Cyclic Multiqueue Systems with Two Priority Classes and Exhaustive Service

Jacqueline GIANINI, Faculty of Administration, University of Ottawa, Ottawa, Canada

David MANFIELD, Bell-Northern Research, Ottawa, Canada

A cyclic multiqueue system consists of several stations in which messages are enqueued for transmission, and which are served sequentially in cyclic order by a single server. The arrivals at each queue are independent Poisson processes, and the transmission times are generally distributed. There is also a non-zero switchover time from one station to the next, which is also generally distributed. Messages at each queue can be either of two priority levels: priority 1 (low) or priority 2 (high), and polling occurs either at low priority, in which case both priority 1 and priority 2 messages can be transmitted, or at high priority, in which case only priority 2 messages are transmitted. The service discipline considered is the exhaustive service discipline: the performance, as measured by the expected delay for high and low priority messages, is evaluated. Part of the analysis is approximate, and simulation results are presented to validate the approximation.

1 INTRODUCTION

Cyclic multiqueue polling models arise naturally in computer communication networks and in the control structures of distributed switching systems [1-3]. This is because Round-Robin type polling is a good way of ensuring fair and guaranteed access to the interconnection medium for multiple distributed processors. Above and beyond this, in real time switching applications, there often exists a further requirement for the interprocessor communication to support multiple priority classes in order to maintain correct operation [3-8]. For example, in a large distributed packet switching system, fault recovery or routing update traffic need to be transmitted with priority over regular data traffic.

Multiple classes of priority add another level of complexity to the service discipline imposed by the Round-Robin polling of stations. The generic multiqueue model is depicted in figure 1.1. The queues hold messages waiting to gain access to the communications medium, represented by the server. Each station has a set of priority queues corresponding to the set of message priority classes supported by the server. Within a given priority level, each station is sequentially polled for...
messages. In general there is switching overhead associated with polling each queue even if there is no message to send. When all messages at a given priority are exhausted, the priority of the server drops to the highest priority that has a message to send, and polling continues at this new level. Conversely, if while polling at a given level a message of a higher priority arrives somewhere in the system, then the server priority jumps non-preemptively up to the priority of this new message, and polling continues at that level.

The polling service discipline at each station and priority can take many forms. It ranges from exhaustive or gated service at each queue [5,7], to various forms of non-exhaustive service where in the simplest case no more than one message is taken from each station at one polling visit [8]: in this paper we examine the exhaustive form of service discipline. In section 2, a brief overview of previous work is given in order to introduce some of the techniques which will be used in the subsequent analysis. Section 3 contains the definitions and notation for the multiqueue model with two priority classes, for which the delay performance analysis is presented in section 4; at this point, an approximation is required for part of the analysis. Numerical results with supporting results from computer simulations are given in section 5.

2 BACKGROUND

Our approach to the performance analysis of the two-priority multiqueue system will be based principally on work presented in references [9,10] for the single priority multiqueue system.

Hashida [9] considered both the exhaustive and the gated service disciplines. He began by observing that if there are $n$ messages enqueued in station $i$ at the instant of polling station $i$, then for the exhaustive service, the duration of the poll of station $i$ is the sum of $n$ busy periods of service times at station $i$. From this observation it is possible to relate the joint moment generating function of all queue lengths at the instant of polling station $i$ to the moment generating function of the same queue lengths at the instant of polling station $i + 1$. By differentiation, a set of linear equations yields the first two moments of the length of each queue at the polling instants.

The next step consists in observing that the number of messages enqueued in station $i$ at the instant of polling station $i$ is equal to the number of messages that arrived at station $i$ since the end of the last poll of station $i$. If the arrival process at each station is Poisson, queue $i$ can then be considered as an $M/G/1$ queue with server vacation time, where the server vacation time is the time elapsed since the end of the last poll of station $i$. Treating queue $i$ as an $M/G/1$ queue with server vacation time yields first the moment generating function of the length of queue $i$ at message departure times, and then the Laplace transform of the delay in station $i$. The mean delay in station $i$ can be obtained as a function of the first two moments of the server vacation time, which in turn are related to the first two moments of the queue lengths at polling instants.

A comprehensive summary of this procedure as well as of other multiqueue systems can be found in [10]. Eisenberg [11] partly reproduces Hashida's results for the exhaustive service discipline, but he considers more general polling schemes, where any station can be polled more than once in a cycle, thus giving certain stations greater access to the server.

3 TWO-PRIORITY MODEL

Consider a multiqueue system with $N$ active stations, each of which transmits variable length messages at two priority levels; each station will then have two queues: a low priority queue (priority = 1) and a high priority queue (priority = 2). Stations are polled sequentially and in a cyclic order by a single server. Each queue is served on a First-In, First-Out basis, and the service discipline is exhaustive. A station can be polled at high or at low priority: a low priority poll occurs only if there are no high priority messages present anywhere in the ring at the polling instant; otherwise, the station is polled at high priority, and only high priority messages are transmitted during that pass of the server.

The queuing model for the two-priority token ring is as follows: arrivals at station $i$, priority $j$ ($i = 1, 2, \ldots, N; j = 1, 2$) are independent Poisson processes with parameters $\lambda_{ij}$; message transmission times (service times) are arbitrarily distributed random variables $H_{ij}$, and we denote:

$$h_{ij} = E(H_{ij}), \quad h_{ij}^{(3)} = E(H_{ij}^2), \quad i = 1, \ldots, N, \quad j = 1, 2.$$  

We also denote:

$$\rho_j = \lambda_j h_{ij}, \quad i = 1, \ldots, N, \quad j = 1, 2,$$

$$\rho_i = \rho_1 + \rho_2, \quad i = 1, \ldots, N, \quad \rho = \sum_{i=1}^{N} \rho_i.$$  

$\rho_i$ represents the traffic offered by station $i$ at priority $j$, while $\rho_i$ is the total traffic offered by station $i$. $\rho$ is the total ring utilisation, and we assume that $\rho < 1$, in order to insure that the system is stable.

Moreover, the overhead time needed to switch from station $i$ to station $i + 1$ is modelled by a random variable $U_i$, we assume that the random variables $U_1, \ldots, U_N$ are independent, and we define:

$$u_i = E(U_i), \quad u_i^{(3)} = E(U_i^2), \quad i = 1, \ldots, N.$$
is a token ring system, this delay $U_i$ corresponds to the propagation delay between stations $i$ and $i+1$, plus the latency within station $i$ caused by bookkeeping operations. The total latency is $U = U_1 + \cdots + U_N$, and we denote by $u$ its expectation: $u = \sum_{i=1}^{N} u_i$.

The mean delay analysis in cyclic queues is related to the study of cycle times, or times between successive polls of a given station. In the two-priority model, since a station is polled at low priority only if there are no high priority messages enqueued anywhere on the ring at the polling instant, we define a supercycle time, or time between successive low priority polls of a given station. We define the cycle time $C_i$ as the time between successive polls of station $i$, and the supercycle time $C'_i$ as the time between two successive low priority polls of station $i$. These two definitions are illustrated in Figure 3.1.

![Figure 3.1: Cycle Time and Supercycle Time](image)

By standard methods (see for example the monograph by Takagi and Kleinrock [10]), it can be shown that:

$$E(C_i) = \frac{\sum_{i=1}^{N} \lambda_i t_i}{1 - \sum_{i=1}^{N} \lambda_i} = \frac{\lambda_i}{1 - \lambda_i}.$$  \hfill (3.1)

Our analysis of the mean delay for low priority messages is based on characteristics of the supercycle time, much in the same way that the mean delay analysis in the single priority model is based on a study of the cycle time.

During a high priority poll with exhaustive service discipline, transmission continues at station $i$ until the high priority queue is empty. The "work" or amount of service generated by a single high priority message enqueued at the polling instant in the high priority queue of station $i$ is a busy period $B_i$ with Laplace transform given by:

$$\Phi_{B_i}(s) = \Phi_{H_{11}}(s + \lambda_2 - \lambda_1; \Phi_{B_i}(s)).$$  \hfill (3.2)

we can also derive the first two moments of $B_i$:

$$E(B_i) = \frac{h_2}{1 - \rho_2,} \quad E(B_i^2) = \frac{h_2^2}{1 - \rho_2^3}$$  \hfill (3.3)

The duration of a poll at high priority is the sum of as many busy periods $B_i$ as there are high priority messages enqueued at station $i$ at the polling instant.

During a low priority poll, however, transmission continues until both the high priority and the low priority queues are empty. The service discipline within each queue is First-In, First-Out, but high priority messages have non-preemptive priority over low priority messages. The work generated by a single low priority message enqueued at station $i$ at a low priority polling instant

is a modified busy period, which is initiated by a low priority message, but in which all arriving messages at either of the two priority queues are transmitted until both queues are empty. Let $B'_i$ be the length of this "modified busy period"; $B'_i$ begins with a low priority service, and continues with high priority services until the high priority queue becomes empty again. If any low priority messages have arrived in the meantime, this extended service is repeated until both the high and the low priority queues are empty. Let $C'_i$ be the length of an extended service at low priority (or time until the high priority queue becomes empty again); we have:

$$\Phi_{B'_i}(s) = \Phi_{H_{11}}(s + \lambda_1 - \lambda_1; \Phi_{B'_i}(s)),$$  \hfill (3.4)

with

$$\Phi_{B'_i}(s) = \Phi_{H_{11}}(s + \lambda_2 - \lambda_2; \Phi_{B'_i}(s)),$$  \hfill (3.5)

where $\Phi_{B'_i}(s)$ is as defined in (3.2); the mean and the second moment of $B'_i$ can be obtained as follows:

$$E(B'_i) = \frac{h_2}{1 - \rho_1'}, \quad E(B'_i^2) = \frac{h_2^2}{1 - \rho_1'^3} + \frac{\lambda_2 h_2}{(1 - \rho_1')^3}.$$  \hfill (3.6)

Since $B'_i$ is the length of service generated by a single low priority message enqueued at station $i$ at a low priority polling instant, the duration of a low priority poll is the sum of as many modified busy periods $B'_i$ as there are low priority messages enqueued in station $i$ at the low priority polling instant of station $i$.

4 ANALYSIS OF THE EXHAUSTIVE SERVICE SYSTEM

Our aim in this section is to find an approximation to the mean delay for both types of priority messages at each station. To do this, we are going to adapt the procedure followed by Haigh [5] as presented by Takagi and Kleinrock [10]. In section 4.1, we begin by determining the joint moment generating function of all queue lengths at polling instants. The first two moments of the length of both priority queues at station $i$ at the polling instant of station $i$ are used in section 4.2 to calculate the mean delay of messages in station $i$. These first two moments depend on the expected number of low priority messages present at station $i$ at an arbitrary polling instant of station $i$, but this quantity cannot be determined directly from the moment generating functions, and is the subject of an approximate analysis, in section 4.3.

4.1 Number of messages at polling instants

Let us define:

- $L_{2i} = $ number of type 2 (high priority) messages present in station $i$ at instant $t$;
- $L_{1i} = $ number of type 1 (low priority) messages present in station $i$ at instant $t$; and
- $r_i(n) = n$-th polling instant of station $i$.

Also, let:

- $L_{2i}^0 = $ number of type 2 messages present in station $i$ at an arbitrary polling instant of station $i$, and
- $L_{1i}^0 = $ number of type 1 messages present in station $i$ at an arbitrary low priority polling instant of station $i$.

If $F_t$ is the joint moment generating function of $(L_{2i}(t), L_{2i+1}(t), \ldots, L_{N_{2i+1}}(t), L_{21}(t), L_{22}(t), \ldots, L_{N_{21}}(t))$ at time $t = r_i(n)$, where $n$ is sufficiently large for steady state to have been achieved, we can relate $F_t(x_1, \ldots, x_n, y_1, \ldots, y_n)$ to $F_{11}(x_1, \ldots, x_n, y_1, \ldots, y_n)$ by noting that the service
time for station \( i \) is the sum of \( L_{i,j}(r_i(n)) \) busy periods \( B_j \) (defined in (3.2)), provided \( L_{i,j}(r_i(n)) \neq 0 \) for some \( j \), whereas if \( L_{i,j}(r_i(n)) = 0 \) for all \( j = 1, \ldots, N \), the service time for station \( i \) is the sum of \( L_{i,j}(r_i(n)) \) modified busy periods \( B'_j \) (defined in (2.4) and (3.5)). Hence, by arguments similar to those presented in Takagi and Kleinrock [10], we have:

\[
F_{\hat{t}}(x_1, \ldots, x_N, y_1, \ldots, y_N) = \Phi_{V_0} \left( \sum_{j=1}^{N} \left( \lambda_j x_j + \lambda_j y_j \right) \right) \times \left[ F_i(x_1, \ldots, x_{i-1}, \Phi_{V_i} \left( \sum_{j=i+1}^{N} \left( \lambda_j x_j + \lambda_j y_j \right) \right), x_{i+1}, \ldots, y_N \right] - F_i(0, \ldots, 0, y_1, \ldots, y_N) + F_i(0, \ldots, 0, y_1, \ldots, y_{i-1}, \Phi_{V_i} \left( \sum_{j=i+1}^{N} \left( \lambda_j x_j + \lambda_j y_j \right) \right), y_{i+1}, \ldots, y_N) \right]
\]  

(4.1)

Now, let

\[
f_i(x_j) = \frac{\partial}{\partial x_j} F_i(x_1, \ldots, x_N, y_1, \ldots, y_N) \bigg|_{x_1=\ldots=x_N=y_1=\ldots=y_N=1} = E(L_{i,j}(r_i(n)))
\]

and

\[
g_{0,i}(x_j) = \frac{\partial}{\partial y_j} F_i(0, \ldots, 0, y_1, \ldots, y_N) \bigg|_{x_1=\ldots=x_N=0} = E(L_{i,j}(r_i(n)))
\]

be, respectively, the expected number of high priority messages and the expected number of low priority messages present at station \( j \) at the instant when station \( i \) is polled; in particular, \( f_i(x) = E(L_{i,j}) \) is the expected number of high priority messages found by the server at an arbitrary polling instant of station \( i \); we can then define

\[
g_{0,i}(x) = \frac{\partial}{\partial y_j} F_i(0, \ldots, 0, y_1, \ldots, y_N) \bigg|_{y_1=\ldots=y_N=0} = E(L_{i,j}(r_i(n)))
\]

and the expected number of low priority messages found by the server at an arbitrary polling instant of station \( i \) is:

\[
E(L_{i,j}) = E(L_{i,j}(r_i(n))) \bigg|_{I_{0,i} = 0} = g_{0,i}(x).
\]

By differentiating equation (4.1) with respect to each of the \( x_i \)’s and each of the \( y_i \)’s, we obtain a set of \( 2N^2 \) equations in \( f_i(x_j) \) and \( g_{0,i}(x_j) \) (for \( i = 1, \ldots, N \)):

\[
f_{i+1}(x_j) = \lambda_j x_j + f_i(x_j) + \frac{\lambda_j h_i}{1 - \rho_i} f_i(x_j) + \frac{\lambda_j h_{0,i}}{1 - \rho_i} g_{0,i}(x_j) \quad (j \neq i);
\]

\[
g_{i+1}(x_j) = \lambda_j x_j + g_i(x_j) + \frac{\lambda_j h_i}{1 - \rho_i} f_i(x_j) - g_{0,i}(x_j) \quad (j \neq i).
\]

(4.2)

(4.3)

Note that we only have \( 2N^2 \) equations in \( 2N^2 + N \) variables: these equations do not suffice to determine all \( f_i(x_j) \)’s, \( g_i(x_j) \)’s and \( g_{0,i}(x_j) \)’s, but by summing equations (4.3) with respect to \( i \), we obtain:

\[
g_{0,j}(x_j) = \lambda_j \frac{1 - \rho_j}{1 - \rho_i} \left[ \sum_{i=1}^{N} u_i + \sum_{i=1}^{N} \frac{h_{0,2}}{1 - \rho_i} f_i(x_j) + \sum_{i=1}^{N} \frac{h_{0,1}}{1 - \rho_i} g_{0,i}(x_j) \right].
\]

(4.4)

and (4.4), together with (4.2), form a system of \( N^2 + N \) equations in \( f_i(x_j) \) and \( g_{0,i}(x_j) \), \( (i, j = 1, \ldots, N) \), which has as its unique solution:

\[
f_i(x_j) = \lambda_j \frac{1 - \rho_i}{1 - \rho_j} \left[ \sum_{k=1}^{i-1} u_k + \sum_{k=i+1}^{N} \frac{h_{0,2}}{1 - \rho_k} f_k(x_j) \right] + g_{0,i}(x_j),
\]

\[
g_{0,i}(x_j) = \lambda_i \frac{1 - \rho_i}{1 - \rho_j} \left[ \sum_{k=1}^{i-1} u_k + \sum_{k=i+1}^{N} \frac{h_{0,1}}{1 - \rho_k} g_{0,k}(x_j) \right],
\]

where \( c = E(C_i) = \frac{1}{1 - \rho} \) is the expected cycle length (see (3.1)).

From (4.2) we can also obtain:

\[
g_i(x_j) = \lambda_i \left[ \sum_{k=1}^{i-1} u_k + \left( \sum_{k=i+1}^{N} \frac{h_{0,2}}{1 - \rho_k} f_k(x_j) \right) + g_{0,i}(x_j) \right],
\]

but we are still unable to calculate \( g_i(x_j) \) (for \( i = 1, \ldots, N \)); an approximate method for calculating \( g_i(x_j) \) will be given in section 4.3.

In order to find expressions for the second moments of \( (L_{i,j}(r_i(n)), L_{i,j}(r_i(n)), \ldots, L_{i,j}(r_i(n))) \), \( (L_{i,j}(r_i(n)), L_{i,j}(r_i(n)), \ldots, L_{i,j}(r_i(n))) \) times \( t = r_i(n) \), we define:

\[
f_i(x_j, k) = \frac{\partial^2}{\partial x_j \partial x_k} F_i(x_1, \ldots, x_N, y_1, \ldots, y_N) \bigg|_{x_1=\ldots=x_N=y_1=\ldots=y_N=1} = E(L_{i,j}(r_i(n)))
\]

and

\[
g_{0,i}(x_j, k) = \frac{\partial^2}{\partial y_j \partial y_k} F_i(0, \ldots, 0, y_1, \ldots, y_N) \bigg|_{y_1=\ldots=y_N=0} = E(L_{i,j}(r_i(n)))
\]

Note that \( f_i(x_j, k) = E(L_{i,j}(r_i(n) - 1)) \), and that \( g_{0,i}(x_j, k)/p_j = E(L_{i,j}(r_i(n) - 1)) \), and the second moments of the queue lengths at polling instants can be obtained by differentiating (4.1) twice. We obtain a system of \( 3N^2 \) equations in \( f_i(x_j, k), g_i(x_j, k), h_i(x_j, k) \) and \( g_{0,i}(x_j, k) \) (for \( i = 1, \ldots, N \)); this system of \( 3N^2 \) equations in \( 3N^2 + N^2 \) variables can be reduced by summing with respect to \( j \), yielding a system of \( 3N^2 \) equations in \( f_i(x_j, k), h_i(x_j, k) \) and \( g_{0,i}(x_j, k) \) (for \( i = 1, \ldots, N \)). This system, though large, could then be solved to obtain \( f_i(x_j, k) \) and \( g_{0,i}(x_j, k) \) (for \( i = 1, \ldots, N \)), the new equations being as follows:

\[
0 = \lambda_j x_j + \sum_{i=1}^{N} u_i f_i(0, x_j) + \frac{\lambda_j h_{0,2}}{1 - \rho_i} f_i(0, x_j) + \sum_{i=1}^{N} \frac{h_{0,1}}{1 - \rho_i} g_{0,i}(x_j) \quad (j \neq i);
\]

\[
g_{0,i}(x_j) = \lambda_i \frac{1 - \rho_i}{1 - \rho_j} \left[ \sum_{k=1}^{i-1} u_k + \sum_{k=i+1}^{N} \frac{h_{0,2}}{1 - \rho_k} f_k(x_j) \right] + \sum_{i=1}^{N} \frac{h_{0,1}}{1 - \rho_i} g_{0,i}(x_j) \quad (j \neq i).
\]

(4.5)

(4.6)
where:

\[ \delta = \begin{cases} 1 & \text{if } j \neq k, \\ 0 & \text{if } j = k; \end{cases} \]

\[
0 = \lambda_1 \lambda_k \sum_{i=1}^{N} u_i^{(2)} + \lambda_1 \sum_{i,p,k} u_i \left[ g_i(k) + \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_2 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_3 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_4 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_5 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_6 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_7 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right]
\]

and

\[
0 = \lambda_2 \sum_{i=1}^{N} u_i^{(2)} + \lambda_2 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_3 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_4 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_5 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_6 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right] + \lambda_7 \sum_{i,p,k} u_i \left[ \frac{\lambda_i h_{12}}{1 - \rho_{12}} f_i(k) + \frac{\lambda_i h_{21}}{1 - \rho_{21}} g_i(0)(i) \right]
\]

In the symmetrical case, where all stations have the same input rates (\( \lambda_{11} = \ldots = \lambda_{N1} \), and \( \lambda_{12} = \ldots = \lambda_{N2} \), and the same service time distribution, and where all switching overhead times have the same distribution, we have:

\[
g_i^{(0)}(i,i) = \frac{\lambda_i}{1 - \rho_i} \left[ \left( N - 1 \right) \lambda_i \right] \left[ \left( N - 1 \right) \lambda_i \right] \frac{1}{1 - \rho_i} \left( N + 1 \right) u_i c
\]

and by differentiation, we obtain:

\[
E(W_{i,j}) = \frac{E(B_i)}{2(\lambda_i + \mu_i)^2} + \frac{\lambda_i h_{12}}{2(1 - \rho_{12})^2(1 - \rho_{12})} - \frac{\lambda_i h_{21}}{2(1 - \rho_{12})^2(1 - \rho_{12})} \frac{1}{1 - \rho_{12}} E(B_i^2),
\]

with \( E(B_i) \) and \( E(B_i^2) \) as calculated in (3.3) and (3.6), respectively. We note that these expressions still contain two unknown parameters, \( \rho_i \) and \( g_i(0) \); we shall find an approximation to these in section 4.3.

### 4.2 Mean delay

To find an expression for the mean delay, we proceed as Eisenberg [11] and Hashida [9], and consider the queuing system in station \( i \) as an \( M/G/1 \) queue with server vacation time. Let us first consider the low priority queue: the supercycle \( C_i \), introduced in section 3, can be separated into a service time at low priority \( T_i \), and an interarrival time, or server vacation time \( A_i \), which lasts until the next low priority poll of station \( i \). Figure 4.1 illustrates these definitions.

\[
\text{FIGURE 4.1: SERVICE TIME AND INTERARRIVAL TIME AT LOW PRIORITY}
\]

Let \( S_i \) be the length of an extended service at low priority, as defined in (3.5); the moment generating function \( G_{S_i} \) of the low priority queue length at extended service completion times can be obtained as:

\[
G_{S_i} = E_{1}^{\left( 1 - e^{-\lambda_i E(S_i)} \right)} \frac{\lambda_i}{1 - \rho_i} \Phi_{S_i} \left( \frac{\lambda_i}{1 - \rho_i} e^{-\lambda_i E(S_i)} \right) - 1
\]

and the Laplace transform of the waiting time of messages in the low priority queue is then:

\[
\Phi_{S_i} = \frac{G_{S_i}}{1 - \Phi_{S_i}} = \frac{1}{\Phi_{S_i}} - \frac{1}{\Phi_{S_i} - \lambda_i E(S_i)} - \frac{1}{\phi_{S_i}(\lambda_i - E(S_i))}
\]

by differentiation, we obtain:

\[
E(W_i) = \frac{E(B_i)}{2(\lambda_i + \mu_i)} + \frac{\lambda_i h_{12}}{2(1 - \rho_{12})^2(1 - \rho_{12})} - \frac{\lambda_i h_{21}}{2(1 - \rho_{12})^2(1 - \rho_{12})} \frac{1}{1 - \rho_{12}} E(B_i^2),
\]

Let \( f_i^{(0)}(i) \) be equal to \( E(x_1, \ldots, x_N, y_1, \ldots, y_N) \) with \( x_k = 0 \) for \( k = 1, \ldots, N \), \( y_k = 1 \) for \( k \neq i \), and \( y_i = y \). Finally, we note that since the number of low priority messages present at the instant of a low priority poll of station \( i \) is equal to the number of low priority messages that have arrived during \( A_i \) the Laplace transform of \( A_i \) is:

\[
\Phi_{A_i}(s) = \Phi_{S_i}^{(0)}(1 - s/\lambda_i)
\]

and by differentiation, we have:

\[
E(A_i) = \frac{1}{\lambda_i} \int f_i^{(0)}(i) \, dt = \frac{1}{\lambda_i} \frac{1}{\mu_i} \times \frac{1}{\rho_i}
\]
To calculate $E(A_i^1)$ and $E(A_i^2)$, we can relate the Laplace transform of $A_i$ to the moment generating function of $L_{i2}$ by noting that the number of high priority messages present at an arbitrary polling instant of station $i$ is equal to the number of high priority messages that arrived during $A_i$:

$$
\Phi_i(s) = \Phi_i(1 - s/\lambda_{i2})
$$

where $\Phi_i(1 - s/\lambda_{i2})$ is equal to $E(X_{i1}, \ldots, X_{iN}; Y_{i1}, \ldots, Y_{iN})$ with $x_k = 1$ for $k \neq i$, $x_i = 1 - s/\lambda_{i2}$, and $y_k = 1$ for $k = 1, \ldots, N$; finally, we have for high priority messages:

$$
E(A_i^1) = \frac{f_i(t)}{\lambda_{i2}} (1 - \rho_{i2}), \quad E(A_i^2) = \frac{f_i(t)}{\lambda_{i2}} \frac{h_{i2}^{(2)}}{1 - \rho_{i2}},
$$

and using (4.13) and (4.14) in (4.12), we have for high priority messages:

$$
E(W_{i2}) = \frac{1}{1 - \rho_{i2}} \frac{f_i(t)}{\lambda_{i2}} \frac{h_{i2}^{(2)}}{1 - \rho_{i2}} \frac{1}{2} \frac{h_{i2}^{(2)}}{1 - \rho_{i2}} + \frac{\lambda_{i1} h_{i2}^{(2)}}{2(1 - \rho_{i2})^2}.
$$

4.3 The approximation for $g_i(t)$

From expression (4.5), we could calculate $g_i(t)$ in the symmetric case if $g_i(t)$ was known; we have seen however that equations (4.2) and (4.3) do not suffice to determine $g_i(t)$. Let us instead remember that $g_i(t)$ is equal to $E(L_{i2}(t))$ when $t$ is an arbitrary polling instant of station $i$; if $g_i(t)$ is then equal to the expected number of low priority messages that have arrived at station $i$ since the end of the last low priority polling of station $i$. If $V_i$ is the time elapsed since the end of the last low priority polling of station $i$, at an arbitrary polling instant of station $i$, we have:

$$
g_i(t) = \lambda_{i1} E(V_i).
$$

To calculate the expectation of $V_i$, we must look at a regenerative cycle for which the regeneration points are the instants at which all stations are empty of both high and low priority messages. We define $X_i$ to be the number of cycles in a supercycle. By standard regenerative arguments, and assuming that $X_i$ is a stopping time with respect to the sequence of cycle times, we have $E(C_i) = E(X_i)\mu_i$, and since we know that $E(C_i) = \epsilon/\mu_i$ (see (4.9)), it follows that $E(X_i) = 1/\mu_i$. Also by regenerative arguments, we have:

$$
E(V_i) = E(X_i^1) + E(X_i^2) - \frac{E(X_i^1)}{2 E(X_i^2)} - \frac{E(X_i^2)}{E(X_i^2)} - \frac{E(X_i^1)}{E(X_i^2)}.
$$

Now, we have seen in (4.8) that $E(X_i^1) = (\rho_{i1}/(1 - \rho_{i1})) (\epsilon/\mu_i) = (\rho_{i1}/(1 - \rho_{i1})) E(X_i)$; in order to express the relationship between $X_i$ and $T_i$, we shall assume that $E(X_i) = (\rho_{i1}/(1 - \rho_{i1})) c E(X_i^1)$, and (4.16) becomes:

$$
E(V_i) = E(X_i^1) + E(X_i^2) - \frac{E(X_i^1)}{2 E(X_i^2)} - \frac{E(X_i^2)}{E(X_i^2)} - \frac{E(X_i^1)}{E(X_i^2)} c
$$

and, since by (4.15) $g_i(t) = \lambda_{i1} E(V_i)$,

$$
g_i(t) = \lambda_{i1} \left( E(X_i^1) + \frac{1}{2} \frac{E(X_i^1)}{1 - \rho_{i2}} \right).
$$

The second part of our approximation deals with estimating $E(X_i^1)$ and $E(X_i^2)$. We present here the approximation for the symmetric case only, although the same method extends easily to
the general case. Let us begin by estimating $p_i$: to do this, we consider an arbitrary polling instant, and we calculate the expected time elapsed since the last poll of station $j$ for $j = 1, \ldots, N$. If we denote by $t_j$ this expected time, then the probability that station $j$ has no high priority messages enqueued at the instant of polling station $i$ is approximated by $e^{-\lambda_j t_j}$, and the probability that all stations are empty of high priority messages at the instant of polling station $i$ is approximated by

$$p_i = \prod_{j=1}^{N} e^{-\lambda_j t_j} = e^{-\sum_{j=1}^{N} \lambda_j t_j},$$

which in the symmetric case reduces to

$$p_i = e^{-\lambda t \sum_{j=1}^{N} t_j}.$$

Now, to calculate $t_j$ for $j = 1, \ldots, N$, we note that the expected time spent polling station $j$ in an arbitrary cycle is $\rho_j c$, which in the symmetric case equals $\rho c$ for all $j = 1, \ldots, N$, while the overhead time associated with station $j$ is $u_j$, which equals $u$ in the symmetric case. We have thus:

$$t_j = N u_j + (N - 1) \rho_j c$$

$$t_{j+1} = (N - 1) u_j + (N - 2) \rho c$$

$$\vdots$$

$$t_{N-1} = u_N,$$

and by solving equations (4.18),

$$\sum_{j=1}^{N} t_j = N^2 u + N(N - 1) \rho c = e \left( N + 1 - 2N \rho \right),$$

which yields:

$$p_i = e^{-\lambda t \left( N^2 u + N(N - 1) \rho c \right)},$$

and, since $E(X_i) = 1 / p_i$,

$$E(X_i) = \frac{1}{p_i} = e^{\lambda t \left( N + 1 - 1N^2 \right)}.$$

It would be tempting now to assume that $X_i$ is a geometric random variable with parameter $p_i$; it is clear however that the first cycle in a supercycle is short, because at the beginning of a supercycle there are no high priority messages enqueued anywhere in the ring, and consequently, fewer high priority messages are served during the first cycle than on average. Consequently, the poll immediately following a low priority poll is more likely to be at low priority than if it was preceded by a high priority poll. The supercycles are thus either very short ($X_i = 1$) or very long, and the second moment of $X_i$ is larger than that of a geometric random variable.

Let $q_i$ be the probability of polling station $i$ at low priority if the previous poll was also at low priority, and let $e_k$, $(k = 1, \ldots, N)$ be the expected duration of the poll of station $i + k$ during the first cycle of a supercycle. We first note that $e_k$ is equal to $E(T_{i+k})$, where $T_{i+k}$ is the duration of the low priority poll. In order to calculate $e_k$ for $k = 1, \ldots, N - 1$, we observe that the number of high priority messages enqueued in station $i + k$ at the instant of polling station $i + k$ is equal to the number of high priority messages that have arrived since the beginning of the low priority poll of station $i$; let $r_k$ be that number: $r_k = \lambda_1 (s_0 + \cdots + s_{k-1} + ku_k)$, and the time spent serving high priority messages during a high priority poll of station $i + k$ is the sum of $r_k$ busy periods of high priority services, and its expectation is $r_k \lambda_2 (1 - \rho_{12})$. As for the low priority messages, we assume that the average time spent polling station $i + k$ at low priority is approximately the same in all cycles, and is consequently equal to $\rho c / (1 - \rho_{23})$. We obtain a set of recursive equations for $s_k$:

$$s_0 = \frac{\rho c}{1 - \rho_{23}}$$

$$s_1 = \frac{\rho c + 1}{1 - \rho_{12}}$$

$$s_{N-1} = \frac{\rho c + 1}{1 - \rho_{12}}$$

and an estimate of $q_i$ in the symmetric case is:

$$q_i = e^{-\lambda t \left( N^2 u + N(N - 1) \rho c \right)} = \frac{1}{p_i} = e^{\lambda t \left( N + 1 - 1N^2 \right)}.$$

To estimate the second moment of $X_i$, we shall assume that $X_i$ is a linear combination of two geometric random variables:

$$X_i = \alpha X_{ia} + (1 - \alpha) X_{ib},$$

where $X_{ia}$ is geometric with parameter $\rho_{ia}$ and $X_{ib}$ is geometric with parameter $\rho_{ib}$. We have the constraints $E(X_i) = 1 / p_i$, and $P(X_i = 1) = q_i$, and we shall also assume that $\alpha / \rho_{ia} = (1 - \alpha) / \rho_{ib} = 1 / 2p_i$. We can now maximise $E(X_i^2)$, subject to these constraints, and we obtain:

$$E(X_i^2) = \frac{1}{p_i^2} \left( 2 \frac{q_i}{p_i} - 1 \right),$$

and we can now use (4.20) and (4.19) in (4.17) to estimate $\rho(t)$.

5 NUMERICAL RESULTS

In this section we present a few numerical examples with comparisons to computer simulations, in each example, the values of $W_{11}$ and of $W_{22}$, as calculated using our analysis, are compared with the results of a GPSS simulation, given as 95% confidence intervals. The accuracy of the approximation is measured by the relative error, which is the absolute difference between the calculated and the simulated delays, divided by the simulated delay.

Our first set of results (see table 5.1) is for a system consisting of 8 stations; the switching overhead times are constant and equal to unity, and the service times at both priorities are exponentially distributed with a mean equal to 10. The high priority utilisation $\rho_2$ is fixed at 0.20, while we let the low priority utilisation $\rho_1$ increase.

Our second set of results (see table 5.2) is again for a system with 8 stations. The switching overhead times and the service durations are as in the previous example, but here the low priority utilisation $\rho_1$ is equal to 0.05 of the total utilisation $\rho$. In this example, as in our first example, the accuracy of the approximation is quite good, the relative errors remaining well under 10% at utilisations of up to 80%.

For our last example (see table 5.3), we consider a system consisting of 20 stations, with the same service time distributions and the same switching overhead times as in the previous example.
<table>
<thead>
<tr>
<th>$p$</th>
<th>Low Priority</th>
<th>High Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_{sim}$</td>
<td>$W_{calc}$</td>
</tr>
<tr>
<td>.20</td>
<td>0 ± 0</td>
<td>0</td>
</tr>
<tr>
<td>.40</td>
<td>19.20 ± .03</td>
<td>18.96</td>
</tr>
<tr>
<td>.50</td>
<td>34.40 ± 0.09</td>
<td>34.35</td>
</tr>
<tr>
<td>.70</td>
<td>52.95 ± 4.20</td>
<td>51.26</td>
</tr>
<tr>
<td>.75</td>
<td>67.43 ± 3.69</td>
<td>65.50</td>
</tr>
<tr>
<td>.80</td>
<td>93.37 ± 13.24</td>
<td>88.13</td>
</tr>
</tbody>
</table>

**TABLE 5.1**

$N = 8, h_{12} = h_{21} = 10, u_i = 1, u_i^{(x)} = 1, \rho_1 = 0.2, \rho_2 = \rho - \rho_2$, Markovian services

<table>
<thead>
<tr>
<th>$p$</th>
<th>Low Priority</th>
<th>High Priority</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$W_{sim}$</td>
<td>$W_{calc}$</td>
</tr>
<tr>
<td>.20</td>
<td>0.20 ± 0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>.40</td>
<td>19.61 ± 1.05</td>
<td>19.79</td>
</tr>
<tr>
<td>.60</td>
<td>49.86 ± 4.03</td>
<td>47.33</td>
</tr>
<tr>
<td>.70</td>
<td>89.97 ± 7.45</td>
<td>83.14</td>
</tr>
<tr>
<td>.75</td>
<td>126.93 ± 8.11</td>
<td>118.80</td>
</tr>
<tr>
<td>.80</td>
<td>191.21 ± 16.34</td>
<td>187.67</td>
</tr>
</tbody>
</table>

**TABLE 5.2**

$N = 8, h_{12} = h_{21} = 10, u_i = 1, u_i^{(x)} = 1, \rho_1 = 0.40, \rho_2 = 0.60, \rho$, Markovian services

Because of the large number of stations, combined with the long switching overhead times (the time between polls of station i when no message is sent during a cycle is equal to 20, or two times the service duration), the low priority delays are extremely large, even at relatively low utilizations. Nevertheless, we observe that the approximation is quite accurate, the relative errors remaining well below 10% in all cases except for the low priority delay at 80% utilization; in this case, however, the low priority delay is so huge (20 times the average service duration) that it would be unacceptable in any realistic application.

6 CONCLUSIONS

In this paper we have presented a method for analysing multiqueue systems with two priority classes. This method is exact, except for the determination of the average number of low priority messages present at any station at the time that station is polled. This method could be extended to systems with three or more priorities, but because it involves solving large systems of linear equations, such an extension would be of little practical use. As a consequence, there is some need for an approximate method permitting to solve the multiple priority problem.

REFERENCES


Computer Science.

20, 440-451.

Ajmone Marsan, A., 243
Albanese A., 173
Auber, J., 423
Barotz, A., 405
Becker, M., 423
Borgonovo, E., 161
Brown, P., 39
Buzzoni, P., 193
Cadorin, E., 181
Capellini, G., 193
De Marca, J.R.B., 453
De Marca, J.R.B., 453
De Moraes, L.F.M., 283
De Nitto Persone, V., 211
Dias, D., 43
Eaton, C.K., 107
Fayolle, G., 155, 407
Fdida, S., 441
Feldman, P.M., 139
Fratta, L., 173
Gall, H.R., 327
Gianini, J., 513
Gnoa, L., 405
Goyal, A., 13
Grillo, D., 211
Guarneri, M.R., 193
Hantler, S.L., 327
Hasegawa, T., 227
Hin, S.D., 471
Hsiao, W.L.Y., 343
Houw, C.E., 377
Huitema, C., 89
Iliadis, I., 359
Kemani, P., 405
Körner, U., 441
Kuehn, P.J., 471
Lebourie, B., 89
Lazar, A.A., 343
LeMaster, R.J., 97
Levin, E., 107
Li, V.K., 139
Lien, T.C., 359
López Valverde, A.N., 283
Livio, A., 23
Mach, F., 193
Manfield, D., 513
Mork, J.W., 71
Perros, H.G., 441
Prost, J-P., 423
Pruchal, P., 3
Rasberg, Z., 57
Restagno, D., 243
Rouillet, G., 23
Rubin, I., 315
Runkosz, M., 393
Sandoval, R., 393
Schacham, N., 57
Shapiro, G., 441
Silvester, J.A., 125
Simonian, A., 39
Szpankowski, W., 297
Takagi, H., 263
Takahashi, K., 227
Takine, T., 227
Taylor, B.A., 327
Thomas, R., 155
Toft, F., 23
Tran-Gia, P., 497
Troncone, S., 243
Wang, J.L., 125
Yang, K.W., 71
Young, B., 107
Zhang, Z., 315