Appl. Math. Inf. Sci. 6 No. 2S pp. 453S-461S (2012)

Applied Mathematics & Information Sciences
An International Journal

A Proportional Differentiation Model Based on Service Level

Ki-Seok Choi
Department of Industrial & Management Engineering, Hankuk University of Foreign Studies, Yongin, 449-791, Korea
Corresponding author: Email: kchoi96@hufs.ac.kr

Received June 22, 2010; Revised March 21, 2011; Accepted 11 June 2011
Published online: 1 January 2012

Abstract: In this paper, we consider a proportional differentiation model based on the performance metric of service level while most existing researches on the proportional differentiation use other metrics such as average waiting time and packet loss probability. We use a metric called unfill rate to measure the service level of the traffic classes. As an implementation method of the proportional differentiation model, we suggest a time-dependent priority scheduling which expresses the priority of the traffic class as a linear function. We find out how to adjust the priority function parameters to achieve the intended proportional differentiation through the asymptotic analysis of the queue waiting time distribution in a two-class queueing system with a Poisson arrival process. Numerical experiments show that the scheduling method is effective for implementing the proportional differentiation model especially when the required service levels are high.

Keywords: Proportional Differentiation, Differentiated Service, Time-Dependent Priority

1 Introduction

Proportional differentiation is a type of relative differentiated services approach for Internet applications. The relative differentiated services approach groups the network traffic into several classes and orders them based on their priority. For a selected performance measure, the network operator tries to make sure that a higher-priority class experiences better service than a lower-priority one in terms of the measure. With proportional differentiation, performance measures should be proportional to the differentiation parameters. Suppose m_i is a performance measure for class i. The proportional differentiation states that

$$\frac{m_i}{m_j} = \frac{\delta_i}{\delta_j} \tag{1.1}$$

where δ_i is the differentiation parameter for class i.

There are several literature on proportional differentiation models. Dovrolis and Ramanathan [1] compare various approaches for differentiated services and describe scheduling methods for several proportional differentiation models. Dovrolis et al. [2] consider delay differentiation, which uses the average queueing delay as the performance measure, and how the proportional delay differentiation can be approximated by several schedulers. Leung et al. [3] study a two-class case where both arrival processes are Poisson. They find several properties on scheduling parameters for proportional delay differentiation between the two classes.

Most studies on proportional differentiation and their analytical results consider a differentiation based on average waiting time, which is called proportional delay differentiation. Although average waiting time is the most common metric in both theory and practice, there are other interesting performance measures in certain circumstances. For example, in call centers, one of the important performance measures is service level which is defined as a proportion of customer calls answered in a specific time referred to as acceptable waiting time (AWT) [4]. This kind of performance measure

is hard to analyze because it requires information not just on the average of waiting time but on its distribution. In this paper, we consider a performance measure called *unfill rate* $u_i(t)$, which is defined to be the probability that the total waiting time R^i of class i is longer than AWT t;

$$u_i(t) := P\{R^i > t\} . \quad (1.2)$$

Since the unfill rate is defined as the probability that the waiting time is longer than AWT t, one can get the service level by taking out the unfill rate from 100%. Note that the unfill rate decreases as AWT t increases. In other words, when the customers are willing to wait for a longer time (larger t), the unfill rate (service level) becomes lower (higher). Using the performance measure of the unfill rate, we consider a proportional differentiation model based on service level and suggest a scheduling discipline which effectively implements it.

In the next section, we describe the scheduling discipline used in this paper to implement the proportional differentiation. Section 3 shows analytical results on the performance measure and suggests how to use the scheduling discipline to achieve proportional differentiation between classes. After giving numerical examples illustrating how effectively the scheduling discipline works in Section 4, we end this paper with concluding remarks in Section 5.

2 Time-Dependent Priority

To implement proportional differentiation, many researchers consider the time-dependent priority (TDP) scheduler, which is a non-preemptive packet scheduling discipline increasing the priority of a packet with its waiting time. There are many versions of TDP scheduling and we refer the reader to Essafi and Bolch [5] for details on them. In this paper, we use a TDP scheduler whose priority function $q_i(t)$ is defined as follows; if a tagged class i packet arrives at time τ_i, then its priority at time $t \geq \tau_i$ is

$$q_i(t) = r_i + t - \tau_i . \quad (2.1)$$

Parameter r_i determines the priority between classes and a higher-priority class is assigned with a larger r_i.

Although it is from a higher-priority class, a packet does not have higher priority over every packet from a lower-priority class. Since the priority increases with the elapsed time in the system, lower-priority class packets have higher priority than higher-priority class packets which have not spent much time waiting in the system. Figure 2.1 shows a plot of the priority function (2.1). Assume that class i has a higher priority than j ($r_i > r_j$). If a class i packet arrives at τ_i, it always has a higher priority over a class j packet which has not arrived before it ($\tau_j \geq \tau_i$). In addition, its priority $q_i(t)$ is higher than a class j packet which has arrived before it if $q_j(t) < q_i(t)$, i.e., $\tau_j > \tau_i - (r_i - r_j)$.

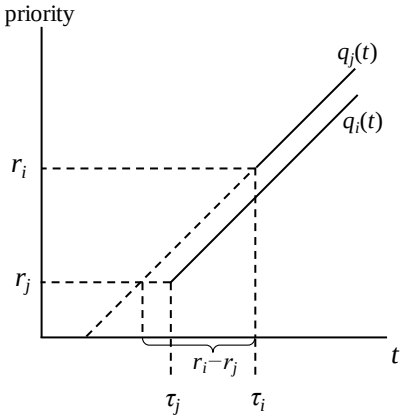


Figure 2.1: Priority of a class i packet is set to r_i when it arrives at τ_i then increases linearly with time t.

3 Proportional Differentiation

In this section, we consider a two-class queueing system with an i.i.d. interarrival time process $\{A_n\}$ and an i.i.d. service time process $\{B_n\}$. Class 1 is assumed to have higher priority than class 2 and their proportional differentiation is stated by the ratio of the unfill rates:

$$\frac{u_1(t)}{u_2(t)} = \frac{\delta_1}{\delta_2} \quad (3.1)$$

To implement the unfill rate differentiation, we consider the TDP priority function (3.1) with r1 > r2. With the priority function, a class 1 packet has higher priority than class 2 packets whose arriving time τ2 is earlier than the class 1 packet arrival time τ1 but within an interval of $\Delta_r := r_1 - r_2$ (i.e., $\tau_1 - \Delta_r < \tau_2 < \tau_1$).

In order to analyze the waiting time distribution, first we compare the service orders under the TDP scheduling with the packet arrival orders. Figure 3.1 illustrates that class 1 packets have higher priority over a number of class 2 packets which have arrived before them. Under the TDP scheduling with priority function (3.1), the class 1 packet which has arrived at τ_1 (called "τ_1-class 1 packet") could be served before the class 2 packet which has arrived between $\tau_1 - \Delta_r$ and τ_1 (called "$[\tau_1-\Delta_r,\tau_1]$-class 2 packet"). Since the TDP scheduling is non-preemptive, the class 1 packet cannot be served before the class 2 packet whose service has already started. In other words, the τ_1-

class 1 packet cannot "catch up" the $[\tau_1-\Delta_r,\tau_1]$-class 2 packet which starts to be served before τ_1. As for those class 2 packets having arrived before $\tau_1 - \Delta_r$, the τ_1-class 1 packet has a lower priority and cannot catch them up at all.

Compared with the well-known First-In-First-Out (FIFO) policy where the service order is the same as the arrival order, the TDP scheduling can reduce the waiting time of a class 1 packet by as much as the interarrival time whenever it catches up a class 2 packet. Let N denote the number of the $[\tau_1-\Delta_r,\tau_1]$-class 2 packets. Hence, N is the maximum number of the class 2 packets that the τ_1-class 1 packet can catch up. In the example of Figure 3.1, N is equal to 2 (the number of white circles between $\tau_1 - \Delta_r$ and τ_1). Under FIFO policy, the τ_1-class 1 would be served at the 5-th position among the 5 packets in Figure 3.1 because the packets are served according their arrival orders.

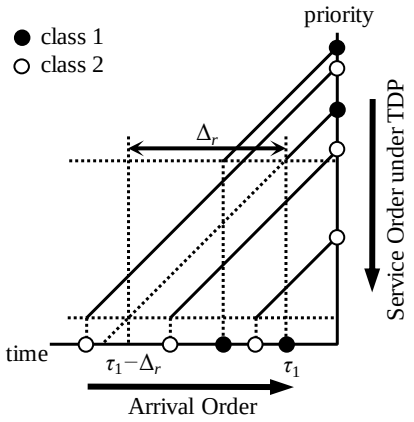


Figure 3.1: Under TDP policy the service order can be different from the arrival order. The class 1 packet arriving at τ_1 would be served before the class 2 packet whose arrival time is in $[\tau_1 - \Delta_r, \tau_1]$ unless it starts to be served before τ_1.

Instead, under the TDP scheduling, it would catch up two class-2 packets and be served at the 3-rd position if none of those class-2 packets do not start to be served until τ_1. The difference of waiting time of the 3-rd and 5-th positions can be as large as twice the interarrival time. Thus, we can see that the TDP scheduling can reduce the waiting time of the τ_1-class 1 packet by $\sum_{n=1}^{N} A_n$ at most.

If we make further assumptions regarding the input process that the packet arrival follows a Poisson process and the percentage of the class 1 traffic is $0 < \theta < 1$, then N has a Poisson distribution with mean $(1 - \theta)\,\Delta_r$. And, we have the following result on the unfill rate of class 1. The proof can be found in Appendix A.

Theorem 1 *If there exists $\gamma > 0$ satisfying*

$$\phi_B(\gamma) + \phi_A(-\gamma) = 0 \qquad (3.2)$$

where ϕ_X denotes the cumulant generating function: $\phi_X(\nu) := \log \mathrm{E}[e^{\nu X}]$, then for a constant C which does not depend on Δ_r or θ

$$\lim_{t\to\infty} e^{\gamma t} u_1(t) = Ce^{-\gamma(1-\theta)\Delta_r}. \qquad (3.3)$$

Theorem 1 suggests that the unfill rate for high-priority customers can be approximated as follows;

$$u_1(t) \approx Ce^{-\gamma[t+(1-\theta)\Delta_r]} \qquad (3.4)$$

We can use the similar arguments as for Theorem 1 to obtain a result on the low-priority class as well. Let M denote the number of class 1 packets which arrive during time slot $[\tau_2, \tau_2 + \Delta_r]$. This number represents how many class-1 packets can catch up the class-2 packet which has arrived at τ_2 (called "τ_2-class 2 packet"). As one class-1 packet catches up the τ_2-class 2 packet, the waiting time of the τ_2-class 2 packet increases by as much as the class-1 packet service time. Thus, under the TDP scheduling, the waiting time of the τ_2-class 2 packet can be increased by $\sum_{n=1}^{M} B_n$ at most.

Under the assumptions that the packet arrival follows a Poisson process and the percentage of the class 1 traffic is θ, M has a Poisson distribution with mean $\theta\Delta_r$. And, we have the following result on the unfill rate of class 2. Its proof is given in Appendix B.

Theorem 2 *If there exists $\gamma > 0$ satisfying (3.2) and $\phi_B(2\gamma) < \infty$, then with the same constant C as in (3.3)*

$$\lim_{t\to\infty} e^{\gamma t} u_2(t) = Ce^{\gamma\theta\Delta_r}. \qquad (3.5)$$

Theorem 2 suggests that the unfill rate for low-priority customers can be approximated as follows;

$$u_2(t) \approx Ce^{-\gamma(t-\theta\Delta_r)}. \qquad (3.6)$$

Using the above asymptotic results on the unfill rates, we examine how the network service provider can control the service level differentiation between the two traffic classes. From Theorem 1 and 2, we can easily get the following result regarding the proportion of the unfill rates.

Corollary 1 *If there exists $\gamma > 0$ satisfying (3.2), then*

$$\lim_{t\to\infty} \frac{u_1(t)}{u_2(t)} = e^{-\gamma\Delta_r}. \qquad (3.7)$$

The right-hand side of (3.7), called the *asymptotic ratio*, gives a hint on how to proportionally differentiate between the high- and low-priority classes. To achieve the proportional differentiation given by (3.1), the priority function parameter r_1

and r_2 can be selected such that their difference $\Delta_r = r_1 - r_2$ satisfies the following equation:

$$e^{-\gamma\Delta_r} = \frac{\delta_1}{\delta_2}, \qquad (3.8)$$

in other words

$$\Delta_r = -\frac{1}{\gamma}\ln\frac{\delta_1}{\delta_2}. \qquad (3.9)$$

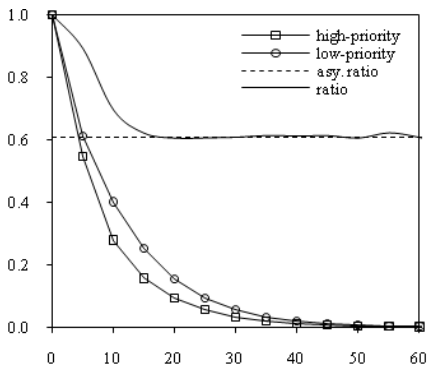
Since the asymptotic ratio in (3.7) is no larger than 1, the unfill rate of high-priority class $u_1(t)$, is no greater than that of low-priority class $u_2(t)$. It becomes 1 only if $\Delta_r = 0$, which makes $q_1(t) = q_2(t)$ and virtually no priority difference between the classes. Note that the asymptotic ratio $e^{-\gamma\Delta_r}$ does not include θ, the percentage of the high-priority class traffic. Thus, Corollary 1 implies that the percentage of the high- and low-priority class traffics has little influence on the proportion of their unfill rates when the AWT t is large enough and, in other words, a high service level is required.

4 Numerical Experiments

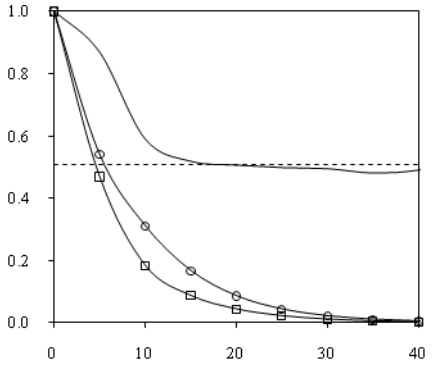
In this section, we conduct numerical experiments to verify the results on the proportional differentiation discussed in the previous section. The total traffic arrival rate λ is assumed to be 0.9. We try several service time distributions with mean 1. The priority parameter gap Δ_r is set to 5 and we consider two different combinations of the traffic classes (θ = 0.3, 0.8). Figure 4.1 and 4.2 plot the unfill rates observed by simulation for both classes and their ratio as well as the asymptotic ratio defined in (3.7) for θ = 0.3 and 0.8, respectively.

They show that the ratio of the unfill rates $u_1(t)/u_2(t)$ converges to the asymptotic ratio $e^{-\gamma\Delta_r}$ as AWT t increases. In every case, the gap between the observed and asymptotic ratios becomes smaller when AWT t is long enough to make the unfill rates under 20%. The numerical results also show that the approximation of $u_1(t)/u_2(t)$ by the asymptotic ratio works better for a service time with a larger squared coefficient of variation c_B^2. When the service time has small variation, e.g., $c_B^2 < 1$ as in (b) of Figure 4.1 and 4.2, it is rarely observed that the waiting time is longer than AWT, which makes inefficient to verify an asymptotic result through simulation.

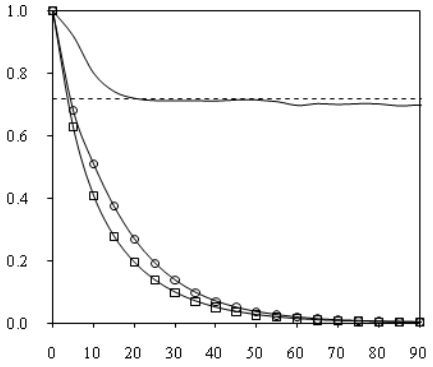
The unfill rates of both high- and low-priority classes increase as θ increases from 0.3 (Figure 4.1) to 0.8 (Figure 4.2). But, their ratio seems to remain at the same level close to the asymptotic ratio when AWT t is large. As Corollary 1 implies, the



(a) Exponential Service Time



(b) Gamma Service Time (c_B^2 = 0.5)

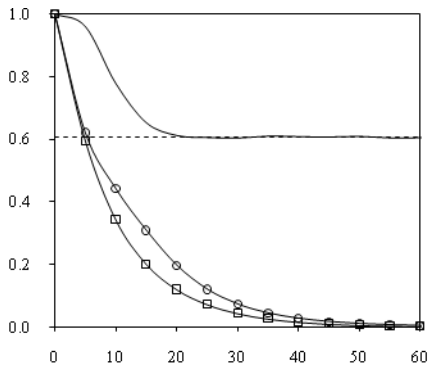


(c) Gamma Service Time (c_B^2 = 2)

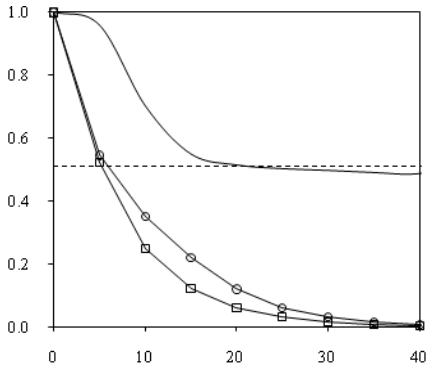
Figure 4.1: Ratio of unfill rates (θ = 0.3)

percentage of high- and low-priority classes has little effect on the ratio of the service levels when the service levels are high enough.

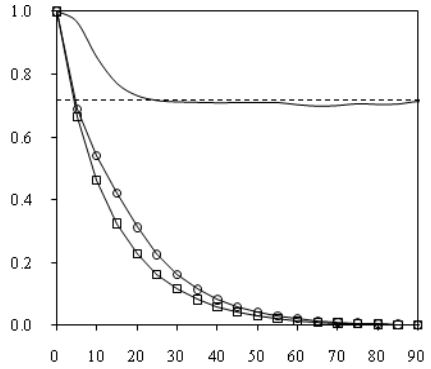
Corollary 1 also suggests that the network service provider can control the relative service levels between traffic classes through parameter Δ_r. We now conduct another numerical experiment and demonstrate how well the asymptotic ratio approximates the actual ratio of the unfill rates for different values of Δ_r. The service time is assumed to have an exponential distribution and other parameters are set to the same as in Figure 4.1 and



(a) Exponential Service Time



(b) Gamma Service Time ($c_B^2 = 0.5$)



(c) Gamma Service Time ($c_B^2 = 2$)

Figure 4.2: Ratio of unfill rates ($\theta = 0.8$)

4.2. Table 4.1 lists the experiment results such as $u_1(t)$, $u_2(t)$, and their ratio for two different AWTs (t = 20, 30). We choose AWTs such that the unfill rates for both classes are less than 20%. Table 4.1 shows that in most cases the actual ratio $u_1(t)/u_2(t)$ observed by simulation is very close to the asymptotic ratio $e^{-\gamma\Delta_r}$ suggested in Corollary 1. With a shorter AWT (t = 20), the gap between the actual and asymptotic ratios gets bigger as Δ_r increases (γ = 0.1 in this example). On the other hand, as t increases (t = 30), the asymptotic ratio approximates more closely the proportional

Table 4.1: Actual ratio of the unfill rates observed by simulation along with the asymptotic ratio

	t = 20			t = 30			
Δ_r	$u_1(t)$	$u_2(t)$	$u_1(t)/u_2(t)$	$u_1(t)$	$u_2(t)$	$u_1(t)/u_2(t)$	$e^{-\gamma\Delta_r}$
1	0.128	0.141	0.911	0.045	0.049	0.916	0.905
2	0.119	0.145	0.821	0.042	0.051	0.825	0.819
3	0.111	0.149	0.741	0.039	0.052	0.743	0.741
5	0.096	0.158	0.603	0.034	0.056	0.601	0.607
7	0.084	0.167	0.501	0.029	0.060	0.490	0.495
10	0.071	0.175	0.406	0.024	0.066	0.361	0.368

differentiation of the unfill rates. In order to make the unfill rate of the high-priority class as low as, say, 60% of that of the low-priority class, the service provider can find a proper value of Δ_r using (3.9). In this example, $-\ln(u_1(t)/u_2(t))/\gamma = -\ln(0.6)/0.1 \approx 5.1$ and the numerical results in Table 4.1 show that, with $\Delta_r = 5$, actually the observed ratio is close to 60% ($u_1(t)/u_2(t)$ = 0.603, 0.602, 0.601 for t = 20, 25, 30, respectively).

6 Concluding Remarks

In this paper, we suggest a proportional differentiation model based on a performance measure of unfill rate. As the implementing method of the proportional differentiation, we suggest a TDP scheduling discipline with priority parameter $\{r_i\}$. Since it is difficult to measure the exact performance under the TDP scheduling discipline, we derive asymptotic results on the waiting time distribution and approximate the ratio of the unfill rates. We demonstrate that the approximated ratio can be used effectively to control the intended proportional differentiation between the high- and low-priory traffic classes when the packet arrival follows a Poisson process. The approach used in this paper can be classes. It requires additional analysis on the service order and waiting time of the packets from each class, which is left for future study. Also it might be verified by more numerical experiments whether the TDP scheduling discipline suggested in the paper remains effective when there are many traffic classes. Examining other types of TDP priority functions is another possible area for further researches.

Acknowledgements

This work was supported by Hankuk University of Foreign Studies Research Fund.

Appendix A

We use the following notations for the poof of Theorem 1 in this section and Theorem 2 in Appendix B.

H: set of arrival indices of high-priority class packets

L: set of arrival indices of low-priority class packets

R_n : total waiting time of the *n*-th arriving packet

W_n : waiting time in the queue of the *n*-th arriving packet

On : service order of the *n*-th arriving customer under the TDP scheduling discipline

B'_n : service time of the *n*-th departing packet

W'_n : waiting time in the queue of the *n*-th arriving packet under FIFO policy

Under the TDP scheduling discipline, the *n*-th arriving packet is served in the order of O_n, which would be different from its arrival order *n*. If the *n*-th arriving packet is class 1, the difference between W_n and W'_{O_n} is equal to the interarrival time between the O_n-th and *n*-th arriving packets;

$$W_n = W'_{O_n} - \sum_{i=O_n}^{n-1} A_i \quad \text{for } n \in H. \tag{A.1}$$

We define another notation related to the high-priority class service level;

$$G_n := n - \left| \{1 \le k < n : \sum_{i=k}^{n-1} A_i \le \Delta_r, k \in L\} \right| \quad \text{for } n \in H. \tag{A.2}$$

As mentioned in Section 3, a high-priority packet can catch up with only the low-priority packets which have arrived at most Δ_r time units earlier. The last term in (A.2) denotes the number of those low-priority packets. Thus, G_n means the earliest service order that the *n*-th arriving packet could take when it is in the high-priority class. Hence, for $n \in H$

$$G_n \le O_n \le n. \tag{A.3}$$

If $O_n > G_n$, it means the G_n-th service has started before the *n*-th packet arrives;

$$W'_{G_n} < \sum_{i=G_n}^{n-1} A_i \quad \text{for } n \in H. \tag{A.4}$$

And, before getting served, the *n*-th packet needs to wait at most for the packets in service order of G_n through $O_n - 1$ to finish their service;

$$W_n \le \sum_{i=G_n}^{O_n-1} B'_i \quad \text{for } n \in H. \tag{A.5}$$

Now, we show that for $n \in H$

$$0 \le P\{R_n > t\} - P\{W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\} \le P\{\sum_{i=1}^{n-G_n+1} \overline{B}_i > t\} \tag{A.6}$$

where $\{\overline{B}_n\}$ is i.i.d. and has the same distribution as $\{B'_n\}$.

$$\begin{aligned} P\{R_n > t\} &= P\{W_n + B_n > t\} = P\{W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\} \\ &= P\{O_n = G_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\} + P\{O_n > G_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\} \\ &= P\{W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\} \\ &\quad + P\{O_n > G_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\} - P\{O_n > G_n, W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\}. \end{aligned} \tag{A.7}$$

From (A.4) and (A.5), we have

$$\begin{aligned} 0 &\le P\{R_n > t\} - P\{W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\} \\ &\le P\{O_n > G_n, W'_{O_n} - \sum_{i=O_n}^{n-1} A_i + B'_{O_n} > t\} \le P\{\sum_{i=G_n}^{O_n-1} B'_i + B'_{O_n} > t\} \\ &\le P\{\sum_{i=1}^{n-G_n+1} B'_i > t\}. \end{aligned} \tag{A.8}$$

From the Markov's inequality,

$$P\{\sum_{i=1}^{n-G_n+1} B'_i > t\} \le e^{-\nu t} E\left[\exp\left(\nu \cdot \sum_{i=1}^{n-G_n+1} B'_i\right)\right] = e^{-\nu t} E\left[\exp(\phi_B(\nu)(n-G_n+1)\right] \tag{A.9}$$

for an arbitrary $\nu \ge 0$. From the definition of G_n in (A.2), we have that $n - G_n$ has the same distribution as $\min\{n, \overline{N}\}$ where $\overline{N}$ is defined as follows;

$$\overline{N} = \left| \{k > 0; \sum_{i=1}^{k} \overline{A}_i \le \Delta_r, k \in L\} \right| \tag{A.10}$$

and $\{\overline{A}_n\}$ is an i.i.d. sequence which has the same distribution as $\{A_n\}$. Thus, for $\nu \ge \gamma$ with finite $\phi_B(\nu)$,

$$e^{\nu t} P\{\sum_{i=1}^{n-G_n+1} \overline{B}_i > t\} \le e^{-(\nu-\gamma)t} E[\exp(\phi_B(\nu)(\overline{N}+1)] \to 0 \quad \text{as } t \to \infty. \tag{A.11}$$

Since the right-hand side of the inequality in (A.11) does not dependent on *n*, it converges to 0 uniformly as $t \to \infty$. With (A.6), this implies that

$$e^{\nu t}\left(P\{R_n > t\} - P\{W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\}\right) \to 0 \quad \text{as } t \to \infty. \tag{A.12}$$

In order to complete the proof, we need to show that for some constant *C*

$$\lim_{n\to\infty} e^{\nu t} P\{W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\} \to C \exp(-\gamma(1-\theta)\Delta_r)$$

as $t \to \infty$. (A.13)

Let $X_k = B'_k - A_k$ and define $\{S_n\}$ as

$$S_0 = 0 \text{ and } S_n = \sum_{k=1}^{n} X_k . \quad (A.14)$$

If $\lambda E[B'_1] < 1$, then W'_n converges weakly to a random variable W' which has the same distribution as $\max_{k\geq 0} S_k$ [6]. Using the result in the proof of Theorem 1 of Glasserman and Wang [7], we conclude $(W'_{G_n}, n - G_n)$ converges in distribution to $(W', \bar{N})$ and W' is independent of $\bar{N}$ and $\{\bar{A}_n\}$.

From the above argument with new notations, we have that

$$\lim_{n\to\infty} P\{W'_{G_n} - \sum_{i=G_n}^{n-1} A_i + B'_{O_n} > t\} = P\{W' - \sum_{i=1}^{\bar{N}} \bar{A}_i + \bar{B} > t\} \quad (A.15)$$

where $\bar{B}$ is a random variable having the same distribution as $\{B'_n\}$ and independent of W', $\bar{N}$ and $\{\bar{A}_n\}$. Using $\{S_n\}$ in (A.14), we define $\tau = \inf\{n \geq 1 : S_n > T\}$ with $T = t + \sum_{i=1}^{\bar{N}} \bar{A}_i - \bar{B}$. Then,

$$P\{W' - \sum_{i=1}^{\bar{N}} \bar{A}_i + \bar{B} > t\} = P\{\max_{n\geq 0} S_n > t + \sum_{i=1}^{\bar{N}} \bar{A}_i - \bar{B}\} = P\{\tau < \infty\} . \quad (A.16)$$

We use exponential twisting [6]. Specially we use γ-twisting of $\{X_n\}$ and $\bar{B}$ and (−γ)-twisting of $\{\bar{A}_n\}$, and denote the use of a twisted measure in computing expectations by $\tilde{E}$. Note that τ is a stopping time for $\{X_n\}$ and $\bar{N}$ is a stopping time for $\{\bar{A}_n\}$. From Theorem XII.4.1 of Asmussen [6], we have that

$$P\{\tau < \infty\} = \tilde{E}\left[\prod_{i=1}^{\tau} \exp(-\gamma X_i + \phi_X(\gamma)) \cdot \prod_{i=1}^{\bar{N}} \exp(-\gamma \bar{A}_i + \phi_A(-\gamma)) \cdot \exp(-\gamma \bar{B} + \phi_B(\gamma)); \tau < \infty\right] \quad (A.17)$$

where the semicolon inside the expectation indicates that the expectation is evaluated over the event after the semicolon. Using the definition of γ, we reduce the above equation further;

$$P\{\tau < \infty\} = \tilde{E}\left[\exp\left(-\gamma S_\tau + \gamma \sum_{i=1}^{\bar{N}} \bar{A}_i - \beta \bar{N} - \gamma \bar{B} + \beta\right); \tau < \infty\right]$$
$$= e^{-\gamma t} e^{\beta} \tilde{E}\left[\exp\left(-\gamma(S_\tau - T) - \beta \bar{N}\right); \tau < \infty\right] \quad (A.18)$$

where $\beta = \phi_B(\gamma)$. Since a cumulant generating function is convex [8] and $\phi_X(0) = 0$,

$$\tilde{E}[X_i] = E\left[e^{\gamma X_1} X_1\right] = \phi'_X(\gamma) > 0 , \quad (A.19)$$

and thus the event of $\{\tau < \infty\}$ has probability one. The random variable T is independent of $\{S_n\}$ and $T \to \infty$ as $t \to \infty$. From Corollary 8.33 of Siegmund [9], we have that

$$C_1 = \lim_{t\to\infty} \tilde{E}\left[e^{-\gamma(S_\tau - T)}\right] = \tilde{E}\left[e^{-\gamma Z}\right] \quad (A.20)$$

where Z is a ladder variable. Since the distribution of Z is independent of $\{\bar{A}_n\}$, we have

$$\lim_{t\to\infty} e^{\gamma t} P\{W' - \sum_{i=1}^{\bar{N}} \bar{A}_i + \bar{B} > t\} = C_1 e^{\beta} \tilde{E}\left[e^{-\beta \bar{N}}\right] = C\tilde{E}\left[e^{-\beta \bar{N}}\right] \quad (A.21)$$

with $C := C_1 e^{\beta}$. After (−γ)-twisting, $\{\bar{A}_n\}$ has an exponential distribution with mean $1/(\lambda + \gamma)$. It means that under the twisted measure $\bar{N}$ has a Poisson distribution with mean $(\lambda + \gamma)(1 - \theta)\Delta_r$. Thus,

$$\tilde{E}\left[e^{-\beta \bar{N}}\right]$$
$$= \sum_{k=0}^{\infty} e^{-\beta k} \exp\left(-(\lambda+\gamma)(1-\theta)\Delta_r\right)\left((\lambda+\gamma)(1-\theta)\Delta_r\right)^k / k!$$
$$= \exp\left(-(\lambda+\gamma)(1-\theta)\Delta_r(1-e^{-\beta})\right) = \exp\left(-\gamma(1-\theta)\Delta_r\right) \quad (A.22)$$

and we have (A.13).

Appendix B

If the n-th arriving packet is class 2, the difference between W_n and W'_n is equal to the total service time of the n-th through $(O_n - 1)$-th departing packets; following

$$W_n = W'_n + \sum_{i=n}^{O_n - 1} B'_i \quad \text{for } n \in L . \quad (B.1)$$

For the analysis of the low-priority class service level, we introduce a new variable $\{F_n\}$, which is similar to $\{G_n\}$ in (A.2) for the high-priority class. For the n-th arriving packet which has low priority (i.e. $n \in L$),

$$F_n := n + \left|\{k > n : \sum_{j=n}^{k-1} A_j \leq \Delta_r, k \in H\}\right| \quad \text{for } n \in L . \quad (B.2)$$

The meaning of F_n is the last service order that the n-th arriving packet could take when it is in the low-priority class 2.

If $O_n < F_n$, there exists at least one high-priority packet which arrives within Δ_r time units after the n-th packet arrived but cannot catch up with it. This means the low-priority packet started to get served before the high-priority packet arrives. Its waiting time in the queue must have been less than Δ_r ($W_n < \Delta_r$). From (B.1), the following inequality holds;

$$W'_n + \sum_{i=n}^{O_n - 1} B'_i < \Delta_r \quad \text{for } n \in L . \quad (B.3)$$

Now, we show that for $n \in L$

$$0 \leq P\{W'_n + \sum_{i=n}^{F_n - 1} B'_i + B'_{O_n} > t\} - P\{R_n > t\} \leq P\{\sum_{i=1}^{F_n - n} \bar{B}_i > (t - \Delta_r)/2\}$$

(B.4)

where $\{\overline{B}_n\}$ is i.i.d. and has the same distribution as $\{B'_n\}$.

$$P\{R_n > t\} = P\{W_n + B_n > t\} = P\{W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\}$$

$$= P\{O_n = F_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\}$$

$$+ P\{O_n < F_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} + B'_{O_n} > t\}$$

$$= P\{W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\}$$

$$+ P\{O_n < F_n, W'_n + \sum_{i=n}^{O_n-1} B'_i + B'_{O_n} > t\}$$

$$- P\{O_n < F_n, W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\}. \quad \text{(B.5)}$$

From (B.3), we have

$$\begin{aligned} 0 &\le P\{W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\} - P\{R_n > t\} \\ &\le P\{O_n < F_n, W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\} \\ &\le P\{\Delta_r + B'_{O_n} + \sum_{i=O_n+1}^{F_n-1} B'_i + B'_{O_n} > t\} \quad \text{(B.6)} \\ &\le P\{\sum_{i=n}^{F_n-1} B'_i > (t-\Delta_r)/2\} \\ &= P\{\sum_{i=1}^{F_n-n} \overline{B}_i > (t-\Delta_r)/2\}. \end{aligned}$$

From the Markov's inequality,

$$P\{\sum_{i=1}^{F_n-n} \overline{B}_i > (t-\Delta_r)/2\} \le e^{-(\nu t-\Delta_r)/2} E\left[\exp(\phi_B(\nu)(F_n-n)\right] \quad \text{(B.7)}$$

for an arbitrary $\nu \ge 0$. From the definition of F_n in (B.2), we have that $F_n - n$ has the same distribution as $\overline{M}$, which is a Poisson random variable defined as follows;

$$\overline{M} = \left| \{k > 0; \sum_{i=1}^{k} \overline{A}_i \le \Delta_r, k \in H\} \right| \quad \text{(B.8)}$$

and $\{\overline{A}_n\}$ is an i.i.d. sequence which has the same distribution as $\{A_n\}$. Thus, for $\nu \ge 2\gamma$ with finite $\phi_B(\nu)$,

$$e^{\gamma t} P\{\sum_{i=1}^{F_n-n} \overline{B}_i > (t-\Delta_r)/2\} \le e^{-(\nu/2-\gamma)t+\Delta_r/2} E[\exp(\phi_B(\nu)\overline{M}] \to 0$$

as $t \to \infty$. (B.9)

Since the right-hand side of the inequality in (B.9) does not dependent on n, it converges to 0 uniformly as $t \to \infty$. With (B.4), this implies that

$$e^{\gamma t}\left(P\{W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\} - P\{R_n > t\} \right) \to 0$$

as $t \to \infty$. (B.10)

In order to complete the proof, we need to show that for some constant C

$$\lim_{n\to\infty} e^{\gamma t} P\{W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\} \to C\exp(\gamma\theta\Delta_r)$$

as $t \to \infty$. (B.11)

Using a similar method as in the proof of Theorem 1, we can show that $(W'_n, F_n - n)$ converges in distribution to $(W', \overline{M})$ and W' is independent of $\overline{M}$. Thus,

$$\lim_{n\to\infty} P\{W'_n + \sum_{i=n}^{F_n-1} B'_i + B'_{O_n} > t\} = P\{W' + \sum_{i=1}^{\overline{M}+1} \overline{B}_i > t\} \quad \text{(B.12)}$$

where $\{\overline{B}_n\}$ is i.i.d. with the same distribution as $\{B'_n\}$ and independent of W', $\overline{M}$ and $\{\overline{A}_n\}$. Using $\{S_n\}$ in (A.14), we define $\tau' = \inf\{n \ge 1 : S_n > T'\}$ with $T' = t - \sum_{i=1}^{\overline{M}+1} \overline{B}_i$. Then,

$$P\{W' + \sum_{i=1}^{\overline{M}+1} \overline{B}_i > t\} = P\{\max_{n\ge 0} S_n > t - \sum_{i=1}^{\overline{M}+1} \overline{B}_i\} = P\{\tau' < \infty\}. \quad \text{(B.13)}$$

With γ-twisting of $\{X_n\}$ and $\{\overline{B}_n\}$, we have that

$$P\{\tau' < \infty\} = \tilde{E}\left[\prod_{i=1}^{\tau'} \exp(-\gamma X_i + \phi_X(\gamma)) \cdot \prod_{i=1}^{\overline{M}+1} \exp(-\gamma \overline{B}_i + \phi_B(\gamma)); \tau' < \infty \right]. \quad \text{(B.14)}$$

Using the definition of γ, we reduce the above equation further;

$$\begin{aligned} P\{\tau' < \infty\} &= \tilde{E}\left[\exp\left(-\gamma S_{\tau'} - \gamma \sum_{i=1}^{\overline{M}+1} \overline{B}_i + \gamma(\overline{M}+1) \right); \tau' < \infty \right] \\ &= e^{-\gamma t} e^{\beta} \tilde{E}\left[\exp\left(-\gamma(S_{\tau'} - T') + \beta\overline{M} \right); \tau' < \infty \right] \end{aligned} \quad \text{(B.15)}$$

With the twisted measure, $\tilde{E}[X_i] = E[e^{\gamma X_1} X_1] = \phi'_X(\gamma) > 0$ and thus the event of $\{\tau' < \infty\}$ has probability one. The random variable T' is independent of $\{S_n\}$ and $T' \to \infty$ as $t \to \infty$. From Corollary 8.33 of Siegmund [9], we have that

$$C_1 = \lim_{t\to\infty} \tilde{E}[\exp(-\gamma(S_{\tau'} - T'))] = \tilde{E}\left[e^{-\gamma Z}\right]. \quad \text{(B.16)}$$

Since the distribution of Z is independent of $\{\overline{B}_n\}$, we have that

$$\lim_{t\to\infty} e^{\gamma t} P\{W' + \sum_{i=1}^{\overline{M}+1} \overline{B}_i > t\} = C_1 e^{\beta} \tilde{E}\left[e^{\beta\overline{M}}\right] = C\tilde{E}\left[e^{\beta\overline{M}}\right]$$

(B.17)

with $C := C_1 e^{\beta}$. Note that the constant is the same as in Theorem 1.

Since $\overline{M}$ is independent of both $\{X_n\}$ and $\{\overline{B}_n\}$,

$$\tilde{E}\left[e^{\beta\overline{M}}\right] = E\left[e^{\beta\overline{M}}\right] = \sum_{k=0}^{\infty} e^{\beta k} \exp\left(-\lambda\theta\Delta_r\right)\left(\lambda\theta\Delta_r\right)^k / k!$$
$$= \exp\left(-\lambda\theta\Delta_r(1-e^{\beta})\right) = \exp\left(\gamma\theta\Delta_r\right) \quad \text{(B.18)}$$

and we have (B.11).

References

[1] C. Dovrolis and P. Ramanathan, A case for relative differentiated services and the proportional differentiation model, IEEE Network, Vol.13, (1999), 26-34.

[2] C. Dovrolis, D. Stiliadis and P. Ramanathan, Proportional differentiated services: delay differentiation and packet scheduling, IEEE/ACM Trans. Networking, Vol.10, No.1, (2002), 12-26.

[3] M. K. H. Leung, J. C. S. Lui and D. K. Y. Yau, Characterization and Performance Evaluation for Proportional Delay Differentiated Services, Proceeding of the IEEE International Conference on Network Protocols, Nov 2000, Osaka, Japan.

[4] L. Essafi and G. Bolch, Time dependent priorities in call centers, Internat. J. of Simulation, Vol.6, No.1-2, (2005), 32-38.

[5] L. Essafi and G. Bolch, Performance Evaluation of Priority Based Schedulers in the Internet, Proceedings of the 17th European Simulation Multiconference, June 2003, Nottingham, UK.

[6] S. Asmussen, Applied Probability and Queues, Wiley, New York, 1987.

[7] P. Glasserman and Y. Wang, Leadtime-inventory trade-offs in assemble to order systems, Operations Research, Vol.46, No.6, (1998), 858-871.

[8] M. Kendall, Advanced Theory of Statistics, Vol.2, 5th Ed., Oxford, New York, 1987.

[9] D. Siegmund, Sequential Analysis : Tests and Confidence Intervals, Springer, New York, 1985.

Ki-Seok Choi received B.S. and M.S. degrees in industrial engineering, respectively from Seoul National University, Korea in 1991, and from KAIST, Korea in 1993, and Ph.D. degree in industrial and systems engineering from Georgia Institute of Technology, US in 2003. He is currently an Associate Professor at Department of Industrial and Management Engineering, Hankuk University of Foreign Studies, Korea. Previously he worked at Electronics and Telecommunication Research Institute and Samsung Data Systems. His research interests are in telecommunication networks and quality of services.