Partition-Based Admission Control in Heterogeneous Networks for Hard Real-Time Connections [*]

Anirudha Sahoo and Wei Zhao
Department of Computer Science,
Texas A&M University,
College Station, Texas 77843-3112,
e-mail: {asahoo, zhao}@cs.tamu.edu

Weijia Jia
Department of Computer Science,
City University of Hong Kong,
Kowloon, Hong Kong
email: wjia@cs.cityu.edu.hk

Abstract

Efficient and accurate connection admission control (CAC) is essential in hard real-time communication. In a hard real-time system, every admitted connection must meet its deadline, otherwise the consequences may be catastrophic. Accurate delay analysis and proper amount of resource allocation are pivotal to the overall performance of the system. Both are more difficult to deal with in heterogeneous networks than in homogeneous networks. We address the issues of connection admission control in heterogeneous network. In particular, we design and analyze CAC methods based on partitioning deadlines. We demonstrate that this kind of partition-based CAC methods are simple, efficient, and scalable. Simulation data showed that systems with this kind of CAC methods can perform effectively in a wide range of system and payload parameters.

Keywords: Heterogeneous Networks, Connection Admission Control, Real-Time Communication, Resource Allocation.

1 Introduction

Much research has been done in Real-time communication. In real-time communication, messages are delay sensitive and have deadlines associated with them. Real-time messages have to be delivered before their deadlines so as to avoid disastrous consequences. Real-time communications have been studied for CSMA/CD [12], token ring [19], FDDI [1, 14], slotted ring [16] and ATM and point-to-point networks [15]. Most of these studies are based on homogeneous networks. However, a lot of existing communication networks are heterogeneous in nature, consisting of different homogeneous subnetworks. Our study in this paper will be concentrated on hard real-time connection in heterogeneous networks, specifically in ATM-Based Heterogeneous Network (ABHN). We will mainly deal with ABHN consisting of FDDI and token ring as the legacy LAN segments in the network.

Hard real-time connections have strict deadlines associated with them. Once such a connection is admitted to the network, it cannot miss its deadline. So connection admission control (CAC) plays a vital role in hard real-time systems. CAC is the process by which the quality of service (QoS) requirement of an incoming connection is evaluated to decide whether to accept or reject it. To request for a connection, an application submits a request with a description of the traffic it is going to send and the QoS it expects from the network for the connection. For hard real-time connections, deadline is the most important QoS. To guarantee that a connection meets its deadline, the CAC has to determine the worst case end-to-end delay that a connection may encounter. Delay analysis is thus an integral part of the CAC. Apart from delay analysis, the CAC has to decide how much resource should be allocated to the connection so that the new connection meets its QoS requirement without violating QoS requirement of the existing connections. At the same time, the amount of resources allocated to the new connection should not reduce the chances of admitting a future connection.

CAC for heterogeneous network is more complex than that for homogeneous network. This is because heterogeneity of the network makes delay analysis more complicated. Furthermore, heterogeneity gives rise to additional issues in resource allocation:

- *Non conservative allocation*: In a single shared media LAN (e.g., FDDI), minimum resource is usually allocated as long as the QoS is guaranteed. But in a heterogeneous network, if the minimum resource is allocated, then the delay of the connection will be too tight. So the disturbance

---

[*] This work was partially sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, grant number F49620-96-1-1076, by Texas Higher Education Coordinating Board, grant number 999903-204 and by research strategic grant (No. 7000664-540) of City University of Hong Kong.

created by a future connection may make this connection violate its QoS requirement. Hence many future connections may be discarded. This will adversely affect the system performance.

- *Balanced bandwidth usage on LAN segments*: Resources should be allocated in a proportional manner so that the LAN segments are kept relatively balanced with respect to each other. If the allocation is not proportional, then the LAN with less resource will soon become a bottleneck, and thus may bring the system performance down.
- *Bandwidth mismatch between LAN segments*: In a heterogeneous network, there may be a (huge) mismatch of bandwidth or capacity between different LANs. For example, an FDDI has 100Mb/s bandwidth while a 802.5 token ring has only 16Mb/s bandwidth. When such LANs are involved in a connection, the LAN with higher capacity may overwhelm the one with lower capacity with large amount of data. This may lead to congestion in the lower capacity LAN.

In [3], a CAC was developed for an FDDI-ATM-FDDI heterogeneous network. This CAC used *feasible region* method to admit a connection. In this method, a region for feasible resource allocation was first identified. That is, any allocation of resources in this region for the new connection satisfies the QoS requirement of the new as well as the existing connections. The key step is then to carefully select an operational point in the feasible region. An advantage of the feasible region method is that it is an *optimal* method in the sense that it will always admit a new connection as long as any other method can do so. But this method is relatively complex and cannot scale well to any new LAN technology that may be introduced to heterogeneous network.

In this paper, we will discuss a new method of connection admission control which uses *deadline partition* approach. With this method, the deadline of a connection is partitioned into several sub-deadlines, one for each LAN segment. Then, the resources in the LAN segments are allocated according to the sub-deadlines. This method overcomes the drawbacks of the feasible region approach. In particular, the new method is simpler and more efficient in terms of computational complexity. It also provides a good scalability in comparison with the feasible region method. For example, the feasible region method proposed in [3] was designed for a heterogeneous network with FDDI LAN segments. If a network contains a token ring LAN segment, then a new way to analyze the feasible region has to be developed. Our new deadline partition method will not have such problem. The deadline partition method uses resource allocation methods developed for individual LAN segments and hence is applicable to any heterogeneous network as long as its LAN segments have been individually analyzed.

2 Related Work

A good amount of work has been done in the design and analysis of real-time communication networks. Determining delay bounds has been the main thrust of most of the research done in real-time communications. Worst case delay bounds have been analyzed in homogeneous networks in [1, 7, 15, 16, 17, 19]. Both shared media networks (e.g., IEEE 802.5, FDDI, and DQDB) and point-to-point networks (e.g., ATM) have been studied for real-time communications over homogeneous networks.

Worst case delays have been derived for IEEE 802.5 token ring networks [19], slotted ring networks [16], and FDDI networks [1]. Various connection admission criteria for different scheduling policies have been studied for connection-oriented networks. Earliest Deadline First scheduling for wide area networks has been analyzed by Ferrari and Verma in [7]. Deterministic delay bounds in networks have been discussed in [5]. Decomposition approach has been used to compute delay in a connection-oriented packet-switched networks in [4]. ATM network was decomposed and its basic servers were analyzed in [17]. A feasible region method of allocating resources for FDDI-ATM-FDDI network has been proposed in [3].

3 ATM Based Heterogeneous Network (ABHN)

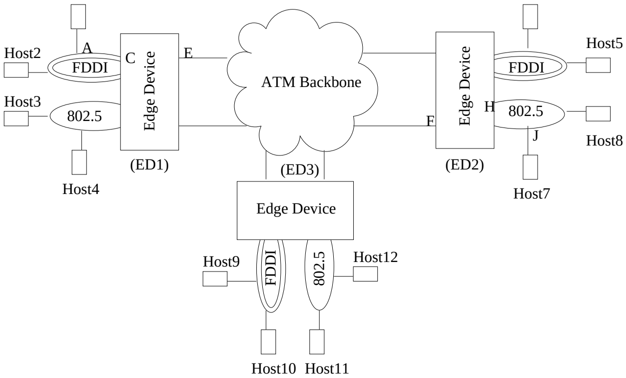

Figure 1: An exemplary ATM-based Heterogeneous Network

In this section we will give a brief description of ATM-based Heterogeneous Network (ABHN) and its various components. An ABHN is a high performance network with high bandwidth capability at the backbone. ABHN has been accepted by many industries as a platform to migrate from router-based to switch-based heterogeneous network. Figure 1 shows an example of an ABHN network. For our analysis, we will only use legacy LANs of FDDI and 802.5 token ring. There are four major components in an ABHN. 1) Hosts, 2) Legacy LANs, 3) Edge devices and 4) ATM backbone. Hosts are connected to legacy LANs such as ethernet, 802.5 token ring and FDDI. These legacy LANs are, in turn, connected to edge devices.

Edge devices act as interface between legacy LANs and the ATM backbone and does packet mapping from one medium to another. The edge device on the sending side of the connection receives packet from the connected legacy LAN. The packet is handled by the edge device in different ways, depending on where the destination LAN is located. The location of destination LAN falls into three cases:

1) It may be on the same LAN as the sender. A connection from Host1 to Host2 in Figure 1 falls into this case. The edge device acts as a Media Access Control (MAC) layer and handles the packet according to the MAC layer protocol of the destination LAN. For example, if the destination LAN is a token ring, then the edge device will buffer the packet until it receives the token. After receiving the token, it will transmit the packet.

2) It may be on a different LAN than the sender, but connected to the same edge device (e.g., connection from Host1 to Host3 in Figure 1). In this case, the edge device replaces the header of the packet with a new header that conforms to the destination LAN specification. The packet is then transmitted according to the MAC protocol of the destination LAN.

3) It may be on a different LAN and different edge device than the sender (e.g., connection from Host1 to Host7 in Figure 1). The connection, in this case, crosses the ATM backbone. After receiving a packet, the edge device on the sender's side (ED1) may apply LLC encapsulation [9] (if LANE interface is not used), switch it to appropriate port and then segment the packet to ATM cells which are then transmitted to the ATM backbone. The edge device on the receiver's side (ED2) receives the cells from the ATM backbone and reassembles them. It then strips off the LLC encapsulation and transmits the packet according to the MAC protocol of the destination LAN.

For a detailed description of the working of an edge device please refer to [2].

IEEE 802.5 token ring is a ring based network. A Media Access Control (MAC) protocol determines how real-time messages can be scheduled for transmission in a ring network. A MAC protocol basically determines the *access arbitration* and *capacity control* policy for a network. Access arbitration policy dictates when a node can access the medium, whereas capacity control policy decides how long a node can access the medium. In a IEEE 802.5 network, a special bit pattern called token moves around the ring. Node possessing the token can only transmit packet onto the ring. The IEEE 802.5 specification follows a *Priority Driven Protocol (PDP)* based arbitration policy, in which messages are assigned priority and a node with highest priority message can only possess the token. A message is divided into fixed size frames. Capacity control is achieved by a token holding timer. When a node possesses the token, it can transmit frames until its token holding timer expires, after which, the node has to relinquish the token and transmit it onto the ring. For more details about IEEE 802.5 specification please refer to [10].

We will not discuss details of operation of FDDI network for real-time application due to space limitation, interested readers may refer to [11].

The backbone of the ABHN is an ATM network. An ATM network is a set of ATM switches interconnected by physical links. These links are very high speed links and in most cases are made of fiber. Data are sent over in units of fixed size packets called cells. Cells destined for the same link are switched and multiplexed in the switch. Hence a link can carry cells pertaining to different connections. For our analysis we assume that the ATM switches follow a FIFO multiplexing scheme. FIFO multiplexing has been implemented in most of the commercial ATM switches. For more detailed description on ATM please refer to [8].

4 Partition-Based Connection Admission Control

For heterogeneous networks, we had proposed feasible region based CAC in our study in [3]. In this study, we are proposing a partition-based CAC which is much simpler and can scale well to the new technologies that may be added to future heterogeneous networks. In this section, we will explain the operational details of this method.

4.1 Basic Steps

Based on deadline partition method, our CAC follows four basic steps:

1. *Partitioning the deadline:* The original deadline of a connection is partitioned into several subdeadlines, one for each LAN segment in the connection. Some heuristic method is used to partition the deadline. A LAN segment is then analyzed as though it has to meet its assigned subdeadline. Thus, the beauty of this method is that it allows a LAN segment to be analyzed in an isolated manner, so that existing analysis for the LAN can still be used. This is much simpler than

analyzing the complex heterogeneous network as a whole, with its original deadline.

2. *Allocating resources:* After the original deadline is partitioned into sub-deadlines, one for each LAN segment in the path of the connection, resources in each LAN segment are allocated by taking its assigned sub-deadline into account. So any resource allocation method available for the LAN segment can be used without any modification except that the assigned sub-deadline has to be used in place of the original deadline. This is a big advantage of the partition-based method in terms of scalability, because no modification is necessary to the basic method of resource allocation when a new LAN technology is introduced to a heterogeneous network.
3. *Computing worst case end-to-end delay:* For hard real-time system, worst case end-to-end delay of a connection has to be found out to provide guarantee to the connection. Using the resource allocated in step 2, the worst case end-to-end delay of the new and the existing connections are calculated. Notice that the delays of the existing connections are calculated again, because the new connection might change the worst case delay of some existing connections if it shares network resources with any of them.
4. *Testing admissibility:* With all the above analysis at hand, the final step is to test if the new connection satisfies the conditions for admissibility. The new connection has to satisfy two conditions: 1) The worst case end-to-end delay of the new connection should be less than its deadline, 2) The worst case end-to-end delay of all the existing connections should also be less than their respective deadlines. The new connection is rejected if any of these conditions is violated.

Note that the basic steps used in partition-based CAC are quite different from that of feasible region method. In feasible region method, the first step is to compute a region for feasible resource allocation for the current state of the network. There is only one condition that needs to be satisfied for a new connection to be admitted: the feasible region should not be empty. Hence the second step is to test if the feasible region is not empty. If the connection is admitted (feasible region was not empty), then the third step is to allocate proper amount of resources from the available resources.

4.2 Partition Methods

In this section, we will formally present the three different deadline partition methods we have used for ABHN. Let n be the number of legacy LAN segments in the new connection $M_{l,m}$ (from Host l to Host m). Let these LAN segments be identified by $LAN_1, LAN_2, \ldots, LAN_n$. We define *partition factor* γ_i, $0 \le \gamma_i \le 1$, for LAN_i as the fraction of the total deadline ($D_{l,m}$) assigned to it i.e.

$$D^i_{l,m} = \gamma_i \; D_{l,m} \tag{1}$$

where $D^i_{l,m}$ is the sub-deadline assigned to LAN_i.

But the sum of all the sub-deadlines should be equal to the total deadline, i.e.

$$\sum_{i=1}^{n} D^i_{l,m} = D_{l,m} \; . \tag{2}$$

$$\text{Hence, } \sum_{i=1}^{n} \gamma_i = 1 \; . \tag{3}$$

Depending on the partition method used, γ_i for LAN_i will be determined in different ways. Then the sub-deadline $D^i_{l,m}$ is calculated using (1). This new sub-deadline is used while allocating resource from LAN_i. We have devised three different partition methods that can be applied to heterogeneous networks.

- *Equal partition method:* In this method, the original deadline of the new connection is partitioned equally among the legacy LAN segments. So the partition factor of LAN_i is given by

  $$\gamma_i = \frac{1}{n} \; . \tag{4}$$

  This is a very simple partition method and it does not require any state information of the network (e.g., load on the LAN segments). This method may work well in a symmetric ABHN, where both legacy LAN segments are of the same capacity and are more or less evenly loaded. But it may perform poorly in an asymmetric ABHN, because assigning equal sub-deadlines to both high and low capacity LANs may result in unbalanced resource allocation and hence may not satisfy the issues mentioned about resource allocation in heterogeneous network.
- *Utilization-based partition:* In this method, the original deadline is divided among the LAN segments such that the sub-deadlines are proportional to the utilization of the respective LANs. Thus, if a LAN has a higher utilization than the other, then it will have a larger sub-deadline than the other. This method addresses the bandwidth balancing problem by allocating resources in a proportional manner. But this method is more complex than equal partition method, because it has to keep track of the runtime utilization of each LAN segment as connections are admitted and released. So it is a dynamic allocation method which takes current load on the LAN into consideration while allocating resources. Hence, it may

perform better than the equal partition method when the LAN segments are not evenly loaded.

Let $u_1, u_2, \ldots, u_n$ be the utilization of the LANs $LAN_1, LAN_2, \ldots, LAN_n$ respectively. Then the partition factor for LAN_i is given by

$$\gamma_i = \frac{u_i}{\sum_{j=1}^{n} u_j} . \quad (5)$$

- *Available bandwidth based partition:* In this method, the deadline is divided among the LAN segments so that the sub-deadlines are inversely proportional to the available bandwidth of the the respective LANs. Thus a LAN with higher available bandwidth will be given lower sub-deadline than a LAN with lower available bandwidth. So this method also addresses bandwidth balancing problem and hence may perform better than the equal partition method, particularly when the LAN segments are not equally loaded. Like the utilization-based partition method, this method is a dynamic allocation method and has to keep track of available bandwidth of all the LAN segments.

  Let $AB_1, AB_2, \ldots, AB_n$ be the available bandwidth of the LANs $LAN_1, LAN_2, \ldots, LAN_n$ respectively. Then the partition factor for LAN_i is given by

$$\gamma_i = \frac{\sum_{j=1}^{n}(AB_j) - AB_i}{(n-1)\sum_{j=1}^{n} AB_j}, \qquad n > 1 . \quad (6)$$

  When assigning sub-deadline for a LAN segment, all other LAN segments are thought of as a single LAN segment with available bandwidth equal to the sum of their individual available bandwidths. Then the deadline is divided in an inversely proportional manner. Hence is the equation (6).

  Note that the above partition formulas work for a network in which a message may travel through an arbitrary number of LAN segments (i.e., n). For the networks we are of concern in this study (as shown in Figure 1), $n = 2$. In the rest of this paper, we assume that this holds.

4.3 Resource Allocation

Let us assume that the deadline (D) of a connection has been partitioned into two parts:

$$D = D_S + D_R \quad (7)$$

where D_S and D_R are sub-deadlines for the sender and receiver LAN segments, respectively. The partition may be done by any of the methods mentioned in section 4.2.

We now discuss how the resources in the LAN segments should be allocated. Let $R_{l,m}$ be the bandwidth requirement of the new connection $M_{l,m}$ (from Host l to Host m) being considered for admission. For an FDDI LAN, the resource is represented by synchronous bandwidth. Let H_S be the synchronous bandwidth allocated to the connection if its sender LAN is an FDDI, with target token rotation time of $TTRT_S$. We propose

$$H_S = \frac{D_S \cdot R_{l,m}}{\lfloor \frac{D_S}{TTRT_S} - 1 \rfloor} . \quad (8)$$

Similarly, let H_R be the synchronous bandwidth allocated to the connection if its receiver LAN is an FDDI, with target token rotation time of $TTRT_R$. Then

$$H_R = \frac{D_R \cdot R_{l,m}}{\lfloor \frac{D_R}{TTRT_R} - 1 \rfloor} . \quad (9)$$

This scheme of synchronous bandwidth allocation can be derived intuitively from the flow conservation principle. Between the arrival of a message and the expiration of its deadline (D_S time units later), a node will have at least $\lfloor \frac{D_S}{TTRT_S} - 1 \rfloor H_S$ synchronous bandwidth available [1]. Also, during D_S time units, $(D_S \cdot R_{l,m})$ can loosely be regarded as load on the node. Thus the allocated synchronous bandwidth is just sufficient to handle the load. Hence (8) and (9) follow.

For an 802.5 token ring LAN segment, no explicit resource needs to be allocated. But a priority is assigned to each message. The 802.5 MAC protocol guarantees a non-preemptive priority-driven policy. Thus, a problem equivalent to resource allocation is priority assignment. We propose to use the earliest deadline first (EDF) to assign priority. This assignment scheme has been extensively studied [13, 20]. The only modification we have to make is to use the sub-deadline obtained from the partition process, rather than the original deadline.

Other LAN segments can be dealt with in a similar manner. The reader now may realize one of the advantages of our partition-based CAC method: We do not have to re-invent resource allocation methods for individual LAN segments. All the existing methods originally developed for real-time communication over a particular local area network can still be used in our method, only the deadline has to be substituted by the sub-deadline obtained from the partition.

4.4 Delay Computation

In this section we will briefly describe how the worst case end-to-end delay can be calculated for an FDDI-ATM-TokenRing heterogeneous network. For an ABHN consisting of a variety of network components, delay analysis is not an easy task. We have used decomposition method to analyze an ABHN connection. In this method, a connection is decomposed into a sequence of servers each of which is analyzed separately. A complex server may further be decomposed into simpler servers to make the analysis tractable. Each server is provided with a description of input traffic and the service discipline used to calculate worst case delays at the server. Output from the server is also calculated so that it can be used as input traffic for the next server. Finally, the worst case delay of the connection is obtained by summing up all the individual delays at each server.

Although the method looks simple, its effectiveness depends on the proper modeling of the server and description of traffic. It is very important that the impact of the server on connection and its service discipline be well understood. The input traffic description should accurately represent the actual traffic of the connection at every point in the network and it should be simple and efficient. So we have used *the maximum rate function* $\Gamma(I)$ as traffic descriptor, which is defined as the maximum arrival rate in any interval of length I (in bits/second). $\Gamma(I)$ has been successfully used as traffic descriptor in ATM networks [17] and in FDDI-ATM-FDDI networks [3], for real-time applications.

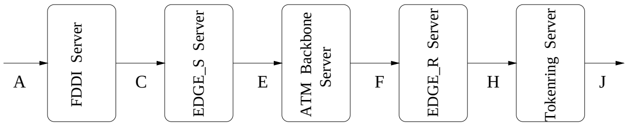

Figure 2: The decomposed servers on a connection path

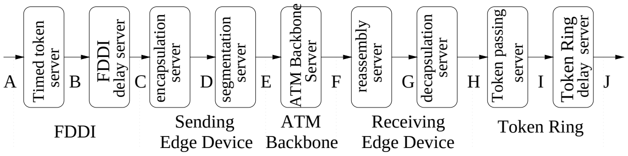

Figure 3: Second level decomposition of servers on a connection path

Now to illustrate how the decomposition method can be applied to our ABHN, consider a connection between Host1 on an FDDI LAN and Host7 on a 802.5 LAN in Figure 1 passing through points A, C, E, F, H, and J. So it encounters a FDDI server, a EDGE_S server, the ATM backbone, a EDGE_R server and a TokenRing server in that order (Figure 2). But this one level decomposition does not result in a structure simple enough for analysis. So some of the servers are further decompsed as shown in Figure 3.

The FDDI server can be further decomposed to a *timed token server* and a *FDDI delay server*. The timed token server is the one that is responsible for getting access to the FDDI ring. It serves FDDI frames according to the timed token protocol of FDDI. The propagation delay of a frame on the FDDI segment is represented as the constant FDDI delay server. Similarly, other servers are further decomposed as shown in Figure 3. For more details on the second level of decomposition please refer to [18].

With the above decomposition, the message delay across this connection can be written as

$$d_{messg} = d_{fddi} + d_{edge\_s} + d_{atm} + d_{edge\_r} + d_{token} \quad (10)$$

where $d_{fddi}, d_{edge\_s}, d_{atm}, d_{edge\_r}, d_{token}$ are the delays encountered in the compound servers FDDI, EDGE_S , ATM backbone, EDGE_R and token ring server respectively. These delays can further be decomposed according to the second level of decomposition to correspond to delays of each server shown in Figure 3.

If d^{wc}_{fddi}, $d^{wc}_{edge\_s}$, d^{wc}_{atm}, $d^{wc}_{edge\_r}$, d^{wc}_{token} are the worst case values of d_{fddi}, d_{edge_s}, d_{atm}, d_{edge_r}, d_{token} respectively, then the worst case end-to-end delay suffered by a message in this connections is given by

$$d^{wc}_{messg} = d^{wc}_{fddi} + d^{wc}_{edge\_s} + d^{wc}_{atm} + d^{wc}_{edge\_r} + d^{wc}_{token}. \quad (11)$$

Notice that we are analyzing delay of an FDDI-ATM-TokenRing connection. Delays of other types of connections can be analyzed in a similar manner.

Due to space limitation we will not go into details of how the worst case delays of (11) at each server are calculated. The details can be found in [18].

4.5 Admissibility Test

Now we will formally present the admissibility test used in our partition-based CAC. Let R_S and R_R be the resources allocated to the sender and the receiver LAN respectively using methods described in section 4.3. Note that these resources can be the synchronous bandwidth for FDDI LAN or assigned priority for token ring LAN. Let $M_{l,m}$ be the new connection with a deadline of $D_{l,m}$. Let $\mathcal{M}$ be the set of identifiers of the connections which are currently active in the network i.e.

$$\mathcal{M} \;=\; \{(p,q) \mid M_{p,q} \text{ is currently active}\}. \quad (12)$$

Let $M_{p,q}$ be an existing connection i.e. $(p,q) \in \mathcal{M}$, with deadline $D_{p,q}$. Let $d^{wc}_{p,q}(R_S, R_R)$ be the worst case end-to-end delay of $M_{p,q}$ when resources R_S and R_R are allocated to $M_{l,m}$ and $d^{wc}_{l,m}(R_S, R_R)$ be the corresponding delay of connection $M_{l,m}$. These delays can be found by appropriately using (11). Connection $M_{l,m}$ is admitted only if all the existing as well as the new connection meet their deadlines, i.e. if both of the following conditions are satisfied:

1. For every $(p,q) \in \mathcal{M}$, $d^{wc}_{p,q}(R_S, R_R) \;\leq\; D_{p,q}$ (13)
2. For $M_{l,m}$, $d^{wc}_{l,m}(R_S, R_R) \;\leq\; D_{l,m}$ (14)

5 Performance Evaluation

5.1 Simulation Model and Performance Metric

To obtain performance data, we have simulated an ABHN similar to the one shown in Figure 1. For the network simulated, there are three edge devices, each of which is connected to four 802.5 token rings and one FDDI ring. The edge devices are connected to each other via an ATM switch. The capacity of each FDDI LAN is 100 Mbps and that of each token ring is 16 Mbps. The ATM links are of 155 Mbps. Connection requests have a Poisson arrival with a rate λ. All admitted connections have a life time which is exponentially distributed with mean $\frac{1}{\mu}$. The source traffic of a connection is periodic with length C, period P, and deadline D. We have used three configurations to carry out our experiment:

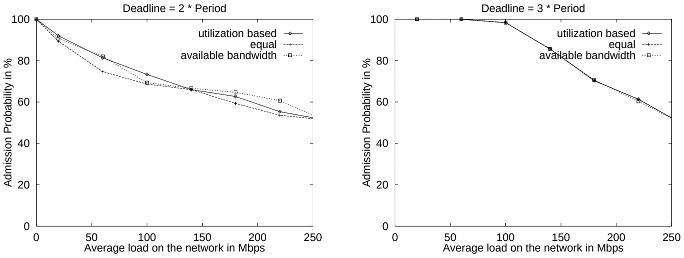

Figure 4: Performance data for Configuration 1

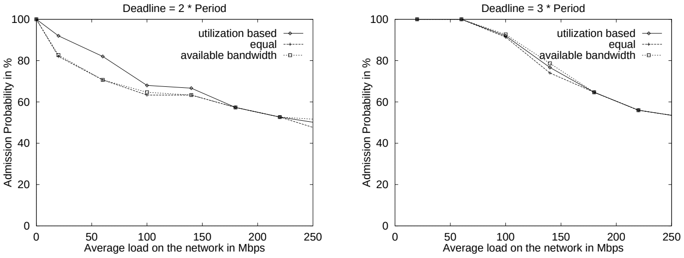

Figure 5: Performance data for Configuration 2

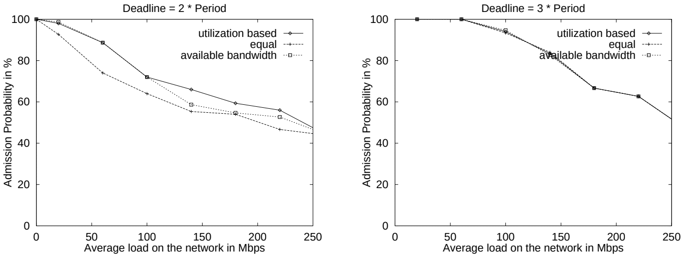

Figure 6: Performance data for Configuration 3

1) *Configuration 1:* In this configuration, the token rings do not participate. So this is an FDDI-ATM-FDDI network. Each of the three FDDIs has the same probability of being chosen as the sender or receiver.
2) *Configuration 2:* The hardware configuration is the same as in Configuration 1. However, while choosing the receiver, one of the FDDI ring is made a *hot spot*. That is, it has a higher probability of being chosen as receiver. This configuration is used to simulate a scenario where an FDDI may be overloaded. For example, a server is connected to an FDDI and many clients are trying to access data from the server.
3) *Configuration 3:* In this configuration, both FDDI rings and token rings participate. Similar to Configuration 2, an FDDI is chosen to be the hot spot on the receiver's site.

We used *admission probability* (AP) as the performance metric for our experiment. It is defined as the ratio of total number of admitted connections to total number of connections requested. Average load on the network is calculated as the average number of connections in the network multiplied by the average bandwidth requirement of a connection (in Mbps).

5.2 Performance Data and Observation

Figures 4, 5 and 6 show the performance curves for the three configurations mentioned above. For Configuration 1, since all the FDDIs have equal probability of being chosen, all the FDDIs will be more or less equally loaded. So the partition factors found by the three methods are almost equal. Hence the difference between APs for the three methods are very small (Figure 4) when deadline is twice the period. This difference is almost indistinguishable when a higher deadline of three times the period is used. Also, the APs are higher than the previous case. This is as expected, since higher deadline will allow more connections to be admitted and the small difference in the partition factors of the three methods are offset by a larger deadline.

For Configuration 2 (refer to Figure 5), one of the FDDIs is chosen as hot spot, so the FDDI segments are not equally loaded. So utilization-based partition performs better than the others, because this method takes uneven utilization (load) of the LANs into account while partitioning deadline. As in case of Configuration 1, the performance of the three methods become very close to each other when the deadline increases to three times the period.

In case of Configuration 3 (refer to Figure 6), both the utilization-based and available bandwidth based methods outperform equal partition methods when the deadline is tight (twice the period). This is so because, both of these methods consider load of the LAN segments while allocating resources. Whereas equal partition method considers all the LAN segments as equally loaded. As the deadline becomes loose (three times the period), performance of all the methods seems to close in, for the reason stated before.

6 Conclusion

We have proposed a deadline partition-based connection admission control mechanism for hard real-time connection in heterogeneous networks. This method, though not optimal, is a much simpler method compared to the feasible region method proposed in [3]. It is also scalable to any new technology that may be introduced to existing heterogeneous network. We have proposed three different methods for partitioning the deadline and shown their performance in ABHN. The performance of partition-based connection admission control is also satisfactory in terms of admission probability.

From our simulation results, we have found that when the deadline is tight, it is better to use utilization or available bandwidth based partition method for allocating resources. If the deadline is not tight, the network may perform equally well with equal partition method. But if there are hot spots in the network, then the performance may be better with utilization or available bandwidth based partition method.

This work involved guaranteeing deadline of connections at the MAC layer. We will extend this project to provide guarantee at the application layer for hard real-time embedded systems. We have envisioned to do it by incorporating our CAC for heterogeneous network into our NetEx software suite [6]. We plan to evaluate the performance of our CAC in a real world ATM-based heterogeneous network in the test bed of our Distributed Computing Laboratory.

References

[1] Agrawal G., Chen B., Zhao W., Davari S., "Guaranteeing Synchronous Message Deadlines in High Speed Token Ring Networks with Timed Token Protocol" - *Proc. of the ICDCS*, pages 468-475, June 1992.

[2] Baseline Text for MPOA, ATMF/95-0824r7 - *Draft from ATM Forum Technical Committee*, May 1996.

[3] Chen B., Sahoo A., Zhao W., Raha A., "Connection-Oriented Communications for Real-Time Applications in FDDI-ATM-FDDI Heterogeneous Networks" - *Proc. of ICDCS*, May 1997.

[4] Cruz R. L., "A Calculus for Network Delay, Part I: Network Elements in Isolation" - *IEEE Transactions on Information Theory*, 37(1):114–131, January 1991.

[5] Cruz R. L., "A Calculus for Network Delay, Part II: Network Elements in Isolation" - *IEEE Transactions on Information Theory*, 37(1):132–141, January 1991.

[6] Devalla B., Li C., Sahoo A., Zhao W., "Connection-Oriented Real-Time Communication for Mission Critical Applications - An Introduction to NetEx A Portable and Efficient Tool Kit" - *Proc. of NAECON*, June 1997.

[7] Ferrari D., and Verma D. C., "A Scheme for Real-Time Channel Establishment in Wide-Area Networks" - *IEEE Journal on Selected Areas in Communications*, SAC-8(3):368–379, April 1990.

[8] Ginsburg D., "ATM: Solutions for Enterprise Internetworking" - *Addison Wesley*, 1996.

[9] Heinanen J., "Multiprotocol Encapsulation over ATM Adaption Layer 5" - *Network Working Group Request for Comments 1483*, July 1993. Technical Committee.

[10] IEEE standard 802.5-1989, "Token Ring Access Method and Physical Layer Specifications" - *The Institute of Electrical and Electronic Engineers, Inc.*, 1989.

[11] Jain R., "FDDI Handbook: High-Speed Networking Using Fiber and Other Media" - *Addison Wesley*, 1994.

[12] Kurose J.F., Schwartz M., Yemini Y., "Controlling Window Protocols for Time-Constrained Communication in a Multiple Access Environment" - *Proc. of the 8th IEEE International Data Communication Symposium*, 1983.

[13] Lim C.C., Yao L., Zhao W., "A Comparative Study of Three Token Ring Protocols for Real-Time Communications" - *Proc. of ICDCS*, pages 308-317, May 1991.

[14] Malcolm N., Zhao W., "Use of The Timed-Token Protocol for Real-Time Communications" - *IEEE Computer*, 27(1):35–41, January 1994.

[15] Mehra A., Indiresan A., and Shin K.G., "Resource Management for RealTime Communications : Making Theory Meet Practice" - *Proceeding of RTAS '96, 1996.*

[16] Mukherjee S., Saha D., Saksena M. C., and Tripathi S. K., "A bandwidth allocation scheme for time constrained message transmission on a slotted ring LAN" - *Proc. of the IEEE RTSS*, pages 44–53, December 1993.

[17] Raha A., Kamat S., and Zhao W., "Guaranteeing End-to-End Deadline in ATM Networks" - *Proc. of ICDCS*, pages 60-68, May 1995.

[18] Sahoo A., Zhao W., "Partition-Based Connection Admission Control for ATM-Based Heterogeneous Network" - *Technical Report, Department of Computer Science, Texas A&M University*, August 1997.

[19] Strosnider J. K., "Highly Responsive Real-Time Token Rings" - *PhD thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University*, 1988.

[20] Yao L., Zhao W., "Performance of an Extended IEEE 802.5 Protocol in Hard Real-Time Systems" - *Proc. of the IEEE INFOCOM '91*, pages 469-478, April 1991.