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, Rechenzent. 17 (1975), H. 1, S. 18 – 27
Manuskriptversion: 26. 3. 1974

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

The system is arranged with a system of two unlimited queues, with Poisson traffic 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 multi-server queuing 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 composables of fate in the first stage. Through 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 Übertragungsmassen und zahlreichen weiteren Systemen führt in vielen Fällen auf Systeme oder Subsysteme mit seriell angeordneten Warteschlangen.


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 queuing 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 arrangements 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).

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 \( \lambda \). 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 \( T_M(i, 1, 2) \) of a call in both stages are independent of each other and negative exponentially distributed with distribution functions (d.f.)

\[
P(T_M(i, 1, 2) \leq t) = F_M(i, 1, 2)(t) = 1 - e^{-\mu_i t}
\]

and means

\[
E(T_M(i, 1, 2)) = \frac{1}{\mu_i}
\]

The traffics offered \( A_j \) are defined by

\[
A_j = \sum_{\lambda_j} \lambda_j
\]

and the utilizations

\[
\eta_j = \frac{A_j}{\mu_j}
\]

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

\[
\lambda < \min(\mu_1, \mu_2)
\]

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

\[
T_W = \text{the random waiting time not including service } T_M \text{ and } T_P = \text{the random flow time in stage } 1
\]

1.3 Known results

It is well known, that the output of the first stage is a Poisson process with mean output rate \( \lambda \) (Burke [1], [3], and others), so that each single stage may be computed completely according to the formulae for the M/M/1 queueing system, d. 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 \( n_3 \geq 1 \) included.

Besides \( T_M(i, 1, 2), T_W, T_P \) and the number of calls met upon arrival in stage \( i \) \( (X_M(i, 1, 2)) \), also the number of calls left behind in the first stage \( X_M(i, 1, 2) \) and the departure 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 \( n_3 = 1 \) are dependent. Using the theorem of Jackson and the virtual delay in stage 2, he showed by explicit calculation that the probabilities of waiting

\[ T_W \]
is identical with the probability that $j = x - z$ calls arrive during $t_0$:
The probability that a call with concrete waiting time $T_0 = t_0$ exists and with $x > 0$ calls in the waiting area, 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 $t_0 (t_0 > 0)$ has met $z (z > 0)$ calls upon arrival in the whole stage (this occurs with probability $p_0 (t_0, i, 0)$), exactly $j = n + 1$ calls must be served until its service begins. Since the departure of the last of them coincides with the end of $T_0$, 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 negatively 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 $T_0 = t_0$ of a test-call begins. From the state probabilities $p_0 (x)$ at time $T$ (called starting probabilities), which are independent of the required service time $t_1$, the state probabilities $p_0 (x) / x$ at time $(T + t_1)$ (called meeting probabilities) are determined. (Prefixes $c$ means that the conditional probability related to a special call, where $c = 0$ refers to a call with $T_0 = t_0$ and $c = 1$ to a call with $T_0 > 0$ of known or unknown duration.)

2.3 Starting probabilities in stage 2

Using the theorem of Jackson [3], it is obvious that the state probabilities $p_0 (x)$ at the arrival of a nonwaiting call in stage 1 is independent and according to the absolute state probabilities:

$$p_0 (x) = p_0 (x)$$

(7)

where

$$p_0 (x) = \sum_{n=0}^{\infty} \frac{A^2}{x^1} \sum_{n=0}^{\infty} \frac{p_0 (0, A^2)}{n^2 + 1} x \geq n$$

(8)

with

$$p_0 (0, A^2) = 1 - A^2 + A^2 e^{-A^2}$$

(9)

which is the absolute probability of waiting in the second stage according to the second formula of Erlang, cf. e.g. Sykstil [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 $t_0 (t_0 > 0)$ 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 $r_2$ the input process of the second stage is Poisson (shown in 2.2), so the state probabilities of the second stage at time $T$ will also be distributed according to the absolute values. So

$$p_2 (x) = \sum_{n=0}^{\infty} \frac{A^2}{x^1} \sum_{n=0}^{\infty} \frac{p_0 (0, A^2)}{n^2 + 1} x \geq n$$

(10)

It is obvious that (10) is also valid when the concrete value of $T_0$ is unknown, or when a concrete service time $T_0 = t_0$ is presupposed to a test-call.

2.4 Meeting probabilities in stage 2

Let $p_1 (t_1)$ be the probability that $i$ calls leave the second stage during the service time $T_2 = t_1$ of a test-call. As long as the stage is fully occupied $(X_2 = m)$, the rate of the whole stage to serve one call will be $p_2 (x) = x^c (1 - c)$ (cf. fig. 5). So the meeting probabilities are

$$P (X_2 = m|x, T_2 = t_1) = \sum_{n=0}^{\infty} p_2 (x) / x$$

(11)

where

$$p_2 (x) = \sum_{n=0}^{\infty} \frac{A^2}{x^1} \sum_{n=0}^{\infty} \frac{p_0 (0, A^2)}{n^2 + 1} x \geq n$$

(12)

is the probability of Poisson events during $t_1$. By integration we will get the meeting probabilities for corresponding test-calls with unknown service time $T_2$,

$$p_2 (x) = \int_0^{\infty} p_2 (x) / x$$

(13)

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

$$P (X_2 = m|x, T_2 = t_1) = \frac{p_2 (x)}{x}$$

(14)

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

$$P (X_2 = m|x, T_2 = t_1) = \frac{p_2 (x)}{x}$$

(15)

Similarly, we obtain the meeting probabilities for calls with $T_2 = t_0$ of known or unknown duration ($\alpha = t_0$):

$$p_2 (x) = \sum_{n=0}^{\infty} \frac{A^2}{x^1} \sum_{n=0}^{\infty} \frac{p_0 (0, A^2)}{n^2 + 1} x \geq n$$

(16)

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

$$P (X_2 = m|x, T_2 = t_0) = (1 - \frac{A^2}{x}) \sum_{n=0}^{\infty} \frac{A^2}{x^1} \sum_{n=0}^{\infty} \frac{p_0 (0, A^2)}{n^2 + 1} x \geq n$$

(17)
The total probability of waiting \( W \) is the probability that a call has to wait somewhere in the system.

\[
W = P(T_w > 0) = P(T_w > t) + P(T_w \leq t) = P(T_w > t) + \int_0^t f(t) dt
\]

where \( f(t) \) is the probability density function of the waiting time.

3.2 Total probability of waiting

The total waiting time for a call in the system is given by

\[
W = \sum_{i=1}^{n} W_i
\]

where \( W_i \) is the waiting time for the \( i \)-th call.

3.3 Distribution function of the total waiting time

The distribution function of the total waiting time is defined as

\[
F_W(t) = P(W \leq t) = \int_0^t f_W(u) du
\]

where \( f_W(u) \) is the probability density function of the total waiting time.

3.4 Correlation and error considerations

The correlation between the waiting times of different calls can be approximated using the formula

\[
E[W_i W_j] = \sum_{k=1}^{n} \rho_{ij} \sigma_i \sigma_j
\]

where \( \rho_{ij} \) is the correlation coefficient between \( W_i \) and \( W_j \) and \( \sigma_i \) and \( \sigma_j \) are the standard deviations of \( W_i \) and \( W_j \) respectively.

In conclusion, the total waiting time for a call in the system is a function of the waiting times of individual calls, and the correlation between the waiting times of different calls can be approximated using the formula above.
The correlation coefficient

\[ r(T_{X_1}, T_{X_2}) = \frac{\text{cov}(T_{X_1}, T_{X_2})}{\sqrt{\text{var}(T_{X_1})} \sqrt{\text{var}(T_{X_2})}} \]

which is in case of \( \text{var}(T_{X_1}) = \text{var}(T_{X_2}) \) identical with the relative increase of variance, when (positive) covariation is taken into account, as shown in fig 9 for \( m=1 \). For comparison, a simple example for \( m > 1 \) is also depicted.

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

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

In this chapter the conditional meeting probabilities

\[ P(X_{2,2+1} = x | X_{1,2+1} = y_1, y_2) \]

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 succeeding calls (FIFO), the random walk diagram is a directed graph without loops (fig. 10).

A general flow-stage \( f_j \) is related to a considered call is defined such that \( i_1 \) calls are in the first and \( i_2 \) in the second stage, where succeeding 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

\[ p_1 = \frac{e_1}{e_1 + e_2}, \quad p_2 = \frac{e_2}{e_1 + e_2} \]

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

Each call starts its walk at the starting flow-stage \( f_j \) and moves on a certain path with certain probability to the absorbing flow-state 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; it is obvious, how \( T_{XX}, T_{XY}, T_{YX}, T_{YY} \) and \( T_{YX} \) are reflected in this diagram.

Let be:
- \( p_e(t, e) \) the flow-state probability, that the walk of a call from \( f_j \) to 0, 0 touches \( t_i, e, j_i \), \( t_i, e \) implicitly understood;
- \( d_i \) the number of \( 'p'-transitions \) of a path (p-distance), \( i = 1, 2 \);

4.2 Flow-state probabilities

For a fixed starting flow-stage \( f_j \) the probability of a path to a fixed flow-stage 1, \( x \) depends on its number \( d_{ps} \) of flow-stages with idle time for stage 2. For \( j_1 - 1 < x < j_1 - 1 + d_{ps} \) 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 \( 0 < x < j_1 - 1 \) there are paths having \( d_{ps} = 0, 1, \ldots, d_{ps} \) idle states. If \( k(d_{ps}) \) is the number of paths from flow-stage \( f_j, f_{j_1} (j_1 > 1, j_1 > 0) \) to a flow state 1, \( x > 1, x \geq 0 \), which touches \( d_{ps} \) idle states, so it holds:

\[ p_{t_x}(x) = \sum_{d_{ps} = 0}^{x - 1} k(d_{ps}) p_{t_{ps}}^1 \]

with \( d_{ps} = j_1 - 1 - d_j, d_{ps} = j_1 - 1 - d_j + x, d_{ps} = j_1 - x \).

To determine \( k(d_{ps}) \), the following definition and lemma are made (fig. 11):

4.3 Conditional meeting probabilities

If a call starts its walk with flow-stage \( f_j, f_{j_1} \) upon its arrival \( x_1 = j_1 - 1 \) respectively \( x_2 = j_2 \) calls already have been in the system. So with

\[ p(x_1, x_2) = p(x_1) p(x_2) \quad \text{for } x_1 > 0 \]

(50) to (52) yield

\[ p_X(x_2, x_3) = p_{t_x}(x) + \sum_{d_{ps} = 0}^{x_2 - 1} p_{t_{ps}} 1 \left( 2 x_2 + x_2 - x_1 \right) \]

with \( p_{t_{ps}} 1 \) the same as \( p_{t_1} \), \( x_2 \geq x_1 \), \( x_3 \geq 0 \). It is shown that \( p_X(x_2, x_3) \) 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 \( x_1 \) and service rate \( x_2 \) (or \( x_1 \) 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 (determination of 'time'-dependence where 'time' is represented by the ordinal number of the arriving call).

This kind of consideration was made by Takacs [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

\[ p_{f_x}(x_1, x_2) = p_{t_x}(x) \sum_{d_{ps} = 0}^{x_2 - 1} \left( 2 x_2 + x_2 - x_1 \right) p_{t_{ps}} 1 \]

by complete induction, agreement between (55) and (54) with \( x_2 = 0 \) can be shown.

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

\[ p_X(x_1) = \sum_{x_2=0}^{x_1} P(X_{2,2x} = x_2 | X_{1,2} = x_1), \quad x_1 \geq 0 \]

(56)

So the general formula is

\[ p_X(x_1) = \left( 1 - A_p \right) p_{t_x}(x) + \sum_{d_{ps} = 0}^{x_2 - 1} \left( A_p \right) p_{t_{ps}} 1 \]

\[ \left( 2 x_2 + x_2 - x_1 \right) \]

with \( x_2 = x_1 \), \( x_2 = 0 \).

The sum over \( x_1 \) should be replaced by \( 0 \) if \( x > x_1 \).
In fig. 12 these meeting probabilities are plotted and compared with the absolute meeting probability \( p_d(x) \). In this example, where the second stage is shorter than the first stage \( x_s < x_l \), it is vividly shown, how momentary variations (traffic peaks) in the first stage are continued later on in the second stage.

\[
E(x|x_l) = E(X_{1,2} | X_{1,1} = x_l) = \sum_{x_l} x \cdot p_d(x|x_l)
\]

which are shown in figs. 13 and 14.

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.

\[
E(x|x_l) = \frac{1}{A_2} \sum_{x_l} x \cdot p_d(x|x_l)
\]

which is for \( x_l = 0 \) identical with \( p_m(x) \) according to (15), derived in chapter 2.

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

\[
P_l(x|x_l) = (1 - A_2)^{-1} \left( \frac{1}{A_1 + A_2 - A_1 A_2} \right)^{x_l + 1}
\]

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

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.

\[
E(x|x_l) = \frac{1}{A_2} \sum_{x_l} x \cdot p_d(x|x_l)
\]

which is for \( x_l = 0 \) identical with \( p_m(x) \) according to (15), derived in chapter 2.

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

\[
P_l(x|x_l) = (1 - A_2)^{-1} \left( \frac{1}{A_1 + A_2 - A_1 A_2} \right)^{x_l + 1}
\]

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 \( n_2 > 1 \) 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

References


Dr.-Ing. Hans-Joachim Dreyer

Vor genau zwei Jahren haben wir die Leser der eR von dem Ausscheiden einer unserer frühesten Redaktionsmitglieder, Dr. phil. Hans Kaufmann, unterrichtet. „Zeitschriften“ - so schrieben wir damals - „sind lebende Gebilde, die ihre charakteristische Prüfung durch diejenigen Kräfte erhalten, die planend und ordnend ihre Träger sind.“

Im Wissenschaftlichen Zeitschriftenwesen sind diese „Träger“ fast ausschließlich ehrenamtliche Mitarbeiter, die ein gutes 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.