Preemption-Distance Priorities in Real-Time Computer Systems

By **Ulrich Herzog**

A Report from the Institute for Switching and Data Technics, University of Stuttgart

1. General Remarks

For operating real-time computer systems it is necessary to serve special pressing demands with preemptive priorities. On the other hand, there are a lot of demands being less urgent which do not justify preemption. Sometimes even preemptive priorities are nonsensical (e. g. if the preempting demand is more important than the interrupted one; however, the time spent on interrupting is greater than the remaining rest-service time of the low priority demand).

Therefore, most of real-time computer systems serve the various demands by a mixed mode [1, 2]. E. g. the new Electronic Data Switching System for Data Communication of the Federal German Post Office [3, 4] is using for the input/output control of the storage unit a reasonable combination of preemptive and non-preemptive priorities, which may be called "priorities with variable preemption distance" and for short "preemption-distance priorities".

Preemption-distance priorities are controlled such, that demands of the priority class p ($p = 1, 2, \ldots, P$; class 1 most urgent) only interrupt demands of priority classes less or equal to $(p + \xi)$. The so called preemption-distance ξ can be chosen between 1 and P (cf. *Fig. 1*). After interruption, service is resumed at the point where it was left off.

This paper deals with the calculation of the expected waiting time, the expected system-response time and the expected number of items in the system waiting simultaneously for service for each priority class p. The investigated system may be abbreviated by $(M(p)/E_k(p)/1)$ with preemption-distance priorities, i. e. we assume Poisson-input and Erlang-distributed processing times of different order and different mean values $b(p) = 1/\mu(p)$ per each priority class (cf. *Fig. 2*). Hence, e. g. processing time for demands of some priority classes may be constant ($k = \infty$), for others negativ exponential ($k = 1$) and for the remaining ones Erlang-distributed between these two boundary values ($0 < k < \infty$). Demands of the same priority class are served in the order of their arrival (FIFO).

If we choose the preemption-distance ξ equal to 1 or P we yield the well known formulas for systems with pure preemptive or pure non-preemptive priorities, respectively.

By reason of simplicity, the preemption-distance ξ has been chosen equally for all priority classes p. When introducing "empty" priority classes (call rate zero) between the actual ones, it is possible to vary the real preemption-distance for each actual priority class individually. Therefore it is possible to generate arbitrary sequences of preemptive and non-preemptive priorities. The only mathematically treated combination of preemptive and non-preemptive priorities, known from literature [5] is included (some classes, interrupting all demands of lower priority and some classes, which do not interrupt any demand; examples cf. *Fig. 3* and *Fig. 4*).

2. Recursive Equation for the Expected Waiting Time for all Demands

Supposedly, an arbitrarily chosen demand of priority class p enters the system. The expected system-response time (time spent in the system, waiting and being processed) for this demand is composed of the following five terms:

a) the expectation $b_R (\leqq p + \xi - 1)$ of the remaining rest-service time for demands of the priority classes 1 to $(p + \xi - 1)$ present at its time of arrival in the server and not being interrupted by the considered p-demand,

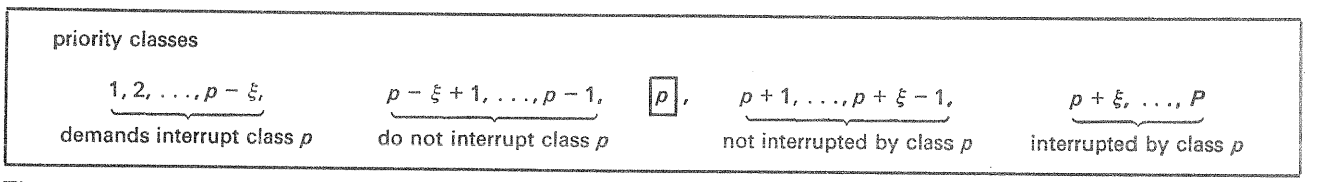

Fig. 1.
Preemption-distance priorities. (Examples see Figrs. 3 and 4).

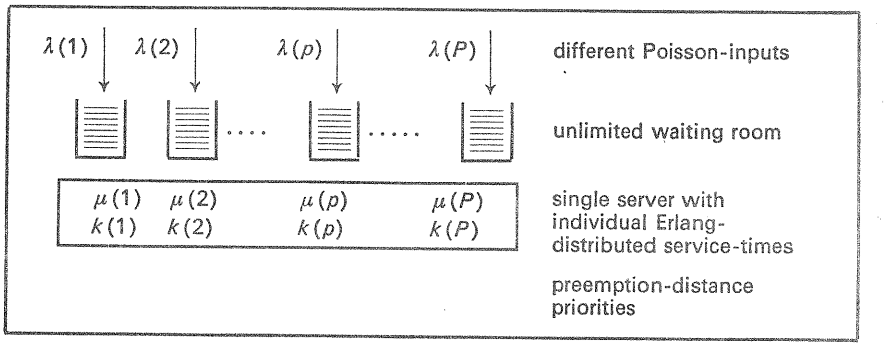

Fig. 2.
The investigated system:
$(M(p)/E_k(p)/1)$
with preemption-distance priorities.

b) the expected time $w_{\mathrm{I}}(p)$ necessary to serve demands of the priority classes 1 to p waiting in the system at its time of arrival,

c) its expected time in service $b(p)$,

d) the expected time $w_{\mathrm{II}}(p)$ necessary to serve demands of preemptive priority classes 1 to $(p - \xi)$ which enter the system, while the considered p-demand is still in the system,

e) the expected time $w_{\mathrm{III}}(p)$ necessary to serve demands of the non-preemptive priority classes $(p - \xi + 1)$ to $(p - 1)$ which enter the system, while the considered p-demand is still in the system, however before its last interruption.

Determining these five terms it is possible to find after some simple transformations the recursive equation for the expected waiting time of the considered p-demand; cf. Eqn. (1).

new demand of class	does not interrupt service of class	interrupts service of class	actual preemption-distance
1	1, 2	3, 4, 5, 6	2
2	1, 2, 3	4, 5, 6	2
3	1, 2, 3, 4	5, 6	2
4	1, 2, 3, 4, 5	6	2
5	1, 2, 3, 4, 5, 6	–	2
6	1, 2, 3, 4, 5, 6	–	2

Fig. 3.
Example for preemption-distance priorities (number of priority classes $P = 6$; preemption-distance $\xi = 2$).

$$t_{\mathrm{w}}(p) = \left\{\sum_{i=1}^{p-1} A(i)\cdot t_{\mathrm{w}}(i) + \sum_{i=1}^{p} \frac{b(i)}{2}\left[1 + \frac{1}{k(i)}\right]\cdot A(i) + \sum_{i=p+1}^{p+\xi-1} \frac{1}{\lambda_{\mathrm{u}}(i)\cdot b(i)}\cdot\left[b(i) - \frac{1}{\lambda_{\mathrm{u}}(i)}\left\{1 - \left(\frac{k(i)}{\lambda_{\mathrm{u}}(i)\cdot b(i) + k(i)}\right)^{k(i)}\right\}\right] + \right.$$
$$\left. + b(p)\sum_{i=1}^{p-\xi} A(i) + \left[b(p) - \frac{1}{\lambda_{\mathrm{u}}(p)}\left\{1 - \left(\frac{k(p)}{\lambda_{\mathrm{u}}(p)\cdot b(p) + k(p)}\right)^{k(p)}\right\}\right]\cdot \sum_{i=p-\xi+1}^{p-1} A(i)\right\}\left\{1 - \sum_{i=1}^{p} A(i)\right\}^{-1} \qquad (1)$$

with

$\lambda(i)$: call rate for priority class i $(i = 1, 2, \ldots, p, \ldots, P)$.

$\lambda_{\mathrm{u}}(p) = \sum_{i=1}^{p-\xi} \lambda(i)$: total call rate for priority classes, interrupting service of priority class p.

$\mu(i) = 1/b(i)$: service (termination) rate for demands of priority class i.

$b(i) = 1/\mu(i)$: mean service time for demands of priority class i.

$A(i) = \lambda(i)\cdot b(i)$: offered traffic for priority class i.

$k(i)$: order of the Erlang-distributed service times for priority class i.

Eqn. (1) allows to calculate recursively the expected waiting time for all priority classes p, starting with the most important priority class one:

$$t_{\mathrm{w}}(1) = \frac{\sum_{i=1}^{\xi} \frac{b(i)}{2}\left[1 + \frac{1}{k(i)}\right]\cdot A(i)}{1 - A(1)}. \qquad (2)$$

3. Explicite Solution for the Expected Waiting Time

It should be mentioned at this point that an explicite solution for the expected waiting times has been found, too (using the methods of Lagrange [6] for the determination of particular solutions of inhomogeneous systems of equations). However, as the above presented recursive equation is extremely practical for computer evaluations, this explicit formula shall be published — with a detailed derivation — later on.

4. Expected System-Response Time and Expected Number of Demands Waiting for Service

Obviously, the expected system-response time for demands of priority class p, i. e. the interval between the epoch at which they enter the system to the epoch at which service is completed, is given by

$$t_{\mathrm{A}}(p) = t_{\mathrm{w}}(p) + b(p). \qquad (3)$$

The expected number of demands waiting simultaneously for service is given by

$$\Omega(p) = \lambda(p)\cdot t_{\mathrm{w}}(p). \qquad (4)$$

new demand of class	does not interrupt service of class	interrupts service of class	actual preemption-distance
1	1	3, 5, 6	1
3	1, 3	5, 6	1
5	1, 3, 5, 6	–	2
6	1, 3, 5, 6	–	2

Fig. 4.
Preemption-distance priorities with "empty" priority classes (4 actual priority classes, 2 empty classes with $\lambda_2 = \lambda_4 = 0$; preemption-distance $\xi = 2$).

5. Conclusion

Introducing the preemption-distance ξ for queuing systems with priorities it is possible to describe uniformly service-strategies, most important for real-time computer systems. It is shown how to determine the expected waiting time, the expected system-response time as well as the expected number of demands waiting simultaneously for service, for each priority class p. Investigations concerning the remaining characteristic values (e. g. waiting time distribution) as well as other types of service time distributions are just under work and will be published at some future time. The well known formulas for queuing systems with pure preemptive or pure non-preemptive priorities are boundary values of the above presented formulas (1) and (2).

* * *

Acknowledgement: The author wishes to express his thanks to Professor Dr.-Ing. A. Lotze, Dipl.-Ing. W. Krämer and Dipl.-Ing. M. Langenbach-Belz for reading this paper and valuable discussions.

References

[1] *Martin, J.:* Design of real time computers. Prentice Hall, Englewood Cliffs, N. J., 1967.

[2] *Graef, M.; Greiller, R.; Hecht, G.:* Datenverarbeitung im Realzeitbetrieb. München: R. Oldenbourg Verlag 1970.

[3] *Gabler, H.:* Technik des Elektronischen Datenvermittlungs-Systems EDS. Jahrb. elektr. Fernmeldewesens 22 (1971) pp. 296 to 337.

[4] *Goslau, K.; Bacher, A. et al.:* EDS — A new electronic data switching system for data communication. Nachrichtentechn. Z. 22 (1969) pp. 444 to 463.

[5] *Chang, W.:* Queuing with nonpreemptive and preemptive-resume priorities. Operations Res. 13 (1965) pp. 1020 to 1022.

[6] *Meschkowski, H.:* Differenzengleichungen. Göttingen: Vandenhoeck u. Ruprecht 1959.

[7] *Balachandran, K. R.:* Parametric priority rules: An approach to optimization in priority queues. Operations Res. 18 (1970) pp. 526 to 540.

[8] *Cobham, A.:* Priority assignment in waiting line problems. Operations Res. 2 (1954) pp. 70 to 76 and 3 (1955) p. 547.

[9] *Gaver, D. P.:* On Priority type disciplines in queueing. Proc. Symp. on Congestion Theory (1964), The University of North Carolina Press, Chapel Hill, 1965.

[10] *Jaiswal, N. K.:* Priority queues. New York and London: Academic Press 1968.

[11] *Takács, L.:* Priority queues. Operations Res. 12 (1964) pp. 63 to 74.

(Manuscript received: November 22, 1971)