MULTILEVEL PROCESSOR-SHARING QUEUEING MODELS
FOR TIME-SHARED SYSTEMS

L. Kleinrock and R. R. Muntz
University of California
Los Angeles, California USA

ABSTRACT

Scheduling algorithms for time-shared computing facilities are considered in terms of a queueing theory model. The extremely useful limit of "processor-sharing" is adopted, wherein the quantum of service shrinks to zero; this approach greatly simplifies the problem. A class of algorithms is studied for which the scheduling discipline may change for a given job as a function of the amount of service received by that job. These multilevel disciplines form a natural extension to many of the disciplines previously considered.

Solved for is the average response time for jobs conditioned on their service requirement. Explicit solutions are given for the system M/G/1 in which levels may be first-come-first-served (FCFS) or feedback (FB) in any order; in addition, the round-robin (RR) may be used at the first level. An integral equation is developed which defines (but does not as yet provide a solution for) the RR case at arbitrary level. The special case in which RR is used at the last level is solved under the condition that the service time behave exponentially for this last level.

Examples are described for which the average response time is plotted. These examples display the great versatility of the results and demonstrate the flexibility available for the intelligent design of discriminatory treatment among jobs (in favor of short jobs and against long jobs) in time-shared system design.

1. INTRODUCTION

Queueing models have been used successfully in the analysis of time-shared computer systems since the appearance of the first applied paper in this field in 1964 [1]. The recent survey [2] of such analytical work attests to this fact. One of the first papers to consider the effect of feedback in queueing systems was due to Takacs [3].

Generally, an arrival enters the time-shared system and competes for the attention of a single processing unit. This arrival is forced to wait in a system of queues until he is permitted a quantum of service time; when this quantum expires, he is then required to join the system of queues to await his second quantum, etc. The precise model for the system is developed in Section 2. In 1967 the notion of allowing the quantum to shrink to zero was first studied [4] and is referred to as "processor-sharing." As the name implies, this zero-quantum limit provides a share or portion of the processing unit to many customers simultaneously; in the case of round-robin (RR) scheduling [4], all customers in the system simultaneously share (equally or on a priority basis) the processor, whereas in the feedback (FB) scheduling [5] only a set of customers with the smallest attained service share the processor. We use the term processor-sharing here since it is the processing unit itself (the central processing unit of the computer) which is being shared among the set of the customers; the phrase "time-sharing" will be reserved to imply that customers are waiting sequentially for their turn to use the entire processor for a finite quantum. In studying the literature one finds that the obtained results appear in a rather complex form and this complexity arises from the fact that the quantum is typically assumed to be finite as opposed to infinitesimal. When one allows the quantum to shrink to zero, giving rise to a processor-sharing system, then the difficulty in analysis as well as in the form of results disappears in large part; one is thus encouraged to consider only the processor-sharing case. Clearly, this limit of infinitesimal quantum is an ideal and can seldom be reached in practice due to considerations of swap time; nevertheless, its extreme simplicity in analysis and results brings us to the studies reported in this paper.

The two processor-sharing systems studied in the past are the RR and the FB [4,5]. Typically, the quantity solved for is the expected response time conditioned on the customer's service time; response time is the elapsed time from when a customer enters the system until he leaves completely serviced. This measure is especially important since it exposes the preferential treatment given to short jobs at the expense of the long jobs. Clearly, this discrimination is purposeful since it is the desire in time-shared systems that small requests should be allowed to pass through the system quickly. In 1969 the distribution for the response time in the RR system was found [6]. In this paper we consider the case of mixed scheduling algorithms whereby customers are treated according to the RR algorithm, the FB algorithm, or first come first served (FCFS) algorithm, depending upon how much total service time they have already received. Thus, as a customer proceeds through the system obtaining service at various rates he is treated according to different disciplines; the policy which is applied among customers in different levels is that of the FB system as explained further in Section 2. This natural generalization of the previously studied processor-sharing systems allows one to create a large number of new and interesting disciplines whose solutions we present.

*This work was supported by the Advanced Research Projects Agency of the Department of Defense (DARPA-15-69-C-0285).
2. THE MODEL

The model we choose to use in representing the scheduling algorithms is drawn from queuing theory. This corresponds to the many previous models studied \([1, 2, 4, 5, 6, 7, 8]\), all of which may be thought of in terms of the structure shown in Fig. 2.1. In this figure we indicate that new requests enter the system of queues upon arrival. Whenever the computer’s central processing unit (CPU) becomes free, some customer is admitted into the service facility for an amount of time referred to as a quantum. If during this quantum, the total accumulated service for a customer equals his required service time, then he departs the system; if not, at the end of his quantum, he cycles back to the system of queues and waits until he is next chosen for additional service. The system of queues may order the customers according to a variety of different criteria in order to select the next customer to receive a quantum. In this paper, we assume that the only measure used in evaluating this criterion is the amount of attained service (total service so far received).

![Diagram of feedback queuing model](image)

**Figure 2.1. The Feedback Queuing Model**

In order to specify the scheduling algorithm in terms of this model, it is required that we identify the following:

a. The customer interarrival time distribution. We assume this to be exponential, i.e.,

\[ P([\text{interarrival time} \leq t]) = 1 - e^{-\lambda t} \quad t \geq 0 \quad (2.1) \]

where \( \lambda \) is the average arrival rate of customers.

b. The distribution of required service time in the CPU. This we assume to be arbitrary (but independent of the interarrival times). We thus assume

\[ P([\text{service time} \leq x]) = B(x) \quad (2.2) \]

Also assume \( 1/\mu = \text{average service time} \).

c. The quantum size. Here we assume a processor-shared model in which customers receive an equal but vanishingly small amount of service each time they are allowed into service. For more discussion of such systems, see \([4, 6, 7]\).

d. The system of queues. We consider here a generalization and consolidation of many systems studied in the past. In particular, we define a set of attained service times \( a_i \) such that

\[ 0 = a_0 < a_1 < a_2 < \ldots < a_N < a_{N+1} = \infty \quad (2.3) \]

The discipline followed for a job when it has attained service, \( \tau \), in the interval

\[ a_{i-1} \leq \tau < a_i \quad i = 1, 2, \ldots, N + 1 \quad (2.4) \]

will be denoted as \( D_i \). We consider \( D_i \) for any given level \( i \) to be either: FIRST COME FIRST SERVED (FCFS); PROCESSOR SHARED-FIFO (FB); or ROUND-ROBIN PROCESSOR SHARED (RR). The FCFS system needs no explanation. The FB system gives service next to that customer who so far has least attained service; if there is a tie (among \( K \) customers, say) for this position, then all \( K \) members in the tie get served simultaneously (each attaining useful service at a rate of \( 1/K \) sec/sec), this being the nature of processor sharing systems. The RR processor sharing system shares the service facility among all customers present (say \( J \) customers)

giving attained service to each at a rate of \( 1/J \) sec/sec.

Moreover, between intervals, the jobs are treated as a set of FB disciplines. See Fig. 2.2. For example, for \( N = 0 \), we have the usual single-level case of either FCFS, RR or FB.

![Diagram of attained service intervals with disciplines](image)

**Figure 2.2. Intervals of Attained Service, with Disciplines, \( D_i \)**

For \( N = 1 \), we could have any of nine disciplines (FCFS followed by FCFS, ..., RR followed by RR); note that FB followed by FB is just a single FB system (due to the overall FB policy between levels).

As an illustrative example, consider the \( N = 2 \) case shown in Fig. 2.3. Any new arrivals begin to share the processor in a RR fashion with all other customers who so far have less than 2 seconds of attained service.

Customers in the range of \( 2 < \tau < 6 \) may get served only if no customers present have had less than 2 seconds of service; in such a case, that customer (or customers) with the least attained service will proceed to occupy the service in an FB fashion until they either leave, or reach \( \tau = 6 \), or some new customer arrives (in which case the overall FB rule provides that the RR policy at level 1 preempts). If all customers have \( \tau > 6 \), then the "oldest" customer will be served to completion unless interrupted by a new arrival. The history of some customers in this example system is shown in Fig. 2.4. We denote customer \( n \) by \( C_n \). Note that the slope of attained service varies as the number of customers simultaneously being served changes. We see that \( C_6 \) required 5 seconds of service and spent 14 seconds in system (i.e., response time of 14 seconds).

![Diagram of customer history](image)

**Figure 2.4. History of Customers in Example**

\[ \text{341/2} \]
So much for the system specification. We may summarize by saying that we have an M/G/1 queueing system\* model with processor sharing and with a generalized multilevel scheduling structure.

The quantity we wish to solve for is

\[ T(t) = E[\text{response time for a customer requiring a total of } t \text{ seconds of attended service}] \]  

(2.5)

We further make the following definitions:

\[ T_i(t) = E[\text{time spent in interval } i \text{ (} a_{i-1}, a_i \text{) for customers requiring a total of } t \text{ seconds of attended service}] \]  

(2.6)

We note that

\[ T_i(t) = T_i(t') \text{ for } t, t' > a_i \]  

(2.7)

Furthermore, we have, for \( a_{k-1} < t < a_k \), that

\[ T(t) = \sum_{i=1}^{k} T_i(t) \]  

(2.8)

Also, we find it convenient to define the following quantities with respect to \( B(t) \) truncated at \( t = x \):

\[ E_x = \int_{0}^{x} dB(t) + x \int_{x}^{\infty} dB(t) \]  

(2.9)

\[ T_x = \int_{0}^{x} t^2 dB(t) + x \int_{x}^{\infty} t^2 dB(t) \]  

(2.10)

\[ \rho_x = \frac{\lambda E_x}{\mu CT_x} \]  

(2.11)

\[ W_x = \frac{T_x}{T(t)} \]  

(2.12)

Note that \( W_x \) represents the expected work found by a new arrival in the system M/G/1 where the service time distribution is \( B(t) \) truncated at \( x \).

3. RESULTS FOR MULTILEVEL QUEUING SYSTEMS

We wish to find an expression for \( T(t) \), the mean system time (i.e., average response time) for jobs with service time \( t \) such that \( a_{i-1} < t < a_i \), i.e., jobs which reach the \( i \)th level queue and then leave the system. To accomplish this it is convenient to isolate the \( i \)th level to some extent. We make use of the following two facts.

1. By the assumption of preemptive priority of lower level queues (i.e., FB discipline between levels) it is clear that jobs in levels higher than the \( i \)th level can be ignored. This follows since these jobs cannot interfere with the servicing of the lower levels.

2. We are interested in jobs that will reach the \( i \)th level queue and then depart from the system before passing to the \( (i+1) \)st level. The system time of such a job can be thought of as occurring in two parts. The first portion is the time from the job's arrival to the queueing system until the group at the \( i \)th level is served for the first time after this job has reached the \( i \)th level. The second portion starts with

\*M/G/1 denotes the single-server queueing system with Poisson arrivals and arbitrary service time distribution.
in virtual time to be bulk arrivals at instants generated from a Poisson process with parameter λ.

Consider a job that requires \( t = a_{i-1} + \tau \) seconds of service \((0 < \tau < a_i - a_{i-1})\). Let \( a_1 \) be the mean real time the job spends in the system until its arrival (at the end of the lower level busy period) at the first level. Let \( a_2(t) \) be the mean virtual time the job spends in the higher level queue.

\( a_1 \) can be calculated using delay cycle analysis. The initial delay is equal to the mean work the job finds in the lower level on arrival plus the \( a_{i-1} \) seconds of work that it contributes to the lower level. This initial delay is expanded by new jobs arriving at the lower level by a factor of \( 1/(1 - \rho_{a_{i-1}}) \) (see [9]). Therefore

\[
a_1 = \frac{1}{1 - \rho_{a_{i-1}}} \left\{ w_{a_{i-1}} + a_{i-1} \right\} \tag{3.3}
\]

If \( a_2(t) \) is the mean virtual time the job spends in the higher level, we can easily convert this to the mean real time spent in this level. In the virtual time interval \( a_2(t) \) there are an average of \( \lambda a_2(t) \) lower level busy periods that have been ignored. Each of these has a mean length of \( \bar{T}_{a_2(t)} \cdot \frac{1}{1 - \rho_{a_{i-1}}} \).

Therefore, the mean real time the job spends in the higher level is given by

\[
a_2(t) + \lambda a_2(t) \frac{1}{1 - \rho_{a_{i-1}}} = \frac{a_2(t)}{1 - \rho_{a_{i-1}}} \tag{3.4}
\]

Combining these results we see that a job requiring \( t = a_{i-1} + \tau \) seconds of service has mean system time given by

\[
T(a_{i-1} + \tau) = \frac{1}{1 - \rho_{a_{i-1}}} \left\{ w_{a_{i-1}} + a_{i-1} + a_2(t) \right\} \tag{3.5}
\]

The only unknown quantity in this equation is \( a_2(t) \). To solve for \( a_2(t) \) we must, in general, consider an \( M/G/1 \) system with bulk arrivals and RR processor sharing. The only exception is the case of RR at the first level which has only single arrivals. Since the higher level queues can be ignored, the solution in this exceptional case is the same as for a round-robin single level system with service time distribution truncated at \( a_1 \). Therefore, from [8] we have for the first level

\[
T(t) = \frac{t}{1 - \rho_{a_{i-1}}} \quad 0 < t < a_1 \tag{3.6}
\]

Let us now consider the bulk arrival RR system in isolation in order to solve for the virtual time spent in the higher level queue, \( a_2(t) \) which we temporarily write as \( a(t) \).

3.3.1. The Bulk Arrival, RR Model

We approach this problem by first considering a discrete time system with quantum size \( q > 0 \). We assume that arrivals and departures take place only at times that are integral multiples of \( q \). For small \( q \) any continuous distribution can be approximated. By letting \( q \) approach zero equations for continuous time systems can be found.

Let \( n(q) \) be the mean number of jobs in the system with \( i q \) seconds of attained service.

\[
a'(t) = \int_0^{t} n(x) \frac{1 - B(x + t)}{1 - B(x)} dx + \lambda a \int_0^{t} a'(x) \left[ 1 - B(t - x) \right] dx + b(1 - B(t)) \tag{3.19}
\]
From Kleinrock and Coffman [7] we also have
\[ n(x) = \lambda a [1 - B(x)] a'(x) \]  
(3.20)

Substituting for \( n(x) \) we have
\[ a'(t) = \lambda a \int_0^t a'(x) [1 - B(x + t)] dx \]
\[ + \lambda a \int_0^t a'(x) [1 - B(t - x)] dx \]
\[ + 1 + b [1 - B(t)] \]  
(3.21)

This integral equation defines \( a'(t) \) for the case of bulk arrival to a RR processor-sharing M/G/1 system. Unfortunately no general solution has been found for this equation in terms of \( B(t) \). However, for exponential service time the equation can be solved.

3.3.1a. M/M/1 with Bulk Arrival. In this case
\[ B(t) = 1 - e^{-\mu t} \]  
(3.22)

Therefore the Eq. (3.21) becomes
\[ a'(t) = \lambda a \int_0^t a'(x) e^{-\mu (x+t)} dx \]
\[ + \lambda a \int_0^t a'(x) e^{-\mu (t-x)} dx \]
\[ + 1 + be^{-\mu t} \]  
(3.23)

From Eq. (3.20) we obtain
\[ \lambda a \int_0^t a'(x) e^{-\mu (x+t)} dx \int_0^\infty n(x) e^{-\mu x} dx = \bar{n} \int_0^\infty e^{-\mu x} dx \]
\[ = \frac{\bar{n} \mu}{\lambda a} \]  
(3.24)

where \( \bar{n} = E[\text{no. found in system by a new arrival}] \). Using this expression for the first term on the right-hand side of Eq. (3.23) and taking Laplace transforms we obtain
\[ s a'(s) = \frac{\bar{n} (s + b + 1) + \mu}{s (s + \mu - \lambda a)} \]  
(3.25)

Inverting, we get for \( t \geq 0 \) (observing that \( a'(0) = \bar{n} + b + 1 \),
\[ a'(t) = \frac{1}{1 - \rho} + \frac{\bar{n} + b + 1}{1 - \rho} [1 - \rho] - \frac{1}{1 - \rho} e^{\mu (1 - \rho) t} \]  
(3.26)

where
\[ \rho = \frac{\lambda a}{\mu} \]  
(3.27)

Here, \( \bar{n} \) and \( b \) are unknown quantities which need not be solved for directly. Instead we combine them in a new unknown \( C = \bar{n} + b - \frac{\rho}{1 - \rho} \) and we obtain
\[ a'(t) = \frac{1}{1 - \rho} + Ce^{\mu (1 - \rho) t} \]  
(3.28)

Integrating and using the initial condition that \( a(0) = 0 \) we get
\[ a(t) = \frac{t}{1 - \rho} + \frac{C}{\mu (1 - \rho)} [1 - e^{\mu (1 - \rho) t}] \]  
(3.29)

In the next section we will evaluate the constant \( C \) and calculate \( \bar{a} \) for a multilevel queuing system with RR at the last level where the service time distribution may be general up to this level, but must be exponential in this last (semi-infinite) level; i.e., \( B(x) \) must have an exponential tail and we denote this system by M/GM/1. The same method can be used to complete the solution for a single level queue with bulk arrivals.

3.3.1b. M/GM/1 with Bulk Arrival. Returning to our discussion of the two level queuing system with breakpoints at \( a_{l-1} = a_1 \), we had Eq. (3.5)
\[ T(a_{l-1} + \tau) = \frac{1}{1 - \rho a_{l-1}} \left\{ -a_{l-1} + a_{l-1} + a_2(t) \right\} \]  
(3.5)

where \( a_2(t) \) was the mean virtual time spent in the higher level queue. But in virtual time this is the solution for the bulk arrival case just studied. The study in the general case M/G/1 led to an integral equation, Eq. (3.21), for which no more explicit solution was obtained. However, in the case of an exponential distribution, we have the solution given in Eq. (3.29). Thus, we may permit RR at the first level (see Eq. (3.6)) in M/G/1 or at the last level in M/GM/1. In the latter case, we therefore consider the equivalent two-level system in which the breakpoints \( a_{l-1} = a_1 \) are restricted to \( a_N \) and \( a_t \), respectively.

Thus, for the case \( t = a_N + \tau \) we have from Eq. (3.29) that
\[ a_2(t) = a(t) = \frac{\tau}{1 - \rho} + d[1 - e^{\mu (1 - \rho) \tau}] \]  
(3.30)

where \( C = \mu (1 - \rho) d \). Therefore, from Eq. (3.5),
\[ T(a_N + \tau) = \frac{1}{1 - \rho a_N} \left\{ a_N + a_t + \frac{\tau}{1 - \rho} + d[1 - e^{\mu (1 - \rho) \tau}] \right\} \]  
(3.31)

In addition to the constant \( d \) we also need to solve for \( \bar{a} \), the mean size of the bulk arrivals to the RR queue, since this is contained in \( \rho = \frac{\lambda a}{\mu} \). This we do for the general case \( a_{l-1} = a_1 \). \( \bar{a} \) is just the mean number of jobs that arrive during a lower level busy period and require more than \( a_{l-1} \) seconds of service. Therefore \( \bar{a} \) must satisfy the equation
\[ \bar{a} = \lambda \bar{a}_{c_{l-1}} \bar{a} + [1 - B(a_{l-1})] \]  
(3.32)

In this equation \( \lambda \bar{a}_{c_{l-1}} \) is the mean number of jobs that arrive during the service time of the first job in the busy period. Since each of these jobs in effect generates a busy period, there are an average of \( \lambda \bar{a}_{c_{l-1}} \) arrivals to the upper level queue due to these jobs. The second term is just the average number of times that the first job in the busy period will require more than \( a_{l-1} \) seconds of service.

Clearly then
\[ \bar{a} = \frac{1 - B(a_{l-1})}{1 - \rho a_{l-1}} \]  
(3.33)

Now we may complete the solution for \( t = a_N + \tau \) by solving for \( C \) by conserving the mean work in the system. Since the single server works at a constant rate as long as there is any work in the system, the amount of work in the system at any time is independent of the service disciplines and system levels. It follows immediately that the mean work in the system is a constant (depending only on \( \lambda \) and the service time distribution). The mean work in the system is given by \( \bar{w} \) (see Eq. (2.12)).

We also have from Eq. (3.20) that \( n(t) = \lambda [1 - B(t)] T^*(t) \) where \( n(t) \) is the mean number of jobs in the system with \( t \) seconds of attained service. The mean remaining service requirement for a job which has already received \( t \) seconds of service is given by
\[ \int_t^\infty \frac{\lambda x B(x)}{1 - B(t) - t} \]  
(3.34)

Therefore the mean work in the system is also given by

341/5
\[ W_m = \int_0^\infty n(t) \left( \int_t^\infty \frac{w(t)}{1 - \frac{B(t)}{1 - \frac{B(t)}{1 - \frac{B(t)}{1 - \frac{B(t)}}}}} - t \right) \, dt \]  
\[ \text{(3.35)} \]

or
\[ W_m = \int_0^\infty \lambda (1 - B(t)) T(t) \left( \int_t^\infty \frac{w(x)}{1 - \frac{B(t)}{1 - \frac{B(t)}{1 - \frac{B(t)}}}} - t \right) \, dt \]  
\[ \text{(3.36)} \]

With no loss of generality, we may assume that the RR discipline is the lower level queue discipline. In this case we have from Eqs. (3.6), (3.31) and \( W_m \) from Eq. (2.12) that
\[ T(t) = \begin{cases} \frac{t}{1 - \rho c_N} & 0 < t < a_N \\ \frac{1}{1 - \rho c_N} \left( \frac{\mu}{\lambda T_c_c} + \frac{a_N}{\lambda} \right) + \frac{d(1 - e^{-\lambda t})(t - a_N)}{\lambda} & t > a_N \end{cases} \]  
\[ \text{(3.37)} \]

Using this expression for \( T(t) \) in Eq. (3.36) we can solve for \( d \). Since \( B(x) \) is arbitrary in the range \( 0 < t < a_N \) and exponential for \( t > a_N \), the solution is not expressible in a compact form. When \( B(x) \) is exponential over \( 0 < x < a \) = the solution is simplified. In particular for all \( t > a_N \)
\[ \int_t^\infty \frac{w(x)}{1 - \frac{B(t)}{1 - \frac{B(t)}}} - t = \frac{1}{\mu} \]  
\[ \text{(3.38)} \]

Therefore
\[ W_m = \int_0^\infty \lambda (1 - B(t)) T(t) \frac{1}{\mu} \, dt = \frac{\lambda}{\mu} \int_0^\infty e^{-\lambda t} T(t) \, dt \]  
\[ \text{(3.39)} \]

Now using the Eq. (3.37) for \( T(t) \) we have
\[ W_m = \frac{1}{\mu} \left( \frac{1 - e^{-\mu a_N}}{\mu(1 - \rho c_N)} + e^{-\mu a_N} \right) \]  
\[ \frac{2(1 - \rho c_N)^2}{\mu(1 - \rho c_N)^2} \]  
\[ + \frac{\mu(1 - \rho)(1 - \rho c_N)}{\mu(1 - \rho)(1 - \rho c_N)} \]  
\[ \frac{1 - e^{-\mu a_N}}{\mu(2 - \rho)(1 - \rho c_N)} \]  
\[ \text{(3.40)} \]

Setting \( W_m = \frac{2}{\lambda T_c_c} = \frac{1}{\mu} \frac{1}{1 - \lambda / \mu} \) we essentially have a solution for \( C \). The solution is illustrated later in the examples section.

4. EXAMPLES

In this section we demonstrate through examples the nature of the results we have obtained. Recall that we have given explicit solutions for our general model in the case M/G/1 with processor sharing where the allowed scheduling disciplines within a given level may be either FCFS or FB. Moreover, for this general system we allow RR at level 1. That is, for the case M/\( \mu \lambda \),
\[ D_k = \begin{cases} \{ \text{RR, FCFS, FB} \} & i = 1 \\ \{ \text{FCFS, FB} \} & i = 2, 3, \ldots, N + 1 \end{cases} \]  
\[ \text{(4.1)} \]

Furthermore, in the case M/GM/1 we permit
\[ \text{RR, FCFS, FB} \]
\[ i = 1 \]
\[ \{ \text{FCFS, FB} \} \]
\[ i = 2, 3, \ldots, N \]
\[ \{ \text{RR, FCFS, FB} \} \]
\[ i = N + 1 \]  
\[ \text{(4.2)} \]

That is, we also permit RR at the highest level if \( B(x) \) is of exponential form in the interval \( a_N \times x \).

We begin with three examples from the system M/M/1. As mentioned in Section 2, we have nine disciplines for the case \( N = 1 \). This comes about from Eq. (4.2) where we allow any one of three disciplines at level 1 and any one of three disciplines at level 2. As we have shown, the behavior of the average conditional response time in any particular level is independent of the discipline in all other levels. In Fig. 4.1 we show the behavior of each of the three disciplines for the system \( N = 1 \).

![Figure 4.1. Response Time Possibilities for \( N = 1 \), M/M/1, \( \mu = 1 \), \( \lambda = 0.75 \), \( s_i = 2 \)](image)

In this case we have assumed \( \mu = 1 \), \( \lambda = 0.75 \), and \( a_1 = 2 \). Note that for the case M/M/1 we have from Eqs. (2.9), (2.10) and (3.33) the following:
\[ \text{Note that for the case M/M/1 we have from Eqs. (2.9), (2.10) and (3.33) the following:} \]
\[ E_{c_1} = \frac{1}{\mu} (1 - e^{-\mu a_1}) = 0.065 \]  
\[ \text{(4.3)} \]
\[ t^{-1} = \frac{2}{\mu^2} \left[ 1 - e^{-\mu a_1} - \mu a_1 e^{-\mu a_1} \right] = 1.19 \]  
\[ \text{(4.4)} \]
\[ \tilde{a} = e^{-\mu a_1} / (1 - \lambda \tilde{E}_{c_1}) = 0.385 \]  
\[ \text{(4.5)} \]

Also, for the parameter values chosen, we have \( C = 2.42 \).

From Eq. (3.1) we see that the response time for the FB system is completely independent of the values \( a_i \) and therefore the curve shown in Fig. 4.1 for this response time is applicable to all of our M/M/1 cases. Note the inflection point in this curve which results in a linear growth for response time as \( t \rightarrow \infty \) (a phenomenon not observable from previously published figures). As can be seen from its defining equation, the response time for FCFS is linear regardless of the level; the RR system at level 1 is also linear, but as we see from this figure and from Eq. (3.37) the RR at level \( N + 1 \) is non-linear. Thus one can determine the behavior of any of our nine possible disciplines from Fig. 4.1. Adiri and Ari-Itzhak considered the case (FB, RR) [12].
Continuing with the case M/M/1, we show in Fig. 4.2 the case for $N = 3$ where $D_1 = RR$, $D_2 = FB$, $D_3 = FCFS$, and $D_4 = RR$. In this case we have chosen $\alpha_1 = 1$ and $\mu = 1$, $\lambda = 0.75$. We also show in this figure the case FB over the entire range as a reference curve for comparison with this discipline. Note (in general for M/M/1) that the response time for any discipline in a given level must either coincide with the FB curve or lie above it in the early part of the interval and below it in the latter part of the interval; this is true due to the conservation law [10].

Also shown in this figure is a dashed line corresponding to the FB system over the entire range. Clearly, one may select any sequence of FB and FCFS with duplicates in adjacent intervals and the behavior for such systems can be found from Fig. 4.3. It is interesting to note in the general M/G/1 case with $D_1 = FCFS$ that we have a solution for the FB system with finite quantum $q_1 = a_1 - a_{1-1}$ where preemption within a quantum is permitted!

For our last example we choose the system M/E/1 with $N = 1$. In this system we have

$$d(x) = (2\mu)^2 x e^{-2\mu x} \quad x \geq 0$$

as shown in Fig. 4.4. We note that the mean service time here is again given by $1/\mu$; the second moment of this distribution is $3/2\mu^2$. We calculate

$$\bar{x}_{a_1} = \frac{1}{\mu} - \frac{1}{\mu} e^{-2\mu} \left(1 + 2\mu a_1 + 2(\mu a_1)^2\right)$$

We choose the system $N = 1$ with $D_1 = RR$ and $D_2 = FCFS$. For the cases $a_1 = 1/2\mu$, $1/\mu$, $2/\mu$, $4/\mu$ with $\mu = 1$ and $\lambda = 0.75$ we show in Fig. 4.5 the behavior of this system.
This figure demonstrates again the kind of behavior obtainable from our results as one varies the appropriate system parameters; once again one may choose to discriminate in a variety of ways in favor of the short jobs and against the longer jobs.

5. CONCLUSION

Our purpose has been to generalize results in the modeling and analysis of time-shared systems. The class of systems considered was the processor-sharing systems in which various disciplines were permitted at different levels of attained service. The principle results for M/G/1 are the following: (1) The average conditional response time at level i is independent of the queuing discipline at all other levels; (2) the performance for the FB discipline at any level is given by Eq. (3.1); (3) the performance for the FCFS discipline is linear with t within any level and is given by Eq. (3.2); (4) the performance for the RR discipline at the first level is well-known (8) and is given by Eq. (3.6); (5) an integral equation for the average conditional response time for RR at any level (equivalent to bulk arrival RR) is given by Eq. (3.21) and remains unsolved in general.

For M/GN/1 (exponential tail for t > a_N) we have the performance for RR at the last level as given by Eq. (3.37).

Examples are given which display the behavior of some of the possible system configurations. From these, we note the great flexibility available in these multilevel systems.

References