Deadline-Oriented Servicing: Waiting-Time Distributions

B. WALKE AND W. ROSENBOHM

Abstract—An infinite queue single server model is considered where requests arrive from independent Poisson streams and demand service according to arbitrary distribution functions which may be different for different requests. Associated with each request is an urgency number which, together with request's time of arrival, defines a deadline for beginning its service. This relative urgency discipline has at its two limiting cases the first-come first-serve and head of the line discipline. In [1] the mean waiting time is computed approximately and close bounds are derived there. Here we present simulation results, derive close approximations for the tails of the waiting-time distribution functions and compare them to those of the two limiting cases.

Index Terms—Computer performance evaluation, queuing analysis.

I. 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 short note we will consider the case of requests arriving at random (not predictable) times. Given the 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 ($1 \leqslant i \leqslant N$) arrive at a rate λ_i from a Poisson process, have an arbitrary service-time b_i according to a distribution function $F_i(t) = P(b_i \leqslant t)$ 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 ω_i. Without any loss of generality we order the types i of requests such that

$$0 \leqslant \omega_1 < \omega_2 < \cdots < \omega_i < \cdots < \omega_N \tag{1}$$

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

Fig. 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 (nonpreemptive 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

$$q_i(t) = \omega_i - t + T_i, \tag{2}$$

where T_i is the arrival time and ω_i the urgency. If we define the waiting time $w_i(t)$ at time t as

$$w_i(t) = t - T_i \tag{3}$$

Manuscript received December 6, 1978; revised August 25, 1979. This work was partially supported by the 3rd EDP Program of the Federal Government of West Germany under Contract DV 0812248.

The authors are with AEG-TELEFUNKEN, Research Institute of Ulm, Ulm, West Germany.

0098-5589/80/0500-0304$00.75

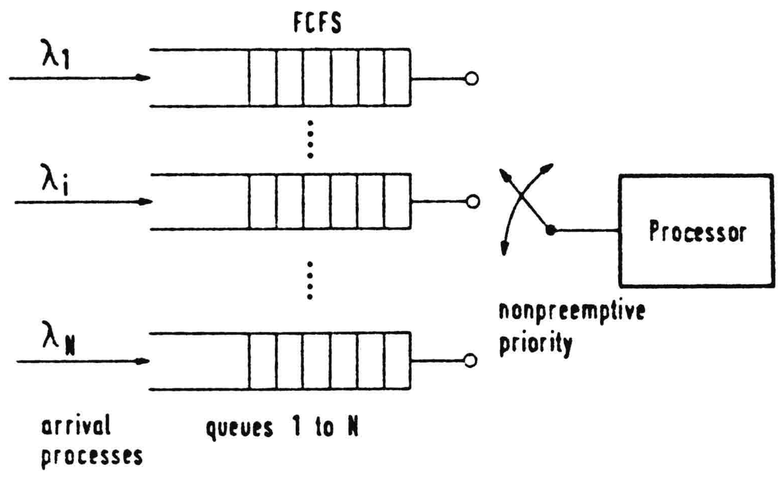

Fig. 1. Model of a real-time computer-control system, λ_i = arrival rate.

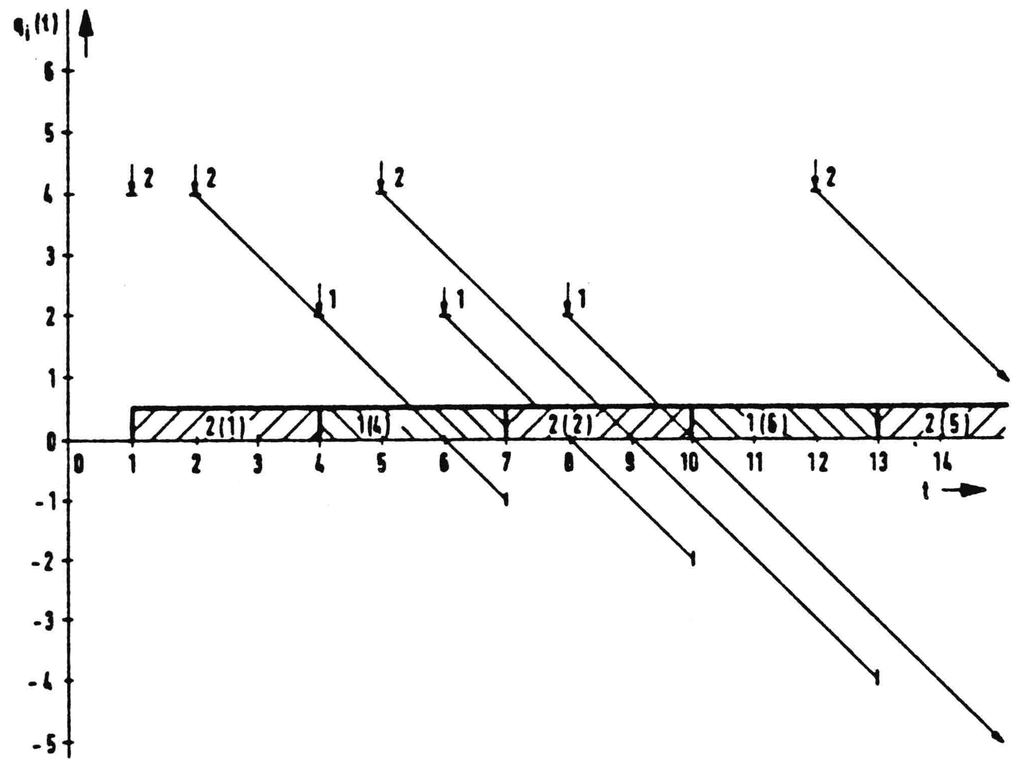

Fig. 2. Example showing the dynamic priority $q_i(t)$ over time t ($N = 2$, $\omega_1 = 2$, $\omega_2 = 4$, equal and constant service times). Times of arrival of type-i requests are marked with an indexed arrow $\downarrow i$.

then the current priority is

$$q_i(t) = \omega_i - w_i(t) \tag{4}$$

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 uniformly. Requests of the same type are processed in the order they arrive [first-come first-serve (FCFS)]. Since requests are serviced according to individual urgencies and arrival times, we speak of a relative-urgency (RU) discipline. One possible interpretation is that a request's deadline is reached if its waiting time $w_i(t)$ equals ω_i.

Fig. 2 shows an example with $N = 2$ types of requests with urgencies $\omega_1 = 2$, $\omega_2 = 4$, and constant and equal service times $\beta_1 = \beta_2 = 3$. The ordinate shows the actual dynamic priority $q_i(t)$ and the abscissa shows the time t. At its arrival time $t = T_i$ each request has a priority $q_i(0) = \omega_i$ because $w_i(t) = 0$. The priority of a waiting request increases linearly [the value of $q_i(t)$ decreases] with its waiting time $w_i(t)$. 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.

At $t = 1$ a busy period starts. The next decision is needed at $t = 4$ when service of the first request is finished. At this time both the type-2 request which arrived at $t = 2$ and the type-1 request which just has arrived have the same priority. Such ties are broken by servicing the most urgent type, i.e., type 1. At $t = 7$ the next decision is needed. At this time the type-2 request which arrived at $t = 2$ will be elected, although a type-1 request is waiting also. This follows from a comparison of the actual priorities at $t = 7$. Apparently less urgent requests are preferred if they have waited long enough.

In addition to the nonpreemptive discipline (RU-NPRE) in Section IV we will consider a few examples of a preemptive discipline (RU-PRE). The following abbreviations are used throughout:

w_i, b_i	random variables: waiting and service time of a type-i request;
d.f.	distribution function;
$W_i(t) = P(w_i \leqslant t)$	waiting-time d.f. of a type-i request;
$W_i^{(r)}$, $W_i^{(1)} = W_i$	rth moment of $W_i(t)$;
$\lambda = \Sigma_{j=1}^{N} \lambda_j$	total arrival rate of all requests together;
$W(t) = \lambda^{-1} \Sigma_{i=1}^{N} \lambda_i W_i(t)$	common waiting time d.f. of all requests together;
$F_i(t) = P(b_i \leqslant t)$	service-time d.f.;
$\beta_i^{(r)}$, $\beta_i^{(1)} = \beta_i$	rth moment of service time d.f.;
$\rho_i = \lambda_i \beta_i$	offered traffic of type-i requests;
$\rho_{\leqslant i} = \Sigma_{j=1}^{i} \rho_j$	offered traffic of types 1 to i;
$\rho = \rho_{\leqslant N}$	total offered traffic;
$F(t) = \lambda^{-1} \Sigma_{i=1}^{N} \lambda_i F_i(t)$	common service-time d.f. of all requests together;
$\beta^{(r)} = \lambda^{-1} \Sigma_{i=1}^{N} \lambda_i \beta_i^{(r)}$	rth moment of $F(t)$;
$C^2 = \beta^{(2)}/\beta^2 - 1$	squared coefficient of variance of $F(t)$;
$\overline{\omega} = \rho^{-1} \Sigma_{i=1}^{N} \rho_i \omega_i$	mean weighted urgencies;
$\text{var}(\omega) = \Sigma_{i=1}^{N} (\rho_i/\rho) \omega_i^2 - \overline{\omega}^2$	variance of weighted urgencies;
$M/A/1/B$	model with Markovian (= Poisson) arrival process, a distinct service-time d.f. ($A = D$ deterministic, $A = M$ Markovian, $A = G$ general), one server, and service discipline B;
HOL	head of the line (static nonpreemptive priority) discipline.

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 Fig. 1 applies, except that there is no urgency ω_i.

If one is interested in the probability of missing deadlines, then one should know the waiting-time d.f. $W_i(t) = P(w_i \leqslant t)$. From this the probability that the waiting time w_i is not greater than a given time t can be obtained.

We will consider the model in Fig. 1 only in a state of equilibrium, which occurs only when the total offered traffic $\rho < 1$. In disciplines with static priorities, a stationary equilibrium is achieved for requests with high priorities, sometimes even when $\rho > 1$. However, we will not consider this case here.

II. Waiting-Time D.F.'s in the RU and HOL Disciplines; Examples

Let us assume $N = 4$ different types of requests, each with its arrival rate λ_i, urgency ω_i, and service-time d.f. $F_i(t)$. Since the waiting-time d.f. cannot be calculated, we will consider two examples which were simulated. We used an $M/D/1$ model with all requests having a constant service time.

Example 1: $\rho_i = \rho/N$, $\beta = 2.39$, $C = 0.81$, $\overline{\omega} = 30$, var $(\omega) = 125$

i	1	2	3	4
β_i	1	3	5	7
ω_i	15	25	35	45

Example 2: corresponds to Example 1 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
ω_i	15	25	35	45

Note that our findings stated in the rest of this paper are also valid for quite different load parameters of our model. This can be seen from [4] where results of $M/G/1$ and $M/M/1$ models are presented.

From earlier investigations, cf. [1] and [2], it is known that the smaller the variance of the urgency numbers is, the less the mean waiting times w_i of the RU-NPRE discipline differ from the mean waiting time

$$W_{\mathrm{FCFS}} = 1/2\lambda\beta^{(2)}/(1-\rho) \tag{5}$$

in the first-come first-serve discipline. On the other hand if the variance var (ω) is very large, the RU-NPRE discipline produces mean waiting times approximating those of the head of the line (HOL) discipline, a discipline with static nonpreemptive priorities

$$W_{i\,\mathrm{HOL}} = \lambda\beta^{(2)}/[2(1-\rho_{\leqslant i})(1-\rho_{<i})]. \tag{6}$$

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.

Assume $p(w_i)$ to be the equilibrium probability that a type-i request must wait no longer than w_i units of time for service. For $0 \leqslant f \leqslant 1$, define

$$Q_i(f) = \inf\{w_i | p(w_i) \geqslant f\}. \tag{7}$$

This is the waiting time corresponding to fractile f in the cumulative waiting-time d.f. of type-i requests. $Q_i(f)$ usually is called the f-quantile.

A. Observations from a Simulation Model

From a simulation model we have made three observations.

First observation: 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. This seems to indicate that for every $M/G/1$/RU-NPRE model we have

$$\lim_{f \to 1} [Q_i(f) - Q_j(f)] = \omega_i - \omega_j, \qquad (i, j = 1, 2, \cdots N). \tag{8}$$

As the probability f approaches 1, the distance between the waiting-time d.f.'s for type-i and type-j requests asymptotically approaches the difference in their urgencies ω_i and ω_j within the accuracy obtainable by simulation. For $f < 1$, this distance is smaller than the difference $\omega_i - \omega_j$.

Second observation: the waiting-time d.f.'s in the RU discipline only depend on the differences of the urgency numbers, i.e., for both parameter sets $\{\omega_1, \omega_2, \cdots, \omega_N\}$ and $\{\omega_1 + r, \omega_2 + r, \cdots, \omega_N + r\}$ (r real number) the same d.f.'s appear.

Third observation: the waiting-time d.f. in the FCFS discipline runs parallel to the tails of the waiting-time d.f.'s in the RU discipline. This and the two other observations seem to be valid for every $M/G/1$/RU model. They are the basis for an approximate analytical computation of the waiting-time d.f.'s in the RU discipline.

An example: Consider an example with $p = W_i(t) = p = W_i(t) = 0.985$ and $\rho = 0.75$. From Fig. 3 it can be seen that the p-quantile of a type-i request, for which the abbreviation $Q_i(p)$ just has been introduced, equals the urgency number ω_i. Apparently for any other probability $f (f \geqslant p)$ the following approximation holds:

$$Q_i(f) - Q_j(f) = w_i - w_j \approx \omega_i - \omega_j. \tag{9}$$

Observe that $f = P(w_i \leqslant Q_i(f)) = P(w_j \leqslant Q_j(f))$.

Very large waiting times of different type requests differ by the constant difference in their urgencies. Smaller waiting times differ by less. The property formulated in (8) implies that for very large waiting times w_i the probability $P(w_i \leqslant t)$ of meeting deadlines is practically independent of the type of request, i.e., all requests are handled with equal fairness.

If the condition

$$\min_i \{\omega_i\} >> \max_i \{\beta_i\}, \qquad (1 \leqslant i \leqslant N),$$

is fulfilled, then the probabilities of meeting the deadlines $P(w_i \leqslant \omega_i)$ are about the same, regardless of type.

B. Dependency on the Total Offered Traffic

Fig. 3 shows simulation results of the waiting-time d.f.'s for Example 1. Two different total offered traffics $\rho = 0.75$ and 0.85 are assumed. If the total offered traffic is increased, the probabilities for meeting a given set of waiting times $\{\omega_1, \omega_2, \omega_3, \omega_4\}$ are decreased. Our observations, mentioned above, only can be made for such offered traffics for which, in fact, the RU discipline does not degenerate to the HOL discipline. Such degeneration appears if the probability of more than one request being waiting is small enough to prevent a nonurgent request to become (during its wait) more urgent than an urgent request which arrived later.

C. Dependency of the Waiting-Time d.f.'s on the Variance of the Urgencies

From Fig. 4 waiting-time d.f.'s in the FCFS and RU disciplines can be seen. We used two sets of urgency numbers having a different variance var (ω). If var (ω) is small the resulting dotted curves are very much closer to the waiting-time d.f. in the FCFS discipline than is the case for a four times larger variance (dashed curves).

D. Dependency on the Attachment of a Set of Mean Service Times to a Set of Urgency Numbers

Assume the model in Fig. 1 in the HOL discipline and define the common-mean waiting time W, cf. (6),

$$W = \lambda^{-1} \sum_{i=1}^{N} \lambda_i W_{i\,\mathrm{HOL}}. \tag{10}$$

The common-mean waiting time results from weighting and summing over the individual mean waiting times of priority

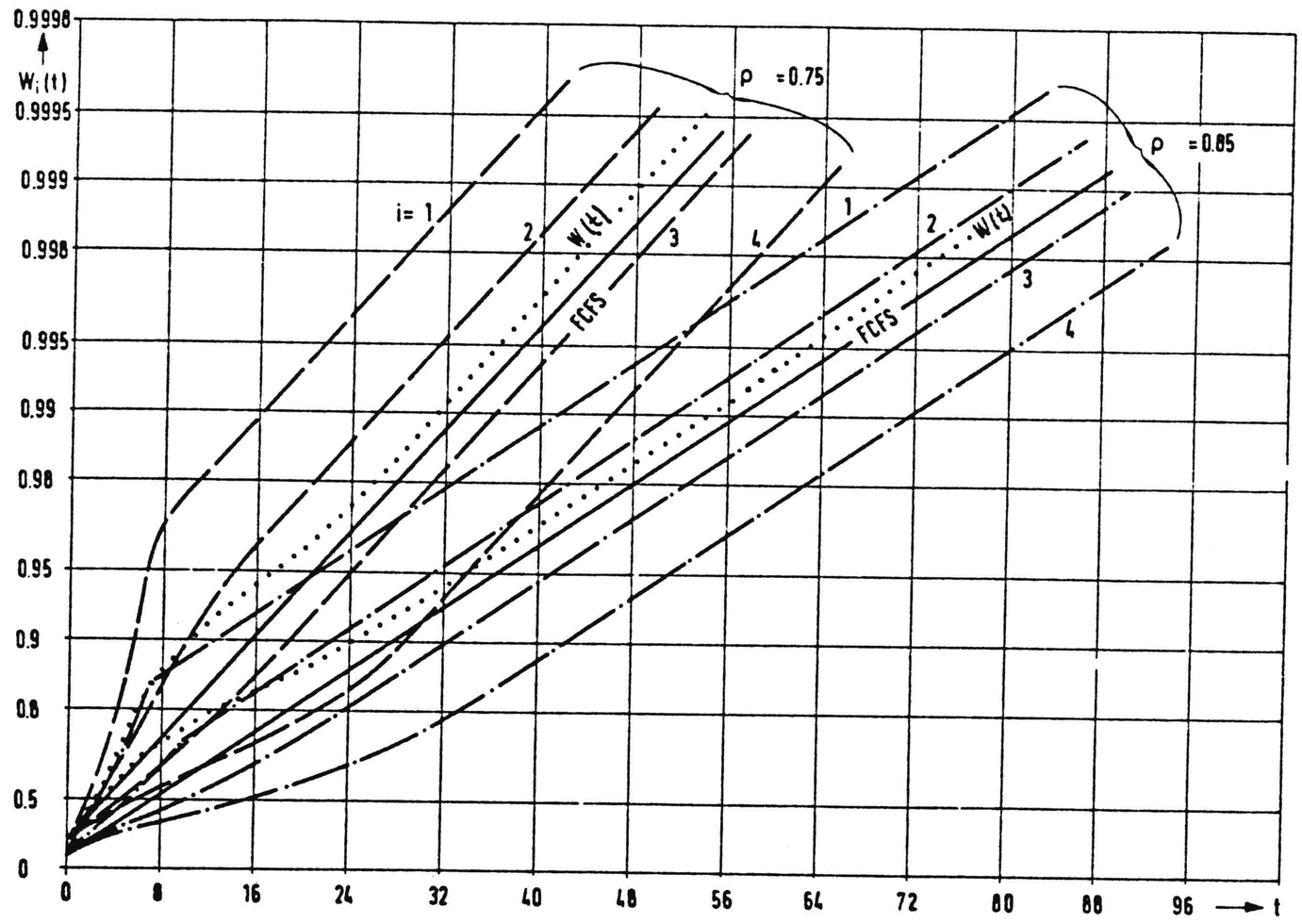

Fig. 3. Simulated waiting-time d.f.'s in the FCFS and RU disciplines, cf. Example 1. The type i and total offered traffic ρ are parameters. The solid and dashed curves result from the FCFS and RU disciplines, respectively. The dotted curve shows the common waiting-time d.f. $W(t)$.

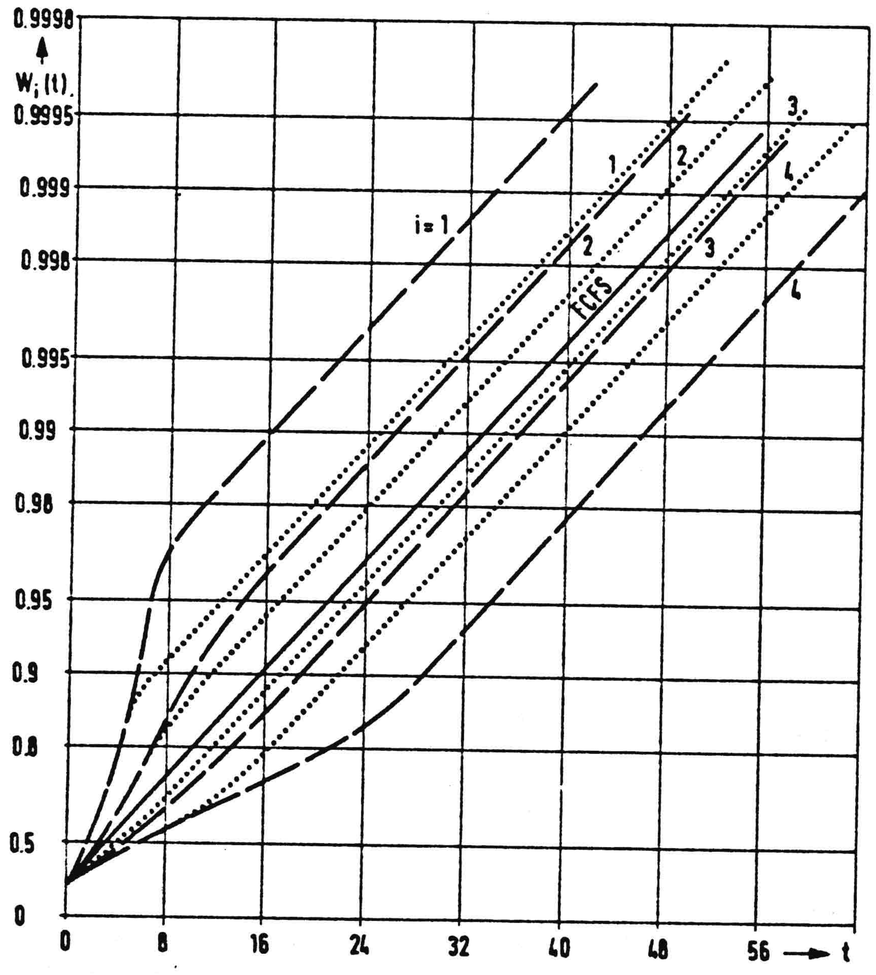

Fig. 4. Waiting-time d.f.'s in the FCFS and RU disciplines, cf. Example 1. The dashed and dotted curves appear for a large and small variance of the urgency numbers, respectively. The total offered traffic is $\rho = 0.75$.

levels i. W equals W_{FCFS}, cf. (5), only if all mean service times β_i are equal. This is not the case in Examples 1 and 2.

From [3] it is known that W can be minimized in a static priority model by satisfying the condition

$$\beta_1 \leqslant \beta_2 \leqslant \cdots \leqslant \beta_i \leqslant \cdots \leqslant \beta_N, \tag{11}$$

where 1 is the highest priority. The kind of service-time d.f.'s themselves do not matter; only (11) must be satisfied. Since W and W_{FCFS} are the means of the corresponding waiting-time d.f.'s $P(w \leqslant t)$ and $P(w_{FCFS} \leqslant t)$, respectively, then for the case of unequal β_i's, if (11) is satisfied, the waiting-time d.f. $P(w \leqslant t)$ must be better in some sense than $P(w_{FCFS} \leqslant t)$.

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

$$w = \lambda^{-1} \sum_{i=1}^{N} \lambda_i w_{iHOL}.$$

It follows that depending on how well or badly the relation

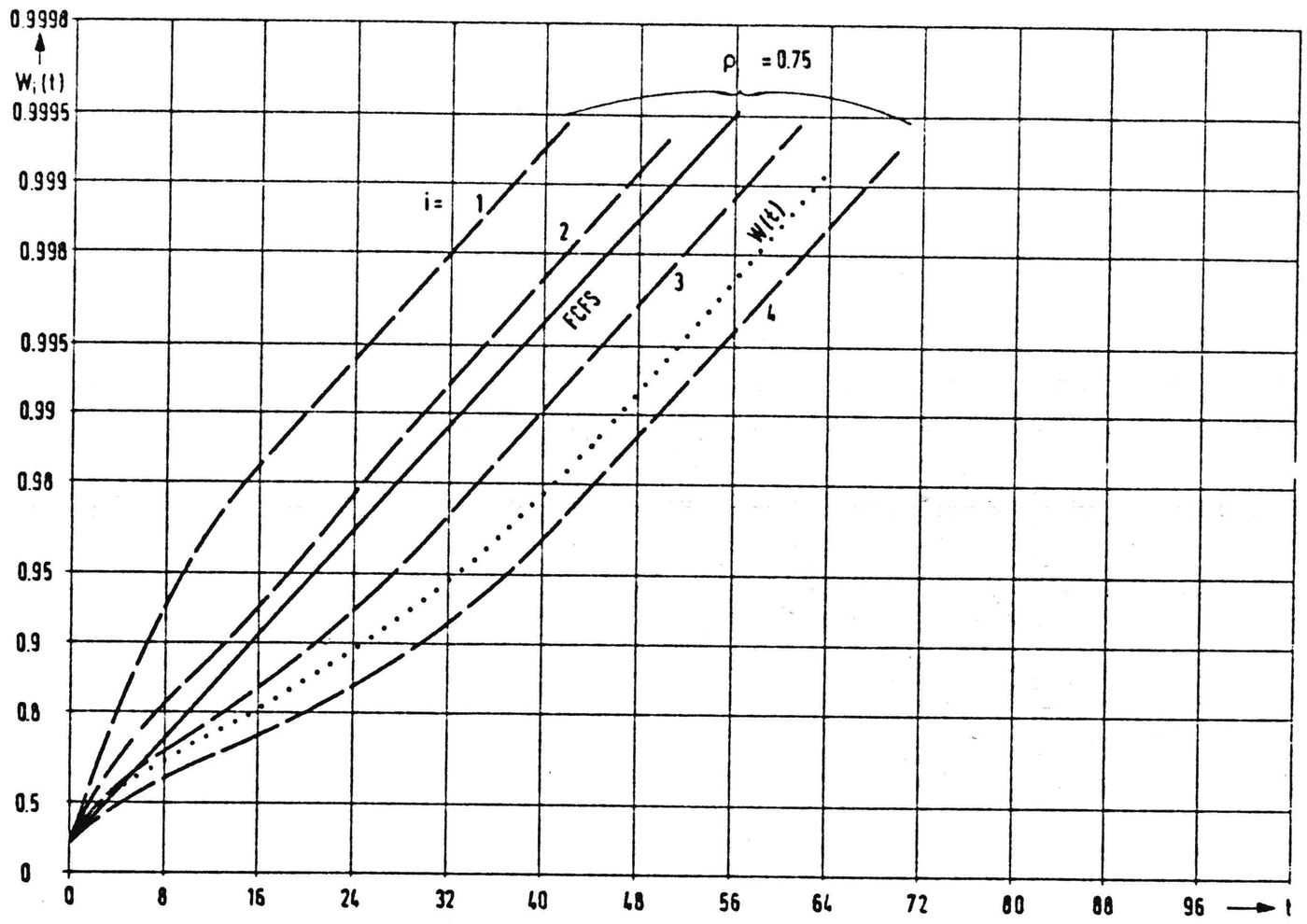

Fig. 5. Waiting-time d.f.'s in the FCFS and RU disciplines, cf. Example 2. The type i and total offered traffic ρ are parameters. The solid and dashed curves result from the FCFS and RU disciplines, respectively. The dotted curve shows the common waiting-time d.f. $W(t)$.

in (11) is observed, the d.f. $P(w_{\text{FCFS}} \leqslant t)$—which is independent of priority assignment—has a worse or better position in the set of curves $P(w_{i\,\text{HOL}} \leqslant t)$, 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. In the next step define the common waiting-time d.f. in the RU discipline.

$$W(t) = \lambda^{-1} \sum_{i=1}^{N} \lambda_i P(w_i \leqslant t).$$

Now it can be argued that assigning small urgency numbers to customers with small mean service times β_i and large ω_i's to customers with large β_i's results in a bad position of $P(w_{\text{FCFS}} \leqslant$) compared to $W(t)$. Just this urgency number assignment minimizes the common-mean waiting time in the RU discipline and optimizes the common waiting-time d.f. $W(t)$. An inverse urgency number assignment results in the least favorable set of curves $P(w_i \leqslant t)$ compared to $P(w_{\text{FCFS}} \leqslant t)$.

The optimal urgency number assignment just introduced is used in Example 1, cf. Fig. 3. It can be seen that $P(w_{\text{FCFS}} \leqslant t)$ (the curve indexed FCFS) has a bad position compared to the set of curves $W_i(t)$. This corresponds to a common waiting-time d.f. $W(t)$ being preferable to $P(w_{\text{FCFS}}) \leqslant t)$.

The opposite is true for Example 2: there, the smaller β_i is, the larger the urgency ω_i. In Fig. 5, consequently, $P(w_{\text{FCFS}} \leqslant t)$ has a good position compared to the set of curves $W_i(t)$.

E. The RU and HOL Disciplines Compared

The RU-NPRE discipline may be thought of as an alternative to the HOL discipline. From [1] and [2] it is known that the greater the variance of the urgency numbers, the more the mean waiting time W_i in the RU discipline differs from W_{FCFS}, and the more $W_{i\,\text{HOL}}$ is approximated. Now, depending on the set of ω_i's, the question arises how the waiting-time d.f.'s in the RU and HOL disciplines will differ. First we recall that only the variance of the urgency numbers plays a role. Our observation is that the waiting-time d.f.'s in the HOL and RU disciplines are very similar for "small" waiting times, cf. Fig. 6. But what a small waiting time is heavily depends on the total offered traffic ρ and the variance of the ω_i's.

One substantial difference between the HOL and the RU disciplines 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 ω_i's [or a comparable set, cf. the second observation following (8)] with nearly the same probability, the corresponding probabilities $P(w_i \leqslant \omega_i)$ may differ by factors in the HOL discipline. In the RU discipline all requests are handled on an equal basis while the opposite is true for the HOL discipline.

By choosing the urgency numbers, it is possible in the RU discipline to relate the waiting-time d.f.'s to one another in a prescribed way which is impossible in the HOL discipline.

III. Approximation of the Waiting-Time D.F.'s

A. Previously Known Results

Back in 1962 Jackson [2] considered a model with discrete time and a nonpreemptive RU discipline. The service times of all requests were taken from one geometric distribution, and the requests themselves were taken from a Bernoulli arrival process. The individual urgency ω_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 the model considered in this paper 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

$$\lim_{f \to 1} [Q_i(f) - Q_j(f)] = \omega_i - \omega_j \qquad \text{[cf., (8)]} \tag{12}$$

and

$$\lim_{f \to 1} [Q_i(f) - Q(f)] = \omega_i - \bar{\omega}'. \tag{13}$$

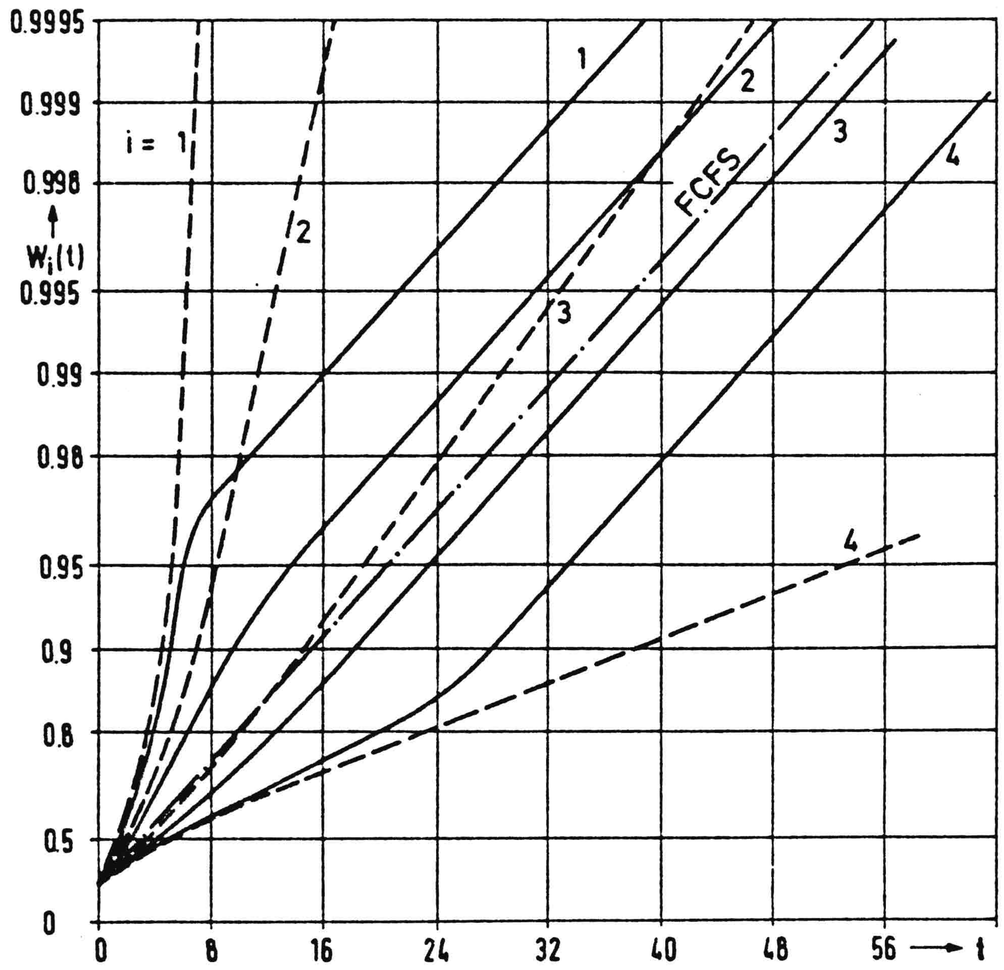

Fig. 6. Waiting-time d.f.'s in the FCFS, RU (solid lines), and HOL (dashed lines) disciplines. Parameters of Example 1, $\rho = 0.75$.

$\bar{\omega}'$ is defined by an expression which for model $M1$ may be approximated by the mean weighted urgencies $\bar{\omega}$, cf. list of abbreviations. $Q(f)$ is the f-quantile of the waiting-time d.f. of a single-queue FCFS model.

B. Approximation of $\bar{\omega}'$ in (13)

Our simulation results support the suspicion that (12) and (13) seem to be generally applicable, regardless of the service-time d.f.'s assumed. We have found that the difference between the waiting-time d.f. in the FCFS discipline and the curves, constructed from (12) using $\bar{\omega}$ instead of $\bar{\omega}'$, always remains small. Only differences up to 9 percent have been observed [4]. However, these differences are considerably larger than those observed by Jackson in 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.

Our observation is that the difference $(\bar{\omega} - \bar{\omega}')$ is, depending on the example, greater or smaller than zero. The reason for this can be plausibly explained as the dependancy of the waiting-time d.f.'s in the RU discipline on the urgency number assignment being optimal in Example 1 and worst in Example 2 (see Section II-D).

Note that the load to the RU model is the same in both Examples 1 and 2. Hence $P(w_{FCFS} \leqslant t)$ is the same for both examples which is not the case for the curves $W_i(t)$ and $W(t)$. Only the assignment of the same set of urgency numbers is different, which is precisely the reason why $\bar{\omega}'$ differs from $\bar{\omega}$. Our observation is that $\epsilon = \bar{\omega} - \bar{\omega}'$ for all the examples studied was always very small (a few percent of $\bar{\omega}$). Therefore we decide to accept (13) with $\bar{\omega}'$ substituted by $\bar{\omega}$ as an approximation which leads to

$$\lim_{f \to 1} [Q_i(f) - Q(f)] = \omega_i - \bar{\omega}. \tag{14}$$

C. Approximations of the Waiting-Time d.f.'s in the FCFS and RU Discipline

From Figs. 3, 4, and 5 it can be seen for our two examples that the waiting-time d.f.'s, when plotted semilogarithmically, run as parallel straight lines from about $f = w_i(t) = 0.98$ on.

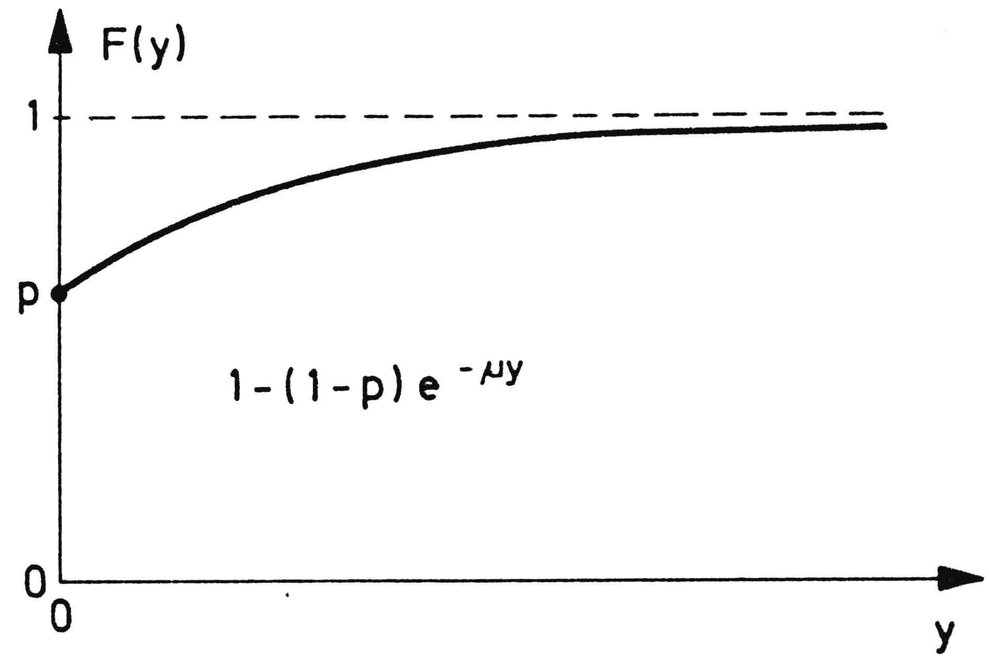

Fig. 7. A degenerated exponential d.f.

We are now going to approximate these functions through proper formulas.

1) *The waiting-time d.f. in the FCFS discipline:* The waiting-time 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 d.f.'s found by simulation, especially in the tail, can be approximated very well by a degenerated exponential d.f., cf. Fig. 7,

$$F(y) = 1 - (1 - p)\, e^{-\mu y}. \tag{15}$$

The rth moment of a degenerated d.f. is

$$\gamma^{(r)} = r!(1 - p)/\mu^r.$$

By setting equal the first two moments of the waiting-time d.f. in the FCFS discipline and the degenerated d.f., namely, $W_{FCFS} = \gamma^{(1)}$ and $W^{(2)}_{FCFS} = \gamma^{(2)}$, it is possible to compute the two unknown parameters p and μ of (15). Thereby a degenerated d.f. is defined having the same first two moments as the waiting-time d.f. in the FCFS discipline. Using (5) and

$$W^{(2)}_{FCFS} = 2W^2_{FCFS} + \lambda\beta^{(3)}/[3(1 - \rho)] \tag{16}$$

if

$$C^2 = W^{(2)}_{FCFS}/W^2_{FCFS} - 1 \tag{17}$$

is the squared coefficient of variance we find

$$p = \frac{C^2 - 1}{C^2 + 1} \quad \text{and} \quad \mu = \frac{2}{W_{FCFS}(1 + C^2)}. \tag{18}$$

Equations (15) and (18) define an approximation of $P(w_{FCFS} \leqslant t)$, whose first two moments are correct and which belongs to the family of the degenerated exponential d.f.'s

$$P(w_{FCFS} \leqslant t) \approx P(t) = 1 - (1 - p)\, e^{-\mu t}. \tag{19}$$

In semilogarithmic scaling this function appears as a straight line.

We are now going to study the quality of (19) as an approximation for $P(w_{FCFS} \leqslant t)$. Fig. 8 shows simulated (+) and approximated waiting-time d.f.'s in the FCFS discipline using the parameters of Examples 1 and 2. (We have mentioned already that both examples should produce the same waiting-time d.f. in the FCFS discipline.) All our investigations, cf. [4], showed that, especially in the tails of the waiting-time d.f.'s, simulation and approximation fit together very well.

2) *The tail of the waiting-time d.f. in the RU discipline:* In our first approximation, cf. (14), we assumed the tails of the waiting-time d.f.'s in the FCFS and RU discipline to run parallel. In our second approximation we introduced (19) to ap-

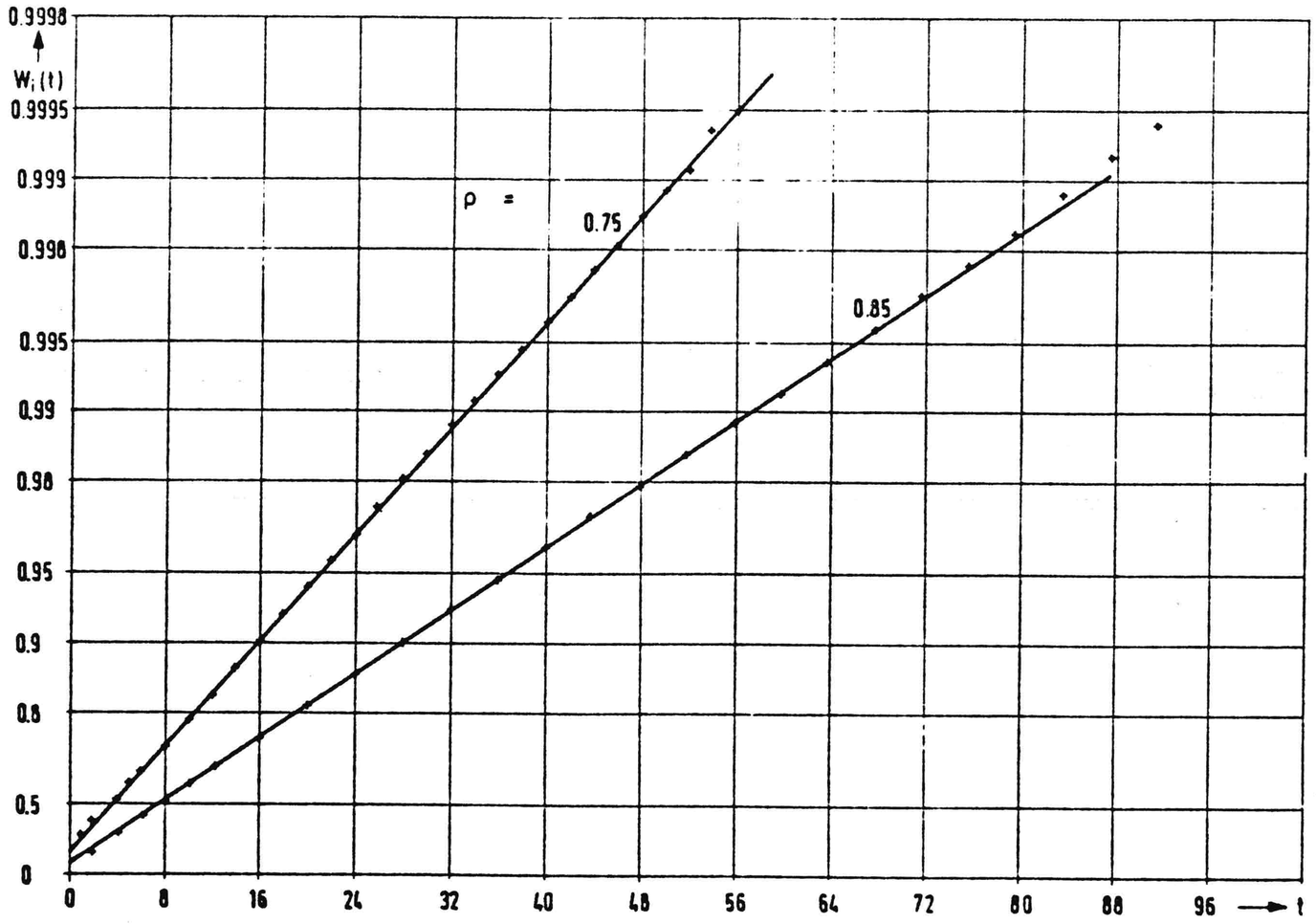

Fig. 8. Waiting-time d.f.'s in the FCFS discipline, cf. Examples 1 and 2. Approximations by means of (19) (lines) and simulation results (+) are shown. Parameter is the total offered traffic.

proximate the tail of the waiting-time d.f. in the FCFS discipline. Combining these two results, we find as the third approximation in this paper a close formula to describe the tails of the waiting-time d.f.'s in the RU discipline.

From (14) it follows immediately that for every pair of f-quantiles $(f \to 1)$ $Q_i(f) = w_i$ and $Q_{\mathrm{FCFS}}(f) = w_{\mathrm{FCFS}}$ of the waiting-time d.f.'s in the RU and FCFS disciplines, respectively, one may write

$$w_i - w_{\mathrm{FCFS}} = \omega_i - \bar{\omega},$$

cf. (9). Recalling (19)

$$P(w_{\mathrm{FCFS}} \leqslant t) \approx 1 - (1 - p)\, e^{-\mu t}$$

and inserting

$$w_{\mathrm{FCFS}} = w_i - \omega_i + \bar{\omega}$$

we have

$$P([w_i - \omega_i + \bar{\omega}] \leqslant t) \approx 1 - (1 - p)\, e^{-\mu t}$$

$$P(w_i \leqslant t) \approx 1 - (1 - p)\, e^{-\mu(t - \omega_i + \bar{\omega})}. \qquad (20)$$

This function defines the tails of the waiting-time d.f.'s in the RU-NPRE discipline.

A comparison of (19) and (20) reveals that they differ by a factor. We aim at an expression for the waiting-time d.f.'s in the RU discipline which has the form of a degenerated exponential d.f., cf. (15). Instead of (20) we write

$$P(w_i \leqslant t) \approx 1 - (1 - p_i)\, e^{-\mu t} \qquad (21)$$

and find, by setting equal the right-hand sides of (20) and (21),

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

Note that p and p_i are the probabilities which arise for $t = 0$ in (15) and (21), cf. Fig. 7. From simulation experiments we know that (21) is a good approximation, if either the total offered traffic is large, $\rho > 0.6$, or the urgency numbers have a small variance, or both. Otherwise, the d.f.'s in the RU discipline appear to be similar to those of the HOL 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.

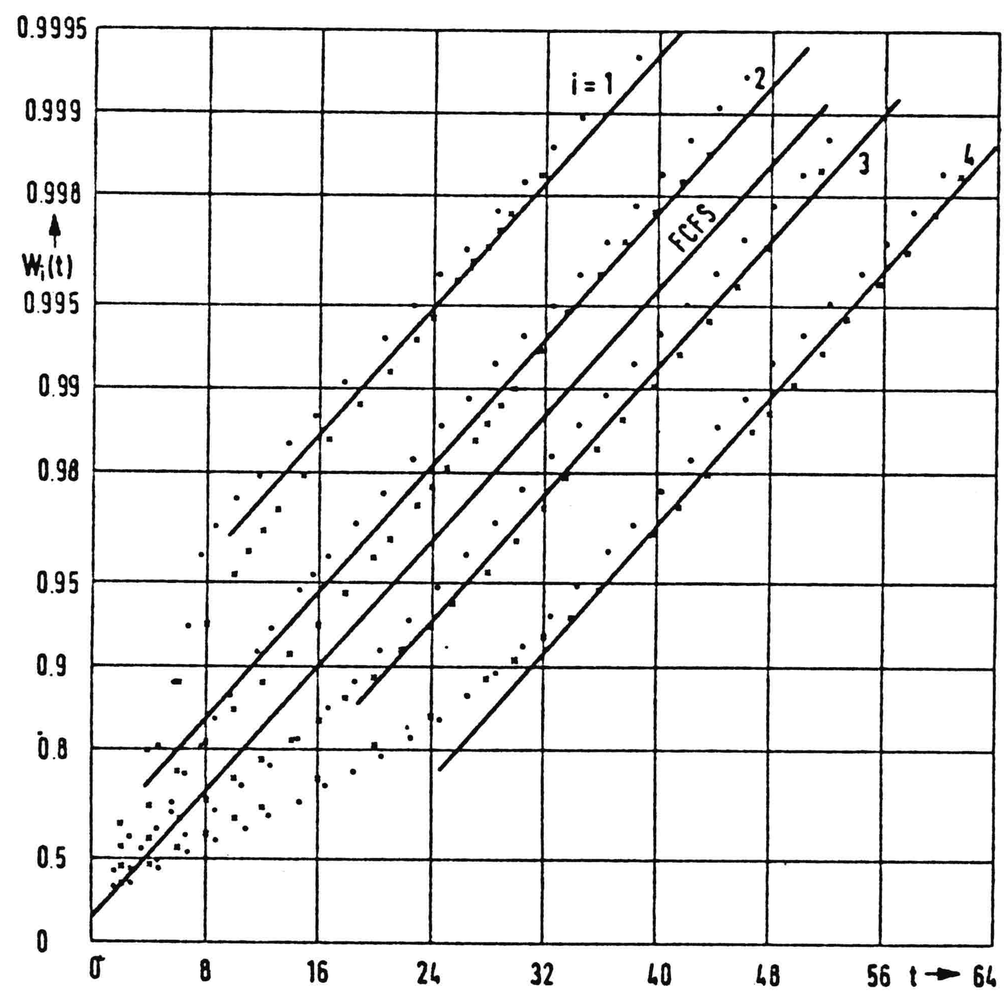

Fig. 9. Waiting-time d.f.'s in the FCFS and RU disciplines, cf. Examples 1 and 2, approximated by means of (21), (lines) and simulation results (·, X). Additionally, the FCFS approximation is shown. The total offered traffic is $\rho = 0.75$.

Fig. 9 shows both simulation results (·, X) and approximate waiting-time d.f.'s computed from (21), which appear as straight lines for the examples defined in Section II. It can be seen that for large waiting times, simulation and approximation agree very well. Deviations are typically below the 10 percent range.

In Fig. 9 we once again can study the results of a different assignment of a set of urgency numbers ω_i to a set of mean

service times β_i as defined by Examples 1 and 2. The approximate waiting-time d.f.'s are derived from the d.f. of the FCFS discipline and therefore are the same for both Examples 1 and 2. As a consequence of the optimal urgency number assignment as defined in Section II-D, which is realized in Example 1, the real d.f.'s ($\cdot$) compare favorably to the computed ones (lines). On the other hand the d.f.'s (X) of Example 2, in which the ω_i are especially poorly assigned to the β_i, are less favorable than the computed ones. Note that the approximation by (21) does not take into account the urgency-number assignment. This should be the reason for the deviation between simulated and computed d.f.'s observed.

3) *The waiting-time d.f.'s in the RU discipline for small waiting-times:* Equation (21) does not apply to small waiting times. In [1] we used a simple approximation of the waiting-time d.f. in the RU discipline to compute the corresponding mean waiting time W_i of type-i requests

$$P(w_i > t) = \rho e^{-\rho t/W_i}. \tag{22}$$

This simple approximation only depends on the total offered traffic ρ and W_i. Fig. 10 shows this approximation (lines) compared to simulation results (X) for Example 1. The most unsatisfactory approximation appears to be for type-1 requests which was not needed in [1]. It can be seen that (22) in general cannot be called a good approximation.

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

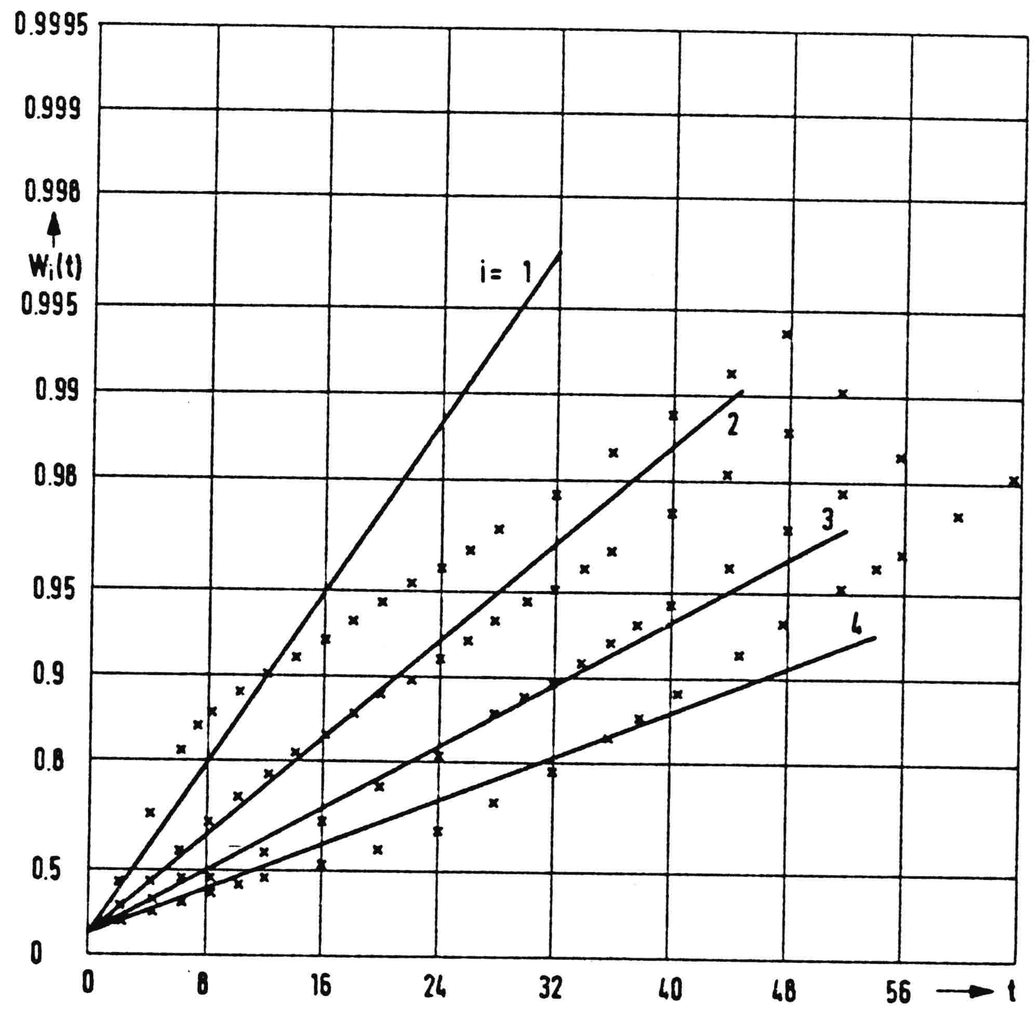

Fig. 10. Waiting-time d.f.'s in the RU discipline, cf. Example 1, approximated by means of (22), (lines) and simulation results (X). The total offered traffic is $\rho = 0.85$.

D. Mean Waiting Time of Requests Which Have Missed Their Deadlines

From (21) the approximate probability of a type-i request missing its deadline (at $t = \omega_i$), presuming a sufficiently large t, can be computed

$$P(w_i > \omega_i) \approx (1 - p)\, e^{-\mu\bar{\omega}}.$$

The waiting-time d.f. of requests missing their deadlines is approximately given by

$$P(w_i \leqslant t \mid t > \omega_i) = [P(w_i \leqslant t) - P(w_i \leqslant \omega_i)]/[1 - P(w_i \leqslant \omega_i)]$$

$$P(w_i \leqslant t \mid t > \omega_i) = 1 - e^{\mu\omega_i} e^{-\mu t}.$$

Apparently a degenerated exponential d.f. arises whose mean

$$E(w_i \mid t > \omega_i) = E_i = e^{\mu\omega_i}/\mu$$

only depends on μ and ω_i. Note that μ can be computed from (18). The approximate mean waiting times E_i of requests having missed their deadlines differ approximately in the factor e^{ω_i}.

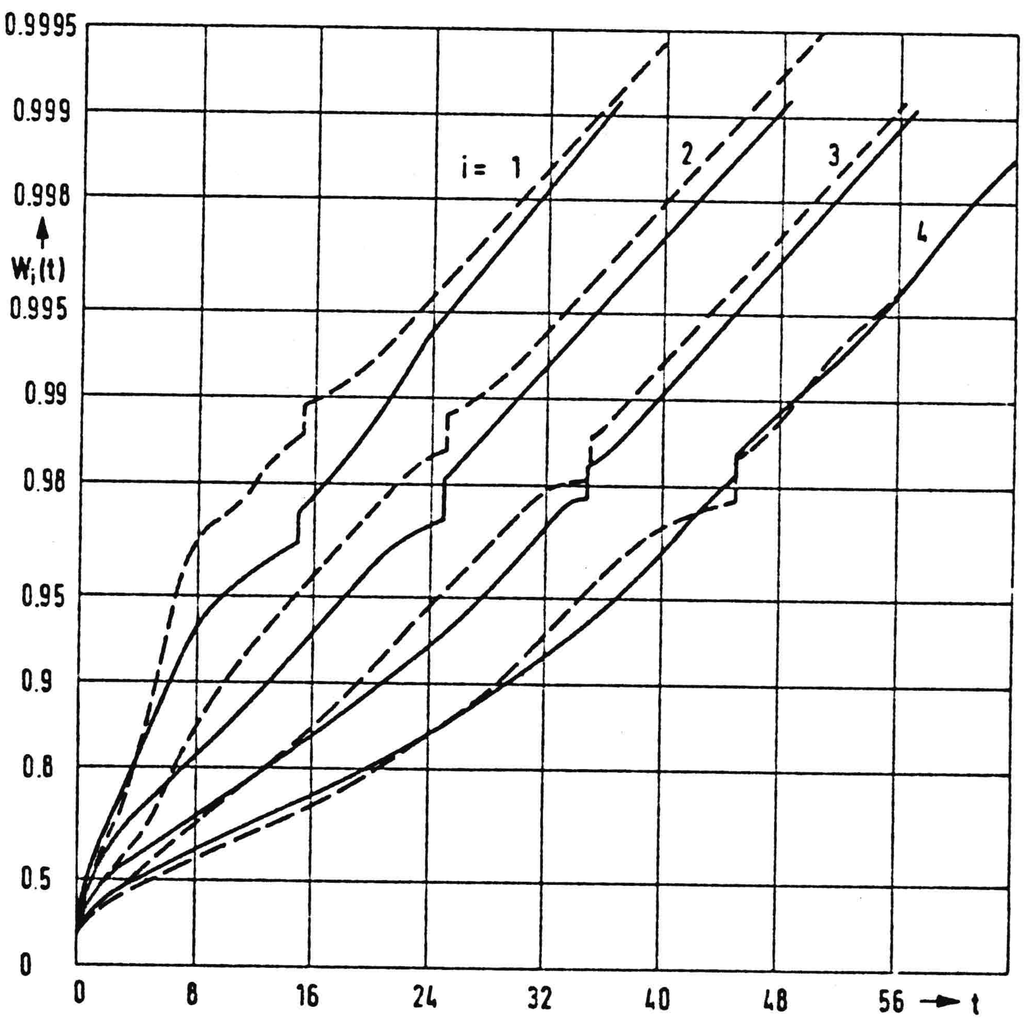

Fig. 11. Simulated waiting-time d.f.'s in the RU-PRE discipline for Examples 2 (solid lines) and 1 (dashed lines). The total offered traffic is $\rho = 0.75$. The jumps appear at $t = \omega_i$.

IV. Some Typical Properties of the Waiting-Time D.F. in the Preemptive RU Discipline

In the preemptive RU (RU-PRE) discipline a type-i request may be interrupted (preempted) by a type-j request if $w_i < \omega_i$ and $w_j = \omega_j$. 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 (4) with $w_i(t)$ being the total time such a request remains in the system without being serviced. The urgency ω_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 nonpreemptive and preemptive versions of the RU discipline a request's deadline may be defined to be reached if $w_i(t)$ equals ω_i. If the waiting time w_i of a request being serviced is less than its voluntary waiting time ω_i, while another type-j request has already waited its voluntary time ω_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 request whose deadline has been reached.

The waiting-time d.f.'s in the RU-PRE discipline (presuming the data of Examples 1 and 2) shown in Fig. 11 were determined by simulation. Comparable results using the RU-NPRE discipline are shown in Fig. 9. It can be seen that for the same total offered traffic distinct differences in the waiting times of

both disciplines arise. Especially worth noting is the jump at $t = \omega_i$ 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 waiting-time d.f.'s in the tail run parallel, as was the case for the RU-NPRE discipline (cf. [6]). For all examples considered the slope of these tails in the RU-PRE and RU-NPRE disciplines was not found to be the same (the same offered traffic ρ is 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., $C_i > 1$. For small C_i's our simulation supports the conjecture that interruptions of requests in favor of others, having a larger expected service time, is disadvantageous for meeting deadlines. The deadlines of some types of service requests are then met better with the RU-NPRE discipline than with the RU-PRE discipline.

References

[1] B. Walke, "Improved bounds and an approximation for a dynamic priority queue," in *Proc. 3rd Int. Symp. Modeling and Perform. Eval. Comput. Syst.*, Bonn, Germany, Oct. 1977. Amsterdam, The Netherlands: North-Holland, 1978, pp. 321–346.

[2] J. R. Jackson, "Waiting time distributions for queues with dynamic priorities," *Naval Res. Logist. Quart.*, vol. 9, pp. 31–46, Mar. 1962.

[3] R. W. Conway, W. L. Maxwell, and L. W. Miller, *Theory of Scheduling*. Reading, MA: Addison-Wesley, 1967.

[4] B. Walke and W. Rosenbohm, "Waiting-time distributions for deadline oriented serving," in *Proc. 4th Int. Symp. Modeling and Perform. Eval. Comput. Syst.*, Vienna, Austria, Feb. 6–8, 1979. New York: North-Holland, 1979.

[5] E. B. Veklerov and I. M. Dukhovnyi, "Queueing systems with time-limited categorical priority," *Probl. Inform. Transm.*, vol. 12, pp. 73–79, Jan.–Mar. 1976.

[6] H. M. Goldberg, "Jackson's conjecture on earliest due date scheduling," Dep. Finance and Management Sci., Univ. Alberta, Edmonton, Alberta, Canada, Sept. 1978.