Measuring, Modelling and Evaluating Computer Systems,
H. Beilner and E. Gelenbe, (eds.)
© North-Holland Publishing Company (1977)

IMPROVED BOUNDS AND AN APPROXIMATION FOR A DYNAMIC PRIORITY QUEUE.

Bernhard Walke
AEG-TELEFUNKEN, Research Institute,
Elisabethenstraße 3, D-7900 Ulm-Donau,
West Germany

An infinite source model M/G/1 is considered and analyzed for a dynamic (nonpreemptive) priority discipline: Arriving customers at a queue are assigned urgency numbers. The customer with smallest urgency-number-plus-arrival-time is served next. This service discipline has the effect of distinguishing between customers with different priorities while at the same time taking into account the undesirability of having low priority customers wait too long [1,2,3].

The upper and lower bounds on equilibrium mean waiting times of reference [3] are distinctly improved, and an approximation is introduced which yields good agreement with simulation results. From a heavy traffic solution, new insights are gained into typical properties of this dynamic priority discipline. Examples are given to compare results of this discipline with results of the well-known static priority discipline.

1. THE MODEL

In many real-time computer applications service requests must be processed within specified time limits. A strict observation of such limits is impossible if service requests occur at random. In this case a high (given) probability for meeting all deadlines is required. These deadlines result from the sum of arrival times of individual requests and their permissible wait times. Usually this permissible wait time is called relative urgency v.

Im Skript wird ω statt r verwendet!

We consider the model in Fig.1 in which i types of service requests ($1 \le i \le N$) arrive in independent Poisson arrival processes with parameters λ_i. Service times s_i of i-type requests are i.i.d. random variables with general d.f. $F_i(t) = \mathrm{Prob}(s_i \le t)$ and finite second moment. An i-type service request is characterized by a relative urgency ν_i. All the urgencies are related by the condition

$$0 \le \nu_1 < \nu_2 < \dots \nu_i < \dots < \nu_N. \qquad (1.1)$$

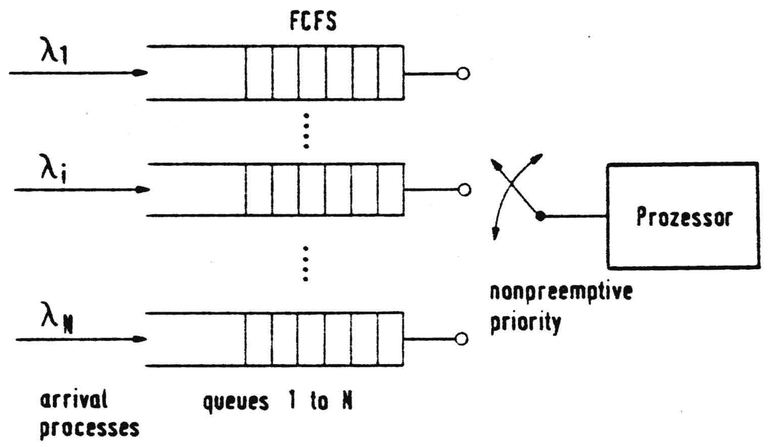


Fig. 1 Model with N queues.

Let the priority $q_i(t)$ of an i-type service request depend on its urgency ν_i, the time of its arrival T_i and its waiting time w_i in the model according to the formula:

$$q_i(t) = \nu_i - t + T_i . \qquad (1.2)$$

Waiting time w_i at time t is defined as

$$w_i = t - T_i . \qquad (1.3)$$

From this relationship we have the dynamic priority of i-type requests at time t to be

$$q_i(t) = \nu_i - w_i . \qquad (1.4)$$

Waiting-time dependent priorities are called dynamic priorities. Service requests in our model are processed by the server without interruptions. Whenever the server has completed a service request, a new request with highest priority $q_i(t)$ (smallest value of $q_i(t)$) is then serviced. This service discipline is known as nonpreemptive relative-urgency (RU) discipline [1,3].

Fig. 2 shows an example with N=2 types of request with urgencies $\nu_1=2$, $\nu_2=4$ and constant and equal service times $\beta_1 = \beta_2 = 3$. The ordinate shows the actual dynamic priority $q_i(t)$ and the abszissa shows the time t. At the arrival time $t=T_i$ each request has a priority $q_i(0) = \nu_i$ because $w_i=0$. The priority of a waiting request increases linearly (the value of $q_i(t)$ decreases) with its waiting time w_i. Service requests arrive at times t = 1,2,4,5,6,8,12. Their individual priority functions are shown. The server occupation by a service request is shown by means of a hatched bar with intervals denoted by $i(T_i)$, the type i and arrival time T_i of an individual request. In cases of equal priority (e.g. for t=4), the most urgent request is serviced first.

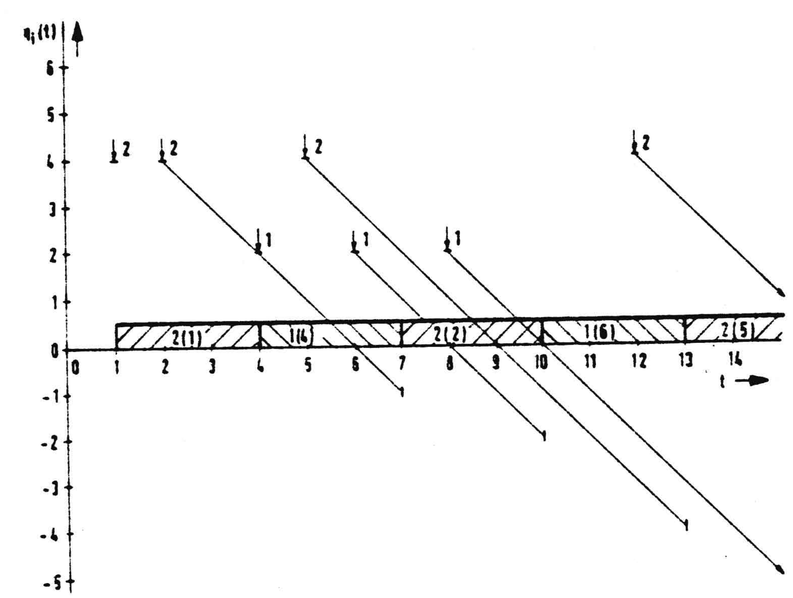


Fig. 2 Example showing the dynamic priority $q_i(t)$ over time t ($N=2$, $\nu_1=2$, $\nu_2=4$, $\beta_1=\beta_2=3$, constant service times). Times of arrival of i-type requests are marked with an indexed arrow $\downarrow i$. The server occupation by an i-type request, which arrived at time T_i, is shown by the hatched bar with intervals denoted by $i(T_i)$.

An analytic solution for the mean waiting time in the RU discipline is known only for a discrete time model (M1), with service times of all requests following the same geometric d.f. and a Bernoulli input process [1,2]. In addition to this exact solution, upper and lower bounds on the mean waiting time are known for a more general model M/G/1 with a Poisson input process and general service time d.f.'s. But these bounds are unsatisfactory.

Two remarkable properties of the RU discipline should be mentioned: (1) If the variance of all urgencies approaches zero, the FCFS discipline is approximated; (2) if the variance of all urgencies approaches infinity, the HOL (head of the line) discipline [4] is approximated. The HOL discipline usually is also called nonpreemptive static-priority discipline.

In section 2 of this paper an approximate computation of the mean waiting time W_i of i-type requests is introduced, and in section 3 improved bounds for the mean waiting time are derived.

The following abbreviations will be used throughout:

W_i	equilibrium mean waiting time of an i-type request
$\rho_i = \lambda_i/\mu_i$	offered traffic of i-type requests
ν_i	urgency of an i-type request
$\rho_{\le i} = \sum_{j=1}^{i} \rho_i$	offered traffic of all j-type requests together ($1 \le j \le i$); probability for a non-idle server if $i=N$
$\lambda_{\le N} = \sum_{i=1}^{N} \lambda_i$	total arrival rate
$F_i(t)$	service-time d.f. of an i-type request
$F_{\le N}(t) = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i F_i(t)$	weighted service-time d.f.
$\beta_{\le N}^{(r)} = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i \beta_i^{(r)}$	r-th moment of $F_{\le N}(t)$
$\beta_{\le N}^{(1)} = \beta_{\le N}$	

$c_{\le N}^2 = \beta_{\le N}^{(2)} / \beta_{\le N}^2 - 1$	squared coefficient of variance of $F_{\le N}(t)$
$W_o = \lambda_{\le N} \beta_{\le N}^{(2)} / 2$	equilibrium mean time to finish a request in service just after arrival of a service request
$W_{FCFS} = W_o/(1-\rho_{\le N})$	equilibrium mean waiting time in the FCFS discipline
$W_{\le N} = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i W_i$	common equilibrium mean waiting time of all service requests

2. APPROXIMATE WAITING-TIME COMPUTATION

We assume that $\rho_{\le N} < 1$ and that equilibrium conditions are satisfied and consider the arrival of a p-type request ($1 \le p \le N$), the tagged request. Upon its arrival, the expected number, $E(n_i)$, of i-type request present in the queue (not being serviced), is [8]

$$E(n_i) = \lambda_i W_i . \tag{2.1}$$

Let f_{ip} represent the expected fraction of these i-type requests which are serviced before the tagged request is. The expected number, $E(m_i)$, of i-type requests which arrive during the expected waiting time, W_p, of the tagged request is

$$E(m_i) = \lambda_i W_p . \tag{2.2}$$

Let g_{ip} represent the expected fraction of these i-type requests which are serviced before the tagged request is.

The mean waiting time W_p of the tagged request can be composed of three components: (1) W_o (see abbreviations above) and two products each consisting of the mean number of i-type requests which will be serviced before the tagged request, multiplied by the mean service time β_i of such requests. The first product (2) consists of those requests which are already waiting when the tagged request arrives and the second product (3) of those requests which arrived after the tagged request but are serviced before it. From these observations, we are able to set up a set of N simultaneous equations, one for each value of p, as follows:

$$W_p = W_o + \sum_{i=1}^{N} E(n_i)\, f_{ip}\, \beta_i + \sum_{i=1}^{N} E(m_i)\, g_{ip}\, \beta_i. \qquad (2.3)$$

All i-type requests already waiting when the tagged request, p, arrives and satisfying $i \leq p$ are serviced before it. From this observation it follows that

$$f_{ip} = 1 \quad \text{for all } i \leq p. \qquad (2.4)$$

All i-type requests which arrive after the tagged request, p, and satisfy $i \geq p$ are serviced after it. From this observation it follows that

$$g_{ip} = 0 \quad \text{for all } i \geq p. \qquad (2.5)$$

Using this information and solving for W_p in eq.(2.3), we obtain

$$W_p = [W_o + \sum_{i=1}^{N} \rho_i W_i + \sum_{i=p+1}^{N} \rho_i W_i f_{ip}]/(1 - \sum_{i=1}^{p-1} \rho_i g_{ip}). \qquad (2.6)$$

Up to this point the derivation in this section is identical to a derivation published by Kleinrock [9] for another dynamic priority discipline.

Derivation of an expression for $E(n_i)f_{ip}$

Let us assume that the tagged request arrives at time t=0. Its priority upon arrival is $q_p(t) = v_p$ (cf. Fig.3). Since w_p is its waiting time, the value of its attained priority at the time it is accepted for service is $q_p(t=w_p) = v_p - w_p$, cf. eq.(1.2). The slope of the priority function is constant and does not depend on the type i of a request. From this it follows that all i-type requests, which have arrived before the tagged request and have been waiting for a time w_i, will be serviced before the tagged request, if and only if their priority is

$$q_i(t=0) < v_p \quad \text{with} \quad q_i(t=0) = v_i - w_i. \qquad (2.7)$$

From this it follows that the arrival time, T_i, of an i-type request (i>p) must satisfy the condition

$$T_i < v_p - v_i. \qquad (2.8)$$

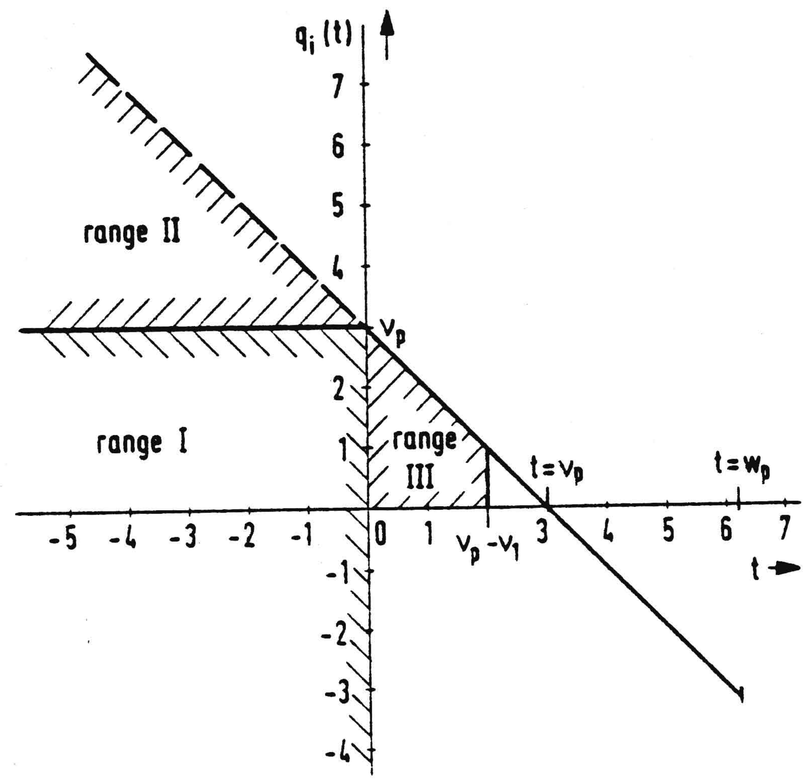


Fig. 3 Priority diagram

Otherwise it will be serviced after the tagged request. The condition in eq.(2.7) can be satisfied by all i-type requests with $i \leq p$ (region I in Fig.3), already considered in eq.(2.4). Other requests which also satisfy this condition are those with urgencies $v_i > v_p$, which satisfy the condition in eq.(2.8) (region II in Fig.3).

The real waiting time $w_i(T_i)$ of less urgent requests (i>p), still waiting at time t=0 and being serviced before the tagged request is

$$|T_i| \leq w_i(T_i) < \infty. \qquad (2.9)$$

The mean number of i-type requests still waiting at time t=0 and being serviced before the tagged request is

$$E(n_i)f_{ip} = \lambda_i \int_{-\infty}^{v_p - v_i} P(t \leq w_i(t) < \infty)\, dt,$$

which is the same as

$$E(n_i)f_{ip} = \lambda_i \int_{v_i-v_p}^{\infty} P(t \le w_i(t) < \infty)dt \qquad (2.10)$$
$$\text{with } i > p.$$

In this expression, $\lambda_i dt$ is the mean number of i-type requests that arrived during the time interval(-t-dt,-t) and P() is the probability that a request, which arrived in that interval, spends at least t time units in the queue. The integration limits in eq.(2.10) result from the eqs.(2.8) and (2.9). Eq.(2.10) can be rewritten as

$$E(n_i)f_{ip} = \lambda_i \int_{v_i-v_p}^{\infty} [1-P(w_i \le t)]dt - \lambda_i \int_{v_i-v_p}^{\infty} [1-P(w_i < \infty)]dt$$

$$= \lambda_i \int_{v_i-v_p}^{\infty} [1-P(w_i \le t)]dt.$$

Observing that

$$E(w_i) = W_i = \int_0^{\infty} [1-P(w_i \le t)]dt$$

we have

$$E(n_i)f_{ip} = \lambda_i W_i - \lambda_i \int_0^{v_i-v_p} P(w_i > t)dt \qquad (2.11)$$
$$\text{with } i > p.$$

Unfortunately the probability $P(w_i>t)$ is unknown. We now introduce two different approximations for the probability $P(w_i>t)$, and are then able to compute the integral in eq.(2.11).

Approximation 1: The probability of waiting is, presuming non-preemptive prioritites, the same for all types of requests [4] and given by

$$P(w_i > 0) = \rho_{\le N}.$$

Remember that the offered traffic, $\rho_{\le N}$, is exactly the probability for the server being non-idle. The assumption that all waiting times, w_i, are greater than the difference $(v_i - v_p)$ results in

$$P(w_i > [v_i - v_p]) = \rho_{\le N} \qquad (2.12)$$

which is called approximation 1 and yields

$$E(n_i)f_{ip} = \lambda_i[W_i - \rho_{\le N}(v_i - v_p)]. \qquad (2.13)$$

In this assumption the service discipline for requests, already waiting when the tagged request arrives is FCFS. Approximation 1 may be used to describe the heavy-traffic condition, $\rho_{\le N} \longrightarrow 1$, of our model. This follows from w_i approaching infinity under this condition.

Approximation 2: Presuming the waiting time of waiting requests being negative exponentially distributed with mean $W_i/\rho_{\le N}$ results in

$$P(w_i > t) = \rho_{\le N}\, e^{-\rho_{\le N} t/W_i}. \qquad (2.14)$$

This waiting time d.f. results from the M/M/1 model using the FCFS discipline. Our model approximates the FCFS model, if the variance of all urgencies approaches zero. The quality of this approximation is discussed in the literature [5] and it has been found to be acceptable.
The factor, $\rho_{\le N}$, in eq.(2.14) exactly describes the probability $P(w_i > 0)$. Using the approximation 2, eq.(2.11) can be written as

$$E(n_i)f_{ip} = \lambda_i W_i\, e^{\rho_{\le N}(v_p-v_i)/W_i}. \qquad (2.15)$$

Derivation of an expression for $E(m_i)g_{ip}$:

Let us assume once more that the tagged request of type p arrives at time t=0. Its priority upon arrival is $q_p(t) = v_p$, cf. Fig. 3. In looking for g_{ip}, we must calculate how many i-type requests arrive after t=0, but before $t=w_p$ and attain a priority $q_i(w_p) \le v_p - w_p$.

Such requests will be serviced before the tagged request. Apparently this condition can be satisfied only by those i-type requests (i < p) which arrive during the time interval $[t=0,\ t=(\nu_p-\nu_i)]$, cf. region III in Fig. 3. This follows from the assumptions made in the eqs. (1.1) and (1.2). The mean number $E(m_i)g_{ip}$ of i-type requests arriving after the time t=0 but being serviced before the tagged request p is given from

$$E(m_i)g_{ip} = \lambda_i \int_0^{\nu_p-\nu_i} P(w_p > t)dt \qquad \text{with } i < p. \tag{2.16}$$

Once again the probability $P(w_p > t)$ is not known, cf. eq.(2.11). Again, we use the same two approximations and obtain for

<u>Approx. 1</u> with $\quad P(w_p > [\nu_p-\nu_i]) = \rho_{\le N}$ (2.17)

which corresponds to eq. (2.12), an expression

$$E(m_i)g_{ip} = \lambda_i(\nu_p-\nu_i)\,\rho_{\le N} \tag{2.18}$$

<u>Approx. 2, eq.(2.14)</u>: $\quad E(m_i)g_{ip} = \lambda_i W_p(1-e^{\rho_{\le N}(\nu_i-\nu_p)/W_p})$. (2.19)

The mean number, E(n), of requests waiting or being serviced is

$$E(n) = \sum_{i=1}^{N} [E(n_i)+\rho_i].$$

Applying eq.(2.1) we obtain

$$E(n) = \sum_{i=1}^{N} \lambda_i(W_i+\beta_i). \tag{2.20}$$

2.1 <u>Application of the heavy traffic approximation 1</u>

From eqs.(2.6), (2.13), (2.18) we have

$$W_p = (W_o + \sum_{i=1}^{p} \rho_i W_i + \sum_{i=p+1}^{N} \rho_i[W_i-\rho_{\le N}(\nu_i-\nu_p])/$$

$$[1-\rho_{\le N} \sum_{i=1}^{p-1} \rho_i(\nu_p-\nu_i)/W_p]$$

$$W_p = W_o + \sum_{i=1}^{N} \rho_i[W_i-\rho_{\le N}(\nu_i-\nu_p)].$$

Using the conservation law [6]

$$\sum_{i=1}^{N} \rho_i W_i = \rho_{\le N}W_o/(1-\rho_{\le N}) \tag{2.21}$$

and setting $\quad W_{FCFS} = W_o/(1-\rho_{\le N})$

we find

$$W_p = W_{FCFS} + \rho_{\le N} \sum_{i=1}^{N} \rho_i(\nu_p-\nu_i) \quad \text{for } \rho_{\le N} \longrightarrow 1. \tag{2.22}$$

From this heavy-traffic solution, two interesting results can be derived:

$$(W_j-W_{j-1})/\rho^2_{\le N} = \nu_j - \nu_{j-1}, \qquad j>1,\ \rho_{\le N} \longrightarrow 1, \tag{2.23}$$

and, defining the mean urgency $\bar{\nu}$ by

$$\bar{\nu} = \rho^{-1}_{\le N} \sum_{i=1}^{N} \rho_i\,\nu_i, \tag{2.24}$$

$$(W_{FCFS} - W_j)/\rho^2_{\le N} = \bar{\nu} - \nu_j, \quad j \ge 1,\ \rho_{\le N} \longrightarrow 1. \tag{2.25}$$

Both results, eqs. (2.23) and (2.25), are concerned with distance measures between mean waiting times. The close relationship to a similar distance measure is noticeable:
Let $Q_j(f)$ and $Q(f)$ be defined as f-quantiles of the waiting-time d.f. of j-type requests and of requests of a FCFS model,

respectively. Jackson [7] has shown that the waiting-time d.f.'s of j-types requests in the model M1, defined in section 1 above, have two properties:

$$Q(f) - Q_j(f) \longrightarrow \bar{\nu} - \nu_j \text{ for } f \longrightarrow 1, \quad 1 \leq j \leq N \tag{2.26}$$

and

$$Q_j(f) - Q_{j-1}(f) \longrightarrow \nu_j - \nu_{j-1}, \text{ for } f \longrightarrow 1, \; j > 1. \tag{2.27}$$

An adequate continuous-time model corresponding to this discrete-time model M1 is a model M/M/1 with RU discipline. The right hand sides of eqs. (2.25) and (2.26) and also eqs. (2.23) and (2.27) are the same. Assuming that Jackson's discrete-time results can be applied to the continuous-time model M/M/1 (which in fact was proposed by Jackson [7] himself), it appears that large mean waiting times and large waiting times of i- and j-type requests in the M/M/1 model differ from each other by the difference of their urgencies $\nu_i - \nu_j$. We might suspect that this should also be true for our model M/G/1. This suspicion is investigated and confirmed in [5].

Eq.(2.12) assumes that all requests arriving before the tagged request of type p are serviced before it, while eq.(2.17) assumes that all i-type requests (i < p) arriving after it, unless its service is started, are also processed before it. In reality not all of these requests are serviced before the tagged request. From this observation one may suspect that, independent of $\rho_{\leq N}$, two relations should hold

$$(W_j - W_{j-1})/\rho^2_{\leq N} \leq \nu_j - \nu_{j-1}, \qquad j > 1 \tag{2.28}$$

$$(W_{FCFS} - W_j)/\rho^2_{\leq N} \leq \bar{\nu} - \nu_j, \qquad j \geq 1. \tag{2.29}$$

A proof of these two relations was not found, but for a number of different examples, these relations have been found to be true.

2.2 Application of the approximation 2

The waiting-time d.f. of i-type requests is assumed to follow a degenerated exponential d.f., eq.(2.14), with mean W_i and squared coefficient of variance

$$c^2 = (2 - \rho_{\leq N})/\rho_{\leq N}.$$

The second moment is approximated by

$$W_i^{(2)} = W_i^2(c^2 + 1),$$

which follows directly the assumption in eq.(2.14). Remember that the upper integration limit in eqs.(2.11) and (2.16) is less than infinity, namely $t = |\nu_i - \nu_p|$. From this it follows that the approximation of the waiting-time d.f. of i-type requests, by the assumption in eq.(2.14), is only of interest for an interval of the waiting time, w_i, given by $0 \leq w_i \leq |\nu_i - \nu_p|$. It also should be mentioned that the probability $P(w_1 > t)$ is not used in our computation, which means that the waiting-time d.f. of 1-type requests may not be approximated.

From eqs.(2.6), (2.15) and (2.19) we can compute an approximation for W_p which is valid over the whole range of $\rho_{\leq N}$

$$W_p = \frac{W_o + \sum_{i=1}^{p} \rho_i W_i + \sum_{i=p+1}^{N} \rho_i W_i \, e^{\rho_{\leq N}(\nu_p - \nu_i)/W_i}}{1 - \sum_{i=1}^{p-1} \rho_i (1 - e^{\rho_{\leq N}(\nu_i - \nu_p)/W_p})}, \text{ for } 0 \leq \rho_{\leq N} < 1$$

and using eq.(2.21) we have

$$W_p[1 - \sum_{i=1}^{p-1} \rho_i (1 - e^{\rho_{\leq N}(\nu_i - \nu_p)/W_p})] =$$

$$W_{FCFS} - \sum_{i=p+1}^{N} \rho_i W_i (1 - e^{\rho_{\leq N}(\nu_p - \nu_i)/W_i}). \tag{2.30}$$

The mean waiting time, W_p, can be computed by solving a system of N simultaneous transcendent equations which can be solved recursively beginning with p=N. For each value of p the unique zero value, W_p, has to be found by numeric methods.

2.3 Examples and comparisons of approximation and simulation results for W_p.

If the variance of all urgencies approaches infinity, our model approximates the HOL discipline. The mean waiting time of requests with priority i for this static priority discipline is [4]

$$W_{iHOL} = \frac{W_o}{(1-\rho_{\leq i})(1-\rho_{\leq i-1})}. \qquad (2.31)$$

The mean waiting time of our model, with all requests having the same urgency, is the same as results from a FCFS model, W_{FCFS}. Let us consider

Example 1: N=4, $\rho_i = \rho_{\leq N}/N$, negative exponential service-time d.f.'s with means $\beta_1=1$, $\beta_2=3$, $\beta_3=5$, $\beta_4=7$, $\beta_{\leq N}=2.4$, $\beta^{(2)}_{\leq N}=19.1$, $C_{\leq N}=1.52$. The urgencies are as given in Fig.5.

It should be mentioned here that the mean waiting times, W_i for this example (with ρ_i being the same for all i) do not denpend on the manner in which the ν_i's are associated with the β_i's. Any other functional dependency (e.g. the same urgencies associated with the inverse order of the β_i's, $\beta_1=7$, $\beta_2=5$, $\beta_3=3$, $\beta_4=1$) yields the same results for the W_i's. Only the common mean waiting time, $W_{\leq N}$, depends on it.

For a static priority model, the mean waiting times, W_i, differ more, the greater the parameter $\rho_{\leq N}$ (cf. Fig.4) is. Even in the case of saturation ($\rho_{\leq N} \geq 1$) the mean waiting times of high priority requests remain small. For a dynamic priority model, it was alread mentioned that the mean waiting times W_i should differ very little from the mean waiting time W_{FCFS} if the variance of the urgency numbers ν_i is small, and should approach the mean waiting times W_{iHOL} if this variance is large. Upper and lower bounds on the times W_i are given by W_{4HOL} and W_{1HOL}.

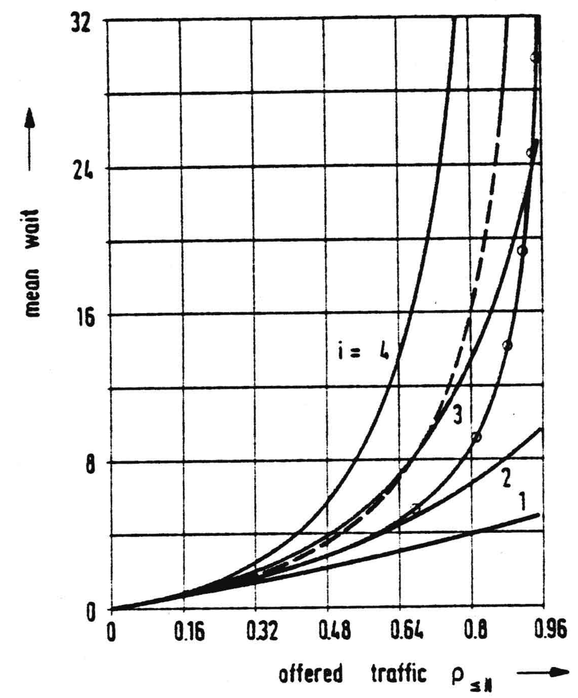
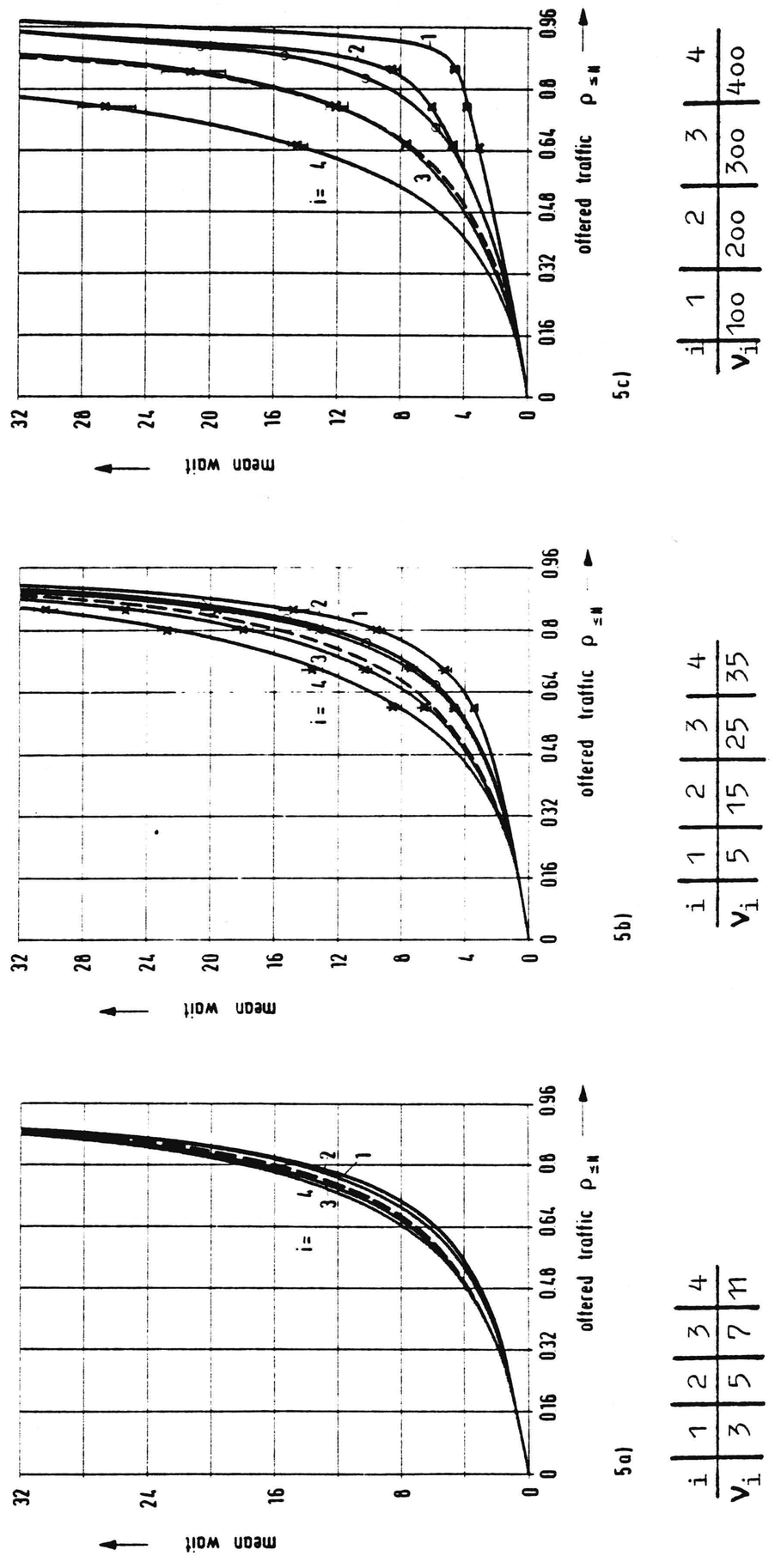


Fig. 4 Mean waiting times over $\rho_{\leq N}$ for example 1. The lines result from a static priority model, the broken line results from the FCFS model. The common mean waiting time $W_{\leq N}$ is shown by the line marked with circles.

2.3.1 Results for the approximation 2

The results in Fig. 5 show the mean waiting time, W_i, as computed from eq.(2.30) for the parameters in example 1. In Fig. 5a only small differences $\nu_i - \nu_{i-1}$ are presupposed; therefore, the mean waiting times, W_i, are close to the curve for W_{FCFS}. In Fig. 5b a greater difference, $\nu_i - \nu_{i-1} = 10$, is assumed which results in a greater deviation of the W_i's from W_{FCFS}. In Fig. 5c a much greater difference, $\nu_i - \nu_{i-1} = 100$, is assumed and from this example the degree of the approximation of the W_{iHOL}-curves can be observed by comparing Fig. 4 to Fig. 5c. The static and dynamic priority models differ in that all mean waiting times of the latter model approach infinity if the offered traffic, $\rho_{\leq N}$, approaches 1 whereas this is not the case for the static priority model.



i	1	2	3	4
ν_i	3	5	7	11

i	1	2	3	4
ν_i	5	15	25	35

i	1	2	3	4
ν_i	100	200	300	400

Fig. 5: Approximation 2 of the mean waiting times of example 1. The lines result from the dynamic priority model; the broken lines result from the FCFS model. The common mean waiting time $W_{\leq N}$ is shown by the line marked with circles. Simulation results, marked by x, are shown for a 95 % level of confidence. Three different sets of urgency numbers ν_i are assumed for the three different figures a), b) and c).

This example shows that the relation $\nu_i \gg \beta_i$ should be observed (e.g. $\nu_i = 5\,\beta_i$); otherwise, the mean waiting times would differ only very little from the time W_{FCFS}. The common mean waiting time, $W_{\leq N}$, is shown by the curve marked with circles. In example 1 the mean service times, β_i, obey the relation $\beta_i < \beta_{i+1}$. In this case the common mean waiting time is minimized in a static priority model [6] (cf. Fig.4). In a dynamic priority model the curve $W_{\leq N}$ approaches the curve W_{FCFS} closer and closer as the variance of the urgency numbers ν_i becomes smaller and smaller. In this example $W_{\leq N}$ is greater in the dynamic priority model than in the static priority model. If we had chosen the opposite ordering of the β_i's ($\beta_1=1$, $\beta_2=3$, $\beta_3=5$, $\beta_4=7$), the dynamic priority model would be superior with respect to a minimal $W_{\leq N}$. Simulation results with a 95 % level of confidence shown in the figures 5b, c confirm our approximation results very well. Other examples show the same tendency (cf.[5]). It should be mentioned that our approximate solution also takes into account that the service-time d.f.'s may have different coefficients of variance (cf. the computation of W_o).

2.3.2 Results for the approximation 1

Approximation 1 yields a heavy traffic solution for the mean waiting time W_i. In just this case, $\rho_{\leq N} \longrightarrow 1$, the models with static and dynamic priorities differ greatly, as was shown in section 2.3.1. If we use approximation 1 [eqs. (2.12), (2.17)] with decreasing offered traffic, starting from $\rho_{\leq N} \longrightarrow 1$, until the mean waiting time, W_i, becomes smaller than $(\nu_i - \nu_{i-1})$ — a somewhat arbitrary limit — then we obtain for our example 1 the curves shown in Fig. 6a, b at the rigth-hand sides of the hatched areas. A comparison with the results in Fig. 5 reveals that the upturn of the curves to the pole at $\rho_{\leq N} = 1$, which is typical for dynamic priorities, is approximated very well. However the individual curves in Fig. 6 differ too much from each other if $\rho_{\leq N} < 1$; the improved results of approximation 2 in Fig. 5 differ remarkably less from the waiting time W_{FCFS} and from each other.

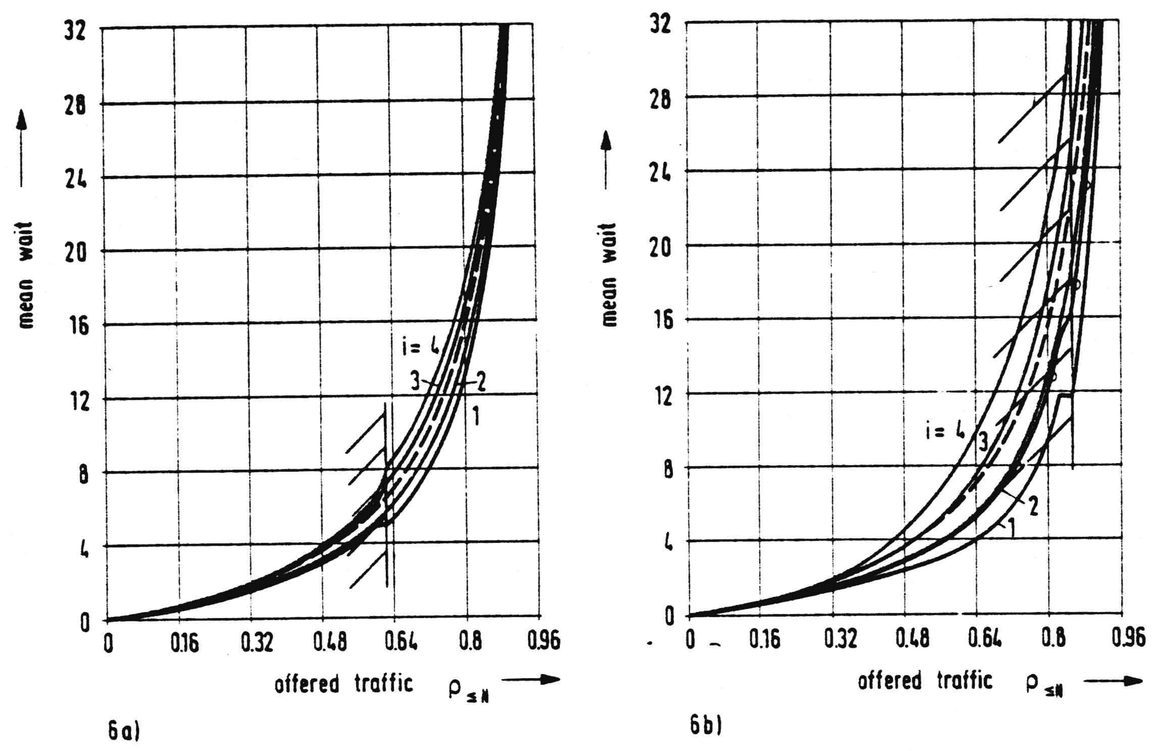


Fig. 6: Mean waiting times of example 1. On the left and right hand side of the limit marked by ⋮ results of the approximation 2 and 1, respectively, are shown. The urgency numbers in Fig. a) and Fig. b) are the same as in the Figs. 5a, 5b.

This comparison shows also that the single curves W_i in Fig. 6 differ more from the curve W_{FCFS}, than they do for the improved approximation 2 in Fig. 5. With regards to the distances $(W_i - W_{FCFS})$, the results of approximation 1 in Fig. 6 can be viewed as bounds: Each curve for W_i is in reality closer to curve W_{FCFS} than that resulting from approximation 1. This statement is already formulated in the relations eqs. (2.28) and (2.29).

3. BOUNDS FOR THE MEAN WAITING TIME

The results of both approximations introduced in section 2 differ more or less from the unknown exact results depending on the parameters of our model. It can be shown by simulation that the approximation results of section 2.3.1 are very good. One example of this was given in Fig. 5.
However, if the exact results are unknown, it is generally advantageous to have close upper and lower bounds. Bounds are known from

a paper by Holtzmann [3]. However, these bounds cannot be called close bounds. These bounds can be improved distinctly as will be shown in section 3.1. First, we will discuss the bounds from literature [3] by means of an example. The abbreviations of section 1 apply here and steady state equilibrium conditions are assumed to hold. From [3] we repeat the results for two upper limits $\bar{W}_p$, $\bar{\bar{W}}_p$ and two lower limits $\underline{W}_p$, $\underline{\underline{W}}_p$ for the mean waiting time W_p of p-type requests in a dynamic priority model:

$$W_p \leq \bar{W}_p = W_{FCFS} - \sum_{i=p+1}^{N} \rho_i \int_0^{\nu_i-\nu_p} g(t)dt + \sum_{i=1}^{p-1} \rho_i \, h(\bar{W}_p, \nu_p-\nu_i), \quad (3.1)$$

$$W_p \geq \underline{W}_p = W_{FCFS} - \sum_{i=p+1}^{N} \rho_i \min\{\bar{W}_i, \nu_i-\nu_p\} + \sum_{i=1}^{p-1} \rho_i \int_0^{\nu_p-\nu_i} g(t)dt, \quad (3.2)$$

$$W_p \geq \underline{\underline{W}}_p = \rho_p^{-1} \left[\rho_{\leq N} W_{FCFS} - \sum_{i=1, i\neq p}^{N} \rho_i \bar{W}_i\right], \quad (3.3)$$

$$W_p \leq \bar{\bar{W}}_p = \rho_p^{-1} \left[\rho_{\leq N} W_{FCFS} - \sum_{i=1, i\neq p}^{N} \rho_i \underline{W}_i\right] \quad (3.4)$$

with $\quad g(t) = \rho_{\leq N} - \lambda_{\leq N} \int_0^t [1-F_{\leq N}(\tau)]d\tau$

$$h(\bar{W}_p, \nu_p-\nu_i) = \begin{cases} \nu_p-\nu_i, & \text{if } \nu_p-\nu_i \leq \bar{W}_p \\ \bar{W}_p[1+\ln(\nu_p-\nu_i)-\ln\bar{W}_p], & \text{if } \bar{W}_p < \nu_p-\nu_i. \end{cases} \quad (3.5)$$

For our example 1 these equations result in the broken-line curves shown in the figures 7a to 7d. Only the least upper bounds $\bar{W}_p$ or $\bar{\bar{W}}_p$ and the greatest lower bounds $\underline{W}_p$ or $\underline{\underline{W}}_p$ are shown. Some upper bounds are limited by W_{4HOL} and all the lower bounds are limited in a large range of $\rho_{\leq N}$ by W_{1HOL}, both being mean waiting times of a static priority model, cf. eq. (3.31). The difference between upper and lower bounds is too large to be satisfactory.

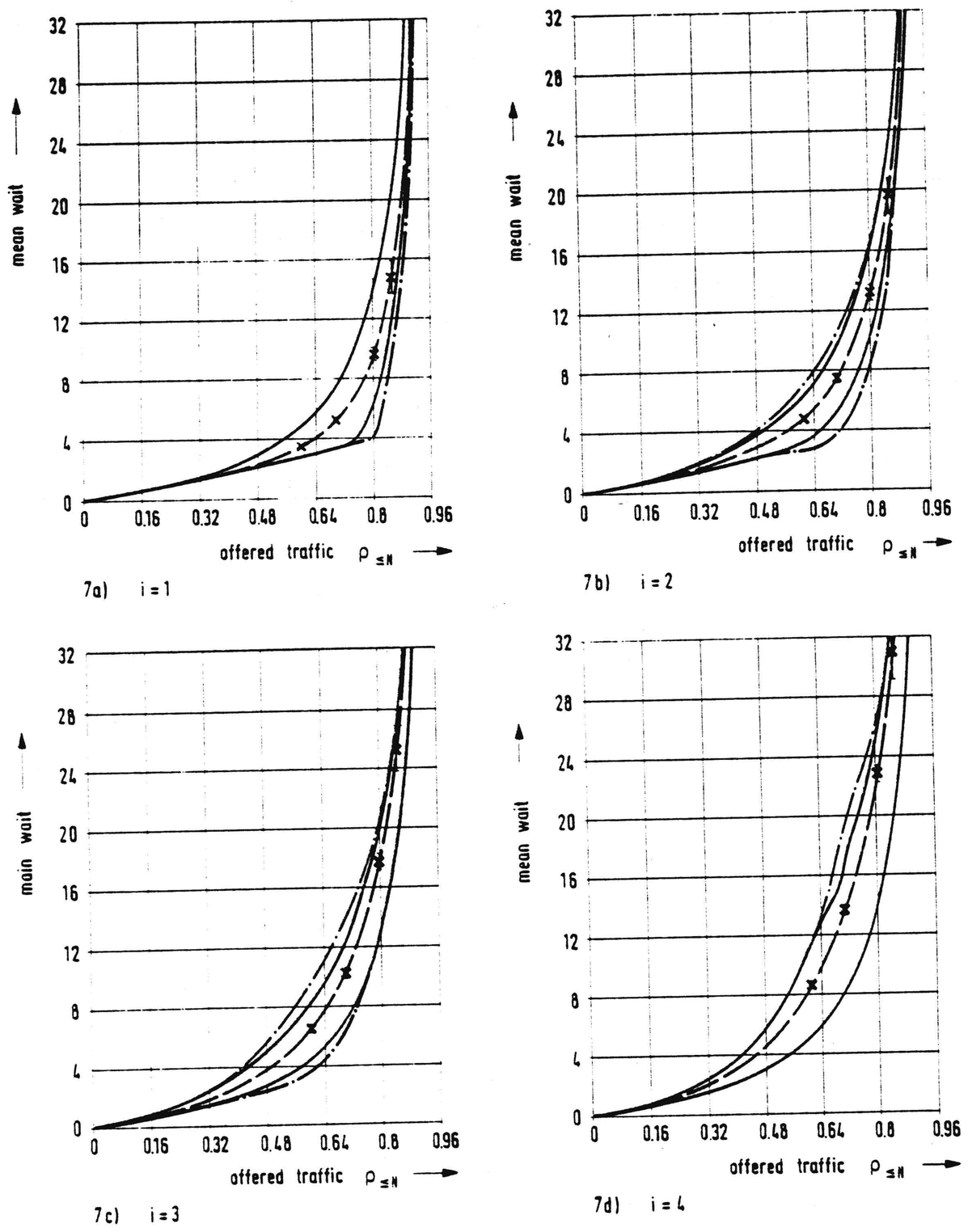


Fig. 7: Holtzmann bounds (centre lines), improved bounds (solid lines) and approximation 2 (broken line) for the mean waiting time W_i of example 1. The ν_i's are as in Fig.5b. Simulation results, marked by x, are shown for a 95 % level of confidence.

3.1 Improved bounds

The upper bound $\bar{W}_p$ is computed by Holtzmann [3] using two relations which can be sharpened. We consider the computation of $E(m_i)g_{ip}$, cf. eq.(2.16), from

$$E(m_i)g_{ip} = \lambda_i \int_0^{\nu_p-\nu_i} P(w_p > t)dt. \qquad (3.6)$$

The probability $P(w_p > t)$ was approximated successfully in section 2, eq.(2.14).
Now we are not looking for an approximation to this probability, but rather for a close upper limit for it. From the considerations which lead to eq. (2.14) it follows that independent of the value of t the following relation must hold:

$$P(w_p > t) \leq \rho_{\leq N}. \qquad (3.7)$$

Instead of this, Holtzmann uses the relation $P(w_p > t) \leq W_p/t$ and derives from this an upper limit

$$P(w_p > t) = \begin{cases} 1 & \text{for} \quad 0 \leq t \leq W_p \\ W_p/t & \quad\quad W_p < t. \end{cases} \qquad (3.8)$$

This function, $P(w_p > t)$, is shown over t as a broken line in Fig. 8 for an example with $W_i = 3$, $\rho_{\leq N} = 0.5$. Considering eq.(3.7), we may write down a sharper bound than that given in eq. (3.8):

$$P(w_p > t) = \begin{cases} \rho_{\leq N} & \text{for} \quad 0 \leq t \leq W_p/\rho_{\leq N} \\ W_p/t & \quad\quad W_p/\rho_{\leq N} < t. \end{cases} \qquad (3.9)$$

This improved bound is also shown in Fig. 8 as a centre line (-.-.-). Both limits are indentical if the relation $W_p/\rho_{\leq N} < t$ holds; otherwise the new limit is an improvement. It is known that the real waiting time d.f. starts for t=0 at $P(w_p > t) = \rho_{\leq N}$ and that the centre line is an upper limit for $t > 0$. Using the improved limit, eq. (3.8), in computing $E(m_i)g_{ip}$, eq.(3.6), the shaded area in Fig. 8 is partially or completely omitted, depending on W_p. Instead of eq. (3.5) we may now write

$$h(\bar{W}_p, \nu_p-\nu_i) = \begin{cases} \rho_{\le N}(\nu_p-\nu_i), & \text{if } (\nu_p-\nu_i)\rho_{\le N} \le \bar{W}_p \\ \bar{W}_p[1+\ln(\rho_{\le N}(\nu_p-\nu_i)/\bar{W}_p)], & \text{if } (\nu_p-\nu_i)\rho_{\le N} > \bar{W}_p. \end{cases} \tag{3.10}$$

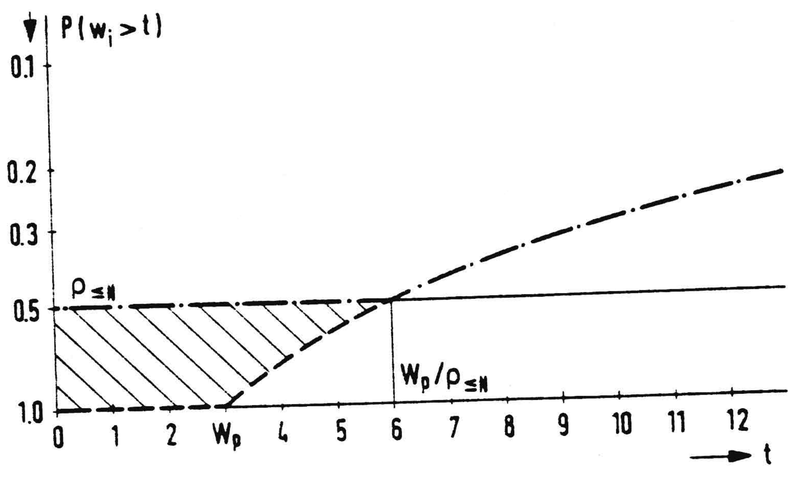


Fig. 8: Comparison of two bounds for the waiting-time d.f. of i-type requests for the parameters $\rho_{\le N} = 0.5$, $W_i = 3$.

For the numeric calculation of $\bar{W}_p$, it is advantageous to write instead of eqs. (3.1) and (3.10):

$$\bar{W}_p = W_{FCFS} - \sum_{i=p+1}^{N} \rho_i \int_0^{\nu_i-\nu_p} g(t)dt + \bar{W}_p \sum_{i=j}^{p-1} \rho_i[1+\ln \frac{\rho_{\le N}(\nu_p-\nu_i)}{\bar{W}_p}]$$

$$+ \rho_{\le N} \sum_{i=1}^{j-1} \rho_i(\nu_p-\nu_i) \tag{3.11}$$

with j such that $\bar{W}_p < \rho_{\le N}(\nu_p-\nu_i)$ for $1 \le j \le p-1$.

Again, we have a set of N simultaneous transcendent equations for the unknown variable, W_p. The calculation of W_p can be greatly simplified, with only a small deterioriation in the bound, if eq. (3.7) is used inplace of eq. (3.9).

$$\bar{W}_p' = W_{FCFS} - \sum_{i=p+1}^{N} \rho_i \int_0^{\nu_i-\nu_p} g(t)dt + \rho_{\le N} \sum_{i=1}^{p-1} \rho_i(\nu_p-\nu_i). \tag{3.12}$$

This upper limit $\bar{W}_p'$ can also be used to watch over the computational results of $\bar{W}_p$ from eq.(3.11). The relation $\bar{W}_p' \ge \bar{W}_p$ must hold.

Eq.(3.9) can also be used to improve the result for $E(n_i)f_{ip}$, cf. eq.(2.11). Using eq.(2.11) and eq.(3.9) we have

$$E(n_i)f_{ip} \ge \lambda_i[W_i - \rho_{\le N}(\nu_i-\nu_p)] \tag{3.13}$$

instead of $$E(n_i)f_{ip} \ge \lambda_i[W_i-(\nu_i-\nu_p)], \tag{3.14}$$

which is used in [3]. Using eq.(3.13), an improved lower bound $\underline{W}_p$ for the mean waiting time W_p results

$$W_p \ge \underline{W}_p = W_{FCFS} - \sum_{i=p+1}^{N} \rho_i \min\{\bar{W}_i, \rho_{\le N}(\nu_i-\nu_p)\} + \sum_{i=1}^{p-1} \rho_i \int_0^{\nu_p-\nu_i} g(t)dt. \tag{3.15}$$

The bounds $\bar{W}_p$ and $\underline{W}_p$ still may be computed applying the eqs.(3.3) and (3.4) but substituting $\bar{W}_i$ and $\underline{W}_i$ from the eqs. (3.11) and (3.15). The lines in Fig. 7 show the improved bounds. Apparently, some bounds according to literature [3] are improved distinctly. The bound $\underline{W}_4$ is improved only insignificantly in this example. Generally, for a dynamic priority queue with N different types of requests, all bounds, except the bound $\bar{W}_1$ are improved.

In addition to the old and new bounds and the approximation for W_p from eq. (2.30), simulation results are also shown in Fig. 7 and are marked by an x. It may be observed that the difference between the improved bounds, $\bar{W}_p$ and $\underline{W}_p$, remains unsatisfactorily large and that the simulation results confirm our approximation results for W_p which are always enclosed by these bounds.

Tests with general service-time d.f.'s applied to our model have demonstrated that the difference between the old and improved bounds can be substantially greater than shown in Fig. 7.

3.2 Comparison of improved bounds with particularly narrow bounds computed for a special case

Holtzmann [3] discusses the possibility of improving his own bounds under special circumstances and demonstrates this for a simple Example 2: N=2, $\rho_i = \rho_{\leq 2}/2$, negative exponential service-time d.f.'s with parameters $\beta_1 = \beta_2 = 1$, $\beta_{\leq 2} = 1$, $\beta^{(2)}_{\leq 2} = 2$, $C_{\leq 2} = 1$, $\nu_1 = 0$, ν_2 = variabel.

By means of Laplace transforms of waiting times of a static priority model, improved bounds are shown to be computable. The mathematical tools used are highly sophisticated, e.g. the inversion of a transformation required must be performed by an approximation. This technique for computing improved bounds does not seem to be generally applicable.

In Fig. 9, for example 2, the simple Holtzmann bounds, cf. eqs.(3.1) to (3.5), and his special improved bounds, $\bar{\bar{W}}_p$ and $\underline{\underline{W}}_p$, are shown. The unknown mean waiting times are limited to the shaded areas. In addition the general improved bounds, cf. eqs. (3.11) and (3.15), are introduced and shown by a centre line. It can be seen that the new bounds and Holtzmann's special bounds differ only slightly. It is shown that the best possible special bounds known are not substantially better than our improved bounds. The bounds $\bar{W}_1$ and $\underline{W}_2$ remain unchanged.

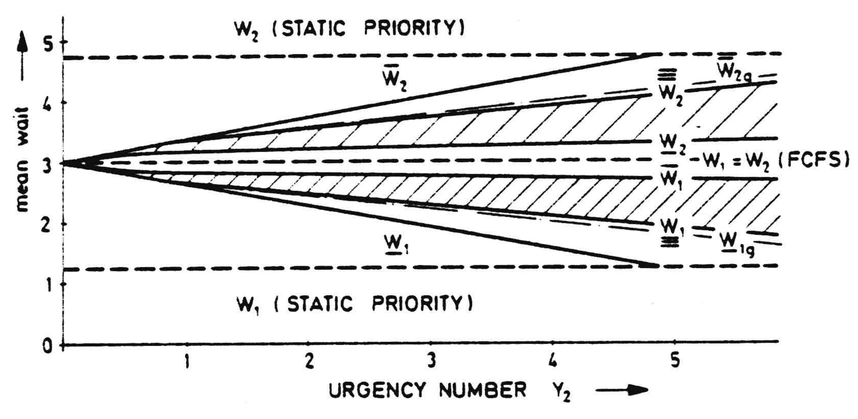


Fig. 9: Simple Holtzmann bounds $\underline{W}_i$ and $\bar{W}_i$, specially improved Holtzmann bounds $\bar{\bar{W}}_2$ and $\underline{\underline{W}}_1$ and generally improved bounds $\bar{W}_{2g}$ and $\underline{W}_{1g}$ (centre line, eqs.(3.11, 3.15) for example 2 and an offered traffic $\rho_{\leq N} = 0.75$.

4. FINAL REMARKS

The model considered in this paper is important for real-time computers in industrial control applications because in such situations scheduling by means of due dates of service requests seems to be much more appropriate than other scheduling algorithms, currently in use. It can be viewed as a disadvantage of our model that only nonpreemptive priorities are permitted. From this results a limitation of the variance of the mean waiting times of different-type requests, possible in this model. This variance can be expected to be somewhat smaller than in the same model using preemptive-resume priorities. Assume that a preemption is permissible only if a waiting service request has and another request being serviced has not reached its dealine. Then the preemptive-priority case can be expected to be solveable approximately in a similar way observing that the probabilities of waiting for service requests with urgencies $\nu_i > 0$ still remain $\rho_{\leq N}$, independently of the type i. The service time d.f.'s of i-type requests, i>1, must be negative exponential and for i=1 of the general type with means β_i.

Acknowledgements

The simulations mentioned in this paper have been performed by Mr. Dipl.-Ing. W. Rosenbohm. I am indebted to him for contributing to these results.
This work was partly supported by the 2nd EDP Program of the Federal Government of West Germany.

References

[1] J.R. Jackson: Queues with dynamic priority discipline Management Science, Vol.8, No.1, Oct. 1961, pp.18-34.

[2] J.R. Jackson: Some problems in queueing with dynamic priorities. Naval Research Logistics Quarterly, Washington D.C,7, Sept.1960, pp.235-249.

[3] J.M. Holtzmann: Bounds for a dynamic priority queue. Oper. Research (USA)1970, pp.461-468.

[4] A. Cobham: Priority assignment in waiting line problems. Oper. Research (USA), 2, 1954, pp. 70-76.

[5] W. Rosenbohm, B. Walke: Properties of the waiting-time distribution of a dynamic priority queue. To be published.

[6] L. Kleinrock: Queueing Systems, Vol.I: Theory John Wiley & Sons, 1975, New York/London/Sydney/Toronto.

[7] J.R.Jackson: Waiting-Time distributions for queues with dynamic priorities. Naval Research Logistics Quarterly, Washington, D.C,9March 1962, pp.31-36.

[8] J.D.C. Little: A proof for the queueing formula $L=\lambda W$. Oper. Research, 9 (1961), pp.383-387.

[9] L. Kleinrock: A delay dependent queue discipline. Naval Research Logistics Quarterly, Washingt. D.C., 11, 1964, No.4, pp.329-341.