Resource Allocation for Multi-Media Communication in Heterogeneous Network

Anirudha Sahoo, Wei Zhao, * Biao Chen, †

1 Introduction

Different from traditional data communication, multimedia communication has two distinct characteristics: 1) It is delay sensitive i.e there is a timing constraint associated with it. If a multimedia data packet is delayed beyond a certain time bound, then the packet may be of little or no use. So multimedia communication requires delay guarantees from the underlying network. 2) Multimedia communication consumes huge amount of network bandwidth. Multimedia applications usually need to transmit an enormous amount of data in a short period of time, thus requiring high bandwidth from the underlying network. For example, a video conferencing application may consume bandwidth of several megabits per second.

These characteristics demand new methodology for resource management in the underlying networks. Much progress has been made for multimedia communication in homogeneous networks such as FDDI, IEEE 802.5 token ring, and ATM. Bandwidth allocation for FDDI network has been addressed in [13, 16]. ATM network has been studied for time constrained applications in [10, 11]. Scheduling policies for ATM switches are analyzed in [3, 4, 5, 6, 8, 14].

But many existing networks used today are heterogeneous in nature, where different types of LANs are connected to form a network to provide the communication services. Unfortunately, The results from the studies of homogeneous networks may not be directly applied to a heterogeneous environment. In a heterogeneous network, resorting to an efficient method of resource allocation for one LAN segment may be good for that LAN segment, but may adversely affect the performance of other LAN segment in the network.

*Department of Computer Science, Texas A&M University, College Station, Texas 77843-3112, Phone: (409) 845-5098, Fax: (409) 847-8578, e-mail: {asahoo,zhao}@cs.tamu.edu

†Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688, phone : (972) 883 2232, email:bchen@utdallas.edu

Hence an integrated resource allocation scheme, in which performance of all the LAN segments in the network is taken into account, should be followed. In this paper, we will discuss issues related to resource management for multimedia applications in a heterogeneous environment.

Multi-media communication is usually connection-oriented. In this kind of communication, resource management is carried out for a new incoming connection in connection admission control (CAC) process. During the CAC process, the system determines whether or not the new connection should be admitted, and if so, how much resource on each legacy LAN should be allocated to the newly admitted connection. Heterogeneity of the network raises three unique issues in this resource allocation:

- *Non conservative allocation:* Generally speaking, in a single shared media network (say, FDDI) the minimum resource should be allocated to a connection as long as its QoS can be satisfied. This method may not be proper in a heterogeneous network. In a heterogeneous network, if the minimum bandwidth is allocated to a connection, then the worst case delay of that connection will be too tight. As a result, the disturbance introduced due to future connections may make this connection to violate its deadline requirement. Thus, this will prevent many future connections from being admitted, causing an adverse effect on the performance of the network.

- *Balanced bandwidth usage on legacy LANs:* Connections go through different legacy LANs in a heterogeneous network. Resources should be allocated out of legacy LANs in a proportional manner so that the network, as a whole, will remain balanced. Otherwise, the LAN with less resource will rapidly become a bottleneck for admissibility of future connections.

- *Dealing with bandwidth mismatch between legacy LANs:* In a heterogeneous network, there may be legacy LANs with a mismatch in their capacity or bandwidth. So when a station belonging to the higher capacity LAN sends data to another station in the lower capacity LAN, the lower capacity LAN may be congested, if it cannot handle the high data rate with which the sender is sending data.

In this paper, we will describe two approaches we have taken in dealing with these issues, namely, the *feasible region approach* and the *deadline partition* approach. With the *feasible region approach*, a region for feasible resource allocation is first identified. That is, any allocation of resources in this region for the new connection will satisfy the QoS requirement of the new as well as the existing connections. The key step is then to carefully select one allocation in the feasible region so that the three issues mentioned above are properly addressed. With the *deadline partition approach*, the deadline of the entire connection is divided into sub-deadlines. Each sub-deadline corresponds to one legacy LAN segment. Resource in each legacy LAN is then allocated according to the sub-deadlines. Each approach has its advantages and disadvantages. The feasible region approach is "optimal" in the sense that a new connection will be admitted as long as any other method can do so. But it involves complex analysis of the entire network. On the other hand, the deadline partition method is a heuristic one. But it is relatively simple and can be easily implemented in a large system.

2 Resource Allocation in ATM-Based Heterogeneous Network

2.1 Network Model

We will focus our study on ATM based heterogeneous network. An ATM-Based Heterogeneous Network (ABHN) is a mixed media network, in which several shared media legacy LANs are interconnected via a backbone of ATM network. The legacy LANs can be token ring, ethernet, or FDDI. They are connected to ATM network via what are known as *edge devices*. Edge devices are switches that convert packets from legacy LANs to ATM cells and vice versa, to provide communication between them. Furthermore, an ABHN network can be either *symmetric* or *asymmetric*. In a symmetric ABHN, the legacy LANs connected on either side of the ATM backbone are of the same type, i.e., the same capacity and the same technology. The examples are FDDI-ATM-FDDI and Tokenring-ATM-Tokenring networks. A typical FDDI-ATM-FDDI network is shown in Figure 1. In an asymmetric ABHN, the legacy LANs connected on either side of the ATM backbone can be of different type. Figure 2 shows an example of an asymmetric ABHN. The bandwidth mismatch problem exists in asymmetric networks.

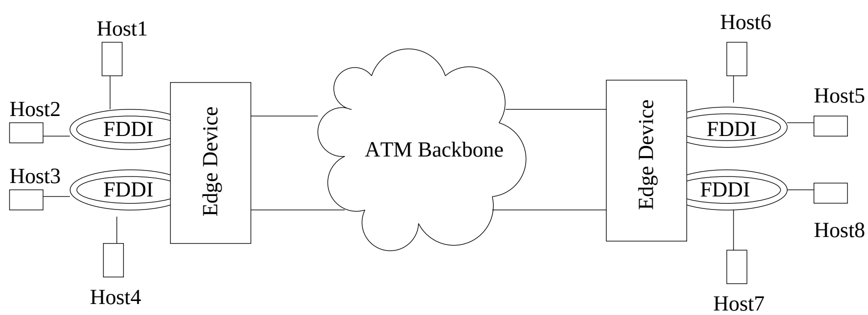


Figure 1: An FDDI-ATM-FDDI network

2.2 Feasible Region Approach

With this approach, for a new connection establishment request, a feasible region of resource allocation is first identified. That is, any resource allocation for legacy LANs that falls in the region will fulfill the QoS requirement of the new as well as all the existing connections. If the feasible region is empty, then new connections cannot be admitted. Otherwise, a particular allocation from the region is selected. In the selection, care should be taken to

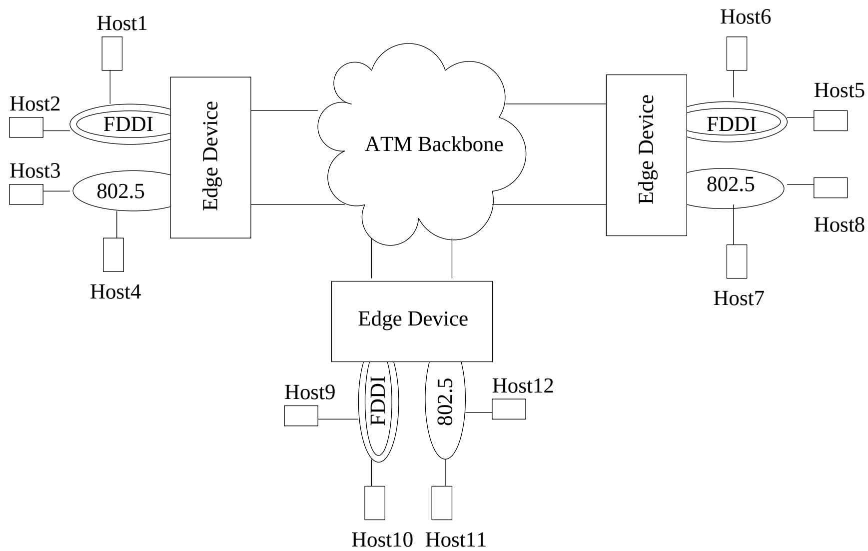


Figure 2: An Example of Asymmetric ATM Based Heterogeneous Network

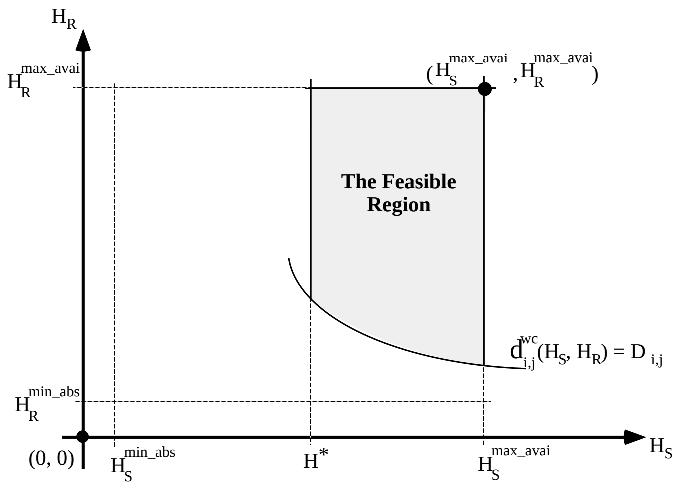
ensure that three issues mentioned in Section 1 (i.e., non conservative allocation, balanced bandwidth usages, and bandwidth mismatch) are properly addressed.

We have taken this approach for resource allocation in symmetric ABHN networks. We now illustrate this approach based on FDDI-ATM-FDDI networks (for an example, refer to Figure 1). The method for Tokenring-ATM-tokenring networks is similar [12]. The resource to be allocated in FDDI-ATM-FDDI network is the synchronous bandwidth of both (sender and receiver side) FDDI LAN segments. Let H_S and H_R be the bandwidth allocated to a new connection on sender and receiver FDDI LANs respectively. Let $\vec{d}(H_S, H_R)$ be the worst case end-to-end delay vector. d_i, the i-th element of this vector, is the worst case end-to-end delay of connection i. Note that the delays vary, depending on the values of H_S and H_R. That is, this vector is a function of H_S and H_R. Let $\vec{D}$ be the deadline vector. That is, D_i, the i-th element of this vector, specifies the deadline of connection i. Both the vectors count the delays and deadlines for the new and existing connections.

Using these notations, the feasible region for allocation H_S and H_R is defined by the following inequality:

$$\vec{d}(H_S, H_R) \leq \vec{D}. \tag{1}$$

That is, any pair of values H_S and H_R that satisfy the above inequality is a feasible allocation. Clearly, to determine the feasible region we need to derive the expression the worst case end-to-end delay as a function of H_S and H_R. To achieve this, we have resorted to a decomposition method [9, 10]. In this method, each component of the network is modeled as a server and worst case delay for a connection is calculated by summing up the worst case delays suffered by the connection at each server. The details of derivation of the worst case delay vector and the discussion of properties of feasible regions are not given here due to space limitation. See [2] for details. Figure 3 shows a typical feasible region



$H_S^{min\_abs}$ is the minimum bandwidth that is absolutely required at the sender FDDI LAN for the new connection. $H_S^{max\_avai}$ is the maximum bandwidth that is available at the sender FDDI LAN when a new connection arrives. $H_R^{min\_abs}$ and $H_R^{max\_avai}$ are the corresponding terms for the receiver FDDI. H^* is the maximum of all the bandwidth allocated to the existing connections at the sender FDDI. $d_{i,j}(H_S, H_R)$ is the worst case delay incurred by the incoming connection and $D_{i,j}$ is the deadline of the new connection.

Figure 3: An Example of Feasible Region for FDDI-ATM-FDDI network

for an FDDI-ATM-FDDI network.

Having determined the feasible region, the next step is to determine the point to be chosen from the feasible region that will represent the amount of resource to be allocated to the two FDDI LANs. As mentioned in the introduction, this is a complex problem. If too much bandwidth were allocated to the requesting connection, a future connection request might be rejected due to insufficient available resource. On the other hand, if too little bandwidth were allocated to the requesting connection, its worst case delay would be very "tight" (i.e., very close to its deadline). Consequently, disturbance introduced by future connections may likely result in the violation of its deadline constraint. Thus, the new connection will be rejected. Hence, excessive or insufficient allocation of resources to a connection may jeopardize the chance of a future connection being admitted. Furthermore, the allocation should result in a relatively balanced usage of bandwidth between the LANs involved. An uneven usage of bandwidth may create bottleneck, causing performance degradation.

We have designed and analyzed a proportional allocation scheme i.e. bandwidth allocated for a new connection is proportional to the available bandwidth of each FDDI LAN. In this way, the allocated bandwidth is neither too conservative nor too excessive. The remaining bandwidth on the two FDDIs will remain relatively balanced. Thus, if the sender FDDI is loaded more than the receiver, less bandwidth will be allocated from the sender FDDI, whereas more bandwidth will be allocated from the receiver FDDI. Extensive performance evaluation shows that this allocation scheme results in good overall network performance. It out-performs several other allocation methods. The details data on performance evaluation is not shown here due to space limitation. See [2] for details.

2.3 Deadline Partition Approach

In this approach the deadline associated with a connection is partitioned into as many sub-deadlines as the number of LAN segments in the path of the connection. Each LAN segment is then assigned a sub-deadline. The resource for each LAN is then allocated accordingly. In this approach, the partition step is critical in addressing the three issues mentioned earlier (i.e., non-conservative allocation, balancing bandwidth usage, and dealing with bandwidth mismatch).

We designed, analyzed, and evaluated several deadline partition methods. We briefly describe them as follows.

- *Equal partition method.* In this method, the original deadline of the new connection is divided equally among the LAN segments in the path of the connection. Obviously this is a very simple partition scheme. This scheme may work well in a symmetric ABHN, because all the LANs are of same capacity. But this scheme may perform poorly in an asymmetric ABHN. In asymmetric ABHN, there may be a mismatch in the capacity of the LAN segments. Giving equal sub-deadlines to both a high and a low capacity LANs may result in an unbalanced resource allocation.

- *Capacity-inverse-proportional partition method.* In this method, the original deadline of the new connection is divided such that the sub-deadlines are inversely proportional to the capacity of the respective LANs. Thus, the LAN segment with lower capacity is assigned a sub-deadline larger than that of a higher capacity LAN segment. That is, this method addresses the bandwidth mismatch problem. Hence, although for a symmetrical ABHN, this scheme is equivalent to the equal partition method scheme, for asymmetric ABHN this scheme should perform better. On the other hand, this method is a static one and hence cannot react properly in a fluctuating load environment of LANs.

- *Utilization-proportional partition method.* In this method the original deadline is divided among the LANs such that the sub-deadlines are directly proportional to the utilization of the respective LANs. Thus, if a LAN has a higher utilization than the other, then this LAN will have larger sub-deadline than the others. It addresses the bandwidth balancing problem in the sense that it assigns sub-deadline according to the runtime utilization of the LAN.

- *Integrated partition method.* This method synthesizes both the capacity-inverse-proportional partition method and utilization-proportional partition method. That is, the deadline is divided such that the sub-deadlines are directly proportional to the utilization and inversely proportional to the capacity of the respective LANs. So this method takes advantage of both the methods while overcoming their weakness. However, this method involves more computation overhead than the others.

We have compared the performance of systems with the four different deadline partition methods. The network studied is an asymmetric one, similar to the one shown in Figure 2. As expected, the integrated partition method turned out to be the best method for all utilization, whereas utilization-proportional scheme faired better than capacity-inverse-proportional scheme at a higher utilization. Capacity-inverse-proportional scheme was better than utilization-proportional for light load condition. Detailed analysis and discussion on performance data are not given here due to space limitations. Interested readers are referred to [12].

3 Final Remarks

We have addressed resource allocation problem in ATM based heterogeneous network for multimedia applications. Resource management in heterogeneous network is much more complex than their homogeneous counterpart. We have identified three issues that are unique to resource allocation in a heterogeneous network (i.e., non-conservative allocation, balanced bandwidth usage, and bandwidth mismatch). We have described two approaches that tackle those issues. Out of these two approaches, the feasible region approach, although optimal, involves complex analysis of the network. Hence it may not be scalable to a large network. The deadline partition method, on the other hand, is a heuristic one but is relatively simple and can be easily implemented in a distributed manner.

Results of our research may be extended to a virtual LAN (VLAN) networking environment. In a VLAN, the hosts may belong to different LAN types and the packets in a VLAN are switched, not routed. So the packet has to travel through different shared and switching media. The commercial availability of Virtual LAN products [7, 15] has added importance to this problem.

References

[1] Gopal Agrawal, Biao Chen, Wei Zhao, Sadegh Davari. "Guaranteeing Synchronous Message Deadlines in High Speed Token Ring Networks with Timed Token Protocol" - In *Proceedings of the 12th IEEE International Conference on Distributed Computing Systems*, pages 468-475, June 1992.

[2] B. Chen, A. Sahoo, W. Zhao, A. Raha. Connection-Oriented Communications for Real-Time Applications in FDDI-ATM-FDDI Heterogeneous Networks. *Technical Report*, Department of Computer Science, Texas A & M University, December 1996.

[3] D. D. Clark, S. Shenker, and L. Zhang. Supporting real-time applications in an integrated services packet network: Architecture and mechanism. *Proc. of ACM SIGCOMM'92*, pg 14–26, Aug. 1992.

[4] A. Demers, S. Keshav, and S. Shenker. Analysis and simulation of a fair queueing algorithm. *Proc. of ACM SIGCOMM'89*, pg 1–12, Sept. 1989.

[5] D. Ferrari and D. C. Verma. A scheme for real-time channel establishment in wide-area networks. *JSAC*, SAC-8(3):368–379, Apr. 1990.

[6] S. J. Golestani. A framing strategy for congestion management. *JSAC*, 9(7):1064–1077, Sept. 1991.

[7] 3Com corporation, "The Virtual LAN Technology Report" - http://www.3com.com/nsc/200374.html.

[8] C. R. Kalmanek, H. Kanakia, and S. Keshav. Rate controlled servers for very high-speed networks. *Proc. of IEEE Global Telecommunications Conf.*, pg 300.3.1–300.3.9, Dec. 1990.

[9] A. Raha, S. Kamat, W. Zhao. Admission Control for Hard Real-Time Connections in ATM LANs *Proc. of INFOCOMM'96*, pg 180–188, March 1996.

[10] A. Raha, S. Kamat, W. Zhao, "Guaranteeing End-to-End Deadlines in ATM Networks", 15th International Conference on Distributed Computing Systems, pages 60-68, May 1995.

[11] A. Raha, "Real-Time Communication in ATM Networks" - PhD thesis, Department of Computer Science, Texas A & M University, 1996.

[12] A. Sahoo, W. Zhao, "Real-Time Communication Over Asymmetric ATM-Based Heterogeneous Network" - Technical Report, Department of Computer Science, Texas A & M University, December 1996.

[13] J.K. Strosnider, J. Lehoczky, L. Sha, "Advanced Real-Time Scheduling Using the IEEE 802.5 Token Ring" - Proc. of the IEEE Real-Time System Symposium, Dec. 1988.

[14] L. Trajkovic and S. J. Golestani. Congestion control for multimedia services. *IEEE Network*, 6(5):20–26, Sept. 1992.

[15] XYLAN Corporation, "AutoTracker Virtual LAN Architecture" - http://www.xylan.com/whitepaper/AUTOTRAC/00_CONT.HTM

[16] L. Yao, W. Zhao, "Performance of an Extended IEEE 802.5 Protocol in Hard-Real-Time Systems" - Proc. of the IEEE INFOCOM '91, April 1991.

[17] Q. Zheng, K.G. Shin, "Synchronous Bandwidth Allocation in FDDI Networks" - Proceedings of the ACM Multimedia '93, page 31-38, Aug. 1993.