Delay Performance of Multi-Service Network with Strict Priority Scheduling Scheme

Hoon Lee*, ***Regular Member***

ABSTRACT

Strict priority scheduling scheme is a good candidate for the implementation of service differentiation in an Internet because of simplicity in implementation and the capability to guarantee the delay requirement of the highest class of traffic. However, it is also known that strict priority starves the lower-class traffic at the cost of prioritizing the higher-class traffic. The purpose of this work is to propose an analytic method which can estimate the average delay performance of Diffserv service architecture and shows that strict priority scheme does not sacrifice the lower class traffic over a diverse condition of the load. From the numerical experiments for three-class Diffserv network we validate our argument that strict priority scheme may be applied to a service differentiation scheme for the future Internet.

Keywords : Performance evaluation, IP network, Differentiated service Internet, Packet scheduling, QoS guarantee

Ⅰ. Introduction

With the development of networking technologies a number of new services such as voice and data services are emerged to IP network. As such, a service differentiation scheme for the packet from different kinds of applications becomes one of the hottest issues in IP network. The typical service differentiation schemes may be the SP (Strict Priority), WFQ (Weighted Fair Queuing), and CBQ (Class-based Queuing) [1]. SP serves the packets from the strict priority order between the different traffic classes (a detailed discussion on SP is given in Section II). WFQ tries to guarantee both the fairness and service differentiation between different traffic classes by setting a certain amount of bandwidth resource to a specific queue, and the amount of the bandwidth is determined by a weight for each traffic class.

A more sophisticated scheme is the hybrid of SP and WFQ scheme in which the EF (Expedited forwarding) class traffic is served with strict priority over the lower classes, via which the packets from the highest class traffic is logically separated from the lower classes of traffic. The remaining classes are served with a certain weight between the classes when there is no packet in the queue of the highest class of traffic.

CBQ tries to solve the problem of SP and WFQ at the same time by giving more freedom to the network operator in realizing the service policy by handing the right to determine the amount of service time over the operator's own service policies.

It is evident that the above two schemes such as WFQ and CBQ are more efficient than the SP scheme in the freedom of policy-enforcement. However, they have a number of drawbacks: First, they are too complex to be operated efficiently in a very fast network if the traffic volume shuffles very quickly. Second, they are not scalable in implementing to the router of large-scale network. Third, and most importantly, the above schemes except the hybrid SP and WFQ scheme can not guarantee the strict delay performance of voice traffic unless a certain limit is set upon the

---

* Changwon National University, Changwon, Korea, 641-773 (hoony@changwon.ac.kr)
논문번호 : KICS2004-10-241, 접수일자 : 2004년 10월 19일

length of the cycle for the service. Otherwise, voice packets may experience too much delay during which the packets from the other classes are served.

However, the requirement for the inter-service time of voice packets is almost constant, whereas it depends heavily on the number of traffic classes as well as the characteristics of the corresponding applications that share the link of a router. Therefore, it is not an easy job to determine the weight for each class of traffic over the operating networks in an optimal manner if one wants to guarantee the strict delay requirement of the voice source.

On the other hand, SP scheme is the most efficient if the focus of the QoS differentiation is put on the guarantee of the delay and delay jitter of the most significant class of traffic in multi-class service network [2], because the priority of service right is set first to the highest priority class, the voice packets. Therefore, SP scheme acts almost in the same manner as that of a dedicated link to voice packets [3]. In addition, SP is very simple to implement in the router. Finally, it is easy to estimate the delay performance of each class of traffic for SP scheme in an analytic manner if a few parameters are known, which is discussed in this work.

SP has some drawbacks: It is not free from the criticism of starvation of the lower classes of traffic, especially if the speed of the outgoing link is low and the offered load of voice traffic is high. However, to the best knowledge of the authors, little work has been done on the rigorous evaluation of the delay performance of SP scheme from diverse points of view such as the effect of the delay under the various conditions of traffic source profile and the different conditions of link capacities.

This paper tries to anatomize the delay performance of the each class of traffic for the SP scheme in detail. In addition, let us have one more purpose in mind, which is to hear carefully the vindication of the SP scheme such that SP scheme is not so negligent in caring for the lower class traffic. In other words, we investigate the delay experienced by packets of lower classes, and confirm that whether SP scheme really starves lower classes of traffic or not under the diverse network and source traffic profiles. Via approximate mathematical analysis and numerical experiments, we will explore whether the argument of the SP scheme is rational or not.

This paper is composed as follows: In Section II SP scheduling scheme in Differentiated Service (DiffServ) framework for the Internet is explained. Section III is devoted to the analysis of the waiting time for the SP scheme. In Section IV, numerical experiment and a discussion on the results is given. Finally, in Section V, the implication of the work is summarized.

Ⅱ. Modeling the SP Scheme

Under the DiffServ architecture, packets are classified as an EF, BBE (Better than Best Effort, which is the same as the AF, the assured forwarding), and BE (Best Effort) PHB (Per Hop Behavior), and let us call each PHB as class 1, 2, and 3, respectively. Each packet is classified at the ingress edge of the network and served with strict priority (SP) from there in, so that packets of the higher class pass the core network with strict priority over the lower class traffic at once packets from each class enter the backbone network. Let us assume that an access router accommodates voice and two types of data packets from a number of connections. At each router composing the end-to-end path of the network, packets are classified into voice and two types of data, and each packet is fed into corresponding buffer, the EF, BBE, and BE buffer. It is assumed that packets generated from a number of input interfaces are distributed into a number of output ports with even distribution. The last assumption renders us to simplify the analysis of a router into a single output port.

It is known that the packet generation processes of multimedia applications such as the web, ftp, and video steam follow self-similar or long-range process with different parameters. However, it is very difficult to model the arrival process of an arbitrary link that accommodates a large number of connections from diverse traffic sources inside a network, and it can be approximated by Poisson process if the number of connections is sufficiently large [4]. Based on this fact, we assume an approximate but a tractable model for the evaluation of performance for the prioritized packet service scheme in a node by using M/G/1 queuing model, where a link is modeled as a single server, the arrival process is Poisson, the service time is generally distributed, and the buffer capacity is sufficiently large. Furthermore, let us model the server as a non-preemptive server with strict priority policy for voice buffer, which faithfully models our packet service architecture.

Using the non-preemptive M/G/1 queuing model with SP scheme in [5], let us describe a procedure for obtaining the waiting time of each class packet by assuming some variables. Packet arrival process from each class of packets is mutually independent, and packets from each class arrive following a Poisson process with mean arrival rate of λ_1, λ_2, and λ_3 for EF, BBE, and BE packets, respectively. The service time of packets from each traffic class follows general distribution with mean service rate of $1/\mu_1$, $1/\mu_2$, and $1/\mu_3$, for EF, BBE, and BE packets, respectively. The variance of the service time of packets from each traffic class is assumed to be σ_1^2, σ_2^2, and σ_3^2, respectively. The mean offered load of the EF, BBE, and BE packets into corresponding buffer is $\rho_1=\lambda_1/\mu_1$, $\rho_2=\lambda_2/\mu_2$, and $\rho_3=\lambda_3/\mu_3$, respectively.

Packet scheduling at the buffer module follows an SP scheme, which operates in the following manner. Initially, server visits an EF buffer. If there exist any packets in EF buffer, the server serves them until the buffer is vacant. Otherwise, the server visits BBE buffer and serves a packet in that buffer. After a service to a BBE packet is finished, and if there is no packet in EF buffer, the server then visits BE buffer, and serves a packet in that buffer. Now the server visits EF buffer and repeats the above operation.

Let us assume that the moving times between the three buffers are so small that they are ignored. When a higher-class packet enters a buffer while a lower class packet is receiving service by the server, it waits in his/her buffer until the server finishes service for the current packet. Therefore, the service scheme is non-preemptive.

Ⅲ. Waiting Time Analysis

Let S_1 be the sojourn time in the system (buffer and server) and W_1 be the waiting time of an EF packet, then the following relationship exists between S_1 and W_1.

$$S_1 = W_1 + \frac{1}{\mu_1}. \tag{1}$$

In order to obtain a formula for W_1, let us define variables Q_1, the expected number of EF packets in EF buffer and R, the expected value of the residual service time of a packet in the server. When a server operates based on SP, the mean waiting time of EF packets can be obtained by using the mean waiting time of a customer for a single class M/G/1 queuing system with vacation, where vacations occur when a server visits BBE and BE buffers in case there is no packet in EF buffer. Therefore, we obtain the following result.

$$W_1 = \frac{Q_1}{\mu_1} + R. \tag{2}$$

From Little's formula [6], (2) is rewritten by

$$W_1 = \rho_1 W_1 + R. \tag{3}$$

If we arrange (3) with respect to W_1, we obtain (4)

$$W_1 = \frac{R}{(1-\rho_1)}. \tag{4}$$

When a packet arrival process follows Poisson distribution, we can apply the PASTA (Poisson arrival see time average) property in the computation of R [7]. From [8], we obtain (5).

$$R = \sum_{k=1}^{3} \lambda_k \frac{E[\tau_k{}^2]}{2}. \tag{5}$$

In (5), τ_k, $k=1,2$, and 3, is the service time of EF, BBE, and BE packet, respectively. $E[\tau_k{}^2]$ is the second moment of τ_k, which may be represented by (6).

$$E[\tau_k^2] = \sigma_k^2 + \frac{1}{\mu_k^2}, k = 1,2,3. \tag{6}$$

Finally, we can obtain the mean value of waiting time for an EF packet in an EF buffer, which is given in (7).

$$W_1 = \frac{R}{(1-\rho_1)}. \qquad (7)$$

In (7), it is assumed that $\rho_1<1$. Note that we can obtain the sojourn time of an EF packet in the system from (1) and (7).

Now let us compute the mean waiting time W_2 for the BBE packets. For the BBE packets, the delay is contributed by the following four components: First, delay due to the mean residual service time of BBE packet under the service. Second, delay due to the packets of EF class when BBE packet arrives. Third, delay due to BBE packets that have been waiting in BBE buffer when a BBE packet arrives to that buffer. Fourth, delay caused by serving all the packets of EF class that will arrive during the total waiting time in BBE buffer. Summing up the four factors, we have the following formula for the waiting time of a packet that enters the BBE buffer.

$$W_2 = R + \frac{X_1}{\mu_1} + \frac{X_2}{\mu_2} + \frac{Z_1}{\mu_1}, \qquad (8)$$

where X_1 and X_2 is the mean number of packets of class 1(EF class) and class 2 (BBE class) in each queue and Z_1 is the mean number of EF packets that arrive while the BBE packets wait in their buffer.

From Little's formula, we obtain $Z_1=\lambda_1 W_2$, $X_1=\lambda_1 W_1$, and $X_2=\lambda_2 W_2$. Then, from (7) and (8), we obtain (9).

$$W_2 = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}, \qquad (9)$$

where $\rho_1+\rho_2<1$.

By repetition, we obtain a formula for the mean waiting time W_3 for the BE packets, which is given in (10).

$$W_3 = \frac{R}{(1-\rho_1-\rho_2)(1-\rho_1-\rho_2-\rho_3)}, \qquad (10)$$

where $\rho_1+\rho_2+\rho_3<1$.

In order to investigate the degree of the degradation of the delay performance of BBE or BE traffic under the DiffServ network compared with that of the conventional Best Effort network, let us compare the performance of mean waiting time for the BBE or BE traffic under two different buffering and scheduling schemes, the SP scheme with separate buffers and the FIFO (First In First Out) scheme with shared buffer. The

performance of the mean waiting time for the BBE or BE traffic is given above, and let us compare those results with that of FIFO scheme.

Under the FIFO scheme, all the traffic from EF, BBE, and BE enter a shared buffer, and they are treated with equal priority. Therefore, we can express the mean waiting time W_{FIFO} for BBE or BE traffic as well as the EF in a single formula from the M/G/1 queue with no priority [6], which is given in (11).

$$W_{FIFO} = \frac{1}{\lambda}\left(\frac{\rho^2}{1-\rho} \times \frac{1+C_S^2}{2}\right). \qquad (11)$$

In (11) $\rho=\rho_1+\rho_2+\rho_3$ and $\lambda=\lambda_1+\lambda_2+\lambda_3$. C_S^2 is the squared coefficient of variation for service time of a packet, and it is defined by $C_S^2=\sigma^2\mu^2$, where σ^2 and μ is the variance and mean value of the service time for the packet, respectively. Note that $\mu=(\mu_1\rho_1+\mu_2\rho_2+\mu_3\rho_3)/\rho$. On the other hand, $\sigma^2=\sigma_2^2=\sigma_3^2$ if we assume that the distribution of packet size of BBE and BE class traffic is the same and the packet size of voice traffic is constant. In case the distribution and packet size of BBE and BE class traffic is different we can compute σ^2 from some manipulation, which is trivial.

Ⅳ. Numerical Experiments and Results

It is usual that the primary objective of an SP scheduling scheme is set to the guarantee of the delay for the highest (EF) class of traffic. However, this work tries to investigate the delay performance of all the classes for the various combination of the traffic load. First, let us investigate the delay of the highest class. After that, let us investigate the delay of the lower class, which is given in the subsequent sections.

4.1 Delay Performance of Highest Class Traffic under SP Scheme

The delay of voice packet can be computed from (5) and (7). Note that the delay of voice packet depends on the profile of the corresponding traffic such as the class 2 and class 3 traffic as well as the class 1(voice) traffic itself. This illustrates that the source of delay in voice packets is composed of two factors as we have argued before: the delay incurred by the voice packets at the same buffer that has arrived before the observed packet at hand and the residual transfer time of lower priority packets. This implicates that the delay of voice packet can be efficiently suppressed to a desired value if one control the ingress of the traffic other than the voice traffic to the link shared by those traffic classes. In order to determine the desired value of the delay for voice packet, one has to know the component of the delay in an end-to-end element of the network.

However, one has to note that, if the type of voice codec and the transmission media are determined, the delay elements of a voice packet in an end-to-end path of the speaker and receiver are almost fixed except the queuing delays incurred in a series of routers through which a voice packet passes. In addition, our purpose in this work is not focused on the anatomization of the delay element in Internet, which is presented at author's other work [9]. Therefore, we only investigate the delay of voice packet in a router under various type of traffic load. For the simplicity in computation, let us assume some mixture of offered load, which is given in Table1.

Table1. Mixture of traffic load Load

Load type	ρ_1	ρ_2	ρ_3	ρ
A	0.1	0.4	0.4	0.9
B	0.3	0.3	0.3	0.9
C	0.5	0.2	0.2	0.9
D	0.7	0.1	0.1	0.9

Note that, in all the cases, the total offered load ρ to the system is assumed to be 0.9, so that the system is relatively heavy. The offered load of the lower two classes are assumed to be the same. We want to investigate how much the delay of voice traffic increases as the total offered load increases if we apply an SP scheme to the voice traffic. This is the first purpose of this experiment. For the comparison of the performance under different link capacity we assumed three cases: 1Mbps, 10Mbps, and 100Mbps.

In order to compare the delay performance let us assume some parameter for the source traffic. The data traffic from BBE and BE classes have packets of variable size with mean *500bytes* and standard deviation of *500 bytes* (header included). It is assumed that the size of voice packet is fixed and is equal to *216bytes* (*160bytes* of payload and *56bytes* of headers) [10].

Fig.1 illustrates the delay performance of voice traffic under various combination of the offered load of lower classes for the three cases of link capacities, 1Mbps, 10Mbps, and 100Mbps. From Fig.1 we found that it is futile to worry about queuing delay of a voice packet for a bandwidth in the order of 100Mbps, which is in the order of a few tens of microseconds. It is also found that the queuing delay of a voice packet over the link of 10Mbps is sufficiently small for all the combination of load type considered in this work. The impact of the offered load of voice traffic to the delay of voice packet is increasing as the offered load of the voice traffic increases, which is expected. The rate of increase is not so steep because the offered load of two data traffic becomes lower as the load of voice traffic increases, which we have assumed in Table1.

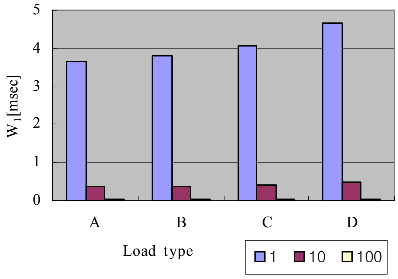

Fig.1. Link capacities and waiting time of voice traffic.

Now let us investigate the effect of the packet sizes of the lower class traffic to the delay of the voice traffic under the assumed non-preemptive SP scheduling scheme. It is very likely that the longer the service time of lower class traffic, the higher the residual service time of the lower class traffic. This causes longer delay to the voice packets. We assumed three cases for packet size, that is, the data traffic from BBE and BE classes have packets of variable size with same mean value of *500bytes, 1000bytes,* and *1,500bytes*. The standard deviation is assumed to be *500 bytes* in all cases. The size of voice packet is fixed to be *216bytes*. The link capacity is assumed to be 1Mbps. Fig.2 illustrates the mean waiting time of voice packets under the different load condition of voice traffic.

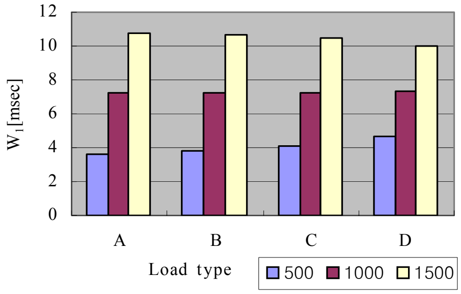

Fig.2. Waiting time of the voice traffic for different packet sizes of data traffic.

As it is expected, the mean waiting time of voice packets increases as the size of data packet increases. This result implies that the queuing delay of voice packet is heavily dependent on the size and residual service time of data packets.

In order to investigate the effect of the residual service time of the packets from the lower class traffic to the waiting time of the voice packet, let us compare the delay of two different systems: One is a shared link with three-class queuing system and SP scheme, and the other is a dedicated and separate link for voice traffic only. The mean waiting time of the SP scheme with three-class queuing system is represented by (7), whereas the delay of a single class queuing system model is given in (11), where the offered load is contributed by only voice packets, so that $\rho=\rho_1$ in (11). The data traffic from BBE and BE classes have packets of variable size with same mean value of *1000bytes*, and the standard deviation is assumed to be *500 bytes*. The size of voice packet is fixed to be *216bytes,* which implies that $C_S^2=0$ in (11). The link capacity is assumed to be 1Mbps. Fig.3 illustrates the mean waiting time of voice packets under the different buffering schemes.

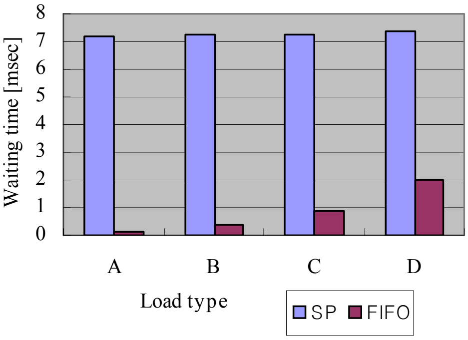

Fig.3. Waiting time of the voice traffic for different buffering schemes.

From Fig.3 we can find that the mean waiting time of voice packet for the single buffer system with FIFO service scheme follows the pattern of conventional exponential increase, whereas the mean waiting time under the multi-buffer system with SP scheme exhibits a smooth increase with much greater delay than that of FIFO system. This phenomenon stems from the mixing nature of the traffic between the three classes of traffic. Anyway, we can find from this result that the delay performance of voice packet for the multi-class queuing system degrades due to sharing of the link capacity with the bursty data traffic, even though a strict priority service scheme is provided to the voice traffic. This result illustrates that the following conventional wisdom of isolation law [11] does not hold at any condition in a network. Rather, we have to say that *Strict priority scheduling does not isolates voice traffic from non-voice traffic*. This implies that the network operator has to be careful in mixing the limited bandwidth (especially, when the bandwidth is not greater than a few mega bits per second) between different classes of traffic if one wants to provide a strict delay performance to the highest class, the voice traffic. Even though the bandwidth becomes thicker, the isolation law would not hold.

4.2 Delay Performance of Lower Class Traffic in SP Scheme

The purpose of this subsection is to estimate the delay performance of the packets of the other classes with respect to the delay performance of the packets of the highest class. In order to compare the delay performance between different classes, let us define a relative value for each class. From (7) and (9), the relative value for the delay of the BBE packet with respect to the EF packet is represented by (12).

$$\frac{W_2}{W_1} = \frac{1}{1-\rho_1-\rho_2}. \qquad (12)$$

Note from (12) that the relative value of the mean waiting time of class2 packet with respect to that of class 1 packet is governed by just the offered load of class1 and 2 packets. Therefore, if the network is dimensioned to satisfy the mean waiting time of the class1 packet, we can easily estimate the mean waiting time of the class2 packets from tuning the offered load of class1 and 2 packets. In other words, if the mean offered load and the delay target of class1 packet is known a priori, one can estimate the maximum allowable offered load of the class2 packet under the condition that the delay target of class1 packet is met, which is very important in the provisioning of the network load profile.

On the other hand, both the BBE and BE traffic is not sensitive to the delay. However, the customers of BBE traffic will be more concerned with the delay performance of the service than that of the BE traffic. Therefore, let us assume some limit on the delay of class2 customers, and observe the delay of class3 customers compared with that of class2 customers. The relative value for the delay of the BE packet with respect to the BBE packet is given by (13).

$$\frac{W_3}{W_2} = \frac{1-\rho_1}{(1-\rho_1-\rho_2-\rho_3)}. \qquad (13)$$

Usually, the delay target of the BE class packet is undeclared. However, one can use (13) in the estimation of the possible mean delay of the BE packet with respect to the mean delay of BBE class packet under the given load profile of class 1, 2, and 3 packets. In order to be more specific, let us assume that the EF class is mapped to a voice traffic, whereas the BBE and BE class is mapped to web browsing and the ftp traffic, respectively. The QoS objectives of all the classes in the proposed mathematical model are defined in terms of delay experienced by a packet at a node. For voice traffic a nodal delay of 4ms is assumed as a satisfactory toll quality [12]. For web browsing traffic, there exists no concrete reference for the objectives of the delay, yet. However, about 8 seconds of round trip delay is the margin (8 seconds rule) for the tolerable delay, so that one-way delay is assumed to be about 4 seconds. Let us observe its value by varying the offered load of BBE traffic. Finally, no SLA (Service Level Agreement) is required for the BE class, so that no limit on the delay of BE traffic is set by users or network. So, we don't have to care about the delay of BE traffic.

However, let us investigate the expected value of BE traffic under the various load conditions in the network. Note, from (5) and (7), that the waiting time of the highest class (W_1) depends on the offered load of all the corresponding classes as well as some other parameters such as the second moments of the service time of each packet.

In [12], a detailed procedure for the determination of the load map between the voice and data traffic under the defined delay requirement of voice packet is given for two-class system. Via numerical experiment, the authors presented the design area for the network link. Therefore, we assume that W_1 can be determined by a combination of the offered load of the corresponding classes.

Fig.4 illustrates the relative value for the mean waiting time of the BBE packet with respect to that of the EF packet when the offered load of class2 traffic is varied. We assumed two cases of offered load for the EF traffic, whose offered load is assumed to be 0.2 or 0.4. We also assumed that the sum of the offered load of the two class of traffic is not greater than 0.8, so that $\rho_1+\rho_2$ is not greater than *0.8*. Note from (12) that W_2/W_1 diverges to infinity as $\rho_1+\rho_2$ approaches to 1.

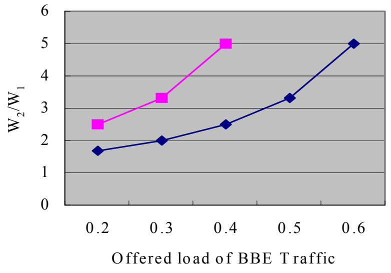

Fig.4. Relative value of waiting time $\frac{W_2}{W_1}$.

As we can find from Fig.4, the mean waiting time experienced by class2 traffic is five times that of class1 traffic when the sum of the offered load of two classes is 0.8. In the real field, the limit on the operational region of the system load is usually not greater than 0.5 or 0.6, so that bandwidth provisioning is newly carried out when the total offered load of a node is greater than 0.6 in an IP network [13]. If we assume that the limit on the operational region of the system load is 0.6, we can observe that the delay of class2 traffic is 2.5 times greater than that of the class1 traffic, which is located as a very satisfactory level to the BBE traffic. Note also that W_2/W_1 increases vary rapidly when the offered load of EF traffic increases.

Fig.5 illustrates the relative value for the mean waiting time of the BE packet with respect to that of BBE packet when the offered load of class3 traffic is varied. We assumed that the offered load ρ_1 of EF class is fixed to be 0.2, because it is estimated that the offered load of EF traffic is usually not so high in a DiffServ network [12]. Let us assume that the offered load ρ_2 of the BBE traffic is assumed to be 0.2 or 0.4. We also assume that the sum of the offered load from the total class of traffic is not greater than 0.9, so that $\rho_1+\rho_2+\rho_3$ is not greater than 0.9.

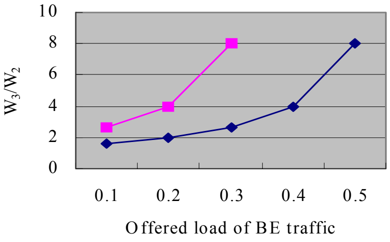

Fig.5. Relative value of waiting time $\frac{W_3}{W_2}$.

Note from Fig.5 that the relative value W_3/W_2 for the mean waiting time of the BE packet with respect to that of BBE packet increases exponentially with the increase of the offered load of class3 traffic. Note also that W_3/W_2 increases exponentially with the increase of the offered load of class2 traffic for a given offered load of class3 traffic. Finally, from Fig.5, we could find that W_3 is 8 times greater than W_2 when the offered load of class 3 traffic is 0.5 under the condition that $\rho_1=\rho_2=0.2$, the worst-case combination of the offered load in this work.

5.3 Comparison of SP scheme and FIFO Scheme

Let us assume that the bandwidth of an output port is *1Mbps*. The source traffic parameters are assumed to be the same as in subsection 1 of this section. Fig.6 illustrates the mean waiting time of the BBE (W_2) and BE (W_3) packets for the SP scheme when the offered loads of BBE and BE traffic are varied. The offered load of EF traffic is assumed to be 0.2. In the figure, the mean waiting time of the BBE and BE packets for the FIFO scheme (W_{FIFO} *is denoted as* W_F *due to some problem in the presentation*) is also shown.

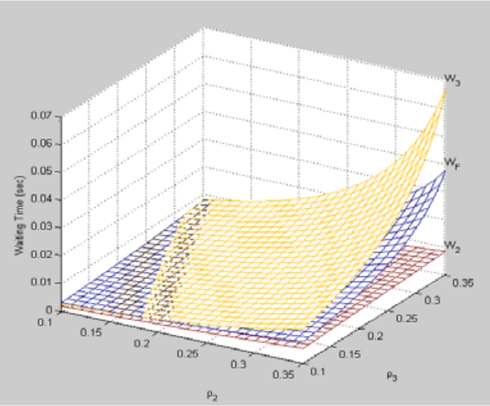

Fig.6. Mean waiting time of the BBE and BE packets

Note that the expected delay of BBE and BE packets under the FIFO scheme is the same, because packets from both classes are buffered to the same queue and they are treated without priority. Note also that W_2 is smaller than W_{FIFO}, which is expected. Therefore, SP is friendly to class2 traffic compared with the current best effort scheme. Only class3 traffic in SP scheme suffers from some additional delay compared with FIFO scheme, because packets in class 3 traffic have to wait until the buffers of class 1 and class 2 traffic are vacant before they are being served.

V. Conclusions

In this work we investigated the delay performance of SP scheduling scheme for the DiffServ architecture in IP network. By using the queuing model, we analyzed and investigated the delay performance of the highest and lower class traffic for the three-class SP scheduling algorithm.

Via numerical experiment, we investigated the delay performance of the class1 traffic, the relative delay performance of class2 and class3 traffic as well as that of class1 and class2 traffic for a diverse set of traffic parameters.

From the experiment, the following conclusions are obtained. First, the delay performance of highest-class traffic is very satisfactory if one adopts an SP scheme for the voice traffic and when the link capacity is in the order of tens of Mbps or higher, under which condition it is futile to worry about the delay performance for the voice traffic. Note however, that the delay performance of highest-class traffic for an SP scheme is dependent on the offered load of the lower class traffic. Therefore, contrary to the conventional wisdom, SP scheme can not definitely isolate the highest-class traffic from the lower classes of traffic, and care must be taken in provisioning the bandwidth resource to lower classes of traffic in a multi-service network.

Second, the delay performance of lower classes of traffic is sufficiently acceptable to the users of data traffic such as the Web browsing and ftp under the low load of EF traffic, and we can argue that SP is not so negligent in the service of lower class traffic. Therefore, we can say that the vindication of SP scheduling scheme is reasonable, and SP is not negligent in caring for the data traffic, especially the class 2 traffic.

In addition, from the comparison of the delay performance of SP scheme and FIFO scheme, we found that the delay performance of the class 2 traffic in SP scheme is better than that of the FIFO scheme when the offered loads of class 2 and class 3 traffic increase, which illustrates that SP scheme is not detrimental to class 2 traffic if a DiffServ architecture is introduced to a multi-service network. The results can be utilized and applied to the design of multi-service network of the future Internet as a baseline for the network resource dimensioning.

The future research area includes the investigation of the relationship between the waiting times of the lower class traffic with respect to the different values of the required target value of the traffic from highest class.

References

[1] G. Armitage, Quality of Services in IP networks, Foundations for a multi-service Internet, Macmillan Technical Publishing, USA, 2000.

[2] M.J. Karam and F.A. Tobagi, Analysis of delay and delay jitter of voice traffic in the Internet, Computer Networks 40 (2002) 711-726.

[3] K. Yamauchi et al., Performance evaluation of a hardware router with QoS control capabilities, Technical Report of IEICE, NS2001-258 (2002-03).

[4] T. Minagawa and T. Kitami, Packet size based dynamic scheduling for assured services in Diffserv network, Tr. of IEICE, Vol.J87-B, No.2, pp.181-189, February 2004.

[5] Hoon Lee Ed., Advances in queuing systems, Lecture notes of Graduate School of IT, Changwon National University, 2001.

[6] L. Kleinrock, Queueing systems, Vol.1; Theory, John Wiley & Sons, 1975.

[7] G. Bolch, S. Greiner, H. de Meer, and K. Trivedi, Queueing networks and Markov chains, John Wiley & Sons, 1998.

[8] P. Nain and D. Towsley, Performance evaluation of computer systems: Lecture notes, 1995.

[9] Hoon Lee and Yong-Chang Baek, Anatomy of Delay for Voice Service in NGN, Proceedings of 2003 IEEK Fall Conference, November 29, 2003.

[10] Y. Kim, Hoon Lee, and K. Lee, Dimensioning Next Generation Networks for Guaranteed Voice Services, Journal of IEEK, Vol.40, TC-12, December 2003.

[11] E. Chi, M. Fu, and J. Walrand, Proactive resource provisioning, Computer Communications 27 (2004) 1174-1182.

[12] Y. Kim, Hoon Lee, and K. Lee, Dimensioning links for NGN VoIP networks, Journal of KICS, Vol.28, No.8B, August 2003.

[13] T. Sugiyama, Bandwidth provisioning for the uncertain WAN link, NIKKEI COMMUNICATIONS, April 12, 2004.

Hoon Lee 정회원

February 1984 B.E. in Electronics from Kyungpook National University

February 1986 M.E. in Communications from Kyungpook National University

March 1996 Ph.D. in Electrical & Communication Engineering from Tohoku University, Japan

February 1986~February 2001 KT R&D Group

March 2001~March 2005 Assistant Professor of Changwon National University

March 2005~ Associate Professor of Changwon National University

Research fields : Teletraffic engineering, network design, performance analysis and provision of QoS, and pricing for high speed telecommunication networks.

Dr. Lee is a member of IEEE, KICS and IEEK.