Performance of Computer Systems
M. Arato, A. Butrimenko, E. Gelenbe (eds.)
©IIASA, North-Holland Publishing Company, 1979

WAITING-TIME DISTRIBUTIONS FOR DEADLINE-ORIENTED SERVING *

B. Walke, W. Rosenbohm
AEG-TELEFUNKEN, Research Institute Ulm,
West Germany

An infinite queue single server model is considered where requests of type i, $1 \leq i \leq N$ arrive from independent Poisson streams and demand service according to arbitrary d.f.'s which may be different for different types of requests. Associated with each type-i request is an urgency number V_i which, together with the request's time of arrival, defines a deadline for beginning its service. Preemption of a request in service is not permitted. This relative urgency (RU) discipline has at its two limiting cases the FCFS and head of the line (HOL) disciplines. In [1] the mean waiting time is computed approximately and close bounds are derived there. Here we derive close approximations for the tails of the waiting time d.f.'s and compare them to those of the HOL and FCFS disciplines. Moreover, some properties of the preemptive version of the RU discipline are presented.

1. INTRODUCTION AND DEFINITION OF MODEL

In real-time computer-control systems it is important that incoming service requests from a running process be completed in time, i.e. within a given time limit. In this paper we will consider the case of requests arriving at random (not predictable) times. Given this randomness it is impossible to guarantee that all finite deadlines for beginning or ending these requested services can be met. The best that can be done is to guarantee a high probability for meeting such deadlines.

Let us consider a computer model with unlimited waiting space for incoming requests which may possess various properties (to be denoted in the following as "types"). Further suppose these type-i requests ($1 \leq i \leq N$) arrive at a rate λ_i from a Poisson process, have an arbitrary service-time b_i according to a distribution function $F_i(t) = P_i(b_i \leq t)$ with a finite second moment, and that each request has a deadline t_i for beginning based on its arrival time T_i and its urgency V_i. Without any loss of generality we order the types i of requests such that

$$0 \leq V_1 < V_2 < \ldots < V_i < \ldots < V_N \tag{1.1}$$

where N is the total number of types and V_i is the urgency of a type-i request. The smaller the value of V_i, the higher the urgency.

Figure 1 shows a model of our system. We will be concerned almost exclusively with the case of requests which, once servicing has begun, may not be interrupted (non-preemptive priority). After a request has been serviced, the request among those waiting to be processed which has the highest dynamic priority $q_i(t)$ is chosen for servicing. The dynamic priority at time t is given by

$$q_i(t) = V_i - t + T_i \ , \tag{1.2}$$

where T_i is the arrival time and V_i the urgency. If we define the waiting time

* This work was partly supported by the 2nd EDP Program of the Federal Government of West Germany.

$w_i(t)$ at time t as

$$w_i(t) = t - T_i \tag{1.3}$$

then the current priority is

$$q_i(t) = V_i - w_i(t) \tag{1.4}$$

and depends linearly only on the current waiting time. As usual small values of $q_i(t)$ indicate a high priority. It is interesting to note that in this model the priorities of all outstanding requests, regardless of type, grow uninformally. This is illustrated in [1] with an example. Requests of the same type are processed in the order they arrived (first come first serve FCFS). Since requests are serviced according to individual urgencies and arrival times, we speak of a relative-urgency (RU) discipline.

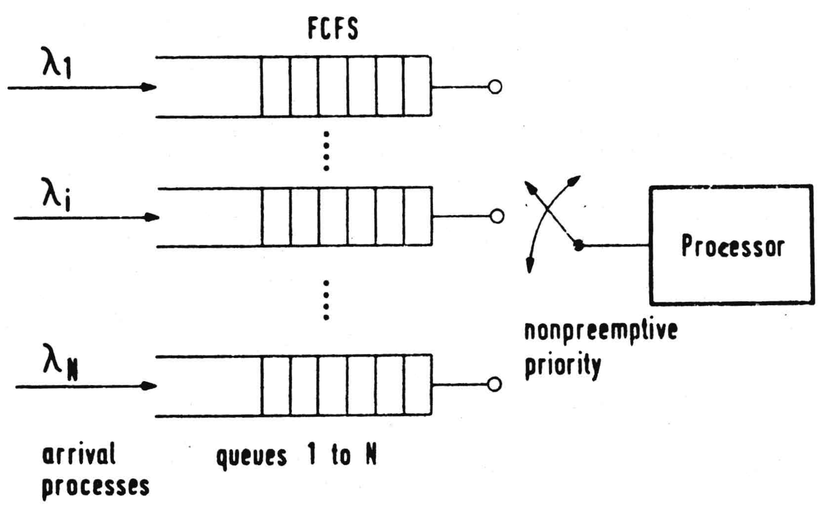

Fig. 1: Model of a real-time computer-control system λ_i = arrival rate.

In addition to the non-preemptive discipline (RU-NONPRE) we will consider a few examples of a preemptive RU discipline (RU-PRE). In the RU-PRE discipline a type-i request may be interrupted (preempted) by a type-j request if $w_i < V_i$ and $w_j = V_j$. The priority $q_i(t)$ of a request remains unchanged once its servicing has begun. But if a preemption occurs, the priority of the interrupted request again follows Eq.(1.4) with $w_i(t)$ being the total time such a request remains in the system without being serviced. The urgency V_i represents the time w_i a request is willing to wait from its arrival T_i to its deadline t_i. For both the non-preemptive and preemptive versions of the RU discipline a request's deadline may be defined to be reached if $w_i(t)$ equals V_i. If the actual waiting time w_i of a type-i request is less than its voluntary waiting time V_i, while another type-j request has already waited its voluntary time V_j, then the type-j request preempts the type-i request. The case i = j is impossible. Requests, whose servicing just barely began on time or which missed their deadlines may not be interrupted. Summarizing, it can be said that in the RU-PRE discipline only requests whose servicing began ahead of schedule may be interrupted and only by requests whose deadline has been reached.

The following abbreviations are used throughout:

w_i, b_i	Random variables: waiting and service time of a type-i request
d.f., d.f.'s, w.t.d.f.	distribution function, d. functions, waiting time d.f.
$W_i(t) = P(w_i \le t)$	w.t.d.f. of a type-i request
$W_i^{(r)}$, $W_i^{(1)} = W_i$	r-th moment of $W_i(t)$
$\lambda_{\le N} = \sum_{j=1}^{N} \lambda_j$	total arrival rate of all requests together
$W_{\le N}(t) = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i W_i(t)$	common w.t.d.f.

$F_i(t) = P(b_i \le t)$	service time d.f.
$\beta_i^{(r)}$, $\beta_i^{(1)} = \beta_i$	r-th moment of service time d.f.
$\rho_i = \lambda_i \beta_i$	offered traffic of type-i requests
$\rho_{\le N} = \sum_{j=1}^{N} \rho_j$	total offered traffic
$F_{\le N}(t) = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i F_i(t)$	common 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 $F_{\le N}(t)$
$C_{\le N}^2 = \beta_{\le}^{(2)} / \beta_{\le N}^{(2)} - 1$	squared coefficient of variance of $F_{\le N}(t)$
$\bar{V} = \rho_{\le N}^{-1} \sum_{i=1}^{N} \rho_i V_i$	mean weighted urgencies
$Var(V) = \sum_{i=1}^{N} \frac{\rho_i}{\rho_{\le N}} V^2 - \bar{V}^2$	variance of the urgency numbers.

The RU discipline uses dynamic priorities based on waiting times. The problem described at the beginning of this paper nowadays is usually treated with a service discipline which recognizes only static priorities. Arriving requests are assigned a static priority i according to their type i. Among the waiting requests the one with the highest priority (lowest i) which has waited the longest of all requests of the same type i (FCFS) is processed first. Here, too, the model in Figure 1 applies, except that there is no urgency V_i.

If one is interested in the probability of missing deadlines, then one should know the waiting time d.f. (distribution function) $W_i(t) = P(w_i \le t)$. From this the probability that the waiting time w_i is not greater than t can be obtained. The properties of static and dynamic priorities with regard to the divergence of the resulting waiting times can be discussed by comparing their waiting time d.f.'s. Such a comparison shows how much the probabilities for meeting given deadlines $t = V_i$ differ between static and dynamic priorities. Since this comparison depends on the type of request considered, the sum of the difference $[W_{istat}(V_i) - W_{idyn}(V_i)]$ between static and dynamic priorities weighted by the probability of occurance $\lambda_i/\lambda_{\le N}$ of type-i requests (which is independent of discipline) is useful. In this manner one obtains a formal figure of comparison V:

$$V = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i [W_{istat}(V_i) - W_{idyn}(V_i)] \quad . \tag{1.5}$$

From this we see that:

if $V < 0$, then a relativ urgency discipline is better, and
if $V > 0$, then a discipline with static priorities is better. (1.6)

From the definition of V we have $-1 \le V \le 1$.

We will consider the model in Figure 1 only in a state of equilibrium, which occurs only when the total offered traffic $\rho_{\le N} < 1$.

In disciplines with static priorities, a stationary equilibrium is achieved for requests with high priorities, sometimes even when $\rho_{\le N} > 1$. However, we will not consider this case here.

2. WAITING TIME D.F.'s FOR VARIOUS DISCIPLINES, EXAMPLES

Let us assume N different types of requests, each with its arrival rate λ_i, urgency V_i, and service time d.f. $F_i(t)$. Since the waiting time d.f. cannot be calculated, we will consider four examples where were simulated.

Example 1: Modell M/M/1, N = 4,

$\rho_i = \rho_{\leq N} / N$, $\beta_{\leq N} = 2.386$, $C_{\leq N} = 1.52$, $\bar{V} = 20$, $Var(V) = 125$.

All requests have a negative exponential distributed service time with mean β_i.

i	1	2	3	4
β_i	1	3	5	7
V_i	5	15	25	35

Example 2: Modell M/D/1, N = 4,

$\rho_i = \rho_{\leq N} / N$, $\beta_{\leq N} = 2.386$, $C_{\leq N} = 0.81$, $\bar{V} = 30$, $Var(V) = 125$.

All requests have a constant service time.

i	1	2	3	4
β_i	1	3	5	7
V_i	15	25	35	45

Example 3: Modell M/G/1, N = 4,

$\rho_i = \rho_{\leq N} / N$, $\beta_{\leq N} = 1.86$, $C_{\leq N} = 2.96$, $\bar{V} = 20.5$, $Var(\bar{V}) = 94.25$.

Service time d.f.'s of type-1 and type-4 requests are hyperexponential of order 2. Requests of type 2 and 3 have constant and negative exponential distributed service times, resp.

i	1	2	3	4
β_i	2.45	1	3	2.45
d.f.	H_2	D	M	H_2
V_i	9	14	25	34

Example 4: Corresponds to example 2 but with inverse order of the mean service time β_i with regard to the type-i.

i	1	2	3	4
β_i	7	5	3	1
V_i	15	25	35	45

The examples given here differ in part markedly in the service time d.f.'s assumed for the individual types of requests and in the coefficients of variance $C_{\leq N}$ of the common service time d.f.

From an earlier investigation (cf. [1] and [4]) it is known that the smaller the variance of the urgency numbers is, the less the mean waiting times W_i of the RU-NONPRE discipline differ from the mean waiting time

$$W_{FCFS} = 1/2\ \lambda_{\leq N}\, \beta^{(2)}_{\leq N} / (1 - \rho_{\leq N}) \tag{2.1}$$

in the first come first serve discipline. (For $\lambda_{\leq N}$, $\beta^{(2)}_{\leq N}$ see the list of abbreviations). On the other hand if the variance $Var(V)$ is very large, the RU-NONPRE discipline produces mean waiting times approximating those of the discipline with static non-preemptive priorities

$$W_{iNONPRE} = \frac{\lambda_{\leq N}\, \beta^{(2)}_{\leq N}}{2(1-\rho_{\leq i})(1-\rho_{\leq i-1})} \quad . \tag{2.2}$$

Laplace-Stieltjes transforms of the waiting time d.f.'s for these extreme cases are known and their first two moments can easily be calculated.

2.1. Waiting time d.f.'s of the RU-NONPRE discipline

Let us start with the results obtained: The simulation shows that the waiting time d.f.'s when plotted semilogarithmically, asymptotically approach parallel straight lines regardless of the example and the individual parameters. Let $p(w_i)$ be the

equilibrium probability that a type-i request must wait no longer than w_i units of time for service. For $0 \leq f \leq 1$, define

$$Q_i(f) = \inf\{w_i \mid p(w_i) \geq f\} \quad . \tag{2.3}$$

(This is the waiting time corresponding to fractile f in the cummulative waiting-time d.f. of type-i requests. $Q_i(f)$ usually is called the f-quantile.)
Then the results of the simulation of all four examples seem to indicate that for every M/G/1/RU-NONPRE model we have

$$\lim_{f \to 1} [Q_i(f) - Q_j(f)] = V_i - V_j \quad . \tag{2.4}$$

As the probability f approaches 1, the distance between the w.t.d.f.'s for type-i and type-j requests asymptotically approaches the difference in their urgencies V_i and V_j within the accuracy obtainable by simulation. For $f < 1$, this distance is smaller than the difference $V_i - V_j$. Figures 2 through 5 show the results of simulation for example 1 through 4. In each case the total offered traffic is a parameter. For the results shown 1.2 million requests were processed by the simulation model. This large number of requests was necessary in order to insure enough samples of rare large waiting times for reasonably reliable results.

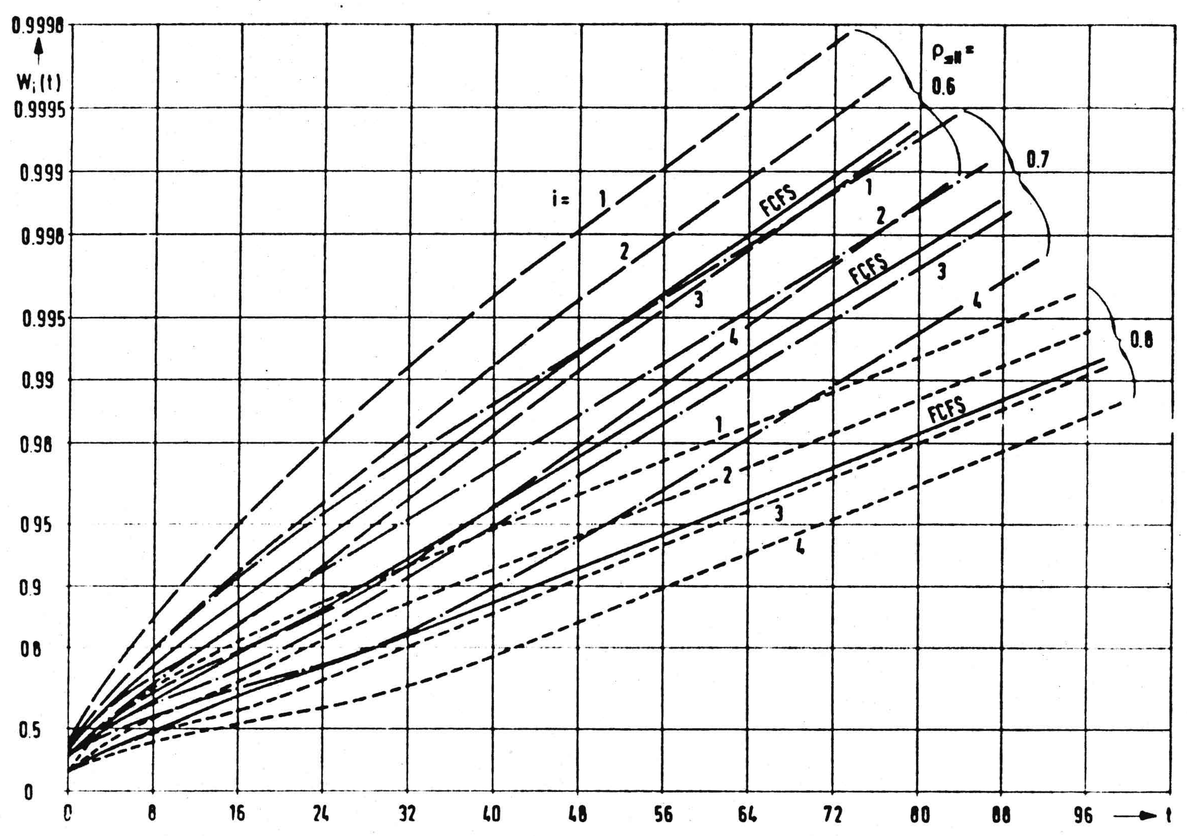

Fig. 2: Waiting-time d.f.'s of the model M/M/1/RU-NONPRE, cf. example 1. The type i and total offered traffic $\rho_{\leq N}$ are parameters. The solid and dashed curves result from the FCFS and RU disciplines, respectively.

Nevertheless, the results shown in Figures 2 through 5 had to be corrected in the tail of the waiting time d.f. (i.e. for large t) by extrapolation. Because of the great rarity of extremely large waiting times in one time-limited simulation experiment where the stationary state could not be completely reached, these large waiting times are underrepresented. Figure 6 shows the true results of example 2 (cf. Fig. 3). The "turning up" of the curves in this figure for large waiting times t is due to the limited number of samples. Experimentally it can be shown that as the number of requests processed per simulation increases, this "turning up" moves to ever greater waiting times. For a given waiting time t, the deviation from the straight line constructed by extrapolation from smaller waiting times becomes less and less.

By comparing the results of the four examples, it can be seen that the waiting time d.f.'s run parallel from about $f = w_i(t) = 0.98$ on when plotted semilogarithmically.

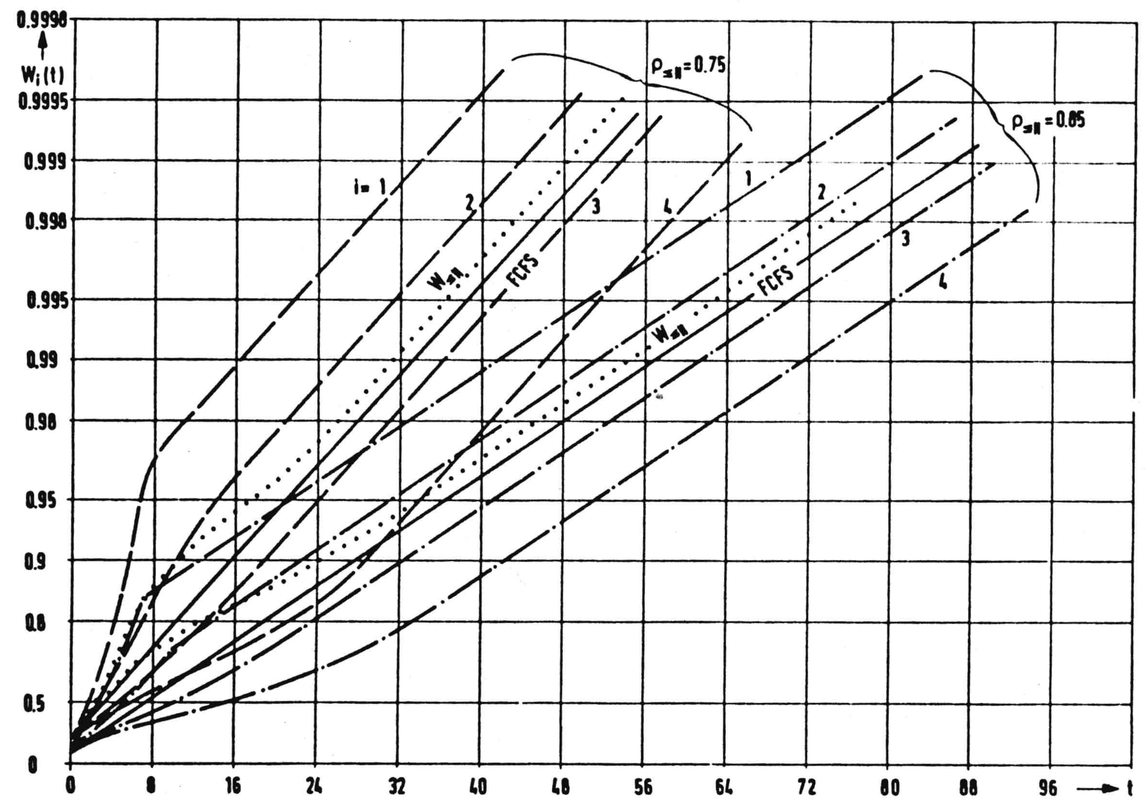

Fig. 3: Waiting-time d.f.'s of the model M/D/1/RU-NONPRE, cf. example 2. The type i and total offered traffic $\rho_{\leq N}$ are parameters. The solid and dashed curves result from the FCFS and RU disciplines, respectively.

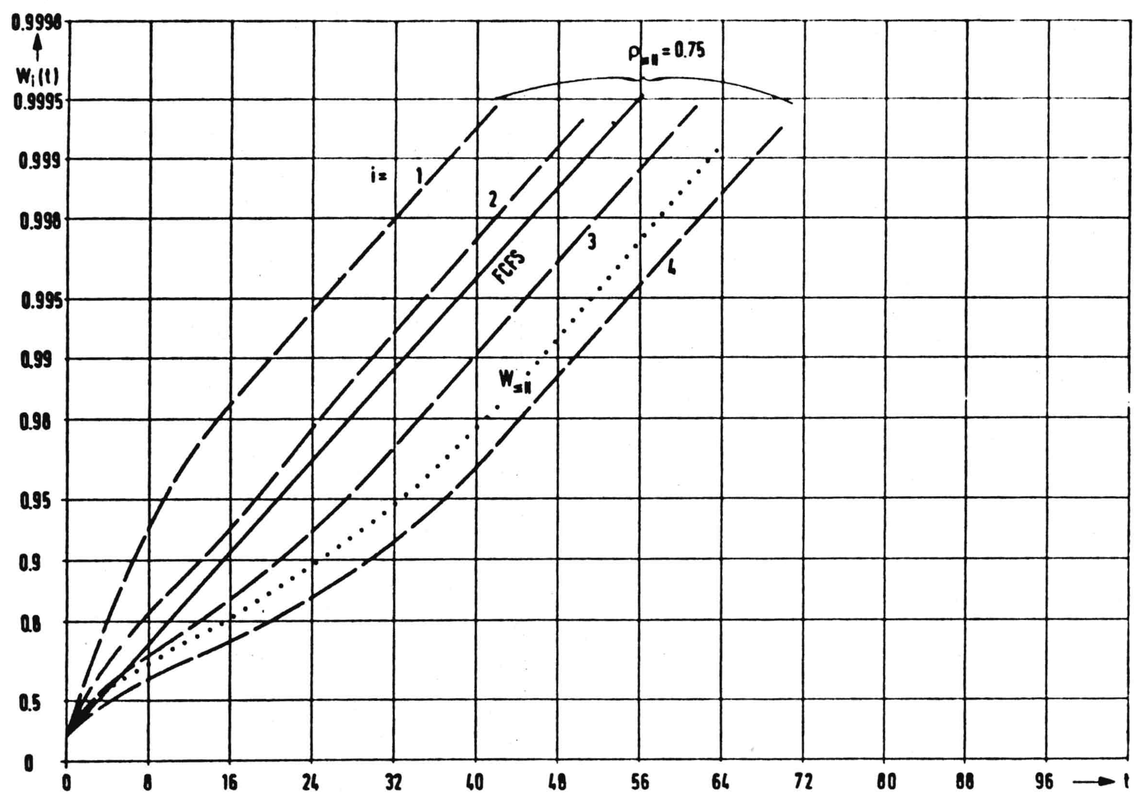

Fig. 4: Waiting-time d.f.'s of the model M/D/1/RU-NONPRE, cf. example 4. The type i and total offered traffic $\rho_{\leq N}$ are parameters. The solid and dashed curves result from the FCFS and RU disciplines, respectively.

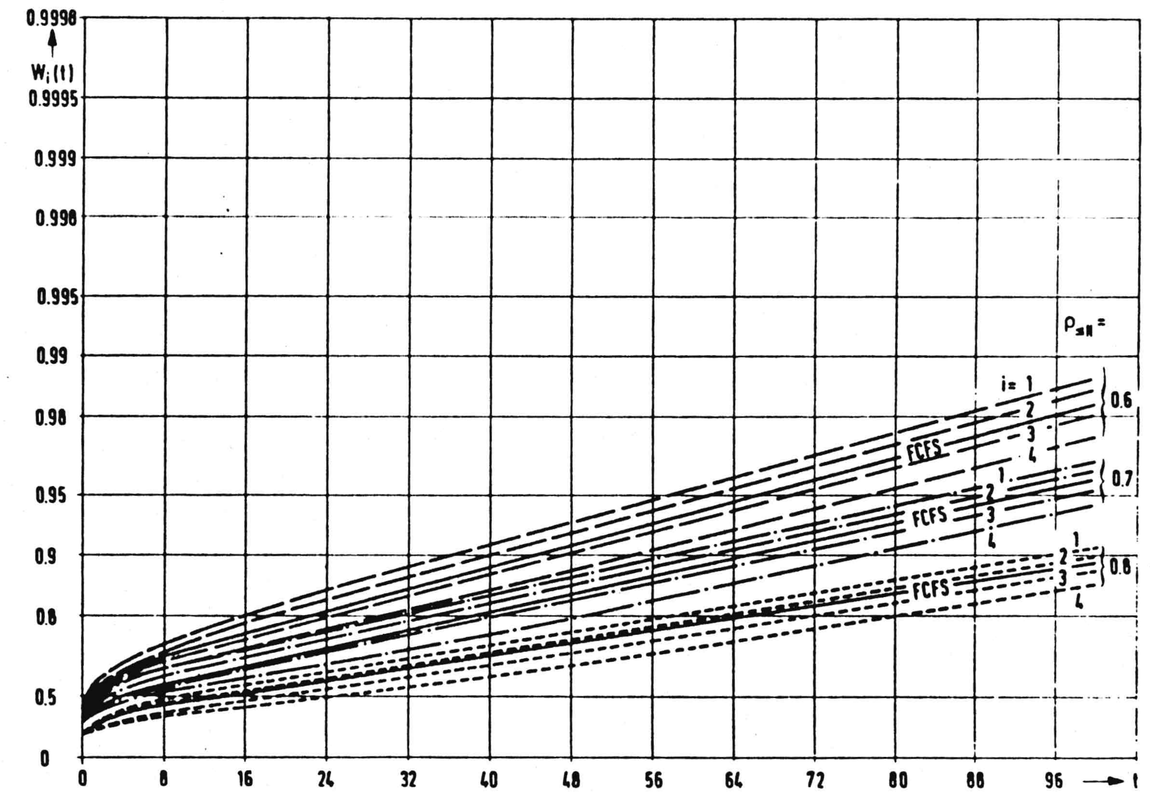

Fig. 5: Waiting-time d.f.'s of the model M/G/1/RU-NONPRE, cf. example 3. The type i and total offered traffic $\rho_{\leq N}$ are parameters. The solid and dashed curves result from the FCFS and RU disciplines, respectively.

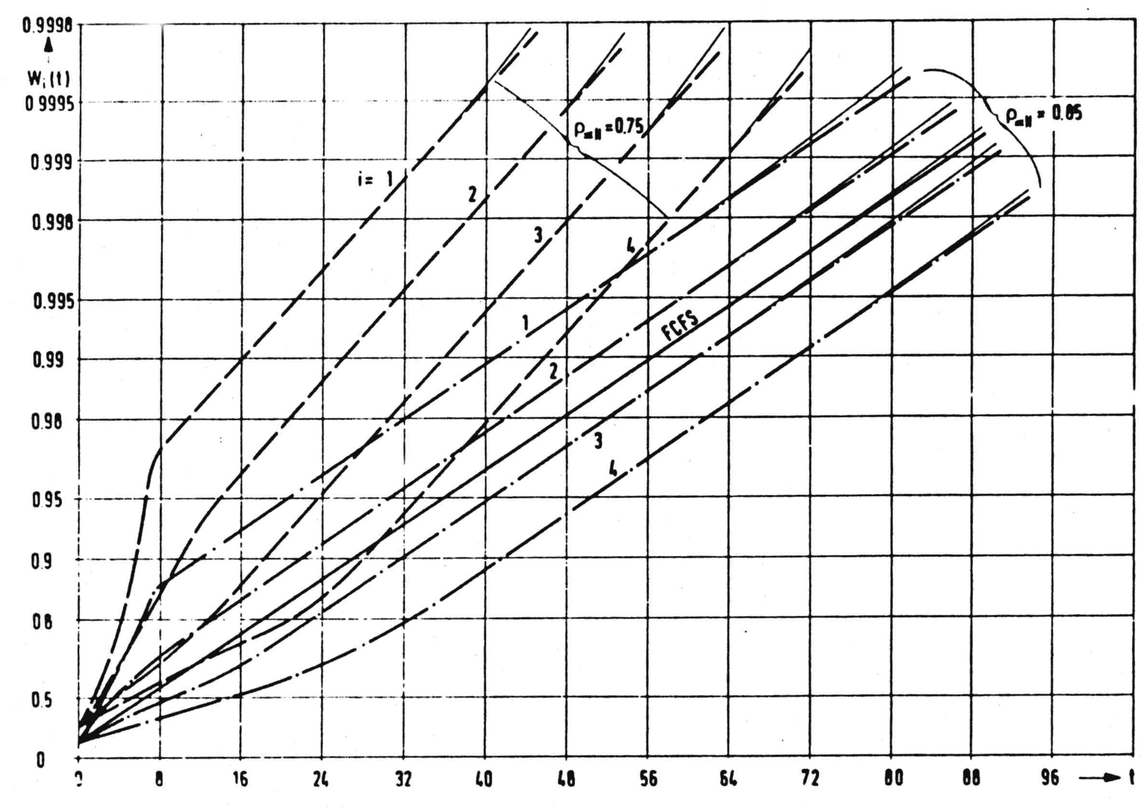

Fig. 6: Extrapolated results and true simulation results for the example 2, cf. Fig. 3.

Very large waiting times of different type requests differ by the difference in their urgencies. Smaller waiting times differ by less. The property formulated in Eq.(2.4) implies that the probability $P(w_i \leq t)$ of meeting deadlines with very large waiting times w_i is practically independent of the type of request, i.e. all requests are handled with equal fairness. If the condition

$$\min_i \{V_i\} \gg \max_i \{\beta_i\} , \qquad (1 \leq i \leq N) . \tag{2.5}$$

is fullfilled, then the probabilities of meeting the deadlines $P(w_i \leq V_i)$ are about the same, regardless of type. Only example 1 (c.f. Fig. 2) does not satisfy this condition. Here $V_1 < \beta_4$ and hence on the average all type-1 requests miss their deadline if they arrive while a type-4 request is being processed. The probability that a type-4 request is being processed is $\rho_{\leq N} / N$ and indeed the results of the simulation show that the difference in probability of meeting their respective deadlines $P(w_1 \leq V_1) - P(w_4 \leq V_4) \geq \rho_{\leq N} / 4$. This example shows that there is not sense in assigning deadlines without taking the condition in Eq.(2.5) into account. Such cases can be handled better by the RU-PRE discipline to be discussed later.

2.1.1. Previously known results

Back in 1962 Jackson [2] considered a model with discrete time and a non-preemptive RU discipline. The service times of all requests was taken from one geometric distribution, and the requests themselves were taken from a Bernoulli arrival process. The individual urgency V_i is computed for every arriving request according to a given distribution. Model M1, which corresponds to this discrete-time model but is time continuous, would have a Poisson arrival process and negative exponentially distributed service times which all originate from a common d.f. Model M1 differs from this model in that different-type requests may have different and arbitrary service-time d.f.'s. Therefore, our model is more general.

In [2] two relations are derived for the discrete time model

$$\lim_{f \to 1} [Q_i(f) - Q_j(f)] = V_i - V_j \qquad \text{(cf. Eq. (2.4))}$$

and

$$\lim [Q_i(f) - Q(f)] = V_i - \bar{V}' . \tag{2.6}$$

$\bar{V}'$ is defined by an expression which for model M1 may be approximated by the mean weighted urgencies $\bar{V}$, cf. list of abbreviations. In Eq.(2.6) Q(f) is the f-quantile of the waiting-time d.f. of a single-queue FCFS model, cf. Eq.(2.3). In Jackson's model and in the corresponding model M1, not only is Eq.(2.4) fulfilled, but also the f-quantiles of type-i requests in the RU-discipline differ from the f-quantile of the FCFS model by the difference $V_i - \bar{V}'$.

2.1.2. Waiting time d.f.'s of models in the RU and FCFS disciplines

Our simulation results of a FCFS model, which can easily be constructed from a model simulating the RU discipline by setting $V_i = r$, ($r \geq 0$, real) support the suspicion that Eq.(2.6) seems to be generally applicable, regardless of the service time d.f.'s assumed. From Fig.'s 2 through 5 it can be observed that $\bar{V}'$ deviates from $\bar{V}$ by no more than 9 %. These deviations are considerably larger than those observed by Jackson for his model. This is due to the fact that unlike Jackson's model the mean service times β_i in our examples are not equal, but different.

The difference $(\bar{V} - \bar{V}')$ is, depending on the example, greater or smaller than zero. The reason for this can be plausibly explained as follows:

We assume the model in Fig. 1 with non-preemptive static priorities and argue using the results for the common mean waiting time (c.m.w.t)

$$W_{\leq N} = \lambda_{\leq N}^{-1} \sum_{i=1}^{N} \lambda_i W_{iNONPRE} . \tag{2.7}$$

The c.m.w.t. results from weighting and summing over the individual mean waiting times of priority levels i. The c.m.w.t. $W_{\leq N}$ equals the m.w.t. W_{FCFS}, cf. Eq.(2.1), only if all mean service times β_i are equal. This is not the case in example 1 through 4.

From [3] it is known that $W_{\leq N}$ can be minimized in a static priority model by satisfying the condition

$$\beta_1 \leq \beta_2 \leq \cdots \leq \beta_i \leq \cdots \leq \beta_N , \tag{2.8}$$

where 1 is the highest priority. The kind of service time d.f.'s themselves do not matter; only Eq.(2.8) must be satisfied. Since $W_{\leq N}$ and W_{FCFS} are the means of the corresponding waiting time d.f.'s $P(w_{\leq N} \leq t)$ and $P(w_{FCFS} \leq t)$, respectively, then for the case of unequal β_i's, if Eq.(2.8) is satisfied, the w.t.d.f. $P(w_{\leq N} \leq t)$ must be better in some sense than $P(w_{FCFS} \leq t)$.

The opposite should be expected if the worst possible priority assignment is chosen, namely $\beta_i \geq \beta_{i+1}$. The w.t.d.f. $P(w_{\leq N} \leq t)$ is simply a weighted combination of the w.t.d.f.'s $P(w_{iNONPRE} \leq t)$ of the static priority model using

$$w_{\leq N} = \lambda_{\leq N}^{-1} \sum_{i=1}^{N} \lambda_i w_{iNONPRE} .$$

It follows that depending on how well or badly the relation in Eq.(2.8) is observed, the w.t.d.f. $P(w_{FCFS} \leq t)$ - which is independent of priority assignment - has a bader or better position in the set of curves $P(w_{iNONPRE} \leq t)$, resulting.

The assignment of urgency numbers to customers with different mean service times β_i has the same effect as the assignment of static priorities: small urgencies correspond to small priority numbers (high priorities) while large urgencies correspond to large priority numbers. Assigning small urgency numbers to customers with small mean service times β_i and large V_i's to customers with large β_i's in the model of Fig. 1 results in a bad position of $P(w_{FCFS} \leq t)$ compared to $P(w_{\leq N} \leq t)$ or compared to the set of curves $P(w_i \leq t)$. Just this urgency number assignment minimizes the common mean w.t. $W_{\leq N}$ and optimizes the c.m.w.t.d.f $P(w_{\leq N} \leq t)$. An inverse urgency number assignment results in the least favorable set of curves $P(w_i \leq t)$ compared to $P(w_{FCFS} \leq t)$. Now it is quite clear that $\bar{V}'$ in Eq.(2.5) cannot be identical to $\bar{V}$ which is approximately the case for equal β_i's.

The optimal urgency number assignment just introduced is used in the example 2 cf. Fig. 3. It can be seen that $P(w_{FCFS} \leq t)$ is less favorable compared to the set of curves $P(w_i \leq t)$ because their common w.t.d.f. $P(w_{\leq N} \leq t)$ is favorable compared to $P(w_{FCFS} \leq t)$. The opposite is true for the example 4, cf. Fig. 4: There the smaller β_i is, the larger the urgency V_i. In Fig. 4 the w.t.d.f. $P(w_{FCFS} \leq t)$ is situated favorable in the set of curves $P(w_i \leq t)$.

Note that the load to the RU model is the same in both examples 2 and 4. Only the assignment of urgencies differ. Hence the w.t.d.f. $P(w_{FCFS} \leq t)$ is the same for both examples which is not the case for the curves $P(w_i \leq t)$ and $P(w_{\leq N} \leq t)$. Only the assignment of the urgency numbers is different, which precisely is the reason why $\bar{V}'$ differs from $\bar{V}$. Our observation is that $\varepsilon = \bar{V} - \bar{V}'$ for all the examples studied was always very small (a few percent of $\bar{V}$). Therefore we decide to accept Eq.(2.6) with $\bar{V}'$ substituted by $\bar{V}$ as an approximation which leads to

$$\lim_{f \to 1} [Q_i(f) - Q(f)] = V_i - \bar{V} . \tag{2.9}$$

2.2. Approximate computation of the w.t.d.f. for the FCFS and RU-NONPRE disciplines

The w.t.d.f. in the FCFS discipline may be computed by inversion of their Laplace transform which can be a very difficult task. For the examples introduced in this paper it can be shown that the w.t.d.f.'s $P(w_{FCFS} \leq t)$ found by simulation espe-

cially in the tail, can be approximated very well by a degenerated exponential d.f. (DM-d.f.)

$$P(\leq t) = 1 - (1-p)\, e^{-\mu t} \quad . \tag{2.10}$$

By setting the first and second moments of $P(w_{FCFS} \leq t)$ equal to the corresponding moments of Eq.(2.10), the parameters μ and p can be shown to be

$$p = \frac{C^2 - 1}{C^2 + 1} \qquad \text{and} \qquad \mu = \frac{2}{W_{FCFS}\,(1 + C^2)} \quad , \tag{2.11}$$

where

$$C^2 = W^{(2)}_{FCFS} / W^2_{FCFS} - 1 \tag{2.12}$$

is the squared coefficient of variance and

$$W^{(2)}_{FCFS} = 2\, W^2_{FCFS} + \frac{\lambda_{\leq N}\, \beta^{(3)}_{\leq N}}{3(1 - \rho_{\leq N})} \quad . \tag{2.13}$$

W_{FCFS} and $W^{(2)}_{FCFS}$ are the first and second moments in the FCFS discipline, respectively. From this we have the approximation

$$P(w_{FCFS} \leq t) \approx (P \leq t) = 1 - (1-p)\, e^{-\mu t} \quad . \tag{2.14}$$

We are now going to study the quality of Eq.(2.14) as an approximation for the w.t.d.f. $P(w_{FCFS} \leq t)$. Instead of computing the first two moments W_{FCFS} and $W^{(2)}_{FCFS}$ we will use the corresponding values found by simulation which somewhat compensates possible inaccuracies of our simulation results. In Fig. 7 through 9 the simulation results for the w.t.d.f. $P(w_{FCFS} \leq t)$ are shown using the parameters of the four examples in section 2. From this figures, in which the total offered traffic is a parameter, it can be seen that simulation (+) and approximation (straight line) fit together very well. This is especially true for the tails of the w.t.d.f.

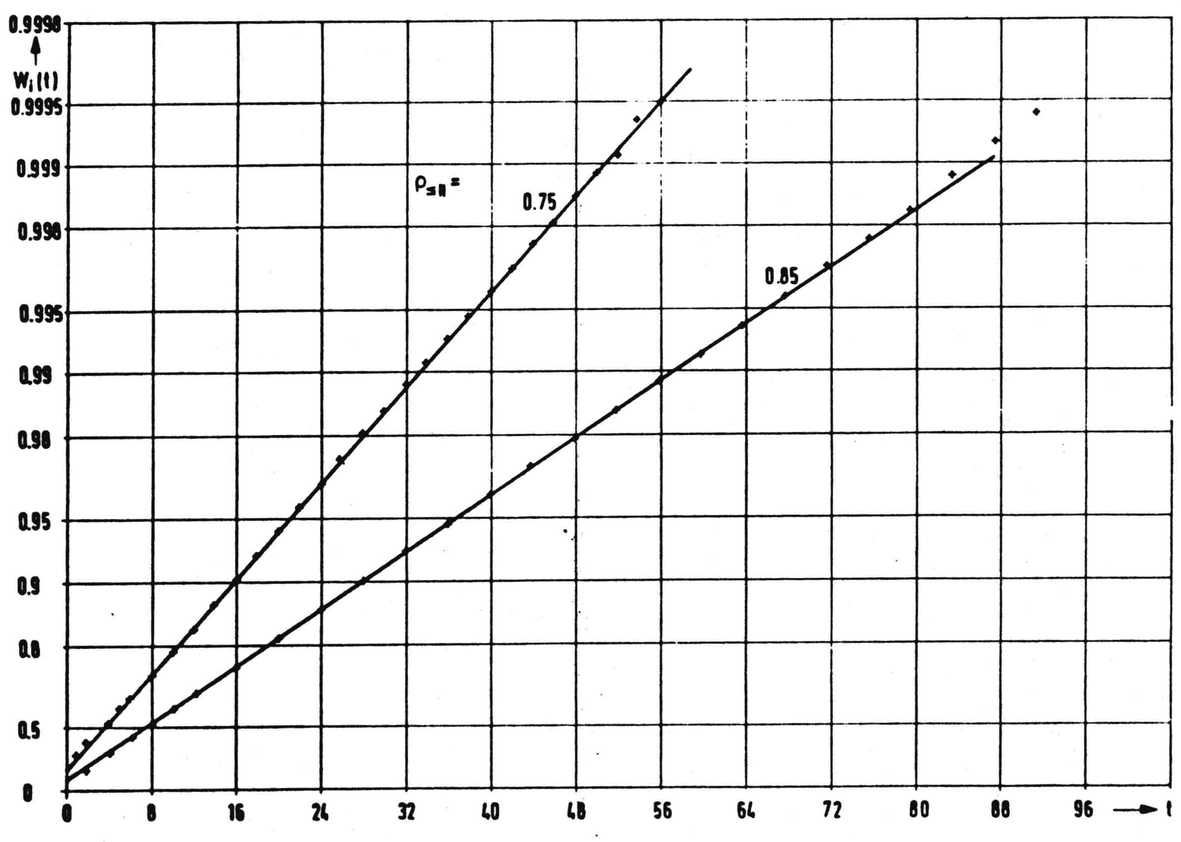

Fig. 7: Waiting time d.f.'s of the model M/D/1/FCFS, cf. examples 2 and 4. Approximations by means of Eq.(2.10) (lines) and simulation results (+) are shown.
Parameter is the total offered traffic.

Our goal is to approximate the tails of the w.t.d.f.'s in the RU discipline using Eq.(2.9). As a starting point we need a good approximation of the tail of the w.t.d.f. in the FCFS discipline. This prerequisite seems to be fulfilled.

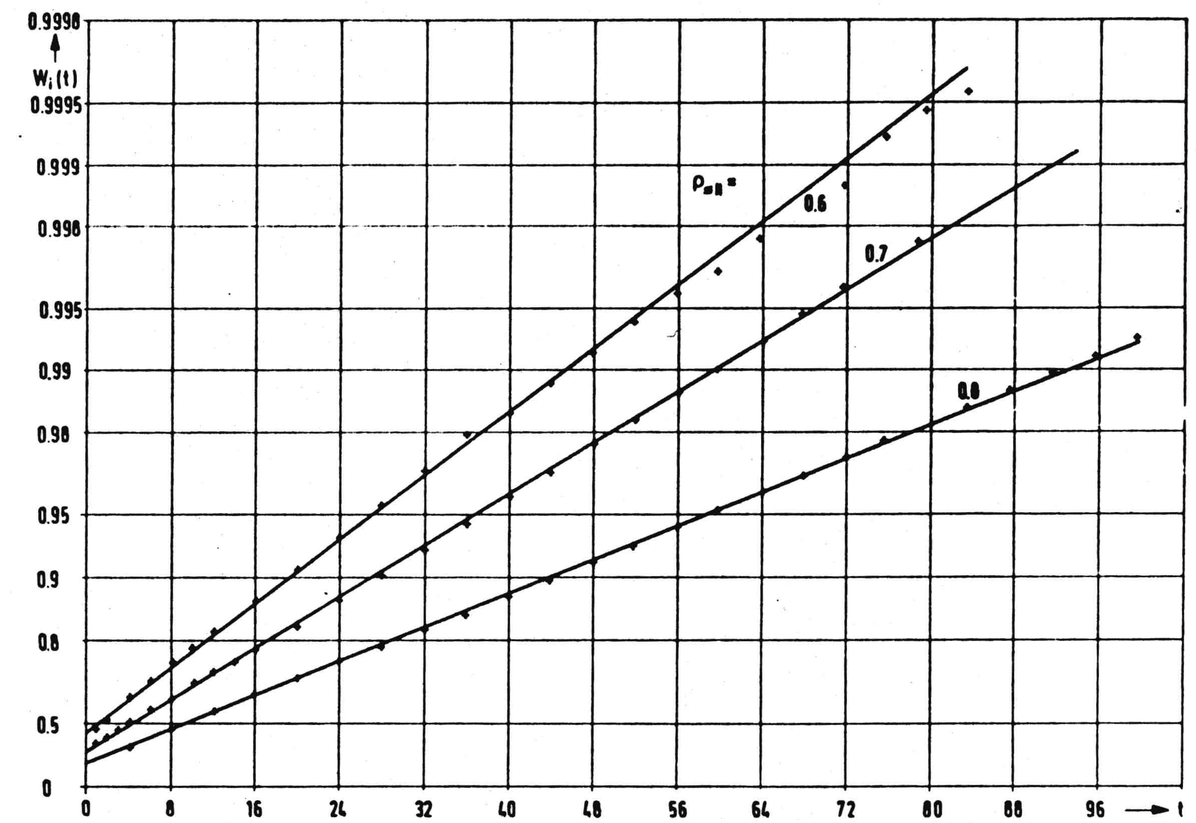

Fig. 8: Waiting time d.f.'s of the model M/M/1/FCFS, cf. example 1. Approximations by means of Eq.(2.10) (lines) and simulation results (+) are shown. The total offered traffic is a parameter.

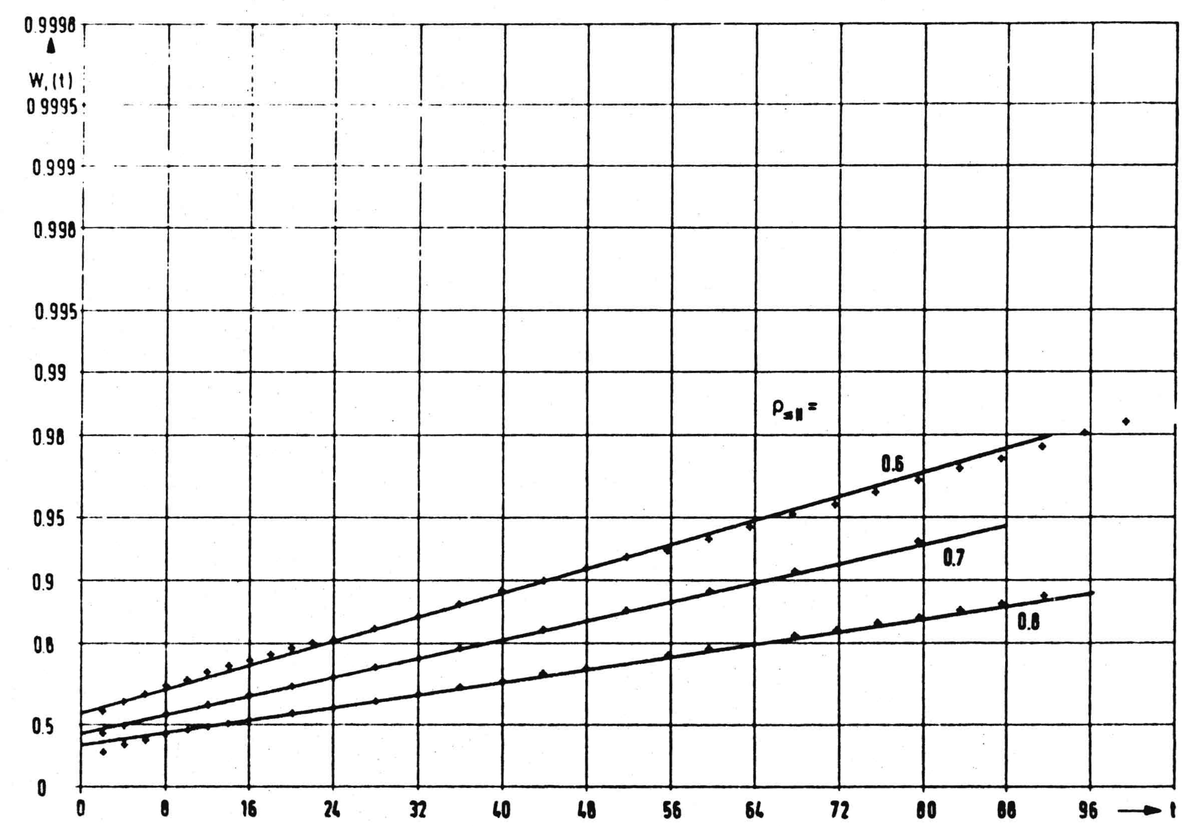

Fig. 9: Waiting time d.f.'s of the model M/G/1/FCFS, cf. example 3. Approximations by means of Eq.(2.10) (lines) and simulation results (+) are shown. The total offered traffic is a parameter.

2.2.1. Approximation of the tail of the w.t.d.f. in the RU-discipline

Using Eq.(2.9) it is possible to compute an expression $P'(w_i > t) = 1 - P'(w_i \leq t)$ which approximates the probability in the RU-NONPRE discipline of a waiting time w_i exceeding a large value of t

$$P'(w_i > t) \approx P(w_{FCFS} > t + \bar{\nu} - \nu_i) \ . \tag{2.15}$$

From experience gained with simulation of the four examples one may call waiting times "large" if the relation $P(w_i > t) < 0.05$ is fulfilled. For medium and small sized waiting times it can be observed that, if

$$\begin{aligned} &\bar{\nu} - \nu_i < 0 \quad \text{then} \quad P(w_i > t) < P'(w_i > t) \\ &\bar{\nu} - \nu_i > 0 \quad \text{then} \quad P(w_i > t) > P'(w_i > t) \ . \end{aligned} \tag{2.16}$$

Using Eqs.(2.9), (2.14) and (2.15) we get as an approximation

$$\lim_{t \to \infty} [P(w_i \leq t)] \approx 1 - (1-p)\, e^{-\mu(\bar{\nu} - \nu_i)}\, e^{-\mu t} \ . \tag{2.17}$$

Eq.(2.17) defines the tails of the w.t.d.f. in the RU-NONPRE discipline. Being adequate to parameter p in Eq.(2.14) we introduce a parameter p_i for type-i requests in the RU discipline and compute, by setting equal the right hand sides of both Eqs.(2.14) and (2.17),

$$p_i = 1 - (1-p)\, e^{-\mu(\bar{\nu} - \nu_i)} \ .$$

By insertion in Eqs.(2.17) we get

$$\lim_{t \to \infty} [P(w_i \leq t)] \approx 1 - (1 - p_i)\, e^{-\mu t} \ .$$

Note that p and p_i are the probabilities which arise for $t = 0$.
From Simulation experiments we know that Eq.(2.17) is a good approximation, if either the total offered traffic is large, $\rho_{\leq N} > 0.6$, or the urgency numbers have a small variance, or both. Otherwise, the w.t.d.f.'s in the RU discipline appear to be similar to those of the static priority discipline which is always the case for small waiting times t. The reason for this is that without the conditions mentioned, there is only a very small probability of missing deadlines for any type of request.
Figures 10 through 13 show both the simulation results (+,•) and approximate w.t.d.f.'s computed from Eq.(2.17) which appear as straight lines for the examples defined in section 2. It can be seen that for large waiting times, simulation and approximation agree very well. Deviations are typically below the 10 % range.

In Fig. 11 we can once again study the results of a different assignment of a set of urgency numbers ν_i to a set of mean service times β_i as defined by examples 2 and 4. The approximate w.t.d.f.'s are derived from the w.t.d.f. of the FCFS discipline and therefore are the same for both examples 2 and 3. As a consequence of the optimal urgency number assignment as defined in section 2.1.2 which is realized in the example 2, the real w.t.d.f.'s (•) compare favorably to the computed ones (lines). On the other hand the w.t.d.f.'s (x) of example 4, in which the ν_i are especially poorly assigned to the β_i, are less favorable than the computed ones. Similar results can be observed from Fig. 10 and 12 where the urgency numbers are assigned optimal. Note that the approximation by Eq.(2.1) does not take into account the urgency-number assignment. This should be the reason for the deviation between simulated and computed w.t.d.f.'s observed.

2.2.2. Approximation of the w.t.d.f. in the RU discipline for small waiting times

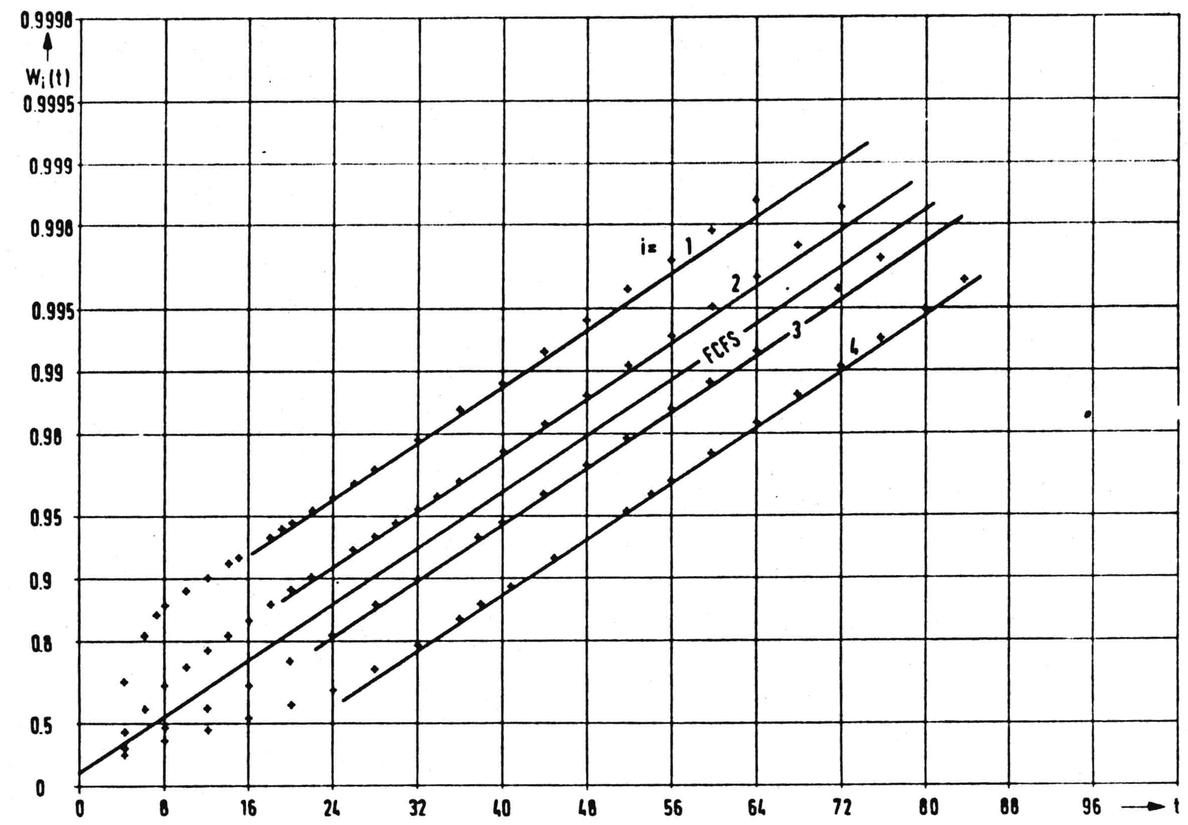

Fig. 10: Waiting time d.f.'s of the model M/D/1/RU-NONPRE, cf. example 2, approximated by means of Eq.(2.17) (lines) and simulation results (+). Additionally the FCFS approximation is shown. The total offered traffic is $\rho_{\leq N} = 0.85$.

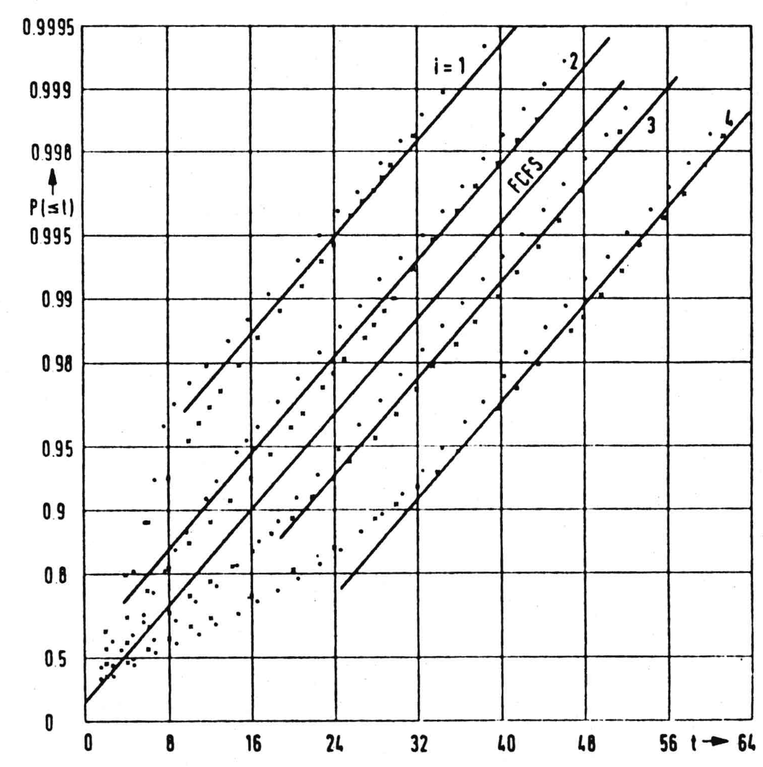

Fig. 11: Waiting time d.f.'s of the model M/D/1/RU-NONPRE, cf. examples 2 and 4, approximated by means of Eq. (2.17) (lines) and simulation results (+, •). Additionally the FCFS approximation is shown. The total offered traffic is $\rho_{\leq N} = 0.75$.

pline to compute the corresponding mean waiting time W_i of type-i requests

$$P(w_i > t) = \rho_{\leq N}\, e^{-\rho_{\leq N}\, t / W_i} \ . \tag{2.18}$$

This simple approximation depends only on the total offered traffic $\rho_{\leq N}$ and W_i. Figures 14 through 16 show this approximation (lines) compared to simulation results (x) for examples 1, 2, and 3. The most unsatisfactory approximation appears to be for type-1 requests which was not needed in [1]. It can be seen that Eq.(2.18) in general cannot be called a good approximation. Even for small waiting times the approximated w.t.d.f. deviates from the simulated w.t.d.f. by up to 20 %.

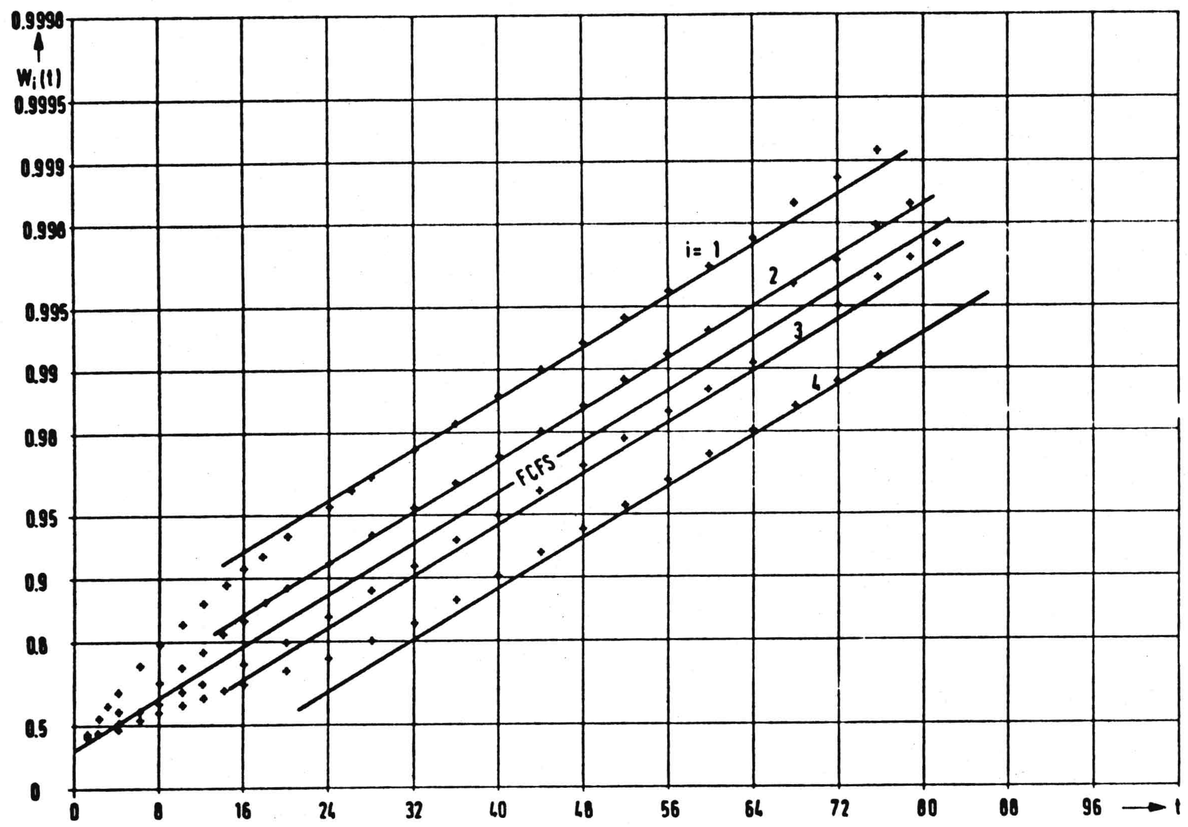

Fig. 12: Waiting time d.f.'s of the model M/M/1/RU-NONPRE, cf. example 1, approximated by means of Eq.(2.17) (lines) and simulation results (+). Additionally the FCFS approximation is shown. The total offered traffic is $\rho_{<N} = 0.75$.

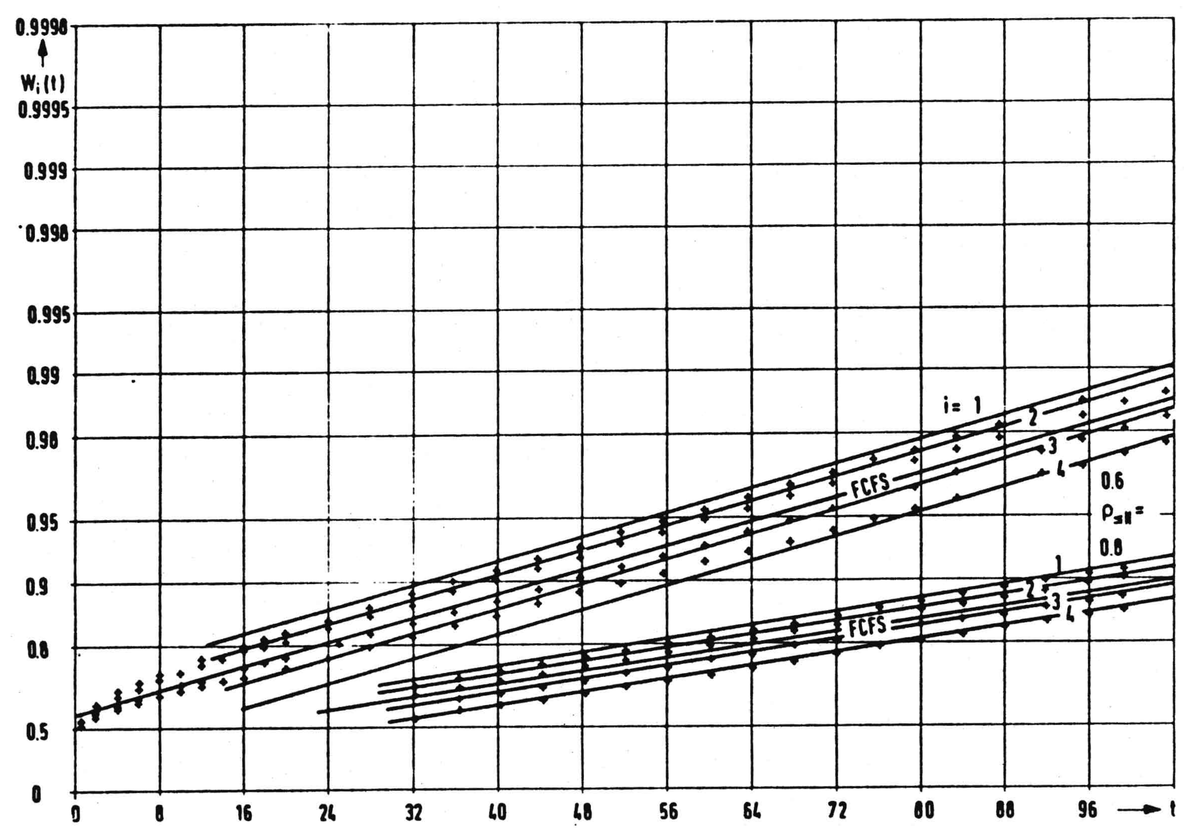

Fig. 13: Waiting time d.f.'s of the model M/G/1/RU-NONPRE, cf. example 3 of section 2, approximated by means of Eq.(2.17) and simulation results (+). Additionally the FCFS approximation is shown. The total offered traffic is a parameter.

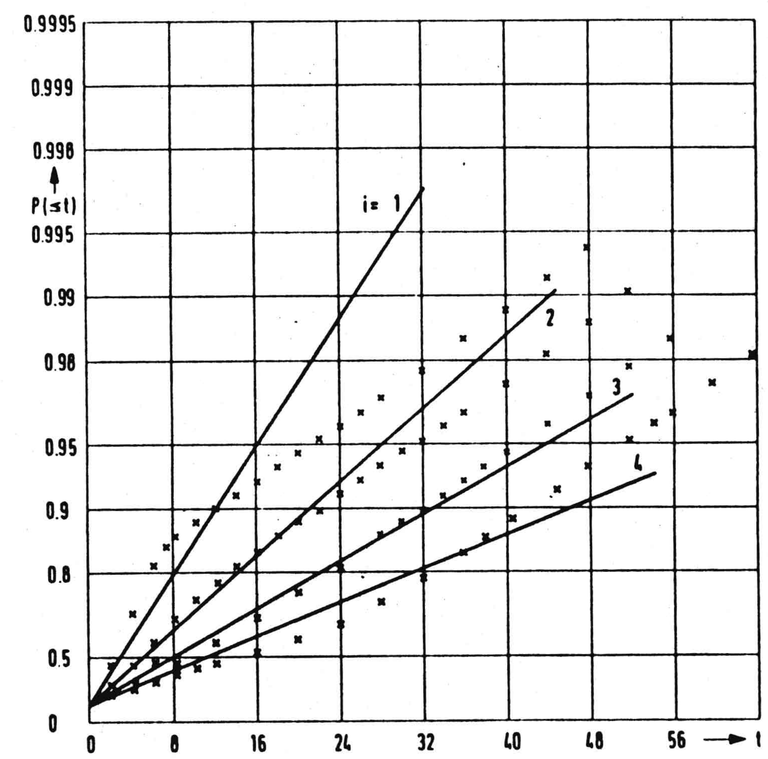

Fig. 14: Waiting time d.f.'s of the model M/D/1/RU-NONPRE, cf. example 2, approximated by means of Eq.(2.18) (lines) and simulation results (x). The total offered traffic is $\rho_{\leq N} = 0.85$.

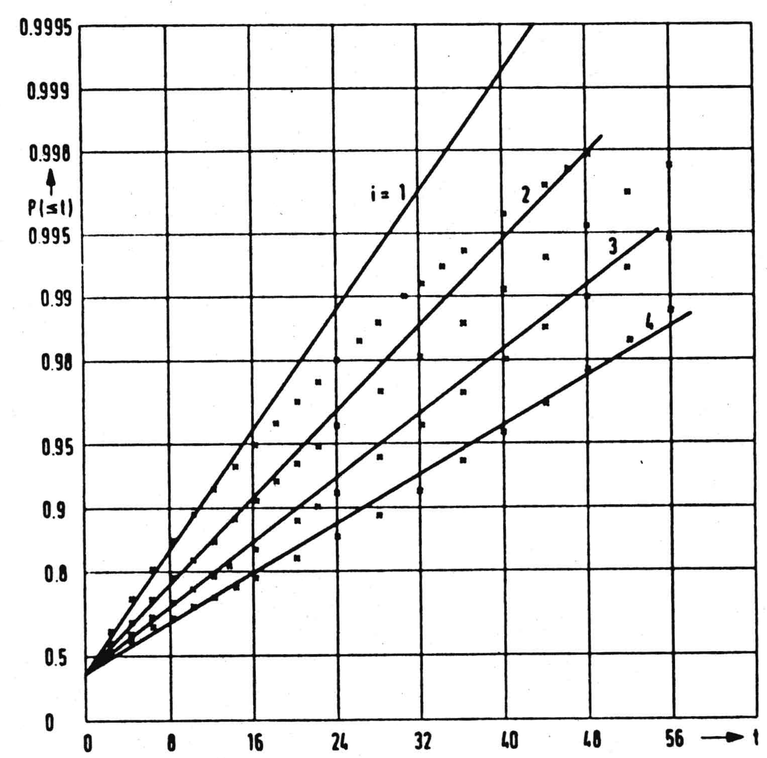

Fig. 15: Waiting time d.f.'s of the model M/M/1/RU-NONPRE, cf. example 1, approximated by means of Eq.(2.18) (lines) and simulation results (x). The total offered traffic is $\rho_{\leq N} = 0.6$.

Nevertheless, the error produced by this approximation has a negligible effect on the computation of W_i, for simulation results agree well with the computed values of W_i, cf. [1].

2.3. Mean waiting time of requests which have missed their deadlines

From Eq.(2.17) the approximate probability of a type-i request missing its deadline (at $t = V_i$), presuming a sufficiently large t, can be computed

$$P(w_i > V_i) \approx (1 - p)\, e^{-\mu \bar{V}} . \tag{2.19}$$

The w.t.d.f. of requests missing their deadlines is approximately given by

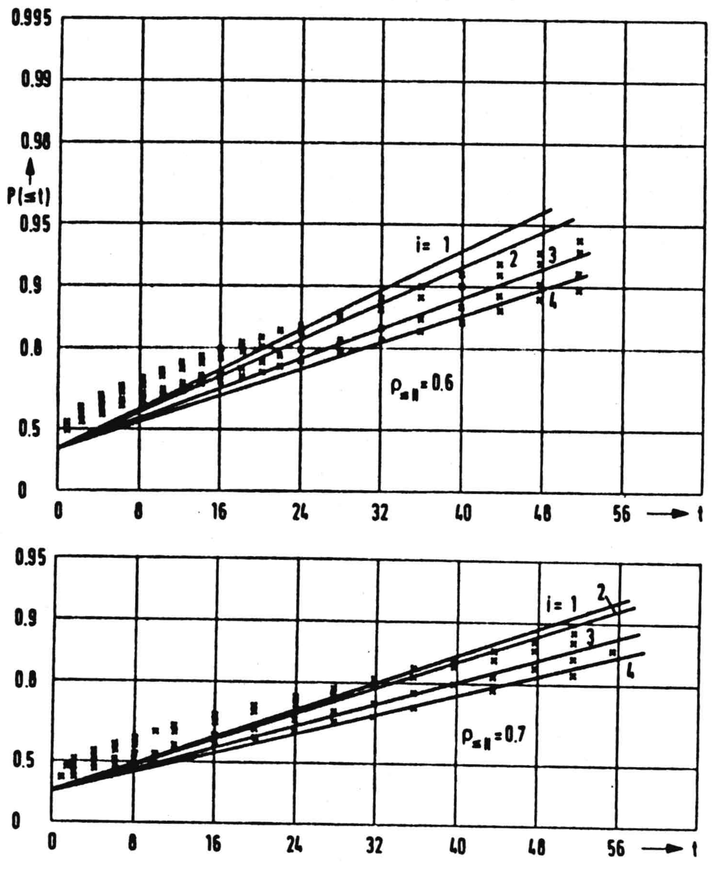

Fig. 16:
Waiting time d.f.'s of the model M/G/1/RU-NONPRE, cf. example 3, approximated by means of Eq.(2.18) (lines) and simulation results (x). The total offered traffic is a parameter.

$$P(w_i \le t \mid t > V_i) = \frac{P(w_i \le t) - P(w_i \le V_i)}{1 - P(w_i \le V_i)}$$

$$P(w_i \le t \mid t > V_i) = 1 - e^{\mu V_i}\, e^{-\mu t} \quad . \tag{2.20}$$

Apparently a DM-d.f. arises whose mean

$$E(w_i \mid t > V_i) = W_i^* = e^{\mu V_i} / \mu \tag{2.21}$$

only depends on μ and V_i. Note that μ can be computed from Eq.(2.11). The approximate mean waiting times W_i^* of requests having missed their deadlines differ approximately in the factor e^{V_i}.

3. COMPARISON OF W.T.D.F.'s OF THE RU-NONPRE AND STATIC PRIORITY DICIPLINES

The RU-NONPRE discipline may be thought of as an alternative to the static priority discipline. From [1] and [2] it is known that the greater the variance of the urgency numbers is the more the mean waiting time W_i in the RU discipline differs from the m.w.t. W_{FCFS} and the more the m.w.t.'s in the static priority discipline $W_{iNONPRE}$ are approximated. Now, depending on the set of V_i's, the question arises how the w.t.d.f.'s in the RU and static priority disciplines will differ. Moreover, it is interesting to know which of those disciplines better observes a given set of deadlines defined by the urgency numbers V_i.
In order to make such an investigation let us define a figure of comparison

$$\ddot{U} = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i P(w_i > V_i) \quad . \tag{3.1}$$

By computing the value of $\ddot{U}$ for different disciplines and a given set of V_i's and λ_i's, we are only able to compare the total probabilities of missing the deadlines given by the V_i's for all type-i requests together.

Recalling that the probabilities needed to compute $\ddot{U}$ are available only from an approximation, Eq.(2.17), we decided to use simulation to clarify this question. For the four examples mentioned, we found regardless of the total offered traffic $\rho_{\le N}$ that neither of the two disciplines is basically superior to the other. Depending on the case, either discipline can produce a lower probability $\ddot{U}$.

Generally speaking it seems that using static priorities and applying a priority assignment according to Eq.(2.8) the common probability $\ddot{U} = \ddot{U}_{stat}$ for missing the set of deadlines given by the V_i's is smaller than $\ddot{U} = \ddot{U}_{dyn}$ using the RU-discipline. The opposite is true if one deviates substantially from the optimal priority assignment, i.e. $\ddot{U}_{dyn} < \ddot{U}_{stat}$ arises.

Recalling from section 1 that

$$W_{istat}(V_i) = P(w_{istat} \le V_i) = 1 - P(w_{istat} > V_i)$$

it is possible to transform the figure of comparison introduced in Eq.(1.5) to

$$V = \lambda_{\le N}^{-1} \sum_{i=1}^{N} \lambda_i \left[P(w_{idyn} > V_i) - P(w_{istat} > V_i) \right]$$

$$V = \ddot{U}_{dyn} - \ddot{U}_{stat} \quad . \tag{3.2}$$

The relations defined by Eq.(1.6) still remain valid.
Simulation results in the static priority discipline of the examples 2 (broken lines) and 4 (solid lines) are shown in Fig. 17. For each example 1 million requests were processed in the simulation model. In example 2 the static priorities are op-

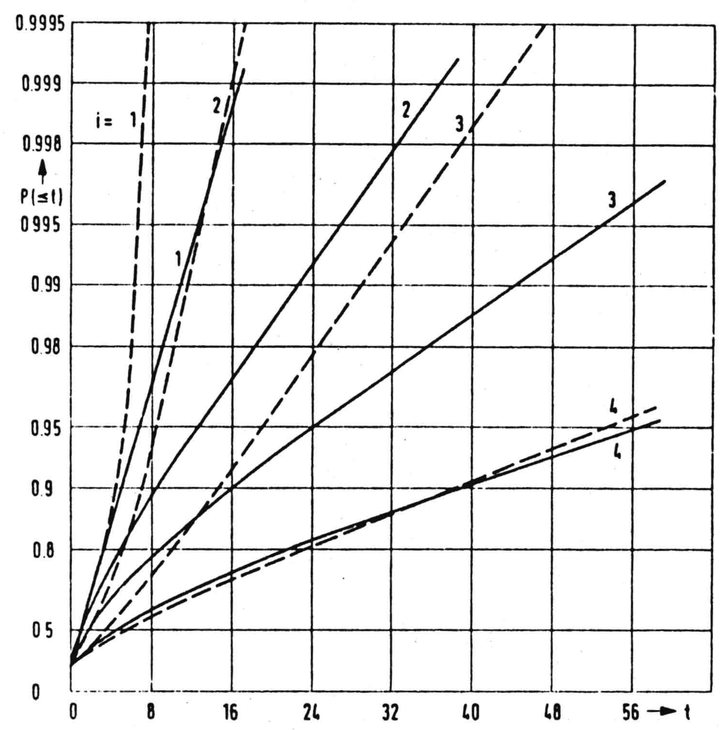

Fig. 17:
Simulated waiting time d.f.'s of the model M/D/1 in the non-preemptive static priority discipline for the example 2 (dashed lines) and 4 (solid lines). The total offered traffic is $\rho_{\le N} = 0.75$.

timally assigned while the opposite is true in example 4. The corresponding w.t.d.f.'s in the RU discipline are already known from Fig.11. For example 2 using the corresponding set of urgency numbers one computes $\ddot{U}_{stat} \approx 0.025 \approx \ddot{U}_{dyn} / 3$ which yields $V \approx 0.05$. For example 4 we find $\ddot{U}_{stat} \approx 0.5 \approx 3\,\ddot{U}_{dyn}$ which yields $V \approx -0.034$. Apparently, neither of the two disciplines can claim to always guarantee the minimum total probability $\ddot{U}$ of missing the deadlines given by the set of the urgency numbers, independent of the parameter set applied.

This observation was also made with other values of $\rho_{\le N}$ and other examples: example 1 with a total offered traffic $\rho_{\le N} = 0.6$ yields $\ddot{U}_{dyn} = 0.172$, $\ddot{U}_{stat} = 0.147$, and

$V = 0.025$, i.e. the static priority discipline is superior.

For the example 3 we found for $\rho_{\leq N} = 0.6$ total probabilities $Ü_{dyn} = 0.215$, $Ü_{stat} = 0.172$; and from this $V = 0.043$ which indicates that here too the static priority discipline is superior.

In the examples considered, V has been observed to appear in the range of +5 % to -3.4 %. Recalling that V expresses mean values of probabilities of missing deadlines for all requests together we have substantial differences between the two disciplines compared. From our observations one can conclude that a priority assignment according to Eq.(2.8) results in $V > 0$, cf. Eq.(1.7).
One substantial difference between the static-priority and the RU-discipline is not taken into account by our comparison using V. This difference can be seen by comparing the individual probabilities of type-i requests in the same discipline: While the RU-discipline meets all deadlines of different requests given by the set of the V_i's with nearly the same probability, this probability may differ by factors in the static priority discipline, presuming the same set of the V_i's. In the RU-discipline all requests are handled on an equal basis while the opposite is true for the static priority discipline.

4. SOME TYPICAL PROPERTIES OF THE W.T.D.F. IN THE PREEMPTIVE RU-DISCIPLINE

The main differences between the preemptive and non-preemptive versions of the RU-discipline were already presented in the paragraph of section 1 preceding the abbreviations. Preemption allows a request being serviced ahead of time to be interrupted by a request whose deadline would otherwise be missed. Interruptions are allowed for this reason only.

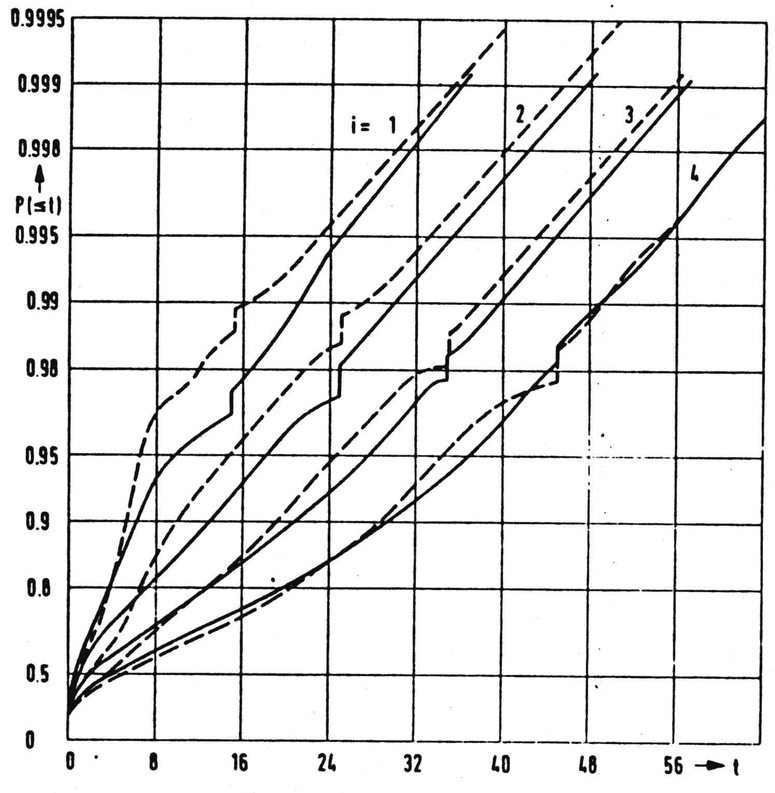

Fig. 18:
Simulated waiting time d.f.'s of the model M/D/1/RU-PRE for the examples 2 (solid lines) and 4 (dashed lines). The total offered traffic is $\rho_{\leq N} = 0.75$. The jumps appear at $t = V_i$.

The waiting-time d.f.'s in the RU-PRE discipline presuming the data of the examples 2 and 4 shown in Fig. 18 and example 3 in Fig. 19 were determined by simulation. Comparable results using the RU-NONPRE discipline are shown in Fig.'s 11 and 13. It can be seen that for the same total offered load, distinct differences in the waiting times of both disciplines arise. Especially worth noting is the jump at $t = V_i$ which results from interrupting prematurely started servicing for such requests which would otherwise miss their deadlines. The size of the jump exactly defines the probability for interrupting other types of requests. It can be observed that the w.t.d.f.'s in the tail in semilogarithmic scaling are parallel, as

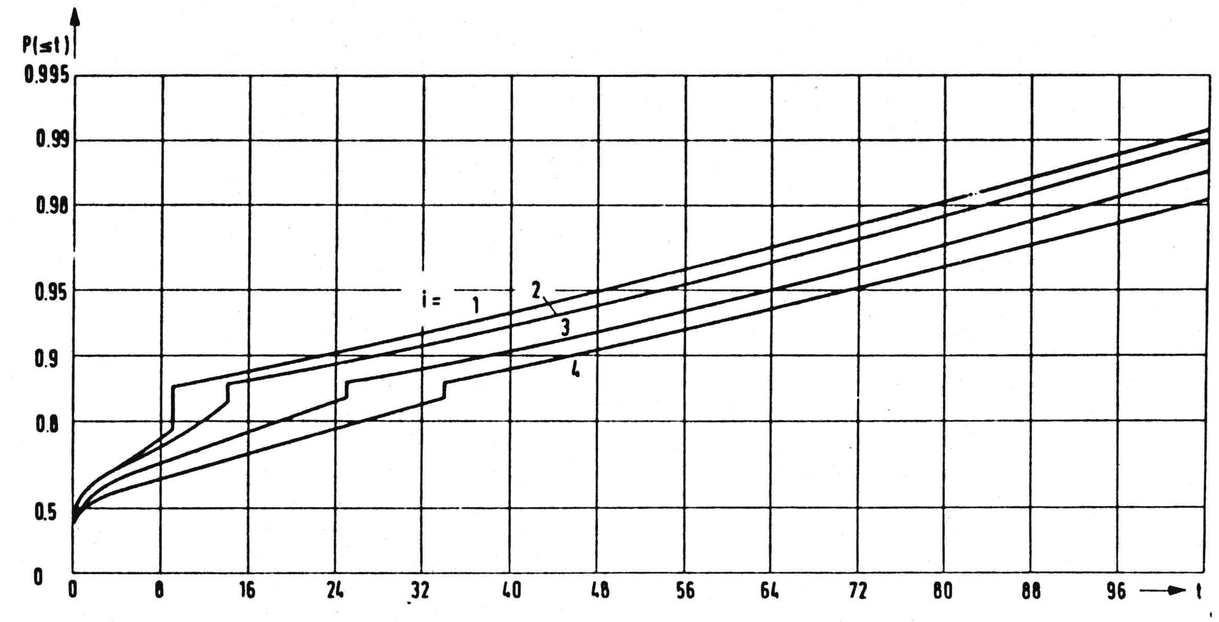

Fig. 19:
Simulated waiting time d.f.'s of the model M/G/1/RU-PRE, cf. example 3. The total offered traffic is $\rho_{\leq N} = 0.6$. The jumps appear at $t = V_i$.

was the case for the RU-NONPRE discipline (cf. [6]). Not for all examples considered the slope of these tails in the RU-PRE and RU-NONPRE disciplines was found to be the same (the same offered traffic $\rho_{\leq N}$ presumed).

It appears that interruptions are useful for increasing the probabilities of meeting deadlines if the coefficients of variance of the service time d.f.'s are large, i.e. $C_i > 1$ (cf. Fig.'s 13 and 19). For small C_i's our simulation supports the conjecture that interruptions of service time in favour of others having a larger expected service time is disadvantageous for meeting deadlines (cf. Fig.'s 11 and 18). The deadlines of some types of service requests are then met better with the RU-NONPRE discipline than with the RU-PRE discipline.

The common probability Ü for missing deadlines with either static priorities or the RU-PRE discipline was determined by simulation. With respect to Ü it can be seen that neither discipline is generally superior to the other, as was the case with non-preemptive priorities. Depending on the example presumed for such a comparison one of both disciplines appears to be superior.

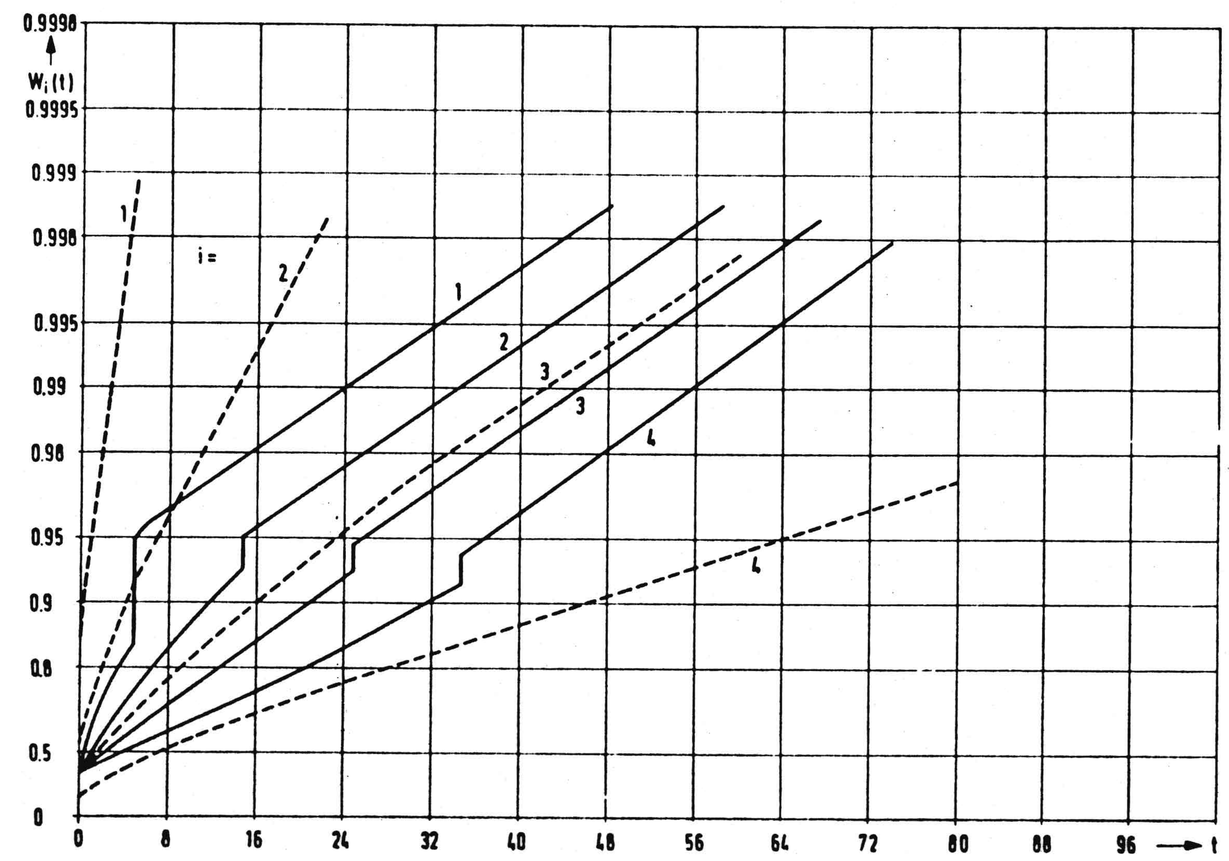

Fig. 20:
Simulated waiting time d.f.'s of the model M/M/1 in both preemptive disciplines: static priorities (dashed lines) and RU (solid lines), cf. example 1. The total offered traffic is $\rho_{\leq N} = 0.6$. The jumps appear at $t = V_i$.

In the example shown in Fig.20 both sets of w.t.d.f.'s in the RU and static priority discipline can be compared. According to Eq.(1.7) we have $V = 0.031$. Upon comparing both versions of the RU-discipline, it was found that in the preemptive version the deadlines are always met somewhat better than in the non-preemptive version (cf. Fig.21). The probability of all types of request meeting their deadlines is, as was the case in the RU-NONPRE discipline, approximately the same.

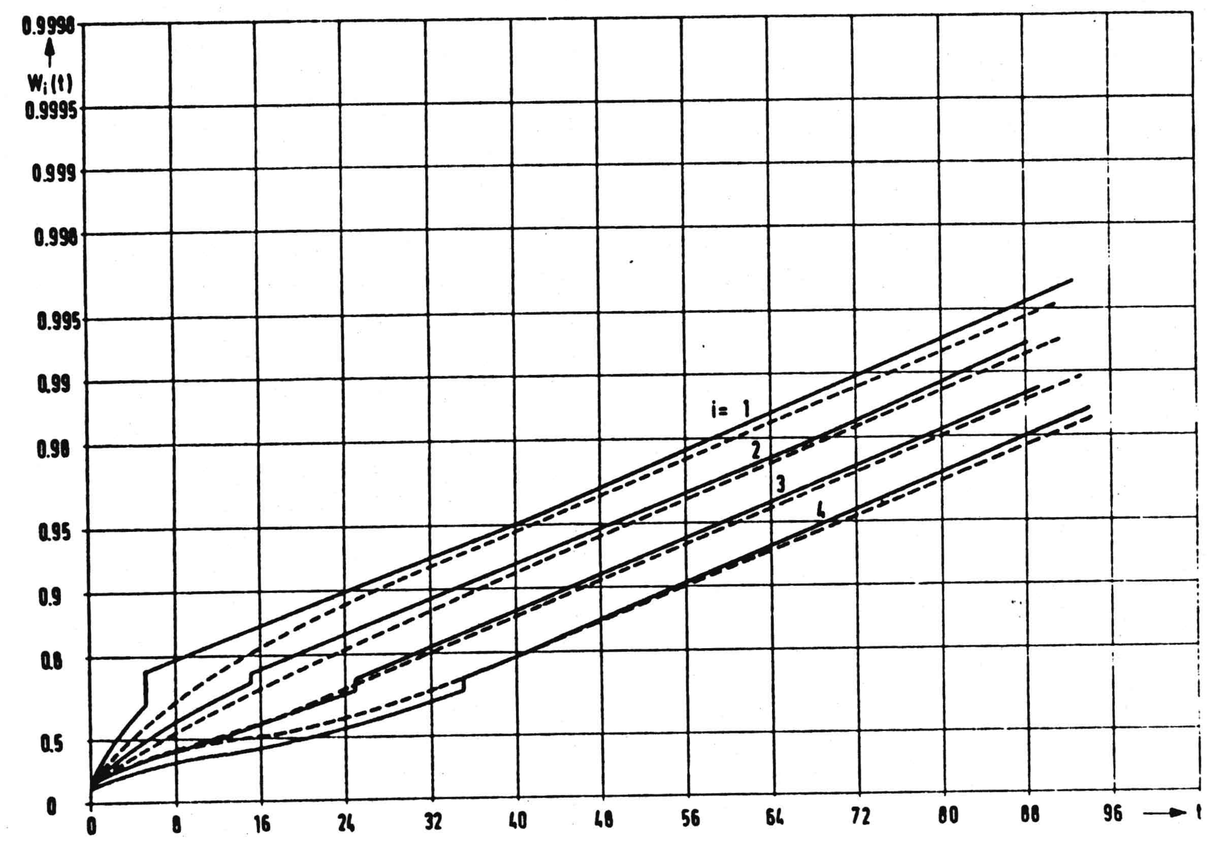

Fig. 21: Simulated waiting time d.f.'s of the model M/M/1 in the preemptive (solid lines) and non-preemptive (dashed lines) versions of the RU discipline, cf. example 1. The total offered traffic is $\rho_{\leq N} = 0.8$.

Literature:

[1] B. Walke: Improved bounds and an approximation for a dynamic priority queue. Third Int. Symp. Modell. and Performance Evaluat. Comp. Syst., Oct. 1977, Bonn, Germany. Proceed. published by North Holland Publishing Comp., Amsterdam, The Netherlands (1978) pp. 321 - 346.

[2] J.R. Jackson: Waiting time distributions for queues with dynamic priorities. Naval Res. Logistic. Quart., March (1962) Washington D.C., 9, pp. 31 - 46.

[3] R.W. Conway, W.L. Maxwell, L.W. Miller: Theory of scheduling. Addison Wesley Publishing Comp., Reading, Mass, (1967).

[4] B. Walke: Realzeitrechner-Modelle, Theorie und Anwendungen, R. Oldenbourg Verlag, Reihe Datenverarbeitung, München (1978).

[5] (not cited) E.B. Veklerov, I.M. Dukhovnyi: Queueing system with time-limited categorical priority. Probl. Inform. Transmiss., Vol.12, No. 1, Jan.-March (1976) pp. 73 - 79.

[6] H.M. Goldberg: Jackson's conjecture on earliest due date scheduling. Departm. Finance & Management Science, Faculty Busin. Administr. & Commerce, University of Alberta, Edmonton, Alberta, Canada T6G 2G1, Sept. (1978).