12

Two-Class Priority Queueing System with Time-Limited Schedule

Tsuyoshi Katayama
Department of Electronics and Informatics
Faculty of Engineering, Toyama Prefectural University
Kosugi, Toyama 939-03, Japan
E-mail: katayama@pu-toyama.ac.jp

Abstract

A flexible priority discipline with time-limited schedule of controllable parameters $(T_1, T_2, \cdots, T_N)$ is presented in this paper, which operates as follows: After the last visit of a single-server at queue n, the server serves messages in queue n, $n = 1, 2, \cdots, N$ until either queue n becomes empty or a timer with time-limit T_n expires, whichever occurs first. In succession, the highest class message present in the system is next served according to the time-limited service. For two-class ($N = 2$), Markovian priority queues with time-limited schedule (T_1, T_2), we determine a generating function of a steady-state, joint queue-length distribution. In the case of $(T_1 = T_2 = \infty)$, this model reduces to the alternating priority queues, while in the case of $(T_1 = \infty, T_2 = 0)$, the ordinary preemptive-resume priority queues, where this priority model is also a limiting case of two queues with alternating service periods first studied by Coffman, Fayolle and Mitrani (1987). Through a generating function approach, we provide Laplace-Stieltjes transforms of distribution functions of the response time and the waiting time of each class, and present numerical examples for the mean performance measures and the mean completion time.

Keywords

Priority queue, flexible priority discipline, time-limited schedule, preemptive-resume discipline, response time, waiting time, completion time, iterative functional equation.

Performance and Management of Complex Communication Networks
T. Hasegawa, H. Takagi & Y. Takahashi (Eds.)

1. INTRODUCTION

A number of priority queueing models have been analyzed for the performance evaluation of communication, computer and manufacturing systems; However, most classical priority disciplines, such as the preemptive/nonpreemptive (head-of-the-line), the shortest/longest-job-first, and the exhaustive-service priority disciplines, have no controllable parameters. A flexible priority discipline with time-limited schedule is defined by a vector of time-limit parameters $\boldsymbol{T} := (T_1, \cdots, T_n, \cdots, T_N)$, $0 \le T_n \le \infty$, and operates as follows: After the last visit of a single-server at queue n, the server serves messages (or customers) in queue n, $n = 1, 2, \cdots, N$ until either queue n becomes empty or a timer with time-limit (also called *maximum-attendance-time*) T_n expires, whichever occurs first. This service discipline is called *a time-limited service*. In succession to the time-limited service for class-n messages, the highest class message present in the system is next served according to the time-limited service, where class 1 is the highest, and class N the lowest. In this paper, we will analyze the simplest but practical two-class, Markovian ($M/M/1$-type) priority queues with time-limited schedule (T_1, T_2) as a special case of the general parameter $\boldsymbol{T}$, which is applicable to the performance analysis of asymmetric half-duplex transmission systems in optical subscriber networks and a multiplexer of voice and data packet transmission used in wideband packet networks. If $T_1 = T_2 = \infty$, then this reduces to the alternating priority discipline, whereas if $T_1 = \infty$ and $T_2 = 0$, it reduces to the ordinary preemptive priority discipline. Setting appropriate timer values, such a flexible priority discipline is effective for performance optimization, and has potential applicability to processing systems with multiple grades of service requirements, e.g. the routing scheme with priority classes used in packet processing systems in the Internet and the broadband ISDN considered in Prycker [15].

Despite flexibility and effectiveness of the time-limited schedule, analytical results have not yet been obtained for the above priority system. However, there have been fruitful results related to polling systems (or cyclic-service systems) and vacation systems with time-limited service: Leung (1994) studied an $M/M/1$-type cyclic-service system with nonpreemptive, time-limited services with the general parameter $\boldsymbol{T}$ using the numerical approach based on discrete Fourier transforms. In the simplest setting, Coffman, Fayolle and Mitrani (1987) derived analytically a generating function of a joint queue-length distribution in an $M/M/1$-type,

alternating service queues with time-limits (T_1, T_2) distributed exponentially by using the boundary-value technique for the first time. Komatsu and Hinomoto (1989) evaluated the mean waiting times in a two-queue model with constant time-limited service ($T_1 = T$) and preemptive/nonpreemptive priority disciplines by using numerical inversion of Laplace transforms. Several analytic approximations for polling systems with time-limited service have been presented by Yue and Brooks (1990), Tangemann and Sauer (1991) and Chang and Sandhu (1994), e.g. Yue et al. analyzed a polling system with high-priority stations controlled by the token holding timer and low-priority stations by the token rotation timer. Various vacation systems with time-limited service have been analyzed by Leung and Eisenberg (1990, 1991), Takagi and Leung (1994), Chiarawongse, Srinivasan and Teorey (1994) and Alfa (1995), e.g. Takagi et al. analyzed a discrete-time vacation model with preemptive-resume, exhaustive time-limited service by using the technique of discrete Fourier transforms. Alfa also analyzed a discrete-time vacation model with Markovian arrival process and phase-type service time distribution using the matrix-geometric method.

The rest of this paper is organized as follows: In Section 2 we describe the model in detail, and give some definitions and notation. In Section 3 we determine a generating function of a steady-state, joint queue-length distribution using a solution of an iterative functional equation. In Section 4 we analyze important performance measures such as the response time (also called system time or sojourn time) and the waiting time in each priority class, and give some numerical examples. In Section 5 we summarize the paper and further research.

2. MODEL AND NOTATION

The two-class priority model with time-limited schedule analyzed in this paper consists of two-parallel queues with infinite capacity waiting rooms, Q_1 and Q_2, for messages of class-1 and class-2, respectively. The arrivals of class-n messages form a Poisson process with rate λ_n, $n = 1, 2$. Messages in Q_1 and Q_2 are served according to the time-limited schedule $(T_1, T_2 \leq \infty)$ as follows: Once starting service of class-2 messages, a single-server serves class-2 messages until either Q_2 becomes empty or a timer with maximum-attendance-time T_2 expires, whichever occurs first. In the latter case, the interrupted service is resumed in the next service period. In succession to the time-limited service with $T_2 \leq \infty$, class-1 messages,

if any, are next served according to the time-limited service with $T_1 \leq \infty$. Here note that class-1 messages are served until Q_1 becomes empty successively, because class-1 messages have priority over class-2 messages at completion of the time-limited service, i.e. the parameter is equivalent to $(T_1 = \infty, T_2 \leq \infty)$. While, if there are no class-1 messages at completion of the time-limited service with T_2, class-2 messages are served according to the next time-limited service with a new value T_2. If there is no message in the system, the server waits for a new arrival. Messages of the same class are served according to the FCFS (first-come first-served) discipline.

Service time H_n, $n = 1, 2$ for class-n messages has an exponential distribution, $H_n(t)$, with service rate μ_n, $n = 1, 2$. The maximum-attendance-time T_2 has also an exponential distribution, $T_2(t)$, with rate $\alpha \geq 0$. The Laplace-Stieltjes transform (LST) and the first moment of $H_n(t)$ are denoted by $H_n^*(s)$ and h_n, $n = 1, 2$, respectively. The LST of the distribution function (DF) $T_2(t)$ is denoted by $T_2^*(s)$. Throughout, we will use

$$H_n^*(s) = \frac{\mu_n}{s+\mu_n}, \qquad h_n = \frac{1}{\mu_n} \qquad n = 1, 2,$$
$$T_2^*(s) = \frac{\alpha}{s+\alpha}, \qquad E(T_2) = \frac{1}{\alpha} \tag{1}$$
$$\rho_n := \lambda_n h_n, \quad n = 1, 2, \qquad \rho := \rho_1 + \rho_2 < 1.$$

Additional notation will be introduced in Sections 3 and 4. Here note that the distribution of a busy period (or the workload process in the system) of our priority model is identical with that of an $M/G/1$ queue with the arrival rate $\lambda := \lambda_1 + \lambda_2$ and the LST for the service time $(\lambda_1 H_1^*(s) + \lambda_2 H_2^*(s))/\lambda$, i.e. the mean service time $h := (\lambda_1 h_1 + \lambda_2 h_2)/\lambda$. Therefore, from the $M/G/1$ queueing theory, that $\lambda h = \rho < 1$ is a necessary and sufficient condition for system stability.

***Remark* 2.1.** The above priority model represents a limiting case of the alternating service queues with $(T_1, T_2 \leq \infty)$ analyzed by Coffman et al. [5]. However, their results require still more a substantial effort to obtain the numerical solution of integral and functional equations, and no waiting time analysis is provided. In contrast, the above model becomes tractable by a classical but different analysis of functional relationships (9) and (10) as discussed in the next section. In addition, the above model does not fall within the class of queues with service

interruptions or breakdowns analyzed by Sengupta [16], because of the state dependent mechanism for switching from one queue to the other. ∎

3. GENERATING FUNCTION ANALYSIS

3.1 Functional Relationships

We first define the steady-state joint probabilities

$$p_n(i, j) := \Pr\{\text{server at } Q_n,\ i \text{ messages in } Q_1 \text{ and } j \text{ messages in } Q_2\} \quad \text{for } n = 1, 2 \text{ and } i + j \geq 1,$$

$$p(0, 0) := \Pr\{\text{there is no message in the system}\},$$

where the number of messages i (or j) includes the one being in service. Here, note that $p_1(0, j) = p_2(i, 0) = 0$ for any $i, j > 0$. We also define its generating functions, for $|x|, |y| \leq 1$,

$$P_1(x, y) := \sum_{i=1}^{\infty} \sum_{j=0}^{\infty} p_1(i, j)\, x^{i-1} y^j, \tag{2a}$$

$$P_2(x, y) := \sum_{i=0}^{\infty} \sum_{j=1}^{\infty} p_2(i, j)\, x^i y^{j-1}. \tag{2b}$$

In addition, define an indicator function $\delta_S = 1$ if S holds and $\delta_S = 0$ otherwise, where S denotes membership in some subset of $\{(i, j): i, j \geq 0\}$. Then, we obtain the following balance equations for the joint queue-length distribution $\{p_n(i, j)\}$:

$$(\lambda_1 + \lambda_2 + \mu_1)p_1(i, j) = \delta_{j>0}\alpha p_2(i, j) + \mu_1 p_1(i+1, j) + \delta_{i>1}\lambda_1 p_1(i-1, j) + \delta_{j>0}\lambda_2 p_1(i, j-1) + \delta_{j=0}\mu_2 p_2(i, 1) + \delta_{i=1}\delta_{j=0}\lambda_1 p(0, 0), \quad \text{for } i \geq 1, j \geq 0, \tag{3a}$$

$$(\lambda_1 + \lambda_2 + \mu_2 + \delta_{i>0}\alpha)p_2(i, j) = \mu_2 p_2(i, j+1) + \delta_{i>0}\lambda_1 p_2(i-1, j) + \delta_{j>1}\lambda_2 p_2(i, j-1) + \delta_{i=0}\mu_1 p_1(1, j) + \delta_{i=0}\delta_{j=1}\lambda_2 p(0, 0), \quad \text{for } i \geq 0, j \geq 1, \tag{3b}$$

$$(\lambda_1 + \lambda_2)p(0, 0) = \mu_1 p_1(1, 0) + \mu_2 p_2(0, 1). \tag{3c}$$

After some algebraic manipulation using (2), (3a)-(3c), we obtain the following functional relationships:

$$xR_1(x, y)P_1(x, y) = \alpha y[P_2(x, y) - P_2(0, y)] - \mu_1 P_1(0, y) + \mu_2[P_2(x, 0) - P_2(0, 0)] + \lambda_1 x p(0, 0), \tag{4}$$

$$y[\alpha + R_2(x, y)]P_2(x, y) = \mu_1[P_1(0, y) - P_1(0, 0)] + \alpha y P_2(0, y) - \mu_2 P_2(x, 0) + \lambda_2 y\, p(0, 0), \tag{5}$$

where

$$R_1(x, y) := \beta(x, y) - \frac{\mu_1}{x}(1\text{-}x), \qquad R_2(x, y) := \beta(x, y) - \frac{\mu_2}{y}(1\text{-}y), \tag{6}$$

and

$$\beta(x, y) := \lambda_1(1\text{-}x) + \lambda_2(1\text{-}y). \tag{7}$$

Putting $x = 0$ and $y = 0$ in (5) and (4), respectively, and using (3c) yield

$$\mu_1 P_1(0, y) = yR_2(0, y)P_2(0, y) + \beta(0, y)p(0, 0), \tag{8a}$$

$$\mu_2 P_2(x, 0) = xR_1(x, 0)P_1(x, 0) + \beta(x, 0)p(0, 0). \tag{8b}$$

From rearranging after substituting (8a) and (8b) for $P_1(0, y)$ and $P_2(x, 0)$ on the right-hand sides of (4) and (5), respectively, we get

$$P_1(x, y) = \frac{1}{xR_1(x, y)[\alpha + R_2(x, y)]}[R_2(x, y)\{xR_1(x, 0)P_1(x, 0) - y(\alpha + R_2(0, y))P_2(0, y)\} + \mu_1 R_2(x, y)P_1(0, 0) + \{\lambda_1 \alpha x - \beta(0, y)(\alpha + R_2(x, y))\}p(0, 0)], \tag{9}$$

$$P_2(x, y) = \frac{1}{y[\alpha + R_2(x, y)]}[y(\alpha + R_2(0, y))P_2(0, y) - xR_1(x, 0)P_1(x, 0) + \mu_2 P_2(0, 0) - \beta(x, 0)p(0, 0)]. \tag{10}$$

These equations are the starting point for our analysis. It is necessary to determine unknown functions $P_1(x, 0)$ and $P_2(0, y)$ and unknown probabilities $P_1(0, 0)$, $P_2(0, 0)$ and $p(0, 0)$ in the numerators on the right-hand sides of (9) and (10). The factors in the denominators on the right-hand sides of (9) and (10), $xR_1(x, y)$ and $y[\alpha + R_2(x, y)]$ are called "*kernels*", the zeros of which in the unit circle play an essential role in what follows.

***Remark* 3.1.** The definition of (2) leads to simplification of (9) and (10) as follows: If we use the ordinary generating function $P_1(x, y)$ defined by

$$P_1(x, y) := \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_1(i, j)\, x^i y^j, \tag{11a}$$

the right-hand sides of (4) and (5) have the terms $P_1(x, 0)$ and

$$P_{1x}'(0, y) := \left[\frac{\partial}{\partial x} P_1(x, y)\right]_{x=0.} \tag{11b}$$

■

3.2 Determination of $P_1(x, 0)$ and $P_2(0, y)$

The kernel appearing in (9) is rewritten as

$$xR_1(x, y) = (\mu_1 + \beta(x, y))[x - H_1^*(s_1 + \lambda_1(1-x))], \quad s_1 := \lambda_2(1-y). \tag{12}$$

Therefore, $x = H_1^*(s_1 + \lambda_1(1-x))$ has only one root $x = \delta_1(y)$ in the unit circle $|x| \le 1$ given by

$$\delta_1(y) := \frac{1}{2\lambda_1}\left[\mu_1 + \beta(0, y) - \sqrt{\{\mu_1+\beta(0, y)\}^2 - 4\lambda_1\mu_1}\right] \tag{13}$$

under the condition $\rho_1 \le 1$ and $|y| \le 1$, see Takács' lemma in [17]. Similarly, the other equation derived from the kernel in (10),

$$y[\alpha + R_2(x, y)]$$
$$= (\mu_2 + \alpha + \beta(x, y))\,[y - H_2^*(s_2 + \lambda_2(1-y))] = 0, \qquad s_2 := \alpha + \lambda_1(1-x), \tag{14}$$

has only one root, $y = \delta_2(x)$, in the unit circle $|y| \le 1$ as

$$\delta_2(x) := \frac{1}{2\lambda_2}\left[\mu_2 + \alpha + \beta(x, 0) - \sqrt{\{\mu_2+\alpha+\beta(x, 0)\}^2 - 4\lambda_2\mu_2}\right]. \tag{15}$$

From the regularity of $P_1(x, y)$ and $P_2(x, y)$, the numerators on the right-hand sides of (9) and (10) should be equal to zero for $x = \delta_1(y)$ and $y = \delta_2(x)$, respectively. Thus, we obtain the following functional relationships between $P_1(x, 0)$ and $P_2(0, y)$: For $|x|, |y| \le 1$,

$$R_2(\delta_1(y), y)[\delta_1(y)R_1(\delta_1(y), 0)P_1(\delta_1(y), 0) - y(\alpha+R_2(0, y))P_2(0, y)]$$
$$+ \mu_1 R_2(\delta_1(y), y)P_1(0, 0) + [\lambda_1\alpha\delta_1(y) - \beta(0, y)(\alpha+R_2(\delta_1(y), y))]p(0, 0) = 0, \tag{16}$$

$$\delta_2(x)(\alpha + R_2(0, \delta_2(x))P_2(0, \delta_2(x)) - xR_1(x, 0)P_1(x, 0) + \mu_2 P_2(0, 0)$$
$$- \beta(x, 0)p(0, 0) = 0. \tag{17}$$

Here, eliminating $P_2(0, \delta_2(x))$ from (16) and (17) after setting $y = \delta_2(x)$ in (16), we have

$$\varphi[f(x)] - \varphi(x) = p(0, 0)g(x), \tag{18}$$

where

$$\varphi(x) := xR_1(x, 0)P_1(x, 0) + \mu_1 P_1(0, 0),$$
$$f(x) := \delta_1[\delta_2(x)], \tag{19}$$
$$g(x) := \beta(x, \delta_2(x)) + \frac{\alpha\beta(\delta_1(\delta_2(x)), \delta_2(x))}{R_2(\delta_1(\delta_2(x)), \delta_2(x))}.$$

Using the iterative scheme (Kuczma et al. [10]), $\varphi(x)$ can be expressed as

$$\varphi(x) = \eta - p(0, 0)\sum_{i=0}^{\infty} g[\sigma_i(x)], \tag{20}$$

where

$$\sigma_0(x) := x, \quad 1 \geq x \geq 0, \quad \sigma_{i+1}(x) := f[\sigma_i(x)], \quad i = 0, 1, 2, \cdots. \tag{21}$$

The constant η in (20) is independent of the sequence $\{\sigma_i(x)\}$. From a boundary condition,

$$\varphi(0) = 0 \tag{22}$$

which follows from (19), the constant η can be determined. Finally, we have

$$\varphi(x) = p(0, 0)G(x), \tag{23}$$

where

$$G(x) := \sum_{i=0}^{\infty} [g(\sigma_i(0)) - g(\sigma_i(x))]. \tag{24}$$

Using (16) and $\varphi(x)$ determined above, we can find the other function $P_2(0, y)$ with unknown probabilities.

3.3 Calculation of Unknown Probabilities

The remaining work for us is to find the unknown probabilities appearing in (9) and (10). Letting $x, y \to 1$ in (9) and (10), and using the normalizing condition, we have simultaneous linear equations with nine unknowns, from which we get

$$\begin{aligned} &p(0, 0) = 1 - \rho, \\ &P_1(0, 0) = (1 - \rho)G(\delta_1(0))/\mu_1, \\ &P_2(0, 0) = (1 - \rho)\{\lambda_1 + \lambda_2 - G(\delta_1(0))\}/\mu_2. \end{aligned} \tag{25}$$

Finally, this completes the formulas to obtain the generating functions $P_1(x, y)$ and $P_2(x, y)$, leading to the following result:

***Theorem* 1.** For $\alpha \geq 0$,

$$P_1(x, y) = \frac{(1-\rho)R_2(x, y)}{xR_1(x, y)(\alpha+R_2(x, y))} \times\left[G(x) - G(\delta_1(y)) - \frac{\alpha\beta(x, y)}{R_2(x, y)} + \frac{\alpha\beta(\delta_1(y), y)}{R_2(\delta_1(y), y)}\right], \tag{26}$$

$$P_2(x, y) = \frac{1-\rho}{y(\alpha+R_2(x, y))}\left[G(\delta_1(y)) - G(x) - \beta(x, y) - \frac{\alpha\beta(\delta_1(y), y)}{R_2(\delta_1(y), y)}\right]. \tag{27}$$

□

4. ANALYSIS OF PERFORMANCE MEASURES

4.1 Mean Response Time and Mean Waiting Time

Let $E(\Theta_n)$ and $E(W_n)$, $n = 1, 2$ denote the expectations of response time Θ_n and waiting time W_n (until beginning service) of class-n messages, $n = 1, 2$, respectively, where the response time means the total time spent by a message in the system, also called system time or sojourn time. Then, we get the following mean delay formulas:

Theorem 2. For $\alpha \neq 0$,

$$E(\Theta_1) = \frac{1}{\alpha} + \frac{h_1}{1-\rho_1} - \frac{(1-\rho)h_1}{\alpha\rho_1(1-\rho_1)} G'(1), \tag{28}$$

$$E(\Theta_2) = \frac{\rho_1\rho_2 h_1 + (1-\rho_1)^2 h_2}{\rho_2(1-\rho)(1-\rho)} - \frac{\rho_1 + \alpha h_2}{\alpha\rho_2} + \frac{(1-\rho)h_1}{\alpha\rho_2(1-\rho_1)} G'(1), \tag{29}$$

where

$$G'(1) = -\sum_{i=0}^{\infty} g'(\sigma_i(1)) \prod_{j=0}^{i-1} f'(\sigma_j(1)) \tag{30}$$

and the null product is unity.

$$E(W_n) = E(\Theta_n) - E(C_n), \qquad n = 1, 2, \tag{31}$$

where

$$E(C_1) := h_1, \tag{32}$$

$$E(C_2) := h_2 + \left[\frac{\rho_1}{\rho_2} - \frac{(1-\rho)h_1}{\rho_2(1-\rho_1)} G'(1)\right] h_2. \tag{33}$$

Proof: Let $L_n(x)$ and $Q_n(x)$, $n = 1, 2$ be the generating functions for the number of class-n messages present in the system and the number of messages waiting in Q_n, $n = 1, 2$, respectively. Then, we get

$$L_1(x) = p(0, 0) + \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_1(i, j)\, x^i + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} p_2(i, j)\, x^i$$

$$= p(0, 0) + xP_1(x, 1) + P_2(x, 1) \tag{34}$$

and

$$L_2(y) = p(0, 0) + P_1(1, y) + yP_2(1, y). \tag{35}$$

Little's result,

$$L_n'(1) = \lambda_n E(\Theta_n), \quad n = 1, 2, \tag{36}$$

and the extensive calculation using L'Hospital's rule for derivative terms obtained by differentiating the right-hand sides of (34) and (35) yield (28) and (29). Similarly, we get

$$Q_1(x) = p(0, 0) + \sum_{i=1}^{\infty}\sum_{j=0}^{\infty} p_1(i, j)\, x^{i-1} + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} p_2(i, j)\, x^i$$
$$= p(0, 0) + P_1(x, 1) + P_2(x, 1). \tag{37}$$

Little's result,

$$Q_1'(1) = \lambda_1 E(W_1), \tag{38}$$

leads to (31) and (32) since we know $P_n(1, 1) = \rho_n$, $n = 1, 2$.

Here, it should be noted that for calculating $E(W_2)$, we can not use the same way used to obtain the above results, since the number of class-2 messages waiting in Q_2 is not j for the state probability $p_1(i, j)$ if a class-2 message is in service-interruption due to the timer expiration. Hence we have to find the following state-probabilities defined as

$$p_1(i, j)_{IR} := \Pr\{\text{server at } Q_1,\ i \text{ messages in } Q_1,\ j \text{ messages in } Q_2 \text{ and a class-2 message is in service-interruption}\},$$

$$p_1(i, j)_{\overline{IR}} := \Pr\{\text{server at } Q_1,\ i \text{ messages in } Q_1,\ j \text{ messages in } Q_2 \text{ and any message of class-2 is not interrupted}\}$$

which satisfy

$$p_1(i, j)_{IR} + p_1(i, j)_{\overline{IR}} = p_1(i, j) \qquad \text{for } i, j \geq 1 \tag{39}$$

and its generating functions, for $|x|, |y| \leq 1$,

$$P_1(x, y)_{IR} := \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_1(i, j)_{IR}\, x^{i-1} y^j,$$
$$P_1(x, y)_{\overline{IR}} := \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_1(i, j)_{\overline{IR}}\, x^{i-1} y^j. \tag{40}$$

From Appendix and (39), we have

$$P_1(x, y)_{IR} = \frac{\alpha y}{xR_1(x, y)} [P_2(x, y) - P_2(\delta_1(y), y)], \tag{41}$$

$$P_1(x, y)_{\overline{IR}} = P_1(x, y) - P_1(x, 0) - P_1(x, y)_{IR}. \tag{42}$$

Consequently we obtain

$$\begin{aligned} Q_2(y) &= p(0, 0) + \sum_{i=1}^{\infty} p_1(i, 0) + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_1(i, j)_{\overline{IR}}\, y^j + \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_1(i, j)_{IR}\, y^{j-1} \\ &\quad + \sum_{i=0}^{\infty}\sum_{j=1}^{\infty} p_2(i, j)\, y^{j-1} \\ &= p(0, 0) + P_1(1, 0) + P_1(1, y)_{\overline{IR}} + \frac{1}{y} P_1(1, y)_{IR} + P_2(1, y) \end{aligned} \tag{43}$$

and

$$Q_2'(1) = L_2'(1) - P_2(1, 1) - P_1(1, 1)_{IR}. \tag{44}$$

Then, applying Little's formula to (44) yields (31) and (33), where we have used

$$P_1(1, 1)_{IR} = \rho_1 - \frac{(1-\rho)h_1}{1-\rho_1} G'(1) \tag{45}$$

which can be obtained from (41) and Theorem 1.

Here, $E(C_n)$, $n = 1, 2$ represents the so-called *mean completion time* for class-n messages. The completion time, C_n, $n = 1, 2$ is defined as the duration of a period that begins from the instant the service of a class-n message starts and ends at the instant the server becomes free to take the next message of class n, i.e. $\Theta_n = W_n + C_n$, $n = 1, 2$ (Jaiswal [7]). □

It is seen from Theorem 2 that the expressions for $E(\Theta_n)$ and $E(W_n)$, $n = 1, 2$ have terms with infinitive sum and infinitive product, however, we can provide delay formulas without infinitive terms only for the following boundary cases, $\alpha = 0$ (alternating-priority discipline) and $\alpha \to \infty$ (preemptive resume discipline), which are directly derived from the previous results of the $M/G/1$ queues (e.g., (4.4a) and (8.28a) in Chap. 3 in Takagi [18]):

***Corollary* 1.** For $\alpha = 0$,

$$E(\Theta_1) = \frac{h_1}{1-\rho_1} + \frac{\rho_2}{(1-\rho_1)(1-\rho)(1-\rho+2\rho_1\rho_2)}[\rho_1\rho_2 h_1 + (1-\rho_1)^2 h_2], \tag{46a}$$

$$E(\Theta_2) = \frac{h_2}{1-\rho_2} + \frac{\rho_1}{(1-\rho_2)(1-\rho)(1-\rho+2\rho_1\rho_2)}[(1-\rho_2)^2 h_1 + \rho_1\rho_2 h_2], \tag{46b}$$

$$E(W_n) = E(\Theta_n) - E(C_n), \qquad n = 1, 2, \tag{47}$$

$$E(C_n) = h_n, \qquad n = 1, 2 \tag{48}$$

and, for $\alpha \to \infty$,

$$E(\Theta_1) = \frac{h_1}{1-\rho_1}, \tag{49a}$$

$$E(\Theta_2) = \frac{\rho_1 h_1 + (1-\rho_1)h_2}{(1-\rho_1)(1-\rho)}, \tag{49b}$$

and

$$E(W_n) = E(\Theta_n) - E(C_n), \qquad n = 1, 2, \tag{50}$$

where

$$E(C_1) := h_1, \quad E(C_2) := \frac{h_2}{1-\rho_1}. \tag{51}$$

Proof: From (23), $G'(1) = \varphi'(1)/p(0, 0)$ and $G''(1) = \varphi''(1)/p(0, 0)$. Using these and setting $x = 1$ after differentiating both sides of (18) one or two times by x, we have, for $\alpha = 0$,

$$\begin{aligned} G'(1) &= \frac{\lambda_1(1-\rho_1)}{1-\rho}, \\ G''(1) &= \frac{2\lambda_1^2\rho_2}{(1-\rho)^2(1-\rho+2\rho_1\rho_2)}[\rho_1\rho_2 h_1 + (1-\rho_1)^2 h_2], \end{aligned} \tag{52}$$

where we have used that $\delta_2(1) = 1$ and $f(1) = 1$. Therefore, from (34), (35), (52) and differentiating $P_n(x, y)$, $n = 1, 2$ with respect to x or y after setting $\alpha = 0$ in

Theorem 1, we obtain (46a) and (46b). Substituting $G'(1)$ for (33) leads to (48). On the other hand, letting $\alpha \to \infty$ in (28) and (29) leads to (49a) and (49b), respectively. By setting $y = 0$ after letting $\alpha \to \infty$ in (26) in Theorem 1, we get

$$xR_1(x, 0)P_1(x, 0) = \varphi(x) - \mu_1 P_1(0, 0) = -(1-\rho)\beta(x, 0), \tag{53}$$

where we have used that $yR_2(\delta_1(y), y) \neq 0$ for $y = 0$. From $\varphi'(1) = (1-\rho)G'(1)$, we have

$$G'(1) = \lambda_1. \tag{54}$$

Therefore, substituting $G'(1)$ for (33) leads to (51). □

***Remark* 4.1.**

(i) Eqs. (28)-(33) in Theorem 2 and (46a)-(51) in Corollary 1 also hold for the work- conserving service discipline beyond the FCFS discipline in each queue.
(ii) The mean response times $E(\Theta_n)$, $n = 1, 2$ in Theorem 2 and Corollary 1 satisfy the workload conservation law, respectively (Wolff [22]),

$$\sum_{n=1}^{2} \rho_n E(\Theta_n) = \frac{1}{1-\rho}(\rho_1 h_1 + \rho_2 h_2). \tag{55}$$

It can also be confirmed that the mean waiting times $E(W_n)$, $n = 1, 2$ satisfy the conservation law,

$$\sum_{n=1}^{2} \rho_n E(W_n^{\#}) = \frac{\rho}{1-\rho}(\rho_1 h_1 + \rho_2 h_2) \tag{56}$$

by setting as

$$E(W_1^{\#}) := E(W_1), \tag{57a}$$

$$E(W_2^{\#}) := E(W_2) + E(C_2) - h_2, \tag{57b}$$

where $\rho_2 E(W_2)$ represents the mean workload associated with waiting messages in Q_2, while $\rho_2(E(C_2) - h_2)$, the mean unfinished workload associated with a message in service-interruption, since the probability of finding a class-2 message in service-interruption is $\lambda_2(E(C_2) - h_2) = P_1(1, 1)_{IR}$, which follows from (33) and (45) for $\alpha \geq 0$. ∎

4.2 LSTs and Higher Moments of the Response Time and the Waiting Time

Let $\Theta_n^*(s)$ and $W_n^*(s)$, $n = 1, 2$ be the LSTs of the distribution functions of the response time Θ_n and the waiting time W_n of class-n messages, respectively. Then, we obtain the following theorem:

***Theorem* 3.** For $\alpha \geq 0$,

$$\Theta_n^*(s) = L_n(1-s/\lambda_n), \tag{58}$$

$$W_n^*(s) = Q_n(1-s/\lambda_n), \qquad n = 1, 2, \tag{59}$$

where $L_n(x)$ and $Q_n(x)$, $n = 1, 2$ are given by (34), (35), (37) and (43), respectively; The mth ($m = 2, 3, \cdots$) moments of the response time and the waiting time are obtained by

$$\begin{aligned} E(\Theta_1^m) &= \frac{1}{\lambda_1^m}[mP_{1x}^{(m-1)}(1, 1) + P_{1x}^{(m)}(1, 1) + P_{2x}^{(m)}(1, 1)], \\ E(\Theta_2^m) &= \frac{1}{\lambda_2^m}[mP_{2y}^{(m-1)}(1, 1) + P_{2y}^{(m)}(1, 1) + P_{1y}^{(m)}(1, 1)] \end{aligned} \tag{60}$$

and

$$E(W_1^m) = \frac{1}{\lambda_1^m}[P_{1x}^{(m)}(1, 1) + P_{2x}^{(m)}(1, 1)], \tag{61}$$

$$E(W_2^m) = \frac{1}{\lambda_2^m}\left[P_{2y}^{(m)}(1, 1) + P_{1y}^{(m)}(1, 1)_{\overline{IR}} + m!\sum_{k=0}^{m}\frac{(-1)^{m-k}}{k!}P_{1y}^{(k)}(1, 1)_{IR}\right],$$

where, for $k = 0, 1, 2, \cdots$,

$$P_{nx}^{(k)}(1, 1) := \left[\frac{\partial^k}{\partial x^k}P_n(x, y)\right]_{x=y=1} \tag{62}$$

and $P_n(x, y)$, $n = 1, 2$, $P_1(x, y)_{IR}$ and $P_1(x, y)_{\overline{IR}}$ are given in Theorem 1, (41) and (42), respectively.

Proof: Since the sample paths of queue-length for each priority class are step functions with upward/downward unit jumps, the generating function for the

number of class-n messages in the system at a class-n message departure is identical to $L_n(x)$, $n = 1, 2$ given by (34) and (35) because of the PASTA property (Wolff [22]) and Finch's departure theorem (or Burke's result in Cooper [6]). Furthermore, from the fact that, under the FCFS discipline in each queue, the number of class-n messages left behind by the departing class-n message is equal to the number of class-n messages that arrive while it has been waiting and in service, we have

$$\Theta_n^*(\lambda_n(1-x)) = L_n(x), \qquad n = 1, 2 \tag{63}$$

which leads to (58). Therefore, from (34) and (35), we get (60). Likewise, from the above argument on the number of class-n messages in the waiting room, we also get

$$W_n^*(\lambda_n(1-x)) = Q_n(x), \qquad n = 1, 2 \tag{64}$$

which leads to (59). Hence, from (37) and (41)-(43), we get (61). □

***Remark* 4.2.** The LST for the completion time of class-2 messages, $C_2^*(s)$, is not given by $C_2^*(s) = \Theta_2^*(s)/W_2^*(s)$, since C_2 depends on W_2. (Recall that at a service starting epoch for a class-2 message, the workload in Q_1 is not always zero as the ordinary preemptive/non-preemptive priority queues). ∎

4.3 Numerical Examples

In getting the numerical results for the mean performance measures using Theorem 2, our main work is the calculation of $G'(1)$ given by (30). Accordingly we need computer programming for the iterative calculation based on (21), however, it has been confirmed that the convergence of the sequence $\{\sigma_i(x)\}$ is very rapid. Table 1 shows values of the mean response times $E(\Theta_n)$, $n = 1, 2$ for the server utilization $\rho = 0.2$ to 0.9 and the mean maximum-attendance-time $E(T_2) = 1/\alpha = 0.01$, 1.00 and 100, where the service times H_n, $n = 1, 2$ and the maximum-attendance-time T_2 are exponentially distributed, $h_1 = h_2 = 1$ and $\lambda_1 = \lambda_2$. Under the same condition with Table 1, Table 2 shows values of the mean completion time $E(C_2)$ as a function of $E(T_2)$ for $\rho = 0.2$ to 0.9. For $E(T_2) = 0.01$, 1.00, 100 and $\rho = 0.2$ to 0.9, $E(W_n) = E(\Theta_n) - E(C_n)$, $n = 1, 2$ are obtained from Tables 1 and 2, where $E(C_1) = h_1 = 1$. In the case of $E(T_2) = 0.0$, the value of

$E(C_2)$ is identical with that of $E(C_2) = h_2/(1-\rho_1)$ given by (51) for the ordinary preemptive-resume priority queues.

It is seen from Tables 1 and 2 that $E(\Theta_n)$ and $E(W_n)$, $n = 1, 2$ can be widely changed by $E(T_2)$, especially in the case of high server utilization. That is, we can select an appropriate value of the controllable parameter $\alpha = 1/E(T_2)$ in order to optimize the mean performance measures.

Table 1 Mean response times $E(\Theta_n)$, $n = 1, 2$ as a function of the server utilization ρ

$E(T_2)$	0.01		1.00		100	
$E(\Theta_n)$	$E(\Theta_1)$	$E(\Theta_2)$	$E(\Theta_1)$	$E(\Theta_2)$	$E(\Theta_1)$	$E(\Theta_2)$
$\rho = 0.2$	1.1122	1.3878	1.1715	1.3285	1.2481	1.2519
0.4	1.2525	2.0809	1.4008	1.9325	1.6590	1.6743
0.6	1.4328	3.5672	1.7237	3.2763	2.4721	2.5279
0.8	1.6733	8.3267	2.2091	7.7909	4.8525	5.1475
0.9	1.8264	18.174	2.5224	17.448	9.3706	10.629

Table 2 Mean completion time $E(C_2)$ as a function of mean maximum-attendance-time $E(T_2)$ for the server utilization $\rho = 0.2$ to 0.9

$E(T_2)$	0.0	0.01	0.1	1.0	10	100	∞
$\rho = 0.2$	1.1111	1.1101	1.1022	1.0604	1.0122	1.0014	1.0000
0.4	1.2500	1.2482	1.2335	1.1508	1.0352	1.0041	1.0000
0.6	1.4286	1.4264	1.4082	1.2951	1.0843	1.0104	1.0000
0.8	1.6667	1.6649	1.6496	1.5424	1.2225	1.0319	1.0000
0.9	1.8182	1.8171	1.8171	1.7342	1.4164	1.0755	1.0000

5. CONCLUSIONS & FURTHER RESEARCH

For the two-class Markovian priority queues $(M_1, M_2/M_1, M_2/1)$ with preemptive-resume, time-limited schedule $(T_1 = \infty, T_2 \leq \infty)$, we have derived the generating function of a joint queue-length distribution, and have obtained LSTs for the response time and the waiting time in each queue. Besides, explicit mean delay formulas have been provided for the performance measures and the completion time. From some numerical examples, we have confirmed the effectiveness of the time-limited schedule.

As the subjects of future research, we may consider (i) three or more priority queues $(N \geq 3)$ with time-limited schedule, (ii) two-class priority queues with general distributions of T_2 and H_n, $n = 1, 2$ and general arrival processes, (iii) the same priority queues with nonpreemptive, time-limited schedule, and (iv) discrete-time priority queues with time-limited schedule. For the first subject, in the case of $N = 3$, we need a solution of a functional equation with *two variables* corresponding to (18), which may be new in the literature, in order to find the joint queue-length distribution. For the second subject, the same approach used for the generating function analysis in this study can be applied to priority models with general maximum-attendance-time distribution, e.g. $M_1, M_2/G_1, M_2/1$, by using the method of supplementary variables. For generalization of the arrival process, we need furthermore the busy period analysis, e.g. Machihara [14] and Takine and Hasegawa [20], by taking account of the fact stated in Remark 4.2, though previous works for the ordinary preemptive priority queues with non-Poisson arrival processes are closely related to this subject. For the third subject, the results of Katayama and Takahashi [8] for priority queues with Bernoulli schedules $[p_1, p_2]$ can be directly applied to a two-class priority model $(M_1, M_2/G_1, G_2/1)$ with general service time distributions and *nonpreemptive*, time-limited schedule $(T_1 = \infty, T_2 \leq \infty)$ distributed exponentially by setting the Bernoulli parameters $[p_1 = 1, p_2 = \Pr\{H_2 < T_2\} = H_2^*(\alpha)]$. Hence, analysis of the multi-class priority queues with nonpreemptive, time-limited schedule $(T_1, T_2, \cdots, T_N)$ distributed exponentially reduces to that of the priority queues with Bernoulli parameters $[p_1, p_2, \cdots, p_N]$. For the fourth subject, we need also to study the discrete-time versions of the above models. Indeed, as the other research direction, we may apply numerical techniques based on the Laguerre-function approximation developed by Leung et al. [11, 12] and discrete Fourier transforms used in [13, 19]

to analysis of our priority queues, and we need also to study the numerical analysis for the results of Coffman et al. [5] obtained by the boundary-value technique as mentioned in Remark 2.1.

Acknowledgments

The author is indebted to Mr. Y. Nakashima and Mr. A. Yamada for performing numerical calculations for Tables 1 and 2, and is also grateful to the anonymous referees of the paper for many valuable comments.

References

[1] A.S. Alfa: "A discrete *MAP/PH/*1 queue with vacations and exhaustive time-lilmited service," *Oper. Res. Letters*, **18** (1995) 31-40.

[2] K.C. Chang and D. Sandhu: "Delay analysis of token-passing protocols with limited token holding times," *IEEE Trans. on Communications*, **COM-42** (1994) 2833-2842.

[3] J. Chiarawongse, M.M. Srinivasan and T.J. Teorey: "The *M/G/*1 queueing system with vacations and timer-controlled service," *IEEE Trans. on Communications*, **COM-42** (1994) 1846-1855.

[4] On-Ching Yue and C.A. Brooks "Performance of the timed token scheme in *MAP*," *IEEE Trans. on Communications*, **COM-38** (1990) 1006-1012.

[5] E.G. Coffman, Tr, G. Fayolle and I. Mitrani: "Two queues with alternating service periods," *Performance '87*, P.-J. Courtois and G. Latouche, Eds. New York: Elsevier North-Holland (1987) 227-239.

[6] R.B. Cooper: "*Introduction to Queueing Theory*," Third edition (Washington, CEEPress Books, The George Washington University, 1990).

[7] N.K. Jaiswal: "*Priority Queues*," (New York, Academic Press, 1968).

[8] T. Katayama and Y. Takahashi: "Analysis of a two-class priority queue with Bernoulli schedules," *J. Oper. Res. Soc. Japan*, **35** (1992) 236-249.

[9] M. Komatsu and Y. Hisamoto: "A composite priority queue with maximum limited of pre-emptive processing time," *Trans. of IEICE*, **J 72-A** (1989) 1700-1703 (in Japanese).

[10] M. Kuczma, B. Choczewski and R. Ger: "*Iterative Functional Equations*," Encyclopedia of Mathematics and its Applications **Vol. 32** (Cambridge University Press, 1990).

[11] K.K. Leung and M. Eisenberg: "A single-server queue with vacations and gated time-limited service," *IEEE Trans. on Communications*, **COM-38** (1990) 1454-1462.

[12] K.K. Leung and M. Eisenberg: "A single-server queue with vacations and non-gated time-limited service," *Performance Evaluation*, **12** (1991) 115-125.

[13] K.K. Leung: "Cyclic-service systems with nonpreemptive, time-limited service," *IEEE Trans. on Communications*, **COM-42** (1994) 2521-2524.

[14] F. Machihara: "On the queue with *PH*-Markov renewal preemptions," *J. Oper. Res. Soc. Japan*, **36** (1993) 13-28.

[15] M. de Prycker: "*Asynchronous Transfer Mode, Solution for Broadband ISDN*," Third edition (Prentice Hall, 1995).

[16] B. Sengupta: "A queue with service interruptions in an alternating random environment," *Oper. Res.*, **38** (1990) 308-318.

[17] L. Takács: "*Introduction to the Theory of Queues*," (New York, Oxford University Press, 1962).

[18] H. Takagi, "*Queueing Analysis: A Foundation of Performance Evaluation Vol. 1, Vacation and Priority Systems, Part 1*," (Elsevier Science Publisher B.V., North-Holland, 1991).

[19] H. Takagi and K.K. Leung: "Analysis of a discrete-time queueing system with time- limited service," *Queueing Systems*, **18** (1994) 183-197.

[20] T. Takine and T. Hasegawa: "The workload in the *MAP/G/*1 queue with state-dependent services: Its applications to a queue with preemptive resume priority," *Stochastic Models*, **10** (1994) 183-204.

[21] M. Tangemann and K. Sauer: "Performance analysis of the timed token protocol of *FDDI* and *FDDI*-II," *IEEE J. on Selected Areas in Communications*, **SAC-9** (1991) 271-278.

[22] R.W. Wolff: "*Stochastic Modeling and the Theory of Queues*," (Prentice-Hall International Editions, 1989).

Appendix Derivation of $P_1(x, y)_{IR}$

We obtain a balance equation for $\{p_1(i, j)_{IR}\}$ as

$$(\lambda_1 + \lambda_2 + \mu_1)p_1(i, j)_{IR} = \lambda_1 p_1(i\text{-}1, j)_{IR} + \lambda_2 p_1(i, j\text{-}1)_{IR} + \mu_1 p_1(i\text{+}1, j)_{IR} + \alpha p_2(i, j) \quad \text{for } i, j \geq 1, \tag{A.1}$$

where the last term corresponds the timer expiration with rate α when the single-server serves a class-2 message. Some algebraic manipulation using (2), (40) and (A.1) yield

$$xR_1(x, y)P_1(x, y)_{IR} = \alpha y\,[P_2(x, y) - P_2(0, y)] - \mu_1 P_1(0, y)_{IR}. \tag{A.2}$$

Since the right-hand side of (A.2) is also zero for $x = \delta_1(y)$ given by (13), it follows that

$$\mu_1 P_1(0, y)_{IR} = \alpha y[P_2(\delta_1(y), y) - P_2(0, y)]. \tag{A.3}$$

Therefore, we obtain $P_1(x, y)_{IR}$ given by (41) from (A.2) and (A.3).