Queueing models of time-sharing systems

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 74-08

Queueing models of time-sharing systems

by

Hubert Keuning

Eindhoven, June 1974

Abstract

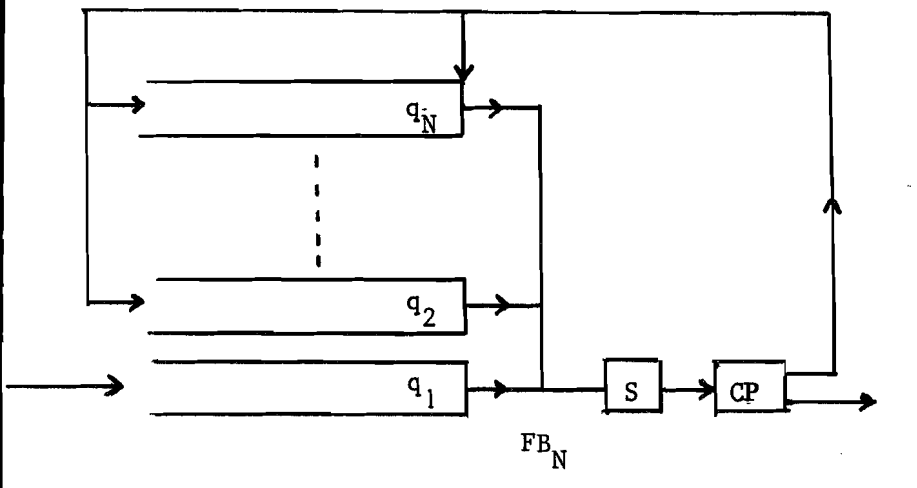
In this paper we analyse an M/G/1 queue with three types of priority rules which give preferential treatment to short jobs over long jobs. Each time a job enters the central processor it receives a quantum of servicetime. When the job is not completed during this quantum it is interrupted, allocated a lower priority and set back in queue. For one priority rule (FB_∞) there are infinitely many priorities. For the other two types of rules (RR_N and FB_N) the number of priorities is finite and they only differ in the treatment of lowest-priority jobs. For the FB_∞ and FB_N rules explicit expressions are obtained for expected delay, for the RR_N rule the expected delays are determined implicitly as a solution of a set of linear equations.

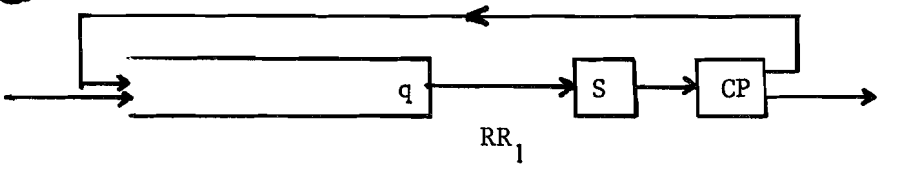
1. Introduction

In a time-sharing system with a single central processor (CP) many users are competing for service. A first-in-first-out (FIFO) treatment of the users is undesirable. The users which require a short processing time have to be serviced relatively fast and independent of system loading. The following scheduling algorithm, which requires no knowledge about the users' processing time demand, is sometimes applied:
Upon arrival in system a job joins the highest-priority queue. The jobs in that queue are FIFO treated and each job gets an amount of processing time to a maximum of q_1. If the job is not completed during this quantum of time it is interrupted and it joins a second queue with lower priority. Otherwise the job leaves the system. The jobs in the second queue are treated in the same way: they get a maximum amount of processing time q_2 (FIFO). If a job is completed during this time it leaves the system, otherwise it is interrupted and joins the next queue with lower priority. And so on. When a quantum of service time is completed the CP attends to the first job in the highest priority non-empty queue. This is called a feed-back algorithm.

The queueing model of systems of this kind consists of a number of parallel queues with a single server. Jobs in queue number one have highest priority, jobs in queue number two second highest priority, and so on. When there are infinitely many priorities we will speak of the FB_∞ priority rule. It is possible that some quantum $q_i = \infty$ and as a consequence the queues with index greater that i will always stay empty (in the case $q_1 = \infty$ we have one ordinary FIFO queue). Nevertheless we call this rule FB_∞. When there are only a finite number of priorities, say N, we will use other names. According to

the FB_N priority rule a job from the last (N-th) queue that is not completed during a quantum returns to the head of that N-th queue. The RR_N priority rule sets such a job back at the end of the last queue (round-robin). In all three disciplines a job in the i-th queue can only start service if the queues 1,...,i-1 are empty, according to the priority rule.
Furthermore we will assume that a constant swap time is needed each time a job enters the CP for service. This swap time is needed to make the data of the job available for use by the CP. The total amount of swap time allocated to jobs is the price we have to pay for using feed-back algorithms instead of the ordinary FIFO treatment.





There is an extensive litterature on this subject. Most of it considers the system under steady state conditions.

- Coffman & Kleinrock [4] analyse the model M/M/1 with RR_1, FB_N and FB_∞ and swap time zero. They also give attention to the limiting case $q \to 0$ (so-called processor shared model) and the case where not all arriving jobs join the first queue but also other lower priority queues. They only give expected delays.

- Heacox & Purdom [5] analyse the model M/M/1 with RR_1 and constant swap time. Besides they analyse the same model with the possibility to allocate a greater quantum of processing time during peakhours to reach a better response time for long jobs (which is against the design philosophy of these algorithms. Using the method of Coffman & Kleinrock they give expected delays and costs.
- For the M/G/1 model with FB_∞ and swap time zero Schrage [9] gives the Laplace-Stieltjes transform of the conditional distribution of the waiting time of a job given its processing time.
- Adiri & Avi-Itzhak determine in [1] for the M/M/1 model with RR_1 the Laplace-Stieltjes transform of the number of jobs in the system and the expected delay of a job. In [2] they give the expected delay of a job given its processing time for the model M/M/1 with RR_N. In both papers they use a positive swap time.
- Wolff [15] treats the model M/G/1 with FB_∞ and RR_1 without swap. He gives expected delays (in RR_1 a set of linear equations).
- Schassberger copies the paper of Wolff in his book [8] and he derives Laplace-Stieltjes transforms for the time dependent distributions in the cases of M/M/1 with FB_∞ and RR_1 and swap time zero.
- Using different approaches Krzesinski [6] and v.d. Weide [13] determine the expected delays in the mode M/G/1 with FB_N and constant swap time.
  Krzesinksi proceeds in the following way. At the arrival of a job requiring k quanta of processing time the queues of the model are combined into two parallel queues: a high-priority queue consisting of the first k queues and a low-priority queue consisting of the other queues (this idea has also been used by schrage, Wolff (FB_∞) and Coffman & Kleinrock).
  Van der Weide splits an arriving job in portions corresponding to the required quanta of processing time. Then he determines the expected delay for each partion.

In this paper we will analyse under steady state conditions the model M/G/1 with FB_∞, FB_N and RR_N and constant swap time. We will use v.d. Weide's method for FB_∞ and FB_N and Wolff's (RR_1) method for the last queue of RR_N (the queues $1,\ldots,N-1$ have the same behavior in FB_∞, FB_N or RR_N).

2. Definitions and Notations

Jobs arrive in the system according to a Poisson process with rate λ. The required processing times of the jobs are independent and identically distributed (i.i.d.) with distribution function F and they are independent of the input process. We assume $\int_0^\infty x\, dF(x) < \infty$. We denote by q_i the i-th quantum, i.e. the maximum amount of processing time which a job may receive when it enters the CP for the i-th time. Let $Q_0 := 0$ and $Q_k := q_1 + \ldots + q_k$ $(k \geq 1)$. If the required processing time of a job is $\underline{t}$ then we split it in $N(\underline{t})$ portions, $N(\underline{t}) := \max\{k \mid Q_{k-1} < \underline{t}\}$, i.e. $N(\underline{t}) = j$ if $Q_{j-1} < \underline{t} < Q_j$. The portions successively amount to $q_1, q_2, \ldots, q_{N(\underline{t})-1}, \underline{t} - Q_{N(\underline{t})-1}$. To every portion a constant swap time S is attached. By service time we will mean the required processing time (of a job or a portion) plus the attached swap time. So for example the service time of a job requiring $\underline{t}$ units of processing time is $\underline{t} + N(\underline{t})S$. The portions of each job receive service in the order of their indices. Portions with the same index of different jobs receive service in the order of their arrival in system.

We establish some more notations: $(i = 1,2,\ldots)$

$$r_i := 1 - F(Q_{i-1}) = \Pr\{\underline{t} > Q_{i-1}\} = \Pr\{N(\underline{t}) \geq i\} .$$

$\underline{T}_i'$:= the service time of the i-th portion of an arbitrary job,

$$\underline{T}_i' = \begin{cases} 0 & \text{if } \underline{t} \leq Q_{i-1} \text{ (prob. } 1 - r_i) \\ \underline{t} + S - Q_{i-1} & \text{if } Q_{i-1} < \underline{t} \leq Q_i \text{ (prob. } r_i - r_{i+1}) \\ q_i + S & \text{if } \underline{t} > Q_i \text{ (prob. } r_{i+1}) . \end{cases}$$

$\underline{T}_i$:= the service time of the i-th portion of an arbitrary job with $N(\underline{t}) \geq i$.

$$\mu_i := \varepsilon \underline{T}_i = \varepsilon[\underline{T}_i' \mid \underline{t} > Q_{i-1}] = \int_{Q_{i-1}}^{Q_i} (t + S - Q_{i-1}) \frac{dF(t)}{r_i} +$$

$$+ (q_i + S)\frac{r_{i+1}}{r_i} .$$

$$\mu_i^{(2)} := \varepsilon T_i^2 = \varepsilon[(T_i')^2 \mid \underline{t} > Q_{i-1}] = \int_{Q_{i-1}}^{Q_i} (t + S - Q_{i-1})^2 \frac{dF(t)}{r_i} +$$

$$+ (q_i + S)^2 \frac{r_{i+1}}{r_i} .$$

In the following we will use results of Stidham [10] and Ross [7] about regenerative processes. We will now summarize some of these results without mentioning precise definitions and conditions.

A regenerative stochastic process is characterized by the existence of random time instants (regeneration points) at which the process probabilistically restarts itself. The times between the regeneration points constitute a renewal sequence, i.e. a sequence of nonnegative, i.i.d. random variables.

1) (Stidham's Th. 1 [10]). Let $\{\underline{v}(t),\ t \geq 0\}$ be a (continuous time) regnerative process, $\{\underline{X}_n\}_{n=1}^{\infty}$ the renewal sequence of the time intervals between the regeneration points with common nonlattice distribution function G and let A be any Borel set. Then we can characterize the stationary version $\{\underline{v}^*(t),\ t \geq 0\}$ of $\{\underline{v}(t),\ t \geq 0\}$ by

$$\Pr\{\underline{v}^*(t) \in A\} = \frac{1}{\varepsilon \underline{X}_1} \int_0^{\infty} \Pr\{\underline{v}(t + x) \in A \mid \underline{X}_1 > x\}(1 - G(x))dx ,$$

for all $t \geq 0$ and this distribution function is independent of t. Moreover

(i) $$\varepsilon[\underline{v}^*(0)] = \frac{1}{\varepsilon \underline{X}_1} \varepsilon\left[\int_0^{\underline{X}_1} \underline{v}(s)ds\right] ,$$

(ii) $$\frac{1}{t} \int_0^t \underline{v}(s)ds \to \frac{1}{\varepsilon \underline{X}_1} \varepsilon\left[\int_0^{\underline{X}_1} \underline{v}(s)ds\right]$$

with probability 1 and in expectation as $t \to \infty$,

(iii) $$\lim_{t \to \infty} \Pr\{\underline{v}(t) \in A\} = \Pr\{\underline{v}^*(0) \in A\} .$$

2) (Stidham's Th. 2 [10]). Let $\{\underline{v}_n, n = 0,1,2,\ldots\}$ be a (discrete time) regenerative process, $\{\underline{N}_j\}_{j=1}^{\infty}$ the integer-valued renewal sequence of the time intervals between the regeneration points with nonperiodic probability distribution. Now we characterize the stationary version $\{\underline{v}_n^*, n = 0,1,\ldots\}$ of $\{\underline{v}_n, n = 0,1,\ldots\}$ by

$$\Pr\{\underline{v}_n^* \in A\} = \frac{1}{\varepsilon \underline{N}_1} \sum_{j=0}^{\infty} \Pr\{\underline{v}_{n+j} \in A \mid \underline{N}_1 > j\} \Pr\{\underline{N}_1 > j\}$$

for $n = 0,1,\ldots$ and this distribution function is independent of n. Moreover

(i) $$\varepsilon[\underline{v}_0^*] = \frac{1}{\varepsilon \underline{N}_1} \varepsilon\Big[\sum_{n=0}^{\underline{N}_1 - 1} \underline{v}_n\Big] ,$$

(ii) $$\frac{1}{n} \sum_{i=0}^{n-1} \underline{v}_i \to \frac{1}{\varepsilon \underline{N}_1} \varepsilon\Big[\sum_{i=0}^{\underline{N}_1 - 1} \underline{v}_i\Big] ,$$

with probability 1 and in expectation as $n \to \infty$,

(iii) $$\lim_{n\to\infty} \Pr\{\underline{v}_n \in A\} = \Pr\{\underline{v}_0^* \in A\} .$$

3) (Stidham's Th. 3 [10]). Let $\{\underline{N}(t), t \geq 0\}$ be a Poisson arrival process for the regenerative process $\{\underline{v}(t), t \geq 0\}$. Let $\{\underline{v}_n, n = 0,1,\ldots\}$ be the imbedded regenerative process at the arrival points of $\{\underline{v}(t), t \geq 0\}$ then

(i) $$\Pr\{\underline{v}_0^* \in A\} = \Pr\{\underline{v}^*(0) \in A\} ,$$

(ii) $$\lim_{n\to\infty} \Pr\{\underline{v}_n \in A\} = \lim_{t\to\infty} \Pr\{\underline{v}(t) \in A\} .$$

The arrivals may be looked upon as exogeneous events which in some way influence the future behavior of $\{\underline{v}(t), t \geq 0\}$. As an example one may consider a queueing system, in which case $\underline{N}(t)$ is the arrival process of customers in the system and $\underline{v}(t)$ might be the work in system or the number of customers in the system. In the last case one may see, provided the limits and expectations exist, the following.

4) $$\int_0^{\underline{X}_1} \underline{v}(s)\,ds = \sum_{i=1}^{\underline{N}_1} \underline{w}_i$$

where $\underline{v}(t)$ is the number of customers in the system at time t and $\underline{w}_i$ the amount of time that the i-th customer spends in the system (waiting time). Define

$$L := \lim_{t\to\infty} \frac{1}{t}\varepsilon\left[\int_0^t \underline{v}(s)\,ds\right]$$

$$w := \lim_{n\to\infty} \frac{1}{n}\varepsilon\left[\sum_{i=1}^{n} \underline{w}_i\right] .$$

Then we can derive the important relation

$$L = \lambda w ,$$

where λ is the arrival rate of customers in the system.

Since we consider an M/G/1 queueing system (Poisson arrival process and general service time distribution) we will frequently use 3) which enables us to look upon the system as stationary on arrival moments. For example in a FIFO M/G/1 queue under stationary conditions any time a customer arrives in the system the (limiting) expected number of customers in system is λw and the (limiting) expected actual delay is equal to the limiting expected virtual delay.

In the following we will mean by limiting probability or limiting expectation the limits of probabilities or expectations as introduced above. At last we define: (i = 1,2,...)

$\rho := \lambda\varepsilon[\underline{t} + N(\underline{t})S]$, the loading factor of the system. We will assume $\rho < 1$, so the system will not be saturated. In this case the expected length of the busy cycle is finite, for the busy cycle in our system is just as long as the busy cycle in an ordinary FIFO M/G/1 queue, where the jobs have a service time $\underline{t} + N(\underline{t})S$. So we may regard the various stochastic processes interesting us - such as the number of jobs in the system,

the number of portions in the j-th queue or the amount of work in the system (virtual delay) - as strictly regenerative processes (see Stidham [10]). Furthermore we will assume that the processing time distribution function F is nonarithmetic (nonlattice) and $\varepsilon[\underline{t}+N(\underline{t})S]^2 < \infty$ so that we can use the above-mentioned results of Ross [7] and Stidham [10].

p_i := the limiting probability the CP is servicing an i-th portion of a job. Applying "L = λw" (or Ross [7] Th. 5.8, or Stidham [10] Th. 5) yields $p_i = \lambda r_i \mu_i$ = the fraction of time that the CP is servicing i-th portions. Furthermore it follows easily with Lebesgue's monotone convergence theorem that

$$\rho = \lambda\varepsilon[\underline{t} + N(\underline{t})S] = \lambda\varepsilon \sum_{i=1}^{\infty} \underline{T}_i' = \lambda \sum_{i=1}^{\infty} \varepsilon\underline{T}_i' = \lambda \sum_{i=1}^{\infty} r_i\mu_i = \sum_{i=1}^{\infty} p_i .$$

p_0 := 1 - ρ.

w_i := the limiting expected delay of an i-th portion of a job with $N(\underline{t}) \geq i$. The delay is measured from the arrival of the job in system until the start of the service to its i-th portion.

3. The FB_∞ priority rule

The queueing system consists of countably many parallel queues with a single server (the CP). The queues are numbered 1,2,... and they have in that order decreasing priority. An arriving job requiring a processing time $\underline{t}$ is divided into $N(\underline{t})$ portions. The first portion joins the end of queue 1, the second joins queue 2, and so on, the last portion joins the end of queue $N(\underline{t})$. The portions $1,\ldots,N(\underline{t})-1$ require a service time $q_1+S,\ldots,q_{N(\underline{t})-1}+S$ respectively, the last portion requires a service time $\underline{t} - Q_{N(\underline{t})-1} + S$. As long as there are jobs in the system the CP admits to service the portion from the head of the highest-priority (lowest-index) nonempty queue. The service times of the portions in the j-th queue (j = 1,2,...) are i.i.d. random variables which have the same distribution as $\underline{T}_j$. The arrival process of portions in the j-th queue is a Poisson process with rate λr_j, which is easily proved.

We will now determine the limiting expected delay w_p of a p-th portion in system until the start of its service.

Consider a p-th portion (A) of a job arriving at the p-th queue (p = 1,2,...). Let the number of portions in the various queues immediately before the arrival be $\underline{n}_1, \underline{n}_2, \ldots$. Simultaneously with portion A the first p-1 portions of the same job arrive in the first p-1 queues. Portion A will have to wait until all portions in the first p queues have received service. Furthermore the CP is possibly servicing some portion at the moment of arrival which has to be finished. Let $\underline{t}_a$ have the limiting distribution of the time necessary to finish the portion in service and let $\underline{t}_b$ have the limiting distribution of the time to service the portions in the first p queues. Set $\underline{t}_0 := \underline{t}_a + \underline{t}_b$. During the time $\underline{t}_0$ new jobs arrive in system, the portions of which join the various queues. Portion A will have to wait until all portions of these new jobs which joined the first p-1 queues have been serviced. Let $\underline{t}_1$ have the limiting distribution of the time needed for service to these new portions. During $\underline{t}_1$ new jobs arrive in system, and so on. Going on in this way we define a sequence of random variables $\{\underline{t}_k\}_{k=1}^{\infty}$. The total waiting time of portion A amounts to $\sum_{k=0}^{\infty} \underline{t}_k$ and

$$(1) \qquad w_p = \varepsilon \sum_{k=0}^{\infty} \underline{t}_k = \sum_{k=0}^{\infty} \varepsilon \underline{t}_k$$

by the monotone convergence theorem of Lebesgue.

The expected service time of the first p-1 portions of an arbitrary job equals

$$\varepsilon \sum_{i=1}^{p-1} \underline{T}_i' = \sum_{i=1}^{p-1} r_i \mu_i .$$

The expected service time of the first p-1 portions of the $\underline{A}(\underline{t}_m)$ jobs arriving during the time $\underline{t}_m$ is by Wald's lemma (Ross [7], Th. 3.6)

$$\varepsilon \underline{t}_{m+1} = \varepsilon \Big[\sum_{j=1}^{\underline{A}(\underline{t}_m)} \sum_{i=1}^{p-1} \underline{T}_{ij}' \Big] = \varepsilon \underline{A}(\underline{t}_m) \sum_{i=1}^{p-1} r_i \mu_i$$

where $\{\sum_{i=1}^{p-1} \underline{T}_{ij}'\}_j$ are i.i.d. random variables which have the same distribution as $\sum_{i=1}^{p-1} \underline{T}_i'$ and $\varepsilon \underline{A}(\underline{t}_m) = \lambda \varepsilon \underline{t}_m$. So

$$\varepsilon \underline{t}_{m+1} = (\sum_{i=1}^{p-1} p_i)\varepsilon \underline{t}_m , \quad m = 0,1,2,\ldots .$$

Applying this equation repeatedly we find

$$\varepsilon \underline{t}_m = (\sum_{i=1}^{p-1} p_i)^m \varepsilon \underline{t}_0 , \quad m = 0,1,2,\ldots .$$

Substitution in (1) yields

$$(2) \qquad w_p = \sum_{k=0}^{\infty} (\sum_{i=1}^{p-1} p_i)^k \varepsilon \underline{t}_0 = \frac{\varepsilon \underline{t}_0}{1 - \sum_{j=1}^{p-1} p_j} ,$$

for $0 \leq \sum_{j=1}^{p-1} p_j \leq \rho < 1$.

So we have to determine the limiting expected delay due to the portions in queue and the portion in service at the moment of arrival, the new arrivals during the waiting time being taken into account by the factor

$$(1 - \sum_{j=1}^{p-1} p_j)^{-1}$$

(as in Cobham [3]).

Firstly $\underline{t}_a$: informally, $\{\underline{T}_{in}\}_{n=1}^{\infty}$ where $\underline{T}_{in}$ is the service time of the i-th portion of the n-th job arriving in system with $N(\underline{t}) \geq i$ is a renewal sequence. If $\underline{s}_i$ has the limiting distribution of the residual life in that renewal process then (see Ross [7], Th. 3.12 or Stidham [10], Th. 6)

$$\varepsilon \underline{s}_i = \frac{1}{2} \frac{\varepsilon \underline{T}_{i1}^2}{\varepsilon \underline{T}_{i1}} = \frac{1}{2} \frac{\mu_i^{(2)}}{\mu_i} .$$

Since $\underline{t}_a = \underline{s}_i$ if the CP is servicing an i-th portion at the arrival of portion A and $\underline{t}_a = 0$ if the system is empty,

$$(3) \qquad \varepsilon \underline{t}_a = \sum_{i=1}^{\infty} r_i \frac{1}{2} \frac{\mu_i^{(2)}}{\mu_i} = \sum_{i=1}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} .$$

A more formal proof can be obtained by using Stidham's [10], Th. 6 or Ross' [7], Th. 5.10.

Secondly $\underline{t}_b$: from the characterization given above it follows that:

$$\underline{t}_b = \sum_{j=1}^{p} \sum_{k=1}^{\underline{n}_j} \underline{T}_{jk} + Q_{p-1} + (p-1)S ,$$

where for each j $\{\underline{T}_{jk}\}_k$ are i.i.d. random variables which have the same distribution as $\underline{T}_j$ and $\underline{n}_j$ is the number of portions in the j-th queue. The first term is due to portions that were already in queue at the arrival of portion A, the second term is due to the foregoing portions of the job to which A belongs. From Stidham's theorems 3 and 1 and "L = λw" we obtain $\Pr\{\underline{n}_j = n\} = \lim_{t\to\infty} \Pr\{\underline{n}_j(t) = n\}$ and $\varepsilon\underline{n}_j = \lambda r_j w_j$. It is clear that for each j all $\underline{T}_{jk}$ are independent of $\underline{n}_j$ so that we get by first conditioning on $\underline{n}_j$

$$(4) \qquad \varepsilon\underline{t}_b = \sum_{j=1}^{p} \varepsilon\underline{n}_j \varepsilon\underline{T}_{j1} + Q_{p-1} + (p-1)S =$$

$$= \sum_{j=1}^{p} p_j w_j + Q_{p-1} + (p-1)S .$$

Substitution of (3) and (4) in (2) according to $\underline{t}_0 = \underline{t}_a + \underline{t}_b$ yields for $p = 1,2,\ldots$

$$(5) \qquad w_p = \frac{1}{1 - \sum_{j=1}^{p-1} p_j} \left[\sum_{i=1}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} + \sum_{j=1}^{p} p_j w_j + Q_{p-1} + (p-1)S \right] .$$

To obtain an explicit solution we need some notational simplifications. We write

$$\delta_\ell := 1 - \sum_{j=1}^{\ell} p_j \quad (\ell = 1,2,\ldots), \quad \delta_0 := 1$$

and

$$R := \sum_{i=1}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} .$$

With this (5) may be written

$$(6) \qquad w_p = \frac{\delta_{p-1}}{\delta_p \delta_{p-1}} \left[\sum_{j=1}^{p-1} p_j w_j + R + Q_{p-1} + (p-1)S \right], \quad p = 1,2,\ldots .$$

Application of (6) to w_{p-1} yields

$$w_p = \frac{\delta_{p-1}}{\delta_p \delta_{p-1}} \Big[\sum_{j=1}^{p-2} p_j w_j + R + Q_{p-1} + (p-1)S + \\ + \frac{p_{p-1}}{\delta_{p-1}} \Big\{ \sum_{j=1}^{p-2} p_j w_j + R + Q_{p-2} + (p-2)S \Big\} \Big] = \\ = \frac{\delta_{p-2}}{\delta_p \delta_{p-1}} \Big[\sum_{j=1}^{p-2} p_j w_j + R + Q_{p-2} + (p-2)S \Big] + \frac{\delta_{p-1}}{\delta_p \delta_{p-1}} (q_{p-1} + S) .$$

Application of (6) to w_{p-2} again yields

$$w_p = \frac{\delta_{p-3}}{\delta_p \delta_{p-1}} \Big[\sum_{j=1}^{p-3} p_j w_j + R + Q_{p-3} + (p-3)S \Big] + \\ + \frac{\delta_{p-2}}{\delta_p \delta_{p-1}} (q_{p-2} + S) + \frac{\delta_{p-1}}{\delta_p \delta_{p-1}} (q_{p-1} + S) .$$

By induction we obtain for $1 \le k \le p$

$$w_p = \frac{\delta_{p-k}}{\delta_p \delta_{p-1}} \Big[\sum_{j=1}^{p-k} p_j w_j + R + Q_{p-k} + (p-k)S \Big] + \\ + \frac{1}{\delta_p \delta_{p-1}} \sum_{j=1}^{k-1} \delta_{p-j} (q_{p-j} + S) ,$$

and for $k = p$

$$w_p = \frac{1}{\delta_p \delta_{p-1}} \cdot R + \frac{1}{\delta_p \delta_{p-1}} \sum_{j=1}^{p-1} \delta_{p-j} (q_{p-j} + S) .$$

So we have the explicit solution

$$(7) \qquad w_p = \frac{1}{(1 - \sum_{j=1}^{p-1} p_j)(1 - \sum_{j=1}^{p} p_j)} \Big[\sum_{j=1}^{\infty} \frac{1}{2} \lambda r_j \mu_j^{(2)} + \\ + \sum_{j=1}^{p-1} (1 - \sum_{i=1}^{j} p_i)(q_j + S) \Big] , \quad p = 1,2,\ldots .$$

Finally the limiting expected delay in system of a job (response time) given its processing time $\underline{t}$ is

$$(8) \qquad \varepsilon[\underline{D} \mid \underline{t}] = w_{N(\underline{t})} + \underline{t} - Q_{N(\underline{t})-1} + S \ ,$$

and the unconditional response time of a job is

$$(9) \qquad \varepsilon\underline{D} = \sum_{j=1}^{\infty} (w_j + \tilde{q}_j)(r_j - r_{j+1}) \ ,$$

where for $j = 1,2,\ldots$

$$\tilde{q}_j := \varepsilon[\underline{T}'_j \mid Q_{j-1} < \underline{t} \leq Q_j] = \int_{Q_{j-1}}^{Q_j} (t + S - Q_{j-1}) \frac{dF(t)}{r_j - r_{j+1}} \ ,$$

the expected service time of the last (j-th) portion of a job.
These results agree with those of Schrage [9] who used different methods.

4. The FB_N priority rule

Now there are N parallel queues with a single server (the CP), priorities decreasing from queue 1 to queue N. In the N-th queue a job receives successive quanta of service time until its completion. At the end of each quantum the job may be preempted because of an arrival of a new job. By the same reasoning as before we divide a job requiring a processing time $\underline{t}$ at the arrival in $N(\underline{t})$ portions which join the various queues. If $N(\underline{t}) \geq N$ then the portions $1,\ldots,N-1$ join the queues $1,\ldots,N-1$ respectively and the other portions join the N-th queue together in the order of their numbering. So the N-th queue contains blocks of portions belonging to the same job, the service times of the blocks being i.i.d. random variables which have the same distribution function as $\underline{\delta}_N$ (see below). We define

$\underline{\delta}'_N$:= the service time of the portions with index N or higher of an arbitrary job, i.e.

$$\underline{\delta}'_N = \sum_{j=N}^{\infty} \underline{T}'_j = \begin{cases} 0 & \text{if } \underline{t} \leq Q_{N-1} \ , \\ \underline{t} + N(\underline{t})S - (Q_{N-1} + (N-1)S) & \text{if } \underline{t} > Q_{N-1} \ . \end{cases}$$

$\underline{\delta}_N$:= the service time of the portions with index N or higher of an arbitrary job with $N(\underline{t}) \geq N$.
It follows easily that

$$\varepsilon\underline{\delta}'_N = \sum_{j=N}^{\infty} \varepsilon\underline{T}'_j = \sum_{j=N}^{\infty} r_j\mu_j \,, \tag{10}$$

and

$$\varepsilon\underline{\delta}_N = \varepsilon[\underline{\delta}'_N \mid \underline{t} > Q_{N-1}] = \sum_{j=N}^{\infty} \frac{r_j\mu_j}{r_N} \,. \tag{11}$$

Furthermore

$$\frac{1}{2}\,\varepsilon(\underline{\delta}'_N)^2 = \frac{1}{2}\sum_{j=N}^{\infty} \varepsilon(\underline{T}'_j)^2 + \sum_{j=N}^{\infty}\sum_{i=N}^{j-1} \varepsilon[\underline{T}'_j\underline{T}'_i] \,.$$

For $j > i$

$$\underline{T}'_j\underline{T}'_i = \begin{cases} 0 & \text{if } \underline{t} \leq Q_{j-1} \\ \underline{T}_j(q_i + S) & \text{if } \underline{t} > Q_{j-1} \end{cases} ,$$

and as a consequence

$$\frac{1}{2}\, r_N\varepsilon\underline{\delta}_N^2 = \frac{1}{2}\,\varepsilon(\underline{\delta}'_N)^2 = \sum_{j=N}^{\infty} \frac{1}{2}\, r_j\mu_j^{(2)} + \tag{12}$$

$$+ \sum_{j=N}^{\infty} r_j\mu_j(Q_{j-1} + (j-1)S - Q_{N-1} - (N-1)S) \,.$$

The waiting time of a portion entering the p-th queue ($p \leq N - 1$) does not depend on the occupation at the arrival of the $p+1^{th},\ldots,N^{th}$ queue nor on the new arrivals during the waiting time in these queues. So the limiting expected delay of such a portion is given by (7) for $1 \leq p \leq N - 1$.

We will now determine the limiting expected delay w_N of a N-th portion joining the N-th queue until the start of its service.
In the N-th queue blocks of portions belonging to the same job arrive according to a Poisson process with rate λr_N which is easily proved. w_N is also the limiting expected delay of a block of portions until the start of its service.

Consider a block of portions (A) arriving at the N-th queue. Application of Cobham's reasoning as before yields (cf. (2))

$$w_N = \frac{\varepsilon \underline{t}_0}{1 - \sum_{j=1}^{N-1} p_j}, \tag{13}$$

where $\varepsilon \underline{t}_0$ is the limiting expected time necessary to finish service to all the jobs in the system at the moment of arrival. We write again $\underline{t}_0 = \underline{t}_a + \underline{t}_b$ where

$\underline{t}_a$:= the time due to finish the portion in service if any at the moment of arrival plus the time due to finish the first block of portions if any in the N-th queue which has received at least once a quantum of service time,

$\underline{t}_b$:= the time due to service of the portions in the queues 1,...,N-1 and of the blocks of portions in the N-th queue which wait for first service and of the first N-1 portions belonging to the same job as block A.

For $\underline{t}_b$ we may write

$$\underline{t}_b = \sum_{j=1}^{N-1} \sum_{k=1}^{\underline{n}_j} \underline{T}_{jk} + \sum_{k=1}^{\underline{n}_N} \underline{\delta}_{Nk} + Q_{N-1} + (N-1)S ,$$

where for each j (j = 1,...,N-1) $\{\underline{T}_{jk}\}_k$ are i.i.d. random variables which have the same distribution as $\underline{T}_j$ and $\underline{n}_j$ is the number of portions waiting in the j-th queue, the $\{\underline{\delta}_{Nk}\}_k$ are i.i.d. random variables which have the same distribution as $\underline{\delta}_N$ and $\underline{n}_N$ is the number of blocks waiting in the N-th queue for first service. By Stidham's [10] theorems 3 and 1 and "L = λw" we obtain (cf. (4))

$$\varepsilon \underline{t}_b = \sum_{j=1}^{N-1} \lambda r_j w_j \cdot \mu_j + \lambda r_N w_N \varepsilon \underline{\delta}_N + Q_{N-1} + (N-1)S = \tag{14}$$

$$= \sum_{j=1}^{N-1} p_j w_j + \lambda r_N w_N \varepsilon \underline{\delta}_N + Q_{N-1} + (N-1)S .$$

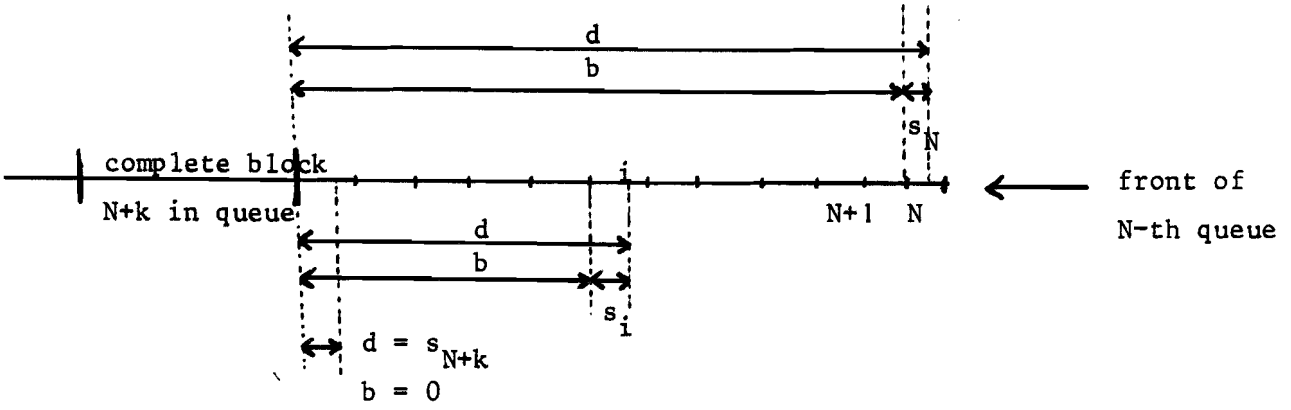
To determine $\varepsilon \underline{t}_a$ we need a somewhat closer investigation. When block A joins the N-th queue two possibilities arise:

a) the CP is servicing an i-th portion, $i \leq N - 1$. This portion has to be finished (time $\underline{s}_i$, $\varepsilon \underline{s}_i = \frac{1}{2} \frac{\mu_i^{(2)}}{\mu_i}$) plus a restblock if any in the N-th queue (time $\underline{b}$). This restblock consists of an integral number of portions of a job of which at least the N-th portion has been serviced.

b) the CP is servicing no portion (system is empty) or a portion out of the N-th queue. The block to which that portion does belong has to be finished (time $\underline{d}$, $\varepsilon \underline{d} = \frac{1}{2} \frac{\varepsilon \underline{\delta}_N^2}{\varepsilon \underline{\delta}_N}$).

Informally we write $\underline{b}$ in terms of $\underline{d}$ and $\underline{s}_j$ ($j \geq N$): (see figure)

$$\underline{b} = \begin{cases} \underline{d} - \underline{s}_j & \text{if the first portion of the restblock is an } j+1\text{-th portion, } j = N, N+1, \ldots \text{ (prob. } p_j (1 - \sum_{\ell=1}^{N-1} p_\ell)^{-1}) \\ 0 & \text{otherwise (prob. } p_0 (1 - \sum_{\ell=1}^{N-1} p_\ell)^{-1}) \, . \end{cases}$$



So

$$\varepsilon \underline{t}_a = \sum_{i=1}^{N-1} p_i \varepsilon \underline{s}_i + (\sum_{i=1}^{N-1} p_i)[\sum_{j=N}^{\infty} \frac{p_j \varepsilon(\underline{d} - \underline{s}_j)}{1 - \sum_{\ell=1}^{N-1} p_\ell}] + (\sum_{i=N}^{\infty} p_i) \varepsilon \underline{d} \, .$$

From (11) we obtain

$$\varepsilon \underline{d} = \frac{1}{2} \frac{\varepsilon \underline{\delta}_N^2}{\varepsilon \underline{\delta}_N} = \frac{1}{2} \cdot \frac{\lambda r_N \varepsilon \underline{\delta}_N^2}{\sum_{j=N}^{\infty} p_j} .$$

Furthermore

$$p_i \varepsilon \underline{s}_i = \lambda r_i \mu_i \cdot \frac{1}{2} \frac{\mu_i^{(2)}}{\mu_i} = \frac{1}{2} \lambda r_i \mu_i^{(2)}, \quad (i = 1,2,\ldots) .$$

This yields

$$(15) \qquad \varepsilon \underline{t}_a = \sum_{i=1}^{N-1} \frac{1}{2} \lambda r_i \mu_i^{(2)} + \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[\frac{1}{2} \lambda r_N \varepsilon \underline{\delta}_N^2 - \left(\sum_{j=1}^{N-1} p_j \right) \sum_{i=N}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} \right].$$

Substituting (14) and (15) in (13):

$$w_N = \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[\sum_{j=1}^{N-1} p_j w_j + w_N \lambda r_N \varepsilon \underline{\delta}_N + Q_{N-1} + (N - 1)S + \right.$$

$$\left. + \sum_{j=1}^{N-1} \frac{1}{2} \lambda r_j \mu_j^{(2)} + \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left\{ \frac{1}{2} \lambda r_N \varepsilon \underline{\delta}_N^2 - \left(\sum_{j=1}^{N-1} p_j \right) \sum_{i=N}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} \right\} \right],$$

and solving for w_N (using (11)):

$$(16) \qquad w_N = \frac{1}{1 - \rho} \left[\sum_{j=1}^{N-1} p_j w_j + Q_{N-1} + (N - 1)S + \sum_{j=1}^{N-1} \frac{1}{2} \lambda r_j \mu_j^{(2)} + \right.$$

$$\left. + \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left\{ \frac{1}{2} \lambda r_N \varepsilon \underline{\delta}_N^2 - \left(\sum_{j=1}^{N-1} p_j \right) \sum_{i=N}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} \right\} \right] .$$

With help of (5) and (7) we find

$$\sum_{j=1}^{N-1} p_j w_j = (1 - \sum_{j=1}^{N-2} p_j) w_{N-1} - \sum_{j=1}^{\infty} \frac{1}{2} \lambda r_j \mu_j^{(2)} - Q_{N-2} - (N-2)S =$$

$$= \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[\left(\sum_{j=1}^{N-1} p_j \right) \sum_{i=1}^{\infty} \frac{1}{2} \lambda r_i \mu_i^{(2)} + \right.$$

$$\left. + \sum_{j=1}^{N-2} (1 - \sum_{i=1}^{j} p_i)(q_j + S) \right] - Q_{N-2} - (N-2)S \ .$$

Substituting this and (12) in (16) we obtain after some algebra

$$(17) \qquad w_N = \frac{1}{(1-\rho)(1 - \sum_{j=1}^{N-1} p_j)} \left[\sum_{j=1}^{\infty} \frac{1}{2} \lambda r_j \mu_j^{(2)} + \sum_{j=1}^{N-1} (1 - \sum_{i=1}^{j} p_i)(q_j + S) + \right.$$

$$\left. + \sum_{j=N}^{\infty} p_j (Q_{j-1} - Q_{N-1} + (j-N)S) \right] .$$

Finally we have to determine the limiting expected delay w_{N+k} of a $N + k$-th portion ($k = 1,2,\ldots$) until the start of its service. Therefore we apply again Cobham's argument which yields (cf. (2))

$$w_{N+k} = w_{N+k-1} + \frac{q_{N+k-1} + S}{1 - \sum_{j=1}^{N-1} p_j} , \quad k = 1,2,\ldots .$$

Using this formula recursive we obtain

$$(18) \qquad w_{N+k} = w_N + \frac{Q_{N+k-1} - Q_{N-1} + kS}{1 - \sum_{j=1}^{N-1} p_j} , \quad k = 1,2,\ldots ,$$

where w_N is given by (17).

The response time of a job given its processing time is (cf. (8)):

$$(19) \qquad \varepsilon[\underline{D} \mid \underline{t}] = w_{N(\underline{t})} + \underline{t} - Q_{N(\underline{t})-1} + S ,$$

the unconditional response time of a job may be calculated from

$$(20)\qquad \varepsilon \underline{D} = \sum_{j=1}^{\infty} (w_j + \tilde{q}_j)(r_j - r_{j+1}) = \sum_{j=1}^{N-1} (w_j - \tilde{q}_j)(r_j - r_{j+1}) +$$

$$+ \sum_{j=N}^{\infty} \tilde{q}_j (r_j - r_{j+1}) + \sum_{k=1}^{\infty} \frac{r_{N+k}(q_{N+k-1} + S)}{1 - \sum_{j=1}^{N-1} p_j} + r_N w_N \,.$$

In (19) and (20) w_j is given by (7) if $j \leq N - 1$, by (17) if $j = N$ and by (18) if $j > N$, $\tilde{q}_j$ is introduced in (9).
These results agree with those of Krzesinski [6] who used different methods.

5. The RR_N priority rule

Just as in the FB_N case there are N parallel queues with a single server (the CP), priorities decreasing from queue 1 to queue N. But now a job in the N-th queue which does not complete its processing during a quantum joins the end of the N-th queue to receive a further quantum of processing time. Again we divide a job requiring a processing time $\underline{t}$ in $N(\underline{t})$ portions. If $N(\underline{t}) \leq N - 1$ then the portions $1,\ldots,N(\underline{t})$ join the queues $1,\ldots,N(\underline{t})$ respectively at the arrival moment of the job in the system as before. But now, if $N(\underline{t}) \geq N$, the $N + k$-th portion ($k = 0,\ldots,N(\underline{t})-N$) joins the end of the N-th queue at the moment the $N + k - 1$-th portion is finished. In the N-th queue we don't have a Poisson arrival process, but we may describe the situation in that queue from time to time.

It is clear that the RR_1 rule is not a specialization of the general RR_N rule. The input process of new jobs in the round-robin queue for the former is Poisson and for the latter not (not even in the case of exponential processing time in consequence of the constant swap time). In the RR_N ($N \geq 2$) system portions only join the round-robin queue at the moment the CP has just finished service to a j-th ($j \geq N - 1$) portion.

The RR_1 rule is analysed by Wolff [15] in an M/G/1 model without swap time, but his results may easily be extended to the case of a constant swap time. We will use his method to analyse the RR_N rule ($N \geq 2$), so from now on we assume $N \geq 2$.

We define:

$\underline{K}_j$ $(j = N,N+1,\ldots)$: the number of j-th portions in the N-th queue. Applying Stidham's [10] theorems and "L = λw" we obtain $Pr\{\underline{K}_j = n\} = \lim_{t\to\infty} Pr\{\underline{K}_j(t) = n\}$ and

$$\varepsilon\underline{K}_j = \lambda r_j(w_j - w_{j-1} - q_{j-1} - S) \tag{21}$$

(see below).

$p_j^{(n)}$ $(j = N-1,N,\ldots)$: the limiting probability the CP is servicing a j-th portion of a job with $N(\underline{t}) \geq j + n$. Applying "L = w" we find

$$p_j^{(n)} = \lambda r_{j+n}(q_j + S) \ . \tag{22}$$

d_j $(j = N,N+1,\ldots) := w_j - w_{j-1}$ (see below).

Just as in the FB_N case the limiting expected delay of a p-th portion $(p \leq N - 1)$ entering the p-th queue is given by (7) for $1 \leq p \leq N - 1$.

In the following we will mean by the waiting time of a N + k-th portion of a job with $N(\underline{t}) \geq N + k$ $(k = 0,1,2,\ldots)$ the time from arrival of the job in system until the start of the service of the N + k-th portion. The limiting expectation of this waiting time is w_{N+k}. The limiting expected delay in queue of that N + k-th portion is $w_{N+k} - w_{N+k-1} - q_{N+k-1} - S$. We will now determine w_{N+k} by deriving a set of linear equations for d_j $(j \geq N)$.

The waiting time of the N-th portion of a tagged job is composed of

(a) the waiting time of the N-1-th portion of the tagged job,

(b) 1. the service time of the N-1-th portion of the tagged job,

2. time needed to service portions which are present in the N-th queue at the arrival of the tagged job in system,

3. time needed to service N-th portions, if any, of jobs of which the N-1-th portions were present in the N-1-th queue at the arrival of the tagged job in system,

4. time needed to service the j+1-th portion, if any, of a job of which the j-th portion was in service at the arrival of the tagged job in system $(j = N-1,N,\ldots)$,

5. time needed to service the N-1-th portions of jobs which did arrive during the waiting time of the N-1-th portion of the tagged job,

(c) time needed to service new arrivals in higher priority queues $1,\ldots,N-1$ during the waiting time (b).

Following Cobham we obtain

$$w_N = w_{N-1} + \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[q_{N-1} + S + \sum_{j=N}^{\infty} (\varepsilon \underline{K}_j) \mu_j + (\varepsilon \underline{n}_{N-1}) \frac{r_N}{r_{N-1}} \mu_N + \right.$$

$$\left. + \sum_{j=N-1}^{\infty} p_j^{(1)} \mu_{j+1} + \lambda r_{N-1} w_{N-1} \mu_{N-1} \right] .$$

The waiting time of the N+1-th portion of a tagged job is composed of

(a) the waiting time of the N-th portion of the tagged job,

(b) 1. the service time of the N-th portion of the tagged job,

2. additional work performed on jobs mentioned in (b) 2 to 5 above,

3. time needed to service new arrivals in the N-th queue (N-th portions) during the time from the start of the service of the N-1-th portion of the tagged job to the start of the service of the N-th portion of the tagged job,

(c) time needed to service new arrivals in the higher priority queues $1,\ldots,N-1$ during the waiting time (b).

So

$$w_{N+1} = w_N + \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[q_N + S + \sum_{j=N}^{\infty} (\varepsilon \underline{K}_j) \frac{r_{j+1}}{r_j} \mu_{j+1} + \right.$$

$$+ (\varepsilon \underline{n}_{N-1}) \frac{r_{N+1}}{r_{N-1}} \mu_{N+1} + \sum_{j=N-1}^{\infty} p_j^{(2)} \mu_{j+2} +$$

$$\left. + (\lambda r_{N-1} w_{N-1}) \frac{r_N}{r_{N-1}} \mu_N + \lambda r_N (w_N - w_{N-1}) \mu_N \right] .$$

Continuing in this manner we obtain

$$w_{N+2} = w_{N+1} + \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \Big[q_{N+1} + S + \sum_{j=N}^{\infty} (\varepsilon \underline{K}_j) \frac{r_{j+2}}{r_j} \mu_{j+2} +$$

$$+ (\varepsilon \underline{n}_{N-1}) \frac{r_{N+2}}{r_{N-1}} \mu_{N+2} + \sum_{j=N-1}^{\infty} p_j^{(3)} \mu_{j+3} + (\lambda r_{N-1} w_{N-1}) \frac{r_{N+1}}{r_{N-1}} \mu_{N+1} +$$

$$+ \lambda r_N (w_N - w_{N-1}) \frac{r_{N+1}}{r_N} \mu_{N+1} + \lambda r_N (w_{N+1} - w_N) \mu_N \Big] ,$$

and by induction for k = 0,1,2,...

$$w_{N+k} - w_{N+k-1} = \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \Big[q_{N+k-1} + S + \sum_{j=N}^{\infty} (\varepsilon \underline{K}_j) \frac{r_{j+k}}{r_j} \mu_{j+k} +$$

$$+ (\varepsilon \underline{n}_{N-1}) \frac{r_{N+k}}{r_{N-1}} \mu_{N+k} + \sum_{j=N-1}^{\infty} p_j^{(k+1)} \mu_{j+k+1} +$$

$$+ (\lambda r_{N-1} w_{N-1}) \frac{r_{N+k-1}}{r_{N-1}} \mu_{N+k-1} +$$

$$+ \sum_{j=0}^{k-1} \lambda r_N (w_{N+j} - w_{N+j-1}) \frac{r_{N+k-1-j}}{r_N} \mu_{N+k-1-j} \Big] .$$

Substitution of (21) and (22), $\varepsilon \underline{n}_{N-1} = \lambda r_{N-1} w_{N-1}$ and $d_j = w_j - w_{j-1}$ gives us the following set of linear equations. For k = 0,1,2,...

$$(23) \qquad d_{N+k} = \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \Big[q_{N+k-1} + S + w_{N-1}(p_{N+k} + P_{N+k-1}) + \sum_{j=N}^{\infty} p_{j+k} d_j +$$

$$+ \sum_{j=0}^{k-1} p_{N+k-1-j} d_{N+j} \Big].$$

Consider first the finite case, i.e. $p_j = 0$ for all $j > K$ for some finite $K > N$. This will occur if all jobs have at most K portions ($F(Q_K) = 1$ or $Q_{K-1} < \infty$ and $q_K = \infty$). Now (23) applies for $d_N,\ldots,d_K$ and we have a finite set of linear equations. In the matrix of coefficients the sum of the absolute values of the nondiagonal elements of each row is

$$\frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[\sum_{j=N}^{N+k-1} p_{j+k} + \sum_{j=N+k+1}^{K-k} p_{j+k} + \sum_{j=0}^{k-1} p_{N+k-1-j} \right] =$$

$$= \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[\rho - \sum_{j=1}^{N-1} p_j - p_{N+2k} \right] < 1 - \frac{p_{N+2k}}{1 - \sum_{j=1}^{N-1} p_j} ,$$

the absolute value of the diagonal element. So the matrix of coefficients is nonsingular and the set of linear equations has a unique solution.

In the infinite case we need some additional arguments. We return for a moment to the original system-description where all jobs pass through the successive queues. Now we may interpret $\underline{K}_j$ ($j \geq N$) into the number of jobs in the system which wait for their j-th quantum of service time. Similarly we may define $\underline{K}_j$ for $j \leq N - 1$ as the number of jobs in the system (in the j-th queue) which wait for their j-th quantum of service time. Furthermore we denote by

v_j: the expected service time required to finish a job with $N(\underline{t}) \geq j$ of which the first j-1 portions are already serviced, and by

$\varepsilon\underline{v}$: the limiting expected virtual delay (total amount of work in system).

According to Stidham [10] and Wolff [14] $\varepsilon\underline{v} < \infty$ if $\varepsilon[\underline{t} + N(\underline{t})S]^2 < \infty$ (which we assumed) and

$$\varepsilon\underline{v} = \sum_{j=1}^{\infty} (\varepsilon\underline{K}_j) v_j + \frac{1}{2} \lambda \varepsilon[\underline{t} + N(\underline{t})S]^2 . \tag{24}$$

We have

$$\varepsilon\underline{K}_j = \lambda r_j (w_j - w_{j-1} - q_{j-1} - (1 - \delta_{j1})S), \quad j = 1,2,...$$

(see (21)), where $q_0 = w_0 = 0$ and δ_{j1} is the Kronecker symbol, and

$$v_j = \sum_{i=j}^{\infty} \frac{r_i}{r_j} \mu_i , \quad j = 1,2,...$$

(see (11)).

Hence we may conclude

$$\sum_{j=1}^{\infty} \lambda r_j (w_j - w_{j-1} - q_{j-1} - S) \sum_{i=j}^{\infty} \frac{r_i}{r_j} \mu_i < \infty ,$$

which yields, if $\sup_i q_i < \infty$, for all $k = 0,1,2,\ldots$

$$\sum_{j=N}^{\infty} p_{j+k} d_j < \infty$$

and these sums are uniformly bounded with respect to k.

Hence it follows from (23) that there are constants A and B, $0 < A < 1$, with

$$d_{N+k} \le A \max_{j<k} \{d_{N+j}\} + B$$

for all $k = 1,2,\ldots$, and $d_N \le B$.

But then is $\sup_k d_{N+k} < \infty$. For suppose $\sup_k d_{N+k} = \infty$ then we can find a j_0 such that

$$d_{N+j} \le \frac{B}{1 - A} < d_{N+j_0}$$

for all $j < j_0$.

Then it follows from (23)

$$d_{N+j_0} \le A \max_{j<j_0} \{d_{N+j}\} + B \le \frac{AB}{1 - A} + B = \frac{B}{1 - A}$$

which is a contradiction. So we may conclude: $\sup_k d_{N+k} < \infty$.

Suppose now $\{a_{N+j}\}_{j=0}^{\infty}$ and $\{b_{N+j}\}_{j=0}^{\infty}$ are bounded solutions of (23), i.e. $\sup_j |a_{N+j}| < \infty$ and $\sup_j |b_{N+j}| < \infty$, then the sequence $\{u_{N+j}\}_{j=0}^{\infty}$ defined by $u_{N+j} := a_{N+j} - b_{N+j}$ is bounded. (23) implies for $k = 0,1,2,\ldots$

$$u_{N+k} = \frac{1}{1 - \sum_{j=1}^{N-1} p_j} \left[\sum_{j=N}^{\infty} p_{j+k} u_j + \sum_{j=0}^{k-1} p_{N+k-1-j} u_{N+j} \right] .$$

Hence

$$|u_{N+k}| \leq \frac{1}{1 - \sum_{j=1}^{N-1} p_j} (\rho - \sum_{j=1}^{N-1} p_j) \sup_j |u_{N+j}| .$$

But

$$\frac{\rho - \sum_{j=1}^{N-1} p_j}{1 - \sum_{j=1}^{N-1} p_j} < 1 ,$$

so $u_{N+j} = 0$ for all j.
Thus in this case if $\{a_{N+j}\}_{j=0}^{\infty}$ is bounded it is the unique bounded solution of (23).

To summarize the infinite case: if $\sup_i q_i < \infty$ then $\{d_{N+j}\}_{j=0}^{\infty}$ is bounded and this sequence is the unique bounded solution of (23).

Remarks.

1. If $p_j = 0$ for all $j \geq N + 1$ then these results are the same as those in the model with the FB_N rule, which is necessary since in that case there is no difference between the two priority rules.
2. In case the processing times of the jobs are exponentially distributed and $q_{N+j} = q_N$ $(j = 1,2,...)$ we are able to determine the solution of (23) explicitly.

Let $F(t) = 1 - e^{-\mu t}$ and $\alpha := e^{-\mu q_N}$ then we may state for $j = 0,1,2,...$:

a) $r_j = e^{-\mu Q_{j-1}}$

b) $\mu_j = S + \frac{1}{\mu} (1 - e^{-\mu q_j})$

c) $\mu_{N+j} = \mu_N$

d) $r_{N+j} = r_N \alpha^j$

e) $p_{N+j} = p_N \alpha^j$ and $\sum_{j=0}^{\infty} p_{N+j} = \frac{p_N}{1 - \alpha}$

f) $p_{N+j+k} = p_{N+j} \alpha^k$ for $k = 0,1,2,...$ and $\sum_{j=N}^{\infty} p_{j+k} d_j = \alpha^k \sum_{j=N}^{\infty} p_j d_j$.

From (24) we obtain

$$(25)\qquad \sum_{i=1}^{\infty} (d_i - q_{i-1} - (1 - \delta_{i1})S) \sum_{j=i}^{\infty} p_j = \varepsilon\underline{v} - \frac{1}{2}\lambda\varepsilon[\underline{t} + N(\underline{t})S]^2 .$$

According to Stidham [10] and Wolff [14 and 15] $\underline{v}$ is invariant under "work-conserving disciplines" (as is the case here), so we may determine $\varepsilon\underline{v}$ from a FIFO queue. For a FIFO M/G/1 queue it is known that

$$\varepsilon\underline{v} = \frac{\lambda\varepsilon[(\underline{t} + N(\underline{t})S]^2}{2(1 - \rho)} .$$

Splitting the left-hand side sum in (25) and using (a) to (f) we get finally

$$(26)\qquad \sum_{j=N}^{\infty} p_j d_j = \frac{p_N}{1-\alpha}[(1-\alpha)(q_{N-1} + S) + \alpha(q_N + S)] +$$

$$+ (1-\alpha)[\frac{\rho\lambda}{2(1-\rho)}\varepsilon(\underline{t} + N(\underline{t})S)^2 - \sum_{j=1}^{N-1} p_j(w_j - Q_{j-1} - (j-1)S) +$$

$$- \frac{p_N}{1-\alpha}(w_{N-1} - Q_{N-2} - (N-2)S)] .$$

Furthermore, on account of (f), the equations (23) may be written as ($k = 0,1,2,\ldots$)

$$(27)\qquad d_{N+k} = \frac{1}{1 - \sum_{j=1}^{N-1} p_j}[q_{N+k-1} + S + w_{N-1}(p_{N+k} + p_{N+k-1}) + \alpha^k \sum_{j=N}^{\infty} p_j d_j +$$

$$+ \sum_{j=0}^{k-1} p_{N+k-1-j} d_{N+j}] .$$

By (27), (26) and (7) we may consider d_N as a known number and for $k \geq 1$ (27) may be written as

$$(28)\qquad d_{N+k} = \frac{1}{1 - \sum_{j=1}^{N-1} p_j}[q_N + S + w_{N-1}p_N(\alpha^k + \alpha^{k-1}) + \alpha^k \sum_{j=N}^{\infty} p_j d_j +$$

$$+ \sum_{j=0}^{k-1} p_N \alpha^{k-j-1} d_{N+j}] .$$

To simplify our notation we denote by

$$x_{N+j} := \frac{d_{N+j}}{\alpha^{j+1}}, \quad j = 0,1,2,\ldots,$$

$$C := \frac{1}{1 - \sum_{j=1}^{N-1} p_j}$$

and

$$D := \frac{1}{\alpha}\left[w_{N-1}p_N\left(1 + \frac{1}{\alpha}\right) + \sum_{j=N}^{\infty} p_j d_j\right].$$

Then (28) changes into

$$x_{N+k} = C\left[\frac{q_N + S}{\alpha^{k+1}} + D + \frac{p_N}{\alpha}\sum_{j=0}^{k-1} x_{N+j}\right], \quad k = 1,2,\ldots. \tag{29}$$

By using (29) recursively we obtain for $\ell = 1,\ldots,k-1$

$$x_{N+k} = \frac{C(q_N + S)}{\alpha^{k+1}}\left[1 + Cp_N \sum_{j=0}^{\ell-1} (\alpha + Cp_N)^j\right] +$$

$$+ CD\left[1 + \frac{Cp_N}{\alpha}\sum_{j=0}^{\ell-1}\left(1 + \frac{Cp_N}{\alpha}\right)^j\right] + \frac{Cp_N}{\alpha}\left(1 + \frac{Cp_N}{\alpha}\right)^{\ell}\sum_{j=0}^{k-1-\ell} x_{N+j}.$$

Substitution of $\ell = k - 1$ and multiplication by α^{k+1} yields for $k = 1,2,\ldots$

$$d_{N+k} = C(q_N + S)\left[1 + Cp_N \sum_{j=0}^{k-2} (\alpha + Cp_N)^j\right] + \tag{30}$$

$$+ C\alpha^k\left[w_{N-1}p_N\left(1 + \frac{1}{\alpha}\right) + \sum_{j=N}^{\infty} p_j d_j\right]\left[1 + \frac{Cp_N}{\alpha}\sum_{j=0}^{k-2}\left(1 + \frac{Cp_N}{\alpha}\right)^j\right] +$$

$$+ Cp_N(\alpha + Cp_N)^{k-1} d_N,$$

where $\sum_{i=N}^{\infty} p_j d_j$ and d_N follows from (26), (27) and (7).

From (30) we may calculate w_{N+m} $(m = 1,2,...)$ since

$$w_{N+m} = \sum_{k=0}^{m} d_{N+k} + w_{N-1} ,$$

and using formula's like (19) and (20) we may obtain the conditional and the unconditional response time of a job.

The results for this special case of exponential processing time agree with those of Adiri and Avi-Itzhak [2] who used different methods which are only employable in this case.

References

[1] I. Adiri & B. Avi-Itzhak: A time-sharing queue, Man. Sci. 15 (July 1969).

[2] I. Adiri & B. Avi-Itzhak: A time-sharing model with many queues, OR. 17 (1969).

[3] A. Cobham: Priority assignment in waiting line problems, OR. 2 (1954).

[4] E.G. Coffman & L. Kleinrock: Feed-back queueing models for time-sharing systems, ACM. J. 15 (1968).

[5] H.C. Heacox & P.W. Purdom: Analysis of two time-sharing queueing models ACM. J. 19 (1972).

[6] A.E. Krzesinski: Stochastic modelling of a time-sharing system, to be published.

[7] S.M. Ross: Applied probability models, Holden-Day (1970).

[8] R. Schassberger: Warteschlangen, Springer-Verlag (1973).

[9] L.E. Schrage: The queue M/G/1 with feed-back to lower priority queues, Man. Sci. 13 (March 1967).

[10] S. Stidham: Regenerative processes in the theory of queues, Adv. Appl. Prob. 4 (1972).

[11] S. Stidham: A last word on $L = \lambda w$, OR. 22 (1974).

[12] L. Takács: Introduction to the theory of queues, N-Y-Oxford University Press (1962).

[13] Th. v.d. Weide: Verwachte doorlooptijden en wachtrijbezettingen in een time-sharing systeem, stageverslag THE (1973).

[14] R.W. Wolff: Work-conserving priorities, J. Appl. Prob. 7 (1970).

[15] R.W. Wolff: Time-sharing with priorities, SIAM. J. Appl. Math. 19 (1970).