Performance Evaluation of a Prioritized Limited Round-Robin System

四方　義昭、華山　宣胤
SHIKATA Yoshiaki and HANAYAMA Nobutane

[Abstract]

Suppose that there are two classes and that an arriving (class-1 or class-2) request encounters n1 class-1 and n2 class-2 requests in a single-server system. We then propose a new prioritized limited round-robin (RR) rule, under which m quanta are individually and simultaneously allocated to class-1 requests in each RR cycle, whereas one quantum is allocated to class-2 requests. Furthermore, the sum of the quanta being included in each RR cycle is kept below a fixed value. An arriving request that cannot be allocated a quantum because of such a restriction is either entered in the service waiting queue (the waiting system) or rejected (the loss system). Practical performance measures, such as the relationship between the mean sojourn time, mean queue waiting time, loss probability, and quantum size are evaluated via simulation. In the evaluation, the sojourn time of an arriving request is evaluated using the requested service time, number of quanta included in each RR cycle, and quantum size. Then, the change in the number of quanta needed before service is complete is reevaluated at the arrival or departure of other class-1 and class-2 requests. Tracking these events and calculations enables us to analyze the performance of the prioritized limited RR rule. In particular, we obtain the most suitable quantum size that minimizes the mean sojourn time for the case where the switching time for each quantum is considered. The proposed RR rule and performance evaluation results are realistic in the server and client type communication system where the time-shared computer is employed as the server system.

Keywords:
prioritized limited RR rule, quantum, processor sharing, sojourn time, performance measures, simulation, loss probability

1. Introduction

Under the Round-Robin (RR) rule, a processor allocates to each request a fixed amount of time, called a quantum, in a fixed order. If the requested service time (the total time required from the processor) is completed in less than the quantum, the request leaves. Otherwise, it is added to the end of the queue of requests waiting for a quantum allocation (called the quantum waiting queue) and waits its turn

to receive another quantum of service. It continues in this fashion until it's requested service time has been obtained from the processor. An arriving request that can be allocated quanta enters the quantum waiting queue and waits for the allocation of the first quantum.

Under the prioritized RR rule where two priority classes (class-1 and class-2) are allowed, the processor allocates m (≥ 2) quanta to each class-1 request and one quantum to each class-2 request in each RR cycle. Here, the RR cycle is a series of quanta cyclically allocated to each request in a fixed order. In such a prioritized RR paradigm, the service ratio for individual requests decreases with the increase in the number of arriving requests. Therefore, in theory, the sojourn time of each request increases to infinity as the number of arriving requests increases. In order to prevent such an increase and develop a realistic model of sharing, the number of requests being allocated quanta can be limited. In such a prioritized limited RR system, the number of quanta that can be included in each cycle is kept below a fixed value (the service-facilitycapacity). Arriving requests that cannot be allocated quanta because of such a restriction are entered into the service waiting queue or rejected.

In this study, we evaluate practical performance measures such as the mean sojourn time of requests, mean waiting time in the service waiting queue, and loss probability of prioritized limited RR systems via simulation. In this simulation algorithm, the sojourn time of an arriving request is first evaluated using the requested service time, number of quantum included in each RR cycle, and quantum size.

Then, the change in the number of quanta needed before service is complete is reevaluated at the arrival or departure of other class-1 or class-2 requests. Tracking these events and calculations enables us to analyze the performance of the prioritized limited RR rule. Moreover, by considering the switching time for each quantum, we obtain the most suitable quantum size for minimizing the mean sojourn time. The proposed RR rule and performance evaluation results are realistic in the server and client type communication system where the time-shared computer is employed as the server system.

The Processor Sharing (PS) rule, an idealization of quantum-based RR scheduling at the limit where quantum size becomes infinitesimal, has been the subject of many papers [1][2]. A limited PS system and a prioritized limited PS system, in which the number of requests receiving service is kept below a fixed value, have also been proposed, and the performance of these systems has been analyzed [3][4]. However, a few explicit results are available for the RR policy. The influence of variable job or quantum size in the presence of switching overhead on the mean sojourn time of the non-limited RR rule has been evaluated [5]-[7]. A limited RR rule has also been proposed, and the performance of this system has been analyzed via simulation [8]. However, practical performance measures of the prioritized limited

RR rule have not been studied. Moreover, the influence that the quantum size or the service-facility capacity may have on the mean sojourn time, the mean waiting time in the service waiting queue, and the loss probability in the prioritized limited RR system have not been investigated.

2. Prioritized limited round-robin rule

2.1 Quantum allocation

Suppose that there are two classes and that an arriving (class-1 or class-2) request encounters n1 class-1 and n2 class-2 requests (including the arriving one) in a single-server system. According to the proposed prioritized limited RR rule, if $m * n1 + n2 \leq C$, m quanta are individually and simultaneously allocated to class-1 requests in each RR cycle, whereas one quantum is allocated to class-2 requests. Otherwise (i.e., $m * n1 + n2 > C$), the arriving request is either queued in the corresponding class service waiting queue (the waiting system) or rejected (the loss system). Here, m (≥ 2) denotes the priority ratio, and C ($\leq \infty$), the service-facility capacity.

In the waiting system, at the end of service for a request, another request is taken from the service waiting queue and is attached to the quantum waiting queue. In the waiting or loss system, an arriving request that can be allocated a quantum (a request that is not put in the service waiting queue, or is not rejected) enters the quantum waiting queue directly. Two queuing methods for adding a request to the quantum waiting queue are possible. In the end-in method, a request is queued to the end of the quantum waiting queue. In the top-in method, it is queued to the top of the quantum waiting queue. However, the difference in the sojourn time between these two methods is small enough to be ignored in comparison with the sojourn time [8]. Therefore, in this study, only the end-in method is considered in the performance evaluation of the system. Moreover, in a prioritized limited RR system in the presence of switching overhead, the most suitable quantum size that minimizes the mean sojourn time may be obtained. Therefore, this suitable quantum size should be studied in various prioritized limited RR systems such as the waiting and loss systems.

2.2 Evaluation technique

(1) Simulation algorithm

Under the RR rule, whenever the service for an arriving request starts or the requested service time of a request is over, the remaining sojourn time of each request currently receiving service is respectively extended or reduced. This extension or reduction of the remaining sojourn time can be calculated using the number of RR cycles that are necessary before the service is completed, the number of each class requests, and quantum size. By logging the change in remaining sojourn time

in the simulation program (Figure 1), performance measures of practical interest such as the mean sojourn time for a request, the mean waiting time in the service waiting queue, and the loss probability may be evaluated. This simulation program uses the variable time increment method, in which the simulation time is skipped until the next event that causes a change in a system state occurs in order to shorten the simulation execution time. Events that can cause a system state change in the simulation of the prioritized limited RR system include the following.

(a) Arrival of a request

An arriving request enters the quantum waiting queue and waits for the allocation of the first quantum. In the evaluation of the sojourn time, the quantum waiting time, i.e., the time that elapses between the addition of the request to the queue and the first allocation of a quantum, has to be considered. This time is obtained from the total number of requests in the quantum waiting queue, consisting of the sum of requests receiving service and requests waiting for the first quantum to be allocated (requests that have not yet received service), and quantum size. The time until the next request arrives is also calculated according to a predetermined distribution such as the exponential distribution, the hyper-exponential distribution, or the Erlang inter-arrival distribution.

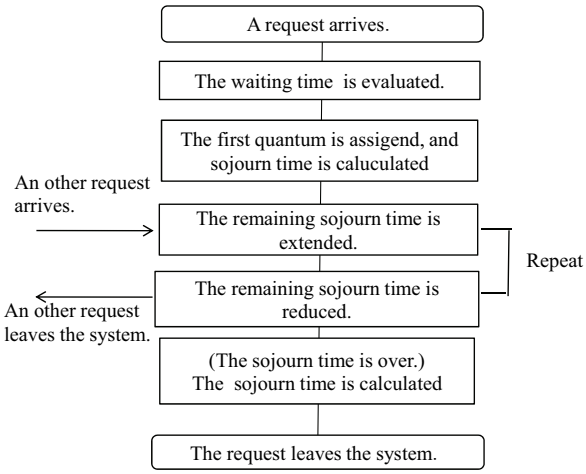

Figure 1. Simulation flow

(b) Start of service

When the waiting time of a request in the quantum waiting queue ends, the first quantum is allocated and the service for the request starts. At the start of the service, the number of quanta required to complete the service is obtained from the requested service time S_r divided by the quantum size qs. The time from the first quantum allocation to the end of service, S_e (the remaining service time), is obtained from the time required to complete the RR cycles plus the portion of quantum left. Each RR cycle includes m quanta allocated to each class-1 request and one quantum allocated to each class-2 request. That is, in the case of the class-2 request

$$Se=floor(Sr/qs)*qs*(m*n1+n2)+Sr-floor(Sr/qs)*qs, \quad (1)$$

and in the case of the class-1 request

$Se=floor(Sr/(m*qs))*qs*(m*n1+n2)+Sr-floor(Sr/(m*qs))*m*qs$ (2)

Here, n1 or n2 represents the number of class-1 or class-2 requests receiving service, and the operation floor(value) returns the next lowest integer value by rounding down if necessary. The requested service time Sr is calculated according to a predetermined distribution.

Each quantum allocated to an arriving request is inserted into the existing RR cycle, as shown in Figure 2. The extension of the remaining service time of each request is obtained by evaluating the number of RR cycles included in the remaining service time. Therefore, the remaining service time of the requests receiving service is extended for the arrival of a class-2 request according to

$S_e=S_o+ceil(S_o/((m*n1+n2)*qs))*qs$ (3)

and for an arrival of a class-1 request as

$S_e=S_o+ceil(S_o/((m*n1+n2)*qs))*m*\ qs$ (4)

Here, the variable So is the remaining service time of each request currently receiving service just before the first quantum is allocated to an arriving request. Further, the function ceil(value) returns the next highest integer value by rounding up if necessary.

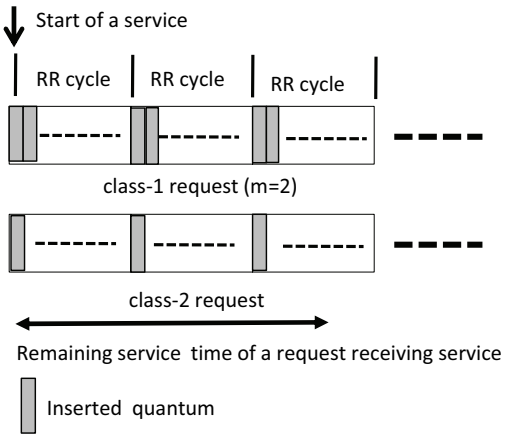

Figure 2 Inserted quantum

(c) End of service

At the end of service for a request, the quantum allocated to this request is removed from the existing RR cycle, as shown in Figure 3. The reduction of remaining service time of each request is also obtained by evaluating the number of RR cycles included in the remaining service time. Therefore, at the end of service for a class-2 request, the remaining service time of each request currently receiving service is shortened by

$S_e=S_o-floor(S_o/((m*n1+n2)*qs))*qs$ (5)

and by

$S_e=S_o-floor(S_o/((m*n1+n2)*qs))*m*qs$ (6)

in the case of the end of service for a class-1 request. Then, in the waiting system, a request in the service waiting queue is removed and put in the quantum waiting queue. The time until the first quantum is allocated is evaluated using the procedure in Section (a).

Tracking these events and calculations enables us to evaluate practical performance measures such as the loss probability, waiting time in the service waiting queue, and mean sojourn time for requests.

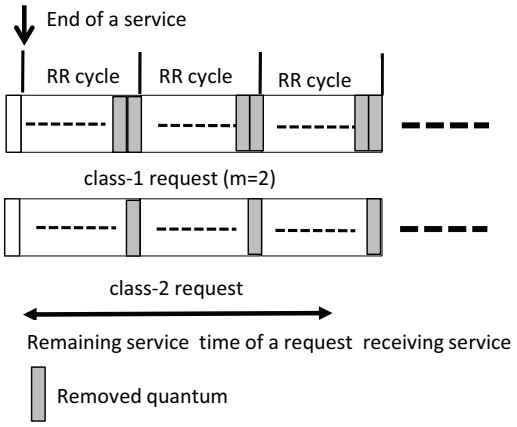

Figure 3 Removed quantum

(2) Simulation clock management

In the simulation program, the simulation clock is controlled by the arrival timer, quantum waiting timer, or service timer of each request that is receiving service. At the arrival of each class request, the time duration until the next arrival of the request is set into the arrival timer according to the predetermined arrival time distribution. Further, the quantum waiting time of each arriving request is set into the quantum waiting timer according to the number of request in the quantum waiting queue. Moreover, the arrival time of each request is memorized in the corresponding variable. At the start of service of each class request, the service time (e.g., remaining sojourn time) of each arriving request is set into the service timer (Equation 1 or 2). The quantum waiting time or service time of each request in the server is evaluated using these data and service end time. In the while loop of this simulation program, the arrival processing, service start process, or service end processing mentioned in the Section (1) is executed on the expiry of one of the arrival timers, quantum waiting timers, or service timers. Moreover, the service timer, quantum waiting timer, or arrival timer with the next smallest value is detected, and the time duration of this timer is subtracted from all the remaining timers. Therefore, in the next while loop this timer expires. Simultaneously, the simulation clock is pushed forward by this time duration in order to skip the insignificant simulation clock. The while loop is repeated until the number of arriving requests attains a predetermined value.

3. Evaluation results

In this evaluation, the priority ratio m is assumed to be 2, and a two-stage Erlang inter-arrival time distribution and two-stage hyper exponential service time distribution are implemented. The mean requested service times and traffic

densities are assumed to have the same value for both the class-1 and class-2 requests, and are assumed to be 1.2 and 0.3, respectively. Evaluation results were averaged over ten simulation runs. About 80,000 requests were produced for each class in each run.

3.1 Loss system

Figure 4 shows the relationship between the loss probability and service-facility capacity in the loss system. The range of markers includes 95% of the reliability intervals obtained from the ten simulation runs. Here, C1 or C2 represents class-1 or class-2 requests, respectively, and QS represents the quantum size. The logarithm of the loss probability increases linearly as the service-facility capacity decreases. The loss probability of the class-1 requests is approximately double that of the class-2 requests. The loss probability when the quantum size is 0.05 increases more rapidly than when the quantum size is 0.2. In general, the influence of the service-facility capacity on the loss probability increases when the quantum size decreases.

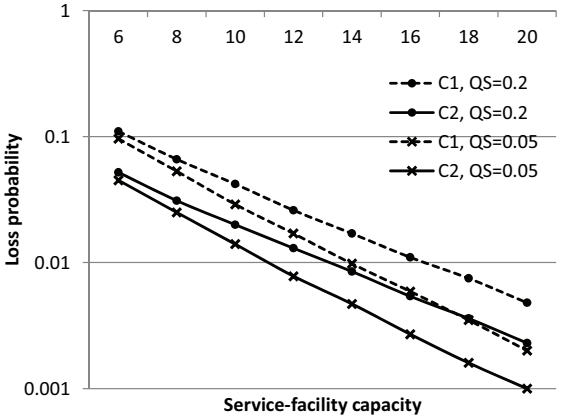

Figure 4, Loss probability versus service-facility capacity

Figure 5 shows the relationship between the loss probability and quantum size. Here, SFC represents the service-facility capacity. The loss probability increases with quantum size. The loss probability when the service-facility capacity is 14 increases more rapidly than when the service-facility capacity is 10. In general, when the service-facility capacity increases, the influence of the quantum size on the loss probability also increases.

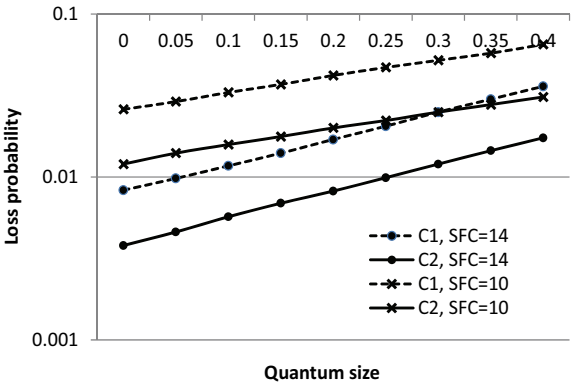

Figure 5, Loss probability versus quantum size

Figure 6 shows relationship between the mean sojourn time and quantum

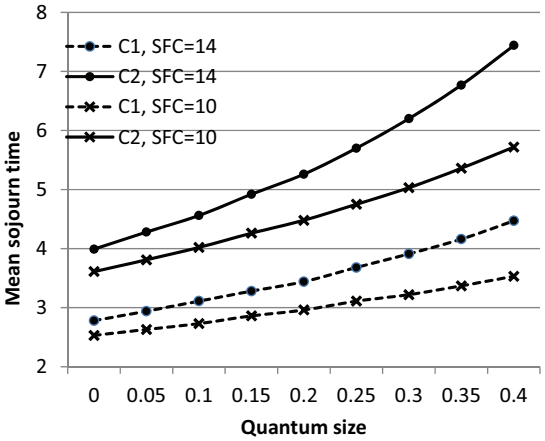

Figure 6. Mean sojourn time versus quantum size

size. Here, the sojourn time is obtained as the sum of the quantum waiting time and the service time. The mean sojourn time increases with quantum size. The mean sojourn time of class-2 requests increases more rapidly than that of the class-1 requests. Moreover, the mean sojourn time when the service-facility capacity is 14 increases more rapidly than when the service-facility capacity is 10.

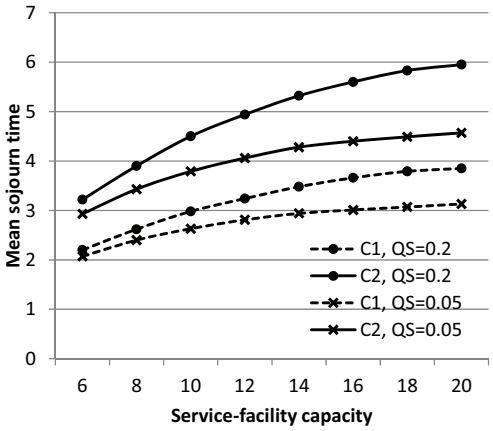

Figure 7. Mean sojourn time versus service-facility capacity

Figure 7 shows the relationship between the mean sojourn time and service-facility capacity. The mean sojourn time increases with an increase in the service-facility capacity. The mean sojourn time of class-2 requests increases more rapidly than that of classs-1 requests.

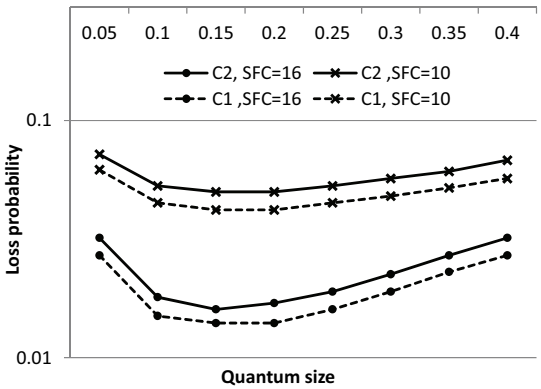

Figure 8. Loss probability versus quantum size

If a quantum is allocated to an arriving class-2 request when the service-facility capacity can afford one quantum, the class-2 request has precedence over a class-1 request. As a result, the loss probability of class-2 requests is smaller than that of class-1 requests, as shown in Figures 4 and 5. A quantum reservation method can reduce the difference in loss probability between class-1 and class-2 requests. In this method, an arriving class-1 or class-2 request is allocated a quantum only when the service-facility capacity can afford more than m quanta. Moreover, for the prioritized limited RR system in the presence of switching overhead, a suitable quantum size that minimizes the mean sojourn time may be obtained. Figure 8 shows the relationship between the loss probability and quantum size in the presence of quantum reservation and switching overhead. Figure 9 also shows the relationship

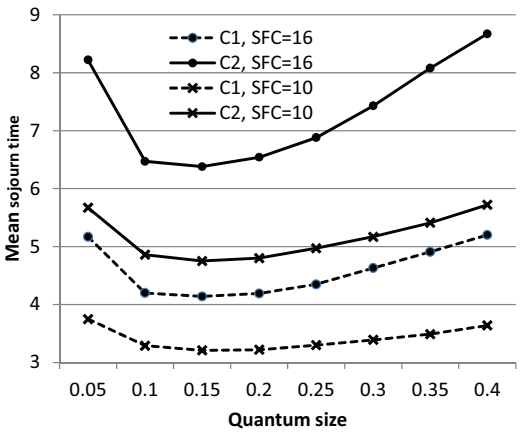

Figure 9. Mean sojourn versus quantum

between the mean sojourn time and quantum size in the presence of the quantum reservation and switching overhead. Here, the switching overhead is assumed to be 0.01. The loss probability or the mean sojourn time decreases with quantum size and reaches a minimum at a quantum size of 0.15.

3.2 Waiting system

Figure10 shows the relationship between the mean sojourn and waiting times in the service waiting queue, and the service-facility capacity in the waiting system. Here, the sojourn time is obtained as the sum of the service waiting queue time, the quantum waiting time, and the service time. The quantum size is assumed to be 0.2. When a class-1 or classs-2 request is moved from the service waiting queue to the quantum waiting queue, the quantum reservation method for the waiting system is also implemented. In this method, when the service-facility capacity becomes able to afford m quanta, a class-1 or class-2 request is moved from the corresponding service waiting queue to the quantum waiting queue. A class-1 request is moved with precedence over a class-2 request. As a result, the mean waiting times in the service waiting queue for both class-1 and class-2 requests becomes almost the same, as shown in Figure 10. The mean sojourn time of each class request increases with the decrease of the service-facility capacity, in contrast to the loss system (Figure 7). This is because the time spent in the service waiting queue increases rapidly when the service-facility capacity decreases. Furthermore, the mean sojourn time of the class-1 request is smaller than that of the class-2 request, but increases

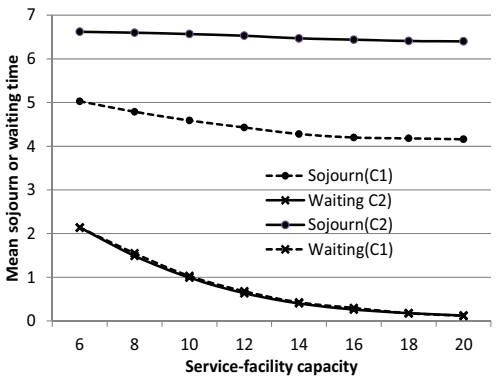

Figure 10, Mean sojourn time versus service-facility capacity

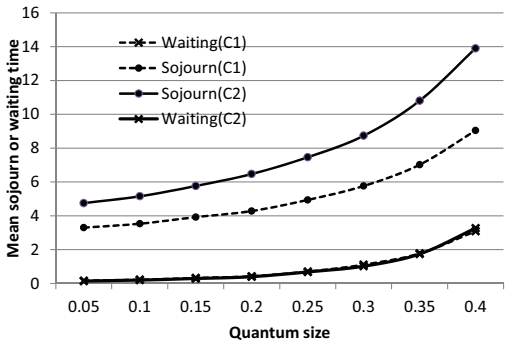

Figure 11, Mean sojourn or waiting time versus quantum size

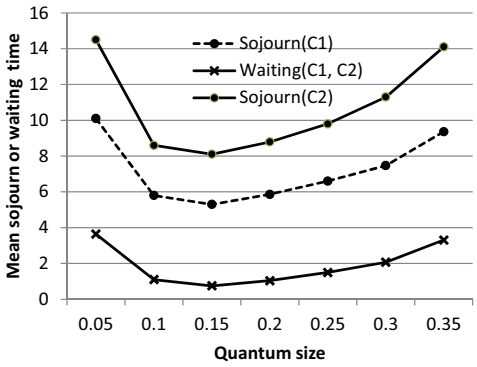

Figure 12. Mean sojourn or waiting time versus quantum size

more rapidly with the decrease of the service-facility capacity.

Figure 11 shows the relationship between the mean sojourn and waiting times in the service waiting queue and quantum size in the presence of quantum reservation and switching overhead. Here, the service-facility capacity is assumed to be 14. The mean waiting times for both class-1 and class-2 requests are the same. The mean sojourn times of both class-1 and class-2 requests increases with quantum size. Moreover, the mean sojourn time of the class-2 request increases more rapidly than that of the class-1 request.

Figure 12 also shows the relationship between the mean sojourn times, the mean waiting time in the service waiting queue, and the quantum size in the presence of the quantum reservation and switching overhead. Here, the service-facility capacity and switching overhead are assumed to be 14 and 0.01, respectively. With a decrease in quantum size, the mean sojourn and waiting times decrease and reach a minimum at a quantum size of 0.15.

4. Conclusion

In order to prevent excessive increases in the sojourn time of each request in a prioritized RR system, a prioritized limited RR system has been proposed. Practical performance measures such as the mean sojourn time in the server, mean waiting time in the service waiting queue, and loss probability, were evaluated using simulation programs. In these programs, the sojourn time of an arriving request is evaluated using the requested service time, number of quanta included in each RR cycle, and quantum size. Then, changes in the number of quanta needed before service is complete are reevaluated at the arrival or departure of other requests. It is clear that the mean sojourn time or loss probability increases in a loss system with an increase in quantum size. Furthermore, the mean sojourn time of each class request decreases with the decrease of the service-facility capacity. On the other hand, in the waiting system, the mean sojourn time of each class request increases with the decrease of the service-facility capacity. It is also clear that in the presence of switching overhead and quantum reservation, the most suitable quantum size that minimizes the mean sojourn time can be obtained. In the future, we intend to study the performance of a prioritized limited Multi-Processor Sharing system, where prioritized and non-prioritized requests share the multi-processor system.

References

[1] L. Kleinrock, "Time-Shared Systems: A Theoretical Treatment", J.A.C.M1.14, 242-261 (1967).

[2] E.Altman, K.Avrachenkov and U.Ayesta, "A survey on processor sharing", Queueing

Syst (2006) 53:53-63
[3] G. Yamazaki and H. Sakasegawa, "An optimal design problem for limited sharing systems", Management Science, vol.33 (8), pp.1010--1019 (1987).
[4] Y.Shikata, W.Katagiri, and Y.Takahashi, "Prioritized Limited Processor-Sharing System with its Performance Analysis", International Conference on Operations Research, August30 - 1, 2011 Zurich
[5] Varun Gupta, "Finding the optimal quantum size: Sensitivity analysis of the M/G/1 round-robin queue", ACM SIGMETRICS Performance Evaluation Review archive Volume 36 Issue 2, September 2008 Pages 104-106
[6] Varun Gupta, Jim Dai, MorHarchol-Balter, and BertZwart, "The effect of higher moments of job size distribution on the performance of an M/G/K queueing system", Technical Report CMU-CS-08-106, School of Computer Science, Carnegie Mellon University, 2008.
[7] Ward Whitt, "On approximations for queues, I: Extremal distributions", AT&T Bell Laboratories Technical Journal, 63:115-138, 1984.
[8] Y.Shikata,"Performance Evaluation of a Limited RR system", WASET International Conference ICCEL2012, June 11-12, 2013, Berlin.