ANALYSIS OF A PRIORITY FEEDER ON A FIFO SERVER

by

Rajesh Mansharamani and Miron Livny

Computer Sciences Technical Report #975

October 1990

Analysis of a Priority Feeder on a FIFO Server

Rajesh Mansharamani   Miron Livny

Computer Sciences Department
University of Wisconsin-Madison.

October 17, 1990

Abstract

The impact of the departure process from a priority queue, on the response time in a subsequent tandem queue is not well understood. In this paper, a model of a priority queue feeding a $FIFO$ queue is used to analyze the response time of both high and low priority jobs in the second queue. For both preemptive and non-preemptive priority, we show that the response time of each class in the $FIFO$ queue is a function of the first two moments of the service times of both queues and the *idle time* in the $FIFO$ queue. These results depict that having priority scheduling in the first queue, can cause high priority jobs to gain in the second queue as well. We also show that when the scheduling discipline in the first queue is non-preemptive, conservation of response time holds in the second queue. We thereby derive bounds for the response times of both high and low priority jobs in the $FIFO$ queue. Knowledge of first and second moments of service times in each queue, is sufficient to bound the response time in the $FIFO$ queue.

1 Introduction

In computer networks messages often have to pass through a number of nodes in sequence. Similarly, in assembly systems, jobs sequentially go through a number of stages. When all jobs are of the same kind, they usually get served in a *first-in-first-out* ($FIFO$) order, in each stage or layer. Sometimes however, jobs may belong to different classes and thus have different response time constraints. For example in real-time systems, like communication networks, deadlines have to be met for voice packets but not for data packets. In order to meet these deadlines, classes with more stringent timing constraints are given higher priority.

Introduction of priority increases the complexity of analyzing the performance of such 'layered' systems. Priority scheduling in the upper layers or stages may affect the response time of jobs in subsequent layers too. Hence the problem that we address is : what are the response times of high and low priority jobs in the stage just below the layer with priority scheduling?

In this paper we consider a $FIFO$ queue that is fed by a priority queue. Previous works dealing with a $FIFO$ queue being fed by an upper level queue, consider the feeder queue to be $FIFO$ [4] [5]. For some special cases of service time distributions in the feeder queue to which the arrival process is a general renewal process, [4] derives bounds on the response time in the second queue. [5] gives useful orderings for response times in the second queue under general service time distributions in the feeder. In this paper, we focus on a $FIFO$ queue being fed by a priority queue, and derive expressions for the response time of high and low priority classes in the $FIFO$ queue. The service time distributions of the servers in both queues are general, and the arrival process to the priority feeder is a Poisson process. On the basis of results from [4], an upper bound for the response time in the $FIFO$ queue is derived, for the case where the feeder has non-preemptive priority scheduling. We also see that response time is 'conserved' in both queues when there is non-preemptive priority in the first, just like the conservation of response time in a single $G/G/1$ queue with non-preemptive priority scheduling [3].

We now lay out the organization of this paper. Section 2 lists the assumptions and notation. In section 3 we derive equations for the response time in the second queue, for both preemptive and non-preemptive scheduling in the first queue. Section 4 presents bounds for high and low priority jobs, when there is non-preemptive priority in the feeder queue. Finally section 5 summarizes this work, and presents the conclusions drawn from it.

2 Assumptions and Notation

All the results in this paper are for a queueing system that consists of a priority queue, Q_1, feeding a $FIFO$ queue, Q_2. The system and its environment are completely specified by the arrival process to the first queue, the distribution of service times in each queue, and the priority queueing discipline in the first queue. Two classes of customers arrive to the priority queue, viz. high priority customers denoted by 'h', and low priority customers denoted by 'l'. Both these classes arrive according to an independent Poisson process [2]; the arrival rate of high priority jobs being λ^h, and that of low priority jobs being λ^l. The service time distribution in both queues can be any general distribution; the mean service times in Q_1 and in Q_2 are $\frac{1}{\mu_1}$ and $\frac{1}{\mu_2}$ respectively. The queueing discipline in Q_1 is head-of-the-line (HOL) priority [3]; scheduling can be either non-preemptive (NP) or preemptive-resume (PR).

We make three major assumptions. First, the total arrival rate to the system, $\lambda = \lambda^h + \lambda^l$, is less than the service rate of either queue, μ_1 and μ_2. Second, we assume no dependency in the service times of jobs in the two queues. Thus, a job makes independent samples for its service time in both queues. Third, it is assumed that each queue is work-conserving [3] (i.e. there is no creation or destruction of work).

Notation	Description
HOL	Head-of-the-line queueing discipline
NP	Non-preemptive HOL
PR	Preemptive Resume HOL
$FIFO$	First In First Out queueing discipline
$x \rightarrow y$	Queueing discipline x in Q_1. Queueing discipline y in Q_2
Q_i	The i^{th} queue in the system (where i is either 1 or 2)
C_{x_i}	Coefficient of variation of the service times in Q_i
λ	Total arrival rate to the system
μ_i	Service rate of server i
ρ_i	Utilization of Q_i
T_i	Response time of a job in Q_i (averaged over all classes)
W_i	Waiting time of a job in Q_i
N_1	Number of customers in Q_1 seen by an arrival
N_2	Number of customers in Q_2 seen by a customer arriving to Q_1
Y_2	Work remaining to be done in Q_2 when there is an arrival of a job,J, to Q_1
V_2	Total work in Q_2 of the jobs ahead of J, in Q_1
I_2	Total idle time in Q_2 during J's system time in Q_1
s_{il}	Service time in Q_l of the i^{th} job ahead of J in Q_1

Table 2.1 Notation

The notation used throughout this paper, is given in Table 2.1. All variables corresponding to high and low priority classes have a superscript 'h' and 'l' respectively. For example, T_2^h is the response time of a high priority job in Q_2. Similarly ρ_2^h is the utilization of high priority jobs in the second queue. (the priority queueing discipline in the first queue will be clear from the context)

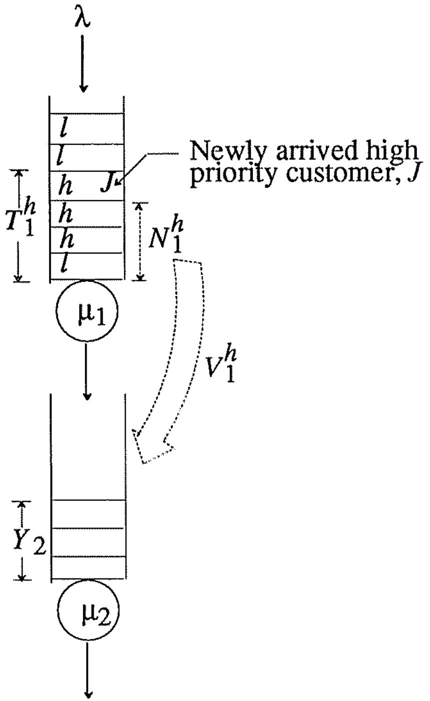

Figure 1. System of Tandem Queues

3 Response Time Equations

3.1 $NP \rightarrow FIFO$

Here we consider the $NP \rightarrow FIFO$ case, where a priority queue with non-preemptive head-of-the-line priority (NP) feeds a $FIFO$ queue. We derive an expression for the mean response time in the $FIFO$ queue for both high and low priority classes. Consider a high priority job, J, which has just arrived to the NP queue. Upon arrival, J sees N_1^h jobs ahead of it in the NP queue, Q_1, and Y_2 as the total work remaining in the $FIFO$ queue, [1] Q_2 (see Figure 1). These N_1^h jobs will join Q_2 before J. Let their total work in Q_2 (sum of service times) be denoted by V_2^h.

We now look at the service time completions in both queues from the instant that J arrives at Q_1. As far as Q_1 is concerned, it will remain busy at least upto the point when J departs from it. In other words, as soon as one job completes in Q_1, that job will be immediately followed by the service of another job, as shown in Figure 2. This need not be the case with Q_2 however. When J arrives, it sees that there is a total work of Y_2 in Q_2. After Y_2 amount of work is done in Q_2 there may or may not be a service of a job immediately thereafter, since that depends on the residual service time of the job currently in service in Q_1.

[1] Since both high and low priority jobs take a random look at Q_2 upon arrival, Y_2 does not have any 'h' or 'l' superscript.

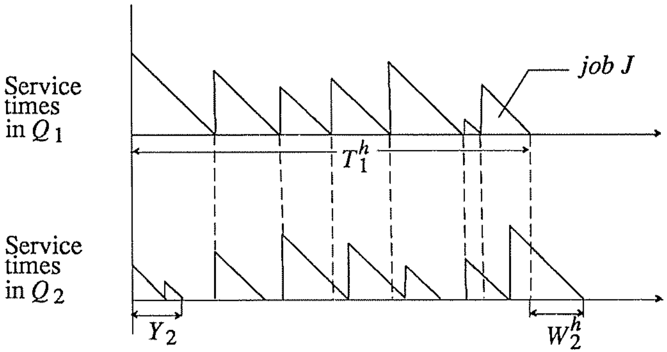

Figure 2. Q_2 idle after Y_2

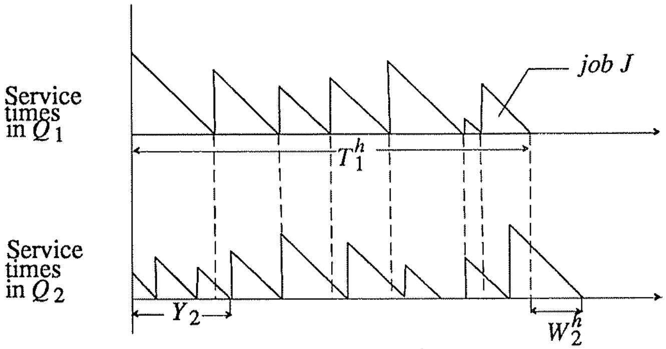

Figure 3. Q_2 busy after Y_2

Figure 2 depicts the situation where Q_2 is idle after Y_2 amount of work completes, and Figure 3 depicts the situation where it is not. Once the Y_2 amount of work completes, the $FIFO$ queue will have to serve V_2^h amount of work (note that there may be idle periods in between, as shown in the above figures). Job J can be served in the $FIFO$ queue, only after this V_2^h amount of work completes in that queue.

When J arrives to Q_2, it sees a waiting time of W_2^h. Let I_2^h denote the total idle time in Q_2 during T_1^h, that is the time that J spends in Q_1; then we can derive the following equation for T_1^h, by equating the horizontal and vertical time components in Figure 2 or Figure 3:

$$T_1^h = Y_2 + V_2^h + I_2^h - W_2^h$$

From this we get :

$$W_2^h = Y_2 + I_2^h + V_2^h - T_1^h \tag{1}$$

which we can write as

$$E[W_2^h] = E[Y_2] + E[I_2^h] + E[V_2^h] - E[T_1^h] \tag{2}$$

Thus far the only known term is $E[T_1^h]$ since Q_1 is an $M/G/1$ queue. So we try to determine the remaining unknown terms.

Let us first try to eliminate Y_2. Since the arrivals to the system follow a Poisson process, they take a random look at the system [2]. Hence, they take a random look not only at Q_1, but also at Q_2. If J sees N_2 jobs in Q_2, the moment it arrives at Q_1, then it follows from Little's Law [2] that

$$E[N_2] = \lambda E[T_2] \tag{3}$$

where $E[T_2]$ is the mean response time of a job in Q_2. We now relate $E[Y_2]$ with $E[N_2]$. From the smoothing property of *conditional expectation* [1] we know that

$$E[Y_2] = E[E[Y_2|N_2]]$$

If r is the residual service time of the job in service, and x_i is the service time of job i in Q_2 where $i = 2, \ldots, N_2$, then

$$Y_2|N_2 = r + x_2 + x_3 + \ldots + x_{N_2}$$

which yields

$$E[Y_2|N_2] = 1_{\{N_2>0\}} \frac{1 + C_{x_2}^2}{2\mu} + \frac{N_2 - 1_{\{N_2>0\}}}{\mu}$$

where $1_{\{N_2>0\}}$ is the *indicator function* [1] of at least one job being present in Q_2. Hence it follows that

$$E[Y_2] = E[E[Y_2|N_2]] = \frac{\rho(1 + {C_{x_2}}^2)}{2\mu} + \frac{E[N_2]}{\mu} - \frac{\rho}{\mu}$$

Combining the above equation with equation (3) we get the following relation between $E[Y_2]$ and $E[T_2]$

$$E[Y_2] = \rho E[T_2] + \frac{\rho({C_{x_2}}^2 - 1)}{2\mu} \qquad (4)$$

The next step would be to evaluate the term $E[V_2^h] - E[T_1^h]$ in equation (2). Customer J sees N_1^h jobs ahead of it in Q_1. V_2^h is the total work of these N_1^h jobs when they join the $FIFO$ queue (i.e. the sum of their service times in the $FIFO$ queue). If the service times of these jobs in Q_2 are s_{i2}, $i = 1, \ldots, N_1^h$, then

$$V_2^h|N_1^h = \sum_{i=1}^{N_1^h} s_{i2}$$

which yields

$$E[V_2^h] = E[E[V_2^h|N_1^h]] = E\left[\frac{N_1^h}{\mu_2}\right] = \frac{E[N_1^h]}{\mu_2}$$

Let N_{q1}^h be the high priority jobs *waiting* in Q_1 ahead of J, when J arrives to that queue. Therefore,

$$N_1^h = N_{q1}^h + 1_{\{N_1>0\}}$$

where $1_{\{N_1>0\}}$ is the indicator function of atleast one job in Q_1. Hence $E[N_1^h] = E[N_{q1}^h] + \rho_1$. Using this in the above expression for $E[V_2^h]$,

$$E[V_2^h] = \frac{E[N_{q1}^h] + \rho_1}{\mu_2}$$

We next use Little's Law for $E[N_{q1}^h]$ to obtain

$$E[N_{q1}^h] = \lambda^h E[W_1^h] = \lambda^h E[T_1^h] - \rho_1^h$$

Substituting this expression in that for $E[V_2^h]$,

$$E[V_2^h] = \rho_2^h E[T_1^h] + \frac{\rho_1^l}{\mu_2} \tag{5}$$

This expresses $E[V_2^h]$ in terms of $E[T_1^h]$. The expression for $E[T_1^h]$ is given in [3]

$$E[T_1^h] = \frac{1}{\mu_1} + \frac{\rho_1(1 + {C_{x_1}}^2)}{2\mu_1(1 - \rho_1^h)} \tag{6}$$

Now most of our equations are set up, and all we need is to substitute them appropriately and rearrange terms. Substituting equations (4), (5), (6) into equation (2),

$$E[W_2^h] = \rho_2 E[T_2] + E[I_2^h] + \frac{\rho_2(1 + {C_{x_2}}^2)}{2\mu_2} - \frac{1-\rho_2}{\mu_1} - \frac{\rho_2}{\mu_2} - \frac{\rho_1(1 + {C_{x_1}}^2)(1 - \rho_2^h)}{2\mu_1(1 - \rho_1^h)} \tag{7}$$

where $E[T_2]$ is the average response time over high and low priority classes in the $FIFO$ queue. Since $E[T_2^h] = E[W_2^h] + \frac{1}{\mu}$,

$$E[T_2^h] = \rho_2 E[T_2] + E[I_2^h] + \frac{\rho_2(1 + {C_{x_2}}^2)}{2\mu_2} - \frac{1-\rho_2}{\mu_1} + \frac{1-\rho_2}{\mu_2} - \frac{\rho_1(1 + {C_{x_1}}^2)(1 - \rho_2^h)}{2\mu_1(1 - \rho_1^h)} \tag{8}$$

A similar analysis can be conducted to derive the response time of low priority jobs in the $FIFO$ queue, $E[T_2^l]$. The main difference here is that instead of the number of jobs in front of an arriving low priority job, we also have to consider the number of jobs that pass it while it is waiting in Q_1. We thereby obtain

$$E[T_2^l] = \rho_2 E[T_2] + E[I_2^l] + \frac{\rho_2(1 + {C_{x_2}}^2)}{2\mu_2} - \frac{1-\rho_2}{\mu_1} + \frac{1-\rho_2}{\mu_2} + \frac{\rho_1(1 + {C_{x_1}}^2)}{2\mu_1(1 - \rho_1)}\left(\rho_2 - \frac{1-\rho_2^h}{1-\rho_1^h}\right) \tag{9}$$

where $E[I_2^l]$ is the mean of the idle time in the $FIFO$ queue, during the response time of a low priority job in Q_1. Since $E[T_2]$ is the average response time of high and low priority jobs in Q_2,

$$E[T_2] = \frac{\rho_2^h E[T_2^h] + \rho_2^l E[T_2^l]}{\rho_2}$$

Combining this equation with equations (8) and (9) above, $E[T_2]$ can be expressed as

$$E[T_2] = \frac{E[I_2]}{1-\rho_2} + \left\{ \frac{1}{\mu_2} + \frac{\rho_2(1+{C_{x_2}}^2)}{2\mu_2(1-\rho_2)} \right\} - \left\{ \frac{1}{\mu_1} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1)} \right\} \tag{10}$$

where $E[I_2]$ is the mean of I_2^h and I_2^l, i.e.

$$E[I_2] = \frac{\rho_2^h E[I_2^h] + \rho_2^l E[I_2^l]}{\rho_2}$$

Note that the last two terms in the expression for $E[T_2]$ correspond to the mean response time in an $M/G/1$ queue.

Using the expression for $E[T_2]$ in equations (8) and (9), we obtain

$$E[T_2^h] = \frac{\rho_2 E[I_2]}{(1-\rho_2)} + E[I_2^h] + \left\{ \frac{1}{\mu_2} + \frac{\rho_2(1+{C_{x_2}}^2)}{2\mu_2(1-\rho_2)} \right\} - \left\{ \frac{1}{\mu_1} + \frac{\rho_1\rho_2(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1)} + \frac{\rho_1(1+{C_{x_1}}^2)(1-\rho_2^h)}{2\mu_1(1-\rho_1^h)} \right\} \tag{11}$$

$$E[T_2^l] = \frac{\rho_2 E[I_2]}{(1-\rho_2)} + E[I_2^l] + \left\{ \frac{1}{\mu_2} + \frac{\rho_2(1+{C_{x_2}}^2)}{2\mu_2(1-\rho_2)} \right\} + \left\{ \frac{1}{\mu_1} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu(1-\rho_1)} \left(\frac{\rho_1^h - \rho_2^h}{1-\rho_1^h} \right) \right\} \tag{12}$$

To compare the mean response time of high priority jobs and low priority jobs, we take the difference of the above two equations,

$$E[T_2^l] - E[T_2^h] = E[I_2^l] - E[I_2^h] + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu(1-\rho_1)} \left(\frac{\rho_2 - \rho_1}{1-\rho_1^h} \right) \tag{13}$$

Equation (13) provides us with insight to the impact that priority in the feeder queue has on the response times in the second queue. First, since both high and low priority jobs have the same service time distribution in the feeder queue, and high priority jobs have a smaller response time there, we can conclude that $E[I_2^l] \geq E[I_2^h]$. Hence, the mean idle period in the $FIFO$ queue during the time that a job spends in the feeder queue is smaller for a high priority job than for a low priority job. Second, as long as $\rho_2 \geq \rho_1$, the last term in equation (13) is non-negative. This implies that for $\rho_2 \geq \rho_1$, $E[T_2^l] \geq E[T_2^h]$; in other words high priority jobs gain over low priority jobs in the second queue even though there is no explicit priority in Q_2! When $\rho_2 < \rho_1$ the last term in equation (13) is negative. Since $E[I_2^l]$ and $E[I_2^h]$ are unknown, it is not known whether high priority jobs gain over low priority jobs when $\rho_2 < \rho_1$. We intuitively expect them to gain even in this case, because by decreasing ρ_2 for a given λ, the arrival process to the $FIFO$ queue does not change. Just changing the mean of the service time distribution in the $FIFO$ queue, (i.e. decreasing ρ_2) should therefore affect only the quantitative behavior and not the qualitative behavior of the system. We have seen that high priority jobs can benefit over low priority jobs in Q_2, w.r.t. response times, even though Q_2 is $FIFO$. The next question to address is whether introducing priority in the feeder queue causes the high priority jobs to benefit over the case where the feeder is a $FIFO$ queue. For this purpose we now move on to examine whether response times are conserved in Q_2.

3.1.1 Conservation of response times in Q_2

In this section we show that having non-preemptive priority (NP) in the feeder queue does not affect the average response time in Q_2 over high and low priority classes as compared to the case where the feeder is a $FIFO$ server. Consider a *priority-blind* gremlin observing the NP queue feeding the $FIFO$ queue. Since the gremlin is priority blind, to him the entire system behaves like two $FIFO$ queues in tandem (the departure process of jobs from the NP queue appears the same to him as if the queue were a $FIFO$ queue). Therefore, $E[Y_2]$ seen by him at arrival instants to Q_1, will be the same as in the $FIFO$ case. Since equation (4) holds for the $FIFO$ case too, it implies that if $E[Y_2]$ is the same for $NP \rightarrow FIFO$ and $FIFO \rightarrow FIFO$ cases, then so should $E[T_2]$. Therefore we have the following conservation of response times in Q_2 :

$$\frac{\rho_2^h E[T_2^h] + \rho_2^l E[T_2^l]}{\rho_2} = E[T_2]|_{FIFO \rightarrow FIFO} \qquad (14)$$

We showed earlier that when $\rho_2 > \rho_1$, high priority jobs have a smaller response time in Q_2 than low priority jobs. Hence by the conservation property, whenever $\rho_2 > \rho_1$, high priority jobs in the $NP \rightarrow FIFO$ case perform better in Q_2 than in the $FIFO \rightarrow FIFO$ case. From our results, however, the existence of a similar benefit when $\rho_2 < \rho_1$ cannot be derived conclusively.

3.2 $PR \rightarrow FIFO$

In the previous subsection response time equations in Q_2 were derived for the case where there was non-preemptive priority in the feeder queue. Since these equations were in terms of the first two moments of the service times and the idle times in Q_2, a natural question to ask is whether this result extends to the preemptive-case as well. In this section we show that this is indeed the case. To obtain equations for the response times in Q_2 in the preemptive case, the procedure is similar to the non-preemptive case. As far as high priority jobs are concerned, the only difference from the non-preemptive case is that they see only high priority jobs in front of them, *including* the one in service. Keeping this in mind, the rest of the analysis proceeds just like in section 3.1. Similarly for the low priority jobs, the high priority jobs that preempt them will also have to be considered. We thereby state the following response time equations without proof :

$$E[T_2^h] \quad = \quad \rho_2 E[T_2] + E[I_2^h] + \left\{ \frac{1-\rho_2}{\mu_2} + \frac{\rho_2(1+{C_{x_1}}^2)}{2\mu_2} \right\} - \left\{ \frac{1}{\mu_1} + \frac{\rho_1^h(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1^h)} \right\} (1-\rho_2^h) \qquad (15)$$

and for low priority jobs

$$\begin{aligned} E[T_2^l] \quad = \quad & \rho_2 E[T_2] + E[I_2^l] + \left\{ \frac{1-\rho_2}{\mu_2} + \frac{\rho_2(1+{C_{x_1}}^2)}{2\mu_2} \right\} - \{ \frac{1}{\mu_1} + \frac{\rho_1^h(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1^h)} \} (1-\rho_2^h) \\ & -\rho_1^h \frac{1-\rho_2^h}{\mu_1} - \frac{\rho_1^l(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1^h)} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1)(1-\rho_1^h)}(\rho_2-\rho_1) \qquad (16) \end{aligned}$$

where as before $E[T_2]$ is the average response time in Q_2 over both classes. In the above two equations we note that the first few terms are almost identical except for the terms $E[I_2^h]$ and $E[I_2^l]$. Just as in the non-preemptive case, we can argue that $E[I_2^h] < E[I_2^l]$. Therefore, over a certain range of ρ_1 and ρ_2, as specified by the last three terms of equation (16), we can conclusively state that high priority jobs gain over low priority jobs. Although from the above equations we cannot infer that this result holds for all values of ρ_1 and ρ_2, we intuitively expect this to be the case because varying ρ_2 w.r.t. ρ_1 does not change the arrival process to the $FIFO$ queue, but merely changes the mean of the service time distribution in Q_2. Hence, just the quantitative behavior should be affected and not the qualititative behavior of the system. Unlike the $NP \rightarrow FIFO$ case, conservation does not hold here, since the arrival process to Q_2 will not be the same as the $FIFO \rightarrow FIFO$ case. The variance of the arrival process, however, will be reduced on account of preemption in the feeder (when $C_{x_2} > 1$), thereby reducing the average response time in Q_2, as shown in [5].

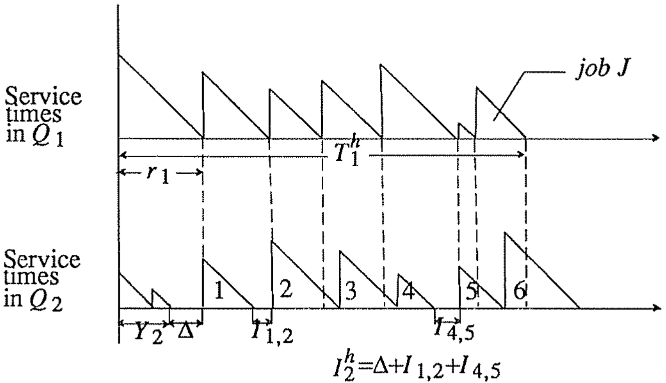

Figure 4. Idle time in Q_2 when $\Delta>0$

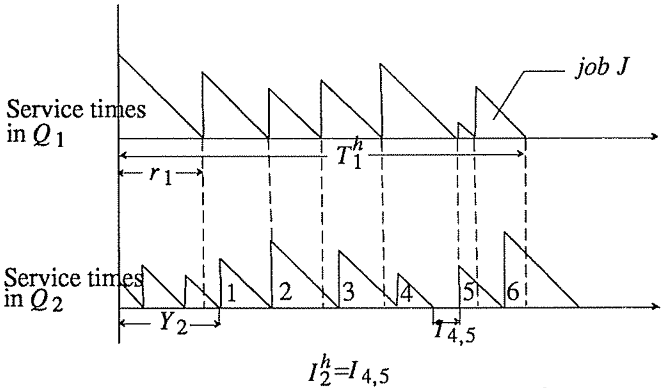

Figure 5. Idle time in Q_2 when $\Delta=0$

4 Response time bounds

In the previous section response time equations were presented for high and low priority classes. These equations contained idle times in Q_2 as unknown factors. In this section we derive upper bounds for these idle times in the case where the feeder queue has non-preemptive priority (NP). These upper bounds can then be used to obtain upper bounds for response times in the $FIFO$ queue.

Let us return to the $NP \rightarrow FIFO$ case. Focussing our attention on equation (8), we notice $E[I_2]$ and $E[I_2^h]$ as unknown factors. The derivation of upper bounds for each of these follows a similar approach; thus we present only the derivation for an upper bound for $E[I_2^h]$. Consider the instant at which a high priority job, J, arrives to Q_1. Let r_1 be the residual service time of the job in service in Q_1, at that instant of time. We compare r_1 with the unfinished work in Q_2, Y_2, to see if there is any idle time in Q_2 during r_1. Let

$$\Delta = (r_1 - Y_2)^+ = max\{0,\ r_1 - Y_2\}$$

Note that by definition, $\Delta \geq 0$. Figures 4 and 5 depict situations in which $\Delta > 0$ and $\Delta = 0$ respectively. Let $I_{i,i+1}$ be the idle time between the service of jobs i and $i+1$, in the $FIFO$ queue. Since job J sees N_1^h jobs ahead of it in the priority queue, these N_1^h jobs will also be served before J in the $FIFO$ queue. Therefore as can be seen from Figure 4,

$$I_2^h | N_1^h = \Delta \ + \sum_{i=1}^{N_1^h} I_{i,i+1}$$

From [4] we find that $E[I_{i,i+1}] = (1 - \rho_2)/\lambda$. Using this in the above equation

$$E[I_2^h|N_1^h] = E[\Delta] + N_1^h \frac{1-\rho_2}{\lambda} \tag{17}$$

Since Q_1 is an $M/G/1$ queue with non-preemptive priority,

$$E[N_1^h] = \rho_1 + \frac{\rho_1^h \rho_1 (1 + {C_{x_1}}^2)}{2(1-\rho_1^h)}$$

Therefore,

$$E[I_2^h] = E[E[I_2^h|N_1^h]] = E[\Delta] + \frac{1-\rho_2}{\mu_1} + \frac{\rho_1^h (1 + {C_{x_1}}^2)(1-\rho_2)}{2\mu_1(1-\rho_1^h)} \tag{18}$$

As $\Delta \leq r_1$ by definition,

$$E[\Delta] \leq E[r_1] = \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1}$$

which can be used in equation (18) to yield an upper bound for $E[I_2^h]$

$$E[I_2^h] \leq \frac{1-\rho_2}{\mu_1} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1} \frac{\rho_1^h(1+{C_{x_1}}^2)(1-\rho_2)}{2\mu_1(1-\rho_1^h)} \tag{19}$$

Now that we have derived an upper bound for $E[I_2^h]$ we are still left with the task of deriving an upper bound for $E[I_2]$. From equations (10), (11), and (12) in section 3.1, and by using the conservation property for the response times in Q_2, it can be shown that

$$E[I_2]|_{FIFO \rightarrow FIFO} = E[I_2]|_{NP \rightarrow FIFO}$$

The derivation of an upper bound for $E[I_2]$ in the $FIFO \rightarrow FIFO$ case proceeds in the same fashion as the derivation of the upper bound for $E[I_2^h]$, thus yielding,

$$E[I_2] \leq \frac{1-\rho_2}{2\mu_1} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1} \left(1 + \frac{1-\rho_2}{1-\rho_1}\right) \tag{20}$$

This bound for $E[I_2]$ yields an upper bound for the response time in Q_2 for the $FIFO \rightarrow FIFO$ case (see equation (10)). In combination with the lower bound for $E[T_2]$ in the $FIFO \rightarrow FIFO$ case given in [5],

$$\frac{1}{2\mu_2} + \frac{\rho_2(1+{C_{x_2}}^2)}{2\mu_2(1-\rho_2)} \leq E[T_2] \leq \frac{1}{\mu_2} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_2)} + \frac{\rho_2(1+{C_{x_2}}^2)}{2\mu_2(1-\rho_2)} \tag{21}$$

We are now in a position to present an upper bound for $E[T_2^h]$. Substituting the two bounds for $E[I_2^h]$ and $E[I_2]$ given by equations (19) and (20), in the expression for $E[T_2^h]$ (equation (8), section 3.1), we obtain

$$E[T_2^h] \leq \frac{1}{\mu_2} + \frac{\rho_2(1+{C_{x_2}}^2)}{2\mu_2(1-\rho_2)} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_2)} - \frac{\rho_1^l(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1^h)} \tag{22}$$

Observe that this bound differs from the upper bound for $E[T_2]$ (equation (21)) in only the last term.

Proceeding in a similar fashion for low priority jobs, we obtain an upper bound for $E[T_2^l]$. The main difference in the derivation of this bound is that we must take into account the number of high priority jobs that pass a low priority job, while it is waiting in Q_1. Thus we state without proof,

$$E[T_2^l] \leq \frac{1}{\mu_2} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1)} + \frac{\rho_1(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_2)} + \frac{\rho_1^h(1+{C_{x_1}}^2)}{2\mu_1(1-\rho_1^h)} \tag{23}$$

Comparing this bound with that for high priority jobs given in equation (22), we observe that the only difference is a sign change in the last term weighted by the utilization of the other class. This shows us how sensitive is the response time in Q_2 to C_{x_1}. Due to conservation of response times, it does not come as a surprise to see that the average of the bounds for the high and low priority jobs is the same as the upper bound for the response times in Q_2 for the $FIFO \rightarrow FIFO$ case (equation (21)).

To see what can be analyzed from these bounds, we first look at the bounds for $E[T_2]$ (equation (21)). As C_{x_2} increases w.r.t. C_{x_1} the lower bound approaches the upper bound. In other words the mean response time in Q_2 over both classes approaches that of an $M/G/1$ queue in isolation. At the same time the upper bounds for $E[T_2^h]$ and $E[T_2^l]$ also reflect the same behavior. This results in two observations : first, priority in the feeder queue does not have much of an impact on response times in Q_2 when $C_{x_2} \gg C_{x_1}$. Second, the response time of both high and low priority classes in Q_2 can be taken as the mean response time of a $M/G/1$ $FIFO$ queue in isolation, when $C_{x_2} \gg C_{x_1}$, because as stated above the average response time over both classes approaches that of a $M/G/1$ $FIFO$ queue, and the upper bound for each class also approaches the mean response time of a $M/G/1$ $FIFO$.

5 Conclusion

This paper examined how priority in a feeder queue, Q_1, affects response times in a subsequent $FIFO$ queue, Q_2. Equations for the response time of both high and low priority jobs in Q_2 were derived for the non-preemptive case. From these equations it could be conclusively shown that when $\rho_2 \geq \rho_1$, high priority jobs gain over low priority jobs in the $FIFO$ queue. In the case when $\rho_1 = \rho_2$, the difference in the response times for both classes, is the amount of time that Q_2 goes idle during the time spent by a job of a corresponding class in Q_1. Thus, the high priority class can benefit in Q_2, even without explicitly having priority there. We expect this result to hold even when $\rho_1 > \rho_2$, but could not derive it from the equations for the response time in the $FIFO$ queue. We next showed that conservation of response times holds in Q_2 under non-preemptive priority in the first queue, just like in a single queue. This helped in obtaining an expression for the response time in Q_2 for the $FIFO \rightarrow FIFO$ case. All these expressions contained idle times in Q_2 as unknown factors. After deriving these results for non-preemptive priority in the feeder queue, preemptive priority in the feeder queue was considered, and similar equations were presented. Conservation of response times, however, does not hold in the $PR \rightarrow FIFO$ case just as it does not hold in a single queue with preemptive priority.

Since these equations contained terms involving idle times in Q_2 which could not be expressed further, upper bounds were derived for the response times in Q_2 for both high and low priority classes in the $NP \rightarrow FIFO$ case. As a by-product, bounds for the $FIFO \rightarrow FIFO$ case were also obtained, but no bounds could be derived for the $PR \rightarrow FIFO$ case. The average of the upper bounds for high and low priority classes matched the upper bound for the $FIFO \rightarrow FIFO$ case. These bounds helped us gain insight into the performance of the $FIFO$ queue fed by a NP feeder, when $C_{x_2} \gg C_{x_1}$.

Acknowledgements

We are very grateful to Thanassis Tsiolis, Mary Vernon, and Rajeev Agrawal for the very useful discussions on this work. We also thank Randy Nelson for providing us with an excellent reference in this area, and V. Srinivasan and Mary Berthold for their valuable comments.

References

[1] Grimmett G., and Stirzaker D., *Probability and Random Processes*, Oxford University Press 1989.

[2] Kleinrock L., *Queueing Systems, Volume I : Theory*, Wiley 1975.

[3] Kleinrock L., *Queueing Systems, Volume II : Computer Applications*, Wiley 1976.

[4] Niu S. C., *Bounds For The Expected Delays In Some Tandem Queues,* Journal Of Applied Probability 17,831-838 (1980).

[5] Niu S. C., *On The Comparison Of Waiting Times In Tandem Queues,* Journal Of Applied Probability 18,707-714 (1981).