Introduction to Flow and Congestion Control

Cambyse Guy Omidyar *Guy Pujolle*

Our readership covers all aspects of telecommunications, but no one is expert in everything. We have prepared this introduction to flow and congestion control to give those readers who are not well-versed in this field a bit of background. In the mid-1970s, computer communications experts were working on how to control the flow of packets in the existing networks. Two of the main contributors at that time were M. Schwartz and L. Kleinrock, and shortly thereafter many others followed. We also recall that for the first time a student named Marek Irland tried to flow control computers at Waterloo University; he later died of cancer in the late 1970s. Today, researchers' headaches are due to the speed of the transport media. Existing networks are going to transport variety of traffic types, and maintain appropriate quality of service for each fraction of the reserved bandwidth on demand. The ATM Forum established standards for Traffic Management, Traffic Management Specification Version 4.0, and the International Telecommunications Union — Telecommunications Standards Sector (ITU-T), formerly the CCITT, adapted I.371 traffic control and congestion control in broadband integrated services digital network (B-ISDN).

What Is Flow Control?

The primary reason for flow control is to reduce the flow of excess traffic. To reduce the flow one can use feedback mechanisms. To redirect the flow one can use routing techniques, and to clean the buffers from excess traffic one can drop the entire bits of a packet on the floor (i.e., early packet discards).

In any case, the principal aspects of congestion control are to reduce the flow of traffic entering the network in order to relieve congestion at a point within the network and to redirect the flow away from those congested nodes that are experiencing unwanted traffic.

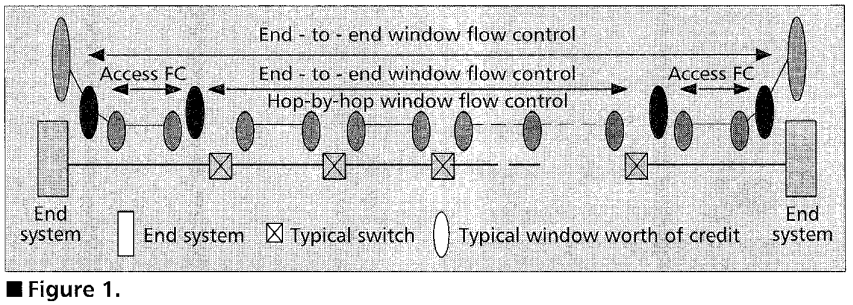
The concept of flow control has been known to many of us for some time. We already experience this control in our lives, from our information superhighway to planetary movements, from macro- to microcells in our body, and from our boss who wants to know and control everything.

What really governs control is its basic rule, which is often formulated in our design algorithms and dictates how to control the flow. Control mechanisms were designed by the researchers at several points and layers in the network, as shown in Fig. 1. For example, source-to-destination flow control is known as end-to-end flow control and is often governed by Transport Protocol window flow control. Node-to-node and/or switch-to-switch flow control is known as hop-by-hop flow control and is a layer two responsibility. Entry-to-exit flow control is layer three flow control; it polices certain edges of the network from incoming traffic. The layers of control are interchangeable: the end-to-end flow control at layer two or below of asynchronous transfer mode (ATM) can assume the end-to-end flow control responsibilities of layer four. The implementation could take place in hardware rather than software so that the switch can catch up with the speed of the incoming and outgoing cells.

Each layer controls different fragments in size. There is a difference between controlling a few large segments and many small fragments. It is like comparing a big truck on a fast lane with a small and efficient car on a highway. Sequencing and keeping track of all the small fragments are often more difficult. Small fragments, like many cells, tend to be switched faster than large packets; thus, their interarrival time at the destination, the time between arrivals, is shorter. One key factor is that ATM can transport voice, video, image, and data simultaneously; in other words, it is a duck who can swim, walk, and fly, but not everything in an efficient manner. Sometimes you wonder how so many control parameters contained in a single algorithm could work together and converge when it is necessary to act.

What Is Congestion Control?

How are congestion phenomena created? A simple answer is that they occur when resources are scarce and highly in demand, and processing and transmission speeds lag behind the speed of incoming traffic. Traditionally, congestion problems in low-speed networks are addressed with techniques that depend on the design of the network layer. For example, congestion is usually controlled in datagram networks by use of "choke packets (CPs)", while in virtual circuit (VC) networks it is frequently solved by use of back pressure (BP). The BP method



■ **Figure 1.**

concentrates on distributing the storage of excess traffic. While the CP method concentrates on reducing the rate of input to the network.

It has been recognized that two principal methods of congestion control exist, passive and active. Passive mechanisms are preventive actions and are implemented at the time of design. One example of such control is oversizing of resources. This is something you often hesitate to do for fear of increasing the cost of the network; however, excess capacity can take care of some of the bursty behavior of traffic when it peaks and persists. Another example is limiting the number of available input and output buffers in a switch. This action will force the network to favor that traffic already in the network over that trying to enter. Similarly, fixed path routing limits the help that routing can provide. Call admission control (CAC) is another example of preventive action (see the article by Harry Perros and Khaled Elsyed in this issue). In active mechanisms the control reacts when congestion is experienced, which is why it is called reactive. Active flow control asks for the estimation of network states, propagation of information, and source traffic reduction. For example, in active control an ATM switch is able to detect and measure the congestion levels by average trunk (VPs and/or Vcs) utilization, average buffer utilization, and/or average queue length on the output trunks local to the switch.

What Is Credit-Based Window Flow Control?

Credit window flow control is a sliding-window flow control which assumes a pair of entities that control data flow between them. For each direction, there is a sending entity (SE) and a receiving entity (RE), and the SE transmits packets with a sequence number and receives acknowledgments (ACKs) from the RE. The RE may stop the SE from transmitting by not acknowledging receiving packets. This will not stop the SE immediately if the number of outstanding packets is smaller than the window size. This technique is used in low-speed packet-switched networks.

What Is Backpressure?

The concept of backpressure is similar to that of fluids flowing down a pipe. When one closes a pipe, the fluid tends to back up the pipe to the point of origin, where no more flow is allowed. The backpressure (BP) method implements per-VC flow control on every link; when a node detects congestion, it ceases to acknowledge packets on VCs that contribute to the congestion. This technique is used in low-speed packet-switched networks. Normally the window or credit on the link FC will be significantly smaller than that of an end-to-end flow control window. When an upstream node (closer to the VC source) sees the link VC window becoming exhausted, it ceases to acknowledge incoming packets on those VCs, and the process continues up the path until the source is throttled. The significant features are that every VC on every link must be individually flow-controlled, and that during congestion each node stores for each congested VC a number of packets or cells equal to the link VC window flow control. This concept applies whether a packet- or cell-switched network is under consideration. The only difference is that the acknowledgment cells do not exist in cell-switched networks.

What Is a Choke Packet?

A choke packet (CP) is a control packet that has been generated at a congested node; it travels against the flow. The significant feature of the design is the use of end-to-end flow control to support congestion control in which flow control is not implemented. Years after it first appeared, this concept was also adapted to flow control in cell -switched technology. In the ATM world, the concept of a choke packet was adapted in VC connections and was named after resource management (RM) cells. The source generates a number of RM cells for so many cells to send; RM cells travel through the uncongested and congested nodes and convey the congested states back to the source. Some of the RM cells also travel to the destinations as well, depending on the technique being implemented. The CP method concentrates on reducing input to the network; it reduces the input immediately, but tends to store the excess traffic in the congested node.

The World of Asynchronous Transfer Mode

In ATM networks, congestion can be short and frequent or long and bursty. However, no matter how the congestion builds up, the control is designed to reduce the flow of traffic entering the network in order to relieve congestion at a point within the network. To deliver the performance ATM promises, congestion control mechanisms used by the network should be simple, effective, and fair.

Two concepts originally captured from CP and BP methods were considered in standard Forums. They were named after credit-based and rate-based flow control.

What Is Credit-Based Control?

The credit-based approach consists of per-link, per-VC window flow control. Each link consists of a sender node (either a source end system or a switch) and a receiver node (either a switch or a destination end system). Each node maintains a separate queue for each VC. The receiver monitors the queue lengths of each VC and determines the number of cells the sender can transmit on that VC. This number is called a "credit." The sender transmits only as many cells as allowed by the credit. If there is only one active VC, the credit must be large enough to allow the whole bandwidth to be utilized at all times and also match the smallest link on the path. The link cell rate can be computed by dividing the link bandwidth by the cell size.

In the later version, the credit-based approach was enhanced to give adaptive credit where each VC only gets a fraction of the round-trip delay's buffer allocation. The fraction depends on the rate at which the VC uses the credit. For highly active VCs, the fraction is larger; for less active VCs, the fraction is smaller. Inactive VCs receive a small fixed credit. If a VC does not use its credit and its observed usage rate over a period is low, it gets a smaller buffer allocation in the next cycle. If a VC becomes active, it may take some time to go through cycles before it can use the full capacity of the link even if there are no other users. The credit-based approach has some implementation complexity at the switches; it requires per-VC queuing.

What Is Rate-Based Control?

The original proposal for a rate-based approach consists of end-to-end control using a single-bit feedback from the network. In the proposals, the switches monitor their queue lengths and, if congested, set the explicit forward congestion indication (EFCI) bit in the cell header. The destinations monitor EFCI bits; if set, the destinations generate an RM cell back to the source. The sources use an additive increase and multiplicative decrease algorithm to adjust the incoming rates. The RM cells are sent only to decrease the rate, but no RM cells are required to increase the rate. The problem is that RM cells may be lost due to heavy congestion in the reverse path, and the sources will keep increasing their load on the forward path. The rate-based approach has been enhanced to include positive feedback as well (i.e., sending RM cells on increase but not on decrease). When RM cells are sent for both increase and decrease, the algorithm is considered bipolar. The current rate-based algorithm not only supports the EFCI control mechanism, it also provides options for explicit rate control (ERC) as well as segmentation of the

flow in the control loop using virtual source and destination (see the ATM Forum Traffic Management Specification and the article by Kerry Fendick in this issue). The ERC limits the source's allowed current rate (ACR) to a specific value fluctuating between the minimum and maximum cell rates agreed to during connection setup. This algorithm is sensitive to feedback round-trip delay.

QUALITATIVE COMPARISON

Rate-based and credit-based approaches work quite differently. The rate-based method concentrates on reducing input traffic to the network, while the credit-based method concentrates on distributing the storage of the excess traffic. Differences are seen in both the processing overhead and the storage required by the algorithm.

The processing overhead in credit-based per-VC link flow control is considerable and present at all times. The rate-based method requires moderate processing. It escalates when congestion is detected and cleared. In credit-based, a per-VC flow control scheme must be implemented at all nodes. To implement per-VC flow control additional bits and pieces are needed; that is, acknowledgments have to be present for every n cells.

The credit-based method stores more cells within the network during congestion than does rate-based, but distributes them over more switches. The rate-based method stores fewer cells, but the storage tends to be concentrated at the congested switch. The congested switch is the best place for the stored cells in order to maintain a smooth flow, but this approach may require more buffer space.

If congestion is short-term and frequent, the rate-based method will generate appreciable overhead; if congestion is infrequent but long-term, the rate-based method will generate little overhead. The same considerations apply to storage: under relatively long-term congestion the credit-based method will tend to load up the network buffers with stored cells, and the user may not be aware of the congestion until after a delay has been experienced. In general, the rate-based method seems best suited to a network in which congestion is relatively infrequent and long-term. This situation could arise in an ATM network if a network busy hour were intense but infrequent, for example, largely due to file and image transfer. Different methods are best suited for different services. If cost and implementation permits, the two methods could coexist at all times.

Researchers are also engaged on how to flow control traffic in wireless networks and over satellites. It sounds like the usual classic problems when traffic is crossing a boundary from a media that has the concept of connection into a connectionless network. Each side offers a different quality of service and speed, from high to low error rates and from low to high error rates. Through the years, flow and congestion control have received a lots of consideration; however, there is still a gap between theoretical solutions and their practical implementations. Recent control designs suffer from embracing too many details which are supposed to take care of the complexities.

Several articles in this issue describe the control mechanisms used in available bit rate (ABR) services and quantify their performance. Kerry Fendick of AT&T Bell Laboratories describes the evolution of the ABR service. K. Kawashima, H. Saito, H. Kitasume, A. Koike, M. Ishizuka, and A. Abe of NTT Telecommunication Networks Labs evaluate the performance of ABR services; and Raj Jain, Shiv Kalynaraman, Rohit Goyal and Sonia Fahmy of the Ohio State University developed a closed-loop rate-based traffic management policy for ABR service. Emilio Leonardi and Fabio Neri of Politecnio di Torino, and Mario Gerla and Prasath Palnati of the University of California in Los Angeles considered high-speed, asynchronous, unslotted wormhole routing and compared two different flow control mechanisms, namely backpressure flow control and deflection routing. V. Catania, G. Ficili, S. Palazzo, and D. Panno of the University of Catania used fuzzy logic techniques in source traffic control; and Harry Perros of North Carolina State University and Khaled Elsayed of Cairo University reviewed CAC schemes. A. Arulambalam of New Jersey Institute of Technology, X. Chen of Bell Laboratories, and N. Ansari of New Jersey Institute of Technology address issues of fair rate allocation for ABR services; and D. Gaiti of Columbia University and the University of Paris VI and Nadia Boukhatem of the University of Versailles proposed a multi-agent system approach which uses adaptive intelligent agents. The guest editors wish to express their thanks to all contributors and to the staff of IEEE Communications Magazine for their support.

ADDITIONAL READING

G. Pujolle, "The Influence of Protocols on the Stability Conditions in Packet-Switching Networks," *IEEE Trans. on Commun.*, vol. 27, no. 3, pp. 611-19, 1979.

M. Pennoti and M. Schwartz , "Congestion Control in Store-and-Forward Tandem Links", *IEEE Trans. on Commun.*, vol. 23, pp 1434-43 , Apr. 1975.

G. P. Caseau,, "Markovian Models Applied to Computer Systems Stability Conditions," *IEEE Trans. Soft. Eng.* vol. 5, 6, pp 631-42, 1979.

C. G. Omidyar and W. Kelly, "Analysis of Congestion Control Protocols in Packet-Switched Networks," IEEE ICC, Amsterdam, The Netherlands, May 14, 1984.

J. Labetoulle and G. Pujolle, "Isolation Method in a Network of Queues", *IEEE Trans. Soft. Eng.*, vol. 6, no. 4, pp. 373-81, 1980.

L. Gerla and L. Kleinrock, "Flow Control: A Comparative Survey", *IEEE Trans. on Commun.*, vol. 28, no. 4, pp 553- , Apr. 1980.

BIOGRAPHIES

CAMBYSE GUY OMIDYAR received the D. SC. in electrical engineering from George Washington University in 1979 and has been an associate professorial lecturer there. He is a science advisor for the U.S. Internal Revenue Service and conducts research on emerging telecommunications technologies, modeling, and performance evaluation at the Research Institute of Illinois Institute of Technology. He was a consulting engineer for NASA on Earth Observing Systems -Mission to Plant Earth and Space Station Freedom, and worked on future NASA telecommunications networks. He worked on a variety of projects, including analysis and evaluation of Strategic Defense Initiative (SDI) battle management, command , control, and communications (BMC3) networking, known as Stars Wars, for the Strategic Defense Initiative Office (SDIO) of the Department of Defense. In the late 1980s he joined Bell Communications Research, where he was a major contributor to the T1M1.5 Lower Layer working group for the Operations System and Network Equipment Interface of the T1 standard committee.

GUY PUJOLLE received the Ph.D. and "Thèse d'Etat" degrees in computer science from the University of Paris IX and XI in 1975 and 1978, respectively. He is currently a professor at the University of Versailles and a member of the PRiSM Laboratory. Previously he was professor at the University of Paris VI and head of the MASI Laboratory under the direction of the CNRS (Centre National de la Recherche Scientifique) from 1981-1993. He was a professor at ENST (Ecole Nationale Supérieure des Télécommunications) from 1979 to 1981, and a member of scientific staff of INRIA (Institut National de la Recherche en Informatique et Automatique), 1974–1979. He is also chairman of the IFIP WG 6.4 on High Performance Networking and a member of the IFIP WG 7.3 on Performance Evaluation. He is also a member of the scientific committee of France Telecom, GMD, CRIM and chairman of the Telecommunication Regulation expert committee at the French Telecom Ministry. His research interests include the analysis and modeling of data communication systems, protocols, high-speed networks, and B-ISDN. He is the author of several books and many papers on diverse aspects of performance analysis and data communication networks.