INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 353-364

Advance and Immediate Request Admission: A Preemptable Service Definition for Bandwidth Brokers

I.T. Okumus, F.U. Dizdar

Ibrahim Taner Okumus
K.Maras Sutcu Imam University, Computer Engineering Department
Avsar Kampusu, 46000, K.Maras, Turkey
okumus@mu.edu.tr

Ferhat Umut Dizdar
Mugla University, Graduate School of Science
Kotekli Kampusu, 48000, Mugla, Turkey

Abstract: Differentiated Services Architecture lacks control level functionalities and Bandwidth Brokers are proposed to fill that gap. In order to provide proper control level functionalities, Bandwidth Brokers need to provide services for both advance requests and immediate requests. There is a tradeoff between preemption of immediate flows and utilization of links. It is important for a resource manager to provide the promised QoS level to a flow without any preemption. In this study, we solve the preemption and the experienced QoS problem by defining a preemptable service and explain how this service works and also show the performance and scalability characteristics of resource manager with the addition of a preemptable service.
Keywords: admission control, IR Flows, AR flows,preemptable forwarding service, quality of services.

1 Introduction

Differentiated Services (Diffserv) Architecture [1] is the de facto standard to provide QoS in IP networks. However Diffserv only specifies the data level functionalities. Control level functionalities are left out from the architectural definition.There are various options to provide control level functionalities of Diffserv. One of the solutions is named Bandwidth Broker (BB) [2], which is essentially an intra-domain resource manager (IDRM) [3] in a Diffserv domain. This manager has admission control, authentication, authorization, inter-BB communication and resource management duties in a Diffserv network.

In the Internet, different applications generate traffic flows. Depending on the application, different flows have different QoS needs. Diffserv architecture defines QoS classes to address the needs of different flows [4] [5]. Again depending on the application, actual traffic generation time will be different. Some applications generate traffic right away and ask for resources to be used immediately. Some applications make advance requests for a traffic that will be generated in the future.

In order to provide QoS for all kinds of applications, IDRM needs to provide service for both advance requests (AR) and immediate requests (IR). In this study, to provide proper QoS for both type of requests we propose a new service called Preemptable Forwarding (PF). We present the details of the architecture with the new PF service and analyze the performance and scalability characteristics of the approach.

Rest of the paper is organized as follows. In the following section, we summarize the previous work in this subject. Section 3 gives the details of the IDRM and also the definition of a PF service. Section 4 provides the analysis results. Comparison of the results with earlier work is given in section 5. We give the concluding remarks in Section 6.

2 Related Work

One of the earlier works in the area of resource planning in advance request agents is provided by Schelen and Pink [6]. In this study, authors tried to reduce the preemption ratio of the IR flows. Authors experimented with different look-ahead times to analyze its effect on preemption rate of IR flows. The look-ahead time is kept constant in time (CLAT). Authors show that when look-ahead time increases, the preemption rate decreases but utilization also decreases.

Lin et.al. [7] proposed to change the look-ahead time depending on the application. In this scheme, authors use prediction methods to calculate the holding time of an IR flow and use this time as the look-ahead time. This scheme requires sensitive prediction techniques which brings high cost to the computing entity.

Ahmad et.al. [8] proposes to use dynamic look-ahead time to solve the problem. Authors argue that look-ahead time is dependent on resource scarcity, IR arrival and release rate, and bandwidth release per IR call. Authors calculate the LAT dynamically (DLAT) by considering these parameters at the time of the calculation. When there is a need for preemption, IR calls that are accepted most recent time are preempted to reduce wasted throughput. However size of these flows is also important. Preempting a recently accepted giant flow will cost more to the network than preempting an old small flow. Also when the AR limit is low and the LAT is high, DLAT model exhibits comparable results with CLAT model.

Degermark et.al. [9] used a model where all flows declare their duration, and admission is based on measurements and future predictions of the traffic load. In this scheme none of the flows are preempted. AR flows are expected to provide the duration of the flow. However assuming all IR calls to provide their duration is not realistic in today's networks.

Greenberg, Srikant, and Whitt [10] proposed a probabilistic approach for AR admission control. In this method, an AR call is admitted based on the call interruption probability. If the probability is below a threshold, request is admitted. This method assumes all AR calls are made for far ahead in time. This scheme allows occasional service disruptions to achieve higher utilization.

Srikant and Whitt [11] used CLT approximation to calculate interrupt probability for resource sharing between AR and IR flows. This method allows multi-class resource scheduling.

Karsten et.al. [12] proposed a policy-based service specification for advance resource reservations. In the study, an IR request is assumed to be nonpreemptable for certain amount of time and preemptable after that duration. During the admission process, IR requests provide the nonpreemptable duration. Also it is possible for a flow to request nonpreemptable for the whole lifetime. However, it is not realistic for a flow to require certain QoS for a limited time and not need that QoS after that. Applications require certain QoS for the whole lifetime. If all IR requests are nonpreemptable, then the network utilization is low. This study does not provide any evaluation result for the suggested scheme.

In order to provide an acceptable service for both AR and IR calls it is important to have both flows accepted in the network and also to have a sensitive preemption scheme to reduce preemption rate of IR flows. When providing QoS, if an IDRM accepts a flow into the network, whether it is AR or IR, flow expects to get the promised QoS through the duration of the flow. Unaware preemption should not be an option in any case. None of the proposed solutions consider user satisfaction and experienced-QoS in their study. In the next section we provide the solution we propose for this problem.

3 Proposed Model: Preemptable Forwarding

In our work, we used IDRM architecture as a resource manager [3]. IDRM is mainly a specialized bandwidth broker. IDRM is responsible for admission control and intra-domain resource management along with other tasks. These two tasks are the important ones for our study.

In order to better manage intra-domain resources, IDRM knows intra-AS topology. IDRM keeps track of the available capacity of individual links in the network. IDRM also keeps track of the currently reserved resources. In this context, to support Advanced Requests (AR), IDRM needs to keep records of start and end times and required capacity of advanced reservation flows. In bandwidth database, IDRM keeps current load of the links and differentiates immediate and advance reserved shares of the bandwidth (Table 1).

Table 1: A sample database state

LinkID	UtilizedBW	ReservedBW
40	5	7
41	5	5

Table 2: Time slot table for reservation states

LinkID	40								
Time Slot	1	2	3	4	5	6	7	8	9
Reserved BW	1	2	4	2	0	1	5	4	0

IDRM also keeps reservation database for AR requests. For the admission of advance requests, IDRM needs to know the available capacity and the reserved capacity of a future time interval. Since time is an analog entity, it is impossible to keep track of the link states in every point in time. Usual way to make time more manageable is to quantify it. We call these quantified time intervals *time slots*. IDRM keeps track of reserved capacity of every time slot through the future. To prevent the scalability problems, the number of time slots into the future needs to be finite. The actual amount of a time slot can be determined according to the network's need. It can be seconds, minutes, hours. A sample state of reservation database is shown in Table 2.

3.1 Admission Control

Admission control methods for IR, AR and PF requests are different. The goal is to reduce the preemption rate of IR flows and to increase the overall throughput as much as possible. In order to achieve this goal, we need to keep a balance between IR and AR flows. If we only accept IR flows, admission control decision is simply based on current available capacity on the path of the flow. Determination of this capacity can be parameter-based or measurement-based which is out of scope of this paper. If we only accept AR flows, then the admission decision is based on the available capacity from the start time to the end time of the request. When we have both IR and AR together, we need to use a mixture of these two admission control methods.

We define LC as the total link capacity, IRLC as the link capacity share of IR flows, CIR_i as the capacity of the i_{th} active IR flow,CAR_i as the capacity of the i_{th} active AR flow, $CIR(t_0)$ as the total IR capacity used by IR flows at time t_0, $CAR(t_0)$ as the total AR capacity used by AR flows at time t_0. $CIR(t_0)$, and $CAR(t_0)$ is calculated as follows:

$$CIR(t_0) = \sum_{i=0}^{n_1} CIR_i, i = 0, 1, ..., n_1 \tag{1}$$

$$CAR(t_0) = \sum_{i=0}^{n_2} CAR_i, i = 0, 1, ..., n_2 \tag{2}$$

Residual capacity at time $t_0$for AR and IR share is calculated as follows:

$$IRLC_r(t_0) = IRLC - CIR(t_0) \tag{3}$$

$$ARLC_r(t_0) = ARLC - CAR(t_0) \tag{4}$$

Admission Control of IR Flows

Considering the preemption and throughput tradeoff, there can be two different admission control approaches for IR admission. In first case, to increase the throughput, IR flows can be allowed to overflow into the AR share of the link capacity. However some IR flows can be preempted if necessary. In the second case, to prevent preemption, IR flows can be limited into their own share of the capacity and not allowed to use the excess AR capacity.

Admission of the first case is based on checking the available capacity on the whole link. If there is enough available capacity at the time of the request, then the IR flow is accepted:

$$c < IRLC_r(t_0) + ARLC_r(t_0) \tag{5}$$

Admission of the second case is based on checking available capacity only on the IR share. If there is enough capacity IR flow is accepted:

$$c < IRLC_r(t_0) \tag{6}$$

Admission Control of AR Flows

Admission control criterion for AR flows is different. AR flows always have priority over IR flows. Admission control decision does not take active IR flow capacity into account. Decision is based on the AR reservations in the duration of the requested flow. If AR flow starts at t_1 and ends at t_2, condition to accept the flow is:

$$ARLC > c + max[C_{AR}(\tau)], \tau = [t_1, t_2] \tag{7}$$

Where ARLC stands for AR Link Capacity share, $max[C_{AR}(\tau)]$ shows the maximum reserved AR capacity in the interval τ. If this condition is satisfied, AR flow is accepted. At time t_1 AR flow will start. It is possible that in between the time of accepting the AR flow and the actual AR flow start time, some IR flows could be accepted. During interval τ, some of the IR flows can be preempted to provide capacity to the pre-accepted AR flows.

Preemptable Forwarding Service

This requirement led us to define a new service type. As we will show in our analysis results, in order to get the most benefit from the network while having both IR and AR flows together, there will be a need to preempt IR flows. Instead of selecting a flow without the consent of the owner, we define a new service called Preemptable Forwarding (PF). In this service type, some IR calls accept the probability of preemption beforehand during the admission control process. From hereon, we refer this flow type as PFIR. For PFIR flows, there is a chance that accepted flow will never be preempted. In this case the experienced QoS will be the same as other IR and AR flows. However if there is a need for preemption, network will select one of the PFIR flows to preempt. None of the regular IR flows will be preempted. Since the user is expecting this preemption and has consented to preemption beforehand, there will not be any user dissatisfaction in terms of the experienced QoS.

This service type is most suitable in cases where there is a limit on both IR and AR flows. Bandwidth capacity is divided between IR & AR flows and none of these flows can overflow

to other flow's share. If there is no capacity available in IR share for a new IR request, PF admission is used.PF flows will be using the excess capacity in the AR share of the link capacity. The amount that can be used by PF flows can be determined beforehand and can be static or dynamic.

Admission Control of PFIR Flows

The admission control decision for PFIR depends on the IR capacity, AR capacity and PF share. If available capacity on the IR share is not enough to accept an IR request, user is prompted to retry for PFIR. PFIR request is accepted only if the residual IR capacity and residual PF share are in total bigger than the requested capacity. Residual PF capacity $PFLC_r$ is dependent on the look-ahead time (LAT), AR reservations in that time interval and the active PFIR flows. Residual AR capacity ($ARLC_r$) is defined as the capacity available on the ARLC and is defined as:

$$ARLC_r = ARLC - max[C_{AR}(LAT)] \tag{8}$$

At time t_0, total PFIR capacity in use is calculated as:

$$PFIR(t_0) = \sum_{i=0}^{n_3} PFIR_i, i = 0, 1, ..., n_3 \tag{9}$$

Residual PF capacity at time t_0 is then defined as:

$$PFLC_r(t_0) = \begin{cases} PFLC - PFIR(t_0) & ARLC_r \geq PFLC \\ ARLC_r - PFIR(t_0) & ARLC_r < PFLC \end{cases} \tag{10}$$

PF admission is based on the residual IR and residual PF capacities:

```
if (c < IRLCr + PFLCr)
 then accept
else reject
```

In order to clarify the admission of PFIR flows, lets take a look at the admission process. An IR request will arrive at the BB via a signaling protocol such as SIBBS [14]. BB will determine the type of the request (IR or AR). If the request is IR, BB performs IR admission procedure. If IR capacity is not enough to accept the request, BB sends back a negative message with PF admission option. This message suggests a PF request and also contains an indicator about the preemption possibility. We define $PFRate(t_0)$ as the PF share usage rate at time t_0:

$$PFRate(t_0) = 1 - \frac{PFLC_r(t_0)}{PFLC}, 0 \leq PFRate(t_0) \leq 1 \tag{11}$$

Average PF rate is calculated using moving average to take into account the previous trends on the PF usage rate. Average PF rate takes values between 0 and 1. 0 indicates that no PF flow is active and PF capacity is available. 1 indicates PF capacity is fully used. Values of variables a and b can be set based on the network needs. In our study, we used values a=0.2 and b=0.8. Average PF rate is calculated as follows:

$$AvgPFRate(t_0) = a * PFRate(t_0 - 1) + b * PFRate(t_0), 0 \leq a, b \leq 1, a + b = 1 \tag{12}$$

Average PF rate shows the possibility of preemption of a PFIR flow in LAT time window. Based on this indicator, user can predict preemption possibility and decide whether or not to use the

PF service. After receiving the negative response and average PF rate, request can be submitted as a PF request.

After PFIR flow is admitted there is a possibility that some AR flows can request a reservation and reclaim the bandwidth currently used by PFIR flows. Then some PFIR flows will be preempted. In this case we propose to preempt latest PFIR flow. A detailed loss-benefit analysis needs to be made on the preemption choices to develop a method to choose the flow to be preempted that will have least harm on the network benefit. However, this study is out of scope of this paper.

4 Simulation Results

Our analysis mainly covers the effects of the admission control methods on throughput and preemption rate. Along with these, to measure the scalability we define benefit and processing load as other measures. 1 Unit benefit is defined as the amount of income network gets from a 1Mbps of flow in 1 sec. If a 100Mbps link is fully utilized for 20 seconds, benefit is 100*20=2000 units. Processing load is the amount of work required to process an incoming request. In order to test this, we calculate the number of accesses to databases by IDRM to produce a response to a request. If an IDRM accesses bandwidth table 2 times and time slot table 4 times that requests' processing load is 6 units.

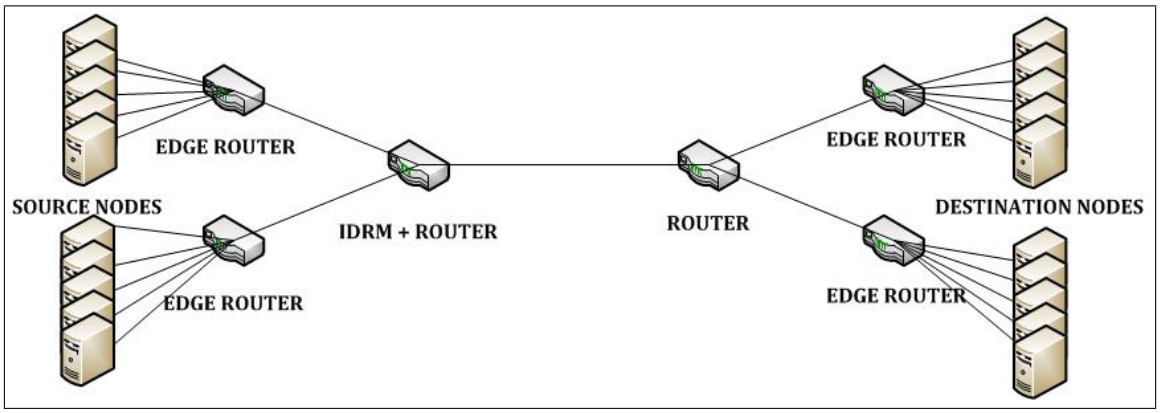

Figure 1: Simulation Topology

IDRM is implemented on ns2 simulation software [15]. We used the topology shown in Figure 1 in our simulation. In this topology, there are 10 source nodes and 10 destination nodes. All the link capacities are 100Mbps. The link between *IDRM+router* and *router* is the main bottleneck link. We assume that all flows are unidirectional from sources to destinations. Flows and reservation requests are entered into the domain from edge routers. Edge routers forward reservation requests to IDRM. IDRM applies admission control methods and sends back an accept/reject answer to the source.

The incoming reservation requests follow Poisson distribution with a mean arrival of 10 requests per time slot. Time slots are used as the time measure and on average 10 request is produced in a single time slot by all the traffic sources. IR & AR limit of these requests are varied depending on the scenario. Every request asks for 1Mbps capacity from the network and flows last for 20 time slots. In all the simulation scenarios, there is a warmup time. Because of the warmup time, first 100 time slots is excluded from evaluation. All results indicate the steady state case.

In order to determine the network behavior and use it as a benchmark for other results, we first tested the network with only IR and with only AR flows.

4.1 IR-only and AR-only flows

In IR-only scenario, sources only make IR requests from the network. There is no limit on the IR flows. These flows can consume the whole link capacity. During the simulation, links are saturated and the network behavior under full capacity is observed. In this case, 29.24% of IR are rejected because of unavailable capacity. Maximum benefit is 20000 units and achieved benefit for this scenario is 19627 units. Average utilization is calculated as 98.13%. Average process-time is 34.17 units and total process time is 70577 units.

In the AR-only case, sources make only AR requests from the network. No limit is imposed on AR flows in terms of the bandwidth use. In this scenario, 38.51% of AR requests are rejected. The difference between IR-only and AR-only case is that, AR-only case works with time slots. That means flows can start only at the beginning of a time slot. However, IR flows can start anytime. This is one of the reasons for high rejection ratio of AR flows. Total network utilization in this scenario is on average 90.73%. Network gained 18146 units of benefit out of 20000 maximum benefit. Total process-time for AR-only case is 166327 units.

Comparing the two cases, AR flows achieve lower network utilization because of the scheduling problems and the slot timing. AR flows achieve lower benefits and also higher process times compared to IR flows.

4.2 IR & AR together, No limit on AR

In this scenario, IR and AR flows appear together in the network. 70% of the requests are IR requests and 30% of the requests are AR requests. Since the holding time for flows is 20 time slots and each flow consumes 1Mbps of bandwidth, resources are saturated in 10 time slots.

Results show that network utilization is 96.62% on average. In case of congestion, IR flows are preempted. 44.2% of the total accepted flows are IR flows and 55.8% are AR flows. 67.9% of IR flows and 0.35% of AR flows are rejected. During the simulation 4.66% of accepted IR flows are preempted. In this IR&AR-no-limit case, when both flows are active in the network, utilization is lower compared to the IR-only case, but it is higher than the AR-only case.

4.3 IR & AR, AR Limited

In this scenario, our goal is to analyze the effect of limiting the AR flows on network utilization. To determine this effect we set the limit of AR flows to 50% in the first simulation and reduced it by 10% for each successive runs.

AR flows can use the limited capacity reserved for them. If there is not enough capacity, then AR flows are rejected. However, IR flows do not have any limit. If there is available capacity in the IR part, flow is accepted immediately. IR flows are preempted in the future when an AR flow asks for the capacity that is being overused by IR flows. Figure 2 shows AR share and accept/reject ratios for both IR and AR flows. Figure 3 shows IR and AR throughput and total throughput for AR-limited scenario. As it can be seen from the figure, as the AR percentage drops, total throughput increases. Network utilization changes from 97.23% for 50% AR share to 98.26% for 10% AR share.

Figure 4 shows the IR drop rate. IR flows are preempted as AR flows move in for their share of the capacity. Figure clearly shows that IR drop rate increases as the AR share increases in the network.

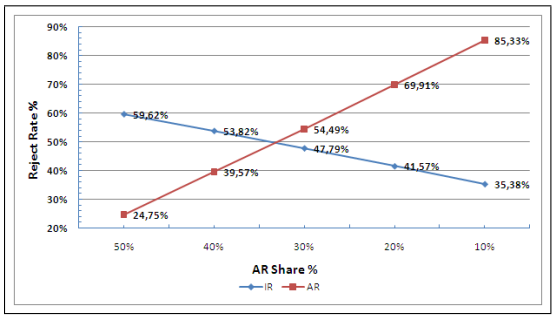

Figure 2: Accept/Reject Rates of IR & AR flows for different limits on AR flows

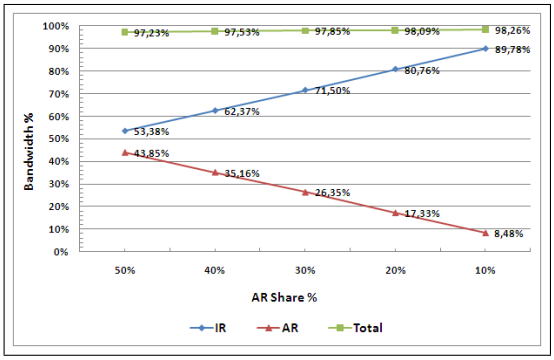

Figure 3: IR, AR bandwidth shares and total throughput for AR limited case.

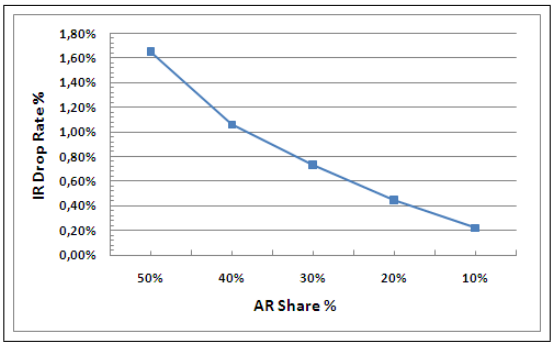

Figure 4: IR Drop Rate for AR limited case

Table 3 shows processing time shares of IR and AR flows for different limits on AR flows. Results show that as the AR limit decreases, total processing time also decreases. This is in accordance with the fact that admittance of AR flows is more costly than admittance of IR flows.

Table 3: Process times for AR limited case

Limit	IRAvg	IRTotal	ARAvg	ARTotal	TOTAL
50%	23	159756	99	279883	439639
40%	25	174704	83	236285	410989
30%	27	190244	66	187944	378188
20%	29	206265	49	140279	346544
10%	32	222212	28	80077	302289

Table 4: Process times for both limited case

Limit(%)	IRAvg	IRTotal	ARAvg	ARTotal	TOTAL
50-50%	21	146917	99	279883	426800
60-40%	23	165047	83	236285	401332
70-30%	26	183177	66	187944	371121
80-20%	29	201307	49	140279	341586
90-10%	31	219347	28	80077	299424

4.4 IR & AR, Both Limited

As we can see from the AR-limited scenario, some IR flows are preempted, which is not desired in a QoS environment. To prevent that, one option is to limit both flows. In this scenario we divide the total capacity between these two flows. We do not allow any of the flow types to use the capacity from the other flow types' share. As in previous case, highest capacity reserved for AR flow is 50%. This limit is reduced by 10% for each simulation run and IR limit is increased by 10% to analyze the effect of the amount of share each flow type gets from the network. We start with 50-50 (IR-AR), then change to 60-40, 70-30 and so on for each successive runs.

Figure 5 shows accept/reject ratios for both limited case. Compared to previous scenarios, IR accept-rate is reduced. The reason is that after using the capacity reserved for IR flows, all IR requests are rejected.

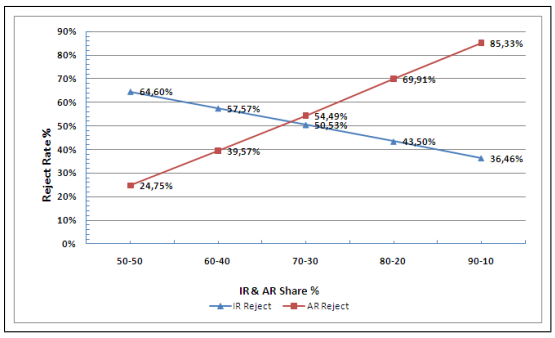

Figure 5: Accept/Reject Rates of IR & AR flows for different limits on both flows

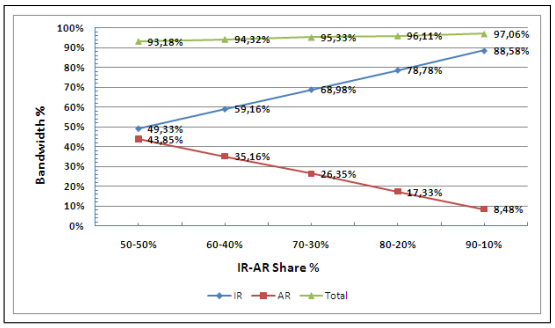

Figure 6: IR, AR shares and Throughput for both limited case

Figure 6 shows IR-AR shares and total throughput for both-limited case. In this scenario, throughput is decreased significantly compared to the base cases and to the AR-limited case. For 50-50 case, throughput is 93.18% and for 90-10 case throughput is 97.06%.

Table 4 shows process times for both limited case. Compared to the AR-limited case, AR process times are same; however IR process time is decreased because less IR flows are accepted into the network.

In this scenario, no IR flows are dropped. Since both flows have their own share and they are bound to it, preemption is not necessary. However, this is achieved by sacrificing from total throughput. This result also confirms that there is a clear tradeoff between throughput and the preemption of IR flows.

Since the main goal of the network provider is to get the maximum benefit from the network, which can be achieved by maximum utilization of the network capacity, we need to find a balance between the preemption and the total throughput. If preemption is inevitable, then the question is which flows should be preempted. In our study we prefer to define a new service called preemptable forwarding (PF). IR flows that are accepted as PF flows will be preempted in case of an overuse.

4.5 Preemptable Forwarding Service with Lookahead Time

In our scheme, in order to balance the throughput and preemption we select to limit the capacities of both IR and AR flows. As we showed in section 4.4, this scheme has low throughput but no IR drops. In order to increase the throughput we allow certain percentage of AR capacity to be used by PF flows if necessary.

We analyzed this scenario on the same topology with the same parameters. In the analysis, we have two parameters to consider. First one is the LAT. How LAT affects the throughput? Second parameter is the PF percentage in the AR region. We analyzed the effect of PF percentage on the throughput. In the simulations we set the IR-AR ratio to 50-50 and changed the PF share from 1% to 15% of the total link capacity and also changed LAT from 1 time slot to 15 time slots.

Figure 7 shows IR lost benefits because of preemption for different PF shares. For low LAT values, IR drop rate is higher for high PF shares. As LAT value increases, lost benefit for all PF shares converge to zero.

Figure 8 shows the total throughput of the network for different LAT values and PF shares. This graph shows that total achieved throughput is higher for lower LAT values and higher PF shares. Again these results show the clear tradeoff between IR drops and the total throughput.

Table 5 gives the utilization values for different PF share and LAT values. Table 6 shows

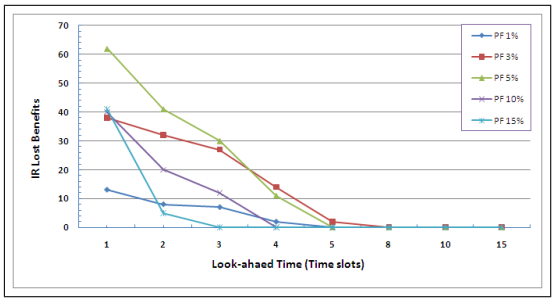

Figure 7: IR Lost benefits

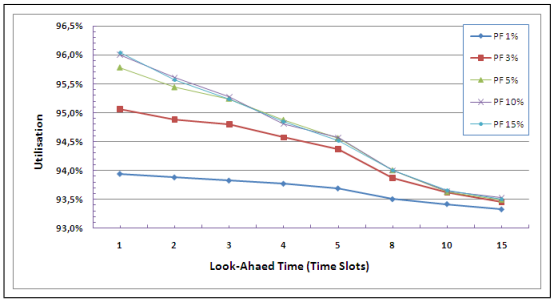

Figure 8: Total throughput

Lost IR benefits for different LAT values and PF shares.

Table 5: Throughput values for different LAT values and PF shares

PF(%)	1	3	5	10	15
LAT	Utilization (%)				
1	93.94	95.07	95.78	96	96.04
2	93.88	94.88	95.44	95.61	95.57
3	93.825	94.80	95.24	95.27	95.23
4	93.77	94.57	94.88	94.81	94.85
5	93.69	94.37	94.56	94.57	94.52
8	93.50	93.88	94.01	94	94.01
10	93.42	93.62	93.63	93.64	93.66
15	93.33	93.46	93.50	93.53	93.5

Table 6: Lost IR benefits for different LAT values and PF shares

PF(%)	1	3	5	10	15
LAT	Lost IR Benefits				
1	13	38	62	40	41
2	8	32	41	20	5
3	7	27	30	12	0
4	2	14	11	0	0
5	0	2	0	0	0
8	0	0	0	0	0
10	0	0	0	0	0
15	0	0	0	0	0

Results on Table 5 and 6 suggest that for 50-50 IR-AR share, best option is to set PF share at 15% and LAT value at 3 time slots. These values give the highest throughput on the network without any preemption. Highest utilization that can be achieved in this scheme is 96.035%. This value is considerably higher than the 50-50 IR-AR both limited case, which has 93.18% utilization.

Processing time for a PF request consist of two components. A PF request is originally an IR request. When there is lack of resources on IR share, request is rejected for resubmission as a PFIR request. PFIR admission is similar to an AR admission. Residual AR capacity is calculated and decision is made depending on that value. So the process is a two pass process. First pass results in rejection of an IR request. Second pass is PF admission. In terms of the processing time, admission of a PFIR request is more costly than both admission of an IR request and also admission of an AR request for the same LAT value. Regular AR admission has a constant LAT value. PFIR processing time increases with the increase of LAT. Table 7 shows average processing times of PFIR requests for different LAT values.

Table 7: Average Processing times for different LAT values

LAT	1	2	3	4	5	8	10	15
Avg. Process Time	82	84	85	87	90	97	103	110

5 Comparison of the Results

Ahmad et.al. [8] proposed a dynamic lookahead time (DLAT) solution for reduce preemption rate at the cost of lower throughput. This study provides results for DLAT and compared the results with the study conducted by Schelen and Pink [6]. Schelen and Pink used constant lookahead time (CLAT) in their study. We will also compare our findings with DLAT and CLAT models.

In our PF model, we used 50%IR and 50%AR share. We compare these results with the same AR limit case in Schelen and Ahmad study. When AR limit is set to 50%, highest normalized throughput achieved by DLAT model is 0.84 with DLAT c=1.0 and highest throughput achieved by CLAT is 0.85 with CLAT 30. These LAT values corresponds to LAT value 2 in our study. With LAT=2, highest utilization in our case is 95.6% (PF %10) and lowest utilization is 93.8% (PF %1). This shows that our scheme performs better than both schemes in terms of the network utilization.

In terms of the preemption rate, our results show the number of preempted flows, while DLAT and CLAT models show preemption probability. We will compare the preemption trends with those studies. Our results indicates that it is possible to prevent preemption by selecting appropriate LAT and PF share values. Also while longer LAT values result in lower preemption, higher LAT values result in higher preemption. DLAT and CLAT models also show the similar trends. Longer DLAT and CLAT values result in lower preemption probability.

As a summary, DLAT model achieves lowest throughput than CLAT and our PF model. PF model performs best in terms of the total throughput. Preventing preemption is necessary in order not to disrupt the QoS of the accepted flows. DLAT model provides this with the cost of throughput. Our model suggest a PF scheme where user pre-agrees a preemption when and if necessary. Also PF scheme can provide low preemption ratios by selecting LAT and PF% values properly. Our proposed scheme results in higher utilization and also in higher user satisfaction in terms of the perceived QoS.

6 Conclusion

In this study we analyzed the admission control method of an IDRM that supports both IR and AR flow types.The main tradeoff in this environment is between preemption of IR flows and the total throughput of the network. As the throughput increases, IR preemption also increases. If we decrease IR drop rate, throughput decreases. Another issue is user satisfaction due to IR drops and also the scheme to select which IR flows to drop in case of a capacity problem.

In order to increase the perceived QoS on the user side, instead of selecting an IR flows among the ones that the network accepted and promised to provide a certain QoS without any interruption, we propose to employ a new QoS class called Preemptable Forwarding (PF). This flow type will be accepted to the network with the condition that the flow will be preempted and not get the promised QoS in case of congestion. Users will be accepting the service with the possibility of a preemption or not get any QoS at all.

We analyzed the effects of the PF share from the total capacity on the IR preemption and the total throughput. As the PF rate increases, total throughput also increases. However, increased PF also causes high IR drop rate.

When we employ lookahead time (LAT) before accepting the PF requests, behavior changes. With high LAT values, IR drop rate is reduced. However, this causes the network utilization to decrease.

Results show that employing a resource manager that uses PF service in admission control of the flows can increase the total throughput and also the user satisfaction in a QoS network.

Bibliography

[1] Blake, S. et.al., An Architecture for Differentiated Services, RFC 2475, 1998

[2] Nichols K. ,Jacobson V. ,Zhang L. , A Two-bit Differentiated Services Architecture for the Internet, RFC 2638, 1999

[3] Mantar H.A., Okumus Ý.T., Hwang J., Chapin S.J., A Scalable Intra-Domain Resource Management Architecture for Diffserv Networks, *Journal of High Speed Networks*, 15, 185-205, 2006

[4] Jacobson V., Nichols K., Poduri K., An Expedited Forwarding PHB, RFC 2598, 1999

[5] Heinanen et.al., Assured Forwarding PHB Group, RFC 2598, 1999

[6] Schelen O., Pink S., Resource Sharing in Advance Request Agents, *Journal of High Speed Networks: Special issue on Multimedia Networking*, 7(3-4):213-228, 1998

[7] Lin Y., Chang C., Hsu Y., Bandwidth Brokers of Instantenaous and Book-ahead Requests for Differentiated Services Networks, *ICICE Transactions on Communication*, E85-B, No.1,278-283, 2002

[8] Ahmad I., Kamruzzaman J., Aswathanarayaniah S., A Dynamic Approach to Reduce Preemption, in *Book-ahead Reservation in QoS-Enabled Networks*, *Computer Communications*, 29(9):1443-1457, 2006

[9] Degermark M. Et.al., Advance reservationss for Predictive Service in the Internet, *Multimedia Systems*, 5(3):177-186, 1997

[10] Greenberg A.G., Srikant R., Whitt W., Resource Sharing for Book -Ahead and Instantaneous -Request Calls, *IEEE/ACM Transactions on Networking*, 7(1):10-22, 1999

[11] Srikant R., Whitt W., Resource Sharing for Book-Ahead and Instantaneous-Request Calls Using a CLT Approximation, *Telecommunication Systems*, 16(3-4):233-253, 2001

[12] Karsten M., Beries N., Wolf L., steinmetz R., A Policy-Based Service Specification for Resource Reservation in Advance, *Proceedings of the International Conference on Computer Communications (ICCC'99)*, Tokyo, Japan, 82-88, Sept 1999

[13] Ahmad I., Kamruzzaman J., Preemption Policy in QoS-Enabled Networks: A Customer Centric Approach, *Journal of Research and Practice in Information technology*, 39(1):61-79, 2007

[14] Adamson A. et.al., QBone Signaling Design Team final Report", Internet2 QBone Sinaling Workgroup, http://qos.internet2.edu/wg/documents-informational/20020709-chimento-etal-qbone-signaling/, Jul 2002

[15] The Network Simulator - ns-2, http://www.isi.edu/nsnam/ns/