ON A MULTI-SERVER QUEUING SYSTEM WITH CONSTANT HOLDING TIME AND PRIORITIES

Sixth International Teletraffic Congress

Munich September 9-15, 1970

Manfred Langenbach-Belz

Technical University Stuttgart

Stuttgart, Federal Republic of Germany

ABSTRACT

This paper deals with a multi-server queuing system with constant holding time and an infinite number of waiting places. The offered traffic is Poissonian and is subdivided into R nonpreemptive priority classes. In a queuing system besides the probability of waiting, the mean waiting times are of main interest. In this paper now a formula is derived for the mean waiting times t_{wr} of each priority class r. With this formula the mean waiting times t_{wr} in a system with priority classes can be calculated by the aid of waiting times t_w obtained from multi-server systems without priorities.
Numerical results are compared with values of an event by event simulation, which was performed on a digital computer.
For the calculation of the mean waiting times t_{wr} in a queuing system with priorities and an arbitrary holding time distribution the same method can be applied as it is shown for constant holding time in this paper.

1. INTRODUCTION

For the following types of queuing systems with priorities there exist formulae for the mean waiting times :

- Single-server system (n=1) and arbitrary holding time distribution from COBHAM /4/.
- Multi-server system (n>1) and negative exponential holding time distribution from COBHAM et al. (see e.g. /3/, /4/).

The formula for a multi-server system and constant holding time derived in this paper completes the above mentioned list.
The manner of solution is the following:
First of all, the mean waiting time in a queuing system without priorities is regarded. Then a single-server system with priorities is considered.
With the aid of a theorem, which was proved for single-server systems in /1/, the mean waiting time of each priority class in single-server systems can be calculated. After this it is shown that the above mentioned theorem can be extended also to multi-server systems. Hence the mean waiting times of the priority classes in multi-server systems are calculated analogously to those of single server systems.

2. THE SYSTEM

2.1 Description of the system

The properties of the system and the offered traffic are as follows:

a) n fully accessible servers (n arbitrarily)
b) infinite number of waiting places
c) the offered traffic is Poissonian with infinite number of sources and has the calling rate λ
d) the offered traffic is subdivided into R nonpreemptive priority classes, where r=1 is the highest priority class and r=R is the lowest priority class (R arbitrarily). The calls of the highest priority class are named 1-calls, the calls of the second priority class are named 2-calls and so on.
e) the waiting calls of each priority class are served with discipline first-come, first-served and they do not leave the system without first being served
f) the holding time T is constant and the same for all priority classes
g) the system is in the statistical equilibrium

2.2 Abbreviations

n	number of servers
T	constant holding time
λ	calling rate
A	$=\lambda\cdot T$ offered traffic
α	$=A/n$
R	number of priority classes
r	priority class r (r=1,2,...,R)
A_r	offered traffic of priority class r

$P(>t)$ waiting time distribution of all calls (probability that the waiting time of a call is greater t)
$P(>0)$ probability of waiting
$W(>t) = P(>t)/P(>0)$ waiting time distribution of the waiting calls
t_W^* mean waiting time of all calls
t_W $= t_W^*/P(>0)$ mean waiting time of the waiting calls
t_{Wr}^* mean waiting time of all r-calls
t_{Wr} $= t_{Wr}^*/P(>0)$ mean waiting time of the waiting r-calls

3. THE QUEUING SYSTEM WITHOUT PRIORITIES

First of all it is useful to consider a queuing system with the properties described in chapter 2.1, but without priority classes. For such a system CROMMELIN has derived formulae for the probability of waiting $P(>0)$, the mean waiting time t_W, and the waiting time distribution $W(>t)$ /2/.
His manner of solution is briefly described in the following:
First, he considered time intervals of the length T and set up the equilibrium equations of the state probabilities $P(j)$ $(j=0,1,\ldots\infty)$. To solve this system of equations, Crommelin made use of the generating function which is defined by

$$\psi(z) = \sum_{j=0}^{\infty} z^j \cdot P(j)$$

Some transformations led to an expression for $\psi(z)$, in which the denominator is of the form

$$1 - z^n \cdot e^{A(1-z)} \qquad (1)$$

For further calculations it was necessary to find the zeros of (1). So the roots β_ν of the following equation must be determined:

$$\beta^n \cdot e^{-A(\beta-1)} = 1$$

$$\text{or} \qquad \beta \cdot e^{-\alpha\beta} = e^{-\alpha} \cdot \sqrt[n]{1} \quad (\text{with } \alpha = A/n) \qquad (2)$$

Crommelin showed that this equation for β has exactly n roots $\beta_0, \beta_1, \ldots, \beta_{n-1}$ such that $|\beta_\nu| \leq 1$. It can be shown easily that $\beta_0=1$ is always a root of equation (2). The other roots $\beta_1, \ldots, \beta_{n-1}$ must be determined by an iterative method.

With these roots of equation (2) and some further calculations Crommelin derived the following formulae for the probability of waiting $P(>0)$ and the mean waiting time:

$$P(>0) = 1 - \frac{n - A}{\prod_{\nu=1}^{n-1} (1-\beta_\nu)} \qquad (3)$$

$$t_W^* = \frac{T}{A}\left[\sum_{\nu=1}^{n-1} \frac{1}{1-\beta_\nu} + \frac{A^2-n^2+n}{2(n-A)}\right] \qquad (4)$$

$$t_W = \frac{t_W^*}{P(>0)} \qquad (5)$$

The most simple multi-server system is a system with n=2 servers. The probability of waiting $P(>0)$ and the mean waiting time t_W^* for this particular system are given by the following formulae:

$$P(>0) = 1 - \frac{2-A}{1-\beta_1}$$

$$t_W^* = \frac{T}{A}\left[\frac{1}{1-\beta_1} + \frac{A^2-2}{2(2-A)}\right]$$

β_1 is a negative real value and the relation between β_1 and $\alpha=A/n$ is shown in the diagram 1.

4. THE QUEUING SYSTEM WITH PRIORITIES

In this system the offered traffic A is subdivided into R priority classes. Each priority class has the offered traffic A_r so that

$$\sum_{r=1}^{R} A_r = A$$

The service discipline is the following:
First, the waiting calls of the priority class r=1 (1-calls) are served. If there are no more waiting 1-calls, then the waiting 2-calls are served and so on. Within each priority class the calls are served with discipline first-come, first-served. The service of any call can not be interrupted by arriving calls of a higher priority class.

4.1 The single-server system (n=1)

For the special case of n=1 (single-server system) the following theorem was proved /1/ :

Theorem: The waiting time distribution function $W_1(>t)$ of the waiting 1-calls (highest priority) is identical with that distribution function, which is obtained, if only the traffic A_1 of the highest priority class is offered to the system.

For the mean waiting time t_W of the waiting calls the following formula holds generally:

$$t_W = \int_0^\infty t \cdot w(t)dt = \int_0^\infty W(>t)dt \quad , (6)$$

$$\text{where} \quad w(t) = -\frac{dW(>t)}{dt}$$

From (6) and the above theorem follows that the mean waiting time t_{W1} of the waiting 1-calls is independent of the calls of the other priority classes. Thus, the mean waiting time t_{W1} can be calculated as in a single-server system without priorities.

To calculate the mean waiting times of the other priority classes, the calls of the priority classes $\leq r$, or $<r$ respectively, are regarded as one group. The mean waiting time of the calls of such a group is independent of their service discipline. Furthermore, with (6) and the theorem follows that the mean waiting time $t_W(\leq r)$, or $t_W(<r)$ respectively, of this group is independent of the calls with lower priorities. So the mean waiting time $t_W(\leq r)$, or $t_W(<r)$ respectively, can be calculated each in the same manner as t_{W1}.

Therefore, the mean waiting time t_{Wr} of the waiting r-calls can be derived from the following equation:

$$\frac{A(<r)}{A(\leq r)} \cdot t_W(<r) + \frac{A_r}{A(\leq r)} \cdot t_{Wr} = t_W(\leq r) \qquad (7)$$

$$\text{with} \quad A(<r) = \sum_{i=1}^{r-1} A_i \quad , \quad A(\leq r) = \sum_{i=1}^{r} A_i$$

Equation (7) holds true because the ratio $A(<r)/A(\leq r)$ of the waiting calls with priority classes $\leq r$ has the mean waiting time $t_W(<r)$ and the ratio $A_r/A(\leq r)$ has the mean waiting time t_{Wr}. From (7) t_{Wr} can be obtained easily as

$$t_{Wr} = \frac{A(\leq r) \cdot t_W(\leq r) - A(<r) \cdot t_W(<r)}{A_r} \qquad (8)$$

$t_W(\leq r)$ and $t_W(<r)$ can be determined as in a single-server system without priorities.

4.2 The multi-server system (n>1)

Section 4.1 dealt with a single-server system. To calculate now the mean waiting times t_{wr} in a system with n>1 it is shown in the following that the theorem of chapter 4.1 is also valid for a multi-server system:

Two separate systems (I and II) are regarded, each with n servers. To each system a traffic A_1 be offered. But to one system (II) an additional traffic A_2 be offered.

System I : Only the traffic A_1 is offered. There are no priority classes. The calls arrive with the time independent calling rate λ_1. If all servers are busy, there exists a blocking interval. During such a blocking interval an arriving call is set in a waiting place. The probability that the new call is set in the waiting place j shall be $W_I(j)$. Then it is

$$P(>0)|_{A_1} = \sum_{j=1}^{\infty} W_I(j)$$

where $P(>0)|_{A_1}$ is the probability of waiting if only the traffic A_1 is offered to the system. The conditional waiting time distribution of that call, which was set in the waiting place j, shall be $W(>t|j)$. Because the waiting calls are always served with discipline first-come, first-served, $W(>t|j)$ is independent of A_1 and $W_I(j)$. Then, the waiting time distribution function $P(>t)$ of all calls is

$$P(>t) = \sum_{j=1}^{\infty} W(>t|j)\cdot W_I(j)$$

The waiting time distribution function $W(>t)$ of the waiting calls is then obtained by division with $P(>0)|_{A_1}$:

$$\boxed{W(>t) = \frac{1}{P(>0)|_{A_1}}\cdot\sum_{j=1}^{\infty} W(>t|j)\cdot W_I(j)} \qquad (9)$$

System II : There are two priority classes. The offered traffic A_1 has the priority r=1 and the additional offered traffic A_2 has the priority r=2. In this system the calls are arriving with the time independent calling rate λ_1 or λ_2, respectively. A blocking interval arises if all servers are occupied by 1-calls or 2-calls in an arbitrary mixture. During this blocking interval a 1-queue, consisting of one or more 1-calls, may be built up and may be served step by step. This "birth" and "death" of the considered 1-queue is independent of already waiting or later arriving 2-calls, which are served always after the waiting 1-calls. Hence, the time dependent behaviour of the considered 1-queue is only a function of the Poisson input process with the calling rate λ_1 and of the termination process of n occupied servers. Because the termination process of the 2-calls and the 1-calls is exactly the same, the existing 1-queue is in no respect influenced by the 2-calls. So the time dependent behaviour of a 1-queue in system II is exactly the same one as in system I. Only the average number of 1-queues, built up per unit of time, depends linearly on the global probability of blocking $P(>0)|_A = f(A_1+A_2,n)$. Thus, the number of 1-queues in system II is increased as against the corresponding number of 1-queues in system I by a factor

$$F = \frac{P(>0)|_A}{P(>0)|_{A_1}}$$

with $\qquad A = A_1 + A_2$

In system II therefore, the probability $W_{II}(j)$ that an arriving 1-call is set in the waiting place j gets

$$W_{II}(j) = F\cdot W_I(j)$$

Owing to the service discipline of the waiting calls the conditional waiting time distribution function $W(>t|j)$ of a 1-call in the waiting place j is independent of A and $W_{II}(j)$. Then, the waiting time distribution function $P_1(>t)$ of all 1-calls is :

$$P_1(>t) = \sum_{j=1}^{\infty} W(>t|j)\cdot W_{II}(j) = F\cdot\sum_{j=1}^{\infty} W(>t|j)\cdot W_I(j)$$

$$P_1(>t) = \frac{P(>0)|_A}{P(>0)|_{A_1}}\cdot\sum_{j=1}^{\infty} W(>t|j)\cdot W_I(j)$$

The waiting time distribution function $W_1(>t)$ of the waiting 1-calls is then obtained by division with the probability of waiting $P(>0)|_A$

$$\boxed{W_1(>t) = \frac{1}{P(>0)|_{A_1}}\cdot\sum_{j=1}^{\infty} W(>t|j)\cdot W_I(j)} \qquad (10)$$

Equations (9) and (10) are identical and so the theorem of chapter 4.1 can be also applied on a multi-server system.

It should be noted that this theorem is valid for an arbitrary holding time distribution, provided that the holding time distribution of each priority class is the same!

By applying the theorem of chapter 4.1 in connection with equation (6) on a multi-server system, the mean waiting time t_{w1} of the waiting 1-calls can be calculated with equations (4) and (5) :

$$t_{w1} = \frac{T}{A_1\cdot P(>0)|_{A_1}}\left(\sum_{\nu=1}^{n-1}\frac{1}{1-\beta_\nu} + \frac{A_1^2-n^2+n}{2(n-A_1)}\right) \qquad (11)$$

with $\quad P(>0)|_{A_1} = 1 - \dfrac{n-A_1}{\prod_{\nu=1}^{n-1}(1-\beta_\nu)}$

For the mean waiting times t_{wr} of the lower priority classes then equation (8) is also available for a multi-server system. With equations (4), (5) and (8) the following formula for t_{wr} is obtained :

$$\boxed{t_{wr} = \frac{T}{A_r}\left[\frac{1}{P(>0)|_{A(\le r)}}\left(\sum_{\nu=1}^{n-1}\frac{1}{1-\beta_{\nu 1}} + \frac{\left(\sum_{\nu=1}^{r}A_\nu\right)^2-n^2+n}{2\left(n-\sum_{\nu=1}^{r}A_\nu\right)}\right) - \frac{1}{P(>0)|_{A(<r)}}\left(\sum_{\nu=1}^{n-1}\frac{1}{1-\beta_{\nu 2}} + \frac{\left(\sum_{\nu=1}^{r-1}A_\nu\right)^2-n^2+n}{2\left(n-\sum_{\nu=1}^{r-1}A_\nu\right)}\right)\right]} \qquad (12$$

In this expression $\beta_{\nu 1}$ are the roots of equation (2) if A is replaced there by $A(\le r)$ and $\beta_{\nu 2}$ are the roots of the same equation if A is replaced there by $A(<r)$.

The probabilities of waiting $P(>0)|_{A(\le r)}$ and $P(<0)|_{A(<r)}$ are determined with equation (3) as

$$P(>0)|_{A(\le r)} = 1 - \frac{n - \sum_{\nu=1}^{r} A_\nu}{\prod_{\nu=1}^{n-1} (1-\beta_{\nu 1})}$$

and

$$P(>0)|_{A(<r)} = 1 - \frac{n - \sum_{\nu=1}^{r-1} A_\nu}{\prod_{\nu=1}^{n-1} (1-\beta_{\nu 2})}$$

The mean waiting time of all r-calls is then

$$t^*_{wr} = P(>0)|_A \cdot t_{wr} \quad .$$

Some results obtained from equation (12) are shown in the diagrams 2-9 and are compared with the corresponding results of an event by event simulation. The values of the simulation are determined with a 95%-confidence interval. As it can be seen from the diagrams, the calculated values for the mean waiting times t_{wr} are situated very well within the 95%-confidence intervals of the simulation.

Attention should be paid to the following problem: In the case of many servers (e.g. n>5) and a small offered traffic A_1 (e.g. $A_1<0.5$) the determination of the roots β_ν (contained in (11)) must be done with high accuracy because in this case a small inaccuracy of β_ν causes a remarkable inaccuracy of t_{w1} (see the small values of A_1 in the diagrams 7-9). From equation (8) it can be seen easily that the mean waiting times of all other priority classes are influenced by this inaccuracy of t_{w1}.

5. CONCLUSION

In this paper the mean waiting times of a queuing system with constant holding time and priorities are studied. A formula for the mean waiting time t_{wr} of each priority class is derived.
With this formula the mean waiting times t_{wr} can be determined with waiting times, which are known from queuing systems without priorities.
The calculated values are situated very well within the 95%-confidence intervals of values, which are found by simulation.

In the case of many servers (e.g. n>5) and a small offered traffic A_1 (e.g. $A_1<0.5$) of the highest priority class the accuracy of the calculated values t_{wr} depends strongly on the accuracy of the iterative determination of the roots β_ν.

Finally it should be noted that the theorem of chapter 4.1 is valid in systems with arbitrary holding time distribution, provided that the holding time distribution of each priority class is the same. If this condition is satisfied, then equation (8) holds true also for the mean waiting times t_{wr} in a system with arbitrary holding time. Thus, the values of t_{wr} in a system with arbitrary holding time can be obtained by the same method as it is shown for constant holding time in this paper.

ACKNOWLEDGEMENT

The author wishes to express his thanks to Prof. Dr.-Ing. A. Lotze and to Dr.-Ing. W. Wagner for many valuable discussions.

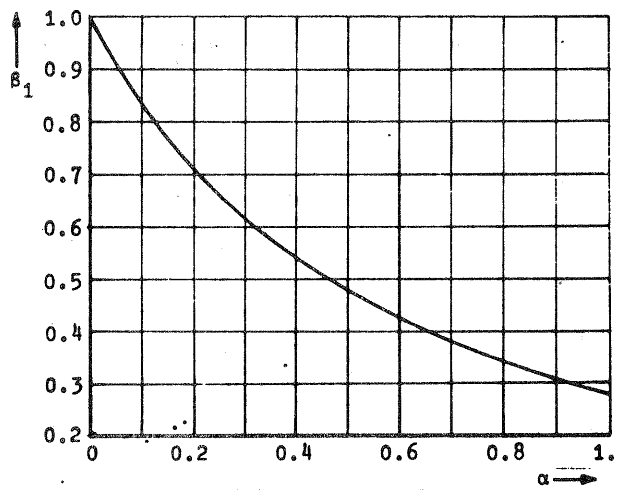

Diagram 1 $\beta_1 = f(\alpha)$ with $\alpha = A/n$ for a queuing system with n = 2

Some examples for the mean waiting times t_{wr}:

In each of the following diagrams it is

simulation value (95%-confidence interval)

calculated value (from formula (12))

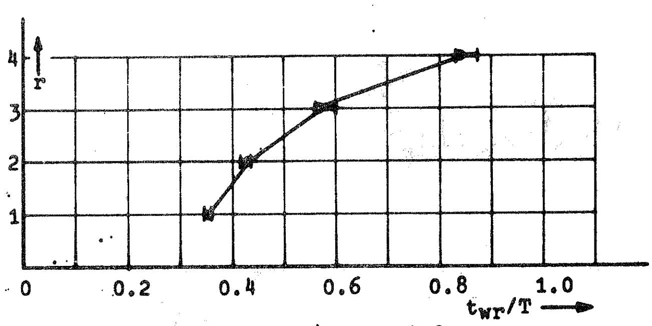

Diagram 2 n = 2, r = 4, A = 1.0
$A_1 = A_2 = A_3 = A_4 = 0.25A$

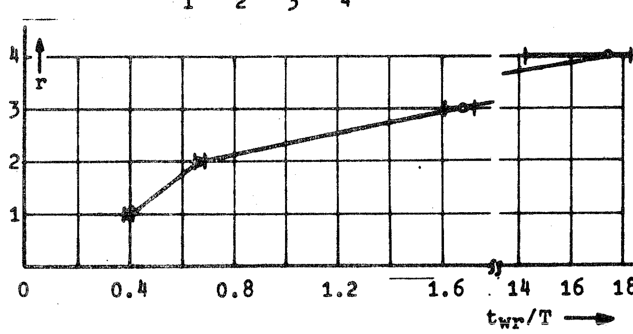

Diagram 3 n = 2, r = 4, A = 1.9
$A_1 = A_2 = A_3 = A_4 = 0.25A$

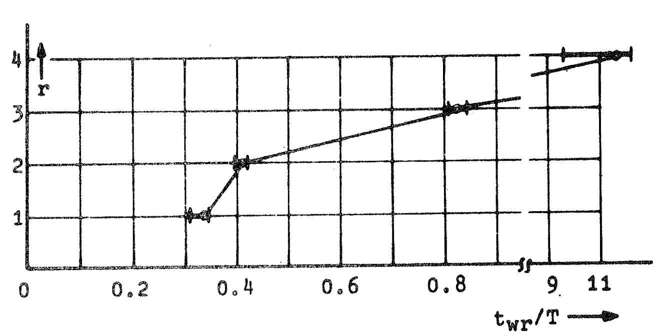

Diagram 4 n = 2, r = 4, A = 1.9
$A_1 = 0.03A$, $A_2 = 0.25A$
$A_3 = 0.32A$, $A_4 = 0.4A$

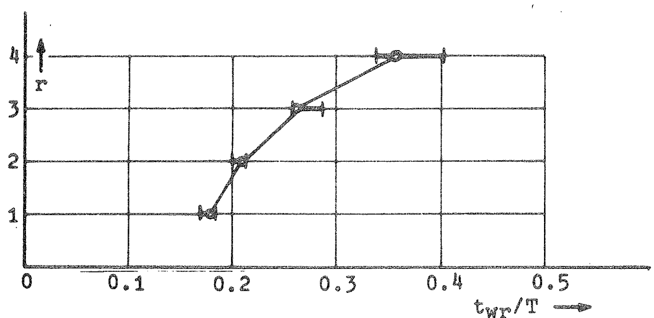

Diagram 5 $n = 5,\ r = 4,\ A = 2.5$
$A_1 = A_2 = A_3 = A_4 = 0.25A$

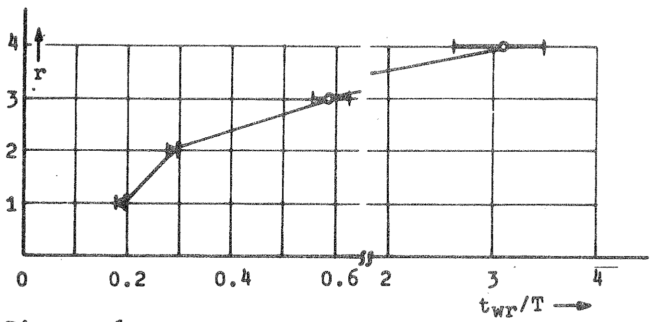

Diagram 6 $n = 5,\ r = 4,\ A = 4.5$
$A_1 = A_2 = A_3 = A_4 = 0.25A$

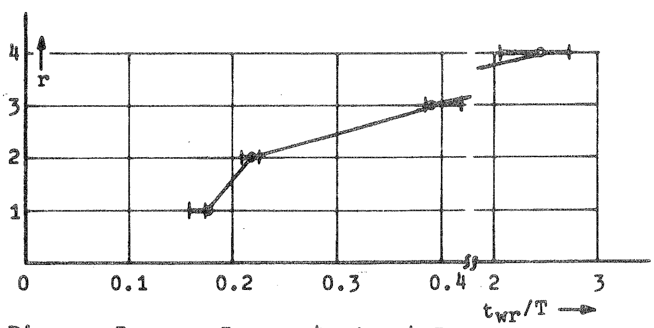

Diagram 7 $n = 5,\ r = 4,\ A = 4.5$
$A_1 = 0.1A,\ A_2 = 0.25A$
$A_3 = 0.3A,\ A_4 = 0.35A$

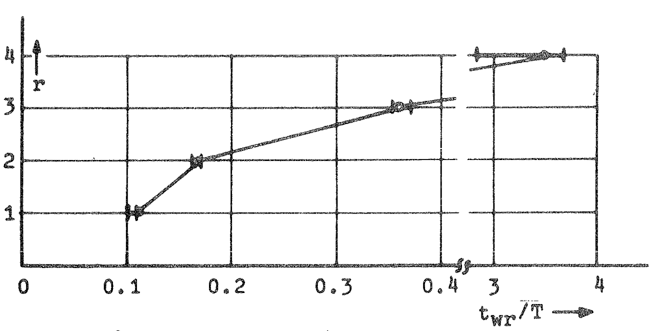

Diagram 8 $n = 10,\ r = 4,\ A = 9.5$
$A_1 = A_2 = A_3 = A_4 = 0.25A$

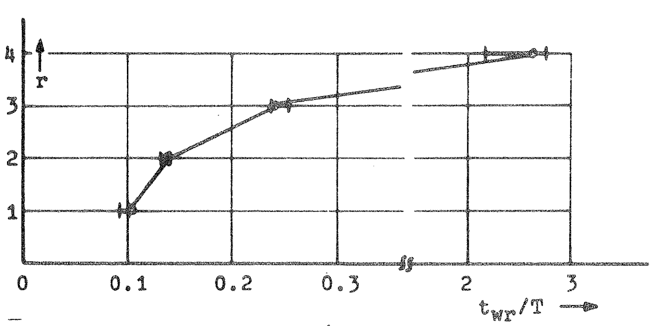

Diagram 9 $n = 10,\ r = 4,\ A = 9.5$
$A_1 = 0.17A,\ A_2 = 0.21A$
$A_3 = 0.27A,\ A_4 = 0.35A$

REFERENCES

/1/ Zimmermann/Störmer — Wartezeiten in Nachrichtenvermittlungen mit Speichern. Oldenbourg, München 1961

/2/ Crommelin — Delay probability formulae when the holding times are constant. P.O. Elect. Engrs. J. 25, 41-50, 1932; 26, 266-274, 1933-34

/3/ Syski — Introduction to congestion theory in telephone systems. Oliver and Boyd, 1960

/4/ Cobham — Priority assignment in waiting line problems. J.Op.Res.Soc.Am. 2,70-76,1954; A correction: 3,547, 1955

/5/ Langenbach-Belz — Wartesystem mit Prioritäten und konstanter Belegungsdauer. Monograph of the Institute for Switching and Data Technics, University Stuttgart, 1968