46

Dimensioning of ATM Networks with Finite Buffers under Call-Level and Cell-Level QoS Constraints

A. Girard
INRS-Télécommunications
16 Place du Commerce, Verdun, (Que) H3E 1H6 Canada. email: andre@inrs-telecom.uquebec.ca, Ph: (514)-765-7832, Fax: (514)-761-8501.

C. Rosenberg
École Polytechnique de Montréal
Département de Génie Electrique et Génie Informatique, PO Box 6079, Succ. "Downtown", Montréal, H3C 3A7 Canada. email: cath@comm.polymtl.ca, Ph: (514)-340-4123, Fax: (514)-340-4562.

Abstract

In broadband networks with guaranteed Quality of Service (QoS), dimensioning becomes an even harder task than in traditional data or circuit networks. The network operator has two resources to manage — the trunk capacity and the switch memory (buffers) — and multiple QoS constraints to respect — call blocking, cell loss ratio and cell delays. We investigate how this new context can affect the design, cost and dimensioning of networks. In particular, we study how much can be gained if we do not restrict ourselves to the two limit cases of a bufferless or an infinite buffer system. We deal first with the single-service case for a single link. We then investigate the multi-service case on a single link. We study the cost of integration when service classes have very different delay constraints. The main results obtained are that finite buffers can have a very significant effect on the optimal cost whenever buffer costs are high, as in satellite systems, especially when a delay constraint is imposed on the problem and that services integration can be very costly when different delay requirements are involved.

Keywords

Broadband Networks, dimensioning, design of networks with finite buffers, quality of service, satellite systems

Broadband Communications P. Kühn & R. Ulrich (Eds.)

1 INTRODUCTION

Designing and dimensioning ATM networks present some new challenges to the network operators. Some of them are related to the fact that these networks have to support multiple services with very different characteristics and others to the fact that these services require commitments on QoS (rather than simply objectives) at the call and cell level (other challenges are dealt with in (Girard & Rosenberg 1997)). The context in which this study has been done is the one of guaranteed bandwidth ATM Transfer Capabilities (ATCs) (ITU-T 1996), i.e., the Deterministic Bit Rate (DBR) and the Statistical Bit rate (SBR). We focus on the SBR ATC which is used to attain reasonable efficiency through statistical multiplexing while providing the stringent QoS guarantees demanded by some VBR applications.

The QoS requirements at the cell level is on Cell Loss Ratio (CLR) and may also be on cell delay (mean delay, maximum delay and/or Cell Delay Variation (CDV)). These requirements put very stringent constraints on the buffer dimensioning. Hence, the network operator has to manage two resources from QoS and cost standpoints, namely the trunk capacity and the switch memory (buffers). This is particularly important for systems where buffers are very expensive or where buffer size is severely limited, as in satellite switches.

A large literature exists on design methods for broadband networks (see (Hui 1988, Ash, Chang & Labourdette 1994, Girard & Lessard 1992, Girard & Zidane 1995) for some references). All these methods are based on the same technique. First, an effective bandwidth is defined for each service class. The multi-service design problem is then equivalent to a multi-rate circuit-switched network design problem. Standard design methods, coupled to fast approximation to the multi-rate blocking formulas, can then be used to obtain a good near-optimal solution.

The three main weaknesses of this approach are 1) that most of the currently proposed effective bandwidth models are based on approximations assuming either no buffer or an infinite buffer, 2) that usually a single cell QoS constraint, namely the cell loss probability is considered and 3) that the underlying source model is not usually in line with the ITU-T traffic descriptors specification (ITU-T 1996).

We use simple models to investigate how finite buffers can affect the design, cost and dimensioning of this new type of networks. We have decomposed the paper in 2 parts. The first one deals with the single-service case and a single buffer. Section 2 formulates the problem in this case and gives some numerical results showing that finite buffers and delay constraints can have a significant cost impact. In the second part, corresponding to Section 3, we study a multi-service model in the single-buffer case. Section 4 presents our conclusions and future work.

2 SINGLE-SERVICE SINGLE LINK CASE

In this section, we consider a single multiplexer or a switch output including buffer and line in which all sources (e.g., connections) are identical. In this way, we can study the effect of the multi-QoS commitments and multi-resource context with a simple call admission control procedure without the added complexity of effective bandwidth (see Section 3). No assumptions are yet made on the source model and the approach that we propose below is general. We define the following parameters:

w the revenue generated by each connected call.
A the parameter λ/μ of the Poisson process describing the arrival of connection requests to the server.
L the call loss probability.
$\overline{L}$ the maximum call loss probability. This is the call-level QoS constraint.
B the call blocking probability. This is the probability that a call cannot be connected on the link at the time of its arrival. In the single-buffer case, this is identical with the call loss probability L. In a network, this may or may not be the case, depending on the call admission and routing techniques used.
C the capacity of the output line, i.e., the rate of the server.
D the maximum transit delay for each cell going through the switch or multiplexer.
$\overline{D}$ the maximum transit delay allowed for cells in the switch. This is one of the 2 cell-level QoS constraints.
P the cell loss probability.
$\overline{P}$ the maximum cell loss probability. This is one of the 2 cell-level QoS constraints.
K the buffer size, expressed in number of cells.
$E(A, N)$ the Erlang-B function for traffic intensity A and N servers.
$\boldsymbol{\beta}$ a vector that describes the source and various QoS parameters.

2.1 Call admission control

Since we want to examine the trade-off between buffer size and capacity (i.e., server rate), we determine the optimal buffer size and rate by solving the problem

$$\min_{K,C} \; z(K,C) \text{ subject to } B(A,K,C,\boldsymbol{\beta}) \leq \overline{L} \text{ and } K,C \geq 0 \tag{1}$$

where $N(K,C,\boldsymbol{\beta})$ is the largest number of calls with parameter vector $\boldsymbol{\beta}$ that can be accepted such that all the cell QoS constraints are met and $z(K,C)$ is some yet unspecified objective function that depends on K and C.

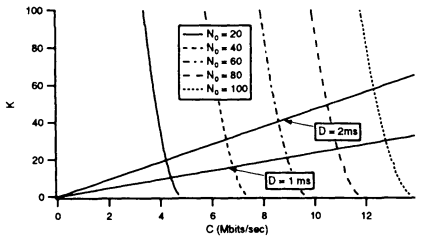

Figure 1 Contours of $N(K, C, \overline{P})$ at different levels N_0. Sources are Class 1

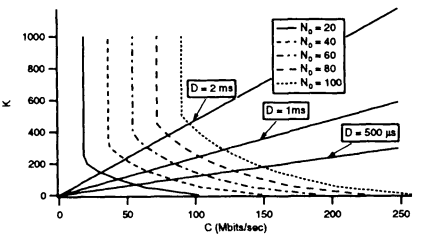

Figure 2 Contours of $N(K, C, \overline{P})$ at different levels N_0. Sources are Class 2

We have an explicit form for the blocking function if we assume that the call arrival process is Poisson.

$$B(A, K, C, \beta) = E[A, N(K, C, \beta)]. \tag{2}$$

The blocking constraints (1) can then be rewritten as $E[A, N(K, C, \beta)] \leq \overline{L}$. which can be further simplified if we define N_0 as the solution of the equation $E(A, N) = \overline{L}$. The constraint (1) on call loss probability can then be expressed, for a given pair $(A, \overline{L})$, simply as

$$N(K, C, \beta) \geq N_0. \tag{3}$$

Eq. (3) takes care of the call loss probability constraint while the set of all cell QoS constraints are present through the parameter β of $N(K, C, \beta)$.

2.2 The problem domain

We can obtain further insight on the problem structure by taking advantage of the particular cell QoS constraints that we have chosen, $\overline{P}$ and $\overline{D}$. We can replace the constraint $N(K, C, \overline{P}, \overline{D}) \geq N_0$ by two decoupled constraints. These constraints still depend implicitly on the source parameters that have not been indicated to simplify the notation. The first is of the form $N(K, C, \overline{P}) \geq N_0$ and deals only with the cell loss QoS constraint. It states that the acceptance region, defined by $N(K, C, \overline{P})$, should be large enough so that the call-level QoS constraints can be met. Since D is the largest delay that each cell can experience going through the switch, we can write $D = K/C$ and the QoS delay constraint can be expressed as $D \leq \overline{D}$ which gives, in terms of buffer size, $K \leq \overline{D}C$. Note that this decoupling of the two cell QoS constraints is possible only because the maximum cell delay is independent of the number of connections in the system.

We now examine the structure of the problem. We have selected two service

Class	PCR (Mbits/sec)	SCR (Mbits/sec)	MBS (cells)	$\overline{P}$
1	0.320	0.064	50	10^{-7}
2	9.	0.9	20	10^{-7}

Table 1 Source parameters for the Worst-Case Traffic (WCT) model

classes characterized by the parameters indicated in Table 1. We do not claim that those values are representative of real services. Instead, they were chosen simply to have very distinct behavior but still within a reasonable range.

First we look at the contours of $N(K, C, \overline{P}) = N_0$ in the (K, C) space (i.e., the isocontour $K(C)$ at level N_0). This is shown on Fig. 1 for Class 1 services. The curves are labeled with values of N_0, each value corresponding to a set of pairs $(A, \overline{L})$. In addition, we have indicated the lines corresponding to the cell delay constraint labeled in milli- or micro-seconds.

We can examine the effect of the source rate on the domain by looking at Fig. 2 where we have assumed that each source operates at a higher rate (Class 2 services). On these graphs, the vertical segment of the curve corresponds to the condition $\rho = 1$, since we do not expect that real networks will operate with utilization higher than 1. As we can see, there is no qualitative difference between the two figures. We have also examined the effect of the cell loss QoS when the value of $\overline{P} = 10^{-9}$. We could observe little qualitative difference in the form of the domains.

2.3 Optimal dimensioning

The optimization problem is then

$$\min_{K,C} z(K, C) \text{ subject to } N(K, C, \overline{P}) \geq N_0 \text{ and } K \leq \overline{D}C \text{ with } K, C \geq 0$$

where N_0 has been calculated from the values of A and $\overline{L}$. We choose to minimize the dimensioning cost, in which case we have $z = C_K K + C_C C$ and the optimization problem is

$$\min_{K,C} z = C_K K + C_C C \text{ subject to } N(K, C, \overline{P}) \geq N_0 \ (\lambda) \text{ and } K \leq \overline{D}C \ (\mu)$$

where the dual variables are indicated with their corresponding constraints. Note that the variables λ and μ defined here have nothing to do with the λ and μ that are used to represent the arrival and service rate for the call process, as used in the definition of A.

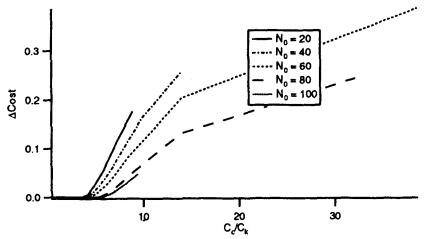

Figure 3 Dimensioning cost of the delay constraint, sources are Class 2, $\overline{D}_2 = 1$ ms

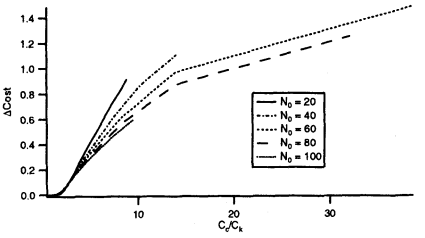

Figure 4 Dimensioning cost of the delay constraint, sources are Class 2, $\overline{D} = 500$ μs

2.4 Cost of the delay constraint

We can gather from Figs. 1 and 2 some information on the structure of the optimal solution as a function of the cost ratio $a \stackrel{\Delta}{=} C_C/C_K$ (here expressed in (\$/Mbits)/(\$/cell buffer space)). When a is small, the transmission cost is small and the optimal solution is at $K = 0$. In that case, the delay constraint is not binding. If a is large, the optimal solution is to choose C as small as possible consistent with the condition $\rho \leq 1$. We call the *trade-off region* the range of values of a where these two conditions are not met. In that range of values, it is possible to trade-off buffers for capacities. The point is that the size of this trade-off region is determined by the variation of the slope of the contour over the range of capacities. We see that for low-bandwidth sources, the trade-off region occurs at relatively high values of a while this region is in a much lower range of values for the high-bandwidth source. This notion will be useful again when we discuss multi-service systems.

We are now in a position to quantify the cost of imposing a maximum delay constraint QoS in addition to the traditional cell loss constraint. The results are presented as a function of the unit cost ratio a. For a given value of a, we can compute C_0, the optimal dimensioning cost without the delay constraint and C_1,the optimal dimensioning cost when the delay constraint is present. We use the ratio

$$\Delta = \begin{cases} (C_1 - C_0)/C_0 & \text{if the delay constraint is active} \\ 0 & \text{otherwise.} \end{cases}$$

We present the case corresponding to Fig. 2 since the other cases are not qualitatively different. The value of Δ is plotted on Fig. 3 as a function of a when the delay is $\overline{D} = 1$ msec. We can note three things on this graph.

First, the trade-off region is small. Near the origin, a is small and the optimal solution is $K = 0$. In that case, the delay constraint is not binding. At the other end, a is large and the optimal solution is to choose C as small as possible. In that case, the buffer would be chosen $K = \infty$ but this is prevented by the delay constraint and the optimal solution is always at the point where the delay constraint is tight. For values above that point, increasing a does

not induce a corresponding change in the buffer size and the cost increment is linear in a.

Second, the cost increase is more important for small call traffic values, i.e., small N_0. Recall that the value of N_0 increases as the call traffic increases, for a given call loss probability value. The third is that the cost increase gets larger as the buffer cost gets relatively less expensive (high a). This result could have been deduced from Fig. 2 directly. What is not so immediate is the size of Δ, in this case as high as 20%.

Similar results are shown on Fig. 4 for a delay of 500 μsec. We see here a substantial increase in cost due to the delay constraint, as high as 140%.

The obvious conclusion is that for homogeneous traffic, adding the delay constraint can have a significant impact on the dimensioning cost of the system. We can also speculate that in a network dimensioning model, this will tend to regroup traffic on large links and eliminate small capacity links, since the delay constraint is much more costly for these small system. This would produce networks with a low-connectivity topology and tend to favor long paths to connect some origin-destination pairs, something that should generally be avoided.

3 MULTIPLE-SERVICE SINGLE LINK CASE

While the results of Sections 2 are interesting in their own right, we are really interested in the effect of finite buffers in the case of multi-service networks, (i.e., in the case of multiple connection types). For multiple call types, we define the additional parameters:

m the number of call types.
$\mathbf{A}$ a vector of offered call traffics whose component A_i, $i = 1, m$ represents the arrival rate of calls of type i.
$\mathbf{x}$ a vector $[x_1, x_2 \ldots x_m]$ where x_i is the number of calls of type i present in the system.
$Q_i(\mathbf{x})$ the i^{th} cell QoS function (CLR, average delay, CDV, etc).
$\overline{Q}_i$ the largest valued permitted for the i^{th} QoS function.

We now assume that we can model the cell process for each call type by a single parameter W_i called the *effective bandwidth* (Ahmadi & Guérin 1990, Rege 1994, Bean 1994, Dziong, Liao & Mason 1991, Choudhury, Lucantoni & Whitt 1994, Gibbens & Hunt 1991, Kelly 1991, Elwalid & Mitra 1993). With this technique, we define

$\mathbf{W}$ the vector of effective bandwidths for all call classes.

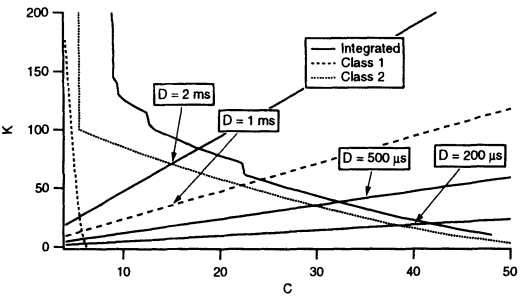

Figure 5 Problem domains for the multi-service system, Class 1 and 2, low call traffic

If we can assume that all effective bandwidths are a multiple of a given base rate C_0, then we can define $N(C)$ the number of circuits for a server rate C and we have $N(C) = \lfloor C/C_0 \rfloor$. In this model, we may have an integer or a fractional number of servers. We may also assume that the effective bandwidths can take any relative values, in which case the number of servers is a real number. This is not really a problem as long as we have continuous approximations (in N) for the blocking function. This is in effect similar to the extension of the Erlang B function to real values of the trunk group. The blocking probability is then defined as

$$B^i = \Pr\left\{\sum_{i=1}^{m} x_i W_i > N(C)\right\} \tag{4}$$

which can be evaluated, for Poisson arrival processes, exactly (Kaufman 1981, Roberts 1981) or approximately (Labourdette & Hart 1990, Gazdzicki, Lambadaris & Mazumdar 1993). Note that to be used in our context, the effective bandwidths should depend now both on the buffer size and the server rate, in addition to the source parameters, which are not indicated here. This is an additional difficulty since most currently known models for effective bandwidth are calculated for a *single* cell QoS constraint, generally the cell loss probability and usually for K assumed either very large or very small. Nevertheless, if we assume that we do have a suitable effective bandwidth model, the trunk group dimensioning problem becomes, for the cost minimization version

$$\min_{C,K} z = C_C C + C_K K \text{ subject to } B^i(\mathbf{A}, \mathbf{W}(C,K), N(C)) \leq \overline{L}^i.$$

3.1 Problem domain

We can see the effect of integration and the interaction with the delay constraints on Fig. 5. First, we have shown the problem domain (i.e., the function

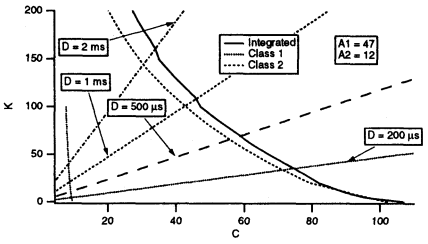

Figure 6 Problem domains, separate and integrated systems, Class 1 and 2, intermediate traffic

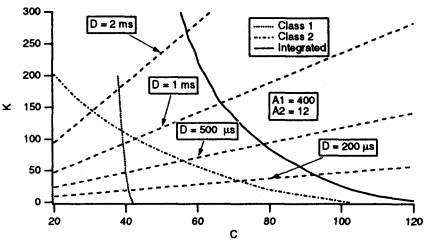

Figure 7 Problem domains, separate and integrated systems, Class 1 and 2, high traffic

$K(C)$) for the two call types separately and for the integrated model. We can see that the wideband service clearly dominates the shape of the integrated domain. Also note the steps in this domain while the corresponding curves for the isolated systems are smooth. This effect is real and probably depends on the discrete changes in values for the effective bandwidths caused by the fact that only integer values of connection numbers are possible.

A more interesting effect is seen when we consider the delay constraints indicated on the figure. Suppose that the two classes have distinct delay constraints $\overline{D}_1 = 200$ μs and $\overline{D}_2 = 2$ ms. In the integrated system, we have no choice but design for the more stringent value of these constraints. The feasible region is then given by the intersection of the $\overline{D} = 200$ μs line and the integrated admission curve. This value will in turn force a high value of bandwidth with a corresponding high cost as soon as the transmission cost is significant with respect to the buffer cost. If, on the other hand, we design each system independently, we can use the intersection of the $\overline{D} = 2$ ms curve with the Class 2 admission curve and the $\overline{D} = 200$ μs intersection with the Class 1 admission curve. Because these curves are very different, there will be substantial savings in having two separate systems instead of a single integrated one.

If, on the other hand, the values were $\overline{D}_1 = 2$ ms and $\overline{D}_2 = 200$ μs, then we would expect that the integrated system would provide savings as compared with the separate systems.

We have also investigated the effect of the call arrival rate on the admission regions. This is shown on Figs. 6–7. On Fig. 6, we see the effect of increasing A_2 while maintaining A_1 constant. The slope of the Class 2 curve varies over a wider range that for the case $A_2 = 2$ which means that the trade-off region will be larger. Also, in that case, the effect of the narrow-band traffic is not very important and the integrated system is driven by the wideband class almost exclusively.

We have then fixed A_2 to 12 and increased the value of A_1 to have a load comparable to the Class 2 load. Figure 7 clearly shows now that *both* classes have an effect on the domain of the integrated system. The narrow-band class

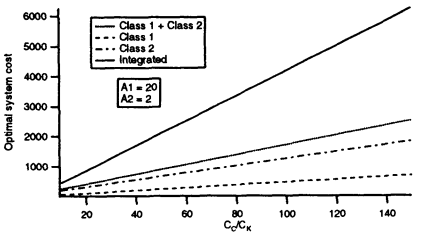

Figure 8 Cost of integration with delay constraints. Low-traffic system, $\overline{D}_1 = 2$ ms, $\overline{D}_2 = 200$ μs

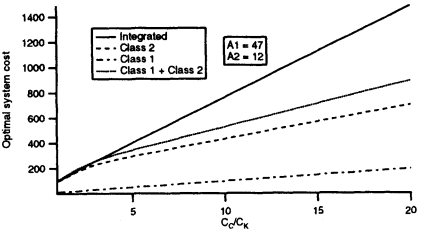

Figure 9 Cost of integration with delay constraints. Medium-traffic system, $\overline{D}_1 = 2$ ms, $\overline{D}_2 = 200$ μs

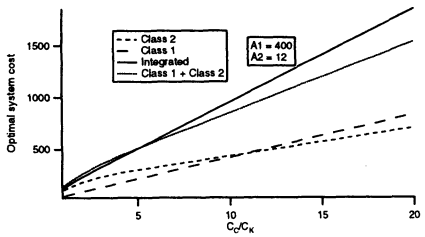

Figure 10 Cost of integration with delay constraints. High-traffic system, $\overline{D}_1 = 2$ ms, $\overline{D}_2 = 200$ μs

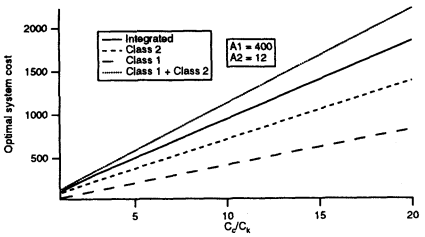

Figure 11 Cost of integration with delay constraints. High-traffic system, $\overline{D}_1 = 200$ μs, $\overline{D}_2 = 2$ ms

seems to act as a limiting factor preventing low values of C while the wideband class pushes up the buffer size.

3.2 Cost of integration

We now investigate the cost of integration when the delay constraints are very different for the two classes. We give on Figs. 8–10 the cost of separate systems designed for the delay constraint of each class separately. We also give the cost of the integrated system designed for the more stringent delay constraint, and the sum of the costs of the individual systems. There are substantial costs incurred by integration for the values of the delays chosen here, although in the separate systems, we would need to build two transmission systems with twice the termination cost.

The second is that the suitability of integration depends, in addition to the cost parameters, also on the arrival rate of call traffic. This is seen on Figs. 9 and 10. In the first case, the advantage of integration occurs for a very small range of values of a and is barely noticeable. In the second case, this range is larger and the benefit of integration is larger over that range. This is another example of the importance of taking into account the interaction of the call and cell processes when deciding on such issues as integration vs segregation.

The higher cost of the integrated system is also heavily dependent on which of the two services has the tightest delay constraint. If we interchange the values of the constraints so that we have $\overline{D}_1 = 200$ μs and $\overline{D}_2 = 2$ ms, we see from Fig. 11 that the multiplexing gain is sufficient to offset the fact that the integrated system must be designed for the smallest delay.

4 CONCLUSIONS

We have investigated here the effect of finite buffers on the optimization of call-level and cell-level QoS-constrained multi-service networks.

We have found that, when there is no delay constraint, the positive cost of the buffer yields a small increase on the total cost which is nearly linear with the cost of the transmission facility. When the presence of the buffer manifests itself in the form of a maximum delay constraint, however, we have shown that this can lead to large increases in the optimal cost (in some cases, up to 100%). We have also noted that this effect is more important for small call traffic values and speculated that this would lead to networks with concentrated traffic and hence sparse topologies.

For the multiple-service case, the main conclusion is that integrating sources on a single server may be more costly than serving them on separate facilities. This effect becomes more important as buffers get more expensive, as is the case for satellites systems. This conclusion, however, is strongly dependent on the delay constraints, the source parameters as well as the call arrival process.

Acknowledgment

This work was partly funded by NSERC Strategic grant No. STR0166996.

REFERENCES

Ahmadi, H. & Guérin, R. (1990), Bandwidth allocation in high-speed networks based on the concept of equivalent capacity, *in* 'Proc. 7th ITC Specialist Seminar'.

Ash, G., Chang, K. & Labourdette, J.-F. (1994), Analysis and design of fully shared networks, *in* Labetoulle & Roberts (1994), pp. 1311–1320.

Bean, N. (1994), Effective bandwidth with different quality of service requirements, *in* V. Iversen, ed., 'Proc. IFIP'94: Integrated broadband communication networks and services', Elsevier Science B.V., pp. 241–252.

Choudhury, G., Lucantoni, D. & Whitt, W. (1994), On the effectiveness of effective bandwidths for admission control in ATM networks, *in* Labetoulle & Roberts (1994), pp. 411–420.

Dziong, Z., Liao, K. & Mason, L. (1991), Buffer dimensioning and effective bandwidth allocation in ATM-based networks with priorities, *in* 'Proc. ITC Specialist Seminar', pp. 154–165.

Elwalid, A. & Mitra, D. (1993), 'Effective bandwidth of general markovian traffic sources and admission control of high-speed networks', *IEEE/ACM Transactions on Networking* **1**(3), 329–343.

Gazdzicki, P., Lambadaris, I. & Mazumdar, R. (1993), 'Blocking probabilities for large multirate Erlang loss systems', *Adv. in Appl. Prob.* **25**, 997–1009.

Gibbens, R. & Hunt, P. (1991), 'Effective bandwidth for multi-type UAS channels', *Questa* **9**, 17–28.

Girard, A. & Lessard, N. (1992), Revenue optimization of virtual circuit ATM networks, *in* 'Proc. Networks'92'.

Girard, A. & Rosenberg, C. (1997), A unified framework for network design with generalized connections, *in* V. Ramaswami & P. Wirth, eds, 'Teletraffic contributions for the information age: Proc. ITC 15', Vol. 2 of *Teletraffic science and engineering*, Elsevier, pp. 319–328.

Girard, A. & Zidane, R. (1995), 'Revenue optimization of B-ISDN networks', *IEEE Transactions on Communications* **43**(5), 1992–1997.

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

ITU-T (1996), 'Recommendation I.371: Traffic and congestion control in B-ISDN'.

Kaufman, J. (1981), 'Blocking in a shared resource environment', *IEEE Transactions on Communications* **29**(10), 1474–1481.

Kelly, F. (1991), 'Effective bandwidths at multi-class queues', *Queueing Systems* **9**, 5–16.

Labetoulle, J. & Roberts, J., eds (1994), *Proc. 14$^{\text{th}}$ International Teletraffic Congress*, Elsevier.

Labourdette, J. & Hart, G. (1990), 'Link access blocking in very large multimedia networks', *Computer Communication Review* **20**(4), 108–117.

Mignault, J. (In preparation), A reference resource allocation method for ATM networks, PhD thesis, Ecole Polytechnique de Montréal, 2900 Edouard-Montpetit, CP 6079, Succ A Montréal, Qué. Canada H3C 3A7.

Mignault, J., Gravey, A. & Rosenberg, C. (1996), 'A survey of straightforward statistical multiplexing models for call access control in ATM networks', *Telecommunication Systems* **5**(1–3), 177–208.

Rege, K. (1994), 'Equivalent bandwidth and related admission criteria for ATM systems—a performance study', *International Journal of Communication Systems* **7**, 181–197.

Roberts, J. (1981), A service system with heterogeneous user requirements: Application to multi-services telecommunications systems, *in* G. Pujolle, ed., 'Performance of Data Communication Systems and their Applications', North-Holland Publishing Co., pp. 423–431.