On the single server retrial queue with priority customers

G.I. Falin[1], J.R. Artalejo and M. Martin

Department of Statistics and Operations Research, Mathematics Faculty, Madrid University – Complutense, Madrid 28040, Spain

Received 3 April 1992; revised 16 March 1993

We consider an $M_2/G_2/1$ type queueing system which serves two types of calls. In the case of blocking the first type customers can be queued whereas the second type customers must leave the service area but return after some random period of time to try their luck again. This model is a natural generalization of the classic $M_2/G_2/1$ priority queue with the head-of-the-line priority discipline and the classic $M/G/1$ retrial queue. We carry out an extensive analysis of the system, including existence of the stationary regime, embedded Markov chain, stochastic decomposition, limit theorems under high and low rates of retrials and heavy traffic analysis.

Keywords: Priority queues; head-of-the-line priority discipline; retrials; limit theorems; stochastic decomposition.

1. Introduction

The so-called retrial queueing systems are characterized by the requirement that customers finding the service area busy must join the retrial group and reapply for service in random order and at random intervals. These models arise frequently in the analysis of telephone and other communications systems. A review of the literature on this topic can be found in Falin [4] and Yang and Templeton [9].

Most retrial queues assume that the input process is homogeneous from the point of view of call characteristics such as the service time and the inter-retrial time distributions. In practice, however, these characteristics may differ widely for different subscriber groups. This leads us to multiclass retrial queues. In a general description we can consider n types of customers. Type i primary customers arrive according to a homogeneous Poisson stream with rate λ_i, the associated retrial intensity is μ_i and the service time distribution function is $B_i(x)$ (for more details see Kulkarni [8] and Falin [3]).

[1] Visiting from: Department of Probability, Mechanics and Mathematics, Moscow State University, Moscow 119899, Russia.

Such models are essentially more difficult than the single class models, so explicit results are available only in a few special cases. Kulkarni and Falin obtained explicit formulas for the first and second moments of the number of sources of repeated calls.

The extreme case of $\mu_i \equiv \mu$ was considered by Hanschke [5]. Recently, Choi and Park [1] investigated a priority retrial queue which, in fact, can be considered as another extreme case of the above multiclass retrial queue by taking $\mu_1 \to \infty$ and $n = 2$.

The latter extreme case is of special interest for practical applications. Khalil et al. [7] have studied this situation in full detail, at a Markovian level, in the context of telephone exchanges serving outgoing and incoming customers. The outgoing calls can be queued whereas blocked incoming calls are initially rejected, but after some random time repeat their demands.

From the point of view of the study of the number of customers in the system, the extreme case $\mu_1 \to \infty$ and $n = 2$ corresponds to a queue with two priority levels. If a priority unit arrives when a non-priority unit is being serviced, it may wait till the non-priority unit completes service, i.e. a variant of postponable (or head-of-the-line) priority discipline is considered.

Choi and Park studied only the distribution of the number of customers in the two groups when $B_1(x) = B_2(x)$. So our main objectives are:

(a) to consider a more general and natural case, assuming different service distributions for both types of customers, and
(b) to study of such a model in more depth. In particular, we investigate stochastic decomposition and asymptotic behaviour of stationary characteristics.

In section 2 we describe the model. The study of the embedded Markov chain at departure epochs and the joint distribution of the waiting line and the orbit is carried out in sections 3 and 4. In section 5 we represent our model as a convolution of two simple random vectors. The investigation of the asymptotic behaviour under heavy-traffic and high and low retrial intensities is undertaken in section 6. Throughout the paper we will show that our results are in agreement with those of Choi and Park [1] and with the well-known results for the classic $M/G/1$ retrial queue and the head-of-the-line priority queue studied by Jaiswal [6].

2. Model description

We consider a single server queueing system at which two different types of primary customers arrive according to independent Poisson streams with rates α and λ, respectively. Demands from the first flow, with rate α, can be identified as priority customers and they are queued after blocking and then served in some discipline such as FCFS or random order. On the other hand, any non-priority customer (those from the second flow) who finds the server busy upon arrival leaves the system immediately, to seek service again at subsequent epochs until he finds the ser-

ver free. The retrial times are assumed to be independent and exponentially distributed with parameter $\mu > 0$.

Both types of customers require a service time with distribution function $B_k(x)$, $k = 1, 2$, where the number "1" is associated with the priority customers. The input flows of primary arrivals, intervals between repeated trials and service times are assumed to be mutually independent.

Let

$$\beta_k(s) = \int_0^\infty e^{-sx}\, dB_k(x), \quad \beta_n^{(k)} = (-1)^n \beta_k^n(0),$$

$$K^{(k)}(y,z) = \beta_k(\alpha(1-y) + \lambda(1-z)), \quad b_k(x) = B_k'(x)/(1 - B_k(x)),$$

$$\sigma = \alpha\beta_1^{(1)}, \quad \rho = \lambda\beta_1^{(2)},$$

where $|y| \leqslant 1$, $|z| \leqslant 1$, $n \in \mathbb{N}$ and $k = 1, 2$.

The state of the system at time t can be described by the Markovian process $X(t) = (A(t), C(t), N(t), \xi(t))$, where $C(t)$ is the number of customers in queue (excluding the customer in service), $N(t)$ is the number of sources of repeated calls (or customers in orbit), $A(t)$ represents the type of the customers in service and $\xi(t)$ is the corresponding elapsed time. We assume that $A(t) = 0$ when no customer is in service at t, hence $A(t) = 0$ implies $C(t) = 0$.

3. Embedded Markov chain

Let η_d be the time of the dth departure. It is easy to see that a sequence of random vectors $X_d = (A(\eta_d - 0), C(\eta_d - 0), N(\eta_d - 0))$ forms a Markov chain, which is the embedded Markov chain for our queueing system. Its state space is $\{1, 2\} \times \mathbb{Z}_+^2$ and its one-step transition probabilities

$$r_{(k,n,m)(l,i,j)} = P\{X_{d+1} = (l, i, j) | X_d = (k, n, m)\}$$

are given by the formulas

$$r_{(k,n,m)(1,i,j)} = \begin{cases} \dfrac{\alpha}{\alpha + \lambda + m\mu} k^{(1)}_{i,j-m} & \text{if } n = 0, \\ k^{(1)}_{i-n+1,j-m} & \text{if } n \geqslant 1, \end{cases}$$

$$r_{(k,n,m)(2,i,j)} = \begin{cases} \dfrac{\lambda}{\alpha + \lambda + m\mu} k^{(2)}_{i,j-m} + \dfrac{m\mu}{\alpha + \lambda + m\mu} k^{(2)}_{i,j-m+1} & \text{if } n = 0, \\ 0 & \text{if } n \geqslant 1, \end{cases}$$

where

$$k_{ij}^{(k)} = \int_0^\infty \frac{(\alpha x)^i}{i!} e^{-\alpha x} \frac{(\lambda x)^j}{j!} e^{-\lambda x} \, dB_k(x)$$

is the probability that i priority and j non-priority units arrive at the system during a service interval of type k.

As usual, the first question to be investigated is the ergodicity of our chain.

THEOREM 1

The embedded Markov chain is ergodic if and only if

$$\sigma + \rho < 1 . \tag{1}$$

Proof

Obviously, condition (1) is necessary for ergodicity. Indeed, since customers cannot be lost, in the steady state carried traffic is equal to offered traffic. But offered traffic is $\sigma + \rho$ and carried traffic is equal to the mean number of busy channels, i.e. to the probability that the channel is busy. This probability is obviously less than 1 and thus $\sigma + \rho < 1$.

Now let $\sigma + \rho < 1$. To establish ergodicity we will use the classic Foster criteria: for an irreducible and aperiodic Markov chain X_d with state space S, a sufficient condition for ergodicity is the existence of a non-negative function $f(s)$, $s \in S$ (so-called test function or Lyapunov function), a positive number ϵ and a finite subset A of the state space S such that the mean drift

$$D_s = E[f(X_{d+1}) - f(X_d) | X_d = s]$$

is finite for all $s \in S$ and $D_s \leqslant -\epsilon$ for all $s \notin A$.

In our case, as the Lyapunov function we consider

$$f(k,n,m) = (\lambda \beta_1^{(1)} + 1 - \rho)n + (\alpha \beta_1^{(2)} + 1 - \sigma)m .$$

Then

$$D_{knm} = \begin{cases} \sigma + \rho - 1 + \dfrac{\alpha + \lambda}{\alpha + \lambda + m\mu} & \text{if } n = 0 , \\ \sigma + \rho - 1 & \text{if } n \geqslant 1 . \end{cases}$$

Let $\epsilon = (1 - \sigma - \rho)/2$ and $M_0 = (\alpha + \lambda)(1 - \epsilon)/\mu\epsilon$. Then, for all states with $k = 1,2$; $n \geqslant 1$; $m \geqslant 0$, we have $D_{knm} = -2\epsilon < -\epsilon$. Besides, for all states with $k = 1,2; n = 0; m \geqslant M_0$, we have

$$D_{knm} = -2\epsilon + \frac{\alpha + \lambda}{\alpha + \lambda + m\mu} \leqslant -2\epsilon + \frac{\alpha + \lambda}{\alpha + \lambda + M_0\mu} = -\epsilon .$$

Applying the above criteria we can guarantee that the chain is ergodic. □

Our second goal is to find the stationary distribution

$$\Pi_{ij}^{(k)} = \lim_{d\to\infty} P\{X_d = (k,i,j)\}.$$

Some information about this distribution can be obtained with the help of Lyapunov's function used to prove theorem 1. Namely, the well-known mean drift relation

$$\sum_{(k,i,j)} D_{kij}\Pi_{ij}^{(k)} = 0$$

becomes

$$\sum_{j=0}^{\infty} \frac{\Pi_{0j}}{\alpha+\lambda+j\mu} = \frac{1-\sigma-\rho}{\alpha+\lambda}, \tag{2}$$

where $\Pi_{0j} = \Pi_{0j}^{(1)} + \Pi_{0j}^{(2)}$.

The following theorem fully describes the stationary distribution of the embedded Markov chain.

THEOREM 2

The stationary distribution $\Pi_{ij}^{(k)}$ has the following partial generating functions $\Pi^{(k)}(y,z) = \sum_{i=0}^{\infty} y^i \sum_{j=0}^{\infty} z^j \Pi_{ij}^{(k)}$:

$$\Pi^{(1)}(y,z) = K^{(1)}(y,z) \times \frac{(\alpha-\alpha h(z)+\lambda-\lambda Q(z))(T(y,z)-z)+(\alpha-\alpha y+\lambda-\lambda T(y,z))(z-Q(z))}{(Q(z)-z)(y-K^{(1)}(y,z))} R(z), \tag{3}$$

$$\Pi^{(2)}(y,z) = K^{(2)}(y,z)\frac{\alpha-\alpha h(z)+\lambda-\lambda z}{Q(z)-z}R(z), \tag{4}$$

where $h(z)$ is the solution of the equation $\beta_1(\alpha-\alpha h(z)+\lambda-\lambda z)-h(z)=0$ in the unit disk $|h|\leqslant 1$ and

$$R(z) = \frac{1-\sigma-\rho}{\alpha+\lambda}\exp\left\{\frac{1}{\mu}\int_1^z \frac{\alpha(1-h(u))+\lambda(1-K^{(2)}(h(u),u))}{K^{(2)}(h(u),u)-u}\,du\right\},$$

$$T(y,z) = K^{(2)}(y,z), \quad Q(z) = K^{(2)}(h(z),z). \tag{5}$$

Proof

Using the above formulas for the one-step transition probabilities of the embedded Markov chain, we get the following set of equations for the stationary distribution $\Pi_{ij}^{(k)}$:

$$\Pi_{ij}^{(1)} = \sum_{n=1}^{i+1}\sum_{m=0}^{j} \Pi_{nm} k_{i+1-n,j-m}^{(1)} + \sum_{m=0}^{j} \Pi_{0m} k_{i,j-m}^{(1)} \frac{\alpha}{\alpha+\lambda+m\mu}, \tag{6}$$

$$\Pi_{ij}^{(2)} = \sum_{m=0}^{j} \Pi_{0m} k_{i,j-m}^{(2)} \frac{\lambda}{\alpha+\lambda+m\mu} + \sum_{m=1}^{j+1} \Pi_{0m} k_{i,j+1-m}^{(2)} \frac{m\mu}{\alpha+\lambda+m\mu}. \tag{7}$$

For the generating functions $\Pi^{(1)}(y,z)$ and $\Pi^{(2)}(y,z)$ eqs. (6) and (7) become

$$y\Pi^{(1)}(y,z) = K^{(1)}(y,z)(\alpha y R(z) + \Pi(y,z) - \Pi(0,z)), \tag{8}$$

$$\Pi^{(2)}(y,z) = K^{(2)}(y,z)(\lambda R(z) + \mu R'(z)), \tag{9}$$

where

$$\Pi(y,z) = \Pi^{(1)}(y,z) + \Pi^{(2)}(y,z) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} y^i z^j \Pi_{ij},$$

$$R(z) = \sum_{j=0}^{\infty} \frac{\Pi_{0j}}{\alpha+\lambda+j\mu} z^j.$$

Note that

$$(\alpha+\lambda)R(z) + \mu z R'(z) = \Pi(0,z), \tag{10}$$

and thus from (8), (9) and (10) we have

$$\Pi^{(1)}(y,z)(y - K^{(1)}(y,z)) = \mu R'(z) K^{(1)}(y,z)(K^{(2)}(y,z) - z) - R(z)K^{(1)}(y,z)(\alpha(1-y) + \lambda(1 - K^{(2)}(y,z))). \tag{11}$$

The key to solve this equation is the following lemma:

LEMMA

If $\sigma + \rho < 1$ then, for each fixed z with $|z| \leqslant 1$, the function $f(y,z) = y - K^{(1)}(y,z)$ has a unique root $y = h(z)$ in the unit disk $|y| \leqslant 1$. The function $h(z)$, $|z| \leqslant 1$, has the following properties:

1) $h'(1) = \lambda\beta_1^{(1)}/(1-\sigma)$.
2) $h''(1) = \lambda^2\beta_2^{(1)}/(1-\sigma)^3$.
3) $z = K^{(2)}(h(z), z)$ if and only if $z = 1$.

Proof of this lemma is standard and we omit it (see [1]).

Note that $h(z)$ is the generating function of the number of non-priority customers arriving in a classic $M_2/G_2/1$ priority queue during a busy cycle of priority customers. In other words, $h(z) = \varphi(\lambda - \lambda z)$, where $\varphi(s)$ is the Laplace–Stieltjes transform of the busy period in an $M/G/1$ queue with arrival rate α and service

time $B_1(x)$. Thus the function $h(z)$ is analytical in the open disk $|z|<1$ and is continuous in the closed disk $|z|\leqslant 1$.

Replacing $y = h(z)$ in (11) we get

$$\mu(K^{(2)}(h(z),z) - z)R'(z) = (\alpha(1-h(z)) + \lambda(1 - K^{(2)}(h(z),z))R(z). \quad (12)$$

As we have noted in the lemma, the coefficient $K^{(2)}(h(z),z) - z$ never vanishes for $z \neq 1$ and besides

$$\lim_{z\to 1-} \frac{\alpha(1-h(z)) + \lambda(1 - K^{(2)}(h(z),z))}{K^{(2)}(h(z),z) - z} = \frac{\lambda(\sigma+\rho)}{1-\sigma-\rho} < \infty. \quad (13)$$

Thus, the function

$$g(z) = \frac{\alpha(1-h(z)) + \lambda(1 - K^{(2)}(h(z),z))}{K^{(2)}(h(z),z) - z} \quad (14)$$

is analytical in the open disk $|z|<1$ and is continuous in the closed disk $|z|\leqslant 1$. Therefore, for $|z|\leqslant 1$ eq. (12) can be solved as follows:

$$R(z) = R(1)\exp\left\{\frac{1}{\mu}\int_1^z \frac{\alpha(1-h(u)) + \lambda(1 - K^{(2)}(h(u),u))}{K^{(2)}(h(u),u) - u}\,\mathrm{d}u\right\}.$$

Using (2) we get the final formula for $R(z)$. Now (3) and (4) follow from (11) and (9), respectively. □

4. Joint distribution of the channel state and the queue length in the steady state

Let

$$P_{0ij} = P(A(t) = 0, C(t) = i, N(t) = j)$$

be the probability that at time t the channel is free; there are i customers in the priority queue and j customers in orbit.

Obviously, $P_{0ij} = 0$ if $i \geqslant 1$. Let

$$P_{kij}(x)\,\mathrm{d}x = P(A(t) = k, C(t) = i, N(t) = j, x < \xi(t) < x + \mathrm{d}x), \quad k = 1,2,$$

be the probability that at time t the channel is occupied by a customer of type k, $k \in \{1,2\}$; elapsed service time is between x and $x + \mathrm{d}x$, there are i customers in priority queue and j customers in the orbit, and

$$P_{kij} = \int_0^\infty P_{kij}(x)\,\mathrm{d}x = P(A(t) = k, C(t) = i, N(t) = j).$$

Introduce also the corresponding partial generating functions:

$$P_0(z) = \sum_{j=0}^{\infty} z^j P_{00j},$$

$$P_k(y,z;x) = \sum_{i=0}^{\infty} y^i \sum_{j=0}^{\infty} z^j P_{kij}(x), \quad k = 1,2,$$

$$P_k(y,z) = \int_0^{\infty} P_k(y,z;x)\,dx = \sum_{i=0}^{\infty} y^i \sum_{j=0}^{\infty} z^j P_{kij}, \quad k = 1,2.$$

THEOREM 3

In steady state the joint distribution of the channel state, the length of the priority queue and the number of sources of repeated calls has partial generating functions:

$$P_0(z) = (1-\sigma-\rho)\exp\left\{\frac{1}{\mu}\int_1^z \frac{\alpha(1-h(u)) + \lambda(1-K^{(2)}(h(u),u))}{K^{(2)}(h(u),u) - u}\,du\right\}, \tag{15}$$

$$P_1(y,z;x) = P_0(z) \times \frac{(\alpha - \alpha h(z) + \lambda - \lambda Q(z))(T(y,z) - z) + (\alpha - \alpha y + \lambda - \lambda T(y,z))(z - Q(z))}{(Q(z) - z)(y - K^{(1)}(y,z))} \times (1 - B_1(x))e^{-(\alpha(1-y)+\lambda(1-z))x}, \tag{16}$$

$$P_2(y,z;x) = P_0(z)\frac{\alpha - \alpha h(z) + \lambda - \lambda z}{Q(z) - z}(1 - B_2(x))e^{-(\alpha(1-y)+\lambda(1-z))x}, \tag{17}$$

where $T(y,z) = K^{(2)}(y,z)$ and $Q(z) = K^{(2)}(h(z),z)$. If in the cases $A(t) \in \{1,2\}$ we neglect the elapsed service time $\xi(t)$ then for the corresponding generating functions we get

$$P_1(y,z) = P_0(z) \times \frac{(\alpha - \alpha h(z) + \lambda - \lambda Q(z))(T(y,z) - z) + (\alpha - \alpha y + \lambda - \lambda T(y,z))(z - Q(z))}{(Q(z) - z)(y - K^{(1)}(y,z))} \times \frac{1 - K^{(1)}(y,z)}{\alpha(1-y) + \lambda(1-z)},$$

$$P_2(y,z) = P_0(z)\frac{\alpha - \alpha h(z) + \lambda - \lambda z}{Q(z) - z}\frac{1 - T(y,z)}{\alpha(1-y) + \lambda(1-z)}.$$

Proof

In a general way we obtain the equations of statistical equilibrium:

$$\frac{\partial}{\partial x}P_{kij}(x) = -(\alpha+\lambda+b_k(x))P_{kij}(x) + \alpha P_{k,i-1,j}(x)(1-\delta_{i0}) + \lambda P_{k,i,j-1}(x)(1-\delta_{j0}), \quad i\geqslant 0; j\geqslant 0; k=1,2, \tag{18}$$

$$P_{1ij}(0) = \alpha P_{00j}\delta_{i0} + \sum_{k=1}^{2}\int_0^\infty P_{k,i+1,j}(x)b_k(x)\,dx, \quad i\geqslant 0; j\geqslant 0, \tag{19}$$

$$P_{2ij}(0) = \begin{cases} 0 & \text{if } i\geqslant 1, \\ \lambda P_{00j} + (j+1)\mu P_{0,0,j+1} & \text{if } i=0, \end{cases} \tag{20}$$

$$(\alpha+\lambda+j\mu)P_{00j} = \sum_{k=1}^{2}\int_0^\infty P_{k0j}(x)b_k(x)\,dx, \tag{21}$$

where δ is Kronecker's function.

For the generating functions $P_0(z), P_1(y,z;x)$ and $P_2(y,z;x)$, eqs. (18)–(21) become

$$\frac{\partial}{\partial x}P_k(y,z;x) = -(\alpha(1-y)+\lambda(1-z)+b_k(x))P_k(y,z;x), \quad k=1,2, \tag{22}$$

$$yP_1(y,z;0) = \sum_{k=1}^{2}\int_0^\infty (P_k(y,z;x) - P_k(0,z;x))b_k(x)\,dx + \alpha y P_0(z), \tag{23}$$

$$P_2(y,z;0) = \lambda P_0(z) + \mu P_0'(z), \tag{24}$$

$$(\alpha+\lambda)P_0(z) + \mu z P_0'(z) = \sum_{k=1}^{2}\int_0^\infty P_k(0,z;x)b_k(x)\,dx. \tag{25}$$

From (22) we find that $P_1(y,z;x)$ and $P_2(y,z;x)$ depend upon x as follows:

$$P_k(y,z;x) = P_k(y,z;0)(1-B_k(x))e^{-(\alpha(1-y)+\lambda(1-z))x}. \tag{26}$$

With the help of (26), from eqs. (23) and (25) we have

$$yP_1(y,z;0) = \sum_{k=1}^{2}P_k(y,z;0)K^{(k)}(y,z) - (\alpha(1-y)+\lambda)P_0(z) - \mu z P_0'(z). \tag{27}$$

Eliminating from (24) and (27) the function $P_2(y,z;0)$ we get

$$\mu P_0'(z)(K^{(2)}(y,z)-z) = P_0(z)(\alpha(1-y)+\lambda(1-K^{(2)}(y,z))) + P_1(y,z;0)(y-K^{(1)}(y,z)). \tag{28}$$

Equation (28) has the same structure as eq. (11) which appeared in the analysis of the embedded Markov chain. Then it can be solved by the same approach.

Namely, if we put $y = h(z)$ in (28) we obtain the following differential equation for $P_0(z)$:

$$\mu(K^{(2)}(h(z), z) - z)P_0'(z) = (\alpha(1 - h(z)) + \lambda(1 - K^{(2)}(h(z), z))P_0(z).$$

The above expression is identical to eq. (12) for the generating function $R(z)$. Thus

$$P_0(z)/R(z) = \text{Constant}.$$

Obviously $P_0(1) = 1 - \sigma - \rho$. Indeed, $1 - P_0(1)$ is the probability that the channel is occupied, i.e. carried traffic. Since customers cannot be lost, in the steady state carried traffic is equal to the offered traffic $\sigma + \rho$. Thus, Const. $= \alpha + \lambda$ and (15) follows from (5).

Now from (28) we can find $P_1(y, z; 0)$ and thus from (26) $P_1(x, y; z)$. This implies formula (16). Similarly, from (24) we find $P_2(y, z; 0)$ and thus from (26) we get (17). □

With the help of generating functions $P_0(z)$, $P_1(y, z)$ and $P_2(y, z)$ we can get various performance characteristics of the system:

a) probability that the channel is occupied by a priority customer (carried priority traffic):

$$P_1 = \sigma;$$

b) probability that the channel is occupied by a non- priority customer (carried non-priority traffic):

$$P_2 = \rho;$$

c) mean number of customers in the priority queue

$$E[C(t)] = \frac{\alpha(\alpha\beta_2^{(1)} + \lambda\beta_2^{(2)})}{2(1 - \sigma)};$$

d) mean number of customers in the non-priority queue

$$E[N(t)] = \frac{\lambda(\alpha\beta_2^{(1)} + \lambda\beta_2^{(2)})}{2(1 - \sigma)(1 - \sigma - \rho)} + \frac{\lambda(\sigma + \rho)}{\mu(1 - \sigma - \rho)}.$$

The mean waiting time for each group of customers can be obtained from $E[C(t)]$ and $E[N(t)]$ via Little's formula.

It should be noted that $E[C(t)]$ and $E[N(t)]$ can be calculated almost automatically with the help of the property of stochastic decomposition, which we will discuss in the next section.

For the special case $B_1(x) = B_2(x)$, replacing $h(z) = K^{(k)}(h(z), z)$, $k = 1, 2$, we find that our results agree with all the results given in Choi and Park [1].

5. Stochastic decomposition

Introduce the random vector (A_μ, C_μ, N_μ), where $A_\mu \in \{0,1,2\}$; $C_\mu \in \mathbb{Z}_+$; $N_\mu \in \mathbb{Z}_+$, with the help of the generating functions $P_k(y,z) = E[y^{C_\mu}, z^{N_\mu}; A_\mu = k]$, $k \in \{1,2\}$ and $P_0(z) = E[z^{N_\mu}; A_\mu = 0]$ (note that $C_\mu = 0$ if $A_\mu = 0$). This vector represents the number of customers in priority and non-priority queues and the type of customer in service at the stationary regime.

Let $(A_\infty, C_\infty, N_\infty)$ be the corresponding vector for the classic $M_2/G_2/1$ priority queue with the head-of-the-line priority discipline.

Introduce also a random variable R_μ with the help of the generating function

$$R_\mu(z) = E[z^{R_\mu}] = \exp\left\{\frac{1}{\mu}\int_1^z \frac{\alpha(1-h(u)) + \lambda(1-K^{(2)}(h(u),u))}{K^{(2)}(h(u),u) - u}\,du\right\}. \quad (29)$$

In fact, R_μ represents the number of customers in orbit given that the server is free.

Then from theorem 3 and well-known results for the classic $M_2/G_2/1$ priority queue [6], we have the following result about stochastic decomposition of the vector (A_μ, C_μ, N_μ).

THEOREM 4

The vector (A_μ, C_μ, N_μ) can be represented as a sum of two independent random vectors. One of them is the vector $(A_\infty, C_\infty, C_\infty)$ and the second is $(0, 0, R_\mu)$:

$$(A_\mu, C_\mu, N_\mu) = (A_\infty, C_\infty, N_\infty) + (0, 0, R_\mu). \quad (30)$$

This result is extremely useful for analysis of the system under consideration.

For example, eq. (30) implies that

$$E[C_\mu] = E[C_\infty],$$
$$E[N_\mu] = E[N_\infty] + E[R_\mu].$$

The values of $E[C_\infty]$ and $E[N_\infty]$ are well-known from the classic theory of priority queues, and $E[R_\mu]$ can be found without difficulty from (29) and (13), so that

$$E[R_\mu] = \frac{\lambda(\sigma+\rho)}{\mu(1-\sigma-\rho)}.$$

It should be pointed out that similar results about stochastic decomposition were establishes for other retrial queues (see [9] and [10]).

6. Limit theorems for high and low retrial intensities and heavy traffic

Although the performance characteristics of the system are available in explicit form, they are, however, cumbersome: the above formulas include integrals of

transforms, solutions of functional equations, etc. However, in some domains of the system parameters we can approximate the steady state distribution by classic distributions such as Gaussian or Gamma distributions. With this goal, in this section we investigate the asymptotic behaviour of the number of customers in the system under limit values of various parameters.

In real situations, subscribers who get a busy signal almost immediately repeat their calls. Therefore, an investigation of the asymptotic behaviour of system characteristics under high retrial intensity is of special interest in practice.

In general, as $\mu \to \infty$, the stationary distribution of a retrial queue converges to a limit, which is usually the stationary distribution of a certain limit system. Intuitively, in our case the limit system is the classical $M_2/G_2/1$ investigated by Jaiswal [6].

This heuristic argument is rigorously established with the help of the stochastic decomposition given in theorem 4. As expected, the marginal distribution of the number of priority customers in queue is independent of parameter μ. Moreover, the distribution of customers in orbit depends on μ through R_μ; but $\lim_{\mu\to\infty} R(z) = 1$, so the limit system for our retrial priority system is the $M_2/G_2/1$ queue with head-of-the-line priority discipline.

The following result about the rate of convergence of the distribution $P_{kij}(\mu)$ of the vector (A_μ, C_μ, N_μ) to the corresponding distribution $P_{kij}(\infty)$ associated with the limit system with the head-of-the-line priority discipline is essentially more interesting.

THEOREM 5

As $\mu \to \infty$ the distance $D = \sum_{k=0}^{2} \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} |P_{kij}(\mu) - P_{kij}(\infty)| = O(1/\mu)$. Moreover, the following inequalities hold:

$$2(1-\sigma-\rho)(1-R_0(\mu)) < D < 2(1-R_0(\mu)), \tag{31}$$

where

$$R_0(\mu) = \exp\left\{-\frac{1}{\mu}\int_0^1 \frac{\alpha(1-h(u)) + \lambda(K^{(2)}(h(u),u))}{K^{(2)}(h(u),u) - u}\, du\right\}. \tag{32}$$

Proof

The proof is based on the stochastic decomposition property. Namely, theorem 4 implies that $P_{ijk}(\mu)$ is a convolution of the distribution $P_{kij}(\infty)$ and $R_m(\mu) = P(R_\mu = m)$, i.e.

$$P_{kij}(\mu) = \sum_{m=0}^{j} P_{kim}(\infty) R_{j-m}(\mu). \tag{33}$$

Thus,

$$P_{kij}(\mu) - P_{kij}(\infty) = P_{kij}(\infty)R_0(\mu) - P_{kij}(\infty) + (1 - \delta_{j0}) \sum_{m=0}^{j-1} P_{kim}(\infty)R_{j-m}(\mu) .$$

Therefore, using (33) we get

$$\begin{aligned} |P_{kij}(\mu) - P_{kij}(\infty)| &< P_{kij}(\infty)(1 - R_0(\mu)) + (1 - \delta_{j0}) \sum_{m=0}^{j-1} P_{kim}(\infty)R_{j-m}(\mu) \\ &= P_{kij}(\infty)(1 - R_0(\mu)) + P_{kij}(\mu) - P_{kij}(\infty)R_0(\mu) \\ &= P_{kij}(\infty)(1 - 2R_0(\mu)) + P_{kij}(\mu) . \end{aligned}$$

Hence, summing over all the states we obtain

$$D < (1 - 2R_0(\mu)) \sum_{k,i,j} P_{kij}(\infty) + \sum_{k,i,j} P_{kij}(\mu) . \tag{34}$$

But both $P_{kij}(\infty)$ and $P_{kij}(\mu)$ are probability distributions. Thus both sums on the right-hand side of (34) are equal to 1, and the upper inequality in (31) follows.

To get estimation from below we use the obvious inequality $|a - b| \geqslant a - b$, so that

$$\begin{aligned} D &\geqslant \sum_{k,i} |P_{ki0}(\mu) - P_{ki0}(\infty)| + \sum_{k,i} \sum_{j=1}^{\infty} (P_{kij}(\mu) - P_{kij}(\infty)) \\ &= (1 - R_0(\mu)) \sum_{k,i} P_{ki0}(\infty) + 1 - \sum_{k,i} P_{ki0}(\mu) - 1 + \sum_{k,i} P_{ki0}(\infty) \\ &= 2(1 - R_0(\mu)) \sum_{k,i} P_{ki0}(\infty) = 2(1 - R_0(\mu)) \frac{(1 - \sigma - \rho)(\alpha - \alpha h(0) + \lambda)}{\lambda\beta_2(\alpha - \alpha h(0) + \lambda)} \\ &> 2(1 - R_0(\mu))(1 - \sigma - \rho) . \end{aligned}$$

Note that $R_0(\mu)$ can be obtained from eq. (29) by putting $z = 0$. □

In the case of $\mu \to 0$ one can prove that an adequate transformation of $N(t)$ leads to a Gaussian distribution. This statement is established in the following:

THEOREM 6

If the $M_2/G_2/1$ retrial queue with head-of-the-line priority discipline is in the steady state and $\beta_2^{(k)} < \infty$, $k = 1, 2$; then as $\mu \to 0$ the number $N(t)$ of customers in orbit is asymptotically Gaussian with mean $\lambda(\sigma + \rho)/\mu(1 - \sigma - \rho)$ and variance $1/\mu(\lambda^2(\alpha\beta_2^{(1)} + \lambda\beta_2^{(2)})/2(1 - \sigma)(1 - \sigma - \rho)^2 + \lambda(\sigma + \rho)/1 - \sigma - \rho)$.

Proof

Let the variable be

$$N^*(t) = \mu^{1/2}\left(N(t) - \frac{\lambda(\sigma+\rho)}{\mu(1-\sigma-\rho)}\right).$$

The characteristic function $E[e^{itN^*}]$ can be expressed in terms of $P_0(z)$, $P_1(1,z)$ and $P_2(1,z)$ as follows:

$$E[e^{itN^*}] = H(e^{it\mu^{1/2}}) \exp\left\{-\frac{it\lambda(\sigma+\rho)}{(1-\sigma-\rho)\mu^{1/2}}\right\},$$

where $H(z) = P_0(z) + P_1(1,z) + P_2(1,z)$. From now on we denote $w = e^{it\mu^{1/2}}$.

Hence, from theorem 3, we have

$$E[e^{itN^*}] = (1-\sigma-\rho)\frac{\alpha(1-h(w)) + \lambda(1-w)}{\lambda(K^{(2)}(h(w),w) - w)} \times \exp\left\{\frac{1}{\mu}\int_1^w g(u)\,du - \frac{it\lambda(\sigma+\rho)}{(1-\sigma-\rho)\mu^{1/2}}\right\}, \tag{35}$$

where the function $g(u)$ is given by formula (14).

If $\mu \to 0$ then $w \to 1$, so from (13) we get

$$\lim_{\mu\to 0}(1-\sigma-\rho)\frac{\alpha(1-h(w)) + \lambda(1-w)}{\lambda(K^{(2)}(h(w),w) - w)} = 1. \tag{36}$$

We turn now to the calculation of the exponent on the right-hand side of (35). Let us transform the argument of the exponential function as follows:

$$\frac{1}{\mu}\int_1^w g(u)\,du - \frac{it\lambda(\sigma+\rho)}{(1-\sigma-\rho)\mu^{1/2}} = \frac{1}{\mu}\int_1^w \left(g(u) - \frac{\lambda(\sigma+\rho)}{1-\sigma-\rho}\right) du + \frac{\lambda(\sigma+\rho)(w-1-it\mu^{1/2})}{\mu(1-\sigma-\rho)}. \tag{37}$$

The second term on the right-hand side of (37) has the limit equal to

$$-\frac{\lambda(\sigma+\rho)t^2}{2(1-\sigma-\rho)}. \tag{38}$$

To calculate the limit of the first term in (38), it is convenient to introduce the function

$$f(\mu) = \int_1^w \left(g(u) - \frac{\lambda(\sigma+\rho)}{1-\sigma-\rho}\right) du. \tag{39}$$

It is easy to see that

$$f(0) = 0, \quad f'(0) = -\frac{t^2\lambda^2(\alpha\beta_2^{(1)} + \lambda\beta_2^{(2)})}{4(1-\sigma)(1-\sigma-\rho)^2}.$$

Thus, as $\mu \to 0$ we have

$$f(\mu) = -\mu \frac{t^2\lambda^2(\alpha\beta_2^{(1)} + \lambda\beta_2^{(2)})}{4(1-\sigma)(1-\sigma-\rho)^2} + o(\mu), \tag{40}$$

and therefore, from (36), (38) and (40) it follows that

$$\lim_{\mu\to 0} E[e^{itN^*}] = \exp\left\{-\frac{t^2}{2}\left(\frac{\lambda^2(\alpha\beta_2^{(1)} + \lambda\beta_2^{(2)})}{2(1-\sigma)(1-\sigma-\rho)^2} + \frac{\lambda(\sigma+\rho)}{1-\sigma-\rho}\right)\right\}$$

and this completes the proof of the theorem. □

It should be noted that if $\alpha = 0$ then theorem 2 agrees with a known result for the $M/G/1$ retrial queue (see [4]).

The case of heavy traffic is more complicated. We assume that heavy traffic means that $\lambda \to (1-\sigma)/\beta_1^{(2)} - o$. Note in $P_1(y,1) + P_2(y,1)$ that the priority queue size converges to a proper distribution, so only the orbit limit behaviour is mathematically interesting. It can be checked that the next theorem agrees with the corresponding result obtained by Falin [2] for the $M/G/1$ retrial queue.

THEOREM 7

If the $M_2/G_2/1$ retrial queue with head-of-the-line priority discipline is in the steady state and $\beta_2^{(k)} < \infty$, $k = 1, 2$; we have

$$\lim_{\lambda\to(1-\sigma)/\beta_1^{(2)}-o} E[e^{-s(1-\sigma-\rho)N(t)}] = (1+as)^{-b},$$

where

$$a = \frac{(1-\sigma)\beta_2^{(2)} + \alpha\beta_2^{(1)}\beta_1^{(2)}}{2\beta_1^{(2)^2}}, \quad b = 1 + \frac{2(1-\sigma)\beta_1^{(2)}}{\mu((1-\sigma)\beta_2^{(2)} + \alpha\beta_2^{(1)}\beta_1^{(2)})},$$

that is, the scaled random variable $N^*(t) = (1-\sigma-\rho)N(t)$ converges weakly to a gamma distribution as $\lambda \to (1-\sigma)/\beta_1^{(2)} - o$.

Proof

The Laplace–Stieltjes transform $E[e^{-sN^*(t)}]$ of the random variable $N^*(t) = (1-\sigma-\rho)N(t)$ can be obtained from $H(z) = P_0(z) + P_1(1,z) + P_2(1,z)$ by putting $z = e^{-s(1-\sigma-\rho)}$, so that

$$E[e^{-sN^*(t)}] = (1-\sigma-\rho)\frac{\alpha(1-h(e^{-s(1-\sigma-\rho)})) + \lambda(1-e^{-s(1-\sigma-\rho)})}{\lambda(K^{(2)}(h(e^{-s(1-\sigma-\rho)}), e^{-s(1-\sigma-\rho)}) - e^{-s(1-\sigma-\rho)})}$$
$$\times \exp\left\{\frac{1}{\mu}\int_1^{e^{-s(1-\sigma-\rho)}} g(u)\,du\right\}, \tag{41}$$

where $g(u)$ is given by (14).

If after some algebra we expand $K^{(2)}(h(e^{-s(1-\sigma-\rho)}), e^{-s(1-\sigma-\rho)})$ into a power series in terms of $1 - e^{-s(1-\sigma-\rho)}$, we have

$$K^{(2)}(h(e^{-s(1-\sigma-\rho)}), e^{-s(1-\sigma-\rho)}) = 1 - \frac{\rho}{1-\sigma}(1 - e^{-s(1-\sigma-\rho)}) + \left(\frac{\lambda^2\beta_2^{(2)}}{2(1-\sigma)^2} + \frac{\alpha\lambda\rho\beta_2^{(1)}}{2(1-\sigma)^3}\right)(1 - e^{-s(1-\sigma-\rho)})^2 + o((1 - e^{-s(1-\sigma-\rho)})^2). \quad (42)$$

With the help of (42), it is possible to calculate the limit of the first factor on the right-hand of (41). It is equal to $(1 + as)^{-1}$.

To find the limit of the exponential term in eq. (41), we must investigate

$$\lim_{\lambda \to (1-\sigma)/\beta_1^{(2)}} \int_1^{e^{-s(1-\sigma-\rho)}} g(u)\, du. \quad (43)$$

By making the change of variable $v = (1 - u)/(1 - e^{-s(1-\sigma-\rho)})$, we put the integral (43) in the form

$$\int_0^1 \frac{\alpha(1 - h(1 - vt)) + \lambda(1 - K^{(2)}(h(1 - vt), 1 - vt))}{1 - vt - K^{(2)}(h(1 - vt), 1 - vt)} t\, dv, \quad (44)$$

where $t = 1 - e^{-s(1-\sigma-\rho)}$.

Under the assumptions of the theorem it can be seen that

$$h(1 - r\epsilon) = 1 - \frac{\lambda\beta_1^{(1)}}{1-\sigma} r\epsilon + \frac{\lambda^2\beta_2^{(1)}}{2(1-\sigma)^3} r^2\epsilon^2 + \epsilon^2 o(1),$$

$$K^{(2)}(h(1 - r\epsilon), 1 - r\epsilon) = 1 - \frac{\rho}{1-\sigma} r\epsilon + \left(\frac{\lambda^2\beta_2^{(2)}}{2(1-\sigma)^2} + \frac{\alpha\lambda\rho\beta_2^{(1)}}{2(1-\sigma)^3}\right) r^2\epsilon^2 + \epsilon^2 o(1),$$

uniformly with respect to $r \in [0, 1]$ as $\epsilon \to 0$.

In particular, for $r = v$ and $\epsilon = 1 - e^{-s(1-\sigma-\rho)}$, we obtain that the integrand in (44) converges uniformly with respect to v, as $\lambda \to (1 - \sigma)/\beta_1^{(2)}$, to the function

$$-\frac{s(1-\sigma)}{\beta_1^{(2)}} \left(1 + \frac{(1-\sigma)\beta_2^{(2)} + \alpha\beta_2^{(1)}\beta_1^{(2)}}{2\beta_1^{(2)^2}} vs\right)^{-1}. \quad (45)$$

Although the basic reasoning used to get (45) is parallel to the case $\alpha = 0$, it is, in comparison, rather more cumbersome. Therefore, some intermediate steps to get (45) have been omitted.

Hence, by solving integral (44) we get that the limit (43) is equal to $(1 + as)^{1-b}$. This completes the proof. □

Acknowledgements

We would like to express our gratitude to the referees for suggestions that improved the quality of the paper.

References

[1] B.D. Choi and K.K. Park, The $M/G/1$ retrial queue with Bernoulli schedule, Queueing Systems 7 (1990) 219–228.
[2] G.I. Falin, A single-line system with secondary orders, Eng. Cybernet. Rev. 17 (1979) 76–83.
[3] G.I. Falin, On a multiclass batch arrival retrial queue, Adv. Appl. Prob. 20 (1988) 483–487.
[4] G.I. Falin, A survey of retrial queues, Queueing Systems 7 (1990) 127–168.
[5] T. Hanschke, The $M/G/1/1$ queue with repeated attempts and different types of feedback effects, OR Spectrum 7 (1985) 209–215.
[6] N.K. Jaiswal, Time-dependent solution of the head-of-the-line priority queue, J. Roy. Stat. Soc. B24 (1962) 91–101.
[7] Z. Khalil, G. Falin and T. Yang, Some analytical results for congestion in subscriber line modules, Queueing Systems 10 (1992) 381–402.
[8] V.G. Kulkarni, On queueing systems with retrials, J. Appl. Prob. 20 (1983) 380–389.
[9] T. Yang and J.G.C. Templeton, A survey on retrial queues, Queueing Systems 2 (1987) 203–233.
[10] T. Yang, M.J.M. Posner, J.G.C. Templeton and H. Li, An approximation method for the $M/G/1$ retrial queue with general service times, to appear in Eur. J. Oper. Res.