Chapter 3

Rate-Based Congestion Control Schemes

3.1 Introduction

We are considering best-effort connectionless packet-switched networks where link capacity is typically fixed. Given that link bandwidth (and hence overall bit rate) is fixed, network operations which deal with bit rates may be useful. For example, if routers can feed back rate information to sources of traffic flows, then the routers can participate in the fair allocation of link capacity. Transport protocols which are rate-based can admit packets into the network uniformly spaced: this helps to prevent short-term congestion. Combined, these techniques can be used as a form of congestion control, by allocating rates to traffic flows which keep network operation at the knee-point of peak power (see Figure 2, pg. 4).

There is less research into the use of rate-based techniques than for non-rated-based techniques in connectionless packet-switched networks. This chapter examines some of this research, and also reviews the use of rate-based techniques in other network architectures such as Asynchronous Transfer Mode (ATM).

3.2 Rate-Based Flow Control

In the late 1980s, several new transport protocols for connectionless networks were proposed: VMTP, NETBLT and XTP. These protocols featured end-to-end rate-based flow control instead of the more traditional sliding window flow control.

3.2.1 NETBLT

In [Clark *et al* 87], Clark et al describe the NETBLT transport protocol. Their previous experience with transport protocol performance indicated two major problems: "a restriction in throughput which arises from the use of windows as a flow control mechanism, and the difficulty in handling timers". They note that "[data] transmission ought to be evenly distributed over a [round-trip time] period to match the receiver's consumption speed. Unfortunately, windows only conveys the latter – *how much* data can be buffered, rather than the latter – *how fast* the transmission should go." The difficulty with timers is, as was noted in Section 2.3.3, due to the high variance in round-trip time. They concluded that:

- Windows and timers perform poorly in synchronizing the end state. To achieve fast resynchronization of the end state, we need better mechanisms than the cumulative acknowledgment, and we must reduce reliance on timers.
- Flow control must be independent from error control. Mixing the two in one mechanism can only make the flow control vulnerable to transmission errors and delays.

The NETBLT transport protocol replaces the traditional window flow control with a rate-based flow control mechanism or, more specifically, a coarse-grained rate-based mechanism: packets are transmitted in *groups*. NETBLT's rate control has two components, a *burst size* in packets which sets the group size, and a *burst interval*. This interval is the period between the transmission start of successive bursts. Selective acknowledgments are used for error control.

At connection setup, an initial burst size and interval is negotiated between source and destination. These are renegotiated by the destination in the selective acknowledgment of the received group[1]. Experimental results over a low round-trip 10Mbps LAN and a high round-trip (1.8 seconds) 1Mbps WAN showed that such a coarse-grained rate-based flow control mechanism showed promise. Clark et al noted, however, that a protocol using rated-based flow control "must be able to set an initial transmission rate based on the available network bandwidth and the speed at which the receiver can process data. ... It is therefore important that support for rate selection exist at the network level."

3.2.2 XTP

The Xpress Transfer Protocol [Sanders & Weaver 90] provides selective acknowledgment error control and group-oriented rate-based flow control in a fashion similar to NETBLT. The aim of XTP's development was to streamline the processing requirements of a transport protocol to the point where it could be "implemented in hardware as a VLSI chip set".

The protocol's headers set a *burst rate*, the maximum number of bytes the receiver will accept per second, and a *burst size*, the maximum number of bytes the receiver will accept per burst. As the aim of XTP is to reduce processing requirements, [Sanders & Weaver 90] gives no further details on its rate-based flow control.

3.2.3 VMTP

The Versatile Message Transport Protocol [Cheriton 86] [Cheriton & Williamson 89] was designed to overcome perceived deficiencies in the performance and functionality of traditional network protocols. Like NETBLT and XTP, VMTP provides selective acknowledgment error control and group-oriented rate-based flow control. VMTP's flow control is more fine-grained than NETBLT or XTP. While packets are still transmitted in bursts, the overall burst rate is controlled by the burst size and the *inter-packet delay*. Without the inter-packet delay, "a client must be prepared to accept an entire packet group at once". Cheriton et al suggest that "if the receiver requests retransmission of every fourth packet, the sender can reasonably increase the inter-packet gap. Moreover, the sender can periodically attempt to reduce the inter-packet gap when no packet loss occurs, to ensure

[1]The mechanism describing this negotiation is not given.

that it is transmitting at the maximum rate the receiver can handle." As with XTP, the immense functionality of VMTP prevents any further description of the rate-based flow mechanism.

3.2.4 Summary

These three studies demonstrate that protocols can be built for connectionless networks where error and flow controls are separated, and that rate-based flow control is feasible. As with all end-to-end protocols, these protocols do not attempt to measure or control congestion *within* the network, but only limit the source's transmission characteristics to that which can be sustained by the destination.

3.3 Available Bit Rate Congestion Control in ATM Networks

Asynchronous Transfer Mode (ATM) networks are characterised as connection-oriented cell-switched networks [Partridge 94]. Cells are fixed in size (53 octets), and virtual circuits are established across the network to propagate the traffic for each connection. ATM provides several service classes to the connections established between sources and destinations: Constant Bit Rate (CBR), Variable Bit Rate (VBR), Available Bit Rate (ABR) and Unspecified Bit Rate (UBR). Of the four service classes, only ABR provides service degradation congestion control [Jain 96], which is the type of congestion control under consideration in this thesis.

With ABR, a source of cell traffic can specify minimum and maximum cell rates at connection establishment. If the connection is granted, then the source can transmit at a rate between the specified minimum and maximum cell rates. However, if the network becomes congested, the source may need to reduce its current cell rate to minimise congestion.

Several rate-based congestion control mechanisms were proposed to control source cell rates for the ABR service. The main proposals are summarised below.

3.3.1 BECN

Early congestion proposals were Forward Explicit Congestion Notification (FECN) [Newman 94] and Backward Explicit Congestion Notification (BECN) [Newman 94]. In BECN, the intermediate switches[2] return special *resource management* (RM) cells to sources of cells if they believe the source is causing congestion. Congestion is identified by monitoring cell queue lengths within the switches. On receipt of an RM cell, a source is obliged to halve its transmission rate.

If no RM cells arrive after a 'recovery period', the source may double its cell rate (without exceeding the maximum cell rate specified at connect time). The recovery period is proportional to the current cell rate: as the cell rate drops, the recovery period shortens.

BECN can be seen to be analogous to the Source Quench scheme, controlling the flow of a specific ATM connection.

[2] ATM switches perform cell routing, and so are analogous to routers in packet-switched networks.

3.3.2 PRCA

Another early proposal for ABR congestion control was a modification of DecBit. The destination returns RM cells to the source if, over a certain period, any Congestion Experienced Bits are set. The source uses Multiplicative Decrease and Additive Increase to control its cell rate.

One problem noted with this approach [Jain 96] is that, under heavy congestion conditions, the RM cells which cause rate decrease may not reach the source, and so the congestion may not be reduced. A modification to the proposal, known as the Proportional Rate Control Algorithm (PRCA) [Hluchyj 94], has the source set every Congestion Experienced Bit on, but leave one in N off. If any RM cells arrive with the Congestion Experienced Bit off, then the destination sends RM cells back to the source, which prompt a source rate *increase*. Again, Multiplicative Decrease and Additive Increase is used to control the source's cell rate. With this approach, the source cannot increase its rate unless RM cells are received.

PRCA was found to have a fairness problem for long-path connections. If the probability of a cell having its Congestion Experienced Bit set by one ATM switch is p, then the probability of it being set over a p-switch path is p^n. Thus, long-path connections have a lower chance of receiving a rate increase RM cell than short-path connections. This is known as the 'beat-down problem' [Bennett & Des Jardins 94].

3.3.3 EPRCA

One problem with binary schemes like BECN and PRCA is that the time to reach an optimum cell rate is long: several iterations of Multiplicative Decrease and/or Additive Increase may be required to bring the source to the optimum cell rate. [Charney *et al* 94] suggested returning an explicit cell rate value to the sources of cells. This rate is set using the 'Fair Share' algorithm described in [Charney 94], and is covered in detail in Section 3.4.

The use of explicit cell rate information has several advantages: switches can monitor for violations of the cell rates, and convergence to the optimum cell rate is shorter than for the binary schemes. The loss of cells carrying explicit cell rate information has less impact than the loss of BECN or PRCA RM cells: the source should receive other cells carrying explicit cell rate information to set its optimum cell rate.

Enhanced PRCA (EPRCA) [Roberts 94] resulted from a combination of PRCA and the explicit cell rate scheme described above. Cell sources send a combination of RM cells and data cells with Congestion Experienced Bits. Switches along the path of a connection may calculate the Fair Share rate for the connection (and place the value in the RM cells), set the Congestion Experienced Bits if the connection is exceeding its fair share (or is causing congestion), or both.

EPRCA allows switches to perform binary-feedback congestion control, explicit rate congestion control, or both.

3.3.4 Summary

Although ATM is a very different network architecture to the connectionless packet-switched networks under consideration, many congestion control techniques are applicable to both. Analogues of both Source Quench and DecBit have been proposed for the ATM ABR service: as well, schemes which provide explicit rate information to sources have been proposed. It may be possible to apply the latter techniques to packet-switched networks. However, this would require such modifications as rate-based Transport protocols, operation with variable-sized packets and the identification of source/destination flows by the Network Layer.

3.4 Charney's Thesis

The EPRCA congestion control scheme depends upon the 'Fair Share' rate allocation algorithm. [Charney 94] describes this algorithm, which is very similar to the RBCC algorithm described in Chapter 6.

The 'Fair Share' algorithm is performed by the routers in a rate-based framework where some packets in a traffic flow contain rate fields, which are adjusted by the algorithm, and then fed back to the traffic source to control its rate. The algorithm effects the requirements of the *Maximin* optimality criterion, described informally in [Charney 94]:

> Consider a network with given link capacities, the set of sessions and fixed session routes. We are interested in such rate allocations that are feasible in the sense that the total throughput of all sessions crossing any link does not exceed the link's capacity. We would like the feasible rate allocation to be fair to all sessions. On the other hand we want the network to be utilized as much as possible.
>
> We now define a fair allocation in the following way. We consider all "bottleneck" links, i.e the link with the smallest capacity available per session. [...] We share the capacity of these links equally between all sessions crossing them. Then we remove these sessions from the network and reduce all link capacities by the bandwidth consumed by the removed sessions. We now identify the "next level" bottleneck links of the reduced network and repeat the procedure. We thus continue until all sessions are assigned their rates.
>
> Such a rate vector is known as *maxmin fair* allocation.

The Fair Share algorithm is described in some detail in [Charney 94] (pages 24-32). The following description is taken from [Jain 96]:

> The fair share is computed using an iterative procedure as follows. Initially, the fair share is set at the link bandwidth divided by the number of active VC's [virtual circuits]. All VCs whose rates are less than the fair share are called "underloading VCs". If the number of underloading VCs increases at any iteration, the fair share is recomputed as follows:
>
> $$Fair\ Share = \frac{Link\ Bandwidth - \Sigma\ Bandwidth\ of\ Underloading\ VCs}{Number\ of\ VCs - Number\ of\ Underloading\ VCs}$$
>
> The iteration is then repeated until the number of underloading VCs and the fair share does not change.

Charney proves that, given an upper bound D on round-trip delay, and the number N of active virtual circuits, the upper bound on convergence time from any initial conditions is no more than $4ND$. Simulations in [Charney 94] show that the bound is more realistically $2ND$.

A comparative discussion on Fair Share and RBCC is delayed until Section 7.6.

3.5 Theoretical Work

One difficulty with congestion control schemes is to verify that they will control congestion over the set of all possible network designs, traffic sources and usage conditions. Some research has been undertaken to prove rate-based congestion control schemes within a limited set of conditions. This research shows that rate-based schemes can control congestion well, and within a short space of time.

[Benmohamed & Meerkov 93] analyse a rate-based control scheme which monitors the occupancy x of router packet buffers against a desired occupancy x^0. Routers periodically calculate an admission rate for sources based on the measured difference between x and x^0, and return the rate to traffic sources, who in turn are obliged to lower their rate to the lowest value obtained.

Analysis is given for a backbone router with one congested outgoing link. They derive formulae to determine a set of fair admission rates for all traffic sources through the congested node:

> If M connections share a transmission link, then each gets $1/M$ of the bandwidth, and if $(M - l)$ connections use less than their share, then the unused portion is equally distributed among the rest.

[Benmohamed & Meerkov 93] note that the interval T between measurements of x affects the performance of the scheme. If T is small, x converges more quickly to x^0 with less buffer overshoot, but also causes higher communication overhead within the network, and more work for the router. As well, the communication delays to the source of each traffic flow, and the round-trip times for each traffic flow, affect the choice of T. Experimental work shows that the congestion scheme does cause x to converge rapidly to x^0 with some oscillation, which may be further reduced by using adaptive techniques.

A similar scheme is proposed by [Fendick *et al* 92], where both the level of occupancy and the arrival rate of packets is measured, and with a desired operating point of $x^0 = 1$. They show that their scheme rapidly converges to the desired operating point, although they noted that the scheme is locally unstable in the region of this operating point. The scheme deals with the effect of both delay due to packet propagation and randomness in the feedback delay.

An alternative rate-based control framework which is end-to-end based is described in [Haas 91]. Here, the source measures the interval between pairs of transmitted packets. The destination also measures this interval as packets arrive. This interval, termed the "virtual delay", is the difference between the interval as measured by the source and destination. The "virtual delay" is averaged over several packets and is returned, along with measured packet arrival rates, to the source. The source uses this data to estimate a level of congestion, β, such that $\beta = 1$ indicates low congestion and $\beta = 0$ indicates high congestion. A step function is used to convert changes in "virtual delay" into values of β.

Sources admit traffic using a scheme like Leaky Bucket [Turner 86] which is controlled by the estimated congestion level. The inter-packet gap is calculated using

$$interpacket\ gap = min(\frac{1}{\beta}, quota) - 1$$

where $quota$ is the interval, in packet transmission time units, between updates of β. When $\beta = 1$, the inter-packet gap is zero, and when $\beta = 0$, the inter-packet gap is $quota$.

The framework allows for many different methods to determine β and the update interval $quota$. Haas notes that the value of $quota$ depends on the network topology and the network traffic, but that the number of "virtual delay" samples per $quota$ should be large enough to minimise statistical noise. A future approach would be to use adaptive techniques to match the values of all the parameters to the network topology and traffic.

Yet another rate-based congestion control alternative, Hop-by-Hop, is described in [Mishra *et al* 96], where service rates of traffic flows are adjusted by routers based on information provided by neighboring routers. Each router monitors its buffer occupancy for all active flows, and periodically informs upstream routers of this information. Upstream routers use this information to compute a *target* sending rate for each flow. Rather than aiming for the ideal of one packet queued per router, Hop-by-Hop keeps a small number of packets queued at each router, which can be transmitted if additional bandwidth becomes available downstream.

Hop-by-hop is analysed over a set of N routers which form a single route: all flows take this single route. Cross-traffic to the flows along the route is also modeled. Analytical results show that the scheme brings flow rates and desired queue occupancies to the desired operating point with no instabilities.

3.5.1 Summary

This theoretical research into rate-based congestion control schemes shows that congestion control can be successfully implemented using rate-based techniques, with fast convergence of network conditions to near the desired operating point. The schemes presented, however, describe simple models or experiments where there is only one congested node, and/or the paths for all traffic sources take the same route. The routers must periodically sample the traffic characteristics of the traffic flows, and the interval between periods affects the performance of the congestion control.

3.6 Non-Service Degradation Control Schemes

Most of the congestion control schemes reviewed so far are service reduction schemes: sources must reduce their transmission rates to reduce congestion. An alternative to this is to identify the existence and extent of network congestion; it is then up to each source to decide what the most suitable response is to the identified congestion conditions.

Williamson & Cheriton [Williamson & Cheriton 91] propose a congestion control schemes which returns a *Loss/Load* curve to the source. This curve gives the probability of packet loss as a function of the source's offered load (as a rate) on the network. This has several advantages over service degradation control schemes:

- The network can provide a predictable level of service without requiring resource reservations, allowing data to be transmitted without connection setup delays.
- The source can use the Loss/Load curve to choose an operating point which is acceptable in terms of throughput and packet loss.

The Loss/Load curve is monotonically increasing: as offered load increases, so will the probability of packet loss. The network is free to provide different Loss/Load curves to different sources, based on an operating policy. The curve can be treated as a 'contract' between the network and the source for a specific time interval.

When the load offered to a router is greater than its available output bandwidth, the router must discard some of the load. There are many different ways to do this; the choice of method is part of the operating policy and affects the Loss/Load curves returned to the sources. The queueing scheme is arbitrary, and the router may use hints on traffic type to bias packet loss (e.g to favour loss-insensitive traffic). The router monitors the rates of traffic flows over an interval using an exponentially weighted average. The Loss/Load curves are calculated and transmitted periodically, or when the load on the router changes significantly. They are not transmitted when packets are lost.

The host receives a number of curves from the routers along the traffic path. It uses the Loss/Load curve for the most congested router, and uses this curve to set its transmission rate. Leaky Bucket [Turner 86] or Virtual Clock [Zhang 91] can then be used for packet admission.

The aim of the Loss/Load congestion scheme is not to eliminate packet loss, but to allow each source to control the amount of loss during transmission. One consequence of the scheme is that, as loss increases as offered load increases, a source's throughput reaches a maximum at a load value before packet loss reaches a probability of 1. Therefore sources which are greedy and transmit above this load value will decrease their overall throughput. Simulation results in [Williamson & Cheriton 91] show that the scheme does give accurate packet loss probabilities, and the Loss/Load curves converge to equilibrium after two to four iterations.

3.7 Summary

Rate-based congestion control schemes form an alternative option to the traditional end-to-end and window-based schemes reviewed in the previous chapter. They are particularly useful in such networks as ATM where connection establishment and admission of data into the network are rate-based.

Rate-based control schemes allow intermediate network nodes to participate in the control of congestion, by monitoring congestion and returning information on congestion to traffic sources. This feedback may be a single bit (BECN), an explicit rate value (PRCA and EPRCA), or even a probability of packet loss (Loss/Load).

Theoretical studies of rate-based congestion control schemes show that they can bring source transmissions to the optimal point of operation within a small number of iterations. In general, these studies have been limited to simplistic network architectures, and may not model global network behaviour.

The next chapter will examine a rate-based framework for congestion control in connectionless packet-switched networks. The framework attempts to address the deficiencies in traditional control schemes via the use of explicit rate feedback to sources of network traffic.