Total waiting time distribution function and the fate of a customer in a system with two queues in series*

Die Verteilungsfunktion der Gesamtwartezeit und das Schicksal einer Anforderung in einem System mit zwei seriell angeordneten Warteschlangen

Elektron. Rechenanl. 17 (1975), H. 1, S. 18–27
Manuskripteingang: 26. 3. 1974

von W. KRÄMER
Lehrstuhl und Institut für Nachrichtenvermittlung und Datenverarbeitung der Universität Stuttgart

The study of traffic flow in modern computers and communication networks and various other technical systems leads in many cases to systems or subsystems with queues arranged in series or tandem.

This paper is concerned especially with a system of two unlimited queues in series, when Poisson traffic is offered to the first stage and the service times in the two stages are independent of each other and negative exponentially distributed. The first stage includes one service-unit, while the second stage is allowed to be a multiserver queueing system.

In this paper, by pursuing a particular customer (call or request) at his walk through the whole system, the fate of a customer in the second stage is determined as a function of all possible components of fate in the first stage. Though the flow times (waiting plus service times) of the same customer in the successive stages are independent, other values of fate (e.g. the waiting times) are not independent.

Die Untersuchung des Verkehrsflusses in Rechnern und Übertragungsnetzen und zahlreichen weiteren Systemen führt in vielen Fällen auf Systeme oder Subsysteme mit seriell angeordneten Warteschlangen.

Dieser Beitrag behandelt ein spezielles 2-stufiges System mit unbegrenzt großen Wartespeichern, einem Poisson-Ankunftsprozeß in Stufe 1 und unabhängigen, negativ-exponentiell verteilten Bedienungszeiten. Die 1. Stufe enthält 1 Bedienungseinheit, in Stufe 2 können mehrere Bedienungseinheiten vorhanden sein.

Durch das Verfolgen von speziellen Test-Anforderungen beim Durchlauf durch das ganze System wird das Schicksal einer Anforderung in Stufe 2 bestimmt als Funktion aller möglichen Komponenten des Schicksals in Stufe 1. Obwohl die Durchlaufzeiten (Warte- plus Bedienungszeiten) derselben Anforderung in beiden Stufen unabhängig sind, trifft dies nicht für andere Schicksalsgrößen (z. B. Wartezeiten) zu.

* Revised version of a paper presented at the 7th Internationl Teletraffic Congress, Stockholm, June 13-20, 1973.

1. Introduction

1.1 General remarks

Modern electronic computers are very complex in structure and in operating strategies. To judge their effectiveness, many aspects have to be considered, including throughput or response time questions, which are tried to be answered by means of queueing theory. Nowadays only such models can be treated with, which either describe the traffic behaviour and/or structures relatively globally or which are models for some subconfigurations. To determine the traffic characteristics within such computer systems models can be made at various levels

- the job or program level
- the instruction level

or in an intermediate level which can be represented in computers with paging by pages or generally by tasks, which could be handled without interrupt of the central processing unit.

Besides arrangements of one or more parallel servers (single stages) there are configurations which can be described by a serial arrangement of queues. The system considered here could be interpreted as a simple computer model for the service of jobs (fig. 1).

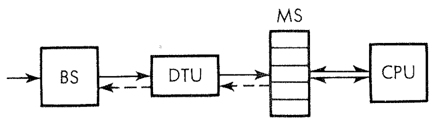


Fig. 1. Computer configuration.
DTU: data transfer unit
CPU: central processing unit (≥ 1 processors)
MS: main storage
BS: backing storage

Related structures can be obtained by analyzing a serial arrangement in the instruction level as there are instruction pipelines etc.

1.2 Description of the system

The system dealt with consists of two unlimited queues arranged in series, where the input process to the first stage is a Poisson process with mean arrival rate λ. The arriving customers, calls or requests, shortly referred to as calls, first are served by a single server and then by one server of the second stage, which is allowed to be a multiserver system (fig. 2).

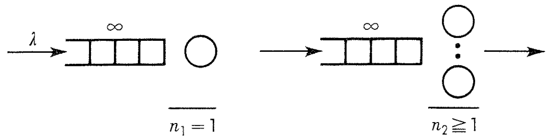


Fig. 2. The system.

The service or holding times T_{Hi} $(i=1, 2)$ of a call in both stages are independent of each other and negative exponentially distributed with distribution functions (d.f.)

$$P(T_{Hi} \leq t) \overset{\text{def}}{=} H_i(t) = 1 - e^{-\varepsilon_i t} \tag{1}$$

and means

$$E(T_{Hi}) = \frac{1}{\varepsilon_i} = h_i \tag{2}$$

The traffics offered A_i are defined by

$$A_i = \frac{\lambda}{\varepsilon_i} = \lambda \cdot h_i \tag{3}$$

and the utilizations

$$\varrho_i = \frac{A_i}{n_i} = \frac{\lambda}{\mu_i} \quad \text{with} \quad \mu_i = n_i \cdot \varepsilon_i \tag{4}$$

Both stages are assumed to be in statistical equilibrium, so that

$$\lambda < \min(\mu_1, \mu_2) \tag{5}$$

Considering a general call at its walk through this system, the following time diagram is obtained:

T_{Wi} is the random waiting time not including service T_{Hi}, and T_{Fi} the random flow time in stage i. The queue disciplines in both stages are, as long as no d.f.'s of the waiting times are concerned, arbitrary, otherwise in order of arrival (FIFO).

1.3 Known results

It is well known, that the output of the first stage is a Poisson process with mean output rate λ (*Burke* [1; 3], and others), so that *each single stage* may be computed completely according to the formulae for the $M/M/n$ queueing system, cf. e.g. *Syski* [11].

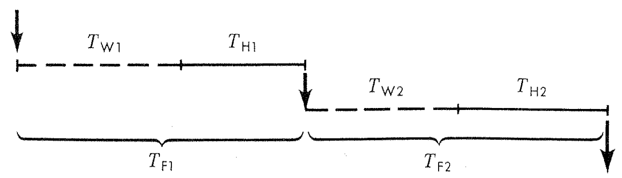


Fig. 3. General time diagram.

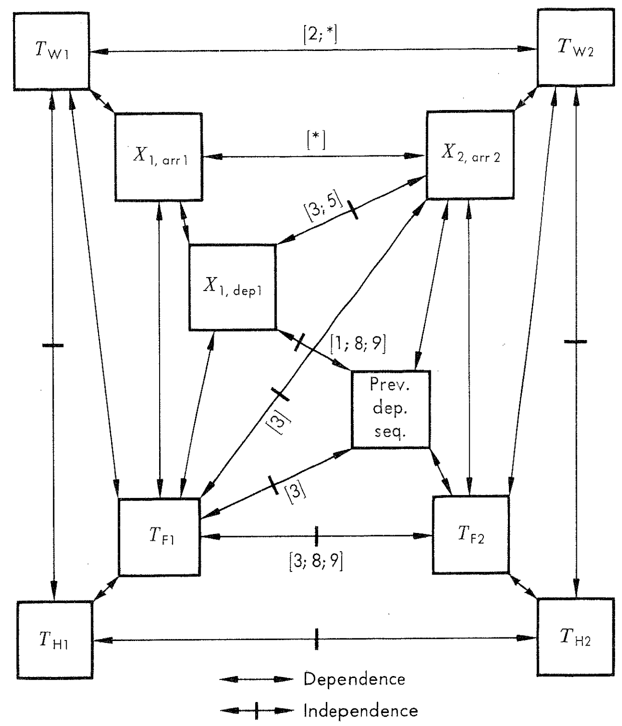


Fig. 4. Relations between fate values.

Jackson [5] proved the state probabilities of the single stages of such a system at the same time to be independent of each other. So it is possible to quote directly the state probabilities of *the whole system.*

Considering *a particular call* at its walk through the whole system, a further group of traffic characteristics may be obtained which will be regarded in this paper in more detail.

Fig. 4 represents a graphical survey of some relations between various fate values (random variables) of a certain arbitrary call in such a two stage system ($n_1 > 1$ included).

Besides T_{Wi}, T_{Hi}, T_{Fi} and the number of calls met upon **arr**ival in stage i ($X_{i,\text{arr}\,i}$), also the number of calls left behind in the first stage ($X_{1,\text{dep}\,1}$) and the **dep**arture sequence of the first stage previous to the departure of the considered call (Prev. dep. seq.) are involved.

The arrows connecting two values showing whether (separately considered) they are dependent or not, are labeled with references to the literature ([*] means basic parts of *this* paper).

Using the concept of reversibility of a Markov chain, the independence of the flow times was shown by *Reich* [9; 10] for single server and by *Burke* [3] for multiserver systems.

Nelson [7] derived by convolution an expression for the distribution function of the total waiting time in a (more generally structured) system, assuming independence of the waiting times in the single stages.

Burke [2] proved that the waiting times of a call in the investigated system with $n_2 = 1$ are dependent. Using the theorem of *Jackson* and the virtual delay in stage 2, he showed by explicit calculation that the probabilities of waiting

$W_i = P(T_{Wi} > 0)$ in the first and second stage are *not* independent. Though the traffic offered to stage 2 is pure chance traffic, the future fate of a call (number $X_{2,\mathrm{arr}\,2}$ of calls met in stage 2 upon arrival there and/or waiting time T_{W2}) is not independent of its previous fate in stage 1.

1.4 Treated problems

The aim of this paper is to investigate and to throw more light upon the dependencies between these various fate values of a call in the two successive stages and to extend known results. Therefore, several test-calls with different and more or less known fate in the first stage are considered.

First, there are calls with special assumptions about the waiting time in the first stage (no waiting, waiting in the first stage of unknown duration, known waiting time (> 0)).

Determining the number of calls met by these test-calls upon arrival in the second stage, the influences of these values to the further fate (waiting or flow time) are obtained. Admitting both calls with known and unknown service time in the first stage, also the influence of the service time in the first stage is shown (chapter 2).

The knowledge of all these dependencies is the precondition for the determination of total fates (total waiting time d.f. with total probability of waiting and mean total waiting time of the waiting calls). Since in a preliminary chapter it is shown, that the additional assumption of a concrete waiting time (> 0) in the first stage has no further effect (called 'limited dependency'), it is possible to determine the d.f. of the total waiting time by convolution of special terms in chapter 3.

In the last chapter, further test-calls are considered, which have found a certain number of predecessors in the first stage. Considering all possible paths in a *random walk* diagram, the number of calls met in a second (single server) stage is determined. It is shown that for the queue lengths no such 'limited dependency' is valid as for the waiting times.

2. Test calls with given waiting time in stage 1

2.1 Output process of M/M/n during concrete waiting time

The aim of this preliminary investigation is to obtain statements about the behaviour of a stationary single stage *M/M/n* system with FIFO during a certain waiting time $T_W = t_0\ (> 0)$ of a test-call. Let $p_w(j, t_0) = P(X_{\mathrm{arr}} = j \mid T_W = t_0)$ be the probability that this test-call has met $j\ (\geq n)$ calls upon its arrival in the whole stage. Applying the theorem of *Bayes* it holds:

$$P(X_{\mathrm{arr}}=j \mid T_W \in [t_0, t_0+\mathrm{d}t]) = \frac{P(X_{\mathrm{arr}}=j)}{P(T_W \in [t_0, t_0+\mathrm{d}t])} \cdot P(T_W \in [t_0, t_0+\mathrm{d}t] \mid X_{\mathrm{arr}}=j) \qquad (6)$$

Inserting known expressions into the right side and making the limit transition $\mathrm{d}t \to 0$, it is obtained

$$p_W(j, t_0) = e^{-\lambda t_0} \cdot \frac{(\lambda t_0)^{j-n}}{(j-n)!} \qquad t_0 > 0,\ j \geq n \qquad (6\mathrm{a})$$

It is obvious that this expression (which is independent of ε) is identical with the probability that $j-n=z$ calls arrive during t_0:

The probability that a call with concrete waiting time $T_W = t_0\ (> 0)$ has met $z\ (\geq 0)$ calls in the waiting storage, is the same as the probability that the same call upon *leaving the waiting storage* (to begin service) leaves $z\ (\geq 0)$ calls behind.

If a call with concrete waiting time $t_0\ (> 0)$ has met $j\ (\geq n)$ calls upon arrival in the whole stage (this occurs with probability $p_w(j, t_0)$), exactly $j-n+1$ calls must be served until its service begins. Since the departure of the last one of them coincides with the end of T_W, exactly $j-n$ calls leave the system during t_0. If we forget the value of j, nevertheless the Poisson distribution (6a) must be fulfilled for reasons of stationarity. This implies that the output intervals during t_0 are negative exponentially distributed with mean $1/\lambda$ (Poisson).

2.2 General way of calculation

The observation of the system starts at time T, when the fixed service time $T_{H1} = t_1$ of a test-call begins.

From the state probabilities $_cp_2'(x)$ at time T^+ (called starting probabilities), which are independent of the required service time t_1, the state probabilities $_cp_2''(x \mid t_1)$ at time $(T+t_1)^-$ (called meeting probabilities) are determined. (Prefix c means that the probability is a conditional probability related to a special call, where $c=0$ refers to a call with $T_{W1}=0$ and $c=1$ to a call with $T_{W1} > 0$ of known *or* unknown duration.)

2.3 Starting probabilities in stage 2

Using the theorem of *Jackson* [5], it is obvious that the state probabilities $_0p_2'(x)$ at the arrival of a nonwaiting call in stage 1 are independent and according to the absolute state probabilities:

$$_0p_2'(x) = p_2(x) \qquad (7)$$

where

$$p_2(x) = \begin{cases} p_2(0) \cdot \dfrac{A_2^x}{x!} & 0 \leq x \leq n_2 \\[2ex] p_2(0) \cdot \dfrac{A_2^x}{n_2!\, n_2^{x-n_2}} & x \geq n_2 - 1 \end{cases} \qquad (8)$$

with

$$p_2(0) \cdot \frac{A_2^{n_2}}{n_2!} \cdot \frac{1}{1-\dfrac{A_2}{n_2}} = E_{2,n_2}(A_2) = W_2 \qquad (9)$$

which is the absolute probability of waiting in the second stage according to the second formula of *Erlang*, cf. e.g. *Syski* [11].

For calls which have to wait in the first stage the following situation prevails:

at time $T - t_0$ the considered test-call with waiting time $t_0\ (> 0)$ arrives in stage 1. Due to *Jackson*'s theorem the state probabilities of the second stage at time $T - t_0$ are identical with the absolute values according (8). During the subsequent time t_0 the input process of the second stage is Poisson (shown in 2.1), so the state probabilities of the second stage at time T^- will also be distributed according to the absolute values. So

$$_1p_2'(x) = \begin{cases} 0 & \text{for } x = 0 \\ p_2(x-1) & \text{for } x > 0 \end{cases} \qquad (10)$$

It is obvious that (10) is also valid when the concrete value of T_{W1} is unknown, or when a concrete service time $T_{H1} = t_1$ is preassigned to a test-call.

2.4 Meeting probabilities in stage 2

Let $p(i, t_1)$ be the probability that i calls leave the second stage during the service time $T_{H1} = t_1$ of a test-call. As long as stage 2 is fully occupied ($X_2 \geq n_2$), the rate of the whole stage to serve one call will be $\mu_2 = n_2 \cdot \varepsilon_2$ (cf. fig. 5). So the meeting probabilities are

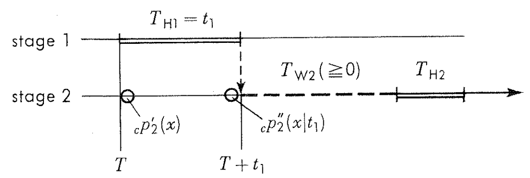


Fig. 5. Time diagram.

$$_cp_2''(x \mid t_1) = \sum_{i=0}^{\infty} p(i, t_1) \cdot {_cp_2'}(x+i) \qquad \begin{matrix} x \geq n_2 \\ c = 0, 1; \end{matrix} \qquad (11)$$

where

$$p(i, t_1) = e^{-\mu_2 t_1} \cdot \frac{(\mu_2 t_1)^i}{i!} \qquad i \geq 0 \qquad (12)$$

is the probability of i Poisson events during t_1. By integration we will get the meeting probabilities for corresponding test-calls with unknown service time T_{H1}:

$$_cp_2''(x) = \int_0^{\infty} {_cp_2''}(x \mid t_1)\, \mathrm{d}H_1(t_1) \qquad c = 0, 1; \qquad (13)$$

For test-calls which have not waited in stage 1 it is obtained from (11)

$$P(X_{2,\mathrm{arr}\,2} = x \mid T_{W1} = 0, T_{H1} = t_1) \overset{\mathrm{def}}{=} {_0p_2''}(x \mid t_1) = p_2(x) \cdot e^{-(\mu_2 - \lambda) t_1} \qquad x \geq n_2 \qquad (14)$$

If the assumption of a certain service time is dropped, we receive by (13) with (1)

$$P(X_{2,\mathrm{arr}\,2} = x \mid T_{W1} = 0) \overset{\mathrm{def}}{=} {_0p_2''}(x) = \frac{\varepsilon_1}{\varepsilon_1 + \mu_2 - \lambda} \cdot p_2(x) \qquad x \geq n_2 \qquad (15)$$

Similarly, we obtain the meeting probabilities for calls with $T_{W1} > 0$ of known or unknown duration ($c=1$):

$$_1p_2''(x \mid t_1) = p_2(x-1) \cdot e^{-(\mu_2 - \lambda) t_1} \qquad x \geq n_2 \qquad (16)$$

$$_1p_2''(x) = \frac{\varepsilon_1}{\varepsilon_1 + \mu_2 - \lambda} \cdot p_2(x-1) \qquad (17)$$

Considering a call of which only the service time in the first stage is known, we get from (14) and (16) by weighting summation

$$P(X_{2,\mathrm{arr}\,2} = x \mid T_{H1} = t_1) \overset{\mathrm{def}}{=} p_2''(x \mid t_1) = \left(1 - \varrho_1 + \frac{\varrho_1}{\varrho_2}\right) \cdot p_2(x) \cdot e^{-(\mu_2 - \lambda) t_1} \qquad x \geq n_2 \qquad (18)$$

2.5 Conditional probabilities of waiting

Summing up the meeting probabilities, the probability of waiting in stage 2 is obtained for the various test-calls:

$$_cP(T_{W2} > 0 \mid T_{H1} = t_1) = \sum_{x=n_2}^{\infty} {_cp_2''}(x \mid t_1) \qquad (19)$$

If service time is unknown:

$$_cW_2 = \sum_{x=n_2}^{\infty} {_cp_2''}(x) \qquad c = 0, 1; \qquad (20)$$

(14) and (16) yield with (19) after some intermediate calculations

$$P(T_{W2} > 0 \mid T_{W1} = 0, T_{H1} = t_1) \overset{\mathrm{def}}{=} {_0P}(T_{W2} > 0 \mid T_{H1} = t_1) = E_{2,n_2}(A_2) \cdot e^{-(\mu_2 - \lambda) t_1} \qquad (21)$$

and

$$_1P(T_{W2} > 0 \mid T_{H1} = t_1) = \frac{n_2}{A_2} \cdot E_{2,n_2}(A_2) \cdot e^{-(\mu_2 - \lambda) t_1} \qquad (22)$$

Considering test-calls with unspecified service times, the following formulae are obtained:

$$P(T_{W2} > 0 \mid T_{W1} = 0) \overset{\mathrm{def}}{=} {_0W_2} = {_0F} \cdot E_{2,n_2}(A_2)$$

with

$$_0F = \frac{\varepsilon_1}{\varepsilon_1 + \mu_2 - \lambda} = \frac{\varrho_2}{\varrho_1 + \varrho_2 - \varrho_1 \varrho_2} \quad (< 1) \qquad (23)$$

$$P(T_{W2} > 0 \mid T_{W1} > 0) \overset{\mathrm{def}}{=} {_1W_2} = {_1F} \cdot E_{2,n_2}(A_2)$$

with

$$_1F = \frac{\mu_2 \cdot \varepsilon_1}{\lambda(\varepsilon_1 + \mu_2 - \lambda)} = \frac{1}{\varrho_1 + \varrho_2 - \varrho_1 \varrho_2} \quad (> 1) \qquad (24)$$

Comparing the two probabilities of waiting, we get the simple relation

$$_0W_2 = \frac{A_2}{n_2} \cdot {_1W_2} \qquad (25)$$

In fig. 6 these conditional probabilities of waiting are plotted versus the utilization $\varrho_2 = A_2/n_2$ of stage 2 with ϱ_2/ϱ_1 as parameter.

All curves for $_0W_2$ are situated below $W_2 = E_{2,n_2}(A_2)$, all curves for $_1W_2$ above.

2.6 Conditional waiting time distribution functions in stage 2

If it is further assumed that a considered test-call must wait in the second stage, for each test-call the same conditional meeting probabilities are obtained as for all waiting calls in stage 2 together:

$$P(X_{2,\mathrm{arr}\,2} = x \mid T_{W2} > 0) = \left(1 - \frac{A_2}{n_2}\right) \cdot \left(\frac{A_2}{n_2}\right)^{x-n_2} \qquad x \geq n_2 \qquad (26)$$

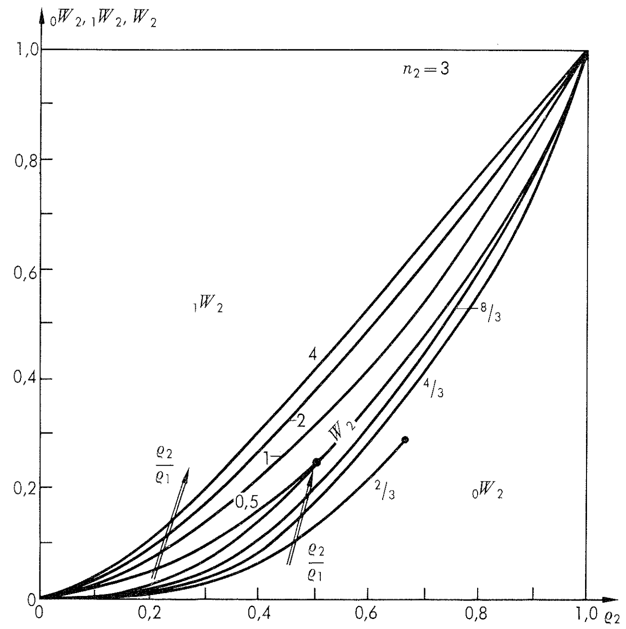


Fig. 6. Conditional probabilities of waiting.

Therefore the following is true:

As soon as a call must wait in the second stage, the number of calls met and therefore the further fate (waiting time, flow time) is independent of the previous fate in stage 1 (waiting and/or service time).

(With the results of chapter 4 it can be shown that this is only true, if no special number of calls met in the first stage $X_{1,\mathrm{arr}\,1}$ (> 0) is assumed.)

If, for example, the queue discipline in stage 2 is FIFO, each of these test calls waiting in the second stage has the same (complementary) conditional d.f. as all calls together:

$$P(T_{W2}>t \mid T_{W2}>0) \overset{\text{def}}{=} W_{2W}(>t)=e^{-(\mu_2-\lambda)t} \tag{27}$$

Dropping the condition that the test-call has to wait in the second stage, it is obtained by (27) with (21) to (24)

$$_cP(T_{W2}>t \mid T_{H1}=t_1)=\frac{E_{2,n_2}(A_2)}{\left(\frac{A_2}{n_2}\right)^c}\cdot e^{-(\mu_2-\lambda)(t_1+t)} \qquad c=0,1; \tag{28}$$

$$_cW_2(>t)={}_cF\cdot E_{2,n_2}(A_2)\cdot e^{-(\mu_2-\lambda)t} \tag{29}$$

Formula (29) for unknown service time in stage 1 yields for $n_2=1$ the same expressions given by *Burke* [2].

From (28) it is possible to determine the fate of a call with a certain required service time in the first stage by weighting summation:

$$\begin{aligned} P(T_{W2}>t \mid T_{H1}=t_1)&=P(T_{W2}>t,\, T_{W1}=0 \mid T_{H1}=t_1)+\\ &+P(T_{W2}>t,\, T_{W1}>0 \mid T_{H1}=t_1)=\\ &=\left(1-\varrho_1+\frac{\varrho_1}{\varrho_2}\right)\cdot E_{2,n_2}(A_2)\cdot e^{-(\mu_2-\lambda)(t_1+t)} \end{aligned} \tag{30}$$

3. Determination of total fates

The independence of the flow times in the two stages proved in the literature causes the total *flow* time d.f. to be a simple convolution of the two single stage flow time d.f.'s, cf. 1.3. Determining the total *waiting* time d.f., the dependencies of the waiting times in the single stages must be taken into account. This is now possible, because in the previous chapter these dependencies not only have been shown to exist but also *completely* determined, for known and for unknown service time as well.

3.1 Total probability of waiting

The total probability of waiting W is the probability that a call has to wait somewhere in the system.

$$\begin{aligned} W&=P(T_W>0)=P(T_{W1}>0)+\\ &+P(T_{W1}=0)\cdot P(T_{W2}>0 \mid T_{W1}=0)=\\ &=P(T_{W1}>0)+P(T_{W2}>0)-P(T_{W1}>0,\, T_{W2}>0) \end{aligned} \tag{31}$$

Using (23) or (24)

$$W=W_1+W_2-\frac{W_1\cdot W_2}{\varrho_1+\varrho_2-\varrho_1\varrho_2}$$

where

$$W_1=\varrho_1=\frac{\lambda}{\varepsilon_1}, \quad W_2=E_{2,n_2}(A_2), \quad \varrho_2=\frac{A_2}{n_2}=\frac{\lambda}{n_2\varepsilon_2} \tag{32}$$

In fig. 7 the total probability of waiting is shown for $n_2=1$. For comparison, the total probability of waiting assuming independence is also depicted:

$$W_{\mathrm{appr}}=W_1+W_2-W_1\cdot W_2 \tag{33}$$

The total probability of waiting always is less than that value obtained by assuming independence of the waiting times, because preferably such calls wait in the second stage, which have already waited in the first stage.

An analogous relation to (31) yields with $P(T_{W2}>0 \mid T_{W1}=0,\, T_{H1}=t_1)$ from (21) a formula describing the dependence of W from the service time in the first stage:

$$P(T_W>0 \mid T_{H1}=t_1)=W_1+(1-W_1)\cdot W_2\cdot e^{-(\mu_2-\lambda)t_1} \tag{34}$$

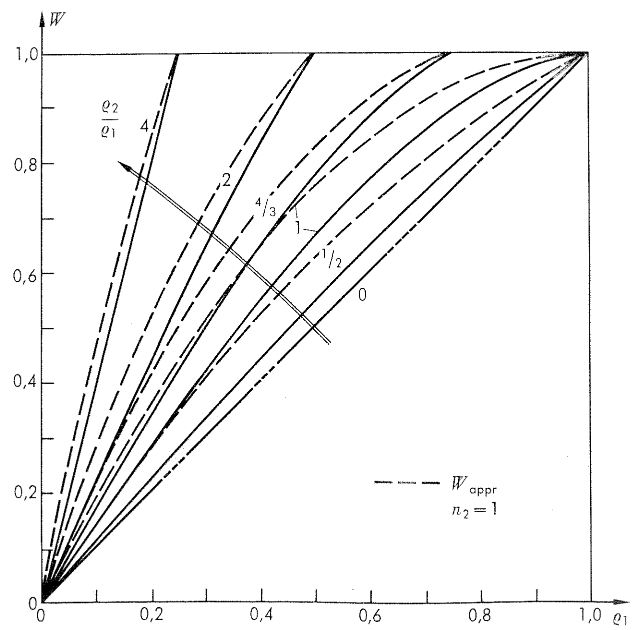


Fig. 7. Total probability of waiting.

3.2 Mean total waiting time of the waiting calls

The mean total waiting time related to *all* calls is (independent of the sequence of the stages) the sum of the mean waiting time in the single stages related to all calls, even if the waiting times are dependent:

$$E(T_W)=E(T_{W1})+E(T_{W2}) \tag{35}$$

So the mean total waiting time of the waiting calls is

$$t_W=\frac{E(T_W)}{W} \tag{36}$$

If, for example, 2 single server stages with $h_1=h_2=h$ are in series ($A_1=A_2=A$) it holds

$$t_W=\frac{2-A}{(1-A)(3-2A)}\cdot 2h \tag{37}$$

3.3 Distribution function of the total waiting time

In a 2-stage system it generally holds

$$\begin{aligned} P(T_W>t)&=P(T_{W2}=0)\cdot P(T_{W1}>t \mid T_{W2}=0)+\\ &+P(T_{W1}=0)\cdot P(T_{W2}>t \mid T_{W1}=0)+\\ &+P(T_{W1}>0,\, T_{W2}>0)\cdot\\ &\cdot P(T_{W1}+T_{W2}>t \mid T_{W1}>0,\, T_{W2}>0) \end{aligned} \tag{38}$$

In this weighted summation of (complementary) conditional d.f.'s all weighting probabilities are known and $P(T_{W2}>t \mid T_{W1}=0)$ was derived in (29). The last d.f. is concerned with calls waiting in both stages. Because it was shown that in the considered system the random values of T_{W1} and T_{W2} are independent of each other if and only if both are > 0, convolution only is allowed under these additional conditions:

$$\begin{aligned} &P(T_{W1}+T_{W2}>t \mid T_{W1}>0,\, T_{W2}>0)=\\ &=P(T_{W1}>t \mid T_{W1}>0) * P(T_{W2}>t \mid T_{W2}>0)=\\ &=e^{-(\varepsilon_1-\lambda)t} * e^{-(\mu_2-\lambda)t} \end{aligned} \tag{39}$$

By the irrelevance of the concrete value of T_{W1} (> 0) it can be shown that

$$\begin{aligned} P(T_{W1}>t \mid T_{W2}=0)&=P(T_{W1}>0 \mid T_{W2}=0)\cdot\\ &\cdot P(T_{W1}>t \mid T_{W1}>0) \end{aligned} \tag{40}$$

So, finally, the d.f. of the total waiting time can be written as:

$$P(T_W>t) \overset{\text{def}}{=} W(>t)=$$

$$=\begin{cases} \left\{\dfrac{\lambda}{\varepsilon_1}-\dfrac{\varepsilon_1-\lambda}{\varepsilon_1+\mu_2-\lambda}\cdot\dfrac{\mu_2}{\varepsilon_1-\mu_2}\cdot W_2\right\}\cdot e^{-(\varepsilon_1-\lambda)t}+ \\ \quad +\dfrac{\varepsilon_1-\lambda}{\varepsilon_1+\mu_2-\lambda}\cdot\dfrac{\varepsilon_1}{\varepsilon_1-\mu_2}\cdot W_2\cdot e^{-(\mu_2-\lambda)t} \\ \qquad\qquad \text{for} \quad \varepsilon_1 \neq \mu_2 & (41) \\ \left\{\dfrac{\lambda}{\mu}+\dfrac{\mu-\lambda}{2\mu-\lambda}\cdot W_2(1+\mu t)\right\}\cdot e^{-(\mu-\lambda)t} \\ \qquad\qquad \text{for} \quad \varepsilon_1=\mu_2=\mu & (42) \end{cases}$$

In case of $n_2=1$ a symmetrical expression in ε_1 and ε_2 can be achieved:

$$\begin{aligned} W(>t)&=\frac{\lambda}{\varepsilon_1}\cdot\frac{\varepsilon_2}{\varepsilon_2-\varepsilon_1}\cdot\frac{\varepsilon_2-\lambda}{\varepsilon_1+\varepsilon_2-\lambda}\cdot e^{-(\varepsilon_1-\lambda)t}+\\ &+\frac{\lambda}{\varepsilon_2}\cdot\frac{\varepsilon_1}{\varepsilon_1-\varepsilon_2}\cdot\frac{\varepsilon_1-\lambda}{\varepsilon_1+\varepsilon_2-\lambda}\cdot e^{-(\varepsilon_2-\lambda)t} \quad \text{for} \quad \varepsilon_1 \neq \varepsilon_2 \end{aligned} \tag{43}$$

If 2 single server stages are in series, the total waiting time d.f. is independent of the sequence of the two stages, i.e. the two stages are *interchangeable* in relation to the total waiting time. This term was used by *Reich* [10] for systems with sequence-independent total flow time distribution function.

It is obvious that the total waiting time d.f. for calls with a preassigned service time in the first stage is obtained by the very same procedure.

3.4 Correlation and error considerations

Assuming independence of the waiting times, the approximate total waiting time d.f. simply is obtained by convolution:

$$W(>t)_{\mathrm{appr}}=P(T_{W1}>t) * P(T_{W2}>t) \tag{44}$$

As a simple example, for 2 equal single server stages ($A_1=A_2=A$, $h_1=h_2=h$) it holds

$$W(>t)_{\mathrm{appr}}=A\left\{2-A+A(1-A)\cdot\frac{t}{h}\right\}\cdot e^{-(1-A)\cdot\frac{t}{h}} \tag{45}$$

The (complementary) d.f. according (42) is

$$W(>t)=\frac{A}{2-A}\left\{3-2A+(1-A)\cdot\frac{t}{h}\right\}\cdot e^{-(1-A)\cdot\frac{t}{h}} \tag{46}$$

In fig. 8 these two d.f.s, which have the same expectation $E(T_W)$, are depicted.

It is easily seen, that the exact curves yield, due to the dependence of the waiting times, a greater variance than the approximated ones.

We have

$$\mathrm{var}(T_W)=\mathrm{var}(T_{W1})+\mathrm{var}(T_{W2})+2\,\mathrm{cov}(T_{W1},T_{W2})$$

with

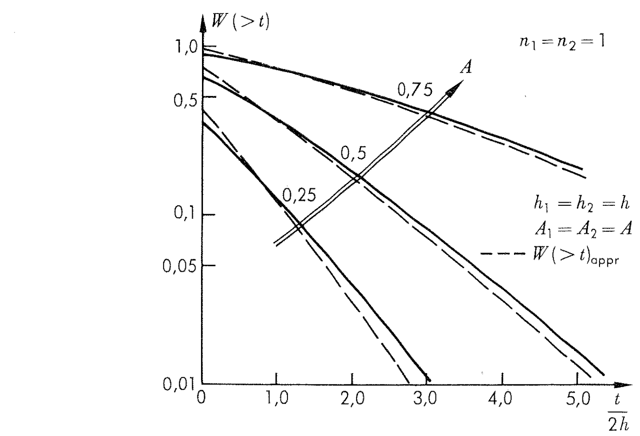


Fig. 8. Total waiting time distribution function.

$$\mathrm{cov}(T_{W1}, T_{W2}) = E(T_{W1} \cdot T_{W2}) - E(T_{W1}) \cdot E(T_{W2}) =$$
$$= \frac{\lambda}{(\varepsilon_1 - \lambda)(\varepsilon_2 - \lambda)} \left\{ \frac{1}{\varepsilon_1 + \varepsilon_2 - \lambda} - \frac{\lambda}{\varepsilon_1 \cdot \varepsilon_2} \right\} \quad (47)$$

The correlation coefficient

$$r(T_{W1}, T_{W2}) = \frac{\mathrm{cov}(T_{W1}, T_{W2})}{\sqrt{\mathrm{var}(T_{W1})} \cdot \sqrt{\mathrm{var}(T_{W2})}}$$

which is in case of var (T_{W1}) = var (T_{W2}) identical with the relative increase of variance, when (positive) covariance is taken into account, is shown in fig. 9 for $n_2 = 1$. For comparison, a simple example for $n_2 > 1$ is also depicted.

Determining the relative errors $(W_{\mathrm{appr}} - W)/W$, which have in principle the same traffic dependencies as the correlation coefficients, we will obtain maximum values decreasing from 33% for $n_2 = 1$ to 4% for $n_2 = 10$.

4. Test-calls with given starting position in stage 1 (random walk)

In this chapter the conditional meeting probabilities

$$P(X_{2,\mathrm{arr}\,2} = x \mid X_{1,\mathrm{arr}\,1} = x_1) \overset{\mathrm{def}}{=} p_2(x \mid x_1)$$

are considered, that is to say, the dependence of the number of calls (queue lengths) met upon the arrivals of the same call in the two stages. In order to obtain explicit results, the case of two single server stages is considered.

4.1 General way of calculation

The walk of a call through the system may be described by a sequence of flow-states, the call is engaged with (path in a *random walk* diagram). Since the fate of the considered call is not influenced by successing calls (FIFO), the *random walk* diagram is a directed graph without loops (fig. 10).

A general flow-state i_1, i_2 related to a considered call is defined such, that i_1 calls are in the first and i_2 in the second stage, where successing calls are irrelevant, but including the call in question. If this flow-state exists, the next event is either the ending of service in the first or second stage with probabilities

$$p_1 = \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2}; \quad p_2 = \frac{\varepsilon_2}{\varepsilon_1 + \varepsilon_2} \quad (48)$$

If a stage is empty, double arrows show that the single-step probability is equal to 1 (reflecting barrier).

Each call starts its walk at the starting flow-state j_1, j_2 and moves on a certain path with certain probability to the absorbing flow state 0, 0 where the call leaves the system. It is not difficult to quote the d.f. of the time a call spends in a flow-state; then it is obvious, how T_{W1}, T_{H1}, T_{W2} and T_{H2} are reflected in this diagram.

Let be:

– $p_F(i_1, i_2)$ the flow-state probability, that the walk of a call from j_1, j_2 to 0, 0 touches i_1, i_2 (j_1, j_2 implicitly understood);

– d_i the number of 'p_i-transitions' of a path (p_i-distance), $i = 1, 2$;

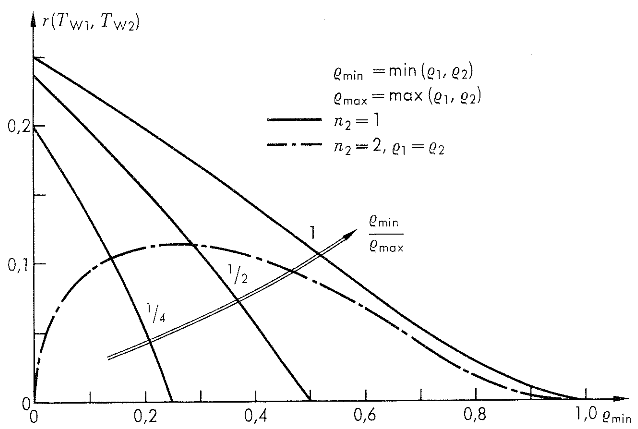


Fig. 9. Correlation coefficients.

– d_0 the number of transitions of a path without alternative.

Then the probability of a certain path is equal to

$$P_{\mathrm{path}} = 1^{d_0} \cdot p_1^{d_1} \cdot p_2^{d_2} \quad (49)$$

and the flow-state probability is equal to the sum of probabilities of all paths leading from j_1, j_2 to the considered flow-state.

If a call uses the transition from $1, x$ to $0, x+1$, it meets exactly x calls upon arrival in the second stage. Thus it is sufficient to calculate the flow state probabilities $p_F(1, x)$ $(x = 0 \ldots j_1 + j_2 - 1)$, from which directly the conditional meeting probabilities for a certain starting pattern $\{x_1, x_2\}$

$$P(X_{2,\mathrm{arr}\,2} = x \mid X_{1,\mathrm{arr}\,1} = x_1, X_{2,\mathrm{arr}\,1} = x_2) \overset{\mathrm{def}}{=} p_2(x \mid x_1, x_2)$$

are obtained. Finally a weighted summation over x_2 yields $p_2(x|x_1)$.

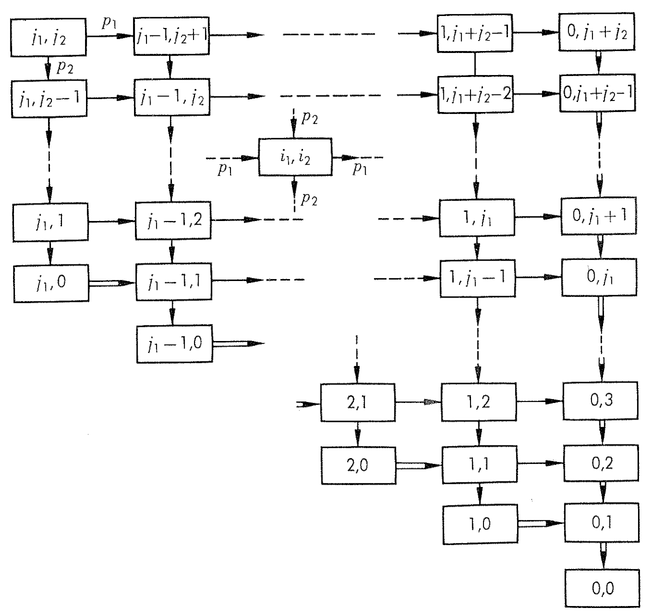


Fig. 10. *Random walk* diagram.

4.2 Flow-state probabilities

For a fixed starting flow-state j_1, j_2 the probability of a path to a fixed flow-state $1, x$ depends on its number d_0 of flow-states with idle time for stage 2. For $j_1 - 1 < x \leq j_1 + j_2 - 1$ there is no path which touches such an idle state. Therefore, all possible paths from a fixed starting position have the same probability and the number of paths is simply to quote. For $0 < x \leq j_1 - 1$ there are paths having $d_0 = 0, 1, \ldots d_{0\max}$ idle states. If $k(d_0, x)$ is the number of paths from flow-state j_1, j_2 $(j_1 > 1, j_2 \geq 0)$ to a flow state $1, x$ $(x > 0)$, which touches d_0 idle states, so it holds:

$$p_F(1, x) = \sum_{d_0 = 0}^{d_{0\max}} k(d_0, x) \cdot p_1^{d_1} \cdot p_2^{d_2} \quad (50)$$

with $d_1 = j_1 - 1 - d_0$, $d_2 = j_1 - 1 + j_2 - x$, $d_{0\max} = j_1 - x$.

To determine $k(d_0, x)$, the following definition and lemma is made (fig. 11):

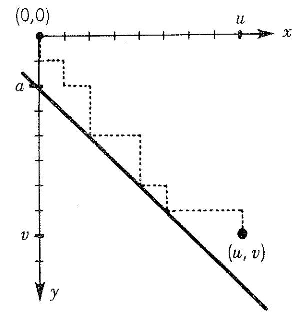


Fig. 11. Directed path.

Let $\nu(d_0, u, v, a)$ be the number of paths leading from (0, 0) to (u, v), which touch but not cross the line $y = x + a$ exactly d_0 times $(a \geq 0,\ u > 0,\ 0 \leq v < u + a)$.

LEMMA:

$$\nu(d_0, u, v, a) =$$
$$= \begin{cases} \dbinom{u+v}{u} - \dbinom{u+v}{u+a} & \text{for} \quad d_0 = 0 \quad (51\,\mathrm{a}) \\[2ex] \dbinom{u+v-d_0}{u+a-1} - \dbinom{u+v-d_0}{u+a} & \text{for} \quad 0 < d_0 \leq v - a + 1 \quad (51\,\mathrm{b}) \end{cases}$$

PROOF: For $a > 0,\ u > 0,\ 0 < v < u + a$, (51 a) is identical with lemma 2 of *Milch* and *Waggoner* [6] and may be proved, also for $a \geq 0,\ u \geq 0,\ 0 \leq v \leq u + a$, with the reflection principle, cf. e.g. *Feller* [4]. (51 b) is for the same conditions part of corollary 2 in [6] and has been proved by the so-called telescope principle. Because of $\nu(d_0, u, v, 0) = \nu(d_0 - 1, u - 1, v, 1)$ for $d_0 > 0$, $a = 0$ is also allowed. Then the validity for $v = 0$ is obvious.

Going back to the random walk diagram with

$$k(d_0, x) = \nu(d_0, j_1 - 1, j_1 - 1 + j_2 - x, j_2) \quad (52)$$

an explicit general formula for the flow-state probabilities $p_F(1, x)$ can be derived.

4.3 Conditional meeting probabilities

If a call starts its walk with flow-state j_1, j_2, upon its arrival $x_1 = j_1 - 1$ respectively $x_2 = j_2$ calls already have been in the system. So with

$$p_2(x \mid x_1, x_2) = p_F(1, x) \cdot p_1 \quad \text{for} \quad x > 0 \quad (53)$$

(50) to (52) yield

$$p_2(x \mid x_1, x_2) = p_2^{x_1 + x_2 - x} \left\{ p_1^{x_1 + 1} \binom{2x_1 + x_2 - x}{x_1} + \right.$$
$$\left. + p_1^x \cdot \sum_{i=0}^{x_1 - x} p_1^i \left\{ \binom{x_1 + x_2 - 1 + i}{x_1 + x_2 - 1} - p_1 \binom{x_1 + x_2 + i}{x_1 + x_2} \right\} \right\} \quad (54)$$

If $x > x_1$ the sum should be replaced by 0.

With

$$\binom{-1}{-1} \overset{\mathrm{def}}{=} 1,$$

it can be shown that (54) generally holds for $x_1, x_2 \geq 0$, $0 \leq x \leq x_1 + x_2$.

These meeting probabilities for a certain starting pattern $\{x_1, x_2\}$ may also be interpreted as state probabilities of a single stage $M/M/1$ with arrival rate ε_1 and service rate ε_2 ($>$ or $\leq \varepsilon_1$) upon the arrival of call number $x_1 + 1$, when at the begin of the arrival process exactly x_2 calls have been in the system. (Investigation of 'time'-dependence where 'time' is represented by the ordinal number of the arriving call.)

This kind of consideration was made by *Takács* [12], who derived with the theory of homogeneous Markov-chains an extensive expression for the bivariate generating function of these higher transition probabilities. For a rather simplifying special case of this function an explicit result was stated, which yields transferred to this problem

$$p_2(x \mid x_1, 0) = \left(\frac{p_1}{p_2}\right)^x \cdot \sum_{i=x}^{x_1} \frac{i}{2x_1 - i} \binom{2x_1 - i}{x_1} \cdot p_1^{x_1 - i} \cdot p_2^{x_1} \quad (55)$$

By complete induction, accordance between (55) and (54) with $x_2 = 0$ can be shown.

By weighting summation over x_2 the wanted conditional meeting probabilities are obtained:

$$p_2(x \mid x_1) = \sum_{x_2 = 0}^{\infty} P(X_{2,\mathrm{arr}\,1} = x_2 \mid X_{1,\mathrm{arr}\,1} = x_1) \cdot$$
$$\cdot\, p_2(x \mid x_1, x_2) \quad (56)$$

So the general formula is

$$p_2(x \mid x_1) = (1 - A_2) \cdot p_2^{x_1 - x} \cdot \sum_{j = \max(0,\, x - x_1)}^{\infty} (A_2 p_2)^j \cdot$$
$$\cdot \left\{ p_1^{x_1 + 1} \binom{2x_1 + j - x}{x_1} + p_1^x \cdot \sum_{i=0}^{x_1 - x} p_1^i \cdot \right.$$
$$\left. \cdot \left\{ \binom{x_1 + j - 1 + i}{x_1 + j - 1} - p_1 \binom{x_1 + j + i}{x_1 + j} \right\} \right\} \quad x, x_1 \geq 0 \quad (57)$$

The sum over i should be replaced by 0 if $x > x_1$.

In fig. 12 these meeting probabilities are plotted and compared with the absolute meeting probability $p_2(x)$. In this example, where the second stage is slower than the first stage ($\varepsilon_2 < \varepsilon_1$), it is vividly shown, how momentary variations (traffic peaks) in the first stage are continued later on in the second stage.

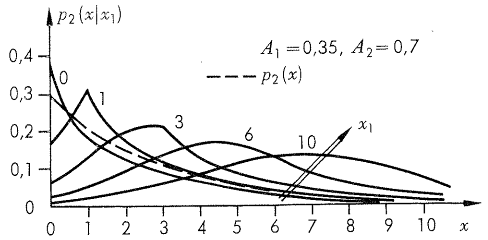


Fig. 12. Conditional meeting probabilities.

If $x > x_1$, no idle situation is possible for stage 2 and (57) can be simplified to

$$p_2(x\,|\,x_1)=(1-A_2)\cdot A_2^{x+1}\cdot\left(\frac{1}{A_1+A_2-A_1A_2}\right)^{x_1+1} \qquad x>x_1\geq 0 \qquad (58)$$

which is for $x_1=0$ identical with ${}_0p_2''(x)$ according (15), derivated in chapter 2.

Equation (57) shows that, for instance, the probability of waiting in the second stage is the higher, the greater the number x_1 of calls met in the first stage. This means that here no such 'limited dependency' is valid as obtained for the waiting times.

4.4 *Further fate values*

The number of calls met in the second stage is completely sufficient for the determination of the further fate in stage 2, as there are the conditional waiting and flow time distributions. E.g. the mean waiting time and the mean flow time can be simply deduced from the conditional expectation

$$E(x\,|\,x_1)\overset{\text{def}}{=}E(X_{2,\text{arr}\,2}\,|\,X_{1,\text{arr}\,1}=x_1)=\sum_{x=0}^{\infty}x\cdot p_2(x\,|\,x_1) \qquad (59)$$

which are shown in figs. 13 and 14.

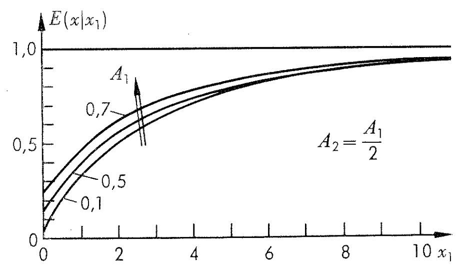


Fig. 13. Conditional expectations in stage 2 ($\varepsilon_2 > \varepsilon_1$).

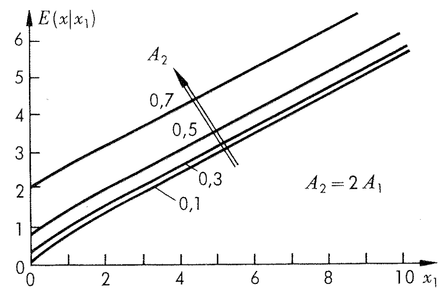


Fig. 14. Conditional expectations in stage 2 ($\varepsilon_2 < \varepsilon_1$).

By these curves the following behaviours (obtained by plausible separate arguments) are confirmed:

If $\varepsilon_2 > \varepsilon_1$, according to a quasistationary behaviour

$$\lim_{x_1\to\infty} E(x\,|\,x_1)=\frac{A_2'}{1-A_2'} \quad \text{with} \quad A_2'=\frac{\varepsilon_1}{\varepsilon_2}, \qquad (60)$$

whereas in case of $\varepsilon_2 < \varepsilon_1$ the expectation value nearly linearly increases with gradient

$$\frac{E(x\,|\,x_1+\Delta x_1)-E(x\,|\,x_1)}{\Delta x_1}\sim 1-\frac{\varepsilon_2}{\varepsilon_1}, \qquad (61)$$

since the probability that the second stage is idle tends to 0 for sufficiently high ε_1 and x_1.

Considering total fates as in chapter 3, besides the total probability of waiting and the inherent mean total waiting time of the waiting calls, the total number of predecessors in stage 1 and 2

$$P(X_{1,\text{arr}\,1}+X_{2,\text{arr}\,2}=x)=\sum_{j=0}^{x}P(X_{1,\text{arr}\,1}=x-j)\cdot p_2(j\,|\,x-j) \qquad (62)$$

further can be determined, which is equivalent to the number of phases a total waiting time is composed of.

Conclusion

For a system with two queues in series the fate of a call in the second stage was determined as a function of all components of the previous fate in stage 1, existing dependencies were illuminated and calculated. Therefore, it is possible to give a more detailed fate prediction of special calls (e.g. with certain required service time).

It was shown that in relation to the waiting times the dependency is limited, so that as practical result the total waiting time d.f. could be determined.

The error made by assuming independence was shown, which cannot always be accepted in this system. Moreover, the dependencies including the correlation coefficient may be considered to give orientation values for corresponding systems, if service time d.f. has no memoryless property as investigated; also it is hoped that these results may give more insight into these problems if the structure is more complicated.

The results given for the total waiting time d.f. are valid for systems with $n_1=1$ and $n_2\geq 1$. In case of $n_1 > 1$ the fate in the second stage depends on the concrete value of the waiting time in the first stage, because of the possibility of overtaking in the first stage.

Finally, it may be noted that it is possible to extend the results of chapters 2 and 3 to a system with several parallel queueing systems in the second stage [13] (tandem queues with group selection).

The author wants to express his thanks to the Deutsche Forschungsgemeinschaft (German Research Society) which supported this work, being part of a research project.

References

[1] *Burke, P. J.*, The Output of a Queuing System. J. Op. Res. 4 (1956), 699–704.

[2] *Burke, P. J.*, The Dependence of Delays in Tandem Queues. Ann. Math. Stat. 35 (1964), 874–875.

[3] *Burke, P. J.*, The Output Process of a Stationary M/M/s Queuing System. Ann. Math. Stat. 39 (1968), 1144–1152.

[4] *Feller, W.*, An Introduction to Probability Theory and Its Applications. Vol. 1, 2nd ed. John Wiley, New York, 1957.

[5] *Jackson, R. R. P.*, Random Queuing Processes with Phase-Type Service. J.R.S.S. Ser. B. 18 (1956), 129–132.

[6] *Milch, P. R., Waggoner, M. H.*, A Random Walk Approach to a Shutdown Queuing System. SIAM J. Appl. Math. 19 (1970) 1, 103–115.

[7] *Nelson, R. T.*, Waiting Time Distributions for Application to a Series of Service Centers. J. Op. Res. 6 (1958), 856–862.

[8] *Reich, E.*, Waiting Times when Queues are in Tandem. Ann. Math. Stat. 28 (1957), 768–773.

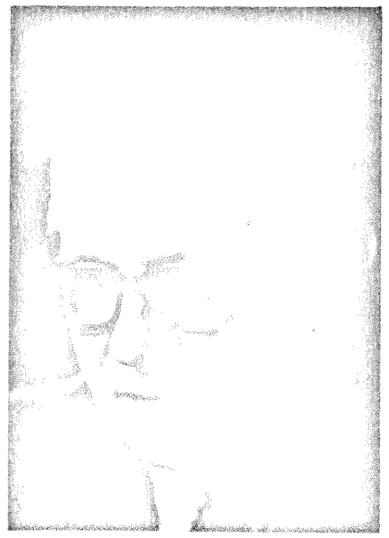
[9] *Reich, E.*, Note on Queues in Tandem. Ann. Math. Stat. 34 (1963), 338–341.

[10] *Reich, E.*, Departure Processes. Proceedings of the Symposium on Congestion Theory (1964), North Carolina, 439–457.

[11] *Syski, R.*, Introduction to Congestion Theory in Telephone Systems. Oliver and Boyd, London, 1960.

[12] *Takács, L.*, Introduction to the Theory of Queues. Oxford University Press. New York, 1962.

[13] *Krämer, W.*, Untersuchung von Systemen mit seriellem Warten. To be published 1974, University of Stuttgart.



Dr.-Ing. Hans-Joachim Dreyer

Vor genau zwei Jahren haben wir die Leser der eR von dem Ausscheiden eines unserer frühesten Redaktionsmitglieder, Dr. phil. nat. *Hans Kaufmann*, unterrichtet. ,,Zeitschriften", so schrieben wir damals, ,,sind lebende Gebilde, die ihre charakterliche Prägung durch diejenigen Kräfte erhalten, die planend und ordnend ihre Träger sind." Im Wissenschaftlichen Zeitschriftenwesen sind diese ,,Träger" fast ausschließlich ehrenamtliche Mitarbeiter, die ein gut Teil ihrer freien Zeit und ihrer Energie der gemeinsamen Unternehmung opfern. So ist es nur natürlich, daß eine Wissenschaftliche Redaktion gelegentliche Veränderungen in Kauf nehmen muß; immerhin ist die Tatsache, daß dies für die eR bis vor kurzem nicht galt, mehr als bemerkenswert.

Heute müssen wir jedoch ankündigen, daß Dr.-Ing. *Hans-Joachim Dreyer*, den wir zur Gründung unserer Zeitschrift geraume Zeit vor dem Erscheinen des ersten Heftes eingeladen hatten, seine Funktion als Mitglied der Wissenschaftlichen Redaktion mit Wirkung vom 1. Januar 1975 aufgegeben hat. *H. J. Dreyer* arbeitete nach Studium und Promotion bis 1956 am Institut für Praktische Mathematik (Prof. *Walther*) in Darmstadt und entwickelte dort als einer der maßgeblichen Kräfte die DERA (Darmstädter Elektronische Rechenanlagen). Danach wechselte er über zu SEL, und seit 1959, d. h. seit ihrem Bestehen, hat er wesentlichen Anteil an der Entwicklung unserer Zeitschrift gehabt. Seine Sachkenntnis, seine allzeit freundlich gewährte Einsatzbereitschaft und seine wohltuende Menschlichkeit werden wir in unseren künftigen Redaktionssitzungen vermissen. Wir haben ihm herzlich zu danken.

R. Oldenbourg