluwer, Boston, 1990).

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

[12] F.P. Kelly, The throughput of a series of buffers. *Advances in Applied Probability* **14**, (1982), 633–653.

[13] F.P. Kelly, Routing in circuit-switched networks: optimization, shadow prices and decentralization. *Advances in Applied Probability* **20**, (1988), 112–144.

[14] F.P. Kelly, Loss networks. *Annals of Applied Probability* **1**, (1991).

[15] J.F.C. Kingman, The heavy traffic approximation in the theory of queues. In *Proceedings of the Symposium on Congestion Theory*, eds. W.L. Smith and W. Wilkinson. (University of North Carolina Press, Chapel Hill, 1965), 137–169.

[16] J.F.C. Kingman, Inequalities in the theory of queues. *Journal of the Royal Statistical Society, Series B* **32**, (1970), 102–110.

[17] S.M. Ross, Bounds on the delay distribution in GI/G/1 queues. *Journal of Applied Probability* **11**, (1974), 417–421.

[18] W.L. Smith, On the distribution of queueing times. *Proceedings of the Cambridge Philosophical Society* **49**, (1953), 449-461.

[19] W. Whitt, Blocking when service is required from several facilities simultaneously. *A.T.&T. Bell Laboratories Technical Journal* **64**, (1985), 1807–1856.