Special reprint
from
AEÜ ARCHIV FÜR ELEKTRONIK UND ÜBERTRAGUNGSTECHNIK
ELECTRONICS AND COMMUNICATION
Band 29 (1975), pp. 53–61

On Multi-Queue Delay Systems with Gradings*

by Paul Kühn**

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

This paper deals with exact and approximate calculation methods for multi-queue delay systems with gradings. The approximate calculation is based on the "Interconnection Delay Formula" which has been adapted to gradings of various type. For the distribution function of waiting time, the exact and an approximate calculation method are reported, the latter includes only the first and second moment. In a final chapter, the efficiency of various grading types is investigated resulting in a suggestion of an optimum grading structure with respect to delay systems.

Über Wartesysteme mit Mischungen und mehreren Eingangs-Warteschlangen

Es wird über exakte und approximative Berechnungsmethoden für Wartesysteme mit Mischungen und mehreren Eingangs-Warteschlangen berichtet. Aufbauend auf der „Interconnections-Warteformel", wird zur approximativen Berechnung realer Mischungen ein Anpassungsverfahren vorgeschlagen. Außer der exakten Berechnungsmethode für die Wartezeitverteilungsfunktion wird eine Näherungslösung angegeben, welche nur die ersten beiden Momente einbezieht. Schließlich wird die Leistung verschiedener Mischungstypen im Wartesystem untersucht und eine optimale Mischungsstruktur vorgeschlagen.

1. Introduction

Delay mechanisms can be found in automatic telephone and data switching systems for reasons of an economic use of centralized devices (servers) as registers, markers, storages, and processors for common control. The connection to these centralized servers is realized by single-stage or multi-stage connecting networks with full or limited access. An arriving request (call) which cannot be served immediately occupies a storage place and waits for service. The additional delay caused by waiting, however, implies disadvantageous effects as dial-tone delays or signal distortions which must be kept properly low by an adequate dimensioning.

In this paper, single-stage connecting systems with limited accessibility are considered where the outgoing servers of several selector multiplies are partially interconnected (grading). Compared with full accessibility, a grading saves crosspoints at the expense of a somewhat lower grade of service. The subjects of the paper are: firstly, the calculation of delays for given grading structures, operating modes, and offered traffics and, secondly, the question of optimum grading structures for graded delay systems.

During the past, gradings have extensively been studied for loss systems, cf. the survey of A. Lotze [1]. Delay systems with fully accessible servers were investigated first by A. K. Erlang [2]. For graded delay systems, an interpolation method was proposed by E. Gambe [3] using results of full accessibility. M. Thierer [4], [5] derived expressions for the probability of waiting and the mean waiting time on the basis of a two-dimensional state description using the combinatorial blocking probability for ideal gradings ("Interconnection Delay Formula IDF"). Combined delay and loss systems with gradings have been investigated by the author [6], [7], [8]. Multi-stage connecting networks with waiting were studied by E. Gambe [9] and L. Hieber [10].

After a more detailed statement of the problems in Chapter 2, the exact calculation of state probabilities, mean values, and the probability distribution function (pdf) of waiting time for multi-queue delay systems with gradings is outlined in Chapter 3. The studies in [4], [5] were focused mainly on mean values (probability of delay, mean queue length, and mean waiting time) for ideal gradings. Based on these investigations, extensions are presented with respect to real gradings and the pdf of waiting time in Chapter 4. In the final Chapter 5, various grading structures are compared with each other by means of exact calculation and simulations from which an optimum grading type for delay systems is derived.

2. Statement of the Problem

A queuing problem can generally be defined by the system structure, the operating mode, the input process, and the service time characteristic. In this chapter, the basic assumptions and problems will be discussed more in detail.

* Revised manuscript of a paper presented at the 7th International Teletraffic Congress (ITC), Stockholm, June 13–20, 1973.

** Dr. P. Kühn, Institut für Nachrichtenvermittlung und Datenverarbeitung der Universität, D-7 Stuttgart 1, Seidenstrasse 36.

2.1. System structure of multi-queue delay systems

2.1.1. General structure

The multi-queue delay system consists of g input queues (grading group queues), each of them is assigned to an input process of calls. The calls are served by n servers which are fully or partially interconnected (commoned). For partially interconnected servers, calls of each group can only hunt k out of n servers (k accessibility). In Fig. 1, an example is given having $n = 8$ servers, accessibility $k = 4$, and $g = 4$ grading groups.

In a pure delay system, the maximum number of storage places s_j in queue j must be sufficiently large such that no loss occurs ($j = 1, 2, \ldots, g$). A combined delay and loss system is generally obtained by limitation of the queues.

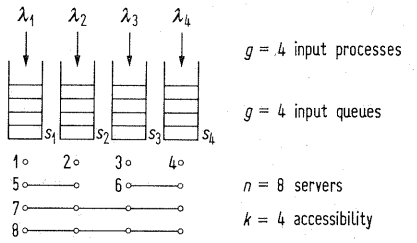

Fig. 1. Multi-queue delay system with limited accessibility.

2.1.2. Types of gradings

The special interconnection scheme (wiring) has an important influence on the efficiency of a grading and has been intensively studied for loss systems, cf. [11]—[21]. Generally, three main wiring methods are applied for the construction of gradings:

- commoning,
- skipping,
- slipping.

Applying these wiring methods on the above example ($n = 8$, $k = 4$, $g = 4$) leads to following gradings, cf. Fig. 2.

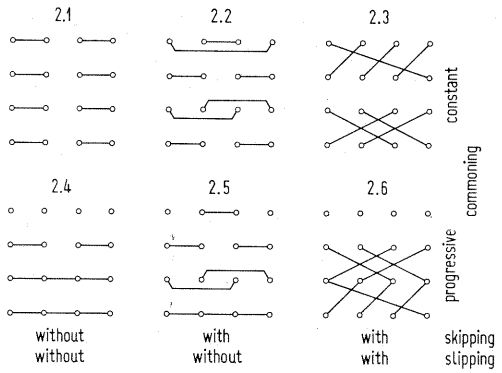

Fig. 2. Types of gradings with various wiring methods Example: $n = 8$, $k = 4$, $g = 4$ ($M = 2$);
2.1. straight homogeneous "grading",
2.2. straight homogeneous grading with skipping,
2.3. homogeneous grading with skipping and slipping,
2.4. straight inhomogeneous grading,
2.5. straight inhomogeneous grading with skipping,
2.6. inhomogeneous grading with skipping and slipping.

Besides the tripel (n, k, g) and the wiring method, the properties of a grading are further characterized by the mean interconnecting number $M = gk/n$ and the matrix (b_{ij}) of the distribution of busies, where b_{ij} denotes the number of interconnections between grading groups i and j [15].

In general, for given n and g the efficiency of a grading increases with k and M. The criteria which were found for efficient gradings in loss systems with respect to the probability of loss and traffic balance hold also for delay systems to a certain extent. Some further criteria, however, have to be considered additionally which originate from waiting.

In the limiting case of full accessibility, all servers are fully interconnected with respect to the g grading groups ($k = n$). Another limiting case is the "Ideal Erlang Grading" having $g = \binom{n}{k} k!$ grading groups. For practice, g assumes too large values. This case is, however, of great theoretical interest since the blocking probability of this grading is exactly known [2].

For practical switching systems, only few types of gradings are applied which have economic advantages as regular construction, simple manufacturing, and easy extension. In Fig. 3, two examples are given, namely a straight inhomogeneous grading ("O'Dell-Grading") of the BPO (Fig. 3.1), and an inhomogeneous grading with skipping and slipping ("Standard Grading") of the German GPO (Fig. 3.2), both with progressive commoning.

For further investigations in this paper, four more types of gradings are shown in Fig. 3: two homogeneous gradings with skipping (Fig. 3.3) or slipping (Fig. 3.4), respectively, and two straight inhomo-

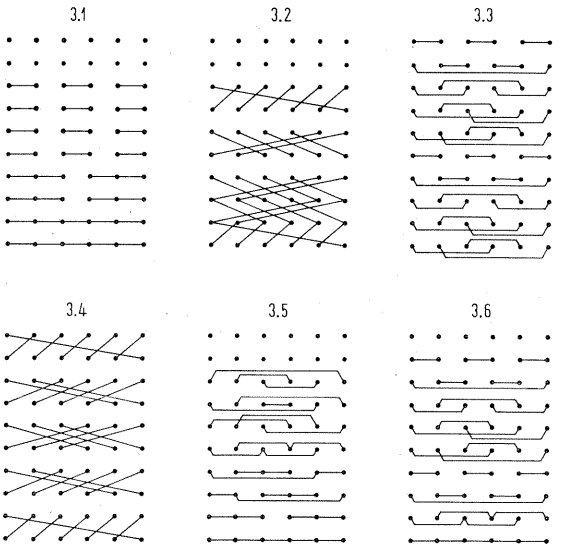

Fig. 3. Practical grading types; example: $n = 30$, $k = 10$, $g = 6$ ($M = 2$);
3.1. O'Dell-grading,
3.2. standard grading
3.3. homogeneous grading with skipping,
3.4. homogeneous grading with slipping,
3.5. high-efficiency grading (loss systems),
3.6. high-efficiency grading (delay systems).

geneous gradings with skipping (Figs. 3.5 and 3.6). Grading Fig. 3.5 represents an individually constructed "high-efficiency grading" with good traffic balance, which is the best one for loss systems, while grading Fig. 3.6 is preferable for delay systems, cf. Chapter 5.

2.2. Operating mode

2.2.1. Hunting disciplines

Servers can be hunted sequentially (with or without homing) or in random order. For gradings with progressive commoning, the sequential hunting method (with homing) is optimal and has, therefore, been assumed for exact calculations and simulations throughout this paper.

2.2.2. Interqueue disciplines

The interqueue discipline controls the service of (non-empty) queues in a multi-queue delay system. For the exact calculation, a general probabilistic discipline was introduced and has already been investigated for special cases as priority, random, and queue lengths-dependent service of queues [6]—[8]. For graded delay systems, the RANDOM service of queues seems to be the most realistic interqueue discipline and will be assumed in this paper together with the idealistic interqueue discipline FIFO (first-in, first-out).

2.2.3. Queue disciplines

The queue discipline controls the service of waiting calls within a queue. In general, the queue discipline has no effect on mean values but on the pdf of waiting time. The influences of queue and interqueue disciplines on mean values and the distribution of waiting times have been studied in [7] by means of exact calculations. In this paper, only FIFO will be assumed.

2.3. Input processes and service times

For the investigations, Markovian properties are assumed, i.e. the interarrival and service times are negative-exponentially distributed, where λ_j the mean arrival rate for arriving calls of group j, $j = 1, 2, \ldots, g$, and ε_i the mean termination rate of server i, $i = 1, 2, \ldots, n$. For the investigations in this paper, only symmetrical conditions are assumed, i.e. $\lambda_j = \lambda/g$, $j = 1, 2, \ldots, g$, $\varepsilon_i = \varepsilon = 1/h$, $i = 1, 2, \ldots, n$, where λ the total mean arrival rate, h the mean service time, and $A = \lambda h$ the "offered traffic".

2.4. Grade of service

The grade of service of a multi-queue delay system may be defined by the main characteristic values as

- probability of waiting,
- probability of loss (finite queues),
- carried traffic,
- mean queue length,
- mean waiting time,
- pdf of waiting time,
- higher moments of the pdf of waiting time.

These characteristic values will be studied under stationary conditions.

3. Exact Calculation of Multi-Queue Delay Systems with Gradings

The exact calculation of multi-queue delay systems with gradings is based on methods of state equations. In principle, the analysis can be performed in two steps [6]—[8]:

i) Solution of a system of linear equations for the stationary probabilities of state,

ii) Solution of a system of linear differential equations for the conditional pdf's (cpdf) of waiting time.

The main traffic values, which characterize the grade of service, can be derived from these values subsequently. In the following two sections, only the fundamental way of solution will be outlined; for a more detailed discussion it is referred to [7].

3.1. The stationary state

3.1.1. Probabilities of state

A system state ξ may be defined by a $(n+g)$-dimensional vector

$$\xi = (\ldots, x_i, \ldots; \ldots, z_j, \ldots), \quad \xi \in \Xi, \tag{1}$$

where $x_i = 0(1)$ if server i is idle (busy), $i = 1, 2, \ldots, n$, and $z_j = 0, 1, \ldots, s_j$ the number of occupied storage places within queue j, $j = 1, 2, \ldots, g$. The set Ξ of system states includes only those states which are physically possible (a queue j can only be built up if at least all accessible servers within grading group j are busy).

The stationary probabilities of state, $p(\xi)$, can be determined from the Kolmogorov-forward-equations considering the service system in equilibrium state

$$q_\xi p(\xi) - \sum_{\pi \neq \xi} q_{\pi\xi} p(\pi) = 0, \quad \xi \in \Xi, \tag{2a}$$

completed by the normalizing relation

$$\sum_{\xi \in \Xi} p(\xi) = 1. \tag{2b}$$

In eq. (2a), $q_{\pi\xi}$ means the coefficient for the transition from state $\pi \neq \xi$ to state ξ, and q_ξ the coefficient for leaving state ξ, where $q_\xi = \sum_{\pi \neq \xi} q_{\xi\pi}$, cf. Fig. 4.

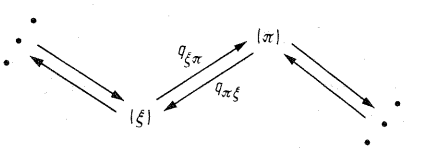

Fig. 4. System state representation with transitions.

The special characteristics of system structure, operating disciplines, and traffic parameters are included in the definitions of system state and transition coefficients which will not be discussed here in detail (cf. [6]—[8]). The numerical evaluation of

the state probabilities is carried out by solution of eqs. (2a, b) using the iterative method of successive overrelaxation.

3.1.2. Characteristic mean values

The most important mean values can be obtained from the probabilities of state by following definitions:

a) probability of waiting for group j:

$$W_j = \sum_{\xi \in \Xi} (1 - \delta_{z_j, s_j}) \prod_{h=1}^{k} (x_{g_{hj}})\, p(\xi)\,, \tag{3}$$

b) probability of loss for group j:

$$B_j = \sum_{\xi \in \Xi} \delta_{z_j, s_j} p(\xi)\,, \tag{4}$$

c) carried traffic on server i:

$$Y_i = \sum_{\xi \in \Xi} x_i\, p(\xi)\,, \tag{5}$$

d) mean queue length of queue j:

$$\Omega_j = \sum_{\xi \in \Xi} z_j\, p(\xi)\,, \tag{6}$$

e) mean waiting time referred to all waiting j-calls:

$$t_{Wj} = \Omega_j / (\lambda_j W_j)\,, \tag{7}$$

where δ_{ij} the Kronecker symbol, and g_{hj} the number of that server which is hunted at step h in group j, $h = 1, 2, \ldots, k$; $j = 1, 2, \ldots, g$.

3.2. Distribution of waiting time

3.2.1. Conditional pdf's of waiting time

For the exact calculation of the waiting time distribution, the waiting process of a test call is considered within the j-th queue. A j-call enters the queue j and starts a waiting process; this process is being "alive" as long as the j-call is waiting and "dies" at that moment when the j-call is selected for service. This waiting process can be constructed from the process of system states by neglecting all those transitions which do not influence the "lifetime" of the j-call under consideration.

For the formal description of the waiting process in queue j, a waiting state ζ_j is introduced which considers all those calls in the system which may have an influence on the waiting time of the considered j-call. ζ_j is built up by the states x_i of all those servers which have no access to group j, and the states of all queues which must be defined dependent on the special queue and interqueue disciplines.

In the simplest case, the interqueue discipline does not depend on the actual lengths of the various queues (e.g., RANDOM-selection of queues). In this case, the waiting state ζ_j can be defined by a $(n - k + g)$-dimensional vector

$$\zeta_j = (\ldots, x_i, \ldots; \ldots, z_\nu, \ldots),\ \zeta_j \in Z_j\,, \tag{8}$$

where $x_i = 0(1)$ if server i is idle (busy), $i \neq g_{hj}$, $h = 1, 2, \ldots, k$, z_ν the number of waiting calls in queue $\nu \neq j$, and z_j the number of predecessors, competitors, or successors of the j-test call in queue j according to whether the queue discipline is FIFO, RANDOM or LIFO, respectively. (For more complicated disciplines cf. [7].) The set Z_j includes all possible waiting states of a j-test call.

The distribution of waiting time for j-calls (T_{Wj}) which met an arbitrary state ζ_j at their arrival is defined by a conditional (complementary) pdf (cpdf)

$$w_j(t \,|\, \zeta_j) = P\{T_{Wj} > t \,|\, \zeta_j\}\,, \quad \zeta_j \in Z_j\,, \tag{9}$$

which are determined from a set of differential equations of the Kolmogorov-backward-type

$$\frac{\mathrm{d}}{\mathrm{d}t} w_j(t \,|\, \zeta_j) = -q_{\zeta_j} w_j(t \,|\, \zeta_j) + \sum_{\eta_j \neq \zeta_j} q_{\zeta_j, \eta_j} w_j(t \,|\, \eta_j)\,, \tag{10}$$

with $\qquad w_j(0 \,|\, \zeta_j) = 1\,; \quad \zeta_j, \eta_j \in Z_j\,.$

In eq. (10), q_{ζ_j, η_j} is the coefficient for transition from waiting state ζ_j to waiting state η_j, and q_{ζ_j} is the coefficient for leaving waiting state ζ_j including "death" of the waiting process with coefficient ε_{ζ_j} according to

$$q_{\zeta_j} = \sum_{\eta_j \neq \zeta_j} q_{\zeta_j, \eta_j} + \varepsilon_{\zeta_j}$$

cf. Fig 5.

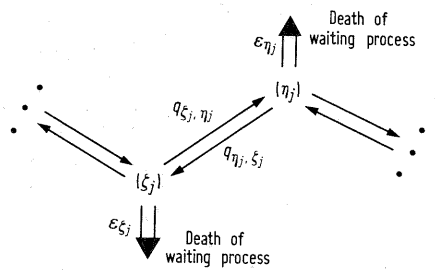

Fig. 5. Waiting state representation with transitions.

For the numerical evaluation of the cpdf's of waiting time, methods of successive power series expansions or approximations using the ordinary moments can be applied, cf. [7].

3.2.2. Total pdf of waiting time and its moments

The total pdf of waiting time can be obtained by averaging over all cpdf's regarding the arrival state probabilities $p(\zeta_j)$:

$$W_j(>t) = P\{T_{Wj} > t\} = \sum_{\zeta_j \in Z_j} p(\zeta_j)\, w_j(t \,|\, \zeta_j)\,, \tag{11}$$

where $W_j(>0) = W_j$, cf. eq. (3).

The r-th ordinary moment m_{jr} of the pdf of waiting time referred to all waiting j-calls is obtained by

$$m_{jr} = -\int_{t=0}^{\infty} t^r \,\mathrm{d}\, \frac{W_j(>t)}{W_j} = \frac{1}{W_j} \sum_{\zeta_j \in Z_j} p(\zeta_j)\, m_{jr}(\zeta_j)\,, \tag{12}$$

where $m_{jr}(\zeta_j)$ the conditional r-th ordinary moments which are obtained from a system of linear equations corresponding to eq. (10). The first moment m_{j1} agrees with the mean waiting time t_{Wj} according to eq (7).

3.3. Numerical example

In this example, a homogeneous grading with $n = 6$ servers, accessibility $k = 4$, and $g = 3$ grading groups will be compared with the corresponding fully accessible system ($n = k = 6$, $g = 3$), both having $s_j = s = 4$ storage places, $j = 1, 2, 3$, with respect to means and distributions of waiting time. The interqueue discipline is RANDOM, the queue discipline is FIFO.

The mean waiting times of waiting calls referred to the mean service time, $\tau_W = t_W/h$, are 0.391, 0.849, and 1.319 for $k = 4$ and 0.326, 0.798, and 1.303 for $k = 6$ for the offered traffics per server $A/n = 0.5$, 1, and 1.5, respectively. This shows that the differences are greatest for small loads ($\tau_W \to 1/k$ for $A \to 0$); on the other hand, the differences vanish for large loads ($\tau_W \to gs/n$ for $A \to \infty$).

Fig. 6 shows the pdf of waiting time referred to all waiting calls, $W(>\tau)/W$, versus the normalized time $\tau = t/h$.

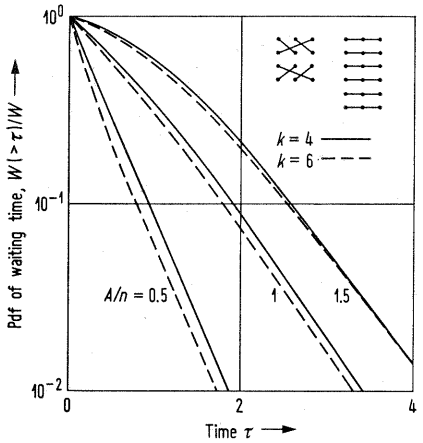

Fig. 6. Pdf of waiting time versus time (exact calculation); parameters: accessibility k, offered traffic per server A/n.

4. Approximate Calculation of Multi-Queue Delay Systems with Gradings

For practical gradings, the number of unknowns is too large for exact calculations so that efficient approximation procedures are necessary. In this chapter, approximation methods are reported for mean values as well as for the pdf of waiting time.

4.1. The stationary state

4.1.1. The "Interconnection Delay Formula" (IDF)

For graded delay systems, M. Thierer [4] suggested a calculation method based on a two-dimensional state description (x, z), where x the number of busy servers, and z the total number of waiting calls. A part of the two-dimensional state space is shown in Fig. 7.

In Fig. 7, $c(x)$ means the blocking probability of the grading in state x, and $r(x, z)$ a conditional probability indicating that no waiting call is served when a server becomes idle. By application of a special symmetry assumption for the statistical equilibrium, cf. Fig. 7, recursive equations were derived which led to explicit expressions for $p(x)$, W, Ω, and τ_W:

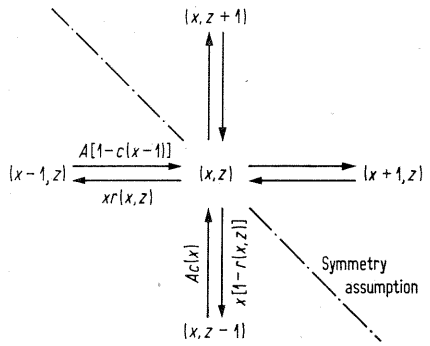

Fig. 7. Two-dimensional system state representation with transitions.

$$p(x) = p(0)\, A^x \prod_{i=0}^{x-1} [1 - c(i)] \Big/ \prod_{i=1}^{x} [i - A\, c(i)], \quad (13)$$

$$W = \sum_{x=k}^{n} p(x)\, c(x), \quad (14)$$

$$\Omega = \sum_{x=k}^{n} p(x) \sum_{i=k}^{x} \frac{A\, c(i)}{i - A\, c(i)}, \quad (15)$$

$$\tau_W = \Omega/(AW), \quad (16)$$

with blocking probability

$$c(x) = \binom{x}{k} \Big/ \binom{n}{k}, \quad k \leqq x \leqq n, \quad (17)$$

according to Ideal Erlang-Gradings [2]. As proved by extensive event-by-event simulations [4], [23], above formulas yield good results for ideal gradings and nonideal gradings with a relatively high mean interconnecting number M. Moreover, in these cases the simulation results turned out to be nearly independent of the interqueue discipline FIFO and RANDOM, respectively.

4.1.2. Adaptation of the IDF to gradings of various type

Simulation results have shown that the results obtained by the IDF are too optimistic for real gradings with small M ($M \approx 2$). Additionally, the more realistic RANDOM-interqueue discipline even increases the results for W, Ω, and τ_W compared to the idealistic FIFO-interqueue discipline in case of gradings with progressive commoning. The influences of the mean interconnecting number M, the grading type, and the interqueue discipline will therefore be taken into account by an adaptation of the IDF (AIDF) to practical gradings as shown in Figs. 3.1, 3.2, and 3.5.

In delay systems with gradings, the storage effect increases the probabilities of blocking occupation patterns compared to loss systems. This effect can be described by a modified blocking probability $c(x, k^*)$ with an adapted nonintegral accessibility $k^* \leqq k$ for delay systems:

$$c(x, k^*) = \binom{x}{k^*} \Big/ \binom{n}{k^*}. \quad (18)$$

This principle has also been applied successfully to loss systems, cf. [22].

Intensive simulation runs for a large number of gradings have shown that k^* depends mainly on k, M, A, the type of grading, and the interqueue discipline. The influences of k^* on W, Ω, and τ_W are indeed slightly different. k^* was adapted such that the mean queue length Ω fitted best possible with simulations which yielded the best overall results. The expressions for k^* are:

$$k^* \approx k - \frac{n^2-k^2}{n^2}\left[\frac{k}{5}\left(\frac{A}{n}\right) + \frac{k-3}{2}\left(\frac{A}{n}\right)^{k/4} + a\frac{k}{5}\left(\frac{A}{n}\right)^4\right]\frac{1}{M-1} \qquad (19a)$$

for standard gradings,

$$k^* \approx k - \frac{n^2-k^2}{n^2}\left[\frac{3}{4}k\left(\frac{A}{n}\right)^{3/2} + a\frac{k}{10}\left(\frac{A}{n}\right)^2\right]\frac{1}{M-1} \qquad (19b)$$

for O'Dell-gradings, where $a = 0$ for the FIFO-, and $a = 1$ for the RANDOM-interqueue discipline. Both formulas hold for $M \geqq 2$.

High-efficiency gradings yield in general slightly better results as standard gradings, but the differences are not much significant so that eq. (19a) can also be used for this case.

For *homogeneous gradings* similar adaptation formulas can be obtained, too. The results for W, Ω, and τ_W, however, are less influenced by the interqueue disciplines FIFO and RANDOM compared to gradings with progressive commoning where RANDOM yields generally worse results than FIFO.

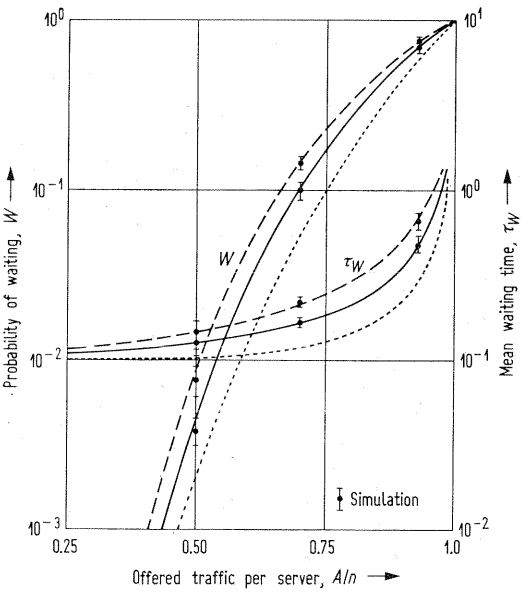

Fig. 8. Probability of waiting and mean waiting time versus offered traffic per server (approximate calculations and simulations); gradings: $n = 60$, $k = 10$,
——— standard grading ($M = 2$),
– – – – O'Dell-grading ($M = 2$),
· · · · · · ideal grading;
interqueue discipline: FIFO.

4.1.3. Numerical example

To demonstrate the efficiency of the adaptation method, some results will be presented. Fig. 8 shows the probability of waiting W and the mean waiting time of waiting calls τ_W (referred to the mean service time) versus the offered traffic per server A/n for a standard grading and an O'Dell-grading having $n = 60$, $k = 10$, $M = 2$ for the FIFO-interqueue discipline. The dotted curves indicate the results obtained by the original IDF for the corresponding ideal grading. The simulation results (with 95% confidence intervals) show the accuracy of the method. The same accuracy was obtained for delay systems with gradings having different mean interconnecting numbers M as well as for the RANDOM-interqueue discipline [24].

4.2. Distribution of waiting time

For the study of waiting time distributions in multi-queue delay systems, two different interqueue disciplines are considered: FIFO and RANDOM. Since the queue discipline is FIFO, there are two operating modes: FIFO/FIFO (F/F) and RANDOM/FIFO (R/F); the first one is identical with "FIFO with respect to all accessible waiting calls".

4.2.1. Approximation by exponential functions

In a first step, the pdf of waiting time can be approximated by exponential functions sufficiently accurate for practical purposes [4], [24]. Compared with simulations, however, there are still certain differences originating from the higher moments of the pdf which are influenced by the operating modes and the grading types, as well. For a more detailed insight, the second and third moments were investigated additionally to the first moment which characterize the pdf of waiting time essentially.

4.2.2. Higher moments of the pdf of waiting time

a) Operating mode F/F.

For full accessibility, the pdf of waiting time is *exponential* and its moments are explicitly known. Extensive simulations showed the following effects in graded delay systems which can also be interpreted by plausible arguments:

- The pdf of waiting time behaves *hypoexponential* with increasing occupancy A/n and increasing ratio n/k.
- The higher moments m_2 and m_3 show only little (if at all) dependence on the mean interconnecting number M.

Based on the first moment $m_1 = \tau_W$, the second moment m_2 can be well approximated for all grading types by

$$\frac{m_2}{m_1^2} \approx \frac{8}{4 + \left[1 - \frac{k}{n}\right]\left(\frac{A}{n}\right)^3}, \qquad (20a)$$

which yields the exact limit value 2 for $A/n \to 0$ or $k \to n$.

b) Operating mode R/F

For full accessibility, only two limiting cases are explicitly known for the pdf of waiting time: $g = 1$, which yields the normal FIFO-queue, and g (or M) $\to \infty$, which is equivalent to a normal RANDOM-queue. By simulations for full and limited accessible servers, the following plausible effects were observed:

- The pdf of waiting time behaves *hyperexponential* with increasing occupancy A/n.
- The hyperexponential behaviour increases with increasing mean interconnecting number M.

The second moment fits well with the following approximation which holds for all grading types including fully accessible systems:

$$\frac{m_2}{m_1^2} \approx \frac{4}{2 - \left(\frac{A}{n}\right)^{n/k}(1 - M^{-2/3})}. \qquad (20\,\mathrm{b})$$

Eq. (20b) yields the exact limits for $A/n \to 0$, or $k = n$ and $M = 1$ or $M \to \infty$, respectively.

4.2.3. Approximation by the gamma distribution

Knowing the first and second moments, the pdf of waiting time can be approximated for both operating modes by a gamma pdf:

$$\frac{W(>\tau)}{W} = 1 - \frac{\gamma(p, b\,\tau)}{\Gamma(p)}, \qquad (21)$$

where $\Gamma(p)$ the complete, and $\gamma(p, z)$ the incomplete gamma functions with

$$p = 1/(m_2/m_1^2 - 1) \quad \text{and} \quad b = p/m_1 .$$

Eq. (21) yields the first and second moments, m_1 and m_2, exactly as they were approximated. Investigations of the third moment of eq. (21), $m_3 = p(p+1)(p+2)/b^3$, have shown that the ratios m_3/m_1^3, obtained from eq. (21) and simulations, differ less than 10% in most cases, respectively.

4.2.4. Numerical example

The accuracy of the approximation method is demonstrated for a standard grading with $n = 30$, $k = 6$, and $M = 2$ for both operating modes F/F and R/F, respectively, cf. Fig. 9. The comparison between calculated and simulated results shows a good agreement.

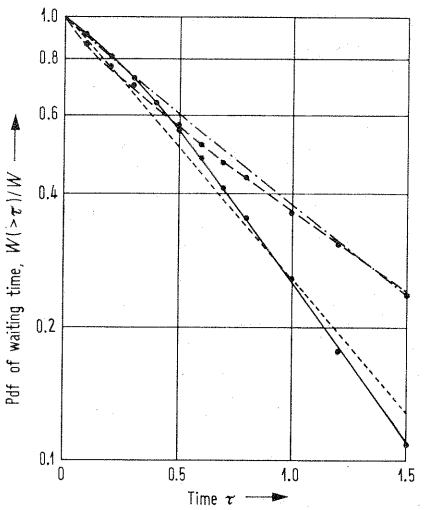

Fig. 9. Pdf of waiting time versus time (approximate calculations and simulations); standard grading: $n = 30$, $k = 6$, $g = 10$ ($M = 2$);
operating modes:
FIFO/FIFO, $A = 27.42$,
——— gamma pdf, ····· exponential pdf;
RANDOM/FIFO, $A = 27.78$,
– – – – – gamma pdf, ·—·—· exponential pdf;
• simulation.

5. Study on Optimum Grading Structures for Delay Systems

5.1. General remarks

In this chapter, the efficiencies of the most important grading types will be compared with each other to find out what grading type yields the best results over the whole range of occupancies. This study refers to a number of investigations which were already made for loss systems, cf. [11]—[21], [25], in order to complete this knowledge with respect to delay systems.

For loss systems and given parameters (n, k, g), two limit theorems were proved exactly [25]: i) for $A \to 0$, the optimum grading type has a maximum number of individual servers, i.e. it consists only of singles and commons, ii) for $A \to \infty$, the optimum grading type is a straight homogeneous grading with skipping. Since these conditions are asymptotically equal in delay systems for $A \to 0$ or $A \to n$, respectively, these results hold also for delay systems. The question is, however, to find out the efficiencies for the actual operating range $0 < A < n$.

5.2. Study by exact calculation

In this study, the six principal wiring methods of Fig. 2 will be studied for a service system with $n = 8$ servers, accessibility $k = 4$, $g = 4$ grading groups, and $s_j = 1$ storage place for each grading group, $j = 1, 2, 3, 4$. The servers are hunted sequentially, the interqueue discipline is RANDOM.

The efficiency of the wiring method can be shown by comparison of the relative probabilities of loss B/B_{ref} versus the occupancy A/n, cf. Fig. 10. (Similar results hold also for the mean total queue length Ω and the total probability of waiting W.)

As shown by Fig. 10, for low occupancies ($0.06 < A/n < 0.4$) the straight inhomogeneous grading with progressive commoning and skipping is best, whereas for higher occupancies ($A/n > 0.4$) the straight homogeneous grading with skipping is best. For higher occupancies, calls queue up and the termination process of all servers determines more and

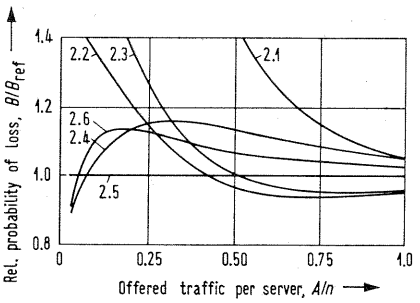

Fig. 10. Efficiency of wiring methods in combined delay and loss systems; relative probability of loss versus offered traffic per server (exact calculation); gradings: $n = 8$, $k = 4$, $g = 4$ according to Fig. 2; interqueue discipline: RANDOM; storage places: $s = 1$ per grading group ($B_{ref} \triangleq$ Fig. 2.5).

more the service quality: in this case, a grading with the best traffic balance is optimal; for given M, the optimum grading is a homogeneous one with a best possible traffic balance. Furthermore, the comparison of Figs. 2.2 and 2.3 shows that for sequential hunting slipping is worse than skipping. The optimum grading for a delay system over the whole range of occupancies should therefore be a grading with certain progression and a considerable homogeneous part with skipping. Grading 2.5 forms a good compromise.

5.3. *Study by simulations*

Another study was performed for the gradings given in Fig. 3 with $n = 30$, $k = 10$, $g = 6$, and $M = 2$ by means of simulations. Gradings 3.1 to 3.5 are well known types, whereas grading type 3.6 was constructed such having certain progression and a large homogeneous part (70%) according to the insight won by the previous study.

Fig. 11 shows the results for pure delay systems by means of the relative mean queue length Ω/Ω_{ref} versus A/n. Again, the gradings with a smooth progression and a good traffic balance (cf. Figs. 3.2 and 3.5) are best for lower occupancies ($A/n < 0.5$), whereas homogeneous gradings (cf. Figs. 3.3 and 3.4) are most efficient for higher occupancies

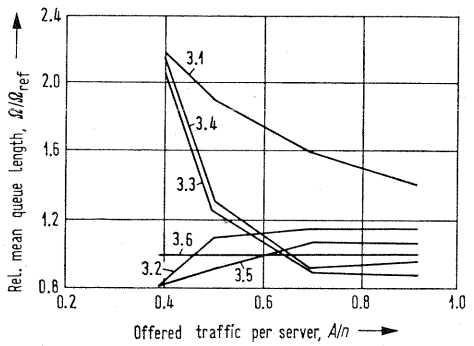

Fig. 11. Efficiency of wiring methods in delay systems; relative mean queue lengths versus offered traffic per server (simulation); gradings: $n = 30$, $k = 10$, $g = 6$ ($M = 2$) according to Fig. 3; interqueue discipline: FIFO; ($\Omega_{ref} \triangleq$ Fig. 3.6).

($A/n > 0.65$). Grading Fig. 3.6 forms a good compromise and can therefore be considered as an optimum grading for delay systems with respect to the whole range of occupancies.

For contrasting, in Fig. 12 the same gradings are compared for loss systems by means of the relative probabilities of loss B/B_{ref} versus A/n. Fig. 12 shows that homogeneous gradings become best not before $A/n = 0.9$ and that grading Fig. 3.5 is the optimum one for loss systems with respect to the whole range of occupancies.

(The given results of Figs. 11 and 12 were obtained by 300000 calls per test run.)

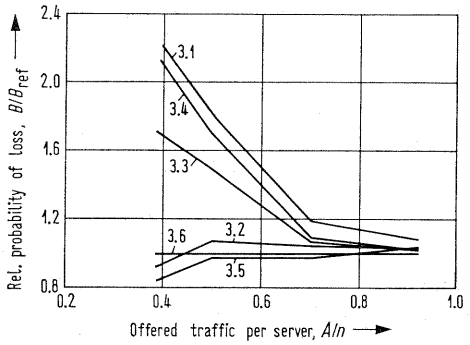

Fig. 12. Efficiency of wiring methods in loss systems; relative probability of loss versus offered traffic per server (simulation); gradings: $n = 30$, $k = 10$, $g = 6$ ($M = 2$) according to Fig. 3 ($B_{ref} \triangleq$ Fig. 3.6).

6. Conclusion

For multi-queue delay systems with gradings, methods have been reported for exact and approximate calculation of mean values and the pdf of waiting time. It has been shown that delay systems with gradings of various type can be calculated sufficiently accurate by the "Interconnection Delay Formula" introducing a modified blocking probability. The pdf of waiting time can be well approximated by a gamma pdf by the aid of its second moment. Finally, it has been shown how the efficiency of a grading is influenced by various wiring methods from which an optimum grading type for delay systems was suggested.

Acknowledgements

The author wishes to express his thanks to Prof. Dr.-Ing. A. Lotze and Dipl.-Ing. G. Kampe for supporting this work and many valuable discussions.

(Received May 2nd, 1974.)

References

[1] Lotze, A., History and development of grading theory. AEÜ **25** [1971], 402—410.

[2] Brockmeyer, E., Halstrøm, H. L. and Jensen, A., The life and works of A. K. Erlang. Transact. Danish Acad. Techn. Sci. No. 2, Copenhagen, 1948.

[3] Gambe, E., A study on the efficiency of graded multiple delay systems through artificial traffic trials. 3. ITC Paris 1961, Doc. 16.

[4] Thierer, M., Delay-tables for limited and full availability according to the Interconnection Delay Formula (IDF). 7th Report on Studies in Congestion Theory. Inst. for Switching and Data Technics, Univ. of Stuttgart, 1968.

[5] Thierer, M., Delay systems with limited availability and constant holding time. 6. ITC München 1970, Congressbook, 322/1—6.

[6] Kühn, P., Combined delay and loss systems with several input queues, full and limited accessibility. 6. ITC München 1970, Congressbook, 323/1—7.

[7] Kühn, P., On the calculation of waiting times in switching and computer systems. 15th Report on Studies in Congestion Theory. Inst. for Switching and Data Technics, Univ. of Stuttgart, 1972.

[8] Herzog, U. and Kühn, P., Comparison of some multiqueue models with overflow and load-sharing strategies for data transmission and computer systems. Symp. on Computer-Communications Networks and Teletraffic. Polytechnic Press of the Polytechnic Inst. of Brooklyn, New York 1972, 449—472.

[9] Gambe, E., Suzuki, T. and Itoh, M., Artificial traffic studies in a two-stage link system with waiting. 5. ITC New York 1967, Prebook, 351—359.

[10] Hieber, L., About multi-stage link systems with queuing. 6. ITC München 1970, Congressbook, 233/1—7.

[11] Longley, H. A., The efficiency of gradings, Part I. Post Off. Elect. Engrs. J. **41** [1948], 45—49.

[12] Elldin, A., On the congestion in gradings with random hunting. Ericsson Technics No. 1 [1955], 35—94.

[13] Lotze, A., Verluste und Gütemerkmale einstufiger Mischungen. Nachrichtentech. Z. **14** [1961], 449—453.

[14] Helms, R. and Kuntze, W., Erhöhung der Leistungsfähigkeit unvollkommener Bündel durch homogene Mischungen. Nachrichtentechnik **12** [1962], 314—320.

[15] Bretschneider, G., Die exakte Bestimmung der Verkehrsleistung kleiner unvollkommener Fernsprechbündel. Nachrichtentech. Z. **16** [1963], 199—205.

[16] Hofstetter, H. and Trautmann, K., Der Einfluß der Mischung auf die Verkehrsleistung der Abnehmerschaltglieder hinter einstufigen Vermittlungsanordnungen. Nachrichtentech. Z. **16** [1963], 635—642.

[17] Rubas, J., Survey of gradings and interconnecting schemes. Telecommunication J. Australia **15** [1965], 120—124.

[18] Stell, F. K., Zweckmäßige Gestaltung von Mischungen. Wiss. Z. Hochschule für Verkehrswesen, Dresden; Part 1: **12** [1965], 271—277, Part 2: **12** [1965], 459—464.

[19] Vanek, N., Ist die homogene Mischung wirklich am besten? Nachrichtentechnik **15** [1965], 213—218.

[20] Cappetti, I., Simulation methods on telephone gradings: analysis of macrostructures and microstructures. 5. ITC New York 1967.

[21] Herzog, U., Lotze, A. and Schehrer, R., Calculation of trunk groups for simplified gradings. Nachrichtentech. Z. **22** [1969], 684—689.

[22] Herzog, U., Calculation of fully available groups and gradings for mixed pure chance traffic. Nachrichtentech. Z. **24** [1971], 627—629.

[23] Kampe, G., Kühn, P. and Ventouris, P.: Mean waiting times and distributions of waiting time in delay systems with real gradings. Inst. for Switching and Data Technics, Univ. of Stuttgart, Monograph No. 371, 1972.

[24] Kühn, P., Waiting time distributions in multi-queue delay systems with gradings. 7. ITC Stockholm 1973, Congressbook, 242/1—9.

[25] Basharin, G. P., Kharkevich, A. D., and Sneps-Sneppe, M. A., Mass service in telephony. Nauka, Moscow, 1968.