## Copyright Notice



Institute of Communication Networks and Computer Engineering
University of Stuttgart
Pfaffenwaldring 47, D-70569 Stuttgart, Germany
Phone: ++49-711-685-68026, Fax: ++49-711-685-67983
Email: mail@ikr.uni-stuttgart.de, http://www.ikr.uni-stuttgart.de

# DELAY PROBLEMS IN COMMUNICATIONS SYSTEMS · CLASSIFICATION OF MODELS AND TABLES FOR APPLICATION

P. Kühn
Institute of Switching and Data Technics, University of Stuttgart
Seidenstrasse 36
D 7000 Stuttgart 1, F.R. of Germany

ABSTRACT

This paper is devoted to the development of delay tables for the dimensioning of communications systems with respect to the grade of service. Firstly, some important delay problems are reviewed influencing the performance of communications systems. Then, the basic delay system models are classified systematically. In the third part, the development of a set of delay tables is outlined which cover the most frequent applications. The theoretical background and the computational methods being used are also referred to shortly. Finally, two examples are given for a straightforward application of the tables as well as for a more complex case requiring a decomposition method in before.

## DELAY PROBLEMS IN COMMUNICATIONS SYSTEMS

Communications systems can be considered as service systems consisting of various resources (trunks, switches, control devices etc.) which are allocated to sources requesting for service. Generally, a larger number of sources share a limited number of resources where the allocation of resources is carried out under the control of the communications system. There are two fundamental mechanisms of resource allocation if there are all eligible resources being occupied at the instant of an arriving request: loss and delay mode. In the loss mode, a request (call) will be rejected immediately whereas in the delay mode, the request is allowed to wait until the resource becomes available. Loss system operation is mainly applied in speech path connecting networks where lost calls can be repeated by subscribers if necessary. Delay system operation is often used for centralized control devices being accessed by a large number of sources and where lost calls cannot be admitted. At the other hand, the involved delay time must not exceed certain limits by a suitable dimensioning.

### Delay Problems

Delay problems are widely spread in communications and computer systems. In this paper, only some principal delay problems can be referred to, cf. Fig.1.

a) Subscriber switching networks (Fig. 1a)
In telephone switching exchanges, subscribers (SUB) are connected to a smaller number of junctors (relay sets RS) through a concentrating subscriber switching network (SSN). Usually, subscribers (sources) may wait for a junctor (resource) in case of blocking. The SSN and the number of junctors must be dimensioned acc. to a prescribed grade of service for a given traffic amount.

b) Access switching networks (Fig.1b)
Centralized control devices are requested only for a short time during call set-up as, e.g., registers (REG) for reception of dialling numbers, markers for path search and switch operation, or translators for routing informations. These devices are switched on through an access switching network (ASN) as, e.g.,a register-finder, operating usually in delay mode. The dimensioning underlies real- time constraints for the additional delay time being involved.

c) I/O-subsystems of switching computers (Fig.1c)
The central processing unit (CPU) of a stored program controlled (spc) switching system interchanges control informations with its peripheral devices (PD) via input/output subsystems either in an interrupt or a clocked operational mode. Its performance is influenced significantly by the I/O-mode, the priority assignment, and the overhead being necessary for I/O-operations.

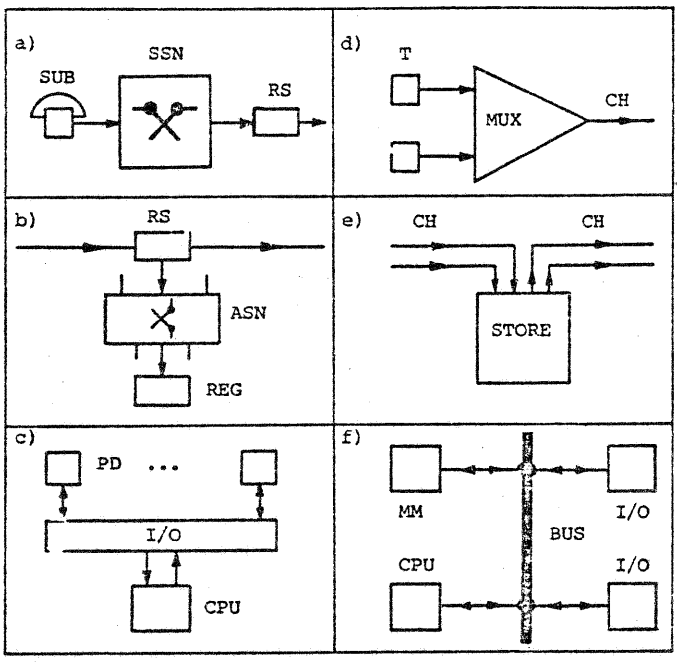



Fig. 1. Delay problems in communications systems

d) Data multiplexers (Fig. 1d)
Data multiplexers (MUX) are used in computer-communications networks to store asynchronously arriving characters of terminals (T) and to multiplex them on a common synchronous data channel (CH). For this purpose, the multiplexer contains a buffer which must be dimensioned with respect to a prescribed upper value of the character loss at a given offered traffic.

e) Store and forward exchanges (Fig. 1e)
In certain types of data exchanges, messages or packets are stored intermediately before they are being forwarded to the neighbor node. The outgoing channels are operated as delay systems. The message delay depends on the channel speed, the signalling mechanism, the routing strategy as well as the control overhead.

f) Switching and general purpose computers (Fig.1f)
Central units of spc switching computers and general purpose computer systems consist of various components as CPU, memory modules (MM), and I/O-channels. The "traffic", i.e. the internal data flow, is switched via a communications system (BUS). The performance of the system can be significantly influenced by additional delays caused by bottlenecks in the various system modules.

### Objective of Delay Tables

For the dimensioning of system resources, queuing theory has proved to be very helpful. Many useful solutions of queuing theory, however, did not yet become applied in practice due to the fact that many solutions are mathematically difficult to understand for a non-specialized user, or that many solutions are rather difficult to evaluate numerically. For these reasons, the solutions of queuing problems are applicable almost only if there is sufficient table support.



Presented at the International Conference on Communications 1977, Chicago.
Conference Record. Vol. I, S. 237-243

During the past, several useful queuing tables have already been published [1-6]. In these tables, usually only one or two different models have been considered. The objective of the newly edited book of delay system tables [9] was to provide a wide spectrum of models differing in structure, arrival processes, service processes, and operating modes enabling a user in most cases to find an adequate model for his special delay problem.

## CLASSIFICATION OF DELAY SYSTEM MODELS

### Modelling

A delay or queuing model is characterized by its structure, its arrival and service processes, and its operating mode. These components are defined in detail by a modelling procedure. Through this procedure, the generation, transfer, delay and service of calls in the real system is mapped to a more or less simplified model consisting of sources, servers, queues, switches, branches and connecting paths. Modelling always involves simplifactions which have to be validated by measurements in practice.

### Delay System Structures

At first, single-stage structures will be considered having one or many servers with full or limited accessibility, and one or many input queues with infinite or finite store capacity. The basic structures are shown in Fig. 2.

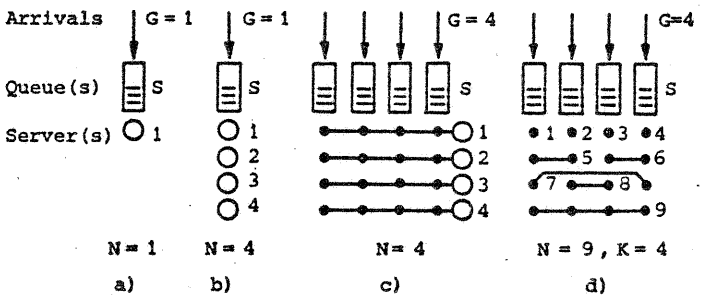



Fig.2. Basic single-stage queuing system structures

a) Single server, single queue structure
b) Many server, single queue structure
c) Many server, many queue structure (full access)
d) Many server, many queue structure (limited access)

The structural parameters are defined as follows:

- N number of servers
- K accessibility ($K \leq N$) = no. of servers accessible from one queue or input group
- G number of queues (input groups)
- S number of waiting places per input queue

In the case of limited accessibility (Fig. 2d), only K out of N servers are accessible from a certain input group. Various interconnection schemes ("gradings") can be distinguished. In the tables [9] only two limiting cases have been considered: 1. "Ideal Erlang Gradings" with $G = \binom{N}{K}$ K! input groups, and 2."Standard Gradings" with the minimum average interconnecting number $M = GK/N = 2$.

In the more general case, the delay system has a multi-stage structure. Basically, there exist two different types of multi-stage structures: firstly, the connection of calls to servers is carried out by a multi-stage switching network with conjugated path selection (link system), or secondly, a call has to pass several stages of service (queuing network). In the tables [9], only single-stage structures have been considered. The analysis of multi-stage systems can also be done by the aid of tables; for this, decomposition methods are used by which the multi-stage system is decomposed into single-stage subsystems, cf. Kampe [8] and Example 2.

### Arrival Processes

The arrival process defines the (random) interarrival times $T_A$ of "calls" being offered to the queuing system. Generally, the arrival process is described by the distribution function (df.) of $T_A$: $A(\leq t) = P\{T_A \leq t\}$. For tabulation, three main types of arrival processes have been considered:

a) single arrivals with constant arrival rate
b) single arrivals with state-dependent arrival rate
c) clocked batch arrivals.

a) Single arrivals with constant arrival rate

Calls arrive individually at the queuing system. The arrival rate $\lambda$ is constant, i.e. independent of the system state. Several types of df.'s of the interarrival time $T_A$ are considered:

| | | |
|---|---|---|
| M | (Markovian) | $A(\leq t) = 1-\exp(-\lambda t)$ |
| D | (Deterministic) | $A(\leq t) = s(t-1/\lambda)$ (Step function) |
| $E_k$ | (Erlangian, order k=1,2,..) | $A(\leq t) = 1-\exp(-\lambda kt)\cdot \sum_{i=0}^{k-1} \frac{(\lambda kt)^i}{i!}$ |
| $H_2$ | (Hyperexponential order 2) | $A(\leq t) = 1-p_1\exp(-t/t_1)-p_2\exp(-t/t_2)$ |
| GI | (General Independent) | $A(\leq t)$ general probability df. |

In case of $H_2$ and GI, the arrival process is further specified by the coefficient of variation $c_A$, where $c_A^2 = VAR[T_A]/E^2[T_A]$.

b) Single arrivals with state-dependent arrival rate

Calls are generated by a finite number Q of uniform traffic sources with call rate $\alpha$ per idle source. The arrival process depends on the momentary number of idle sources and the idle period $T_{IP}$ of an individual source. The idle periods are Markovian according to $P\{T_{IP} \leq t\} = 1 - \exp(-\alpha t)$.

c) Clocked batch arrivals (Fig. 3a)

In this case, calls arrive in batches of size J=0,1,2,... at equidistant time epochs with clock period $t_c$. The distribution of batch size J is assumed to be Poissonian:

$$p_j = P\{J=j\} = \frac{(\lambda t_c)^j}{j!}\cdot\exp(-\lambda t_c), \quad j = 0,1,2,...$$

with mean batch size $E[J] = \lambda t_c$. As short notation, we introduce "CBI" (clocked batch input).

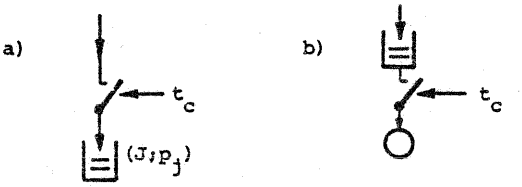



Fig. 3. Clocked batch arrivals (a) and clocked service (b)

### Service Processes

The service process defines the (random) service times $T_S$ of calls with df. $S(\leq t) = P\{T_S \leq t\}$. Two main types of service mechanisms are considered:

a) Immediate service
b) Clocked service

a) Immediate service

After becoming idle, a server can be reoccupied immediately for service. The service times $T_S$ are distributed according to following df.'s:

- M (Markovian)
- D (Deterministic)
- $E_k$ (Erlangian order k)
- $H_2$ (Hyperexponential order 2)
- G (General).

In case of $H_2$ and G, the service process is further specified by the coefficient of variation $c_S$, where $c_S^2 = VAR[T_S]/E^2[T_S]$.

b) Clocked service (Fig. 3b)
Servers with clocked operation can only be occupied at equidistant time epochs 0, $t_c$, $2t_c$,... Calls arriving between two clock epochs have to wait even if there are idle servers. If the number of servers is $N > 1$, all servers start service at the same time. As short notation, we introduce the letter "C" (clocked service).

Operating Modes

The operating mode of a queuing system describes the rules the offered calls are handled by. It can be classified into system operation with respect to delay and loss, hunting modes of servers, queue disciplines, and interqueue disciplines including priority handling.

a) System operation with respect to delay and loss
Tables should include all three basic modes:

- delay system operation
- loss system operation
- combined delay and loss system operation.

In the latter case, a limited waiting room is provided only so that calls may be rejected if all accessible storage places are occupied at the instant of an arrival.

b) Hunting mode
Since all servers are assumed to be homogenous, the hunting mode for idle servers has only an influence on the considered grade of service values in the case of limited access. Generally, sequential hunting with "home position" is assumed. In case of ideal gradings, the results also hold for random hunting.

c) Queue disciplines
The queue discipline defines the rule by which calls waiting within a certain queue are selected for service. Generally, the queue discipline does not influence the mean waiting time but the df. of waiting time. The two most important disciplines were selected for tabulation: FIFO (first-in, first-out) and RANDOM (random selection).

d) Interqueue disciplines
The interqueue discipline defines the rule by which a queue is selected for service. If to each priority class of calls an individual queue is assigned, priority service disciplines are also included. Interqueue disciplines may be:

- NONPREEMPTIVE PRIORITY SERVICE
- PREEMPTIVE PRIORITY SERVICE
- RANDOM SERVICE
- CYCLIC SERVICE
- PROBABILISTIC SERVICE etc.

Tabulation has been carried out only for non-priority disciplines for reasons of volume.

Short Notation of Queuing Models

A slightly generalized scheme of Kendall's short notation will be used to define a queuing model:

$$X_A \;/\; X_S \;/\; N\ (K) - S$$

where

| | |
|---|---|
| $X_A$ | Type of arrival process (M,D,$E_k$,$H_2$,GI,CBI) |
| $X_S$ | Type of service process (M,D,$E_k$,$H_2$,G,C) |
| N | Number of servers |
| K | Accessibility (if limited) |
| S | Number of waiting places (if limited). |

Additional details are given verbally, as for example:

- system operation with respect to delay and loss
- number of sources or queues
- grading types
- queue disciplines
- special features of processes

Examples:

| | |
|---|---|
| $E_k$/M/1 | Delay system, queue discipline RANDOM |
| M/M/N-S | Delay-loss system with many queues |
| M/D/N(K) | Delay system with standard grading |
| CBI/G/1 | Delay system with clocked batch input |

Service System Characteristics

The performance of a service system depends on the system structure, arrival and service processes, and the operating mode. It is described by a number of "characteristic values" which are obtained from a mathematical analysis. The most important values are:

| | |
|---|---|
| $A = E[T_S]/E[T_A]$ | offered traffic |
| $Y = E[X]$ | carried traffic (expectation of the random number X of busy servers) |
| $\Omega = E[Z]$ | mean queue length (expectation of the random number Z of waiting calls) |
| B | probability of loss |
| $W = P\{T_W>0\}$ | probability of delay |
| $w_k = E[T_W{}^k]$ | k-th moment of random waiting times $T_W$ referred to all arriving calls |
| $t_W = E[T_W \mid T_W>0]$ | mean waiting time referred to calls being delayed |
| $W(>t) = P\{T_W>t\}$ | df. of waiting time referred to all arriving calls |
| $c_W = \sqrt{\frac{VAR[T_W \mid T_W>0]}{E^2[T_W \mid T_W>0]}}$ | variation coefficient of waiting time referred to calls being delayed. |

## DEVELOPMENT OF DELAY TABLES

General Objectives

The basic objective of the delay tables was to provide a certain spectrum of models differing in structure, arrival and service processes, and operating modes as well. The final list of models for tabulation has been selected mainly according to arguments of application. From the viewpoint of theory, however, the list of known solutions is rather incomplete. Thus, a number of new solutions has been involved with it. The selection of models has been motivated by the following arguments:

- Providing various types of arrival processes
- Providing various types of service processes
- Consideration of clocked arrivals
- Consideration of clocked service
- Infinite and finite number of sources
- Full and limited accessibility
- Infinite and finite queue capacity
- Single and multi-queue systems
- Two types of queue disciplines
- Applicability to practical problems

Special regard was paid to the df. of waiting times which has been given either explicitly or at least by means of its variation coefficient $c_W$. Standardized df. have been included which allow the approximate construction of the df. of waiting time from its mean and its variation coefficient only.

Outline of Delay Tables

In the following, a set of tables is outlined which has been included in a new book entitled "Tables on Delay Systems" [9]. The tables comprise six major parts:

Part 1: Standardized distribution functions
Part 2: Single server delay systems
Part 3: Many server delay systems (full access)
Part 4: Many server delay systems (limited access)
Part 5: Single server delay-loss systems
Part 6: Many server delay-loss and loss systems

Part 1: Standardized Distribution Functions

| no. | type of df. | order k | variation coefficient c |
|---|---|---|---|
| 1.1 | M | - | 1 |
| | $E_k$ | 2 - 20 | $1/\sqrt{2}$ - $1/\sqrt{20}$ |
| | $H_2$ | 2 | 1.00 - 3.00 |
| 1.2 | Weibull | - | 0.10 - 10.00 |

Part 2: Single Server Delay Systems

| no. | type of model | arrival process | service process | queue discipline |
|---|---|---|---|---|
| 2.1 | M/M/1 | M | M | FIFO, RANDOM |
| 2.2 | M/D/1 | M | D | FIFO, RANDOM |
| 2.3 | $M/E_k/1$ | M | $E_k$, k =2-4 | FIFO, RANDOM |
| 2.4 | $M/H_2/1$ | M | $H_2$, $c_S$=1.25 -2.50 | FIFO, RANDOM |
| 2.5 | D/M/1 | D | M | FIFO, RANDOM |
| 2.6 | $E_k/M/1$ | $E_k$, k = 2-5 | M | FIFO, RANDOM |
| 2.7 | $H_2/M/1$ | $H_2$, $c_A$=1.25 -2.50 | M | FIFO, RANDOM |
| 2.8 | GI/G/1 | GI, $c_A$=0.00 -2.50 | G, $c_S$=0.00 -3.00 | not specified |
| 2.9 | CBI/G/1 | CBI with Poisson batch sizes | G, $c_S$=0.00 -2.00 | not specified |
| 2.10 | M/M/1 finite source | M per idle souce Q = 2 - 50 | M | FIFO, RANDOM |

Part 3: Many Server Delay Systems (Full Access)

| no. | type of model | no. of servers | arrival process | serv. proc. | queue discipline |
|---|---|---|---|---|---|
| 3.1 | M/M/N | N=1-250 | M | M | FIFO, RANDOM |
| 3.2 | M/D/N | N=1-250 | M | D | FIFO, RANDOM |
| 3.3 | GI/M/N | N=1-100 | GI $c_A$=0-2.50 | M | FIFO, RANDOM |
| 3.4 | M/M/N finite source | N=1-100 | M per idle s. Q=2N-10N | M | FIFO, RANDOM |

Part 4: Many Server Delay Systems (Limited Access)

| no. | type of model | access-ibility | no. of servers | grading type | interqueue/ queue discip. |
|---|---|---|---|---|---|
| 4.1 | M/M/N(K) | K=5-50 | N=K-100 | IDEAL, STANDARD | RANDOM/ FIFO |
| 4.2 | M/D/N(K) | K=5-50 | N=K-100 | IDEAL, STANDARD | RANDOM/ FIFO |

Part 5: Single Server Delay-Loss Systems

| no. | type of model | arr. proc. | service process | no. of wait.pl. | queue discipline |
|---|---|---|---|---|---|
| 5.1 | M/M/1-S | M | M | S = 1-10 | FIFO, RANDOM |
| 5.2 | M/D/1-S | M | D | S = 1-10 | FIFO, RANDOM |
| 5.3 | $M/E_k/1$-S | M | $E_k$, k=2-4 | S = 1-10 | FIFO, RANDOM |
| 5.4 | $M/H_2/1$-S | M | $H_2$, $c_S^2$=1.25-2 | S = 1-10 | FIFO, RANDOM |

Part 6: Many Server Delay-Loss and Loss Systems (Full Access)

| no. | type of model | arrival process | serv. proc. | no. of servers | no. of queues | no. of wait.pl. |
|---|---|---|---|---|---|---|
| 6.1 | M/M/N-S | M | M | N=1-100 | G=1 | S =1- 20 |
| 6.2 | M/M/N-S | M | M | N=1-100 | G=1-10 | S =1- 5 |
| 6.3 | M/C/N-S | M | C | N=1- 50 | G=1 | S =N-200 |
| 6.4 | M/G/N-O | M | G | N=1-200 | - | - |
| 6.5 | M/G/N-O finite source | M per idle source Q=2N-10N | G | N=1-100 | - | - |

Theoretical Background and Computation

The treatment of the theoretical background of the tabulated queuing models is beyond the scope of this paper. Here, only the principal methods will be outlined with references to the literature.

The computation of the theoretical results caused a number of problems resulting from extreme parameter combinations, numerical instabilities, and approximations with respect to a wide range of parameter values. Thus, all solutions, exact ones as well as approximate ones, had to be checked by intensive simulation runs with a total of more than $50 \cdot 10^6$ calls. Computations and simulations have been carried out on a Siemens 306 computer.

Part 1: Standardized Distribution Functions

This part provides several types of df.'s as phase type df.'s ($E_k$,$H_2$) and the Weibull df.:

$E_k$ : $$F(t) = P\{T \le t\} = 1 - \exp\left(-\frac{t}{E[T]/k}\right) \cdot \sum_{i=0}^{k-1} \frac{\left(\frac{t}{E[T]/k}\right)^i}{i!}$$

where $k = 1,2,\ldots$ and $c = 1/\sqrt{k} \le 1$.

$H_2$ : $$F(t) = P\{T \le t\} = 1 - p_1 \exp\left(-\frac{t}{t_1}\right) - p_2 \exp\left(-\frac{t}{t_2}\right)$$

where $E[T] = (p_1 t_1 + p_2 t_2)$, $c^2 = \frac{2(p_1 t_1^2 + p_2 t_2^2)}{(p_1 t_1 + p_2 t_2)^2} - 1 \ge 1$.

Weibull: $$F(t) = P\{T \le t\} = 1 - \exp(-[at]^b)$$

where $E[T] = \frac{1}{a}\Gamma\left(1+\frac{1}{b}\right)$, $c^2 = \frac{\Gamma(1+2/b)}{\Gamma^2(1+1/b)} - 1 \gtrless 1$.

The parameters are calculated from E[T] (time base) and given c. The tables are ordered according to the variation coefficient c. The Weibull-df. is tabulated in very small steps of c; it serves as approximant for df.'s of waiting time to be constructed from $t_W$ and $c_W$ in those cases when there was not enough volume to print out the whole df. of waiting time.

Part 2: Single-Server Delay Systems

Delay Systems M/G/1:

Mean values ($W$, $\Omega$, $t_W$, $w_1$) are calculated exactly according to the Pollaczek-Khintchine-formula (imbedded Markov chain analysis)[10-14]. The df. of waiting time is exact in case of M/G/1-FIFO (inverse transformation of the general Pollaczek-Khintchine-formula) and M/D/1-RANDOM (acc. to P.J.Burke[15]). In case of M/M/1, $M/E_k/1$, and $M/H_2/1$ with RANDOM queue discipline, the df. of waiting time is approximate based on the Weibull df. fitting exactly in $t_W$ and $c_W$ [9] ($c_W$ acc. to L.Takács[16]).

Delay Systems GI/M/1 (GI/M/N):

Mean values are calculated exactly in the cases D/M/1, $E_k$/M/1, $H_2$/M/1 according to an imbedded Markov chain analysis[12]. The df. of waiting time is exact (exponentially) in case of FIFO, whereas in case of RANDOM, a Weibull-approximation has been used fitting exactly in $t_w$ and $c_w$[9] ($c_w$ acc. to L. Takács[17]).

Delay Systems GI/G/1 and CBI/G/1:

The solution of GI/G/1 is an approximation based on two moments of the arrival process ($E[T_A]$,$c_A$) and service process ($E[T_S]$,$c_S$) according to W. Krämer and M. Langenbach-Belz[18]. Delay systems of the type CBI/G/1 were reduced exactly to GI/G/1-systems, where the above mentioned solution of GI/G/1 could be used again[18]. The exactly known solutions in case of CBI/D/1[19], CBI/M/1[20] and CBI/$E_k$/1[21] have not been used for reasons of uniformity.

Delay Systems M/M/1 (M/M/N) with a Finite Number of Sources:

Mean values are solved exactly according to F.L. Bauer and H. Störmer[22] (birth and death equations). The df. of waiting time is exact in case of FIFO[22]; in case of RANDOM it is an approximation by exponential sums matching the first, second, and third conditional moments exactly[7].

Part 3: Many Server Delay Systems (Full Access)

Delay Systems M/M/N:

Exact solutions of mean values and df. of waiting time (FIFO) according to Erlang[10-14]. In case of RANDOM, a Weibull-approximation was used matching exactly $t_w$ and $c_w$[7].

Delay Systems M/D/N:

The general theory (C.D. Crommelin[23]) involves N roots of a transcendental equation. The general state equations were solved iteratively yielding the probabilities of state and, subsequently, the mean values straightforwardly[7]. The df. of waiting time can be solved exactly in case of FIFO[23], whereas in case of RANDOM a Weibull-df. was used[7]. The variation coefficient $c_w$ can be given exactly in case of FIFO[7] (using a relationship between the k-th factorial moment of the number of calls in the system and the k-th ordinary moment of waiting times for delay systems of the type M/G/N[13]). In case of RANDOM, an approximation was used for the calculation of $c_w$[7].

Delay Systems GI/M/N:

Generally, the type of GI must be specified explicitly for evaluation[12]. In case of hypoexponential arrival processes, D ($c_A$=0) and a phase-type dr. as series of an $E_k$- and an M-phase ($c_A \leq 1$) has been used[9]. For hyperexponential arrival processes, we have used a $H_2$-df. ($c_A \geq 1$). The analysis is based on an imbedded Markov chain approach as for GI/M/1.

Delay Systems M/M/N with a Finite Number of Sources:

See M/M/1 with a finite number of sources (Part 2).

Part 4: Many Server Delay Systems (Limited Access)

Delay Systems M/M/N(K):

Graded delay systems have been analyzed by two-dimensional birth and death equations using a state-dependent blocking probability of ideal gradings acc. to the "Interconnection Delay Formula" of M. Thierer[24]. For real gradings, this formula was adapted using a modified blocking probability[26]. The variation coefficient of waiting time is an empirical approximation.

Delay Systems M/D/N(K):

The theory of graded delay systems with constant service times is an extension of Crommelin's method being applied to a two-dimensional state description acc. to M. Thierer[25]. Ideal and real gradings can be considered again only by introduction of a proper blocking probability of the grading. The variation coefficient of waiting time is an empirical approximation[7].

Part 5: Single Server Delay-Loss Systems

General theories on the delay-loss system M/G/1-S are either based on the imbedded Markov chain analysis or on the more general supplementary variable method, respectively[12,11]. Both approaches yield the exact mean values ($W,B,\Omega,w_1,t_w$). The supplementary variable method also yields the solution of the df. of waiting time (FIFO) in terms of the Laplace-Stieltjes transform[11]. The inverse transformation, however, may become complicated in certain cases. From this transform, a closed-form expression of $c_w$ has also been derived[27].

Delay-Loss Systems M/M/1-S (M/M/N-S):

In case of FIFO, the df. of waiting time is known explicitly[28]. The exact solution in case of RANDOM involves S roots of an algebraic equation[29]; for tabulation, a two-moments-approximation has been applied with a Weibull-df. fitting exactly in $t_w$ and $c_w$.

Delay-Loss Systems M/D/1-S:

For the FIFO queue discipline, the known Laplace-Stieltjes transform of the df. of waiting time[11] has been transformed into the time domain explicitly[27]. In case of RANDOM, a new solution has been derived by extending Burke's method for the pure delay system M/D/1 with RANDOM queue discipline[15] to a finite space[30].

Delay-Loss Systems M/$E_k$/1-S and M/$H_2$/1-S:

In both cases of FIFO and RANDOM, the df. of waiting time and its higher moments has been derived from multidimensional backward-type equations for the corresponding waiting processes using the method of phases[31].

Part 6: Many Server Delay-Loss and Loss Systems (Full Access)

Delay-Loss Systems M/M/N-S:

In case of a single queue, the solutions are analogously as in case of M/M/1-S. For the multi-queue delay-loss system, an exact (recursive) solution was used which is based on multidimensional state descriptions[7,32]. It holds for the disciplines FIFO, RANDOM, or LIFO with respect to all waiting calls[7].

Delay-Loss Systems M/C/N-S:

The theory of single and multi-server delay-loss systems with clocked service is based on the imbedded Markov chain analysis, cf. N.M.Dor[33] and W.Chu[34]. The equations of state were solved recursively in the cases N = 1 and N = S; for the general case 1 < N < S, an iterative evaluation technique has been used.

Loss Systems M/G/N-O:

For loss systems with Markovian input and general service times, the identical solution holds as in case of exponential service times (Erlang's loss formula in case of an infinite number of sources, and the so-called Engset loss formula in case of a finite number of sources, respectively[10-12]). For finite source input, a distinction is made between call congestion B and time congestion E.

## APPLICATION OF DELAY TABLES

Before using the tables, the user has to find an adequate model. Usually, the modelling procedure yields a rather complex model which is far beyond the scope of standardized tabulation. If there is no direct analysis feasible, the model must either be more simplified or it must be decomposed into submodels of standardized form. Then, the parameters of the model have to be specified. After having found the performance values from the tables, they finally have to be interpreted in terms of the field of application.

The usage of tables will be demonstrated in the following by means of two examples.

Example 1: Dimensioning of registers in a register-controlled switching system

In a register-controlled switching system, centralized registers are connected to requesting relay sets by a one-stage access switching network (ASN). Requests are generated acc.to a Markovian process with arrival rate $\lambda = 1/\text{sec}$. Registers are held for a constant time $t_S = 20\text{sec}$.

a) How many registers are necessary in case of an ASN with full access provided that the mean waiting time of delayed calls ($t_W$) does not exceed 2 sec ?
b) Give the result in case of limited accessibility K = 15 for a standard grading .
c) Give the probability of delay W, the mean waiting time of delayed calls $t_W$, and the probability that a delayed call waits at least t = 10 sec in case of an ASN with full access, N = 25 registers at a prescribed load (offered traffic) of A = 20 Erlangs.
d) What are the results of c) in case of limited accessibility (standard grading, K = 15) ?

Solution:

a) Offered traffic $A = \lambda t_S = 20$ Erlangs, $t_W/t_S < 0.1$

Model M/D/N: (table 3.2) N = 28 registers

b) Model M/D/N(K): (table 4.2) N = 30 registers

c) Model M/D/N:

A/N = 0.80 (table 3.2) W = 0.190
$t_W$ = 2.512 sec

$t/t_S$ = 0.50 (table 3.2) $\frac{W(>t)}{W(>0)} = \begin{cases} 0.0088 & \text{(FIFO)} \\ 0.0412 & \text{(RANDOM)} \end{cases}$

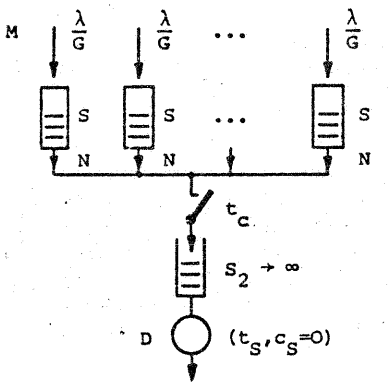
d) Model M/D/N(K):

A/N = 0.80 (table 4.2) W = 0.233
$t_W$ = 2.960 sec
$c_W$ = 0.971 (RANDOM/FIFO)

$t/t_W$=3.38 (table 1.2) $\frac{W(>t)}{W(>0)} \approx 0.032$ .

Example 2: Analysis of delays in a computer-controlled switching system

The I/O-subsystem and the central processor of a computer-controlled switching system may be modelled as a two-stage multi-queue delay-loss system with an intermediate clock mechanism for I/O acc. to Fig. 4.




λ total arrival rate
G peripheral groups
G peripheral buffers
S buffer capacity
I/O system
N max.group size of transf.requ./buffer
$t_c$ clock period
internal queue with unlimited capacity
central processing unit (CPU)

Fig.4. Model of the I/O-subsystem and the central control of a computer-controlled switching system

At the peripheral buffers, control requests arrive acc. to Markovian processes (M) with total arrival rate $\lambda = 400/\text{sec}$ equally distributed over G =10 peripheral groups. Waiting requests are transferred to the internal queue clockwise with clock period $t_c$=10 msec; at each clock instant, only $N \leq 1$ requests can be transferred per buffer. The CPU serves the waiting requests from the internal queue continuously; the service times are constant $t_S$= 2 msec per request including overhead for programme changing (I/O-overhead be neglected).

a) Find the capacity S of the peripheral buffers under the condition that the overflow probability does not exceed a1) 0.0005, a2) 0.000 000 1 ($10^{-7}$).
Find the mean waiting time of delayed requests in the peripheral buffers in case of a1).
b) Find the mean waiting time in the internal queue with respect to all and with respect to all delayed requests passing the CPU.
c) Give the average total flow time $t_F$ with respect to the accepted requests.

Solution:

a) Since the internal queue capacity is unlimited, there is no reaction of the internal system on the peripheral systems. The peripheral buffers can be modelled as M/C/N-S delay-loss systems with N = 1.

Offered traffic per buffer $A_1 = \frac{\lambda}{G} \cdot t_c = 0.4$ Erlangs

Model M/C/1-S: (table 6.3) a1) S = 5 (B = 0.00027)
a2) S = 10

a1) $t_{W1}$ = 8.31 msec

b) The number of requests being transferred into the second stage (internal system) at a clock instant is a function of the state of the first-stage buffers at this instant. The input process of the second stage is no longer recurrent and depends on the first-stage buffer loss and the max. group size N of transferred requests per buffer and clock period.

For an approximate solution, a decomposition method is used. The secondary system is modelled as a delay system of type CBI/D/1 with Poisson-distributed batch sizes. The latter assumption is motivated by the fact of small buffer losses in the first buffer stages and the comparatively large number G; in reality, the maximum batch size is limited by G-N = 10.

Rel.clock period $\tau = t_c/t_S$ = 5

Mean batch size $E[J] = \lambda t_c (1-B) \approx \lambda t_c$ = 4

Offered traffic $A_2 = E[J]/\tau$ = 0.8 Erlangs

Model CBI/D/1: (table 2.9) $t_{W2}$ = 6.984 msec
W = 0.870
$w_{12} = W \cdot t_{W2}$ = 6.078 msec

c) $t_F = t_{W1} + w_{12}$ $t_F$ = 14.388 msec.

Remarks:

Any applied decomposition method must be controlled with respect to its accuracy. Comparisons with simulation of the complete two-stage system yielded the following:

| | Calculation | Simulation |
|---|---|---|
| Peripheral buffers | B = 0.00027<br>$t_{W1}$ = 8.310 msec | B = 0.00036±0.00012<br>$t_{W1}$ = 8.308 ±0.005 msec |
| Internal queue | $t_{W2}$ = 6.984 msec | $t_{W2}$ = 6.170 ±0.153 msec |
| Total system | $t_F$ = 14.388 msec | $t_F$ = 13.400 ±0.159 msec |

The differences between calculation and simulation for the internal queue and for the total system result from two facts: Inaccuracy of the assumed arrival process (CBI with Poisson-distributed batch sizes) and the approximate solution of the secondary system (CBI/D/1).

A more advanced decomposition method has been suggested for general queuing networks considering mean and variance of any process in the network under a recurrence-assumption of processes [35].

## CONCLUSION

In this paper, a set of practice-oriented queuing models has been outlined which cover many cases of application in communications and computer systems. These queuing models have been tabulated and are included in a newly edited table book on delay systems [9]. It has been shown by means of two examples that the grade of service parameters can simply be found and applied for system analysis and design without special knowledge of the queuing theory behind it.


## ACKNOWLEDGEMENTS

The author wishes to express his thanks to Prof. Dr.-Ing. A. Lotze for his steady support and his encouragement for the development of the delay tables. He also thanks his colleagues and all those students who contributed in research and programming.

This work was supported by the Fed. Ministry of Research and Technology of the Fed. Rep. of Germany.


## REFERENCES


(1) L.G.Pack and R.N.Hazelwood, "Finite queuing tables," J. Wiley and Sons, Inc., New York, 1958.

(2) A.Descloux, "Delay tables for finite- and infinite source systems," McGraw-Hill Book Comp., Inc., New York/Toronto/London, 1962.

(3) G.Dietrich et al, "Traffic engineering manual," Standard Elektrik Lorenz AG, Stuttgart, 1966.

(4) M.Thierer, "Delay tables for limited and full availability acc. to the interconnection delay formula (IDF)," 7th Report on Studies in Congestion Theory, Institute of Switching and Data Technics, Univ. of Stuttgart, 1968.

(5) "Telephone traffic theory. Tables and charts," Part 1, Siemens AG, München/Berlin, 1970.

(6) G. Knoblich, "Tabellen für Wartesysteme," Fachberichte der Telefonbau und Normalzeit, Frankfurt/Main, 1, 1975.

(7) G.Kampe and P.Kühn, "Graded delay systems with infinite or finite source traffic and exponential or constant holding time," 8th Internat.Teletraffic Congr.(ITC), Melbourne, 1976, Congressbook, pp. 251/1-10.

(8) G.Kampe, "Analysis of link systems with delay," ICC 77, Internat.Conf.on Communications, Chicago, Ill., 1977.

(9) P.Kühn, "Tables on delay systems." Institute of Switching and Data Technics, Univ. of Stuttgart, 1976 (440 p.).

(10) R.Syski, "Introduction to congestion theory in telephone systems," Oliver and Boyd, Ltd., London, 1960.

(11) J.Riordan, "Stochastic service systems," J. Wiley and Sons, Inc., New York/London, 1962.

(12) R.B.Cooper, "Introduction to queuing theory," The Macmillan Comp., New York, 1972.

(13) D.Gross and C.M.Harris, "Fundamentals of queuing theory," J. Wiley and Sons, Inc., New York/London/Sydney/Toronto, 1974.

(14) L.Kleinrock, "Queuing systems," Vol.1: Theory, J. Wiley and Sons, Inc., New York/London/Sydney/Toronto, 1975.

(15) P.J.Burke, "Equilibrium delay distribution for one channel with constant holding time, Poisson input and random service," BSTJ, vol. 38, pp. 1021-1031, 1959.

(16) L.Takács, "Delay distributions for one line with Poisson input, general holding times and various orders of service," BSTJ, vol. 42, pp. 487-503,1963.

(17) L.Takács, "Delay distributions for simple trunk groups with recurrent input and exponential service times," BSTJ, vol. 41, pp.311-320, 1962.

(18) W.Krämer and M.Langenbach-Belz, "Approximate formulae for the delay in the queuing system GI/G/1," Proc. 8th Internat. Teletraffic Congr. (ITC), Melbourne, 1976. Congressbook, pp. 235/1-8.

(19) M.Langenbach-Belz, "Two-stage queuing system with sampled parallel input queues," El.Rechenanlagen, vol. 17, pp. 71-79, 1975.

(20) H.G.Schwaertzel, "Serving strategies of batch arrivals in common control switching systems," Proc. 7th Internat. Teletraffic Congr. (ITC), Stockholm, 1973, Congressbook, pp. 433/1-4.

(21) H.Weisschuh, "Entwicklung der Systemsoftware für eine rechnergesteuerte PCM-Vermittlungsstelle," to be published.

(22) F.L.Bauer and H. Störmer, "Berechnung von Wartezeiten in Vermittlungseinrichtungen mit kleinen Zubringerbündeln," AEÜ, vol.9, pp.69-73, 1955.

(23) C.D.Crommelin, "Delay probability formulae when the holding times are constant," POEEJ, vol. 25, pp. 41-50, 1932.

(24) M.Thierer, "Delay systems with limited accessibility," Proc. 5th Internat. Teletraffic Congr.(ITC), New York, 1967, Prebook, pp. 203-214.

(25) M.Thierer, "Delay systems with limited availability and constant holding time," Proc. 6th Internat. Teletraffic Congr.(ITC), Munich, 1970, Congressbook, pp. 322/1-6.

(26) P.Kühn, "Waiting time distributions in multi-queue delay systems with gradings," AEÜ, vol. 29, pp. 53-61, 1975.

(27) P.Kühn and E.Lutz, Monogr. No. 431, 1973 (unpublished).

(28) H.Störmer, "Wartezeitlenkung in handbedienten Vermittlungsanlagen," AEÜ, vol. 10, pp. 58-64, 1956.

(29) P.Kühn, "On a combined delay and loss system with different queue disciplines," 6th Prague Conf. on Information Theory, Statistical Decision Functions, Random Processes, Academia, Publ. House of the Czech. Acad. of Sciences, Prague, 1973, pp.501-528.

(30) P.Kühn and A.Röder, Monogr. No. 433, 1974 (unpublished).

(31) P.Kühn and S.Schiessl, Monogr. No. 475, 1975 (unpublished).

(32) P.Kühn, "Combined delay and loss systems with several input queues, full and limited accessibility," AEÜ, vol. 25, pp. 449-454, 1971.

(33) N.M.Dor, "Guide to the length of buffer storage required for random (Poisson) input and constant output rates," IEEE Transact. Electr. Comp., vol. EC-16, pp. 683-684, 1967.

(34) W.W.Chu, "Buffer behavior for Poisson arrivals and multiple synchronous constant outputs," IEEE Transact. on Comp., vol. C-19, pp. 530-534, 1970.

(35) P.Kühn, "Analysis of complex queuing networks by decomposition," Proc. 8th Internat. Teletraffic Congr. (ITC), Melbourne, 1976, Congressbook, pp. 236/1-8.