Iterative Method for Solving M/G/1//N-type Loops with Priority Queues

By ANDRÁS SZÉP

1. Introduction

The range of applicability of queueing models increases every year. Here, closed queueing systems with general service time distributions and several priority dispatching rules are of special interest. However, existing solution methods such as [1] for a special case of exponential servers are cumbersome and cannot be applicable to cases of general distributions. On other hand, an original algorithm [2] to solve $M/G/1//N$-type loops with *FCFS* (first-come-first-served) queues has been recently suggested. Through combining this solution technique with an effective method of decompositions of general servers into exponentials [3] a really well-work method can be synthetized. It would be of practical significant if this technique could be applicable to cases of priority dispatching rules at queues.

In this paper an iterational method is suggested for performance evaluation of closed queueing systems with preemptive resume and non-preemptive priority queues and general service time distributions.

2. Preemptive resume priority queues

Consider a closed queueing system consisting of two service centers, at one of which there is exponential service time distribution and an infinite number of servers (i.e. simple delay type service center), and the other is a Coxian collection replacement for an arbitrary non-exponential server with preemptive resume queueing discipline (see fig. 1). Assume that customers belong to one of R priority classes and

(i) class p customers have priority over class r customers at phase I if $1 \leqq p < r \leqq R$,

(ii) customers whithin the the same priority class follow the *FCFS* queueing discipline at phase I.

At second phase all customers are served simultaneously due to an infinite number of servers but service rates for different classes' customers may vary.

For convenience, we define

R — number of customers classes,

n_p — number of customers in class p $(1 \leqq p \leqq R)$,

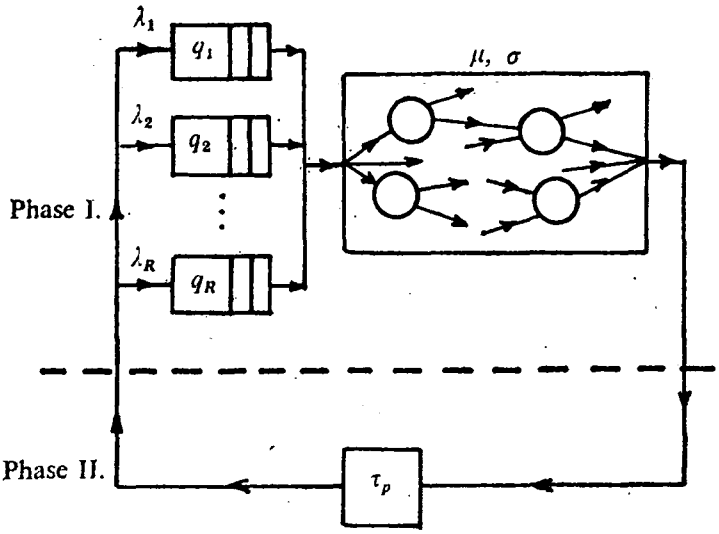

Fig. 1
An $M/G/1//N$-type loop with priority queueing discipline

n_p^* — total number of customers in first p classes,

$$n_p^* = \sum_{r=1}^{p} n_r, \quad \text{for} \quad p = 1, 2, \ldots R, \tag{1}$$

μ — mean service rate at phase I,
σ — standard deviation of service time at phase I,
τ_p — mean service time of customers of class p at phase II, $(1 \leqq p \leqq R)$,
τ_p^* — aggregate mean service time of customers of first p classes at phase II , $(1 \leqq p \leqq R)$,
λ_p — throughput of customers of class p, $(1 \leqq p \leqq R)$,
λ_p^* — aggregate throughput of customers of first p classes

$$\lambda_p^* = \sum_{r=1}^{p} \lambda_r, \quad \text{for} \quad p = 1, 2, \ldots R, \tag{2}$$

q_p — average number of customers of class p at phase I,
q_p^* — aggregate average number of customers of first p classes at phase I,

$$q_p^* = \sum_{r=1}^{p} q_r, \quad \text{for} \quad p = 1, 2, \ldots R, \tag{3}$$

t_p — average elapsed time of customers at phase I for class p, $(1 \leqq p \leqq R)$,
t_p^* — aggregate mean elapsed time of customers of first p classes at phase I, $(1 \leqq p \leqq R)$.

The method suggested in this paper is based on the approach described in [4]. Thus, Little's rule [5] may be applied to the queueing system being considered

$$q_p = \lambda_p t_p, \quad \text{for} \quad p = 1, 2, \ldots R. \tag{4}$$

Several authors (e.g. Jaiswal [6]) noted that aggregate performance characteristics such as mean response time etc. do not depend on the queueing discipline. Moreover, in case of preemptive resume priority disciplines performance characteristics of servicing low priority customers (r) have no effect on servicing high priority customers (p) ($1 \leqq p < r \leqq R$). Therefore, Little's rule may be applied to the aggregate characteristics

$$q_p^* = \lambda_p^* t_p^*, \quad \text{for} \quad p = 1, 2, \ldots R. \tag{5}$$

Having made some simple transformations on (1)—(5) we obtain

$$t_p = \frac{\lambda_p^* t_p^* - \lambda_{p-1}^* t_{p-1}^*}{\lambda_p^* - \lambda_{p-1}^*}, \qquad \text{for} \quad p = 1, 2, \ldots R, \tag{6}$$

assuming $\lambda_0^* = 0$ and $t_0^* = 0$.

The last expression shows the way of successive computation of serving characteristics for customers at all levels of priorities. To attain this it is enough to compute aggregate serving characteristics for the same queueing system as given but with *FCFS* serving discipline at phase I.

There are well known methods and formulas for performance evaluation of closed queueing system with exponential servers. Recently an effective algorithm has been suggested to solve $M/G/1//N$-type loops with *FCFS* queueing discipline in [2].

The above mentioned problem can be solved though difficulties arise. The quantities λ_p and t_p^* are given by Little's formula (5):

$$\lambda_p = \frac{n_p}{\tau_p + t_p}, \quad \text{for} \quad p = 1, 2, \ldots R, \tag{7}$$

and

$$\tau_p^* = \frac{n_p^*}{\lambda_p^*} - t_p^*, \quad \text{for} \quad p = 1, 2, \ldots R. \tag{8}$$

Since t_p does not depend on λ_p linearly, an iterational method can be suggested. The following algorithm contains two embedded iteration processes.

Algorithm 1.

Step 1. Input data — R,

$$\text{and} \quad \|n_1, n_2, \ldots n_R\|, \quad \|\tau_1, \tau_2, \ldots \tau_R\|, \quad \mu, \sigma.$$

Step 2. Initial values $p \leftarrow 0$,

$$\text{and} \quad n_0^* \leftarrow 0, \quad \lambda_0^* \leftarrow 0, \quad t_0^* \leftarrow 0.$$

Step 3. Get next class $p \leftarrow p+1$,

$$\text{take} \quad n_p^* \leftarrow n_{p-1}^* + n_p,$$

$$\text{and first approximation} \quad t_p \leftarrow \frac{n_p}{\mu}.$$

Step 4. Compute

$$\lambda_p \leftarrow \frac{n_p}{\tau_p + t_p},$$

$$\text{and} \quad \lambda_p^* \leftarrow \lambda_{p-1}^* + \lambda_p,$$

$$\text{and first approximation} \quad t_p^* \leftarrow \frac{n_p^*}{\mu}.$$

Step 5. Find next approximation $\tau_p^* \leftarrow \frac{n_p^*}{\lambda_p^*} - t_p^*$.

Step 6. Knowing n_p^*, τ_p^*, μ and σ find a solution of $M/G/1//N$-type loop with *FCFS* queues by the method suggested in [2] and obtain new value for $t_p^{*\prime}$.

Step 7. If $|t_p^{*\prime} - t_p^*| > \varepsilon$ then take $t_p^* \leftarrow t_p^{*\prime}$ and go to step 5., else compute t_p' by formula (6).

Step 8. If $|t_p' - t_p| > \varepsilon$ then take $t_p \leftarrow t_p'$ and go to step 4., else compute $q_p \leftarrow \lambda_p \cdot t_p$.

Step 9. If $p < R$ then go to step 3., else output results —

$$\|\lambda_1, \lambda_2, \ldots \lambda_R\|, \quad \|t_1, t_2, \ldots t_R\|, \quad \|q_1, q_2, \ldots q_R\| \text{ and stop.}$$

(ε — means the error's bound)

It is easy to prove that both iterational processes of this algorithm converge (see fig. 2. and Appendix *A*). The number of elementary operations required for computation is equal to $\sigma(i_1, i_2, N)$, where

$$N = \sum_{r=1}^{R} \sum_{p=1}^{r} n_p,$$

and i_1, i_2-means the number of iterations.

For $\varepsilon = 0.1\%$ the total number of iterations (i_1, i_2) in most cases did not exceed 30, therefore the suggested method for the performance evaluation of closed queueing systems with priorities looks much more efficient than the methods based on calculations of all steady-state probabilities of the system. Note that the number of states of such systems is around $\sim N^R$.

3. Computation speed

A further study of iterational processes in Algorithm 1 indicates that although in most cases they converge rapidly, in case of heavy traffic a relatively large number of iterations may be required (> 100). Therefore we expect it to increase the speed of convergence. This can be achieved by several ways. First, if one chooses a better first approximation, secondly by using more powerful solution searching methods (dichotomic, gradient etc.), and at last by merging two iteration processes. Experiments show that the simultaneous implementation of first and third principle provides the maximum increase of computation performance. Implementation of the second way causes an additional consumption in use of computer resources.

The next algorithm for better convergence was derived for getting performance

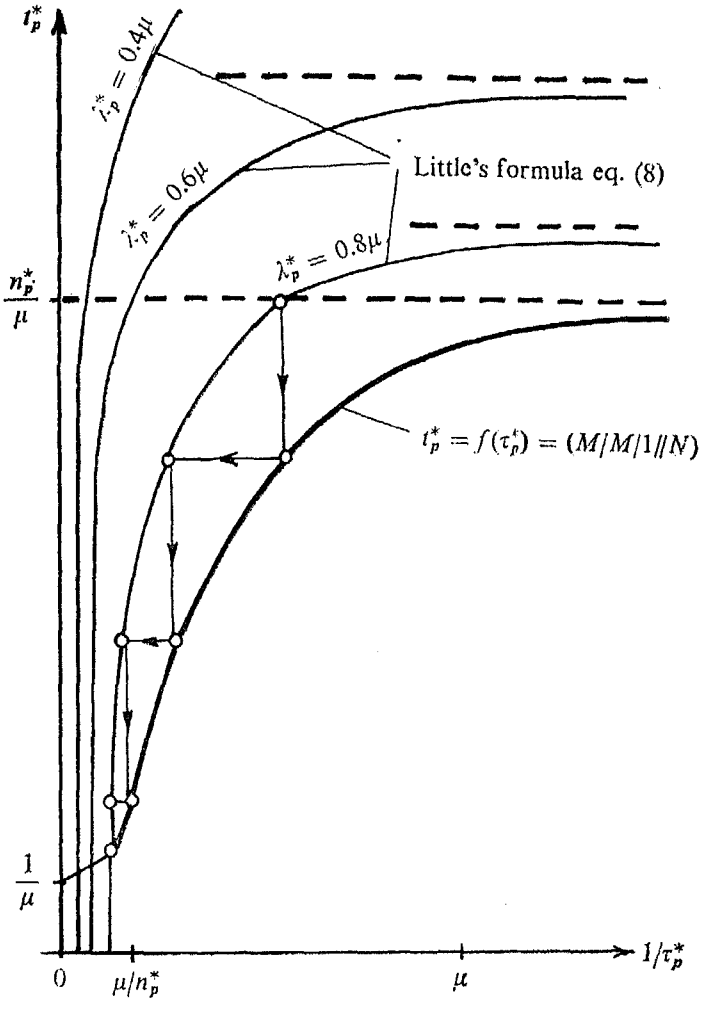

Fig. 2
Dependence of the elapsed times of customers at phase I upon the serving rate at phase II, and an example of convergence

of closed queueing loops with generalized service time distribution and preemptive resume priority dispatching rules.

Algorithm 2.

Step 1. Input data – R,

$$\text{and} \quad \|n_1, n_2, \ldots n_R\|; \ \|\tau_1, \tau_2, \ldots \tau_R\|, \quad \mu, \sigma.$$

Step 2. Initial values $p \leftarrow 0$,

$$\text{and} \quad n_0^* \leftarrow 0, \quad \lambda_0^* \leftarrow 0, \quad t_0^* \leftarrow 0.$$

Step 3. Get next class $p \leftarrow p+1$,

$$\text{and take} \quad n_p^* \leftarrow n_{p-1}^* + n_p.$$

Step 4. Find first approximation for t_p knowing n_p, τ_p, μ and σ by the method of [2], and let $t_p^* \leftarrow t_p$.

Step 5. Compute

$$\lambda_p \leftarrow \frac{n_p}{\tau_p + t_p},$$

$$\lambda_p^* \leftarrow \lambda_{p-1}^* + \lambda_p;$$

$$\text{and} \quad \tau_p^* \leftarrow \frac{n_p^*}{\lambda_p^*} - t_p^*.$$

Step 6. Knowing n_p^*, τ_p^*, μ and σ compute t_p^* by the method of [2].

Step 7. Compute t_p' by the formula (6).

Step 8. If $|t_p' - t_p| > \varepsilon$ then let $t_p \leftarrow t_p'$ and go to step 5., else compute $q_p \leftarrow \lambda_p t_p$.

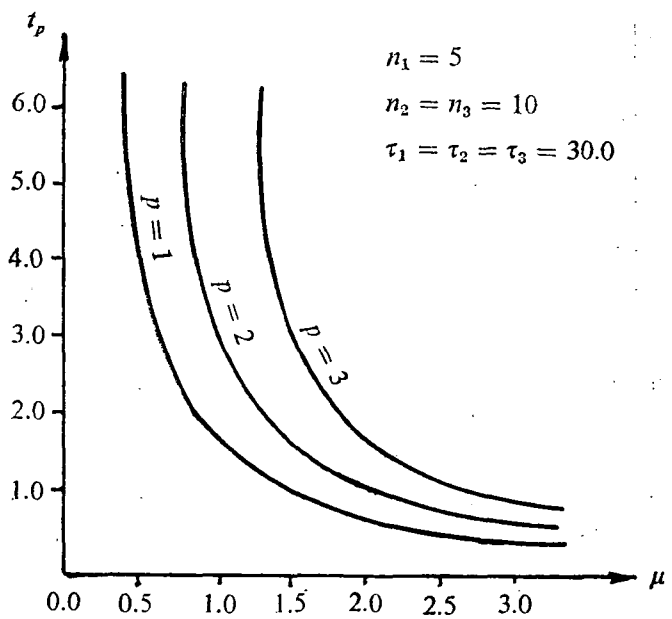

Fig. 3
Dependence of the mean elapsed times of customers upon the service rates at phase I for an $M/G/1//N$-type preemptive resume system

Step 9. If $p<R$ then go to step 3., else output results —

$$\|\lambda_1, \lambda_2, \ldots \lambda_R\|, \quad \|t_1, t_2, \ldots t_R\|, \quad \|q_1, q_2, \ldots q_R\|$$

and stop.

The study of convergence of the above algorithm indicates that 1—10 iterations are sufficient for $\varepsilon=0.1\%$ even in the most extreme cases. Note that for the case of exponential distribution of service times at phase I the mean elapsed times t_p^* can be found by mean-value analysis (MVA) methods (see [7]) at step 6., of Algorithm 1., and at steps 4. and 6., of Algorithm 2.

4. Non-preemptive priority queues

A common approach for solving closed queueing systems with non-preemptive priority queueing discipline is similar to that applied to solve systems with preemptive resume priorities. But it differs since in non-preemptive priority queues the service of customers cannot be interrupted by the arrival of customers with higher priority and they must await releasing of the server. Let us define the mean residual life-time W_p as the mean time which remains until the end of actual service of customer in class p. Then [6]

$$W_p = \frac{\lambda_p \overline{X^2}}{2}, \quad \text{for} \quad p = 1, 2, \ldots R, \tag{9}$$

where $\overline{X^2}$ — means the second moment of the service time distribution. It is clear that the aggregate mean elapsed time of customers in the first p classes at phase I is equal to the aggregate service time of customers in the first p classes in the system with preemptive resume priorities plus the mean residual life-time of all priority classes with number greater than p, i.e.,

$$t_p^*(\text{non-preemptive}) = t_p^*(\text{preemptive}) + \frac{(\lambda_R^* - \lambda_p^*)\overline{X^2}}{2}, \text{ for } p = 1,2,\ldots R \tag{10}$$

where $(\lambda_R^* - \lambda_p^*)$ means the aggregate customers throughput of classes $p+1, p+2, \ldots, R$. Note that λ_R^* (non-preemptive) $= \lambda_R^*$ (preemptive), and the aggregate throughput λ_R^* does not depend on queueing discipline and can be calculated although neither troughputs nor residual lifetimes are known.

The next algorithm is suggested to calculate the performance of closed queueing loops with generalized service time distribution and non-preemptive priority queueing discipline on the basis of Algorithm 2.

Algorithm 3.

Step 1. Do all the claculations of Algorithm 2, and define λ_R^*.

Step 2. Do once more all the calculations of Algorithm 2 with the exception of a correction at step 6, where to the computed value of t_p^* the residual life-times are added

$$t_p^* \leftarrow t_p^* + \frac{(\lambda_R^* - \lambda_p^*)\overline{X^2}}{2}, \quad (p = 1, 2, \ldots R).$$

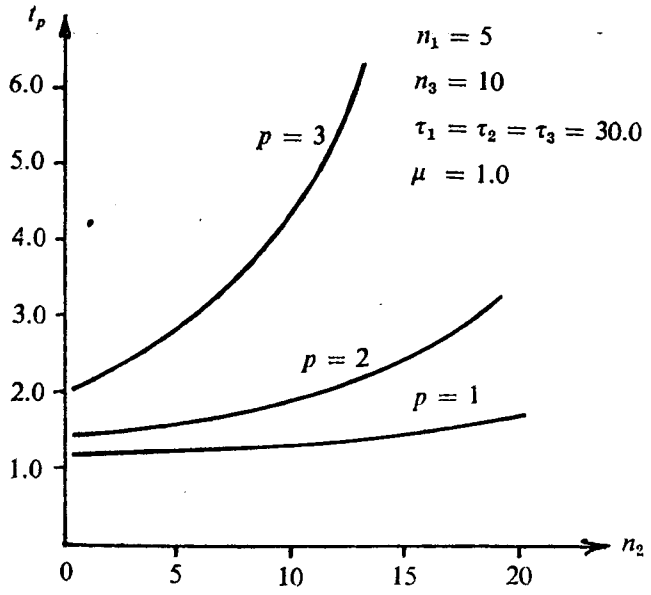

Fig. 4
Dependence of the mean elapsed times upon the number of customers in the second class

This algorithm has the same advantages as the previous one.

In conclusion note that if service rates at phase I for all classes are equal then the suggested method and described algorithms will ensure obtaining theoretically accurate results. If service rates are different for different classes of customers then the aggregate service rates have to be evaluated by

$$\mu_p = \frac{\sum_{r=1}^{p} \lambda_r \mu_r}{\lambda_p^*}, \qquad \text{for} \quad p = 1, 2, \ldots R. \tag{11}$$

Although this way allows for systematic errors in the results of computations, usually evaluated quantities of modelled systems remain within the range of applicability and errors do not exceed 10% [4].

5. Conclusion

Simple algorithms for computing exact mean elapsed times, queue lengths and throughputs of individual customers classes in $M/G/1//N$-type closed preemptive resume and non-preemptive priority systems have been presented for the case when all customers classes have equal mean service times at the non-exponential phase and different at the other phase. It has been shown that algorithms based on the

suggested iterational method converge rapidly and they have unique solutions. For the case of unequal service times an approximation technique has been suggested. The described method and algorithms are efficient for solving large scale application problems.

Acknowledgement

I would like to express my deep gratitude to prof. V. I. Dmitriev for his consulting and valuable advices.

Appendix A

Proof of the convergence of iterational processes and the uniqueness of solution.

On fig. 2, the dependence of elapsed times at phase I upon serving times at phase II is shown. Because the linearity in Little's formula for the convergence of our iterations and uniqueness of the solution it is sufficient to prove that $\frac{\partial t_p^*}{\partial \tau_p^*} > -1$.

The proof will be carried out by induction.

According to MVA (see [7])

$$t_p^*(i) = \frac{1}{\mu}\left(1+q_p^*(i-1)\right) = \frac{1}{\mu}\left(1+\lambda_p^*(i-1)t_p^*(i-1)\right) =$$

$$= \frac{1}{\mu}\left[1+\frac{i-1}{\tau_p^*+t_p^*(i-1)}\,t_p^*(i-1\right] = \frac{1}{\mu}\left[1+\frac{i-1}{\dfrac{\tau_p^*}{1+t_p^*(i-1)}}\right].$$

It is obvious that $\frac{\partial t_p^*(0)}{\partial \tau_p^*} = 0 > -1$.

Let us suppose that $\frac{\partial t_p^*(i-1)}{\partial \tau_p^*} > -1$ then

$$\frac{\partial t_p^*(i)}{\partial \tau_p^*} = -\frac{i-1}{\mu}\left[\frac{t_p^*(i-1)-\tau_p^*\dfrac{\partial t_p^*(i-1)}{\partial \tau_p^*}}{(t_p^*(i-1)+\tau_p^*)^2}\right] >$$

$$> -\frac{i-1}{\mu}\,\frac{t_p^*(i-1)+\tau_p^*}{(t_p^*(i-1)+\tau_p^*)^2} = -\frac{i-1}{\mu(t_p^*(i-1)+\tau_p^*)};$$

because $\tau_p^* \cdot \frac{\partial t_p^*(i-1)}{\partial \tau_p^*} > -\tau_p^*$. But $t_p^*(i-1)+\tau_p^* = \frac{i-1}{\lambda_p^*(i-1)}$ and $\lambda_p^*(i-1) \leqq \mu$ which gives

$$\frac{\partial t_p^*(i)}{\partial \tau_p^*} > -1. \quad \square$$

Abstract

Based on Little's formula an iterational method was derived for the solution of $M/G/1//N$-type closed queueing systems with preemptive resume and non-preemptive priority queues at the general server. Efficient algorithms are outlined and described in detail. Convergence of iterations and uniqueness of solution was proved and also an approximation technique was suggested for the case when service rates differ for different classes of customers at the general server.

Keywords. Closed queueing systems, general distributions, preemptive resume and non-preemptive priorities, iterational method, mean-value analysis.

CENTRAL RESEARCH INSTITUTE FOR PHYSICS
OF THE HUNGARIAN ACADEMY OF SCIENCES
P.O.BOX 49, BUDAPEST, H-1525, HUNGARY

References

[1] A. BRANDWAJN, A finite difference equations approach to a priority queue, Opns. Res. 30. (1982), 74—81.
[2] J. L. CAROLL, A. VAN DE LIEFVOORT, L. LIPSKY, Solutions of $M/G/1//N$-type loops with extensions to $M/G/1$ and $GI/M/1$ queue, Opns. Res. 30 (1982), 490—514.
[3] K. C. SEVCIK, A. I. LEVY, S. K. TRIPATHI, J. L. ZAHORJAN, Improving approximations of aggregated queueing network subsystems, in: Computer Performance (Proc. Int. Symp. on Computer Performance Modeling and Evaluation, New York, 1977, North-Holland publ. co. Amsterdam, 1977), 1—22.
[4] J. P. BUZEN, A. B. BONDI, The response times of priority classes under preemptive resume in $M/M/m$ queues, Opns. Res. 31 (1983), 456—465.
[5] J. D. C. LITTLE, A proof for the queueing formula $L=\lambda W$, Opns. Res. 9 (1961), 383—387.
[6] N. K. JAISWAL, Priority Queues (Academic Press, New York, 1968).
[7] M. REISER, S. S. LAVENBERG, Mean-value analysis of closed multichain queueing networks, J. ACM 27 (1980) 313—322.

Received Nov. 14, 1984.