Time Dependent Priority Scheduling for Guaranteed QOS Systems

S. Chaudhry

Dept of Elec & Comp Engr
Syracuse University
Syracuse, NY 13210

Alok Choudhary

Dept of Elec & Comp Engr
NorthW estern University
Evanston, IL 60208

Abstract

With the advances in server technology, and the advent of fast Gigabit networks, it has become possible to support multi-media applications. To support the requirements for the transmission of isochronous data, the network must provide service guarantees to connections, including minimum bandwidth, packet delay, delay jitter, and loss.

Three factors determine the utilization of the network when providing these services. These are the scheduling algorithm *employed at each switch; the accuracy (tightness) of the* admission control (schedulability condition) *that detects violations to the service guarantees; accuracy of the* input traffic characterization*. In this paper we present a scheduling algorithm, its schedulability condition and implementation. The schedulability condition is free of input traffic characterization and thus any input traffic model can be used. Further, the algorithm is capable of achieving up to maximum efficiency possible at each switch.*

1 Introduction

Emerging Broadband Integrated Services Digital Networks (B-ISDN) must support applications with diverse traffic characteristics and performance requirements. Of the many classes in future B-ISDN networks, variable bit rate multimedia data poses a unique challenge. Multimedia data is isochronous in nature. To convey the information held in a media stream, the media units of the stream must be presented in a time continuum. Thus, this type of data is delay sensitive and the network must support a resource reservation scheme through admission control to allocate network resources to each multimedia stream. However, the burstiness of VBR video traffic makes it difficult to determine the amount of resources required. On the one hand, if resources are reserved at the average rate of the VBR video source, unacceptable delays may result if the source is transmitting at its peak rate. On the other hand, if resources are reserved at peak rate, the network may be under-utilized most of the time.

The initial proposals for establishing guaranteed bandwidth channels by Ferrari et. al. [3, 14] used peak rate to reserve resources. This resulted in low utilization of network resources by VBR streams. In order to increase utilization, Ferrari et. al. [8] investigated admissibility for various scheduling algorithms. Their investigations revealed the the schedulability region of EDF to be significantly larger than those of Static Priority and FIFO. These results were confirmed in [7] which also determined the most important criteria to be traffic characterization model, scheduling algorithm and its associated admission control test.

A traffic model of all the sources receiving the service is required for guaranteed network service. Such a model must be a worst-case characterization of the source to provide an absolute upper bound on a source's packet arrivals. Further, the model must be parameterized so that a source can efficiently specify its traffic characterization to the network. The model should also characterize the traffic as accurately as possible so that the admission control algorithms do not over-estimate the resources required by the connection. Finally, the model must allow policing to enable the network to enforce a source's traffic characterization.

Finding an appropriate traffic characterization that captures the dynamics of the source is a key challenge of incorporating VBR video traffic into networks with guaranteed service. A rich set of literature exists on characterizing VBR video traffic by stochastic processes such as the Markov-Modulated Poisson Process. Other approaches include Auto Regressive Moving Average models, histogram-based modules, and self-similar models. While stochastic models may be used to achieve higher network utilization via statistical multiplexing, they make it difficult to implement a policing mechanism that enforces a stochastic traffic characterization. Moreover, most stochastic models for characterizing video source are too complex for practical implementation.

For these reasons, we base all calculation in terms of a worst case representation. A worst-case representation of a traffic source may be described as follows. If the actual traffic of a connection is given by a function A such that $A[t, t + r]$ denotes the traffic arrivals in time interval $[t, t + r]$, and an upper bound on A can be given by a function $\hat{A}$ if for all times $t \geq 0$ and all $r \geq 0$, the following holds:

$$A[t, t + r] \leq \hat{A}(r) \qquad (1)$$

We refer to a function $\hat{A}(r)$ that satisfies the property in equation 1 as a *traffic constraint function*. We

use $A[t, t+r]$ and $\hat{A}(r)$ in the calculation of the schedulability condition and any parameterized deterministic traffic characterization and constraint function can be directly plugged into the results. For models that bound the tail distribution of arrivals, the results can be used to guarantee statistical bounds.

The other criterion, the scheduling algorithm and its associated admission control test, is important in increasing utilization because VBR traffic has deadlines associated with each packet. Traditional schemes such as FIFO and SP that do not base service schedules on impending deadlines of packets result in poor utilization. Scheduling with real-time constraints requires dynamic algorithms to increase utilization. However, algorithms in this class like EDF and Virtual Clock have considerable scheduling overheads and suffer from flaws in overloaded conditions. We propose a time varying priority scheme which is a generalization of EDF but is able to prioritize not only on the basis of deadline but also on the basis of the loss parameter in QOS guarantees. We also provide a low-cost implementation and admission control tests for our scheduling algorithm.

2 Scheduling Algorithms

In a connection oriented packet switched network, packets from a particular connection traverse the network on a fixed path of switches and links. Each switch has a packet scheduler for each outgoing link. Since only one packet can be transmitted at any time the scheduler maintains a queue containing all packets waiting for transmission. The scheduling discipline for this queue can be divide into two categories: static or dynamic. Here we consider the pros and cons of algorithms in both classes.

The best known static algorithms are First-In-First-Out(FIFO) and Static Priority(SP).

FIFO: FIFO schedulers transmit all packets in order of their arrival. Since the maximum delay in a FIFO scheduler scheduler is the same for all streams $j \in \mathcal{N}$, all streams must have identical delay bound. It is easy to implement with very low overhead and therefore attractive.

SP: In an SP scheduler, each stream $j \in \mathcal{N}$ is assigned a priority p with $1 \leq p \leq P$, where a lower priority index indicates a higher priority. All connections with priority p have the same delay bound d_p, with $d_p < d_q$ if $p < q$. SP maintains one FIFO queue for each priority level, always selecting the first packet in the highest priority FIFO queue for transmission. This offers a choice of different deadlines and is implementable with very low overhead. However, lower priorities may experience starvation and therefore the scheduling conditions are strict in order to avoid such a scenario. This can result in low utilization of the network.

The drawbacks inherent in static schemes have led many to consider dynamic scheduling algorithms for real-time scheduling purposes, such as Earliest Deadline First (EDF) and Virtual Clock.

EDF: With EDF scheduling, each connection $j \in \mathcal{N}$ is assigned a delay bound d_j, where the delay bound maybe different for each connection. An EDF scheduler selects packets for transmission in increasing order of packet deadlines, where the deadlines are calculated as the sum of arrival time and delay bound of the packet. EDF is relatively difficult to implement and entails significant overhead. However, the algorithm posses some very desirable properties, namely:

Theorem 1 *Any sequence that at any instance schedules the packet with earliest due date among eligible packets, is optimal with respect to minimizing maximum lateness [5].*

Theorem 2 *If there exists a schedule that will satisfy deadlines of all packets, then at least one schedule is an EDF schedule [9].*

EDF is optimal with respect to Lateness, but Locke's [11] experiments have shown that the algorithm performs very poorly in overload conditions. This is because it gives highest priority to packets that are close to missing their deadlines.

A typical phenomenon that may happen when the scheduler is overloaded is the *domino effect*, since the first packet that misses its deadline may cause all subsequent packets to miss their deadlines. In such a situation, EDF does not provide any type of guarantee on which packets will meet their deadlines. This is very undesirable behavior since either the admission control has to ensure against this by ignoring the loss parameter leading to under utilization or risk all packets missing their deadlines. A number of heuristic EDF algorithms have been proposed to deal with EDF overloads [13] [4].

Virtual Clock: The Virtual Clock scheduling algorithm assigns each packet a value upon its arrival and transmits packets in increasing order of the value assigned to the packets. Let p_j^i stand for the ith packet of connection j, $A(p_j^i)$ stand for the arrival time of the ith packet of connection j, $V(p_j^i)$ denote the value assigned to the ith packet of connection j, and l_j^i denote the length of the ith packet of connection j. If connection j is assigned a rate r_j then the value assigned to packet i of connection j is given by

$$V(p_j^0) = 0 \qquad (2)$$

$$V(p_j^i) = max A(p_j^i), V(p_j^{i-1}) + \frac{l_j^i}{r_j} \quad i \geq 0 \qquad (3)$$

If $\sum_{j \in \mathcal{N}} r_j$ is less than the capacity of the link then such a scheme guarantees bandwidth to a particular connection, but offers no strict deadline for each packet at a switch. However, such a scheme has been shown to offer end to end delay deadline by effecting a trade-off between bandwidth and buffer space. The ratio of this trade-off is inflexible, and can not be altered due to the nature of the algorithm. This disadvantage becomes apparent in situations where buffering is limited and bandwidth is available or vice-versa. Moreover, it is at least as expensive to implement as EDF.
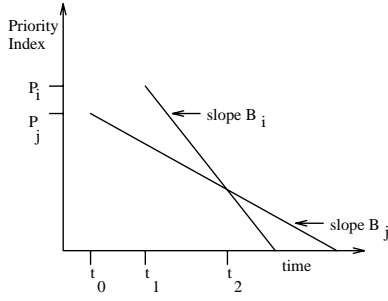
Figure 1: Interaction of priority functions between two connections

3 Time Dependent Priorities

The discussion of EDF has shown that a priority structure is needed to give preferential treatment to the "higher priority" groups at the expense of "lower priority groups". SP is optimal in that it minimizes weighted mean delay (higher priorities have more weight) but only in the absence of time constraints. Since low priority queues can easily experience starvation, admission control in such cases would have to guard against this possibility by over-reserving network resources, leading once again to low utilization.

We thus propose a Time-Dependent Priority (TDP) scheme similar to the scheme proposed in [6]. It will be shown that this scheme covers the spectrum from that discipline which separates priority groups to the greatest possible extent (SP) to the discipline that does not separate them at all (FIFO). This spectrum includes EDF.

In the scheme, a packet from connection j is given a priority index P_j upon arrival. At any time t after arrival and before service, the priority index of the packet is given by $P_j - (t - arrival\ time) * B_j$, where B_j is the rate of decrease of priority index. This is depicted in figure 1. No preemption is allowed and whenever the service facility (link) is free, the packet with the lowest priority index is chosen for transmission. Whenever a tie for highest priority (lowest priority index) occurs the tie is broken by servicing packets with a higher rate of increase of priority first, and in case we still have a tie, the tie is broken by the FCFS rule. We note that connections with a higher slope (B_j) can be labeled as "higher priority" connections while connections with lower initial priority index (P_j) can be labeled as "higher priority" connections.

Figure 1 also shows an example of the manner in which this priority structure allows interaction between priority functions of two connections. Specifically, the first packet (from connection j) will be chosen in preference to the second packet (from connection i) if the link is free between t_0 and t_2; but any time after t_2 the second packet will be chosen in preference.

It should be noted that there can be at most only one interchange between two packets and this property simplifies the analysis. Also, the slopes are linear and no generalization is achieved if the dependence of the

priority index with time was the rth power instead.

It can easily be seen that TDP covers the spectrum of service disciplines. Specifically, if $B_j = 0$ $\forall j$, we have service order governed by SP. And if $P_i = P_j$, $B_i = B_j$ $\forall i, j$, we have a FIFO schedule. Lastly, if $B_i = B_j$ $\forall i, j$, the service discipline is the same as EDF.

4 Mathematical Properties of TDP

Due to space considerations, we provide only the statements of the Theorems here. The complete proofs and details are given in [1].

Consider a switch through which a set connections $\mathcal{N} = \{1, 2, \ldots, n\}$ have an initial priority index P and the rate of increase of priority of a connection j is B_j. Without loss of generality let $B_i \leq B_j$ iff $i < j$. The associated deadline of a packet of connection k is given by $P/B_k \equiv d_k$.

Theorem 3 *A connection k will meet its deadline if*

1. $\forall a \quad such \quad that \quad B_a \quad < \quad B_k$
   $\sum_{j : B_j > B_a} \hat{A}_j ((\frac{B_j - B_a}{B_j})t) \leq t \ for \ 0 < t < d_a$

2. *and* $\sum_{j \in \mathcal{N}} \hat{A}_j (t - d_j) \leq t \ \forall \ t \geq d_a$.

Now for any connection j let its initial priority index be P_j.

Theorem 4 *A connection k will meet its deadline if*

1. $\forall a \ st. \ B_a < B_k$
   $\sum_{j : B_j < B_a \ \wedge \ P_j < P_a} \hat{A}_j (t - d_j) +$
   $\sum_{j : P_j / B_a \leq t \ \wedge \ B_j \geq B_a} \hat{A}_j (t - d_j)$
   $+ \sum_{j : P_j / B_a > t \ \wedge \ B_j \geq B_a} \hat{A}_j (\frac{B_j - B_a}{B_j} t) \ for \ t < d_a$.

2. *and* $\sum_{j \in \mathcal{N}} \hat{A}_j (t - d_j) \leq t \ \forall \ t > d_n$ *otherwise.*

Corollary 5 *If*

1. $\sum_{j : B_j > B_a} \hat{A}_j ((\frac{B_j - B_a}{B_j})t) \leq t \ for \ t_1 < t < d_a$

2. *and* $\sum_{j \in \mathcal{N}} \hat{A}_j (t - d_j) \leq t \ \forall \ t \geq d_a$,

then the maximum delay experienced by a packet of connection k is $d_k + t_1$.

Corollary 6 *If $\forall a$ such that $B_a < B_k$*

1. $\sum_{j : B_j < B_a \ \wedge \ P_j < P_a} \hat{A}_j (t - d_j) +$
   $\sum_{j : P_j / B_a \leq t \ \wedge \ B_j \geq B_a} \hat{A}_j (t - d_j) +$
   $\sum_{j : P_j / B_a > t \ \wedge \ B_j \geq B_a} \hat{A}_j (\frac{B_j - B_a}{B_j} t) \ for \ t_1 \leq t < d_a$.

2. *and* $\sum_{j \in \mathcal{N}} \hat{A}_j (t - d_j) \leq t \ \forall \ t > d_n \ for \ t > d_a$.

then a packet of connection will experience delay of less than $d_k + t_1$.

Figure 2: Node Architecture
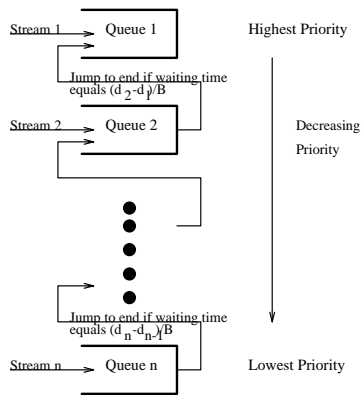

Figure 3: Low Delay Schedulable Region Same Initial Priority Index

Now, if stream k was provided an extra buffer space to hold data needed for transmission in t_1 time units, then condition 1 of Theorems 3 and 4 can be relaxed *i.e.* condition 1 can be checked from time t_1 onwards. On the other hand, if rate of increase of priority for packets of stream k is lowered such that its deadline increases by t_1, conditions after t_1 are relaxed (and to a smaller extent before t_1). Thus, the buffer space can be used to overcome short bursts or long term burstiness according to the expected input traffic characteristics.

5   Impleme ntation

We briefly describe a low cost implementation of TDP. Let there be a set of connections $\mathcal{N} = \{1, 2, \ldots, n\}$. Let a connection j be characterized by P_j, and B $\forall j \in \mathcal{N}$. The deadline of each connection j is then $d_j = P_j/B$. Without loss of generalit y let $P_{i-1} < P_i$ $1 < i \leq n$.

Now a consider a set of $\mathcal{N} = \{1, 2, \ldots, n\}$ queues which are served in static priority order with queue i having higher priority iff $i < j$. Also, if a packet has resided/waited in queue i for $\frac{P_i - P_{i-1}}{B}$ units of time then let it jump to the end of queue $i - 1$. This is depicted pictorially in figure 2.

The service order of the above implementation is the same as prescribed as TDP [10]. A formal proof is omitted but it is easy to see that when a packet jumps to the end of any queue i, its priority index is the same as that of a new arrival to queue i. This ensures that the packet is served before others with higher priority index and after those with lower priority index.

Now, this can be implemented using a timer for each queue and a timestamp mec hanism. Each packet is timestamped upon arriv al. Consider a packet that arrives in queue i at time t_1. When the pac ket moves to the head of the queue at time t_2, the timer is loaded with value equal to $t_1 + \frac{P_i - P_{i-1}}{B} - t_2$. The timer then counts down and upon reaching zero, the packet is moved to the end of queue $i - 1$. The timer is then reloaded with value *timestamp of the packet now at head of queue (arrival time of the packet) minus t_1*. Again the timer counts down to zero, and the process
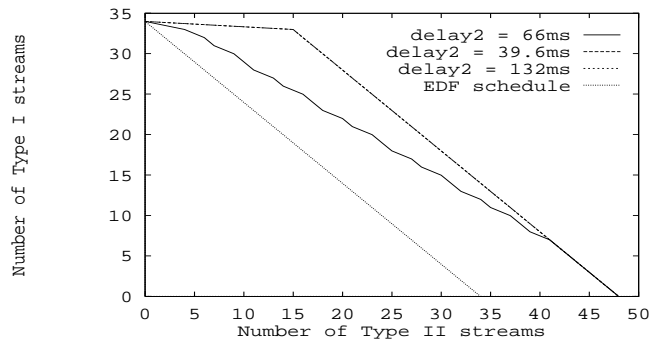
is repeated. If a packet is served before the timer expires, the timer is stopped and the value *timestamp of the current packet at the head of queue - timestamp of the packet being served* is added and the countdown is restarted.

However, if the network consists of switches that employ the BUP Protocol to implemen t guaranteed bandwidth channels as proposed in [2], the need for timestamps and timers can be remo ved. The time spent in a queue by a packet at the head of the queue can be obtained by the number of packets in the queue, since the streams are rate controlled. The details of the algorithms to implemen t TDP and the data structures are not provided here due to space limitations. The reader can refer to [1] for the deatils.

In more recent switch designs, custom silicon is being employed to reduce cost and increase capacity and functionality. Frequently, the queuing function is being implemented with custom silicon and static RAM permitting queues to be implemen ted as linked lists of cells. With this switc h implemen tation, the proposed queues are easily realisable [12].

Any algorithm for guaranteed service must decide before admission if all packets can be successfully scheduled. The admission control is responsible for this, but ideally the admission control (which is bound by the scheduling policy) should come close to optimal. An Optimal algorithm in a bounded case is described in [1]. By assigning a lower slope for higher permissible loss parameter (in verse of weight) specified in the QOS requirements of a stream and extending their deadlines TDP will heuristically delay those packets that would be dropped by the optimal algorithm. However, we need to devise another test to ensure that at most only a permissible proportion of the packets will be delayed beyond their desired deadline.

6   Results

To verify the schedulable region for different VBR traffic conditions and QOS requiremen ts, we took an MPEG encoded mo vie and recorded its frame sizes. Theorem 3, Corollary 5, Theorem 4, and Corol-
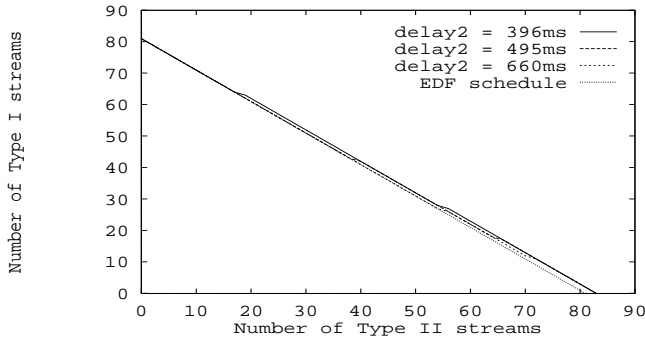
Figure 4: High Delay Schedulable Region Same Initial Priority Index

Figure 5: Low Delay Schedulable Region Different Initial Priority

lary 6 can be used along with an *envelope* $\hat{A}(33 * i)$ = *Maximum sum of contiguous i frames* to check if a configuration is schedulable.

For a stream with an allowable percent of packets that can be delayed beyond a required deadline, we increase its deadline beyond the desired deadline by giving such streams a slower rate of increase of priority. This delays packets from such streams beyond the required deadline when the link is temporarily congested. This is analogous to the optimal algorithm. It allows us to increase the number of streams admitted. However, another test needs to be administered to ensure that the percent of packets delayed beyond their desired deadline adheres to the parameters specified.

Consider two streams with identical service requirements except that one does not allow any packets to be delayed beyond their deadline and the other has 10 percent delayable beyond the deadline. Let us denote the streams as streams 1 and 2 respectively. Let the slope assigned to stream 1 be B_1 and to stream 2 be B_2, and their deadlines be d_1 and d_2 respectively. Let the priority index of a packet from stream 2 that has waited d_1 units of time, the desired deadline, be m ($m > 0$). Further, let us denote the length of time a packet from stream 1 has to wait to obtain a priority index of m as d_n.

Now if $\hat{A}_1(t - d_n) + \hat{A}_2(t - d_1) \leq t$ $\forall t$, then all packets will leave before they attain a priority index less than m. Also, since the *traffic constraint function (envelope)* is concave if $\hat{A}_1(t - d_n) + \hat{A}_2(t - d_1) \leq t$ $\forall t > t_1$ and $\hat{A}_2(t_1) \leq 0.9\hat{A}_2(t_{max})$, then in the worst case only 10 percent of the packets will be delayed such that their priority index is m. This in turn will lead to at most 10 percent of the packets being delayed beyond the desired deadline.

Lastly, if $\hat{A}_1(t - d_1) + \hat{A}_2(t - d_1) \leq t$ $\forall t$ and $\hat{A}_1(\frac{B_1 - B_2}{B_1}t) \not\leq t$ $\forall t$, this test can be ignored. All stream 1 packets will meet their deadline as the delay experienced by these packets is never lengthened by decreasing the rate of increase of priority of stream 2 packets, if initial priority indices are the same.

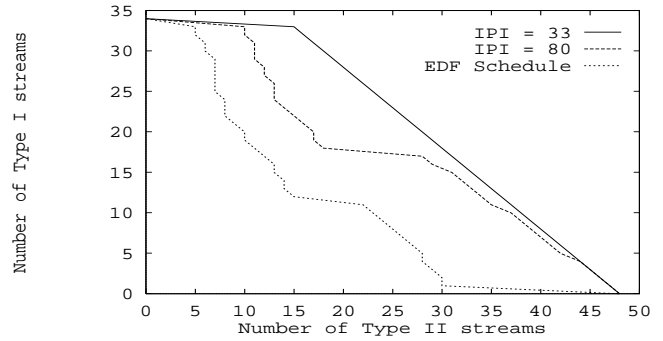The graphs shown in figures 3, and 4 are obtained

for deadlines of 33ms(low delay), and 330ms(high delay) respectively. For various number of streams where no packet was delayable beyond the deadline(Type I stream), streams that allowed 10 percent of the packets to be delayed beyond the desired deadline (Type II streams) were admitted until one of the tests failed. The initial priority index for all streams was kept identical and the rate of increase of priority for Type I streams was 1. The rate of increase of priority for Type II streams was varied to obtain the different curves.

As the delay bound for Type II streams is increased, the service order will start approaching SP because the difference in the rate of priority is large. This will tend to result in starvation, and Type II streams will have to be admitted as Type I streams, yielding an EDF service order. As a consequence, for large delay bounds for Type II streams TDP and EDF share a greater number of common possible configurations, that is, lines indicating the schedulable region overlap at the top left corner of the graphs. Further, if the desired deadlines are small, EDF can not work adequately due to the extreme time constraints, since the burst length is longer than the delay bound. The difference between schedulable configurations of EDF and TDP is then large. On the other hand, if the desired deadlines are very large any scheduling algorithm can fulfill the requirements; and the difference in schedulable regions of EDF and TDP is small with that of TDP being larger.

To study the effects of initial priority, Type II streams given a greater deadline than the desired deadline and a rate of increase of priority identical to that of Type I streams. This would be the result if the test to provide a percent delayable parameter was used with EDF scheduling. The initial priority index was lowered to check the effects on the schedulable region. The delay requirements were 33ms (low delay) and the slope of Type I streams was 1. A delay bound of 132ms was given to Type II streams and the initial priority was varied to obtain the graph shown in figure 5.

As can be seen in figure 5, when the burst length is longer than the delay bound EDF is unable to prior-

itize between streams and the number of connections establishable is small compared to TDP. TDP is able to delay Type II packets during a burst and thus able to establish more connections. As the delay bound becomes larger, the difference in the schedulable region becomes smaller for the same reasons as explained before.

Finally, if Type II streams are experiencing starvation due to difference in rate of increase in priority, extra buffering for Type II streams will increase the number of admitted streams and the gain for the extra buffering will be maximized. Also, if Type I streams are missing their deadlines due to low initial priority index of Type II streams, then providing extra buffer to Type I streams will alleviate the problem. Again, the gains for the extra buffer space are greater than if the space was used to provide a smaller extra buffer for all streams, as opposed to only Type I streams. The information on which stream should receive extra buffer space is obtained by knowing which test failed.

7 Conclusions

In this paper we have proposed a new scheduling algorithm that multiplexes VBR time constrained traffic over a link. Unlike traditional algorithms that either do not guarantee a QOS or do so at the expense of low utilization, our algorithm is able to provide QOS guaranteed even at reasonably high utilization. In fact, it provides close to optimal utilization and is also capable of using the loss parameter in the QOS of a stream to further increase utilization. It overcomes the flaws of EDF, such as *domino effect* by its ability to prioritize streams by providing different rates of priority increase.

We have proposed an implementation of complexity $O(m)$, where m is the number of switch (priority) queues. The complexity is independent of the number of streams admitted. If the number of switch queues is limited, the number of packets schedulable per second is large and the link can be potentially fully utilized. Further, in more recent switch designs, the queuing function is being implemented with custom silicon and static RAM permitting queues to be implemented as linked lists of cells. With this switch implementation, the proposed queues are easily realisable.

Although our implementation was provided for switches that use the BUP protocol, it can be easily modified to operate in switches that use rate-based flow control, as accepted by the ATM forum. This feature makes our scheme even more attractive. Apart from high utilization and easy implementation, TDP also offers greater flexibility since it leaves open the tradeoff between network bandwidth and buffer space.

Under medium delay requirements, the schedulable region obtained under TDP is significantly larger than that obtainable under EDF. As memory costs, the delay bound requirements will increase and we expect that most of the connections will operate at medium delay requirements. Thus our scheme offers high utilization for conditions that are likely to occur in BISDN networks.

References

[1] S. Chaudhry and A. Choudhary. Scheduling algorithms for guaranteed qos systems. *CASE Technical Report*, July 1995.

[2] S. Chaudhry, M. Raziuddin, and A. Choudhary. On guaranteed bandwidth channels. *International Conference on Network Protocols*, pages 47–55, 1995.

[3] Domenico Ferrari and Dinesh Verma. A scheme for real-time channel establishment in wide-area networks. *IEEE Journal on Selected Areas in Communications*, 8(3):368–379, April 1990.

[4] J. R. Haritsa, M. Livny, and M. J. Carey. Earliest deadline first scheduling for real-time database systems. *Proc. Real time Systems Symp.*, pages 232–242, 1991.

[5] J. R. Jackson. Queues with dynamic priority discipline. *Naval Research Quaterly*, 7, September 1960.

[6] L. Klienrock. *Queueing Theory, Vol 2*. John Wiley, 1967.

[7] E. Knightly, D. Wrege, J. Leibeherr, and H. Zhang. Fundamental limits and tradeoffs of providing deterministic guarantees to vbr video traffic. *Technical Report*, 1994.

[8] J. Leibeherr, D. E. Wrege, and Domenico Ferrari. Exact admission control for networks with bounded delay service. *Technical Report*, 1994.

[9] Thomas Leighton, Charles Leiserson, and Ronald Rivest. *Introduction to Algorithms*. The MIT Press, 1991.

[10] Youngho Lim and John Kobza. Analysis of a delay-dependent priority discipline in a multiclass traffic packet switching node. *IEEE INFOCOM*, pages 9A.4.1–9A.4.10, 1989.

[11] C. D. Loche. Best effort decisiion making for real-time scheduling. *Phd. Thesis, Carnigie Mellon Univ*, 1986.

[12] K. K. Ramakrishnan and Peter Newman. Integration of rate and credit schemes for atm flow control. *IEEE network*, March/April 1995.

[13] P. Thambidurai and K. S. Trivedi. Transient overloads in fault-tolerant real-time systems. *Proc. Real time Systems Symp.*, pages 126–133, 1989.

[14] Dinesh C. Verma, Hui Zhang, and Domenico Ferrari. Delay jitter for real-time communication in a packet switching network. In *Proceedings of the Triangle Conference on Communication Software*, pages 35–43. TRICOMM, 1991.