The OSU Scheme for Congestion Avoidance in ATM Networks Using Explicit Rate Indication[1]

Raj Jain, Shiv Kalyanaraman and Ram Viswanathan[2]
Department of Computer and Information Science
The Ohio State University
Columbus, OH 43210-1277
Email: {*jain, shivkuma*}*@cis.ohio-state.edu, ramv@microsoft.com*

Abstract

We propose an end-to-end rate-based congestion avoidance scheme for ABR traffic on ATM networks using explicit rate indication to sources. The scheme uses a new congestion detection technique and an O(1) switch algorithm to provide high thoughput, low queues, fair operation, quick convergence and a small set of well understood parameters.

1 Introduction

Congestion in computer networks occurs whenever the total input traffic is greater than the total output traffic. When congestion occurs, sources of congestion must reduce their traffic. Sources learn about congestion either through feedback from the bottleneck or through source timeout. When a source receives congestion indication, it reduces its rate or window size. The traditional goals of congestion control schemes were to achieve high throughput and low delay. Some schemes like DECbit [5] achieve fairness between sources, while others like TCP slow start do not. A distinguishing feature in the former is that explicit feedback of one bit is used.

The problem of congestion is more important in high speed networks (HSNs) particularly ATM networks. This is because bandwidth has increased but the feedback delay has remained unchanged [3]. Hence, more data can be sent into the network before the sources learn about congestion. In particular, bit based schemes (like DECbit) or timeout schemes (slow start) may take several round trip times to converge to optimal operation.

One advantage of increased bandwidth is that more control information can be sent at the same percentage overhead. Hence, we can have a feedback longer than one bit. Further, HSNs are connection oriented and the network can maintain state about a connection. With accurate feedback, a scheme can achieve quick convergence and fairness without increasing the complexity of switch design. We present one such scheme in this paper, the OSU scheme.

In the OSU scheme, the network provides an explicit *rate indication* to the sources, rather than a single bit feedback. The OSU scheme is an example of a *rate-based* scheme. Rate-based schemes use end-to-end feedback to adjust source rates. Rate-based schemes are

[1] Proc. WATM'95 First Workshop on ATM Traffic Management, Paris, December 1995. Available as http://www.cis.ohio-state.edu/~jain/papers/OSUC_conf_version.ps

[2] Ram Viswanathan is currently with Microsoft Corp., WA

differentiated from *credit-based* schemes which use window (or credit) based flow control for every hop and require per-VC queues at every switch. After much debate, the ATM Forum has adopted the rate-based paradigm for ABR traffic management to allow flexibility to switch designers. A survey of ATM congestion control schemes and the rate vs credit debate may be found in [2].

The OSU scheme is similar to the MIT scheme [1, 7], the first explicit rate indication scheme proposed for ATM ABR service. The OSU scheme has a new congestion detection mechanism and an O(1) switch algorithm whereas the MIT switch algorithm is of O(N) complexity w.r.t. the number of VCs.

2 The OSU Scheme

The OSU scheme requires sources to monitor their load and *periodically* send control cells that contain the source rate information. The switches monitor their own load and use it with the information provided by the control cells to compute a factor by which the source should go up or down. The destination simply returns the control cells to the source, which then adjusts its rate as instructed by the network. The various components of the scheme are described next.

2.1 Control Cell Format

The control cell contains the following the fields relevant to our discussion:

1) Transmission Cell Rate (TCR = 1/inter-cell-transmission-time)

2) The average Offered Cell Rate (OCR, measured at the source)

3) Load Adjustment Factor (LAF, initially 0)

4) Averaging interval (AI, initially 0)

2.2 The Source Algorithm

2.2.1 Control Cell Sending Algorithm

The sources send a control cell into the network every T microseconds. The switches on the path have averaging intervals to measure their load levels (z). These averaging intervals are set locally by network managers. A single value of z is assumed to correspond to *one* OCR value of every source. If two control cells of a source with different OCRs are seen in a single interval (for one value of z), the above assumption is violated and conflicting feedbacks may be given to the source. Hence, the source interval T is set to the maximum of the switch averaging intervals in the path. This value is returned in the AI field of the control cell. The method ensures that a switch sees atmost one control cell from every source per switch interval.

2.2.2 Offered Cell Rate vs Transmission Cell Rate

Transmission Cell Rate is a variable controlling the minimum inter cell time of the source, whereas the average Offered Cell Rate (OCR) is the measured output rate of the source. This distinction between TCR and OCR is shown in Figure 1.

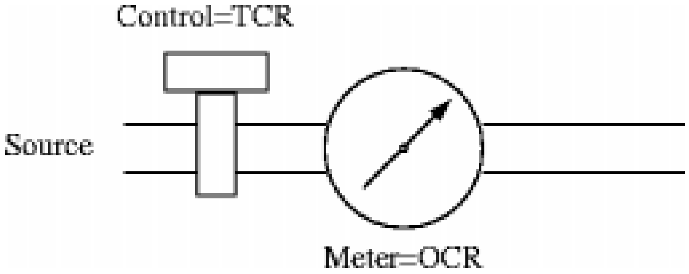


Figure 1: TCR (controlled) vs OCR (measured)

Consistent with the above definition, we require that OCR in cell $\leq$ TCR in cell. However, when TCR has just been reduced, the OCR may have a value between the old TCR and current TCR. Hence, we set

$$\text{TCR in Control Cell} \leftarrow \max\{\text{TCR, OCR}\}$$

2.2.3 Responding to Network Feedback

The network modifies only the LAF and the AI fields of the control cell. As we shall see later, the OCR field is used for switch computation. But, the source has to modify its TCR using the LAF and TCR in the control cell as follows:

$$\text{New TCR} \leftarrow \frac{\text{TCR in Cell}}{\text{LAF in Cell}}$$

$$\text{if (LAF} \geq 1 \textit{ and } \text{New TCR} < \text{TCR) TCR} = \text{New TCR}$$

$$\text{else if (LAF} < 1 \textit{ and } \text{New TCR} > \text{TCR) TCR} = \text{New TCR}$$

When LAF $\geq$ 1, the network is asking the source to decrease its TCR. If *New TCR* is less than the current TCR, the source reduces its TCR to *New TCR*. No adjustments are required otherwise. The other case (LAF $<$ 1) is similar.

2.3 The Switch Algorithm

2.3.1 Measuring The Current Load Level *z*

The switch measures its current load level, z , as the ratio of its “input rate” to its “target output rate”. The input rate is measured by counting the number of cells *received* by the switch during a fixed averaging interval. The target output rate is set to a fraction (close to 100 %) of the link rate. This fraction, called Target Utilization (*TU*), allows high utilization and low queues in steady state. The current load level z is used to detect congestion at the switch and determine an overload or underload condition.

$$\text{Target Output Cell Rate} = \frac{\text{Target Utilization (TU)} \times \text{Link bandwidth in Mbps}}{\text{Cell size in bits}}$$

$$z = \frac{\text{Number of cells received during the averaging interval}}{\text{Target Output Cell Rate} \times \text{Averaging Interval}}$$

2.3.2 Achieving Efficiency

Efficiency is achieved as follows:

$$\text{LAF in cell} \leftarrow \text{Max(LAF in cell, } z)$$

The idea is that if all sources divide their rates by LAF, the switch will have $z = \mathit{1}$ in the next cycle. In the presence of other bottlenecks, this algorithm converges to $z = \mathit{1}$ In fact it reaches a band $1 \pm \Delta$ quickly. This band is identified as an efficient operating region in the next subsection. However, it does not ensure fair allocation of available bandwidth among contending sources.

2.3.3 Achieving Fairness

Our first goal is to achieve efficient operation. Once the network is operating close to the target utilization, we take steps to achieve fairness. The network manager declares a target utilization band (*TUB*), say, 90±9% or 81% to 99%. When the link utilization is in the TUB, the link is said to be operating efficiently. The TUB is henceforth expressed in the U(1±Δ) format, where U is the target utilization (*TU*) and Δ is the half-width of the TUB. For example, 90±9% is expressed as $90(1 \pm 0.1)\%$.

Given the number of active sources, a fair share value is computed as follows:

$$\text{FairShare} = \frac{\text{Target Cell Rate}}{\text{Number of Active Sources}}$$

The number of active sources can be counted in the same averaging interval as that of load measurement. To achieve fairness, we treat the underloading and overloading sources differently. Underloading sources are sources that are using bandwidth less than the FairShare and overloading sources are those that are using more than the FairShare.

If the current load level is z, the underloading sources are treated as if the load level is $z/(1+\Delta)$ and the overloading sources are treated as if the load level is $z/(1-\Delta)$.

$$\text{If (OCR in cell} < \text{FairShare)} \;\; \text{LAF in cell} \leftarrow \text{Max(LAF in cell, } \frac{z}{(1+\Delta)})\}$$

$$\text{else LAF in cell} \leftarrow \text{Max(LAF in cell, } \frac{z}{(1-\Delta)})\}$$

We prove in [8] that this algorithm guarantees that the system consistently moves towards fair operation. We note that all the switch steps are O(1) w.r.t. the number of VCs.

The value of OCR in the cell is corelated to z when the control cell *enters* the switch queue. The value of z may change before the control cell leaves the switch queue. Hence, the OCR in the cell at the time of leaving the queue is not necessarily coorelated with z. Hence, the above computation is done when the control cell enters the queue.

2.4 The Destination Algorithm

The destination simply returns all control cells back to the source.

3 Unique Features of the OSU scheme

3.1 Congestion Avoidance

The OSU scheme is a congestion *avoidance* scheme. As defined in [6], a congestion avoidance scheme is one that keeps the network at high throughput and low delay in the steady state. The system operates at the *knee* of the throughput delay-curve as shown in Figure 2.

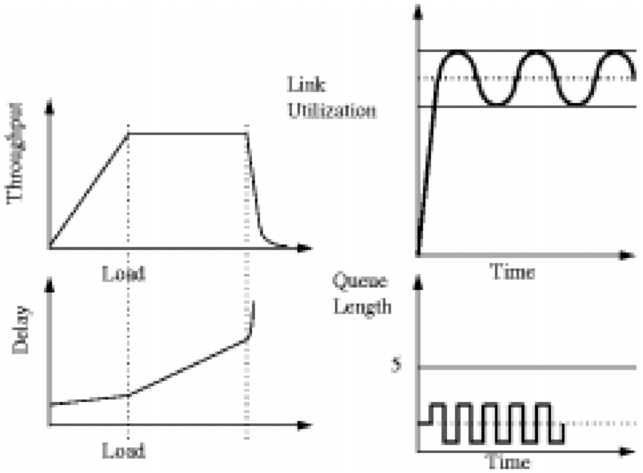


Figure 2: Throughput and delay vs Load

The OSU scheme keeps the steady state bottleneck link utilization in the target utilization band (TUB). The utilization is high and the oscillations are bounded by the TUB. Hence, in spite of oscillations in the TUB, the load on the switch is always less than one. So the switch queues are close to zero resulting in minimum delay to sources.

3.2 Parameters

The OSU scheme requires just three parameters: the switch averaging interval (AI) , the target link utilization (TU) , and the half-width of the target utilization band (Δ).

The target utilization (TU) and the TUB present a few tradeoffs. During overload (transients), TU affects queue drain rate. Lower TU increases drain rate during transients, but reduces utilization in steady state. Further, higher TU also constrains the size of the TUB. A narrow TUB slows down the convergence to fairness (since the formula depends on Δ) but has smaller oscillations in steady state. A wide TUB results in faster progress towards fairness, but has more oscillations in steady state. We find that a TUB of $90\%(1 \pm 0.1)$ used in our simulations is a good choice.

The switch averaging interval affects the stability of z. Shorter intervals cause more variation in the z and hence more oscillations. Larger intervals cause slow feedback and hence slow progress towards steady state.

3.3 Input Rate vs Queue length for Congestion Detection

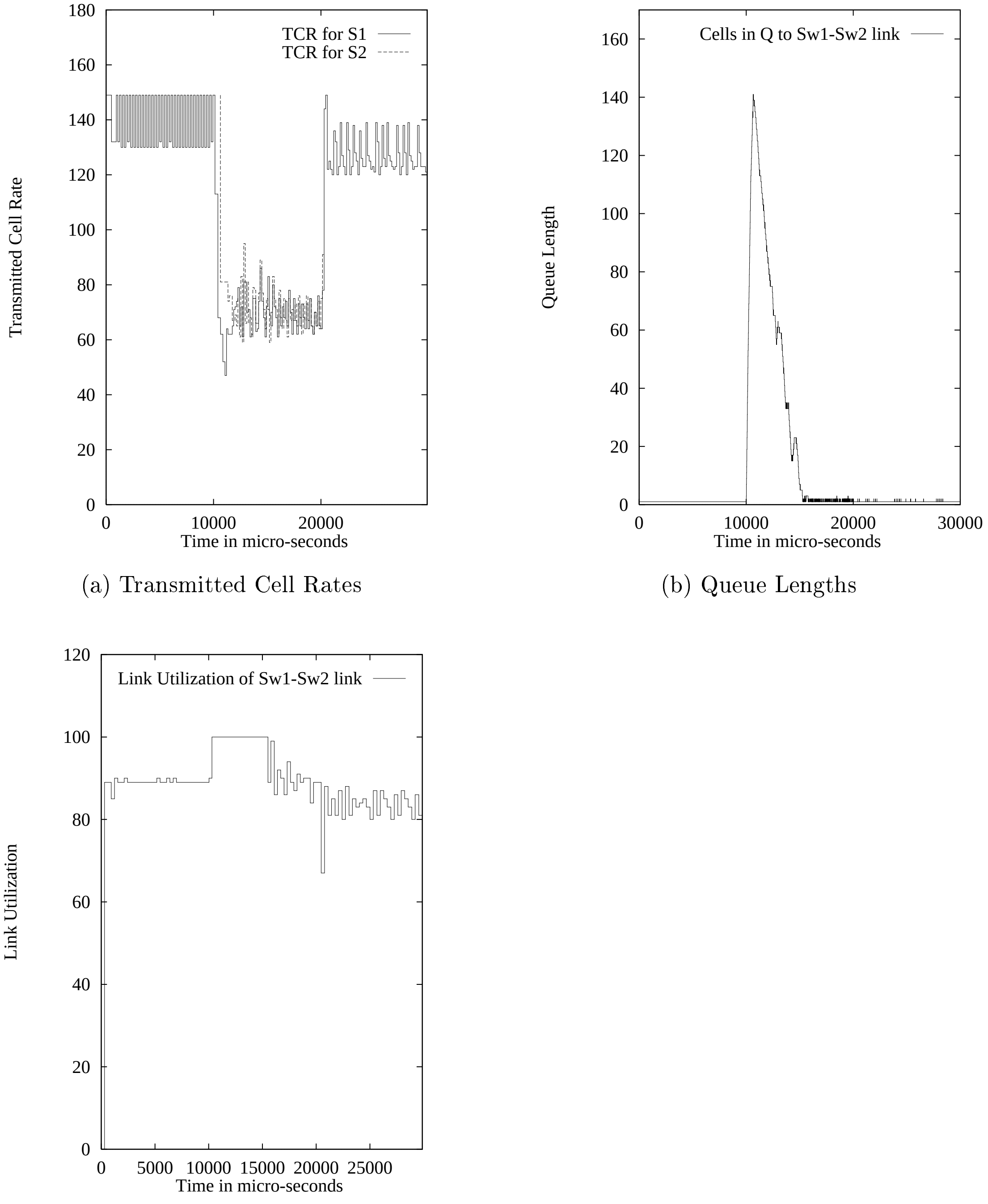
The OSU scheme detects congestion by measuring the current load level based on input rate at the switch queue. Many switch schemes use queue length as the congestion indicator. In window-based control, the sum of the source windows equals the maximum queue length. However, in rate-based control, the sum of the source rates (input rate) may be greater than, equal to, or less than the link output rate for *any* value of queue length. Hence, queue length gives no information about the relation between the current input rate and the ideal rate. Rate-based vs window-based control is further discussed in [3] and [4].

In rate-based control, the ratio of the input and output rates should be less than one for the switch queues to decrease. Our measure z uses this ratio. We aim for $z = 1$ which guarantees that in steady state, the input rate is smaller than the output rate.

4 Transient Source Simulation

The transient simulation consists of one persistant connection which is always active. A second persistant connection which shares one inter-switch link with the first, comes on after one third of the simulation run and goes off at two third of the total simulation time. All links are 1 km long running at 155 Mbps. The averaging interval of 300 μs and a target utilization band of $90(1\pm 0.1)\%$ are used.

This sample configuration tests the steady state as well as the transient response of the scheme. It also shows convergence to fairshares. The TCRs of sources, bottleneck queue length and link utilization are shown in Figure 3. A complete set of simulation configurations and results may be found in [8].



(a) Transmitted Cell Rates

(b) Queue Lengths

(c) Link Utilization

Figure 3: Simulation results for the Transient Source Simulation

5 Summary

We have developed an end-to-end rate-based congestion avoidance scheme for ABR traffic on ATM networks. The scheme uses a new congestion detection technique and an O(1) switch algorithm and achieves the goals of high thoughput, low queues and fair operation with a small set of parameters whose effects are well understood.

References

[1] A. Charny, D. D. Clark, R. Jain, "Congestion Control with Explici t Rate Indication," *Proc. ICC'95*, June 1995.

[2] R. Jain, "Congestion Control and Traffic Management in ATM Networks: Recent advances and a survey," invited submission to *Computer Networks and ISDN Systems*, 1995, also *AF-TM 95-0177*,[3] February 1995.

[3] R. Jain, "Myths about Congestion Management in High Speed Networks," *Internetworking: Research and Experience*, Vol 3, 1992, pp. 101-113.

[4] R. Jain, "Congestion Control in Computer Networks: Issues and Trends," *IEEE Network Magazine*, May 1990, pp. 24-30.

[5] R. Jain, K. K. Ramakrishnan, and D. M. Chiu, "Congestion Avoidance in Computer Networks with a Connectionless Network Layer," *Digital Equipment Corporation, Technical Report, DEC-TR-506*, August 1987, 17 pp. Also in C. Partridge, Ed., *Innovations in Internetworking*, Artech House, Norwood, MA, 1988, pp. 140-156.

[6] R. Jain, "A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-4, No. 7, October 1986, pp. 1162-1167.

[7] Anna Charny, "An Algorithm for Rate Allocation in a Cell-Switching Network with Feedback",MIT TR-601, May 1994.

[8] R. Jain, Shiv Kalyanaraman and Ram Viswanathan, "The OSU Scheme for Congestion Avoidance using Explicit Rate Indication", *OSU Technical Report OSU-CISRC-1/96-TR02*,[4] Dept. of CIS, Ohio State University, January 1996.

[3] AF-TM refers to ATM Forum Traffic Management sub-working group contributions.

[4] All our papers and ATM Forum contributions are available through http://www.cis.ohio-state.edu/~jain/