The Loading Region: A Framework for the Performance Evaluation of Admission Control Policies for High Speed Networks

Ying H. Cao Carey L. Williamson
Department of Computer Science
University of Saskatchewan
Saskatoon, SK, Canada S7N 0W0

Abstract

This paper proposes a simulation framework called the loading region *for the performance evaluation of admission control policies for high speed networks. The loading region provides a means to assess the performance characteristics of an admission control policy in terms of its sensitivity to traffic mix, traffic characteristics, cell loss constraints, and cell delay constraints. The loading region also facilitates the direct comparison of different admission control policies, both graphically and quantitatively, with respect to these criteria.*

In this paper, three different admission control policies are evaluated using simulation to demonstrate the use of the loading region framework. Among the three policies evaluated, the Joint Scheduling and Admission Control policy is the most aggressive, and has the greatest sensitivity to QOS constraints.

1 Introduction

Admission control is a vital component in any congestion control strategy for high speed networks [4]. The purpose of an admission control policy is to decide, at the time of call arrival, whether a new call should be admitted into the network or not. A call is admitted if and only if its quality of service (QOS) constraints can be met without affecting the quality of service of the existing calls in the network.

The difficulty of admission control decision-making arises from the diverse nature of the traffic expected in high speed networks. Bandwidth requirements for calls may range across four or more orders of magnitude (e.g., from tens of kilobits per second for voice traffic, to several megabits per second for data or imaging traffic, to tens or hundreds of megabits per second for video traffic [8]). Furthermore, the service requirements (i.e., the QOS constraints) of the different traffic types also vary [8]. Some traffic is delay-sensitive. Some traffic is loss-sensitive. Some traffic is both. Some traffic is neither. The challenge for an admission control policy is to achieve the highest possible network utilization without violating the QOS requirements of any call.

Many different admission control policies are proposed in the literature [2, 3, 5, 6, 7]. Most policies can be classified as either *schedulable-region-based* or *quality-estimate-based*. The schedulable-region-based policies [5, 6] use information about cell-level scheduling (i.e., service disciplines) at switches and multiplexers to determine the region of acceptable operating load for the switch. A new call is admitted if and only if the resulting load remains within the schedulable region. The quality-estimate-based policies [2, 3, 7] attempt to estimate the impact of a new call on the QOS of existing calls in the network, and admit the new call if and only if the QOS of existing calls will not be violated.

Comparing the performance of different admission control policies is often difficult, for several reasons. First, different policies may have different assumptions about the type of network environment (e.g., LAN, MAN, or WAN) in which they operate, or the type of traffic that they support (e.g., homogeneous vs. heterogeneous). Second, different policies make different assumptions about the statistical characteristics of the traffic they admit. For example, the Effective Bandwidth admission control policy [2] assumes that all traffic sources can be modelled as Poisson or Markov-Modulated Poisson Processes (MMPP). Third, different admission control policies often make different assumptions about what constitutes an acceptable level of QOS (e.g., 10^{-9} cell loss rate for video in [8]). For these reasons, it is often difficult or impossible to make a direct performance comparison between different admission control policies in the literature.

This paper attempts to address this problem by proposing a unified framework called the *loading region* for the performance evaluation of admission control policies. Simulation is used to establish the loading region, which is then used to graphically portray and analyze the behaviour of different admission control policies. The loading region provides a means to assess the performance characteristics of an admission control policy in terms of its sensitivity to traffic mix, sensitivity to traffic characteristics, sensitivity to cell loss constraints, and sensitivity to cell delay constraints. The loading region also facilitates the direct comparison of different admission control policies, both graphically and quantitatively, with respect to these criteria.

The remainder of this paper is organized as follows. Section 2 presents the definition of the loading region, as well as its construction. Section 3 describes the experimental methodology used to demonstrate the loading region concept, along with the three admission control policies evaluated in this paper[1]. Section 4 presents the results of the experiment, and Section 5 summarizes the conclusions from this work.

2 The Loading Region

2.1 The Loading Region Concept

The purpose of the loading region is to indicate the number of calls of heterogeneous traffic that can be admitted into the network by an admission control policy, without violating the QOS constraints of the admitted traffic. The loading region is based on the idea of a schedulable region [5, 6], but is generalized to consider both call-level and cell-level QOS constraints, and normalized traffic flows. If there are only two classes of traffic in the network (say, Class I and Class II), then the loading region has a simple two-dimensional graphical representation, as shown in Figure 1. A network can admit lots of calls of Class I (only), or lots of calls of Class II (only), or any intermediate combination of calls from Class I and Class II that remains within the (acceptable) loading region.

A loading region can be further classified into three sub-regions, as shown in Figure 1: the green, yellow, and red regions. A loading region is called a green region whenever for all calls of all types of traffic admitted to the network, the total peak bandwidth requirements for all calls does not exceed the output link capacity (i.e., deterministic multiplexing is used) and

[1] Additional policies and traffic loads are evaluated in [1].

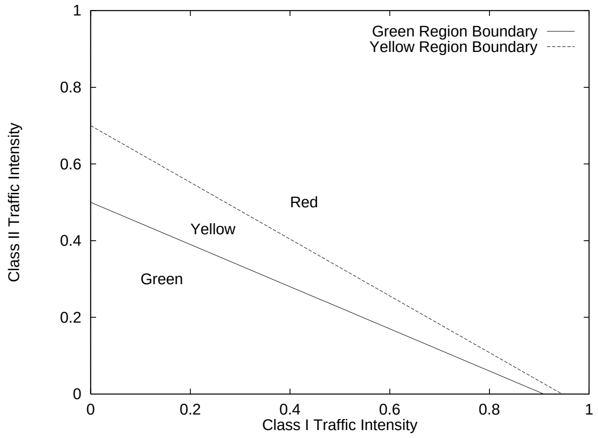

Figure 1: An Example of the Loading Region

the QOS constraints are guaranteed to be satisfied. A loading region is called a yellow region when the total peak bandwidth requirements for all calls exceed the output link capacity (i.e., statistical multiplexing is used) and the QOS constraints can be met for all calls of all types of traffic admitted in the network. When the number of calls admitted in the network falls in the red region, the QOS performance constraints cannot be guaranteed at all.

2.2 Construction of the Loading Region

The loading region is constructed by simulating selected subsets of traffic mix ratios. These traffic mix ratios, in the two-dimensional traffic case, can be thought of as linear trajectories starting from the origin, each with a different slope $\frac{(1-c)}{c}$ (see Figure 2). Successive points are simulated along each trajectory until specific traffic combinations on either side of the loading region boundary are identified (within the statistical limits of simulation, of course), for specific QOS constraints. The resulting boundary points are then connected to construct the (approximate) loading region boundary, for both the green and yellow regions.

More formally, the loading region is calculated as follows. The service rates of the traffic classes are fixed at 1.0 throughout the simulation. The arrival rates of each traffic class (normalized to the load sustainable by Deadline Scheduling [6]) are proportionately raised to increase the total traffic intensity ρ along the intended trajectory, to a point where a violation of the call admission criterion or the cell-level QOS constraints occurs. The resulting point, denoted by t, is called a *detecting point*. The procedure is repeated

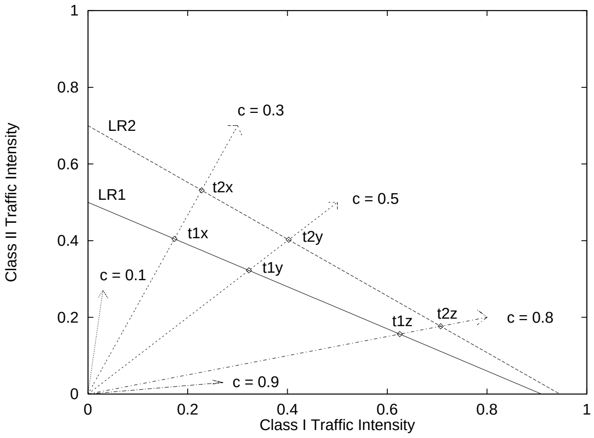

Figure 2: The Construction of the Loading Region

for 21 values of $c = [0.0,\ 0.05,\ 0.10,\ \ldots,\ 1.0]$, forming 21 different detecting points $t_j \in [t_0,\ t_1,\ \ldots,\ t_{20}]$. For example, the points t_x^1, t_y^1, and t_z^1 in Figure 2 are three of the detecting points for loading region LR1, while the points t_x^2, t_y^2, and t_z^2 are three of the detecting points for loading region LR2. The t_j's are then connected by a line to define the boundary of the loading region.

3 Experimental Design

A simulation experiment was designed to demonstrate the use of the loading region framework on three admission control policies from the literature. This section presents an overview of the experimental methodology used.

3.1 Network Model

The simulation model used in the experiments is shown in Figure 3. The ATM network is modeled as a single multiplexer with up to N heterogeneous input sources and one output port. The input and output link capacities are $L = 150$ Mbps. The multiplexer checks the input ports in a round robin fashion. The arriving cells are multiplexed onto the output port, which has three sets of buffers and one server. Each set of buffers is a completely shared buffer with a finite capacity of M cells assigned to a specific class of traffic (e.g., video, voice, and data in Figure 3). The value of M is large enough (say, $M \approx 1000$) for each class of traffic so that no buffer overflow occurs. In other words, cell loss for a class of traffic occurs only when the cell delay for the class exceeds the cell delay boundary, and never due to buffer overflow. The

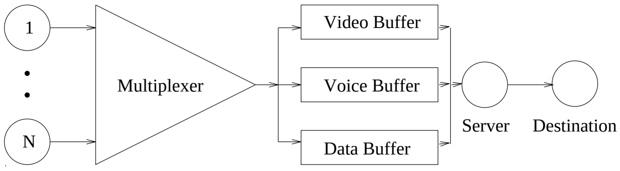

Figure 3: Network Model

processing time at the multiplexer is assumed to be zero. Therefore, cell delay is due only to queueing at the server.

Input cells are serviced by the server, and then forwarded to the destination. If the network server is idle at the time of a cell arrival, the newly arriving cell goes directly to the server. When the server is busy, arriving cells are put into the buffer according to their class of traffic. After a cell has been served, the server transmits the cell to the destination and selects the next cell to be served from the buffers based on the Deadline Scheduling algorithm [6].

3.2 Traffic Models

Two classes of traffic are used in the experiments: Class I for video traffic and Class II for voice traffic. Additional experiments with a third traffic class (data) are presented elsewhere [1].

The video traffic model used in the simulation is based on Hyman's two-state model [5]. The video traffic is represented as a periodic random process that has a fixed frame duration $F = 62.5$ ms, a peak bit rate $P = 10$ Mbps, and an average active period $C = 25$ ms. The active period for each frame, C_{active}, is a random variable uniformly distributed between $C_{min} = 10$ ms and $C_{max} = 40$ ms. During the active period, the source emits cells at rate P. At the end of the active period, the source stops emitting cells for the duration of the frame $(F - C_{active})$. The mean bit rate of the source is:

$$A = \frac{C}{F} P = 4 \text{ Mbps} \qquad (1)$$

The multiplexing of K video calls is done by uniformly distributing frame start times at intervals of F/K ms.

A single voice call (Class II traffic) is modeled as an ON-OFF source [5]. It is characterized by an exponentially distributed active period, in which cells are generated with constant rate $P = 64$ Kbps. C_{on} and C_{off} are two statistically independent random variables with a negative exponential distribution describing the ON-OFF behaviour of the source. $\overline{C_{on}} = 352$ ms and $\overline{C_{off}} = 650$ ms denote the average values of the ON period and the OFF period, respectively [5]. The average rate for one such source is:

$$A = \frac{C_{on}}{C_{on} + C_{off}} P = 22.5 \text{ Kbps} \qquad (2)$$

The average burst length is $B = 53$ cells.

3.3 QOS Constraints

The quality of service constraints for the two types of traffic (Class I and II) are defined by a four-tuple $(\varepsilon^I, \eta^I; \varepsilon^{II}, \eta^{II})$, where ε represents cell delay constraints, and η represents cell loss constraints.

Class I (video) traffic has QOS constraints specified by ε^I and η^I. The maximum tolerable cell delay (in milliseconds) at the multiplexer is ε^I. Class I requires a cell loss probability that is less than or equal to η^I. Typical values for these parameters are $\varepsilon^I = 1$ ms and $\eta^I = 10^{-6}$.

Class II (voice) traffic has QOS constraints specified by ε^{II} and η^{II}. The maximum tolerable cell delay at the multiplexer is ε^{II}, and the maximum tolerable cell loss is η^{II}. Typical values for these parameters are $\varepsilon^{II} = 2$ ms and $\eta^{II} = 10^{-3}$.

3.4 Admission Control Policies

Three admission control policies are evaluated in this paper:

1. **Deterministic Policy**
   The Deterministic control policy operates in deterministic multiplexing mode. A new call with peak bandwidth requirements P_n is admitted to the network if and only if the sum of its peak bandwidth P_n and the peak bandwidth requirements of all existing calls in the network is less than or equal to the link capacity of the network. Otherwise, the call is blocked.

2. **Joint Scheduling and Admission Control Policy**
   In the Joint Scheduling and Admission Control (JSAC) policy [6], a schedulable region based on the Deadline Scheduling algorithm [6] is predefined, measured, and stored with the admission control procedure. Incoming calls are admitted as long as the number of calls in the network does not exceed the boundary of the schedulable region. Otherwise, the call is blocked.

3. **Virtual Bandwidth Policy**
   The Virtual Bandwidth [7] admission control policy is based on virtual cell loss probability estimates. The virtual cell loss probability provides an upper bound estimate of the actual cell loss in the network, computed from the ratio of excess traffic T and average traffic load O, for a bufferless network model. This estimate is then used to calculate the virtual bandwidth of each call [7]. If the available bandwidth in the network is large enough to accommodate the virtual bandwidth for the new call, then the new call is accepted. Otherwise, the call is blocked.

4 Experimental Results

4.1 Loading Regions

The loading regions for the Deterministic Policy, the Joint Scheduling and Admission Control (JSAC) Policy, and the Virtual Bandwidth Policy are illustrated in Figures 4, 5, and 6, respectively. The loading regions are plotted using the method discussed in Section 2. The total number of cells sent by the two types of sources in each simulation run were 14,500,000 cells representing a simulated time of 60 seconds. In addition, five replications of the simulation were carried out for each loading region using different random number seeds. The final loading region plotted is the average from these five simulation runs. Confidence intervals for the region boundary were also determined [1], but are not shown in this paper.

In Figure 4, the green region boundary and the yellow region boundary for the Deterministic Policy are identical (as expected). In other words, the loading region for the Deterministic Policy has only two distinct sub-regions: the green region and the red region. This observation illustrates the (obvious) property that the Deterministic Policy does not allow the calls to operate in statistical multiplexing mode.

Figure 4 also shows that the maximum traffic intensity for video traffic is only 0.54 (when compared to the load sustainable by Deadline Scheduling [6]), and the maximum traffic intensity for voice traffic is only 0.51. These simulation results suggest that the Deterministic Policy is not efficient when the traffic in the network is bursty, as expected.

The loading region for the JSAC Policy, shown in Figure 5, is clearly divided into three sub-regions: the green region which is bounded by the dotted line, the yellow region which is bounded by the solid line, and the red region which is above the solid line. The maximum video and voice traffic intensities are 0.54 and 0.51, respectively, for the green region (the same as the Deterministic Policy). The maximum video and voice traffic intensities are 0.96 and 1.0, respectively,

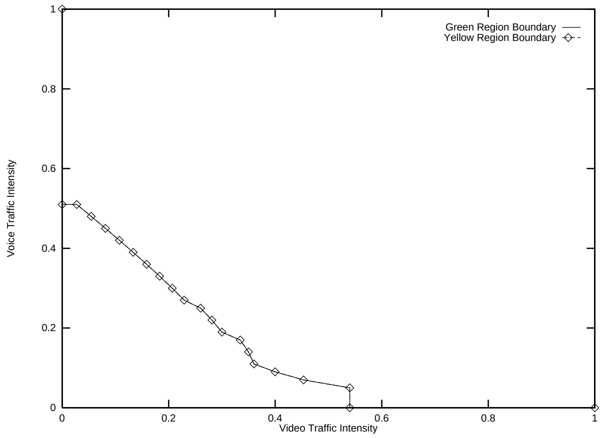

Figure 4: The Loading Region for the Deterministic Policy With $QOS\ SET = (1 \text{ ms}, 10^{-6}; 2 \text{ ms}, 10^{-3})$

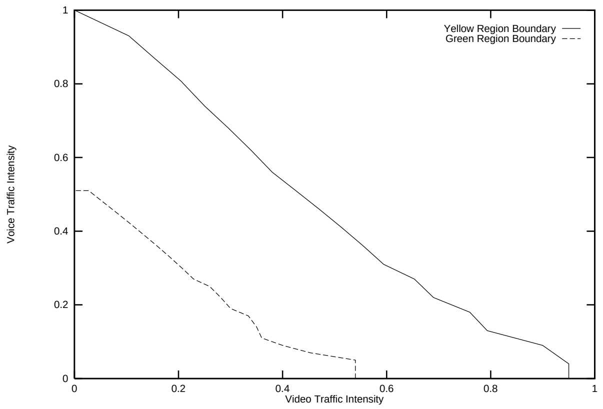

Figure 5: The Loading Region for the Joint Scheduling and Admission Control (JSAC) Policy With $QOS\ SET = (1 \text{ ms}, 10^{-6}; 2 \text{ ms}, 10^{-3})$

for the yellow region. The yellow region for the JSAC Policy represents the amount of statistical gain that can be achieved by statistical multiplexing, while still satisfying cell-level QOS constraints.

The loading region for the Virtual Bandwidth Policy, shown in Figure 6, is also divided into three subregions: the green, the yellow, and the red regions. The yellow region is fairly wide and high, but not as large as that of the JSAC Policy. The maximum video traffic intensity is 0.67, and the maximum voice traffic intensity is 0.95.

4.2 Sensitivity to Cell Loss Constraints

The loading region provides a means to evaluate the sensitivity of an admission control policy to cell-

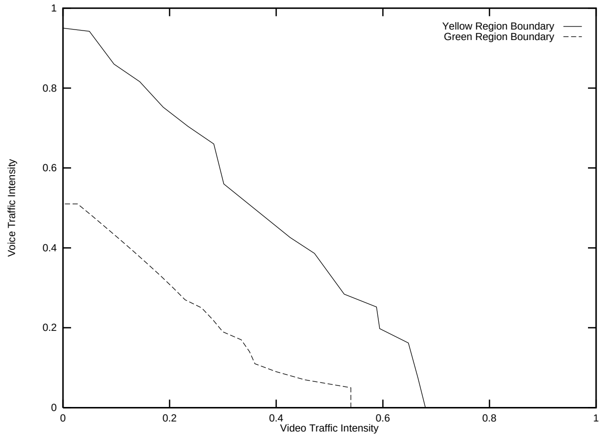

Figure 6: The Loading Region for the Virtual Bandwidth Policy With $QOS\ SET = (1 \text{ ms}, 10^{-6}; 2 \text{ ms}, 10^{-3})$

level QOS constraints, such as cell loss or cell delay. In other words, it can provide answers to questions such as, "If traffic sources had a less stringent cell loss constraint (say, 10^{-4} instead of 10^{-6}), then how much more traffic could be admitted into the network?".

The loading region provides a straightforward means to measure this "elasticity", which in turn provides an understanding of the performance properties of an admission control policy. That is, the elasticity behaviour gives insight into why the loading region boundary for an admission control policy falls where it does: is it because cell loss constraints would be violated if more traffic was admitted, or because cell delay constraints would be violated, or because of other restrictions in the admission control policy that are unrelated to QOS constraints?

Two policies, the JSAC Policy and the Virtual Bandwidth Policy, were found to be sensitive to the cell loss constraint, while the Deterministic Policy was not. Four different values of the cell loss constraint for video traffic ($\eta^I = 10^{-6}, 10^{-4}, 10^{-2}$, and 10^{-1}) were used in the experiments to assess the sensitivity of the policies[2].

Figures 7 and 8 illustrate the simulation results. In both figures, the yellow region increases in size when the cell loss constraint is relaxed. A clear stratification of loading regions (much like contour lines) appears as the cell loss constraint is relaxed.

This observation indicates that both the JSAC Policy and the Virtual Bandwidth Policy are sensitive to

[2] An estimate of the loading region for $\eta^I = 10^{-9}$ is also shown, but simulation run length considerations precluded precise determination of this loading region boundary.

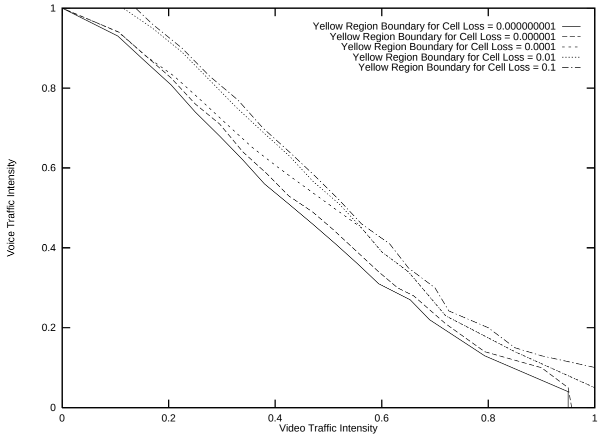

Figure 7: The Sensitivity of the JSAC Policy to Cell Loss Constraints

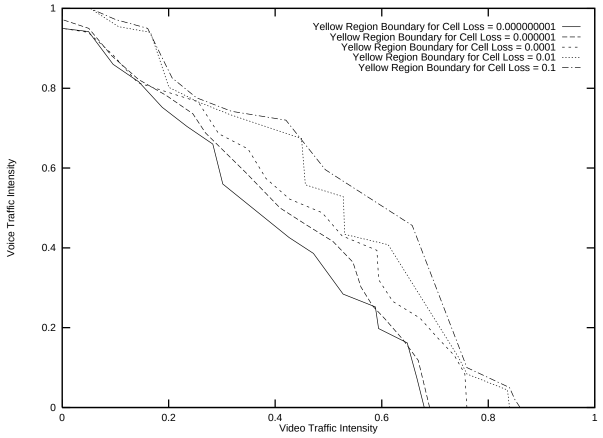

Figure 8: The Sensitivity of the Virtual Bandwidth Policy to Cell Loss Constraints

the cell loss constraint. The Virtual Bandwidth Policy is slightly more sensitive to the cell loss constraint than the JSAC policy, as can be seen by comparing Figures 7 and 8.

4.3 Sensitivity to Cell Delay Constraints

The same argument made previously for cell loss constraints can be applied to cell delay constraints. If traffic classes are delay-tolerant, then more traffic should be allowed into the network. If traffic classes are delay-sensitive, then the admission control policy should be more conservative.

Among the three policies studied in this paper, the JSAC Policy is the only policy that is sensitive to the cell delay constraint. The following sets of QOS constraints were used in the experiments to explore the

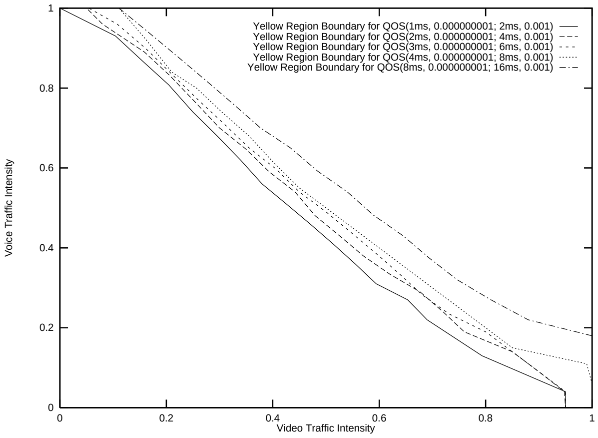

Figure 9: The Sensitivity of the JSAC Policy to Cell Delay Constraints

sensitivity of the JSAC Policy to the cell delay constraint:

$$QOS\ SET_1 = (1 \text{ ms}, 10^{-6}; 2 \text{ ms}, 10^{-3}) \quad (3)$$

$$QOS\ SET_2 = (2 \text{ ms}, 10^{-6}; 4 \text{ ms}, 10^{-3}) \quad (4)$$

$$QOS\ SET_3 = (4 \text{ ms}, 10^{-6}; 8 \text{ ms}, 10^{-3}) \quad (5)$$

$$QOS\ SET_4 = (8 \text{ ms}, 10^{-6}; 16 \text{ ms}, 10^{-3}) \quad (6)$$

Each successive set relaxes the cell delay constraint by a factor of two, while leaving the cell loss constraint unchanged.

Figure 9 shows the yellow regions of the JSAC Policy for the four cell delay QOS settings. When the cell delay constraint is relaxed, the yellow region becomes higher and wider, and the more the cell delay constraint is relaxed, the higher and wider is the yellow region. The changes in the size and shape of the yellow region clearly indicate that the JSAC Policy is sensitive to the cell delay constraint.

4.4 Summary

The results discussed in the previous sections show that the JSAC Policy is the most aggressive policy amongst the three admission control policies. This observation is clear from Figure 10, which shows the yellow regions for all three admission control policies. The Deterministic Policy is the most conservative policy among the three.

All policies are sensitive to the input traffic characteristics[3]. The Virtual Bandwidth Policy is the most sensitive policy to the cell loss constraints. Only the JSAC Policy is sensitive to the cell delay constraint.

[3]Not shown in this paper, but clearly shown in [1].

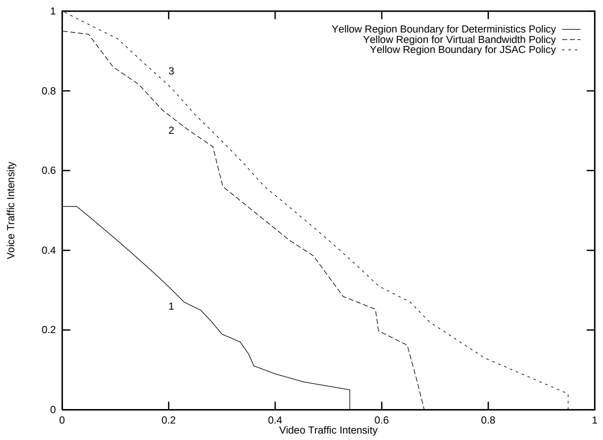

Figure 10: The Comparison of the Yellow Regions for Three Admission Control Policies with $QOSSET = (1\text{ ms}, 10^{-6}; 2\text{ ms}, 10^{-3})$

5 Conclusions

The loading region provides a useful framework for the performance evaluation of admission control policies. The loading region captures the aggressiveness of an admission control policy, by showing the size and shape of the yellow region. The loading region also captures the sensitivity of an admission control policy to QOS constraints and traffic characteristics, by showing the shifting of the region when parameters change.

The different admission control policies evaluated in this paper have very different loading regions. Among the three admission control policies evaluated in this paper, the Joint Scheduling and Admission Control policy is the most aggressive, and has the greatest sensitivity to QOS constraints.

Acknowledgements

Financial support for this research was provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada, through research grant OGP0121969. The authors are also grateful to the anonymous reviewers for their valuable comments on an earlier version of this paper.

References

[1] Y. Cao, "Performance Evaluation of Admission Control Policies for High Speed Networks", M.Sc. Thesis, Department of Computational Science, University of Saskatchewan, May 1994. (Available by anonymous ftp to cs.usask.ca as `pub/discus/thesis-cao.ps.Z`)

[2] A. Elwalid and D. Mitra, "Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High-Speed Networks", *IEEE/ACM Transactions on Networking*, Vol. 1, No. 3, pp. 329-343, June 1993.

[3] J. Ferrandiz and A. Lazar, "Admission Control for Real-Time Packet Sessions", *Proceedings of IEEE INFOCOMM'91*, Bal Harbour, Florida, pp. 553-559, April 1991.

[4] I. Habib and T. Saadawi, "Controlling Flow and Avoiding Congestion in Broadband Networks", *IEEE Communications Magazine*, Vol. 29, No. 10, pp. 46-53, October 1991.

[5] J. Hyman, A. Lazar, and G. Pacifici, "Real-Time Scheduling with Quality of Service Constraints", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, pp. 1052-1063, September 1991.

[6] J. Hyman, A. Lazar, and G. Pacifici, "Joint Scheduling and Admission Control for ATS-Based Switching Nodes", *Proceedings ACM SIGCOMM'92*, Baltimore, MD, pp. 223-234, August 1992.

[7] T. Murase, H. Suzuki, S. Sato, and T. Takeuchi, "A Call Admission Control Scheme for ATM Networks Using a Simple Quality Estimate", *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 9, pp. 1461-1470, December 1991.

[8] G. Woodruff and R. Kositpaiboon, "Multimedia Traffic Management Principles for Guaranteed ATM Network Performance", *IEEE Journal on Selected Areas in Communications*, Vol. 8, No. 3, pp. 437-446, April 1990.