ABSTRACT:
We carry out message delay and system throughput analysis for communication and telecommunications systems with preemptive buffer (PB) policies. Such systems are of particular importance in situations where it is essential to provide short (and at times, bounded) message waiting-time delays, for messages which are selected for service. This is the case for many telephone and telecommunications network systems. This is also the case for many time-critical space, sensor, telemetry and radar communication and processing systems, in which the information contents of a message is associated with a timeliness index, so that the most recent message to arrive contains the most valuable information and thus needs to be given preference for service (within a prescribed acceptable time window).

We analyze and compare the performance of such systems, modeling them as finite capacity M/G/1/N queueing systems, and employing the service/buffering policies FCFS/NPB, FCFS/PB, and LCFS/PB. We evaluate the performance tradeoffs offered by these policies, as they related to the mean and variance of the waiting-times of messages which are eventually served or eventually rejected/preempted.

1. Introduction
We consider a communication or processing/queueing system served by a single channel. Messages are admitted into the system in accordance with a window-based admission-control policy; or, equivalently, the system contains a finite capacity buffer facility, so that a limited number of messages can be held in the system at any time. The selection of messages for admission into the system and the service channel is dictated by the queueing discipline. In turn, the latter is characterized by the underlying service and buffering policies.

The service policy determines the selection of queued messages for service, when the service channel becomes available. Examples include: First-Come First-Served (FCFS), Last-Come Last-Served (LCFS). The buffering policy specifies the messages which are to be admitted and preempted from the buffer, when a new message arrives at a saturated system (i.e., a system in which the buffer facility is full). The most common buffering policy keeps the message already in the system and blocks the newly arriving ones. This policy is called Non-Preemptive-Buffering (NPB). Another alternative buffering policy allocates buffer space to the most recent arrival and preempts the queued message which has waited the longest. This policy is referred to as Preemptive-Buffering (PB).

Many communication systems and telecommunications network applications require the use of finite preemptive-buffering service systems. Such systems are of particular importance in situations where it is essential to provide short message waiting-time delays to those messages which are selected for service. This is the case for many telephone systems (see, for example, [4]), in which messages which are not served or preempted promptly can induce significant (throughput and delay) performance degradations.

This is also the case for many time-critical space, sensor, telemetry and radar communication and processing systems [2], in which the information contents of a message is associated with a timeliness index, so that the most recent message to arrive contains the most valuable information, and thus needs to be given preference for selection for service (within a prescribed acceptable time window).

Such processing systems are studied in this paper. We analyze and compare the performance of the different queueing disciplines for an M/G/1/N queueing system model with a finite buffer. Using a notation XY to designate X and Y as the service and buffering policies, respectively, we consider the following systems: FCFS/NPB, FCFS/PB, and LCFS/PB. The moment generating function, as well as, the mean and the variance, of the waiting-time for served messages, are presented as solutions to a linear-recursive system of equations for the three schemes mentioned above.

Assume the system to be at steady state. Note that if W denotes the (steady-state) waiting time in queue, of an arbitrary tagged message, which may be eventually either served or rejected (preempted, or lost), then we have, under a PB scheme,
\[ E[W_{PB}] = E[W_{PB}|lost] P(lost) + E[W_{PB}|served] P(served), \]
where \( P(\text{ served } ) = 1 - P(\text{ lost } ) \),

and \( P(\text{ served } ) \) and \( P(\text{ lost } ) \) are the steady-state probabilities that a message is eventually served or lost, respectively. Under a NPB policy, we have:
A work conserving discipline (W.C.D.) is defined as a queueing discipline under which the following conditions hold: 1) no service channel is left idle when the buffer is occupied by waiting messages; 2) selection for service is independent of the duration of the requested service-time; 3) no new work is created or destroyed by the employed service procedure. We note that FCFS/PB and LCFS/PB are all W.C.D.'s. Let X denote the steady-state queue size variable. We note that the distribution of X does not depend on the specific work conserving discipline employed, if no preemption from service is permitted. Therefore, P( lost message | arrived in accordance with a Poisson process with intensity \( \lambda \)) is the same for all three disciplines mentioned above. Using Little's formula (see [10]), for a NPB policy, we write

\[
\lambda \mathbb{P}( \text{service} ) = \mathbb{E}[W_{NPB}] = \mathbb{E}[X].
\]

In turn, under a PB scheme, noting that each admitted message and all arriving messages are admitted into the buffer, we write

\[
\lambda \mathbb{P}( \text{service} ) = \mathbb{E}[W_{PB}] = \mathbb{E}[X].
\]

Comparing these equations, we readily conclude that

\[
\mathbb{E}[W_{NPB}] = \mathbb{E}[W_{PB}] = \mathbb{E}[W_{NFB}].
\]

\[
\mathbb{E}[W_{NPB}] = \mathbb{E}[W_{PB}] = \mathbb{E}[W_{NFB}] = \mathbb{E}[X].
\]

so that under both a PB scheme and the average overall message wait-time and the average served-message wait-time values are lower than the corresponding ones attained when a NPB-policy is employed, same for the three disciplines. One can also readily show that, among all W.C.D.'s, E[\( W_{NPB} \)] is minimized when a LCFS/PB scheme is used.

For a FCFS/NPB system, Lavenberg [8] used a model consisting of a two-stage cyclic queueing network with a finite number of messages to derive the moment generating function of the waiting-time as a rational function of the steady-state queue-size probabilities imbedded at departures. As for the waiting-time analysis of a queueing system not employing a NPB policy, Smimova and Feofanov [9] solved the problem for the case of an M/M/c system employing LCFS/PB. Doshi and Heffes [4] analyzed an M/M/1 system operating under FCFS/PB and LCFS/PB. Clare and Rubin [2] analyzed a discrete queueing system with geometric batch arrivals and deterministic service-times, employing FCFS/PB and LCFS/PB. To the best of our knowledge, no work has been presented for the general service-time queueing system employing a PB scheme.

In section 2, we present the queue size analysis. Section 3 includes the waiting time analysis of the three schemes. In section 4, we investigate the limiting case, when the capacity of the system is infinite. Numerical results are presented in section 5, and finally, conclusions are drawn in section 6.

2. Queue Size Analysis

We consider an M/G/1 queueing system. Customers arrive in accordance with a Poisson process with intensity \( \lambda > 0 \). Let \( S_n \) be the time spent in service by the \( n^{th} \) message which is admitted into service. Assume \{ \( A_n, n \geq 1 \) \} to be a sequence of independent identically distributed (i.i.d) random variables. \( B(t) \) is the distribution of the service time, \( B'(t) \) is its Laplace-Stieljes transform (L.S.T), and \( \beta = \mu^{-1} \) is the mean service-time. Let \( A_n \) be the random variable representing the number of messages arriving during the \( n^{th} \) service time. Assume \( \{ A_n, n \geq 1 \} \) to be a sequence of i.i.d random variables. We set:

\[
a_j = P(A_n = j) = \frac{e^{-\lambda j} \lambda^j}{j!}, \quad j=0,1,2,...
\]

\[
a_j = P(A_n \geq j) = \sum_{i=j}^{\infty} a_i, \quad j=0,1,2,...
\]

The waiting room (buffer) capacity (not including the message in service) is equal to \( K \) (messages), \( K < \infty \). The system traffic intensity parameter is defined as \( \rho = \frac{\lambda}{\mu} \).

Let \( T_n, n \geq 1 \) be the instant of the \( n^{th} \) departure from service, \( T_0 \geq 0 \). where \( X_t \) denotes the number of messages in the system at time \( t \). We let \( Z = \{ Z_n, n \geq 1 \} \) represent the system size process embedded at departures, \( Z_n = X_{T_n} \).

We note that \( Z \) is a Markov chain with stationary transition probabilities and finite state space \( S = \{ 0,1,2,...,K \} \). In addition, one readily checks that since arrivals follow a Poisson process, the chain is irreducible and aperiodic. Hence the chain is positive recurrent and the following limits exist:

\[
r_j \triangleq \lim_{n \to \infty} P(Z_n = j) = 0, j \leq K
\]

The set of steady state equations for \( r_j, j=0,1,...,K \) is given as (see [5]):

\[
r_j = r_0 a_j + \sum_{i=1}^{j} r_i a_{j-i}, \quad j=0,1,...,K-1
\]

\[
r_K = r_0 a_K + \sum_{i=1}^{K} r_i a_{K-i}
\]

(2.1)

\[
K \sum_{j=0}^{K} r_j = 1
\]

We consider a tagged message, with the system reaching steady-state. Let \( Y \) be the random variable representing the number of messages in the system prior to the arrival of the tagged message, at steady-state. Let \( \pi_j, j=0,1,...,K+1 \), be the steady state probabilities that an arriving message finds \( j \) messages in the system. Thus we have:

\[
\pi_j = P(Y = j) \quad j=0,1,...,K+1
\]

Then (see [5]) we have,

\[
\pi_j = r_j / (r_0 + \rho), \quad j=0,1,...,K
\]

(2.2)

\[
\pi_{K+1} = 1 - 1/(r_0 + \rho)
\]

Let

\[
p(j) = \lim_{t \to \infty} P(X_t = j) = j=0,...,K+1
\]

Since the arrival process is Poisson, we have (see [3]):

\[
p(j) = \pi_j, \quad j=0,...,K+1.
\]

Since state 0 is recurrent, every message must eventually (in finite time) depart from the system, w.p.1. Hence,
no message stays in systems for an indefinitely long period of time. Therefore, we define the steady-state binary random variable \( L \), designating the loss status of a message, as follows.

\[
L = 1 \quad \text{if the tagged message is lost before it is served,}
\]

and

\[
L = 0 \quad \text{if the message gets eventually served.}
\]

We note that any W.C.D. queueing discipline yields the same distribution of \( L \), \( Y \), \( X \), and \( Z \). For FCFS/NPB service/buffering discipline, we have:

\[
P(L=1) = \pi_{K+1}
\]

so that by (2.2),

\[
P(L=0) = \frac{1}{(\tau_0 + p)}
\]

Hence to compute the distribution of \( L \), it is not necessary to compute the full distribution \( \{ \gamma_j , j=0,1,..,K \} \), but only \( \gamma_0 \) is required. For computing \( \gamma_0 \), one may use a recursive algorithm introduced by Yan (see [11]).

3. Waiting Time Analysis

In this section, we obtain expressions for the moment generating function (m.g.f) and mean of the limiting distribution of the waiting-time of a served message, in systems operating under any one of the three queueing disciplines described in section 1. We let \( W \) denote the steady-state waiting-time random variable, in the queue of a tagged message in the queue (i.e.; the sojourn time in the buffer).

We define the steady-state waiting-time variable \( W \) as follows:

\[
\bar{W} = 0 \quad \text{if the tagged message initiates a busy period (i.e.; finds the system idle at arrival). Otherwise, let } \bar{W} \text{ denote the waiting time of the tagged message measured from the moment the server becomes free for the first time after the tagged message's arrival, and available to select a message for service, (i.e.; a departure due to a service completion occurs).}
\]

\( \text{the instant} \) the tagged message leaves the buffer (either due to its preemption from the buffer or due to its selection for service). This waiting-time variable \( \bar{W} \) corresponds to the usually defined waiting-time duration, if we append it to the residual service-time of the message in service, if any, upon arrival of the tagged message.

We set \( R \) to be the random variable associated with the time to completion of the ongoing service when the tagged message finds the system busy upon arrival. Then, if the message initiates a busy period, we have:

\[
W = \bar{W} = 0
\]

Else,

\[
W = \bar{W} + R
\]

Let \( W(t), \bar{W}(t), \text{ and } R(t) \) be the probability distribution functions of \( W, \bar{W} \text{ and } R \) respectively. Examining the "position" of the tagged message at the epoch of the \( n^{th} \) departure from service following its arrival, we define the sequences of random variables \( \{v_n, n=1,2,..\} \) and \( \{v_n, n=1,2,..\} \) as follows.

If the tagged message initiates a busy period, then \( v_1 = 0 \text{ and } v_1 = 0 \). If the message is blocked, or admitted but preempted prior to the end of the remaining service period \( R \), then,

\[
v_1 = -1 \text{ and } v_1 = -1. \text{ If at the epoch of the } n^{th} \text{ departure from service following his arrival, the tagged message is no longer in the buffer, then we set } v_n = -1 \text{ and } v_n = -1. \text{ Finally, if the message is still in the buffer at the epoch of the } n^{th} \text{ departure from service, then, } v_n \text{ is set to be given equal to the number of messages positioned ahead of itself in the buffer, and } v_n \text{ is defined as the number of messages behind itself in the buffer.}
\]

For all work-conserving disciplines, the following limits exist and are attained almost surely (a.s) (see [8]):

\[
\lim_{n \to \infty} v_n = -1 \quad \text{(the absorbing state)} \quad (3.1)
\]

\[
\lim_{n \to \infty} v_n = -1 \quad \text{(the absorbing state)} \quad (3.2)
\]

3.1. FCFS/NPB Discipline

Consider a FCFS/NPB service-buffering discipline. Customers admitted to the buffer will not leave before being served. Newly arriving messages are blocked when the buffer is saturated. Admitted messages are served in accordance to a FCFS policy. This discipline has been analyzed by several authors, (see for example [8]). We readily have:

\[
\bar{W}(t) = \sum_{n=1}^{K} \pi_n B^{(n-1)}(r) + \pi_0 \quad (3.3)
\]

and

\[
W(t) = \sum_{n=1}^{K} \pi_n [R(t) + B^{(n-1)}(r)] + \pi_0 \quad (3.4)
\]

where \( \{\pi_n, n=0,1,..,K\} \) are given by (2.2); \( B^{(n)}(t) \) is the \( n \)-fold convolution of \( B(t); R(t) \) is the convolution of \( R(t) \) and \( B(t) \), and

\[
R(t) = \mu \int_{0}^{t} (1 - B(x)) \, dx. \quad (3.5)
\]

Note that under this policy, the waiting time of a lost message is always equal to zero. Therefore,

\[
E[W] = E[W \mid L = 0]P(L=0). \quad (3.6)
\]

3.2. LCFS/PB Discipline

Consider a LCFS/PB discipline. This represents a stack buffering system implementation. Newly arriving messages are always admitted and positioned at the top of the stack buffer. In case of overflow, (i.e.; when the buffer becomes oversaturated) messages are removed (preempted) from the bottom of the buffer. An available server, selects for service the message positioned at the top of the stack (see figure 1). Note that under this discipline, \( v_1 \) is equal to the number of arrivals during the remaining service period \( R \), for a tagged message which is still in line at the termination time of this service.

To derive the limiting message waiting-time distribution, the following lemma is first presented. For proof of the lemma, we refer the reader to [8].

**Lemma 1**: Let \( \psi \) be a bounded, or monotone measurable function; for \( 0 < r < K \) and \( 0 \leq j < K \), we have:
We have,
\[
E[\psi(W)(L=0)|\nu_1=r] = \sum_{j=0}^{K-r} \int E[\psi(W+r)(L=0)|\nu_1=r+j] (\lambda j)^{r-j} \frac{1}{j!} e^{-\lambda j} dB(t)
\]  
(3.7)

We have, 
\[
E[\psi(W)(L=0)|Y=0] = E[\psi(W)(L=0)] 
\]  
(3.8)

and 
\[
E[\psi(W)(L=0)|Y>0] = \sum_{k=0}^{K-1} E[\psi(W+r)(L=0)|\nu_1=k] (\lambda k)^{r-k} \frac{1}{k!} e^{-\lambda k} dB(t)
\]  
(3.9)

For \( Y = 0 \), 
\[
E[\psi(W)(L=0)|Y=0] = \psi(0)
\]  
(3.10)

We also have, using Lemma 1, 
\[
E[\psi(W)(L=0)|Y>0] = \int_0^{\infty} E[\psi(W)(L=0)|Y>0, R=r] dR(t)
\]

\[
E[\psi(W)(L=0)|Y>0] = \sum_{k=0}^{K-1} \int E[\psi(W+r)(L=0)|\nu_1=k] (\lambda k)^{r-k} \frac{1}{k!} e^{-\lambda k} dB(t)
\]  
(3.11)

We have thus derived the following theorem:

**Theorem 3.1:** If \( \psi \) is a bounded, or monotone measurable function, we have:
\[
E[\psi(W)(L=0)|\nu_1=0] + \rho \sum_{k=0}^{K-1} \int E[\psi(W+r)(L=0)|\nu_1=k] (\lambda k)^{r-k} \frac{1}{k!} e^{-\lambda k} dB(t)
\]  
(3.12)

Therefore \( E[\psi(W)|L=0] \) can be computed by solving the system of integro-difference equations (3.7) for \( E[\psi(W)(L=0)|\nu_1=r] \) and using eq (3.12). However, when \( \psi(.) \) is a m.g.f or defines any other moment of the waiting time \( W \), we will show that eqs (3.8)-(3.11) reduce to a much simpler system of linear-recursive equations.

### 3.2.1. Moment Generating Function

**Corollary 3.1.1:** The m.g.f. of the waiting-time for the served message is given by:
\[
E[e^{-\lambda W}(L=0)] = r_0 + \sum_{k=0}^{K-1} \sigma_k(s) V_k(s)
\]  
(3.13)

with \( V_k(s) \) being a solution of the following recursive linear system:
\[
V_k(s) = \sum_{j=0}^{K-r} V_{k+j}(s) \theta_j(s) \quad 0 < r < K
\]
\[
V_k(s) = 1 \quad 0 < r < K
\]  
(3.14)

\[
\theta_j(s) = \int_0^{\infty} e^{-\lambda j} \frac{1}{j!} e^{-\lambda t} dB(t)
\]  
(3.15)

and 
\[
\sigma_k(s) \triangleq \int_0^{\infty} e^{-\lambda t} (\lambda j)^k \frac{1}{k!} e^{-\lambda t} dB(t)
\]  
(3.16)

**Proof**

To compute the m.g.f. of \( W \), we set \( \psi(W) = e^{-\lambda W} \), \( Re(s) > 0 \). We note that \( \psi \) is bounded so that the results derived above can be used. Therefore, using eq (3.12), we write:
\[
E[e^{-\lambda W}(L=0)] = r_0 + \rho \sum_{k=0}^{K-1} \int E[e^{-\lambda W}(L=0)|\nu_1=k] (\lambda k)^{r-k} \frac{1}{k!} e^{-\lambda k} dB(t)
\]  
(3.17)

\[
= r_0 + \rho \sum_{k=0}^{K-1} \int E[e^{-\lambda W}(L=0)|\nu_1=k] (\lambda k)^{r-k} \frac{1}{k!} e^{-\lambda k} dB(t)
\]

Hence, by defining:
\[
\sigma_k(s) \triangleq \int_0^{\infty} e^{-\lambda t} (\lambda j)^k \frac{1}{k!} e^{-\lambda t} dB(t)
\]

\[
V_k(s) \triangleq E[e^{-\lambda t}(L=0)|\nu_1=k]
\]

we deduce the stated results. Eq.(3.13) provides for a proper definition of the m.g.f. For a detailed proof, the reader is referred to [8].

### 3.2.2. Mean Waiting Time

We can derive the mean waiting time using Corollary 3.1.1 by means of differentiation of the derived m.g.f. Alternatively, the result is also obtained by setting \( \psi(W) = W \), and using Theorem 3.1.

**Corollary 3.1.2:** The mean waiting-time is given by:
\[
E[W|L=0] = \rho \sum_{j=0}^{K-1} (\sigma_j V_j + e_j)
\]  
(3.18)

where 
\[
\sigma_j = \int_0^{\infty} (\lambda j)^k \frac{1}{j!} e^{-\lambda t} dB(t)
\]  
(3.19)

where 
\[
e_j = \rho \frac{H+1}{\lambda} \sigma_{j+1}
\]

with \( V_k(s) \) satisfying the following linear-recursive system of equations:
\[
\begin{cases}
V_r = \sum_{j=0}^{K-r} (a_j V_{r-j+1} + m_{jr}) & 0 < r < K \\
V_0 = 0
\end{cases}
\]  
(3.20)

where 
\[
m_{jr} = p_{r-1} \frac{j+1}{\lambda} a_{j+1} \quad 0 \leq j \leq K-r
\]

and \( p_j \) satisfies the following system:
\[
\begin{cases}
p_K = \rho \sum_{j=0}^{K} a_j p_{n-j} & 0 < n < K \\
p_0 = 1
\end{cases}
\]  
(3.21)
We note that the sets of equations derived above are linear-recursive systems. One, then, readily concludes, as shown in Theorem 3.1, that sets of equations (3.20)-(3.21) have each a unique solution. The corresponding matrices are upper triangular, Toeplitz matrices (so that they have equal entries on all diagonals). Therefore, we can use a filtering method presented by Levinson[1] to solve these systems in a numerically efficient manner. Under the latter method, the solution requires a number of steps which is of the order of $K^2$, rather than the usual $K^3$ needed for regular solution of linear systems, when the system matrix is of dimension $K$.

The second moment and the variance can also be derived using Theorem 3.1.

3.3. FCFS/PB Discipline

Consider a FCFS/PB service/buffering policy. All newly arriving messages are admitted into the system and positioned at the end of the queue. When the buffer is saturated, the message at the head of the queue is preempted (see figure 2). Since the analysis of this scheme is similar to the analysis of the LCFSPB scheme presented in the previous section, we present only an outline of the derivation of the main results.

Lemma 2: Let $\psi$ be a bounded, or monotone measurable function. Then, we have:

i) for $0 < k < K$, $0 \leq n < K-k$, and $0 \leq j < K-k-n$, we have:

$$E[\psi(W)(L=0)|v_1=k, \tau_1=n, A_1=j]=\left(\frac{\lambda \lambda^j}{j!}\right) e^{-\lambda_1} dB(\gamma)$$

(3.22)

ii) for $0 < k < K$, $0 \leq n < K-k$, and $K-k-n < j < K-n-1$, we have:

$$E[\psi(W)(L=0)|v_1=k, \tau_1=n, A_1=j]=\left(\frac{\lambda \lambda^j}{j!}\right) e^{-\lambda_1} dB(\gamma)$$

(3.23)

Using Lemma 2, we may formulate the following Theorem 3.2. The proof to the lemma and the theorem are provided by [8].

Theorem 3.2: If $\psi$ is a bounded, or a monotone, measurable function, we have:

$$E[\psi(W)(L=0)]=r_0\psi(0)$$

(3.24)

Using Theorem 3.2, we derive the results presented in the next 2 sections.

3.3.1. Moment generating function

Corollary 3.2.1: The m.g.f. of the waiting-time distribution is given by:

$$E[e^{-\lambda W}(L=0)]=r_0+\sum_{k=1}^{K} \sum_{n=0}^{k-1} r_k \sigma_k(s) V(k-1,n,s) + R(s)>0$$

$$+ \sum_{k=1}^{K} \sum_{n=0}^{k-1} r_k \sigma_k(s) V(K-n-1,n,s) + \sum_{n=0}^{K-1} (r_0+p-1) \sigma_n(s) V(K-n-1,n,s)$$

(3.25)

where $\sigma_n(s)$ is given by (3.16) and $V(k,n,s)$ is solution to the following system of equations:

$$V(0,n,s)=1, n \leq K-1$$

$$V(k,n,s)=\sum_{m=0}^{k-1} \theta_m(s) V(k-n-m-1,n+m,s) + 0 < k < K, 0 \leq n < k$$

(3.26)

To prove the above corollary, we set $\psi(W)=e^{-\lambda W}$ and use (3.24). For a detailed proof, we refer the reader to ref [8]. Eq.(3.25) provides for a proper definition of the m.g.f. (see[8]).

3.3.2. Mean waiting time

By means of differentiation of the m.g.f., or by setting $\psi(W)=W$, we obtain the following corollary.

Corollary 3.2.2: The mean waiting-time for the served messages in a system employing FCFS/PB is given by:

$$E[W(L=0)]=\sum_{k=1}^{K} \sum_{n=0}^{k-1} r_k \sigma_k(V(k-1,n)+D_{k,n})$$

$$+ \sum_{k=1}^{K} \sum_{n=0}^{k-1} r_k \sigma_k(V(K-n-1)+D_{K-k,n})$$

$$+ (r_0+p-1) \sum_{n=0}^{K-1} \sigma_n(V(K-n-1,n)+D_{K-n,n})$$

(3.27)

where $\sigma_n$ is as defined previously and,

$$D_{k,n}=p_{k-1,n} + 1 \sigma_{k+1}$$

$$p_{k,n} \text{ satisfies:}$$

$$p_{k,0}=1, 0 \leq n \leq K-1$$

$$p_{k,n}=\sum_{m=0}^{K-k-n} a_m p_{k-1,n+m} + 0 < k < K, 0 \leq n < K-k$$

(3.28)

5C1.5.
\[ V(k,n) \text{ is a solution to the following recursive linear system of equations:} \]
\[ \begin{align*}
V(0,n) &= 0 \\
V(k,n) &= \sum_{m=0}^{K-k} (a_k V(k-1,n+m)+C_{k,m,n}) \\
&+ \sum_{m=n-K+1}^{K-k} (a_k V(k-n-m-1,0)+C_{k,m,n}) \\
&= 0 \text{, } k < K, 0 \leq n < K-k
\end{align*} \]
\[ (3.29) \]

where
\[ C_{k,m,n} = \frac{(m+1)!}{\lambda^m} \]

The proof is similar to the proof of Corollary 3.1.2, using the following definitions:
\[ V(k,n) = E[I(k_0)] \]
\[ p_k = P(L = 0 | \nu_1 = k_1, t_1 = n) \]

We note that the recursive systems of equations (3.28)-(3.29) can be solved in the following order. Solve for \((v,z)=(1,K-1)\) then for \((1,K-2) \ldots (1,0)\) and then for \((2,K-1) \ldots (2,0) \ldots (K-1,0)\).

The second moment and the variance can be derived and \((2,K-1) \ldots (2,0) \ldots (K-1,0)\).

In a similar fashion using Theorem 3.2.

4. Limiting Case: Infinite Buffer

We set \(K = \infty\). Consequently, under any of the three schemes, all messages will eventually be served, so that \(P(L=0)=1\). We then obtain
\[ E[e^{-\nu_1 \nu_1}] = E[e^{-\nu_1 \nu_1}] \]

In addition,
\[ r_j = \pi_j \]
\[ \pi_0 = 1-p \]
We will assume \(p < 1\).

4.1. FCFS/NPB Scheme for Infinite K

All newly arriving messages are admitted and eventually served. The resulting system is an M/G/1 system. The m.g.f of the waiting time is given by the Pollaczek-Khintchine equation (see [3]). Also, by taking limits as \(K\) goes to \(\infty\) in (3.7), we obtain
\[ W(s) = \pi_0 + \sum_{j=0}^{\infty} \pi_j \left( R(s)^j g^{(n-1)}(s) \right). \]
\[ (4.1) \]

4.2. LCFS/PB Scheme for Infinite K

No message preemptions occur, and all messages will eventually be served. The resulting system is an M/G/1 system under a LCFS service discipline. By letting \(K\) go to \(\infty\), system (3.14) becomes
\[ \begin{align*}
V_0(s) &= 0 \\
V_k(s) &= \sum_{n=0}^{\infty} V_{k-1}(s) \theta_n(s) \\
V_0(s) &= 1
\end{align*} \]
\[ (4.2) \]

Define the operator \(D\) as
\[ DV = V_{n+1}, \]
we have:
\[ V_k(s) \left( D - \sum_{j=0}^{\infty} D \theta_n(s) \right) = 0 \]
\[ (4.3) \]

Define
\[ \beta(s) = \sum_{j=0}^{\infty} \theta_n(s)^j \]
\[ (4.4) \]

and let \(x_k, \forall k\) be the solutions in \{Re(s)\geq 0, \mid s \mid<1\}, to
\[ z = \beta(s) \]
\[ (4.5) \]

Then we readily observe that
\[ V_k(s) = C \sum_{\forall \beta} \]
\[ (4.6) \]

where \(C\) is a constant.

We note that
\[ \beta(s) = B'(s+\lambda - \lambda_\beta) \]

so that, (4.5) becomes:
\[ z = B'(s+\lambda - \lambda_\beta) \]
\[ (4.6) \]

We define \(G(t)\) to be the busy period distribution and \(G'(s)\) to be its m.g.f. Since \(p < 1\), there is only one admissible root for (4.6), namely, \(G'(s)\), which is the root with the smallest absolute value in \{Re(s)\geq 0, \mid s \mid<1\} (see [7]). Hence
\[ V_k(s) = C G'(s)^j \]
\[ (4.7) \]

Equation (4.5) is a well known result (see[3]).

4.3. FCFS/PB Scheme for Infinite K

None of the messages are preempted. All messages will eventually be served. The resulting model is an M/G/1 system operating under a FCFS service discipline. By letting \(K\) go to \(\infty\), system (3.26) becomes:
\[ \begin{align*}
V_0(n) &= 0 \\
V_k(n) &= \sum_{m=0}^{\infty} V_{k-1}(n+m) \theta_n(s) \\
V_0(n) &= 1
\end{align*} \]
\[ (4.8) \]

It is readily verified, by induction, that
\[ V_k(n) = (B'(s))^k \]
\[ (4.9) \]

\[ 0 \leq k, n \]

is the solution to (4.6), so that (3.18) becomes
\[ E[e^{-\nu_1 \nu_1}] = (1-p) + \sum_{k=1}^{\infty} \pi_k R(s)(B'(s))^k, \text{Re}(s)\geq 0, \]
\[ (4.9) \]

and
$P(W \leq t) = (1-\rho) + \sum_{k=1}^{\infty} \rho^k R(0)B_{t,k}(0), \quad 0 \leq t,$  
\hspace{1cm} (4.10)

which coincides with the well-known formula for the FCFS M/G/1 system given by (4.1), noting that $\pi_0 = 1-\rho$.

5. Numerical Results

In the application of the above results, we analyze a single-server queueing system with capacity 3, (i.e.; the buffer capacity is 2). For simplicity, we choose the service-time distribution to be exponential. The service rate is $\mu = 10$. We vary the arrival rate $\lambda$ so that the throughput varies.

Figure 3 and Figure 4 present the mean waiting-time under the three policies, FCFS/NPB, FCFS/PB and LCFS/PB, versus the offered load and the carried load, respectively. We consider the waiting-time of served messages only. We observe, in both graphs, that the FCFS/NPB scheme yields the highest mean wait and the LCFS/PB policy attains the smallest mean wait, while under the FCFS/PB policy, an intermediate average message wait is obtained. We also observe that as the traffic increases, the mean wait under a PB policy, tends to zero, while it tends to a level of $0.2$, the maximum wait, in the case of an NPB policy. It is due to the fact that, when the traffic is high, a message arriving to a system operating under PB, is either served soon after the arrival or otherwise preempted. While in the NPB case, the admitted message will have to wait for the preexisting messages to be served first.

In Figure 5, we continue with the example. The graph shows the variance of the wait-time for a served message versus the channel throughput. We observe that the variance for the FCFS/PB scheme, is the smallest, and is very close to the variance attained by the FCFS/NPB policy for low traffic. When the throughput is less than or equal to a value very close to 0.4, under which level, the variance decreases while the throughput becomes larger than 0.4, the FCFS/NPB scheme yields the highest variance value. This is due to the fact that, when the traffic is high, a message arriving to a system operating under PB, is either served soon after the arrival or otherwise preempted. While in the NPB case, the admitted message will have to wait for the preexisting messages to be served first.

Figure 6 and Figure 7 exhibit performance graphs for systems operating under FCFS/PB and LCFS/PB, respectively. The cost criterion (C) is defined as:

$C(\alpha) = E[W|L=0] + \alpha E[W|L=1]$  
\hspace{1cm} where $\alpha$ is an arbitrary real number. Note that for a PB discipline, $E[W|L=1] = 0$, so that we have

$C(\alpha)_{NPB} = E[W|L=0]$  

independently of $\alpha$. Thus, $C$ expresses an average wait cost, when the mean wait of a served message as well as the mean wait of a rejected message are taken into consideration; the latter two wait components are allocated a cost ratio equal to $\alpha$. We, then fix $\alpha = 2$ and vary $\lambda$ and $\mu$ accordingly. We plot the cost function versus the throughput level for various values of $\alpha$. Similar performance behavior is exhibited by FCFS/PB and LCFS/PB. Recalling that for PB disciplines, the cost function

$C(\alpha) = E[W|L=0] + \alpha E[P(L=1)|P(L=0) = 8/7$ (since $\rho = 2$), it is worthwhile to employ a PB scheme. Otherwise, an NPB procedure yields a lower cost value.

In Figure 8, using the same cost functions, we compare the cost performance of the three schemes, FCFS/NPB, FCFS/PB and LCFS/PB. We fix $\alpha = 64/21$, noting that $\alpha = 64/21$ for $\rho = 4$. We fix $\mu = 10$ and vary the offered traffic intensity $\rho$. The graph exhibits the cost function $C(64/21) = E[W|L=0] + 64/21 E[W|L=1]$ versus $\rho$. It is again observed that for $\rho < 4$, systems operating under PB schemes yield better cost performance than systems operating under NPB schemes. In addition, LCFS/PB yields a lower cost than FCFS/PB for such values of $\rho$, (when $\alpha < \alpha^*$). The three schemes yield the same cost value for $\rho = 4$. As $\rho$ becomes larger than 4, the situation is reversed; NPB schemes yield better cost performance, and FCFS/PB has a lower cost than LCFS/PB.

In Figure 9, we fix $\rho = 3$, $\lambda = 50$ and $\mu = 10$. We vary the cost ratio parameter $\alpha$ and compare the cost performance of the three schemes, FCFS/NPB, FCFS/PB and LCFS/PB. The cost due to a system operating under a NPB policy is constant, due to the fact that the unserved messages never wait in such a system. The three lines intersect at the single point corresponding to $\alpha = \alpha^* = 27/13$ (the $\alpha^*$ level for $\rho = 3$). In addition, we observe, as noted previously, that the LCFS/PB scheme yields a lower cost value when $\alpha < \alpha^* = 27/13$, while it leads to a higher cost level for $\alpha > \alpha^*$.

6. Conclusion

In this paper, we have carried-out queue-size and message waiting-time analysis for finite capacity, M/G/1/N-type queueing systems, which employ a preemptive buffering discipline. Results are given for the FCFS/NPB, FCFS/PB and LCFS/PB schemes. We have also evaluated the performance tradeoffs offered by these policies, as they related to the mean and variance of the waiting-times of messages which are eventually served or eventually rejected/preempted.

The following main conclusions are drawn. When no timeliness element is involved, so that the most recent messages to arrive are not particularly more valuable, the FCFS/NPB scheme can yield acceptable performance. In turn, when a timeliness element is imposed, a preemptive buffering scheme is used. Then, in comparing the LCFS/PB and the FCFS/PB schemes, we note that the former one yields a lower average waiting-time for served messages, while the latter one leads to a lower waiting-time variance for served messages and provides a guaranteed maximum wait-time level to each message, whether served or preempted.

Allocating general cost weights to waiting-time delays experienced by served and non-served (preempted or blocked) messages, we have also compared the underlying three schemes in relation to their average cost function. We observe that when the cost ratio $\alpha$ (expressing the ratio between the cost weight of an unserved message and that of a served message) is greater than or equal to, a value very close to 4, under which level, the FCFSPB policy yields the largest variance, while when the throughput becomes larger than 4, the situation is reversed; PB schemes yield better cost performance, and FCFS/PB has a lower cost than LCFS/PB.

Figure 6 and Figure 7 exhibit performance graphs for systems operating under FCFS/PB and LCFS/PB, respectively. The cost criterion (C) is defined as:

$C(\alpha) = E[W|L=0] + \alpha E[P(L=1)|P(L=0) = 8/7$ (since $\rho = 2$), it is worthwhile to employ a PB scheme. Otherwise, an NPB procedure yields a lower cost value.

In Figure 8, using the same cost functions, we compare the cost performance of the three schemes, FCFS/NPB, FCFS/PB and LCFS/PB. We fix $\alpha = 64/21$, noting that $\alpha = 64/21$ for $\rho = 4$. We fix $\mu = 10$ and vary the offered traffic intensity $\rho$. The graph exhibits the cost function $C(64/21) = E[W|L=0] + 64/21 E[W|L=1]$ versus $\rho$. It is again observed that for $\rho < 4$, systems operating under PB schemes yield better cost performance than systems operating under NPB schemes. In addition, LCFS/PB yields a lower cost than FCFS/PB for such values of $\rho$, (when $\alpha < \alpha^*$). The three schemes yield the same cost value for $\rho = 4$. As $\rho$ becomes larger than 4, the situation is reversed; NPB schemes yield better cost performance, and FCFS/PB has a lower cost than LCFS/PB.

In Figure 9, we fix $\rho = 3$, $\lambda = 50$ and $\mu = 10$. We vary the cost ratio parameter $\alpha$ and compare the cost performance of the three schemes, FCFS/NPB, FCFS/PB and LCFS/PB. The cost due to a system operating under a NPB policy is constant, due to the fact that the unserved messages never wait in such a system. The three lines intersect at the single point corresponding to $\alpha = \alpha^* = 27/13$ (the $\alpha^*$ level for $\rho = 3$). In addition, we observe, as noted previously, that the LCFS/PB scheme yields a lower cost value when $\alpha < \alpha^* = 27/13$, while it leads to a higher cost level for $\alpha > \alpha^*$.
a served message) is lower than a threshold value $\alpha^*$ (determined by the prescribed blocking probability, $\alpha^* = P(L=1)/P(L=0)$), the preemptive-buffering policies yield lower cost values. In addition, in this region, the LCFS/PB scheme yields an average cost which is lower than that attained by a FCFS/PB policy. In turn, for $\alpha > \alpha^*$, an NPB policy yields lower average cost values.

REFERENCES

Fig. 3: The Average Served-Message Wait-time Versus the Offered Load $\rho$, for $K = 2$, $\mu = 10$

Fig. 4: The Average Served-Message Wait-Time Versus the Throughput, for $K = 2$, $\lambda = 10$

Fig. 5: The Variance of the Served-Message Wait-time Versus the Throughput, for $K = 2$, $\lambda = 10$

Fig. 6: Cost Function $C(\alpha) = E[W|L=0] + \alpha E[W|L=1]$ Versus The Throughput, under FCFS/PB, for $\rho = 2$, $K = 2$, with $\alpha^* = P(L=1)/P(L=0) = 8/7$

5C.1.9.
Fig. 7: Cost Function $C(\alpha) = \mathbb{E}[W|L=0] + \alpha \mathbb{E}[W|L=1]$ 
Versus The Throughput, under LCFS/PB, 
for $\rho = 2$, $K = 2$, with $\alpha^* = P(L=1)/P(L=0) = 8/7$

Fig. 8: The Cost Function $C(\alpha) = \mathbb{E}[W|L=0] + \alpha \mathbb{E}[W|L=1]$ 
Versus the Offered Load $\rho$, for $K=2$, $\mu = 10$, with $\alpha^* = 56/21$

Fig. 9: The Cost Function $C(\alpha) = \mathbb{E}[W|L=0] + \alpha \mathbb{E}[W|L=1]$ 
Versus $\alpha$, for $\rho = 3$, $K = 2$, with $\alpha^* = 27/13$