A STATIC-PRIORITY QUEUE WITH TIME SLICING

B. WALKE

A. E. G. Telefunken - Ulm - Federal Republic of Germany

ABSTRACT

An infinite queue single-server model with N external preemptive-resume priorities is analysed. Customers with priority i ($1 \leq i \leq N$) arrive according to a Poisson process with intensity λ_i, join the i^{th} priority level and demand service according to a general distribution function. In each priority level the service discipline may be either FCFS or MLFCFS (multi-level FCFS). In the MLFCFS discipline service requests are assigned to different sublevels within a priority level, dependent on the amount of attained service. Between such sublevels the feedback discipline FB_∞ is applied.

We compute the mean waiting time of customers with priority number i, both with and without condition on their service requirement. Examples are discussed for which the mean waiting-time is plotted.

1. INTRODUCTION

In real-time computer-control systems and time-shared computer systems usually service demands of different importance have to be distinguished. Reasons for such importance-dependent handling of service demands may be: different response-time requirements of different subgroups of customers which form the total workload, different importance of customer subgroups or different cost functions of service demands etc. Usually service demands with different importance are assigned different extern priority numbers which are fixed. Whenever a situation arises, in which service demands with different priority request service at the same time, the highest priority (lowest priority number) demand is serviced first.

In this paper we consider service disciplines for only computing facilities in terms of a queueing model. Service demands of priority i are completely described by their interarrival and service times, both being independent random variables.

From theory of scheduling [7], [8] it is known that service demands with a large squared coefficient of variance C^2 (variance over squared mean), namely $C^2 > 1$, must be served by means of an appropriate feedback (FB) discipline to reach a minimum mean waiting time. This fundamental law is violated whenever the FCFS discipline is used to serve demands with $C^2 > 1$. FCFS only is optimal when $C^2 \leq 1$ is satisfied. Recalling that the FCFS discipline is nearly generally used within fixed priority levels of real-time computer-control systems one may, without any knowledge about C^2 in such systems, suspect that something should be done there to clearify and possibly change this situation. One goal of this paper is to promote the investigation process needed by presenting analytic results for different workloads to demonstrate the advantages of time-slicing disciplines applied within extern priority levels. Such workloads differ in the value of C presumed.

Multi-level queueing models with static priorities and FB disciplines are rarely to be found in the literature. A recently published paper by Babad [1] considers a so-called generalized multi-entrance and multipriority M/G/1 time-sharing system. In each priority level of his model systems of queues exist. An arriving customer joins the n^{th} priority level and there the k^{th} queue and is eligible to a finite or infinite quantum of service. If the service requirements of this customer are not satisfied during this quantum, he is transferred to the next lower priority (n+1) and there to the end of the k^{th} queue for additional service. The head of the line discipline is used after finishing a quantum to choose the nex customer to be serviced. Preliminary papers are cited in [1]; all of them can be characterized by a model M/M/1 with static priorities and quantum service. Another paper comparable to Babads paper should be mentioned [2].

It solves for the Laplace-Stieltjes transform of the conditional waiting time. A more recently paper [3], too, considers a model of the same type as in [1]. Typical for all these papers is that a service quantum is not interruptable. Moreover the external priority of a service demand is not consequently observed: demands with lower external priorities may get service before a demand with higher external priority and large service time requirement, due to the FB discipline applied. All these models may be called feedback models with arrivals not only at the first but also subsequent levels.

Our model differs from these models in that 1) the external priority of a service demand is used to guarantee preferential service over all lower priority demands, 2) the FB discipline with finite or infinite quantum size is used only within each priority level, 3) preemption (of the resume type) of service quanta within a priority level and between different priority levels is allowed.

The service discipline within each of the feedback queues is FCFS. Using service quanta of infinite length within each priority level, our model degenerates to the well-known model with ...eemtive resume priorities, first studied by White and Christie [4] and Miller [5].

2. THE MODEL

We consider an infinite queue model of a computer system, cf. Fig. 1, with N extern-priority levels of the preemptive resume type.

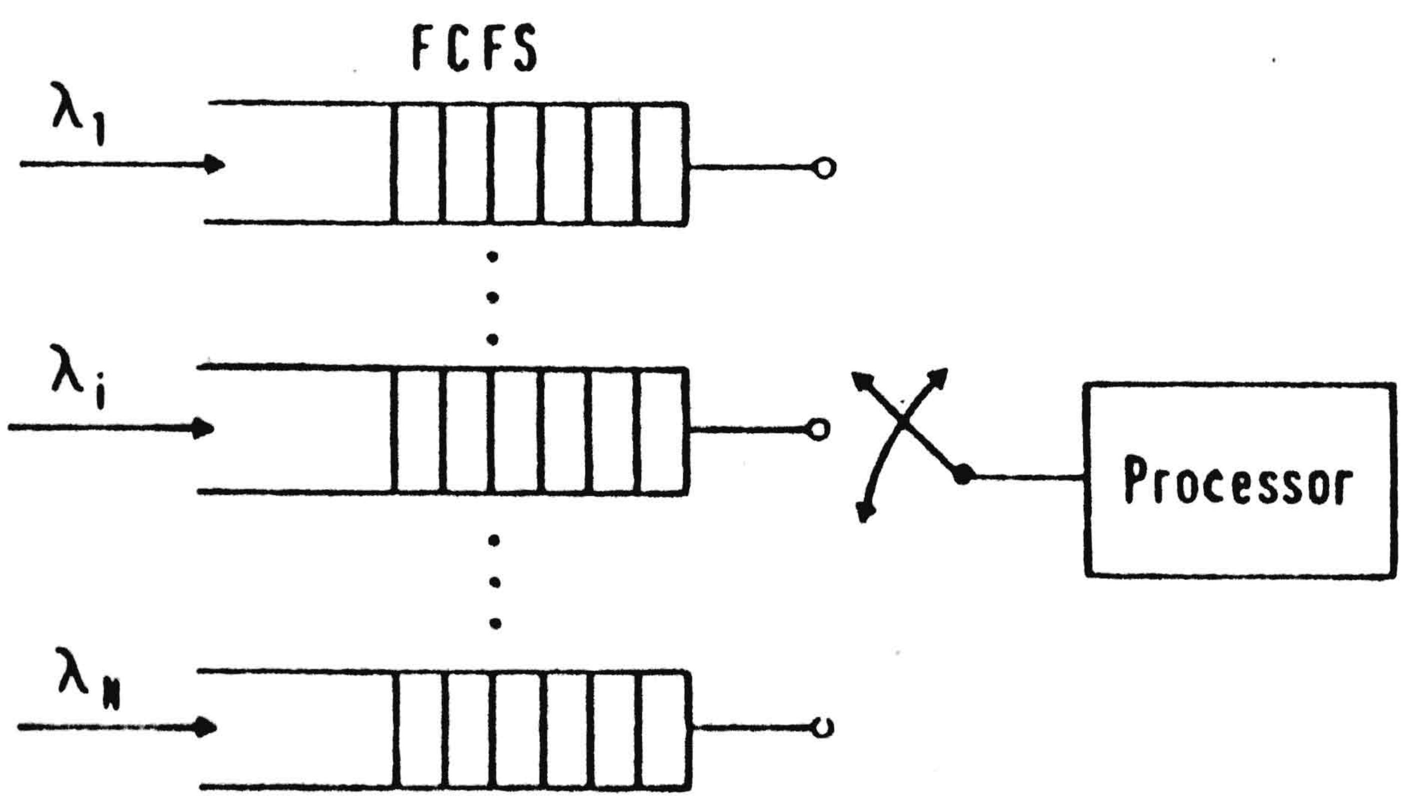

Fig. 1: Model with N queues one for each external preemptive-resume priority level. Within each level the FCFS discipline is applied. λ_i = arrival rate.

Customers with, say, priority number i ($1 \leq i \leq N$) arrive from a Poisson arrival process with parameter λ_i and demand service. Their service times b_i are assumed to be independent and identically distributed with general distribution function (d.f.) and finite second moment. The related arrival processes are assumed to be independent, too. Apparently, we consider a multi-entrance M/G/1 model. Small priority numbers denote high priorities. The service discipline within priority level i is of the feedback type with finite or infinite length service quanta which are interruptable. The functioning of this discipline can be explained by means of the model in Fig. 2 and by Table I. The model in Fig. 2 should be thought of as inserted in the priority level i of the model in Fig. 1, as is shown in Fig. 3. The servie-time d.f. of customers with priority number i is defined by $P(b_i \leq t) = B_i(\leq t)$. We use the simpler model in Fig. 2 to define the multi-level first come first serve (MLFCFS) discipline.

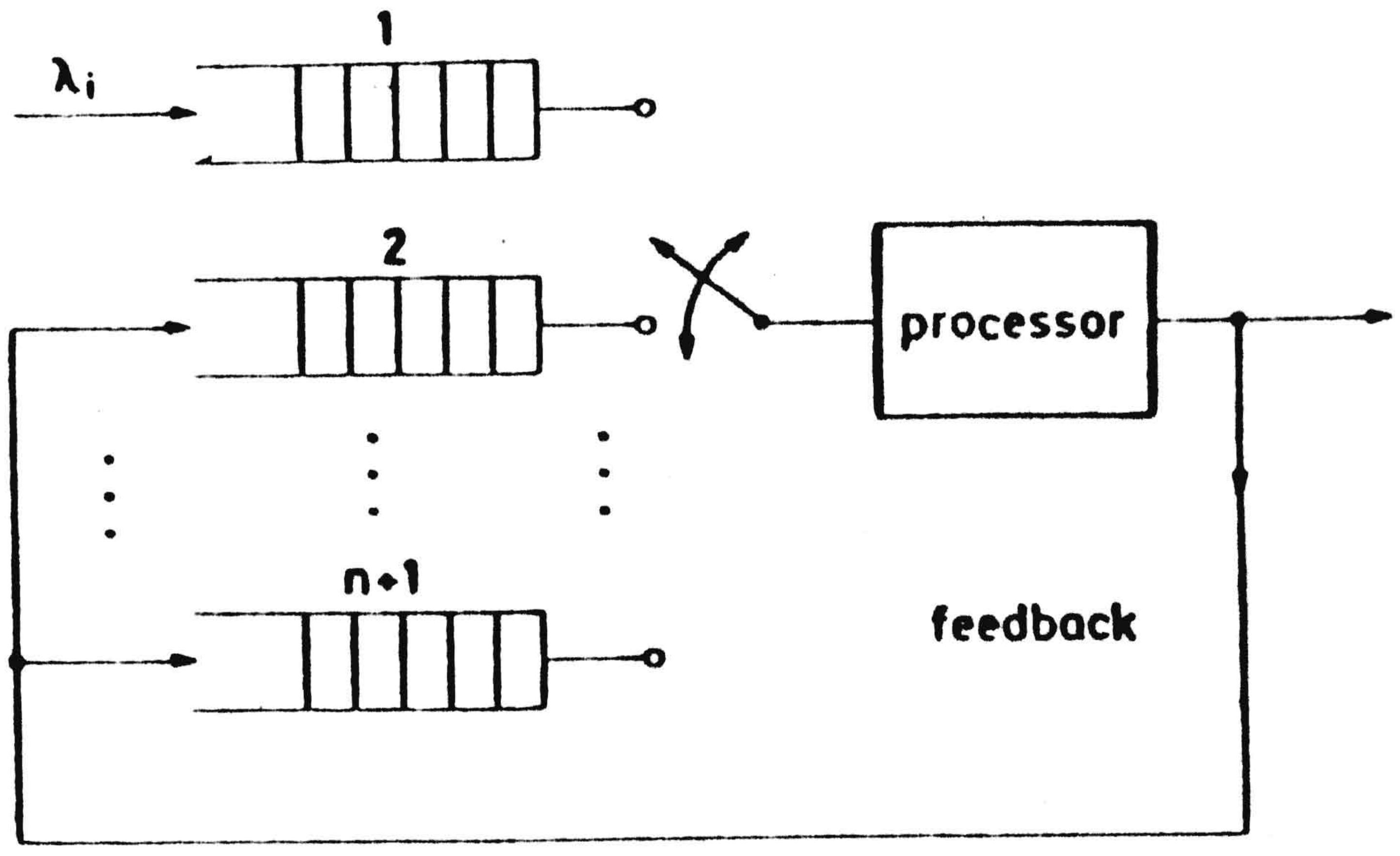

Fig. 2: Model with (n+1) feedback levels and the FCFS discipline applied to the k-th level ($1 \leq k \leq n+1$). The servicing discipline for customers from the arrival process is called MLFCFS (multi-level FCFS).

A new arriving customer joins the queue number 1, where he remains until either his service requirement is fullfilled during the first quantum of size t_{g1} and he leaves the model or he has consumed the total quantum. Customers which need more than the first quantum are fed back, when the quantum has expired, to queue number 2 where the quantum size is $(t_{g2}-t_{g1})$. Customers which need still more service are fed back again etc. In queue number n+1 the quantum size is infinite, cf. Table I.

level of the queueing system	1	2	...	n+1
attained service time is	$\leq t_{g1}$	$\leq t_{g2}$	...	$\leq t_{gn+1} = \infty$

Table I: Attained service time in queue number i of the model in Fig. 2.

The notion "quantum" does not mean that the service in any of such queues may not be interrupted. Instead of this preemptive resume prioritites are applied to give preferential service to high priority (low number) queues of the model in Fig. 2. In each queue the service discipline is FCFS. It can easily be seen that, for an infinite quantum assigned to customers of the first queue, the other queues are obsolete and the simple

FCFS queue arises.

The following abbreviations are used throughout:

$\rho_i = \lambda_i \beta_i$ offered traffic to priority level i,

$\rho_{\le N} = \sum_{i=1}^{N} \rho_i$ offered traffic to priority levels 1 through N, total offered traffic,

$\rho_{<k} = \sum_{i=1}^{k-1} \rho_i$ offered traffic to priority levels 1 through k-1,

$\gamma^{(r)} = \int_0^\infty t^r dP(\le t)$ r^{th} moment of $P(\le t)$,

$W_i^{(r)}, \beta_i^{(r)}$ r^{th} moment of waiting-time and service-time d.f.'s, respectively

$W_i = W_1^{(1)}, \beta_i = \beta_i^{(1)}, \gamma_i = \gamma_1^{(1)}$ means (= first moments)

$B_{\le k}(\le t) = \lambda_{\le k}^{-1} \sum_{i=1}^{k} \lambda_i B_i(\le t)$ weighted service-time d.f. of customers with priorities 1 through k

$\beta_{\le k}^{(r)} = \lambda_{\le k}^{-1} \sum_{i=1}^{k} \lambda_i \beta_i^{(r)}$ r^{th} moment of the weighted service-time d.f. $B_{\le k}(\le t)$

$\sigma^2 = \gamma^{(2)} - \gamma^2$ variance

$c = \sigma/\gamma$ coefficient of variance.

In sections 3 and 4 we compute the mean waiting time in the MLFCFS discipline for arbitrary service-time d.f.'s and in section 5 we apply this discipline by means of three examples.

3. PREVIOUSLY KNOWN RESULTS

Computation of the first two moments of the mean waiting-time of the M/G/1 model in Fig.1, presuming Poisson arrival processes, arbitrary service-time d.f.'s and FCFS scheduling in each priority level, is known [4], [5]. We recall these results shortly in section 3.1. On the other hand the M/G/1 model in Fig.2 has been analysed in the MLFCFS discipline and the first moment of the mean waiting time was found in [6] for an arbitrary number of levels, while the Laplace transform of the waiting-time d.f. for a two-level model was found in [9]. In literature [6] the MLFCFS was called a processor sharing discipline with feedback between levels and FCFS for any level. The corresponding results are recalled in section 3.2.

3.1 Model with preemtive priorities and FCFS in each customer queue

We consider the model in Fig.1. The response time of a test request can be combined of three disjoint portions, namely (1) the waiting-time before service of this request, (2) the individual service-time and (3) the waiting-time which results from interruption periods during which requests with higher priorities are serviced. Usually the last two components are added together and are called completion time. The related moments will be abbreviated by $V_i^{(r)}$ (response), $W_i^{*(r)}$ (first wait) and $T_i^{(r)}$ (completion). The first two moments of the total waiting-time of a customer with priority i is known from [10] to be

$$W_i^{(r)} = V_i^{(r)} - \beta_i^{(r)} = W_i^{*(r)} + T_i^{(r)} - \beta_i^{(r)}, \quad (r=1,2) \qquad (3.1)$$

with

$$W_i^* = \lambda_{\le i} \beta_{\le i}^{(2)} / (2(1-\rho_{\le i})(1-\rho_{<i})) \qquad (3.2)$$

$$W_i^{*(2)} = \lambda_{\le i} \beta_{\le i}^{(3)} / (3(1-\rho_{\le i})(1-\rho_{<i})^2) + (\lambda_{\le i} \beta_{\le i}^{(2)})^2 / (2(1-\rho_{\le i})^2 (1-\rho_{<i})^2) + \lambda_{<i} \beta_{<i}^{2} \lambda_{\le i} \beta_{\le i}^{(2)} / (2(1-\rho_{\le i})(1-\rho_{<i})^3) \qquad (3.3)$$

$$T_i = \beta_i + \beta_i \rho_{<i} / (1-\rho_{<i}) \qquad (3.4)$$

$$T_i^{(2)} = \beta_i^{(2)} / (1-\rho_{<i})^2 + \beta_i \lambda_{<i} \beta_{<i}^{(2)} / (1-\rho_{<i})^3 \qquad (3.5)$$

From Eqs. (3.1), (3.2) and (3.4) we have the mean waiting-time

$$W_i = W_i^* + T_i - \beta_i . \qquad (3.6)$$

The second moment can be computed from Eqs.(3.1), (3.3) and (3.5).

3.2 A simple model in the MLFCFS discipline

We consider the model in Fig.2 where we have k feedback levels ($1 \le k \le n+1$). The attained service time of a service request in level k is as shown in Table I.

In deriving an expression for the mean response time for this model in the MLFCFS discipline, a test request with a service requirement $t_{gk-1} < t \le t_{gk}$) is considered. This request just reaches the k^{th} level and then leaves the model. The following argumentation is used in [6]; which we reproduce here because a similar argumentation is used in section 4. The queue number k can be considered in isolation to some extent and two facts may be used.

1) By assumption of preemptive priority of lower level queues with numbers j, ($1 \le j < k$), i.e. FB discipline between levels, it is clear that requests in levels i>k can be ignored. This follows since these jobs cannot interface with the servicing of the lower levels.
2) The system time (= response time) of our test request can be thought of as occuring in two parts. The first portion is the time from the requests arrival to the first level queue until the k-th level is serviced for the first

time after the test request has reached the k-th level. The second portion starts with the end of the first portion and ends when this request leaves the system. Both the first and the second portions of the request's system time can easily be seen to be unaffected by the service discipline used in levels 1 through k-1. Therefore one can assume any convenient discipline. In fact, all these levels can be lumped into one equivalent level which services customers with a service requirement t, $(0 < t \le t_{gk-1})$, using any service discipline.

From (1) and (2) it follows that the mean response time V(t) of customers can be computed that leave the system from the k-th level by considering a two level system, say M1. The lower level services customers with service requirements between 0 and t_{gk-1} whereas the second level services customers with requirements between t_{gk-1} and t_{gk}. Customers that would have passed to the (k+1)-st level after receiving a service time t_{gk} in the original system are now assumed to leave the system at that point. In other words the service-time d.f. is truncated at t_{gk}. We consider the simpler two level queueing system M1 with FCFS discipline in both levels, the second level corresponding to the k-th level in the original system.

A test customer entering the system M1 will be delayed by the sum of the work currently in both levels plus any new arrivals to the lower queue during the interval this customer is in the system. These new arrivals form a Poisson process with parameter λ and their contribution to the delay is a random variable with moments

$$t^{(r)}_{k-1} = \int_0^{t_{gk-1}} t^r dB(\le t) + t^r_{gk-1} [1-B(\le t_{gk-1})]. \quad (3.7)$$

The traffic offered from these arrivals is λt_{k-1}. The sum of the work currently in both levels exactly equals the mean waiting-time of a single-queue FCFS system, namely

$$W_k' = \lambda t_k^{(2)} / (2(1-\lambda t_k)), \quad (3.8)$$

wheras λt_k is the traffic offered to the levels 1 through k in the original system and t_k and $t_k^{(2)}$ can be computed from Eq.(3.7) by substituting (k-1) by k.

As usual [6] the response time V(t) of a test customer with service requirement t can be combined from W_k', t, and $\lambda t_{k-1} V(t)$ to be

$$V(t) = W_k' + t + \lambda V(t)\, t_{k-1}, \quad (3.9)$$

from which we have

$$V(t) = (W_k' + t) / (1-\lambda t_{k-1}). \quad (3.10)$$

After defining P_k to be the probability of a customer's service requirement to fall between the limits t_{gk-1} and t_{gk}

$$P_k = \int_{t_{gk-1}}^{t_{gk}} dB(\le t), \quad (3.11)$$

and further defining E_k to be the mean service time of a customer whose service requirement is between the limits t_{gk-1}, t_{gk}

$$E_k = \frac{1}{P} \int_{t_{gk-1}}^{t_{gk}} t dB(\le t) \quad (3.12)$$

it is then possible by weighting and summarizing to compute the mean response time of all customers together

$$V = \sum_{k=1}^{n+1} \frac{P_k [W_k' + E_k]}{1-\lambda t_{k-1}}, \text{ with } t_o = 0. \quad (3.13)$$

From this the mean waiting time can be found

$$W = V - \beta .$$

4. NEW RESULTS

We now consider the model in Fig.3 where the queueing system of Fig. 2 is inserted in the system of Fig.1 at an arbitrary priority level i.

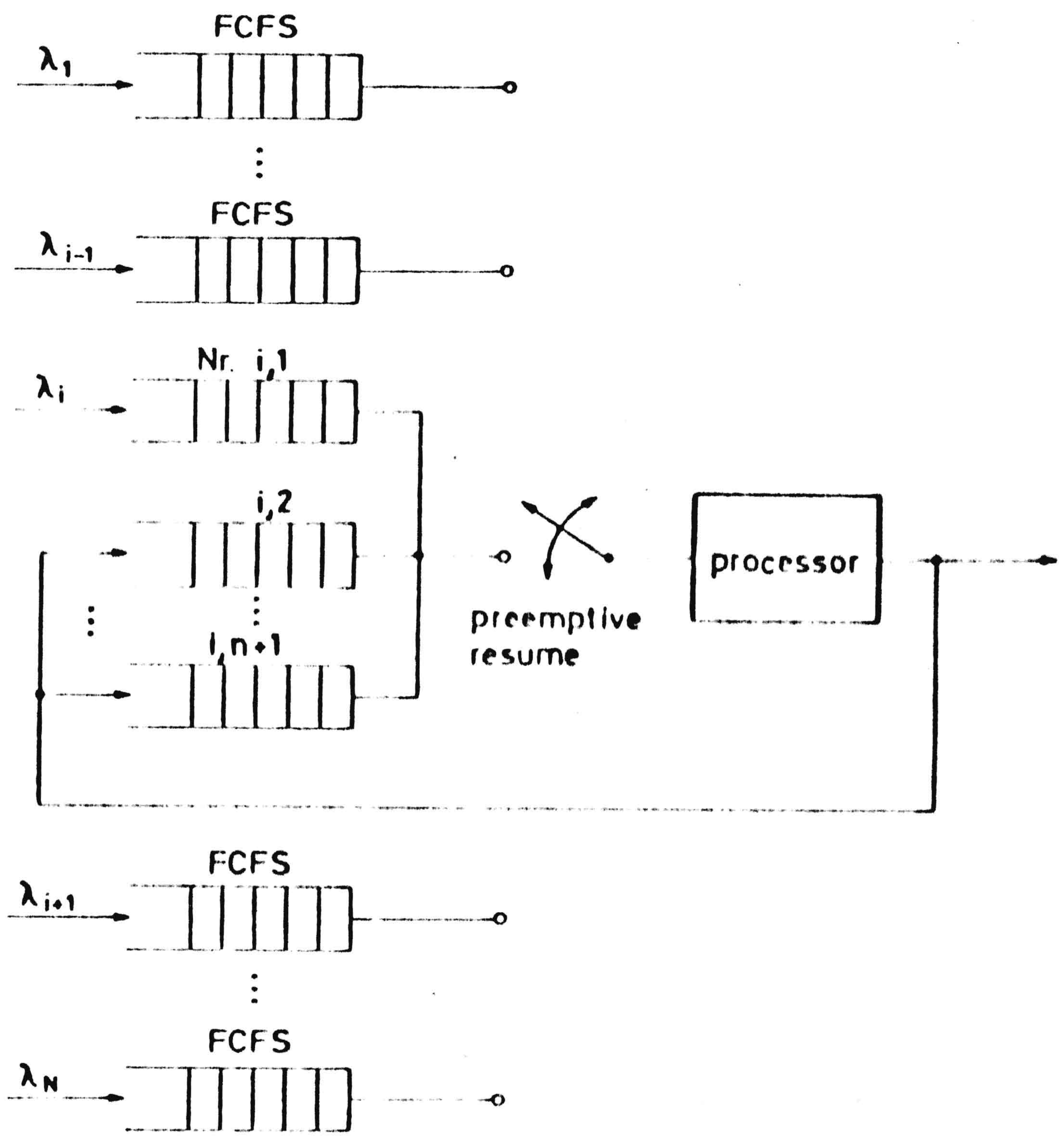

Fig.3: Model with N external preemptive-resume priority levels and the MLFCFS and FCFS disciplines applied to levels i and j≠i, resp. (i,j ∈ N).

The different queues at priority level i are indexed (i,k), ($1 \le k \le n+1$). We are interested in computing the mean waiting-time of a customer of priority level i which leaves the whole system from level (i,k). Inside priority level i the MLFCFS discipline is applied.

The service-time d.f. of customers in the level i is assumed to be $B_i(\le t)$. The service requirement t of a customer being queued in queue (i,k) is limited to be $t_{gi,k-1} < t \le t_{gi,k}$, where the index i refers to the external priority i of the related level.

We now may apply the same arguments used in section 3.2:

1) By the assumption of preemtive priority of lower level queues with numbers j, ($1 < j < k$)
and r ($1 \le r \le i-1$) it is clear that requests with $j > k$
and $r > i$ can be ignored.

2) The response time of a test request with service-time t, which enters the system at queue number (i,1) and leaves the system from queue number (i,k) can be thought of as occuring in two parts.

Both portions (being the same as described in section 3.2, paragraph 2) are unaffected by the service discipline used in levels 1 through (i,k-1). Therefore these levels can be lumped together to form one equivalent level which services all requests with priority numbers 1 through (i-1) and those of i which reach the level (i,k-1).

Now the two level system M1 mentioned in section 3.2 is considered. The lower level services customers with service requirements between 0 and $t_{gi,k-1}$ and all those customers which belong to priority levels 1 through i-1. The second level services customers with service requirements between $t_{gi,k-1}$ and $t_{gi,k}$. The ervice-time d.f. of this level is assumed to be truncated at $t_{gi,k}$. The two levels system M1 is considered with FCFS discipline in both levels, the second level corresponding to the (i,k) -th level in the original system.

We make the following definitions:

$B_{\le i,k}(\le t)$ = weighted service-time d.f. of all customers with priorities 1 through (i-1) and those of priority i, which leave the system from level (i,k).

$$B_{\le i,k}(\le t) = \lambda_{\le i}^{-1}\left[\sum_{j=1}^{i-1} \lambda_j B_j(\le t) + \lambda_i \, P_i(\le t)\right] \qquad (4.1)$$

$$\text{with } P_i(\le t) = \begin{cases} B_i(\le t) & \text{for} \quad 0 \le t \le t_{gi,k} \\ 1 & \text{otherwise} \end{cases}$$

We note that the r^{th} moment of this d.f. is defined by

$$t_{\le i,k}^{(r)} = \lambda_{\le i}^{-1} \left[\sum_{j=1}^{i-1} \lambda_j \int_0^\infty t^r \, dB_j(\le t) + \right.$$
$$\left. + \lambda_i \left\{ \int_0^{t_{gi,k}} t^r \, dB_i(\le t) + t_{gi,k}^r \left[1 - B_i(\le t_{gi,k})\right]\right\}\right]$$

which can be written, using Eq. (3.7),

$$t_{\le i,k}^{(r)} = \lambda_{\le i}^{-1} \left[\sum_{j=1}^{i-1} \lambda_j \beta_j^{(r)} + \lambda_i \, t_k^{(r)} \right]. \qquad (4.2)$$

The remaining course of the computation equals that of section 3.2. Therefore we only give the results here corresponding to Eqs.(3.8), (3.10), (3.11), (3.12), (3.13)

$$W'_{i,k} = (\lambda_{\le i} \, t_{\le i,k}^{(2)}) / (2(1 - \rho_{<i} - \lambda_i \, t_k)). \qquad (4.3)$$

$$V_i(t) = (W'_{i,k} + t)/(1 - \lambda_{\le i} \, t_{\le i,k-1}). \qquad (4.4)$$

$$P_{i,k} = \int_{t_{gi,k-1}}^{t_{gi,k}} d\, B_i(\le t) = B(\le t_{gi,k}) - B(\le t_{gi,k-1}). \qquad (4.5)$$

$$E_{i,k} = 1/P_{i,k} \int_{t_{gi,k-1}}^{t_{gi,k}} t \, d\, B_i(\le t) \qquad (4.6)$$

$$V_i = \sum_{k=1}^{n+1} (P_{i,k}[W'_{i,k} + E_{i,k}])/(1 - \lambda t_{\le i,k-1}). \qquad (4.7)$$

As can easily be seen, by recalling the argumentation leading to the simple two level model M1, these results remain unchanged even if in any of the priority levels $j \ne i$ of the model in Fig. 3 the MLFCFS instead of the FCFS discipline is applied. In fact, all priority levels may be scheduled by means of the MLFCFS discipline without changing the results for the i-th level.

The mean waiting time W_i can be computed from

$$W_i = V_i - \beta_i$$

and the mean waiting time in the MLFCFS discipline, depending on the service requirement t

$$W_i(t) = (W'_{i,k} + t \, \lambda_{\le i} t_{\le i,k-1})/(1 - \lambda_{\le i} t_{\le i,k-1}) \qquad (4.8)$$

with $t_{gi,k-1} < t \le t_{gi,k}$.

Finally, by using the results of section 3.1. for the model in Fig.1, the second moment $W_{i,1}^{(2)}$ of the waiting-time d.f. of customers of priority level i can be computed, which have a service requirement t($0 < t \le t_{gi,1}$). This is possible only if the third moments of all involved service-time d.f.'s, cf. Eq.(4.1), are finite.

5. MEAN WAITING TIME OF CUSTOMERS WITH PRIORITY I

First we recall that for customers with a large coefficient of variance, C>1, an appropriate feedback discipline can be expected to gain a smaller mean waiting time than the FCFS discipline [7], [8]. If $C \leq 1$ than the FCFS discipline is optimal.

Using the results of section 4 we are able to compute the mean waiting time of every priority level i in the model of Fig.1 using either the FCFS or MLFCFS discipline.

It is common practice to approximate service-time d.f.'s with C>1 by members of the hyperexponential d.f. of order n (H_n) family which is defined by

$$P(\leq t) = 1-\sum_{i=1}^{n} w_i\, e^{-\mu_i t};\ \mu_i>0,\ w_i>0,\ \sum_{i=1}^{n} w_i=1. \qquad (5.1)$$

h approximations imply that the FB_∞ discipline with infinite small service quanta must be used to yield the minimum mean waiting time. This is unacceptable for practical applications due to the vast amount of switching overhead associated with the FB_∞ discipline. Therefore in praxis feedback disciplines of the MLFCFS type which are only suboptimal are used to service customers with hyperexponential d.f.. For example problems arise in sizing the quanta to be applied in each of the levels.

5.1 Examples

Throughout this section we assume the model in Fig.3 having N=3 external priority levels and equally to the levels distributed total offered traffic, namely $\rho_i=\rho_{\leq N}/N$. Moreover, we assume the same service-time d.f. for customers of ..l three priority levels. We consider an Example 1 with a H_2 service-time d.f., defined by Eq.(5.1) with parameters w_1=0.9, w_2=0.1, $\mu_1=1.24s^{-1}$, $\mu_2=0.05\ s^{-1}$, t_g=2s, C=3.5.

Using the expressions from section 4 the mean waiting times for both disciplines FCFS and MLFCFS, applied to customers of priority level i ($1\leq i\leq N=3$), can be computed. The corresponding results are shown in Fig .4, where the meain waiting time W_i over the mean service β_i at the ordinate is plotted over the total offered traffic $\rho_{\leq N}$. The lines and broken lines show the results of the FCFS and MLFCFS disciplines, respectively. The case of saturation, $\rho_{\leq N}>1$, is included.
A comparison of the mean waiting times of the FCFS and MLFCFS reveals that for a fixed $\rho_{\leq N}$ the difference is larger the lower the priority is. Next we consider an
Example 2 with the same H_2 d.f. as in example 1

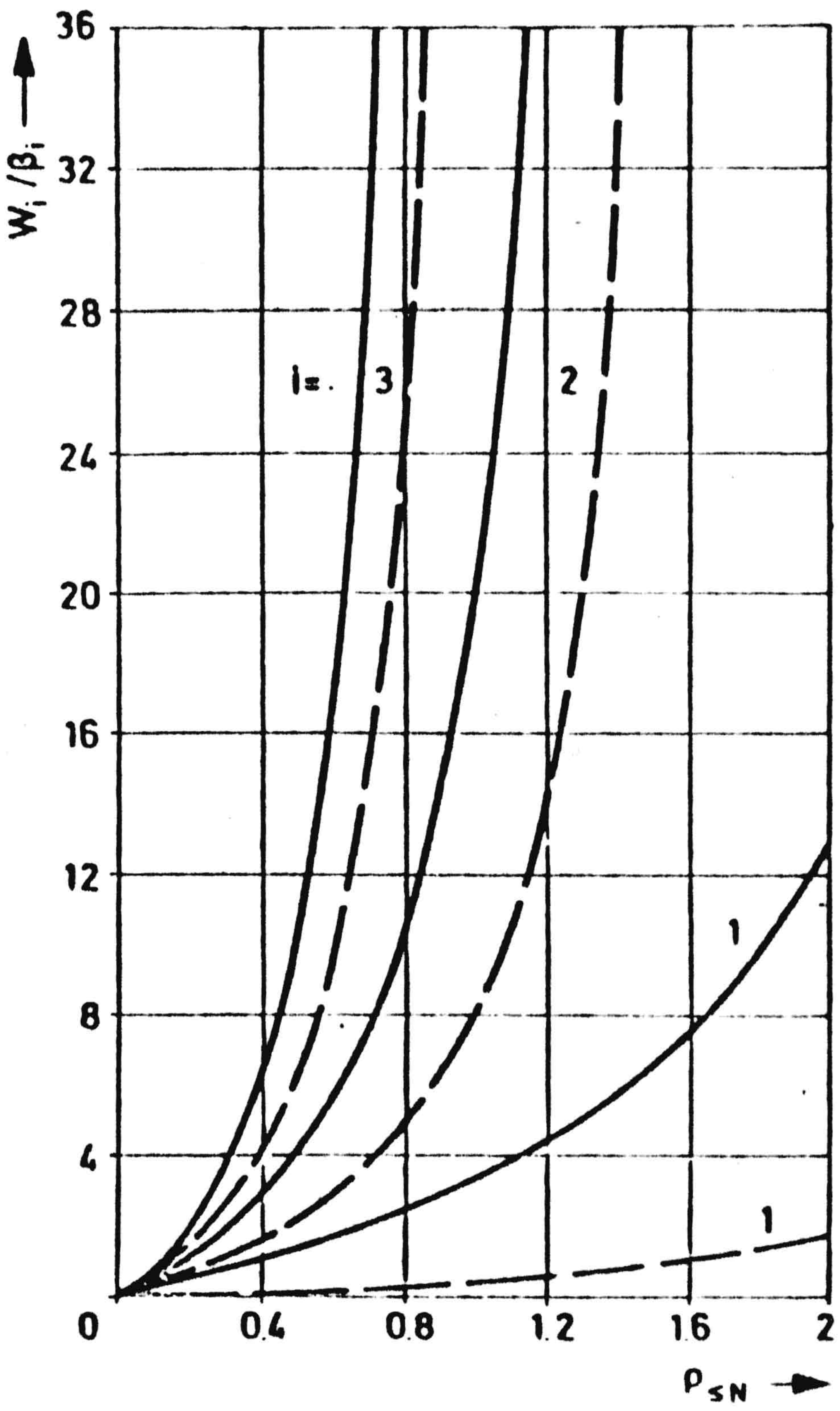

Fig.4: Normalized mean waiting time over total offered traffic for the model in Fig. 3 and the assumptions described in example 1, section 5.1. The curves show results of the FCFS (lines) and MLFCFS (broken lines) disciplines applied to priority level i.

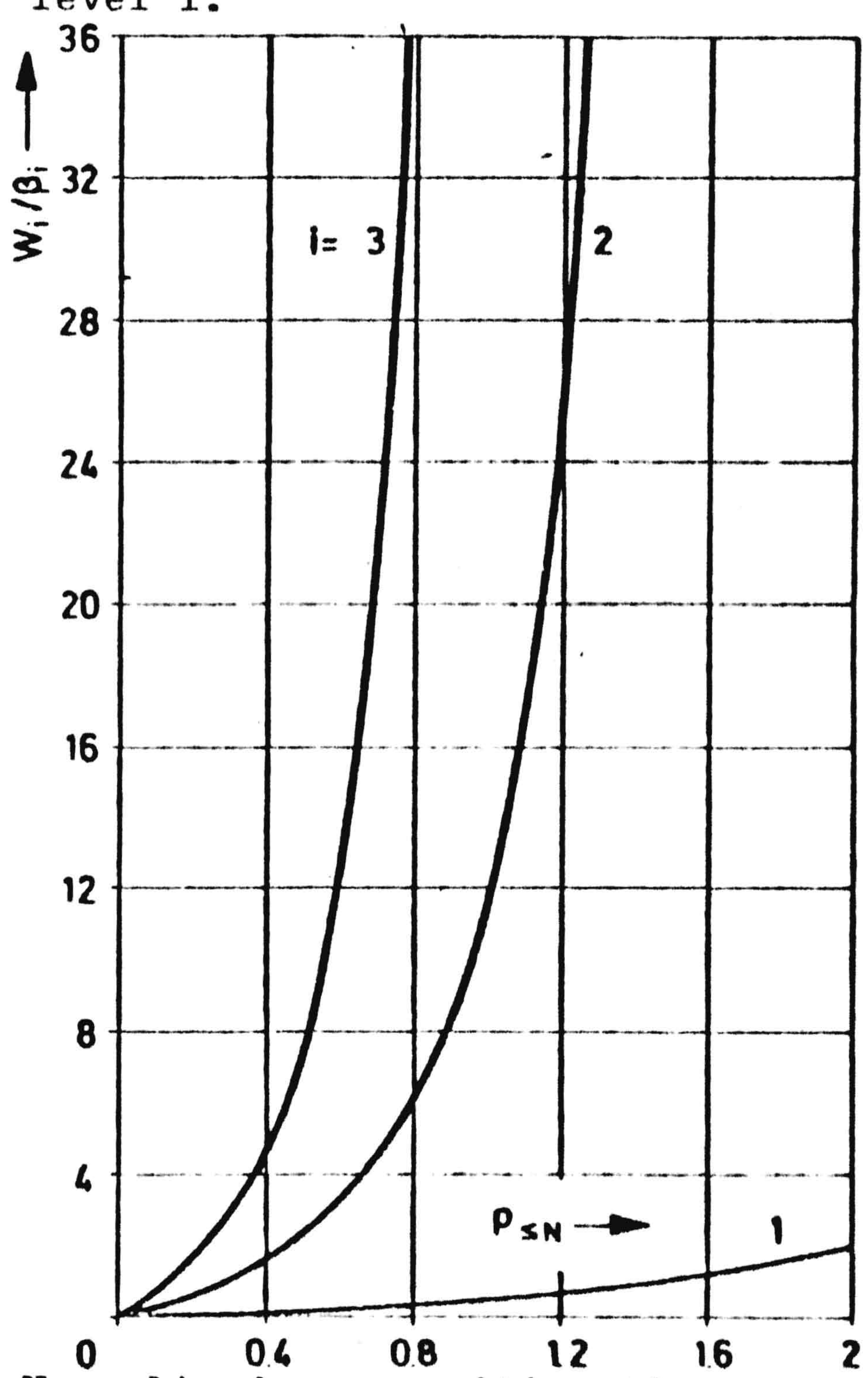

Fig.5: Normalized mean waiting time over total offered traffic for the model in Fig. 3 and the assumptions described in example 2, section 5.1.

in priority levels $j \neq i$ but with a M-d.f. with parameter $\mu_i = 1/\beta_i$, in the priority level i.

In this case the discipline FCFS and MLFCFS both applied to the level i yield the same results which are shown in Fig. 5 for i=1,2,3. A comparison to the results of the first example, cf. Fig. 4, reveals that for a fixed priority level i the broken lines in Fig. 4 always are below the lines in Fig. 5. By this and other similar examples it can be shown that the waiting-time optimal discipline yields a smaller mean waiting time, the greater the coefficient of variance, $C \geq 1$, is.

An expression for the mean waiting time $W_i(t)$ conditioned on the service requirement t of a customer of priority level i has also been computed in section 4. Fig. 6 shows results for the assumptions of example 1: $W_2(t)$ is shown over t for three different total offered traffics ρ_{sN}=0.3, 0.6, 0.9. The jump at t=t=2s is a consequence of the two-level queueing system used for the MLFCFS discipline with the first level having preemptive priority over the second.

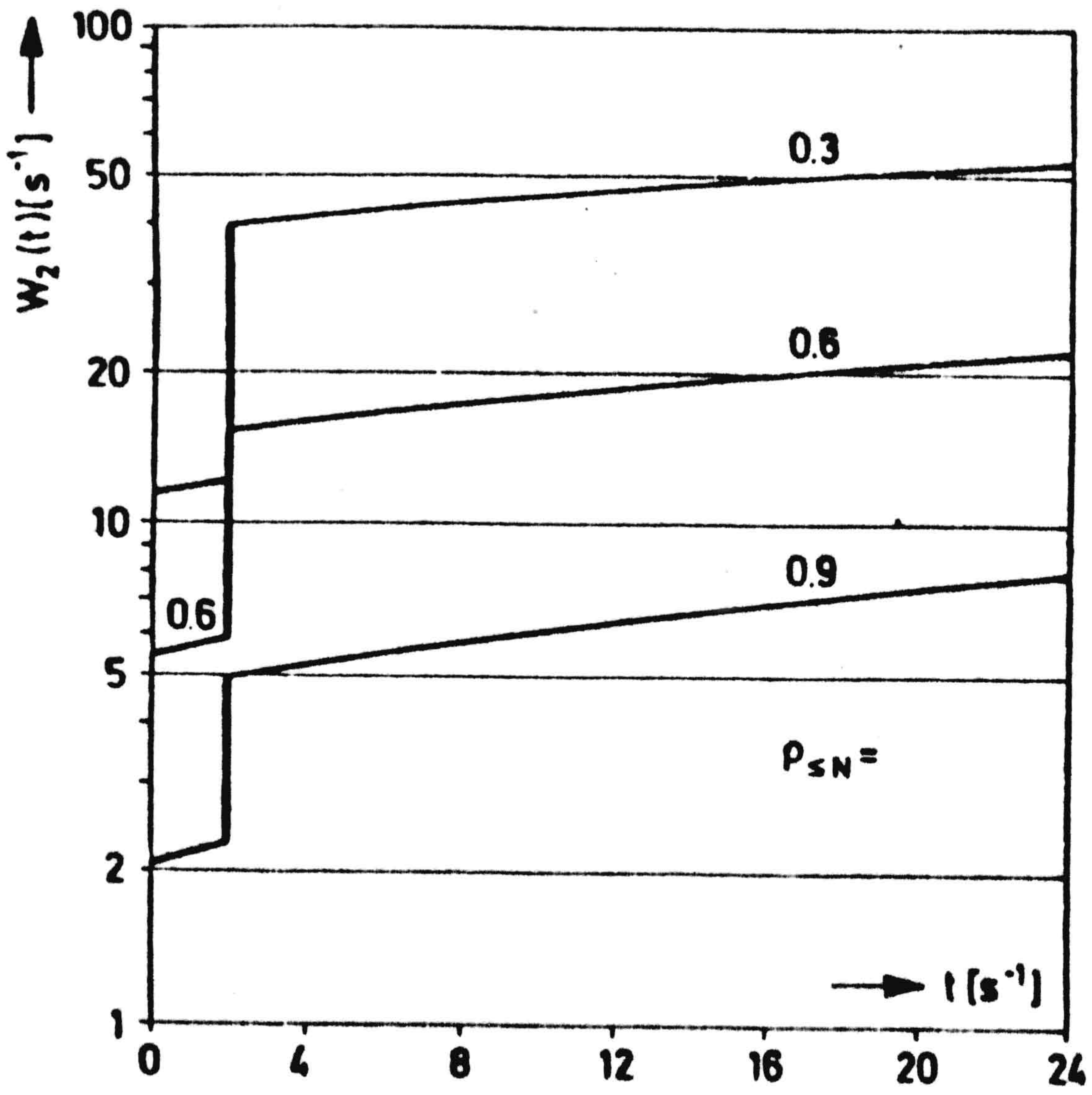

Fig.6: Conditioned mean waiting time over service requirement t of customers of priority level 2 of the model in Fig.3 in the MLFCFS discipline.
Assumptions are those of example 1, section 5.1. Parameter is the total offered traffic.

For some applications the size of the jumps at $t=t_g$ may be undesirable large. Instead of this a more continous increase of $W_i(t)$ dependent on the service time t might be appropriate. This is simply possible by introducing more than two levels in the MLFCFS discipline, cf.

Example 3: The same load as in example 2 but two quanta t_{g1}=0.25s and t_{g2}=5.8s are used. From Fig. 7 it can be seen that the conditioned waiting time now has two jumps at $t=t_{g1}$ and $t=t_{g2}$, each of smaller size than t_g, whose sizes summarized are greater than at t_g (cf. Fig.6). Two such jumps better reach the goal to attain a more continous increase of $W_i(t)$ dependent on t (without large sized jumps), than is the case with only one jump, cf. Fig.6. The mean waiting time W_i for both examples 1 and 3 nearly remains the same.

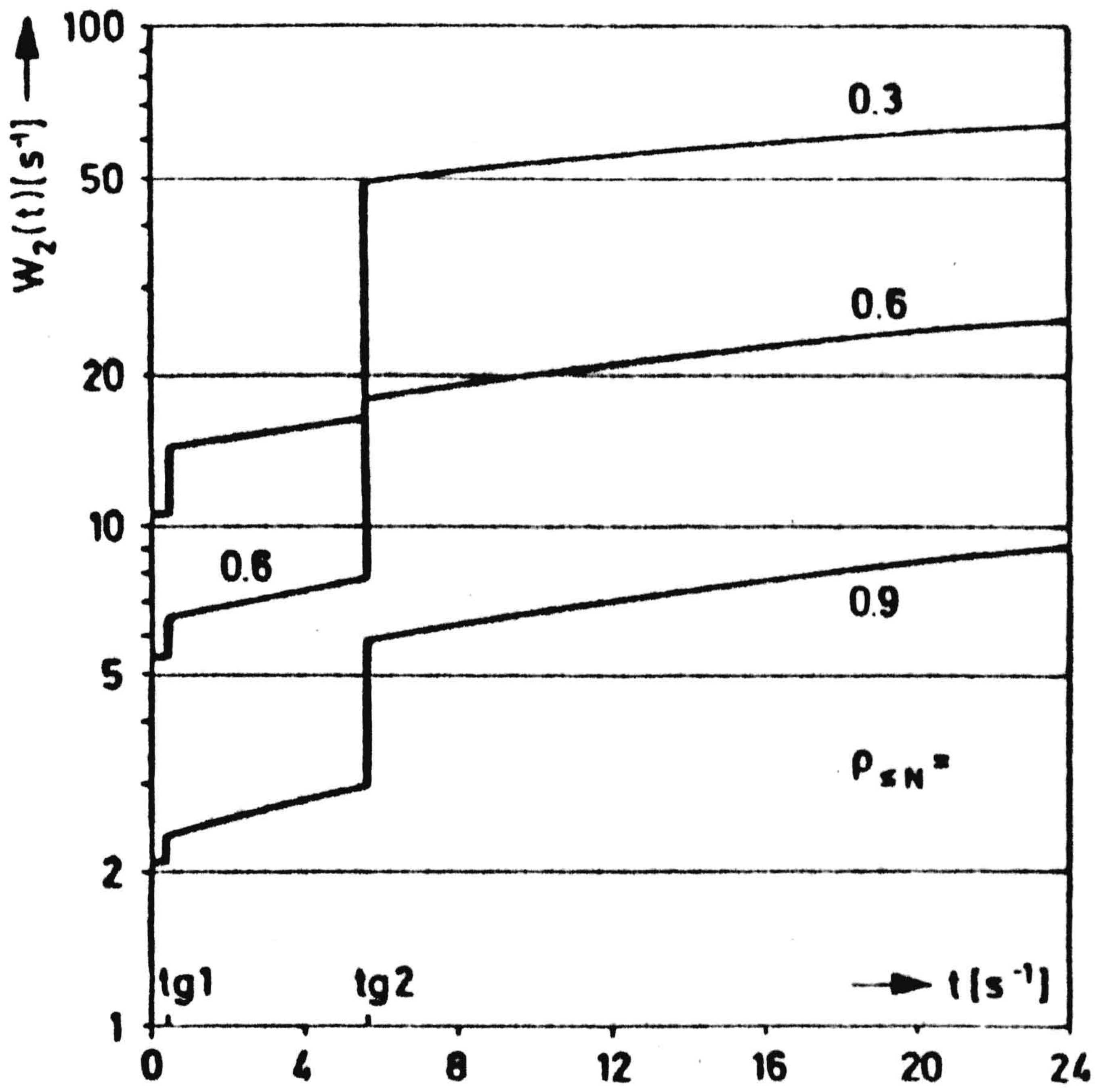

Fig.7: Conditioned mean waiting time over service requirement t of customers of priority level 2 of the model in Fig.3 in the MLFCFS discipline. Assumptions are those of example 3, section 5.1. Parameter is the total offered traffic.

REFERENCES

[1] Babad, J.M.: A generalized multi-entrance time-sharing priority queue. Journ. ACM, Vol22, No.2, April 1975, pp.231-247.

[2] Fujiki, M., Gambe, E.: Conditional waiting time in a multi-level feedback priority queueing model. Electron. and Comm. in Japan, Vol.52-C, No.1, 1969, pp.129-136.

[3] Rosberg, Z., Adiri, I.: Multi-level queues with external priorities. Journ. ACM, Vol. 23, No.4, Oct. 1976, pp.680-690.

[4] White, H., Christie, L.S.: Queueing with preemptive priorities or with breakdown. Oper.Res.(USA), Vol.6, No.1, 1958, pp.79-95.

[5] Miller, R.G.: Priority queues. Annals Math. Stat. 31, 1960, pp.86-103.

[6] Kleinrock, L., Muntz, R.R.: Multi-level processor-sharing queueing models for time-shared systems. 6th Int. Teletraffic Congr., München 1970, Congressbook pp. 314/1-8.

[7] Olivier, G.: Optimum time allocation for queueing data-processing jobs. Elektron. Rechenanl. 9 (1967), No.5, pp.218-224. In German.

[8] Sakata, M., Noguchi, S., Oizumi, J.: Optimal scheduling in single-processor systems. Proc. 3rd Hawaii Int. Conf. Syst. Sienc. 1970, Univ. of Hawaii, pp.708-711.

[9] Guimaraes, C.: A foreground-background time-sharing queue with general service times. Proc.Int.Symp.Operat.Sys., Rocquencourt, France, April 1974, Lect.Notes Comp.Science, 16, Springer Verlag, Berlin, New York 1974.

[10] Chang, W.: Single-server queueing processes in computing systems. IBM-Syst. Journal, No.1, 1970, pp.36-71.