In Proc. Eleventh IEEE Workshop on Computer Communications, Reston, VA (Sept. 1996).

Efficient Resource Management for Satisfying Diversified QoS Guarantees*

Yin Bao <bao@cis.udel.edu>
Adarshpal S. Sethi <sethi@cis.udel.edu>

Department of Computer and Information Sciences
University of Delaware, Newark, DE

Extended Abstract The major task of a network is to provide transmission service to different applications with diversified QoS (Quality of Service) requirements. Various QoS control schemes have been proposed and studied [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. For each application, there is a functional relationship between the supported QoS and the resource needed. Given a limited amount of network resources, usually in terms of buffers and bandwidth, it is desirable to allocate as few resources as possible, as long as the amount of resources is enough to satisfy the requested QoS; while the rest of the resources can be used for other applications. Due to the increasing dynamics and unpredictability of traffic in current network applications, it is necessary to have dynamic resource management to adjust the relationship between the resources needed and the change of traffic characteristics over the lifetime of a transmission to ensure the twin objectives of QoS satisfaction and network efficiency. Meanwhile, many real-time network applications require specific end-to-end delay bound and loss ratio QoS. Thus, QoS control should try to provide *accurate* realization of QoS as close as possible to what is requested, not just to a certain degree.

We adopt the approach of adjusting resource allocation of each flow to achieve the requested QoS. With this approach, a referenced amount of resources is allocated before the start of a transmission, and this amount is adjusted during the lifetime of the transmission based on a set of criteria such as QoS performance and resource utilization. The gain from dynamic resource adjustment is that for each flow, the average resource needed for ensuring requested QoS is smaller than that under static resource allocation, therefore more flows may be accommodated into the network. Here we introduce two dynamic resource management schemes to satisfy a diversified combination of delay bound and loss ratio requirements. In both schemes, the allocated buffer size and service rate for a flow are coordinated so that the admitted packets will be served before their deadlines.

The main idea of **buffer-directed QoS control** is to achieve QoS by dynamically changing the allocated resources based on the observed QoS performance. Before the start of transmission, an upper loss ratio bound L_u and a lower loss ratio bound L_l are chosen according to the loss ratio requirement of this flow L ($L_l < L < L_u$). The dynamic adjustment of resources intends to control the loss ratio performance to be within the $[L_l, L_u]$ interval: if the loss ratio performance is higher than L_u, more resources are allocated to this flow; if it is lower than L_l, some allocated resources are released from this flow. The initial simulations and stochastic analysis show that the observed loss ratio is less likely to reach the boundaries of the $[L_l, L_u]$ interval as time goes on. We call this phenomenon—the slowing down of the change of loss ratio—the *waving effect*. Generally, the waving effect makes the loss ratio performance less responsive to change in

*This work was supported in part by the U.S. Department of the Army, Army Research Laboratory under Cooperative Agreement DAAL01-96-2-0002 Federated Laboratory ATIRP Consortium.

statistics as time goes on. Analysis also shows that due to the waving effect, the minimum loss ratio dropping time from L_u to L_l is linearly increasing with time. This is detrimental because when the observed loss ratio reaches L_u, the amount of resources is increased to a relatively high level, and this amount is reserved until the observed loss ratio drops to L_l. The improved version of this method uses dynamic loss ratio boundaries to deal with the waving effect. The idea is to bring the loss ratio upper bound L_u and lower bound L_l gradually closer to the mean L to enable the resource adjustment to have an equal opportunity to be invoked throughout the lifetime of the transmission even under the waving effect. The simulations conducted on various sources show that in most cases the scheme successfully controls the QoS performance accurately as requested with reasonable amount of resources on average. Meanwhile, the dynamically shrinking boundary effectively eases the damage caused by the waving effect and supplies the flexibility of choosing different control granularities.

OCP_A(OCcuPancy_Adjusting) is a different resource allocation adjustment scheme towards better resource efficiency. In OCP_A, resource adjustment is responsive to the allocated buffer resource utilization as well as QoS requirements. The resource adjustment is based on the observation of the buffer occupancy: the allocated resources are increased if the buffer is full; decreased if the buffer occupancy is low. The buffer occupancy thresholds should be chosen carefully to prevent unnecessary resource adjustment oscillation. Simulations show that though the resource efficiency is good, the frequency of resource adjustment is high, especially for bursty traffic such as real-time video. It is our opinion that frequent resource adjustment, though resulting in good resource efficiency, may not be always preferable for the following reasons. First, it introduces more control overhead. Second, it makes it harder to track the excess unused resources due to the lack of stability. To prevent frequent resource re-adjustment, a confidence period is used for decreasing resources: the resources can be only decreased if the buffer occupancy is *consistently* low for a certain period of confidence time. Simulations have been conducted on various sources to study the effects of different values of confidence time. As a result, it is recommended that the system should always choose the confidence period based on the frequency of the traffic rate change, plus the consideration of the resource efficiency and control overhead trade-off.

Both dynamic resource management schemes are general schemes for QoS control which can be applied onto various types of sources within various network environments. The performance of these schemes for multiple flows sharing the resources is currently under study. We are also studying the effects of combining OCP_A into buffer-directed QoS control method to improve resource efficiency. The future work involves extending the method to satisfy end-to-end QoS. The buffer-directed dynamic QoS control scheme may be also revised to satisfy *instantaneous* loss ratio requirement over shorter periods of time than the transmission lifetime, which is desirable for multimedia applications. This can be done by introducing a window size for the loss ratio performance observation and boundary shrinking.

References

[1] A.Demers, S.Keshav, and S. Shenkar. Analysis and simulation of a fair queueing algorithm. *Internet. Res. and Exper.*, 1, 1990.

[2] Raffaele Bolla, Franco Davoli, Alfio Lombardo, Sergio Palazzo, and Daniela Panno. Adaptive bandwidth allocation by hierarchical control of multiple ATM traffic classes. In *Proc. IEEE INFOCOM '92*, volume 1, page 1A.4.1, 1992.

[3] Geping Chen and Ioannis Stavrakakis. Management of ATM traffic with diversified loss and delay requirements. In *Proc. IEEE INFOCOM '96*, 1996.

[4] David D. Clark, Scott Shenker, and Lixia Zhang. Supporting real-time applications in an integrated service packet network: Architecture and mechanism. *ACM SIGCOMM'92*, (8), 1992.

[5] S.Jamaloddin Golestani. Congestion-free transmission of real-time traffic in packet networks. In *Proc. IEEE INFOCOM '90*, volume 2, page 527, 1990.

[6] S.Jamaloddin Golestani. Self-clocked fair queueing scheme for broadband applications. In *Proc. IEEE INFOCOM '94*, volume 2, pages 636–646, 1994.

[7] Tim Kowk. A vision for residential broadband services: ATM-to-the-Home. *IEEE Network*, 9(5), Sept.-Oct. 1995.

[8] Abhay K.Parekh and Robert G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single-node case. *IEEE/ACM Transactions on Netwroking*, 1(3), June 1993.

[9] Simon S. Lam and Geoffrey G. Xie. Burst scheduling: Architecture and algorithm for switching packet video. In *Proc. IEEE INFOCOM '95*, pages 8.a.1.1–8.a.1.11, 1995.

[10] Aurel A. Lazar and Giovanni Pacifici. Real-time scheduling with quality of service constraints. *IEEE Journal on Selected Areas in Communications*, 9(7), September 1991.

[11] Aurel A. Lazar and Giovanni Pacifici. A seperation principle between scheduling and admission control for broadband switching. *IEEE Journal on Selected Areas in Communications*, 11(4), May 1993.

[12] S.S. Panwar, D. Towsley, and J.K.Wolf. Optimal scheduling policies for a class of queues with customer deadlines to the beginning of services. *Journal of ACM*, 1988.

[13] R.Bolla, F.Danovaro, F.Davoli, and M.Marchese. An integrated dynamic resource allocation schemem for ATM networks. In *Proc. IEEE INFOCOM '93*, volume 3, page 1288, 1993.

[14] R.Chipalkatti, J.F.Kurose, and D. Towsley. Scheduling policies for real-time and non-real-time traffic in a statistical multiplexer. In *Proc. IEEE INFOCOM '89*, pages 774–783, Canada, April 1989.

[15] Dimitrios Stiliadis and Anujan Varma. Latency-rate servers: a general model for analysis of traffic scheduling algorithms. In *Proc. IEEE INFOCOM '96*, volume 1, pages 111–119, 1996.

[16] Lixia Zhang. Virtualclock: A new traffic control algorithm for packet switching netwroks. In *Proc. ACM SIGCOMM '90*, pages 19–29, August 1990.