International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:5, 2013

Performance Evaluation of a Limited Round-Robin System

Yoshiaki Shikata

***Abstract*—**Performance of a limited Round-Robin (RR) rule is studied in order to clarify the characteristics of a realistic sharing model of a processor. Under the limited RR rule, the processor allocates to each request a fixed amount of time, called a quantum, in a fixed order. The sum of the requests being allocated these quanta is kept below a fixed value. Arriving requests that cannot be allocated quanta because of such a restriction are queued or rejected. Practical performance measures, such as the relationship between the mean sojourn time, the mean number of requests, or the loss probability and the quantum size are evaluated via simulation. In the evaluation, the requested service time of an arriving request is converted into a quantum number. One of these quanta is included in an RR cycle, which means a series of quanta allocated to each request in a fixed order. The service time of the arriving request can be evaluated using the number of RR cycles required to complete the service, the number of requests receiving service, and the quantum size. Then an increase or decrease in the number of quanta that are necessary before service is completed is reevaluated at the arrival or departure of other requests. Tracking these events and calculations enables us to analyze the performance of our limited RR rule. In particular, we obtain the most suitable quantum size, which minimizes the mean sojourn time, for the case in which the switching time for each quantum is considered.

***Keywords*—**Limited RR rule, quantum, processor sharing, sojourn time, performance measures, simulation, loss probability.

I. Introduction

UNDER the RR (Round-Robin) rule, a processor allocates to each request a fixed amount of time, called a quantum, in a fixed order. If a requested service time (the total time required from the processor) is completed in less than the quantum, the request leaves; otherwise, it feeds back to the end of the queue of quantum waiting requests (called a quantum waiting queue), waits its turn to receive another quantum of service, and continues in this fashion until its requested service time has been obtained from the processor. In such an RR paradigm, the service ratio for individual requests decreases with an increase in the number of arriving requests. Therefore, in theory, the sojourn time of each request increases to infinity with an increase in the number of arriving requests. In order to prevent such an increase in the sojourn time of each request in an RR paradigm and to develop a realistic model of sharing, a method for limiting the number of requests being allocated quanta is considered. In such a limited RR system, the sum of the number of requests being allocated quanta is kept below a fixed value (called the service number restriction). Arriving requests that cannot be allocated quanta because of such a restriction are entered in the queue of service waiting requests (called a service waiting queue) or rejected. Practical performance measures, e.g., the mean sojourn time of requests, the mean number of requests, and the loss probability of such limited RR systems, are evaluated via simulation. Moreover, in order to clarify the performance of the realistic limited RR rule, two queuing methods of an arriving request are also considered. The performances of the two methods are compared.

In order to evaluate these practical performance measures, we propose a new simulation algorithm. In this simulation algorithm, first, the requested service time of an arriving request is converted into a quantum number. One of these quanta is included in an RR cycle, which means a series of quanta allocated to each request in a fixed order. The service time of such an arriving request can be evaluated using the number of RR cycles required to complete the service, the number of requests receiving service, and the quantum size. Then, at the arrival or departure of other requests, an increase or decrease in the number of quanta that are necessary before service is completed is reevaluated. Tracking these events and calculations enables us to analyze performance for our 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 processor-sharing (PS) rule, an idealization of quantum-based RR scheduling in the limit where the quantum size becomes infinitesimal, has been the subject of many papers [1]-[2]. A limited PS system, in which the number of requests receiving service is kept below a fixed value, has also been proposed, and performance of this system has been analyzed [3]. Moreover, the influence that the variability of the job size or the quantum size in the presence of switching overhead gives to the mean sojourn time in the non-limited RR rule has been evaluated [4]-[6]. However, practical performance measures of the limited RR rule have not been studied. Moreover, the influence that the quantum size or the service restriction number may have on the mean sojourn time, the mean number of requests, or the loss probability in the limited RR system have not been clarified.

II. Limited Round-Robin Rule

A. Quantum Allocation

In the limited RR rule, the sum of the requests being allocated the quantum is kept below a fixed value. An arriving request that cannot be allocated the quantum because of such a service number restriction will be entered in the service waiting queue (called the queuing system) or rejected (called the loss

Prof. Dr. Y. Shikata is with the Faculty of Informatics for Arts, Shobi University, Kawagoe, Saitama 350-1153, Japan (Corresponding author; phone: 81-49-246-5251; fax: 81-49-246-5235; e-mail: shikata@ictv.ne.jp).

system). In the queuing system, at the end of service for a request, another request is taken from the service waiting queue and is allocated the quantum.

Two queuing methods of an arriving request to the quantum waiting queue are considered. In the end-in method, an arriving 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. The waiting time before the first quantum is allocated to an arriving request (called the waiting time) is extended in the end-in method, but the time from the first quantum allocation to the end of service (called the service time) is shortened. In contrast, the waiting time may be shortened in the top-in method, but the service time may be extended. Therefore, the mean sojourn times of these two methods, which are obtained as the sum of the respective mean waiting times and mean service times, have to be compared.

Moreover, in the limited RR system in the presence of switching overhead, the suitable quantum size, which minimizes the mean sojourn time, may be obtained. Therefore, the suitable quantum size has to be studied in various limited RR systems, such as the queuing system, the loss systems, the top-in system, or the end-in system.

B. Simulation Algorithm

In the simulation program 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, is used in order to shorten the simulation execution time. Events that can cause a change in a system state in the simulation of the limited RR system include the following.

1. Arrival of a Request

When a request arrives, the waiting time is obtained, in the case of the top-in method, from the number of requests waiting for the first quantum to be allocated (requests that have not received service yet) depending on the quantum size plus the time before the start of the next quantum after the arrival of the request (Fig. 1). On the other hand, in the case of the end-in method, this time is obtained from the total number of requests in the quantum waiting queue, which is the sum of the requests receiving service and the requests waiting for the first quantum to be allocated, again depending on the quantum size. The time until a next request arrives is also calculated according to a predetermined distribution, e.g., the exponential distribution, the hyper-exponential distribution, or the Erlang inter-arrival distribution.

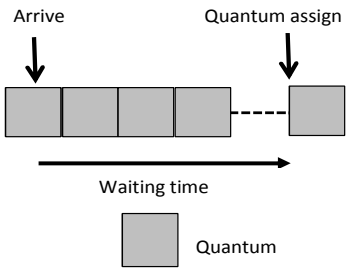

Fig. 1 Waiting Time

2. Start of Service

At the beginning of a service request (the first quantum allocation), first, the quantum number required to complete the service is obtained from the requested service time $S_r$over the quantum length ql.The service time of this arriving request $S_a$is obtained from the time required to complete the RR cycles plus a portion of the left quantum.Each RR cycle includes one of the quantum numbers obtained above. That is,

$$S_a = \text{floor}(S_r/ql) * ql * n + S_r - \text{floor}(S_r/ql) * ql \quad (1)$$

Here, n represents the number of requests receiving service, and floor (value) returns the next lowest integer value by rounding down, if necessary. Sr is also calculated according to the predetermined distribution.

Each quantum allocated to an arriving request is inserted into the existing RR cycle, as shown in Fig. 2. Therefore, the remaining service time of the requests receiving service is extended as

$$S_a = S_o + \text{ceil}(S_o/ql) * ql \quad (2)$$

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

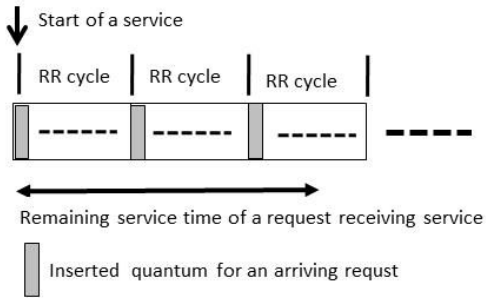

Fig. 2 Inserted quantum

3. 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 Fig. 3.

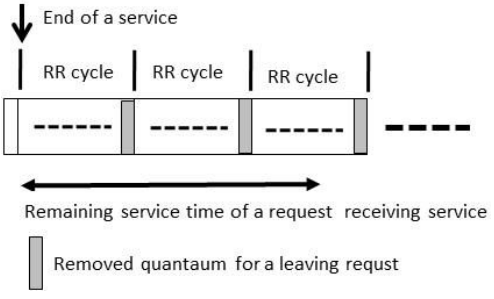

Fig. 3 Removed quantum

Therefore, the remaining service time of each request is shortened as

$$S_a = S_o - \text{floor}(S_o/ql) * ql \quad (3)$$

Then, a request in the service waiting queue in the queuing system is taken, and the time until the first quantum is allocated to this request is calculated, using the procedure shown in Section II.*B*.1.

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

III. Evaluation Results

In the evaluation, the two-stage Erlang inter-arrival distribution and the two-stage hyper-exponential requested service time distribution are considered. Evaluation results are obtained as the average of ten simulation results. About 70,000 requests are produced in each simulation.

A. Non-Limited RR System

Fig. 4 shows the evaluation results for the relationship between the mean sojourn time (shown as round makers) and the quantum size in the non-limited RR system. Here, Ar represents the arrival rate of requests. In this figure the mean waiting time (shown as cross markers), which is included in the mean sojourn time, is also evaluated. 95% reliability intervals obtained from the ten simulation results are included in the range of markers. The evaluation results for the quantum size of 0 are obtained using a simulation program for the limited PS system [7].

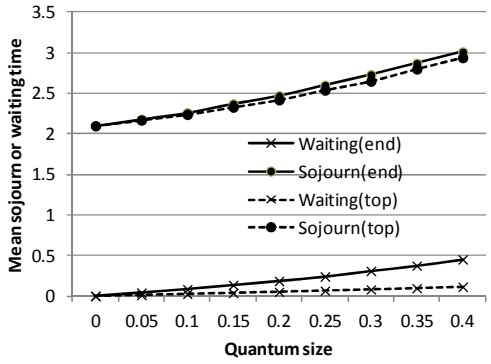

Fig. 4Comparison of sojourn time in non-limited RR system (Ar = 0.6, Sr = 1)

With an increase in the quantum size, the mean sojourn time increases in both the end-in method and the top-in methods. The mean waiting time in the case of the end-in method increases more rapidly than in the case of the top-in method. On the other hand, the mean service time in the case of the top-in method increases more rapidly than in the case of the end-in method. As a result, the mean sojourn times in the case of the end-in method become slightly greater than in the case of the top-in method.

B. Infinite Queuing System

In this queuing system, it is necessary to evaluate the waiting time in the service waiting queue (called the service waiting time) in addition to the sojourn time in the non-limited RR system.

Fig. 5 shows the evaluation results for the relationship between the mean total sojourn time (shown as round markers), which is obtained as the sum of the mean sojourn time defined in Section II.*A* and the mean service waiting time, and the quantum size in the queuing system in which infinite waiting rooms are prepared. This figure also shows the mean service waiting time (shown as cross markers). Here, Rn represents the service restriction number. With an increase in the quantum size, the mean total sojourn time and the service waiting time increase. The mean service waiting time of the top-in method is the same as that of the end-in method. The mean total sojourn time in the case of the top-in method (shown as a dotted line) is slightly less than in the case of the end-in method (shown as a solid line). This is because of the same reason as in the case of the non-limited RR system (see Section III.*A*).

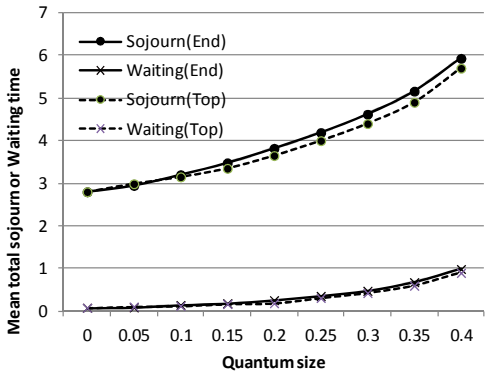

Fig. 5 Comparison of the mean total sojourn or waiting time in the limited RR system (Ar = 0.7, Sr =1, Rn = 9)

Fig. 6 compares the mean sojourn time of the limited RR system with the top-in method (shown as a solid line with round markers) with that of the non-limited RR system (shown as a solid line with cross markers). The mean sojourn time of the limited RR system is less than that of the non-limited RR system. With an increase in the quantum size, the difference in the mean sojourn times between these two systems increases. Fig. 6 also shows the evaluation result of the mean service waiting time of the limited RR system (shown as a dotted line). The total mean sojourn time of the limited RR system is almost the same as the sojourn time of the non-limited RR system.

C. Suitable Quantum Size

In the limited RR system in the presence of switching overhead, the suitable quantum size, which minimizes the mean total sojourn time, may be obtained. Fig. 7 shows the evaluation result for the relationship between the mean total sojourn time and the quantum size in the presence of switching overhead in

the infinite queuing system. Here, Sw represents the switching overhead. With a decrease in the quantum size, the mean total sojourn time decreases. It becomes a minimum at a quantum size of 0.15 (in the case of either the end-in or the top-in method).

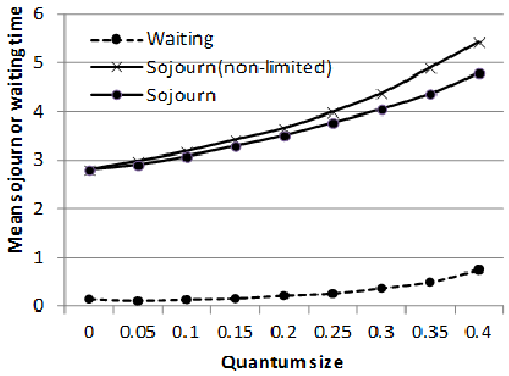

Fig. 6 Comparison of the mean total sojourn or waiting time (Ar = 0.7, Sr =1, Rn = 9)

D. Suitable Service Restriction Number

Fig. 8 shows the evaluation result for the relationship between the mean service waiting time (shown as a solid line) or the mean sojourn time (shown as a dotted line) and the service restriction number in the infinite queuing system with the top-in method. With a decrease in the service restriction number, the mean service waiting time increases. On the other hands, the mean sojourn time decreases. Therefore, the total mean sojourn time level becomes a minimum at a service restriction number of 12 or 14 for an arrival rate of 0.7 or 0.6.

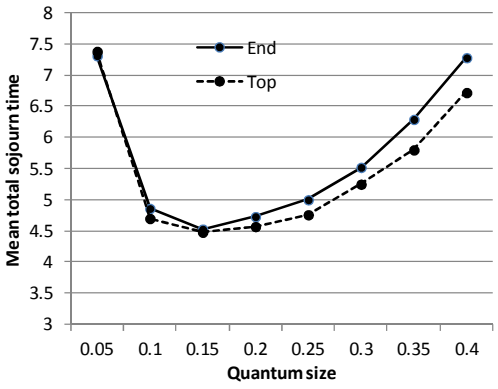

Fig. 7 Suitable quantum size (Ar = 0.7, Sr = 1, Rn = 8, Sw = 0.01)

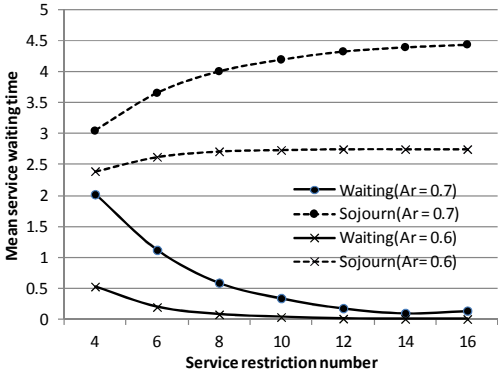

Fig. 8 Suitable restriction number (Ar = 0.7, Sr = 1, Rn = 9, Sw = 0.01)

E. Loss System

Fig. 9 shows the evaluation result for the relationship between the mean number of requests and the quantum size in the loss system in which the service restriction number is 6 (shown as round markers) or 9 (shown as cross markers). Fig. 10 also shows the evaluation result for the relationship between the mean sojourn time and the quantum size. In the case of a service restriction number of 9, the mean number of requests and the mean sojourn time is greater than in the case of the service restriction number of 6. Moreover, the mean number of requests in the case of the end-in method (shown as a solid line) is greater than in the case of the top-in method (shown as a dotted line). This is because the number of requests waiting for the first quantum to be allocated in the quantum waiting queue in the case of the end-in method is much greater than in the case of the top-in method. On the other hand, the percentage of waiting time that consists of the sojourn time is small. Therefore, the mean sojourn time both in the case of the end-in method (shown as a solid line) and in the top-in method (shown as a dotted line) is almost the same.

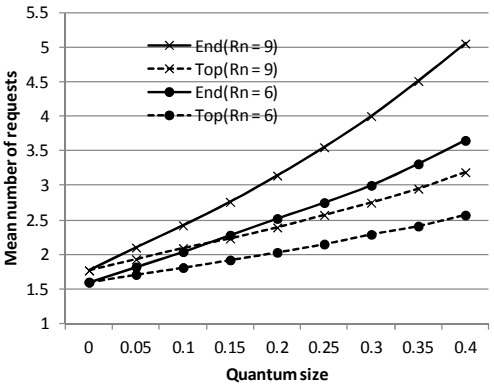

Fig. 9 Mean number of requests in the loss system (Ar = 0.7, Sr = 1)

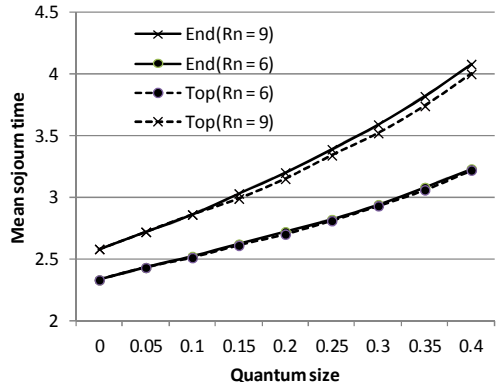

Fig. 10 Mean sojourn time in the loss system (Ar = 0.7, Sr = 1)

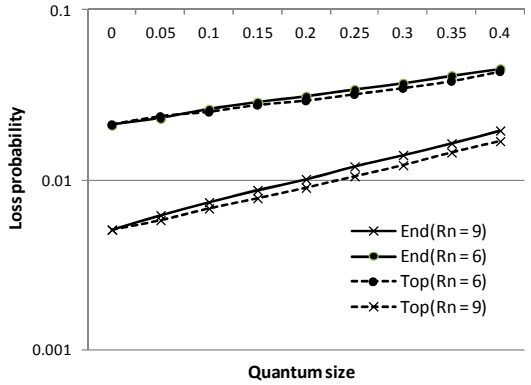

Fig. 11 Loss probability in the loss system (Ar = 0.7, Sr = 1)

Fig. 11 shows the evaluation result for the relationship between the loss probability and the quantum size in the loss system. The logarithm of the loss probability increases linearly with an increase in the quantum size.

Fig. 12 compares the mean sojourn time of the loss system (shown as a solid line) with that of the non-limited RR system (shown as a dotted line). With an increase in the quantum size, the difference in the mean sojourn time between the case of the limited RR system and the case of the loss system increases.

IV. Conclusion

In order to prevent excessive increase in the sojourn time of each request in an RR discipline, we proposed a limited RR system. Practical performance measures, e.g., the mean sojourn time in the server, waiting time in the queue, and loss probability, were evaluated using simulation programs.

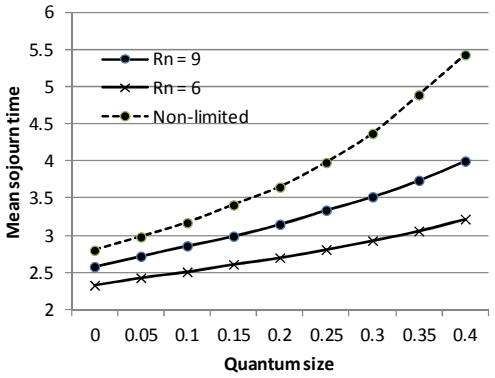

Fig.12 Comparison of the mean sojourn time (Ar = 0.7, Sr = 1)

In these programs, the requested service time of an arriving request is converted to the number of RR cycles, and increases or decreases in the number of these RR cycles is tracked at the arrival or departure of other requests. Moreover, the top-in method and the end-in method are studied as queuing methods of an arriving request to the quantum waiting queue, and the performances of these two methods are compared. It is also clarified that in the infinite queuing system, the suitable quantum size in the presence of switching overhead, or the suitable service restriction number, which minimizes the mean total sojourn time, can be obtained. In the future, we intend to study the performance of a prioritized limited RR system, where prioritized requests and non-prioritized requests share the quantum.

References

[1] L. Kleinrock, "Time-Shared Systems: A Theoretical Treatment", J.A.C.M1.14, 242-261 (1967).
[2] K. Hoshi, Y. Shikata, Y. Takahashi and N. Komatsu, "An Approximate Formula for a GI/G/1 Processor-Sharing System", International Conference on Operations Research, September 1 - 3, 2010 Munich.
[3] G. Yamazaki and H. Sakasegawa, "An optimal design problem for limited sharing systems", Management Science, vol.33 (8), pp.1010--1019 (1987).
[4] 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
[5] Varun Gupta, Jim Dai, MorHarchol-Balter, and BertZwart, "The effect of higher moments of job size distribution on the performance of an M/G/Kqueueing system", Technical Report CMU-CS-08-106, School of Computer Science, Carnegie Mellon University, 2008.
[6] Ward Whitt, "On approximations for queues, I: Extremal distributions", AT&T Bell Laboratories Technical Journal, 63:115–138, 1984.
[7] Y.Shikata, W. Katagiri, and Y. Takahashi,"Performance Evaluation of Prioritized Limited Processor-Sharing System" Y. Shikata, W. Katagiri, and Y. Takahashi, WASET International Conference ICCEL2012, June 11-12, 2012, Copenhagen.

Yoshiaki Shikata received his B.E., M.E., and Ph.D. degrees in electrical engineering from Shizuoka University in 1971, 1974, and 1991, respectively. He joined the Musashino ECL, NTT in 1974. Until 1997, he was engaged in the research and development of satellite switched-TDMA communication systems and Personal Handy-phone System (PHS). Since 1997, he has been responsible for developing the Advanced Intelligent Network (AIN) system. Dr. Shikata is currently a professor at the Faculty of Informatics for Arts, Shobi University, Japan.