****

ATM Forum Document Number: ATM Forum/94-0883

****

Title: The OSU Scheme for Congestion Avoidance using Explicit Rate Indication

****

Abstract:

An explicit rate indication scheme for congestion avoidance in ATM networks is proposed. The sources monitor their load and provide the information periodically to the switches. The switches, in turn, compute the load level and ask the sources to adjust their rates up or down. The scheme achieves high link utilization, low delay, fair allocation of rates among contending sources, provides quick convergence and works for bursty traffic.

****

Source:

Raj Jain, Shiv Kalyanaraman and Ram Viswanathan
The Ohio State University
Department of CIS

Raj Jain is now at Washington University in Saint Louis, jain@cse.wustl.edu http://www.cse.wustl.edu/~jain/

****

Date: September 26-29, 1994

****

Distribution: ATM Forum Technical Working Group Members (Traffic Management)

****

****

INTRODUCTION

This scheme developed at The Ohio State University (OSU) is also an explicit rate indication scheme similar to the MIT scheme [1,2]. However, it does not necessarily require the switches to remember the rates of all VCs. Thus, the minimal storage requirements as well as the computational complexity becomes order one, O(1), that is, the computation or storage does not change as the number of VCs is changed. Also, it uses the exact overload as measured at the switch to determine the allowed rate.

FEATURES OF THE OSU SCHEME

The OSU scheme has the following desirable features:

1. It provides high throughput. The bottleneck utilization can be made close to 90-99%. The target utilization band (TUB) is actually a parameter set by the switch owner.

2. The oscillations are bounded. Once the system enters the TUB, it stays in the TUB trying to achieve fairness

3. The delays are minimum. The steady state queue lengths are close to 1 resulting in minimum possible delay.

4. The OSU scheme is a congestion avoidance scheme in the sense that the scheme provides high throughput and low delay. Also, the network operating point does not become suboptimal as the more memory is added to the switches.

5. The actual overload measured at the switch is used. Thus, any unused capacity, which is not used by sources to whom it has been allocated becomes available for other sources.

6. The scheme WORKS for bursty traffic.

7. The number of parameters has been kept small. The only parameters are the target utilization band and the load averaging interval.

8. The scheme is not very sensitive to parameter values. Slight mistuning of these parameters does not cause instability in the network.

9. The parameters are easy to set. Both target utilization band and averaging intervals have intutitive meaning and can be easily set by unskilled network operators.

10. The scheme requires only order one O(1) computation. In the most basic form of the scheme, only the rate of the current VCs is used and so the computation does not increase as the number of VCs is increased.

11. Bipolar feedback is used. The switches can increase or decrease the rate. Additional round-trip for increase is avoided.

12. Fairness is achieved without any per-VC scheduling, such as, round-robin or fair-queueing.

13. A backward congestion notification (BECN) option is provided. It is not required for proper operation but helps in some cases.

14. A precision fairness computation option, in which, the rates of all sources are used in computing the feedback, in a manner similar to that in the MIT scheme has also been designed. Again, this is not required for proper operation but minimized oscillations.

THE OSU SCHEME

In the OSU scheme, the sources monitor their average load and periodically send control cells that contain the load information. The switches monitor their own load and use it in combination with the information provided in the control cells, compute a factor by which the sources should go up or down. Like the MIT Scheme, the control cell is returned by the destination to the source, which then adjusts its rate as instructed by the network. The key difference betwen the OSU scheme and the other explicit rate schemes is the way the load is measured and rate

adjustment factor is computed.

The Control Cell (RM Cell) contains the following fields:

1. Transmitted cell rate (TCR). This is the inverse of the inter-cell transmission time.
2. The Offered Average Cell Rate (OCR) as measured at the source
3. Rate Adjustment Factor (initially 0)
4. Averaging interval (initially 0)
5. The direction of feedback (backward/forward)
6. Timestamp containing the time at which the control cell was generated at the source

The last two fields are used in the backward congestion notification option and need not be present if that option is not used. Other fields are explained later in this sections.

THE SOURCE ALGORITHM:

The source algorithm consists of three components:
1. How often to send control cells
2. How to measure the offered average cell rate
3. How to respond to the feedback received from the network

These three questions are answered in the next three subsections.

CONTROL CELL SENDING ALGORITHM:

The control cells are sent periodically every T interval. Although it could be done by the cell count, using interval allows the scheme to work on networks with widely varying link speeds. The network manager sets the averaging interval parameter for each switch. The maximum of the averaging interval along a path is returned in the control cell. This is the interval that the source uses to send the control cells.

During an idle interval, no control cells are sent. If the source measures the OCR to be zero, then one control cell is sent, subsequent control cells are sent only after the rate becomes non-zero.

MEASURING THE OFFERED AVERAGE LOAD

Unlike any other scheme proposed so far, each source also measures its own load. The measurment is done over the same averaging interval that is used for sending the control cells. Notice that there are two separate parameters: transmitted cell rate and offered average cell rate. The first is the instantaneous cell rate during burst transmissions. The cells are sent equally spaced in time. The inter-cell time is computed based on the transmitted cell rate. However, the source may be idle in between the bursts and so the average cell rate is different from the transmitted cell rate. This average is called the offered average cell rate and is also included in the cell. Notice that TCR is a control variable (like the knob on a faucet) while the OCR is a measured quantity (like a meter on a pipe).

Normally the OCR should be less than the TCR, except when the TCR has just been reduced. In such cases, the the maximum of current TCR and OCR is put in the TCR field.
In other words,

TCR in Cell <- max{TCR, OCR}

RESPONDING TO NETWORK FEEDBACK

The control cells returned from the network contain a "load adjustment factor" along with the TCR. The current TCR may be different from that in the cell. The source computes a new TCR by dividing the TCR in the cell by the load adjustment factor in the cell:

$$\text{New TCR} = \frac{\text{TCR in the Cell}}{\text{Load Adjustment Factor in the Cell}}$$

If the load adjustment factor is more than one, the network is asking the source to decrease. If the new TCR is less than the current TCR, the source sets its TCR to the new TCR value. However, if the new TCR is more than current TCR, the source is already operating below the network's requested rate and there is no need make any adjustments.

Similarly, if the load adjustment factor is less than one, the network is permitting the source to increase. If the current TCR is below the new TCR, the source increases its rate to the new value. However, if the current TCR is above the new TCR, the new value is ignored and no adjustment is done.

THE SWITCH ALGORITHM

The switch algorithm consists of the following components:

1. How to measure the available capacity
2. How to achieve efficiency
3. How to achieve fairness

These issues and others arising from these are discussed next.

MEASURING THE CURRENT LOAD:

This consists of simply counting the number of cells received during a fixed averaging interval. The interval is set by the network manager. Based on the known capacity of the link, the switch can compute the load level and determine whether it is overloaded or underloaded.

Since running a link at full load generally results in large queues, it is best to target the link utilization at close to but not quite at 100%. To achieve this the network manager selects a target utilization, say 90%. Whenever the input rate is more than 90% of the nominal capacity, the link is said to be overloaded and whenever the utilization is less than 90%, the link is said to be underloaded. The link cell rate when the network is operating at the target utilization is computed:

$$\text{Target Cell Rate} = \frac{\text{Target Utilization X Link bandwidth in Mbps}}{\text{Cell size in bits}}$$

The current load level is then given by:

$$\text{Current Load level} = \frac{\text{Number of cells received during the averaging interval}}{\text{Target Cell Rate X Averaging Interval}}$$

ACHIEVING EFFICIENCY

To achieve efficiency, all we need is to replace the load adjustment factor in each control cell by the maximum of the the

current load level and the load adjustment value already in the cell.

Load Adjustment Factor = max(Load Adjustment Factor in the cell,
Current Load Level in this Switch)

This simple algorithm is sufficient to bring the network to efficient operation within the next round trip. However, the allocation of the available bandwidth among contending VCs may not be fair. To achieve fairness we need to make use of the other information in the control cells as discussed later.

COUNTING THE NUMBER OF ACTIVE SOURCES:

Like the MIT scheme, the switches in our scheme may also remember the rates declared by various sources and use them in computing the fair share. However, there are two differences. First, the rates declared by the sources are "Offered Average Cell Rates (OCRs)" and not the desired cell rates, which may or may not be related to the actual rates. Secondly, in the simplest version of our scheme rates of all sources are not required. All we need is the number of active sources, which can be counted either by counting the number of sources with non-zero OCRs or by marking a bit in the VC table whenever a cell from a VC is seen. The bits are counted at the end of each averaging interval and are cleared at the beginning of each interval.

ACHIEVING FAIRNESS: The TUB Algorithm

In resource allocation, the top priority is to bring the network to efficient operation. Once the network is operating close to the target utilization, we need to take steps to achieve fairness. The network manager declares a target utilization band (TUB), say, 90+-9% or 81% to 99%. Whenever the link utilization is in TUB, the link is said to be operating efficiently. As will be seen later, it is better to express TUB in the U(1+- Delta) format, where U is the target utilization level. For example, 90+-9% is expressed as 90(1+- 0.1)%.

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

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

To achieve fairness, we treat the underloading and overloading sources differently. Underloading sources for our scheme are those sources that are using less than the fair share. While overloading sources are those that are using more than the fair share.

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). Here Delta is the half-width of the TUB. We call this "the TUB algorithm."

If the OCR in the control cell is less than the fair share, the load adjustment factor in the cell is changed as follows:

Load Adjustment Factor = max(Load Adjustment Factor in the cell, z/(1+Delta))

On the other hand, if the OCR in the control cell is more than the fair share, the load adjustment factor in the cell is

adjusted as follows:

Load Adjustment Factor = max(Load Adjustment Factor in the cell, z/(1-Delta))

We have proven that this algorithm guarantees that the system consistently moves towards more fair operation. Also, once inside the TUB, the network remains in the TUB unless the number of sources or their load pattern changes. In other words, TUB is a ``closed'' operating region. These statements are true for any value of Delta less than 0.5.

If Delta is small, as is usually the case, division by 1+Delta is approximately equivalent to a multiplication by 1-Delta and vice versa.

THE DESTINATION ALGORITHM:

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

SIMULATION RESULTS:

We have done extensive simulation testing of the scheme [3]. The results will be presented partly in this forum meeting and then in the November meeting.

OTHER OPTIONS:

The basic scheme as described above is sufficient to bring the network to optimal and fair operation under all circumstances. However, the performance can be improved by a number of extensions. These extensions will be the subject of a future ATM Forum contribution.

REFERENCES

[1] A. Charny, D. Clark, and R. Jain, "Congestion Control with Explicit Rate Indication," ATM Forum Contribution 94-0692, July 1994.

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

[3] R. Jain, S. Kalyanraman, and R. Viswanathan, "The OSU Scheme for Congestion Avoidance in ATM Networks Using Explicit Rate Indication," The Ohio State University, Department of CIS, Technical report, under preparation.