Bandwidth Allocation for IP Traffic Over Satellite Links

Claude Bélisle, Peter Andreadis, Steve Bernier, François Lévesque, and Michel Barbeau*
Communications Research Centre Canada
3701 Carling Avenue
Ottawa, Ontario, Canada K2H 8S2
claude.belisle@crc.ca

Satellites are being deployed for extending the terrestrial Internet infrastructure to remote areas, offering tele-medicine and tele-education as well as e-mail, web browsing, and e-commerce. Satellite bandwidth is, however, an expensive resource. Static apportionment is not cost effective and explicit resource reservation techniques, as in ATM and RSVP protocols, are not exploited widely enough by end user applications to provide solution for efficient satellite bandwidth allocation. To address the vast majority of Internet traffic, a technique to implicitly evaluate the resource requirements is needed at the ground terminal. In this paper, such a technique, based on flow analysis and allowing dynamic bandwidth allocation and quality of service provisioning is discussed.

Index terms: satellite communications, ground terminals, TCP/IP, QoS, quality of service, bandwidth allocation, traffic management.

I. INTRODUCTION

The ubiquitous nature of satellite communications (satcom) makes it an ideal candidate for providing Internet services worldwide. In areas where terrestrial high-bandwidth communications infrastructures are impractical or non-existent, satcom may be the only solution. The advantages of combining the high bandwidth, wide area coverage, reconfigurability, and multicast capabilities of satellites with terrestrial networks provide immense new market opportunities. Fig. 1 shows a network architecture involving a mix of terrestrial and satellite links.

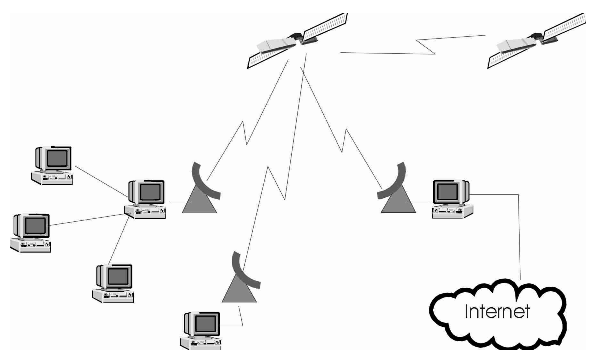

Fig. 1 Typical satcom network architecture.

As shown in the figure, users may be connected to the satellite link directly from their home or mobile terminal, via a local area network or through an Internet gateway. This architecture creates a wide variety of traffic bandwidth requirements by the different satcom ground terminals (GT). Each GT must determine its own bandwidth requirement and must periodically send a message to the master resource controller

* Current Address: Carleton University, School of Computer Science
Ottawa, Ontario, Canada K1S 5B6, barbeau@scs.carleton.ca

(MRC) to request bandwidth or to release it for others to use. The MRC, located either onboard the payload or on the ground, has the responsibility of dynamically apportioning the bandwidth between the different GTs based on the received requests. The apportioning scheme should efficiently allocate the bandwidth while respecting quality of service (QoS) requests required by certain applications.

Efficient bandwidth utilization, which is essential in a satcom environment, and quality of service (QoS), are unfortunately two competing goals. To offer QoS, a dedicated circuit is desirable but this may be to the detriment of bandwidth efficiency, especially for bursty traffic such as compressed video. To optimize the link utilization, bandwidth sharing is preferred. The concept of the virtual circuit, which is connection oriented, is used in the ATM protocol [1] and is being developed for the IP network in the form of the RSVP protocol [2]. However, only a minority of systems is based on the ATM or RSVP protocols. ATM-based satcom services are only in their infancy [3]. Most satellites do not have the necessary air interface to take full advantage of these protocols. The world's largest computer network, the Internet, uses the connectionless IP protocol. To address the vast majority of Internet traffic, a technique to implicitly evaluate the bandwidth requirement at the GT and to manage the traffic flow on the allocated bandwidth is essential.

In this paper, dynamic bandwidth allocation and QoS provisioning on satcom links are discussed in the context of the IP network. Section II deals with requests for dynamic bandwidth allocations. In Section III, the issue of QoS is discussed. In Section IV, traffic management at the GT is discussed. We conclude with section V.

II. REQUESTS FOR DYNAMIC BANDWIDTH ALLOCATION

The GT must send requests to the MRC to obtain or release channel capacity based on its dynamic evaluation of bandwidth requirements. This contrasts significantly with terrestrial networks where bandwidth between nodes (switches or routers) is fixed.

For a connectionless network, such as IP, bandwidth requirements are not explicitly defined by the user application and must be evaluated by the GT. The GT must determine when and how much bandwidth to request. An analogous problem has been addressed by Harita and Leslie [4] for variable rate ATM traffic on an ISDN network. They propose to measure either the incoming cell rate or buffer queue size and to request or release bandwidth when predetermined threshold values are crossed. To avoid unstable situations, average values of queue size and input data rates can be used.

In the ERICA protocol [5], a queue size monitoring approach is also adopted but in this case requests are sent at regular intervals to the MRC. The MRC collects queue size information from every GT and distributes bandwidth to the different GT to minimize network congestion. GTs with longer queue would be allocated more capacity to prevent congestion.

These techniques could be combined with the method proposed by Keshav et al. [6] to determine when bandwidth should be released. Their method is based on estimating a distribution of inter-arrival times of previous incoming data packets and takes into account the cost of holding the bandwidth and of releasing it.

For satellite systems these approaches may bring a problem of reactivity due to the long delays experienced between requests and allocations of bandwidth.

III. QUALITY OF SERVICE PROVISIONING

In order to provide QoS, a mean to prioritize the incoming traffic and report this information to the MRC is necessary. This would prevent a GT filled with low priority traffic to be allocated more bandwidth than a terminal with a smaller traffic volume but of higher priority. The incoming traffic should then be prioritized and high priority traffic should be given precedence over the low priority ones. This concept of precedence should be respected at the MRC when allocating bandwidth to the different GTs, and also at the GT level when selecting traffic to be transmitted on a congested link.

An approach could be to categorize the incoming traffic according to some predetermined rules and set priorities to each category. Traffic categorization may be done using the concept of flows as suggested in [7], [8], and [9]. In addition to dividing the incoming traffic packets according to source - destination address pair, other parameters such as the transport protocol and the application type (usually defined by the UDP/TCP port number) could be monitored to offer better traffic segregation. All incoming packets with similar parameters are then buffered together in a queue and treated as a distinctive traffic flow.

To each flow created, a certain priority can be associated based on a predetermined set of rules. It may be decided for example that traffic from a specific source address should be given top priority or, that priority allocation is application specific (real-time video for example) or some other combination.

The requests sent by the GT to the MRC may now indicate not only bandwidth requirements but also the priority level associated to the request. An algorithm, implemented in the MRC, allocates bandwidth to minimize traffic congestion while respecting the traffic priority requested by each GT. Virtual circuits (VC), with guaranteed bandwidth, can then be allocated to each GT.

On a congested link, a mechanism is required to determine when a flow should be considered no longer active and bandwidth be released. This will prevent a GT from holding the resources allocated to a VC after the communication is over, thus preventing other terminals from obtaining guaranteed bandwidth. This mechanism will also allow reduction in computational time required to service the flow and minimization of the GT buffer space requirements.

Claffy et al. [9] suggested a temporal analysis of the packet arrival. A flow is considered active as long a there are packets meeting the flow specification within an inter-arrival time less than a predetermined timeout value. With this definition, the problem now transposes to the choice of the timeout value. Holding a flow active too long may impact the bandwidth apportionment algorithm, since capacity must be secured to ensure the flow's expected QoS. This may prevent other terminals guaranteed bandwidth for their own flows. Releasing capacity allocated to a flow still active may also have adverse effect since obtaining the allocation again may not be guaranteed, especially on congested links.

An approach where the timeout value is evaluated according to an adaptive policy is thus preferable since it becomes possible to adjust dynamically the value of the timeout based on the current state of a flow. Keshav et al. [6] proposed such a technique based on the incoming packet inter-arrival time. A histogram of inter-arrival time is generated and used to calculate the timeout value that will minimize the cost of releasing a circuit. This releasing cost is function of the cost incurred for maintaining the circuit open and the cost to open a new one. For satellite systems, where propagation delays are long and bandwidth limited, this estimation is of utmost importance.

IV. GT TRAFFIC MANAGEMENT

Upon allocation of bandwidth by the MRC, the GT must perform traffic management and congestion control to service the many flows created. Many scheduling schemes have been proposed to perform this task. Two of the most popular techniques are the round-robin (RR) and weighted-fair-queuing (WFQ) algorithm. In the RR approach, each queue is serviced one after the other, with one packet being removed at a time. This approach could, however, lead to unfair distribution of the bandwidth if packets from the different queues are not of equal size. The WFQ scheduler addresses this problem by retrieving packets based on the time of arrival and their size. The WFQ remains relatively fair even when packets of different sizes are sent.

However, those two popular scheduling techniques are still considered best effort allocation processes. They do not allow QoS provisioning. To allow guaranteed bandwidth to the various flows, a priority-based WFQ scheduler is required. Flows of higher priority are serviced first using a WFQ scheduler.

V. CONCLUSION

In this paper, techniques for efficient dynamic bandwidth allocation and quality of service provisioning for IP traffic on satcom links were addressed. Information about incoming data rate or queue size combined with threshold decision has been suggested. QoS provisioning was addressed by categorizing incoming traffic into flows and assignation of priorities. A priority-based WFQ can be used for traffic management within the GT.

ACKNOWLEDGMENTS

This research is sponsored by the Canadian Space Agency and the Canadian Department of National Defence.

REFERENCES

[1] ATM Forum, http://www.atmforum.com

[2] L. Zhang, S.Deering, D.Estrin, S.Schenker, and D.Zappala,"RSVP: A New Resource Reservation Protocol", IEEE Network, 7(9), pp.8-18, Sept. 1993.

[3] I.F. Akildiz and S.H. Jeong, "Satellite ATM Networks: A Survey", IEEE Communications Magazine, pp.30-43, July 1997.

[4] B.R Harita and M. Leslie, "Dynamic Bandwidth management of primary rate ISDN to Support ATM access", Proceedings of ACM SIG-COMM'89, pp. 197-210, 1989.

[5] R.Jain, S.Kalyanaraman, R.Goyal, S.Fahmy, and R.Viswanathan, "ERICA Switch Algorithm: A Complete Description", ATM Forum/96-1172, August 1996.

[6] S. Keshav, C.Lund, S.Phillips, N.Reingold, and H. Saran, "An Empirical Evaluation of Virtual Circuits Holding Time Policis in IP-over-ATM networks", IEEE journal on Selected Areas in Communications, Oct. 1995.

[7] N. Brownlee, C. Mills, and G. Ruth, "Traffic Flow Measurement: Architecture", IETF RFC-2722, Nov. 1999.

[8] R. Jain and S.A. Routhier, "Packet trains measurement and a new model for computer network traffic", IEEE Journal on Selected Areas in Communications, vol. 4, no.6 pp.986-995, Sept. 1986.

[9] K.C. Claffy, H.W. Braun, and G.C. Polyzos, "A Parametrizable Methodology for Internet traffic Flow Profiling", IEEE Transactions on Selected Areas in Communications, 13(8), October 1995.