IV-1 26

ABSTRACT

THE IMPACT OF QUEUING THEORY ON THE OPTIMIZATION OF COMMUNICATIONS AND COMPUTER SYSTEMS.

by Paul Kühn

Institute for Switching and Data Technics
University of Stuttgart

7 Stuttgart 1, Seidenstrasse 36,
Fed. Republic of Germany
Tel. No. 711 - 2073 2524

The growing amount of telephone and data traffic needs more and more communications and switching facilities. Large computers are built up by various modules causing intensive traffic flows for the exchange of informations. For sufficient handling of the traffic and an economical use of communications, switching, and computer facilities, the system designers have increasingly to take into consideration the traffic characteristics to derive optimum structures and operating strategies for the service facilities.

The paper gives a short review of typical queuing problems in communications and computer systems emphasizing those questions which are most important from the viewpoint of applications. Three examples are reported which refer to various types of optimization criteria characterized by quantitative aspects as well as qualitative aspects. In connection with these examples, some typical analysis methods of queuing theory are outlined to demonstrate their application to practical queuing problems and their impact on the optimization of communications and computer systems if the results can be evaluated suitably. Vice versa, the tasks of actual system design cause a number of new questions to queuing theory to be answered.

THE IMPACT OF QUEUING THEORY ON THE OPTIMIZATION OF COMMUNICATIONS AND COMPUTER SYSTEMS

by Paul Kühn
Institute for Switching and Data Technics
University of Stuttgart, Germany

1. INTRODUCTION

The growing amount of telephone and data traffic needs more and more communications and switching facilities. The modern switching exchanges are generally computer-controlled for technical, economical and administrative reasons, as well. For sufficient handling of the mass traffic and the economical use of the communications and switching facilities, the system designers have to take into consideration the traffic characteristics to derive optimum structures and operating strategies for the service facilities.

The architecture of large computers reveals various independent units as central processing units, memory modules, input/output-channels, and background memories which are connected via a communications and switching system under control of the operating system. Since the traffic flow within a computer system behaves quite statistically, the capacities of the various system components and their operating strategies have to be optimized under similar aspects as in communications systems.

For an efficient optimization of a technical system, methods for analysis and synthesis have to be developed which are suitable for practical applications. This procedure usually starts with the modelling of the service system. The service system model can be analyzed by tools of queuing theory or by simulation. An optimum lay-out, however, requires further studies on different system models considering various structures, operating modes, and costs, as well. For this purpose, it is necessary to find out more general synthesis criteria and to work up the theoretical results by means of curves and tables to put them into practice easily.

The aim of this paper is to point out some typical problems in the communications and computer system design and to demonstrate their interaction with queuing theory.

2. MODELLING AND CRITERIA OF SERVICE SYSTEMS

2.1 General aspects

A service system model (or queuing model) describes the generation and processing of service requests ("calls") under the following general aspects:

- system structures
- operating rules
- input processes
- service processes.

The structure of a queuing model refers to the flow of calls through the system and defines the location and numbers of its components as servers, queues, gates etc. Operating rules are formed by various disciplines concerning the selection of servers, queues or calls as, e.g., hunting, queue and interqueue disciplines, priority strategies, overflow strategies, load-sharing strategies, random branches etc. Input and service processes describe the statistical behavior of interarrival times and service times of the calls, respectively.

Modelling constitutes often a first approximation step by describing complicated actual system structures and events by simplified assumptions. Another critical point is the knowledge of reliable traffic parameters which are often not known exactly during a project phase. Therefore, the accuracy of such a modelling has to be controlled by measurements of the real system or by its simulation.

The criteria of a service system are values characterizing the service quality (grade of service) as, e.g., probabilities of waiting, blocking, and loss, mean queue lengths, carried traffic, means and probability distribution functions (pdf's) of waiting, blocking and response times. Moreover, further important criteria are optimum system structures and operating strategies for high efficiency with respect to the characteristics of the special application.

2.2 Queuing models in communications and computer systems

In this section, some examples will be referred to which represent typical queuing problems in communications and computer systems. It is further aimed to point out the most important questions from the viewpoint of applications (Cited references are not claimed to be exhaustive; for general methods cf. [1-10], special problems cf. [11-50], tables and charts cf. [51-60]).

2.2.1 Single stage service systems

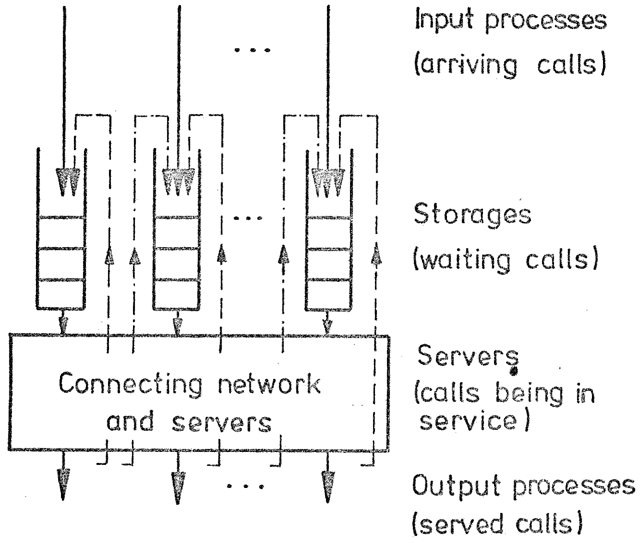

Fig. 1. General single stage service system
——— unidirectional traffic flow
– – – – round-robin traffic flow
—·—·— feedback traffic flow

Fig. 1 shows the general case of a single stage service system which can be found in switching systems for connection of calls with centralized devices as registers, markers, storages, and processors. Calls are often generated by many independent groups of sources assigned to individual storages. The server arrangement is either a single stage connecting network with full or limited accessible servers or a multi-stage connecting network with conjugate switching (link system). The servers can be hunted either sequentially or at random. Queues can be served according to various interqueue disciplines as random selection, cyclic service, priority service, and so on. Service within the queues may be FIFO, RANDOM, LIFO, or other disciplines [11-17].

In computer systems, the server arrangement reduces often to a single server; the traffic flow turns out, however, to be much more complicated as indicated in Fig. 1 by round-robin traffic flow and/or feedback traffic flow for time-sharing processor service strategies. Further strategies are oriented to service times (e.g., shortest-job-first) or can be obtained by introduction of priorities (preemptive, non-preemptive, preemption-delay, preemption-distance), batch service, or sampled batch service [18-25].

Single stage service systems have extensively been studied during the past for various system structures, operating rules, input and service processes, cf. [1-25]. The optimization criteria can be quite different depending on the special application. There are generally two main aims: First, efficient utilization of the (expensive) service equipment and, second, small congestion or small waiting and response times. Both can be achieved by suitable service system structures and efficient operating strategies. In switching systems, usually the number of crosspoints of the connecting network and the capacity of storages should be minimal under meeting the requirements of the grade of service (probabilities of waiting and loss, waiting times). For computer systems, an appropriate operating strategy has usually to be found which guarantees efficient server utilization or small response times (or both).

2.2.2 Systems with queues and servers in series

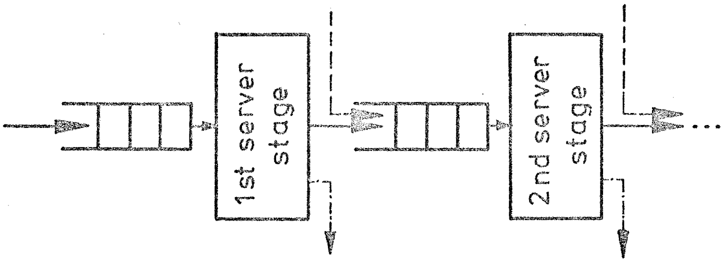

Fig. 2. General service system with queues and servers in series
——— unidirectional traffic flow
– – – – additional feed-in
—·—·— additional branch-off

The analysis of the traffic flow within computer-controlled switching systems and large computers shows that calls are often processed subsequently in successive server stages, as indicated in Fig. 2, (e.g., peripheral storage with pre-processing - central memory and processing unit). Furthermore, additional feed-in and branch-off can appear for calls coming from other pre-processing stages or leaving for other successive processing stages, respectively. Further complications arise when the call-transfers between or within the stages need combined occupations of storage places and servers or when these transfers can only be carried out at certain times, e.g., by sampling according to a sampling clock [26-33].

The optimum lay-out of such systems requires relations between the system parameters (e.g., numbers of servers and storage places, server speeds, clock period, batch-size) and the values characterizing the service quality (e.g., probabilities of waiting and blocking, queue lengths, means and distributions of waiting times, blocking times, flow times etc.). For the processing of dial numbers in switching systems with common control,e.g., the system parameters have to be dimensioned such that the requirements of total flow time through the system (which is related to the "post-dialling delay") are fulfilled under given throughput. Another example is the dimensioning of the capacity of intermediate storages ("buffer storages") in case of the traffic flow between the CPU and an I/O-channel of a computer system: finite capacities cause usually blocking and reduce the total throughput.

2.2.3 Data networks with routing strategies

Data networks have become very important for computer communications. These networks are constructed with respect to economical and safety reasons and operate according to alternate or adaptive routing. By these methods, the traffic is firstly offered to a primary route. In case of overload or breakdown of the primary route, the traffic overflows to a secondary route or even to a third route. The traffic characteristics of such overflow traffics are considerably different compared with normal offered traffics and have, besides the costs, to be taken into consideration for an optimum design of such networks. Methods for optimum design of line-switching as well as message-switching (store- and forward mode) networks have already been suggested and applied, cf. [34-42].

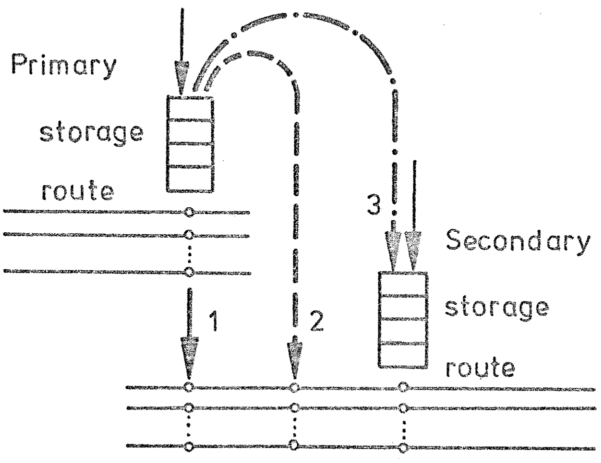

Fig. 3. Overflow strategies for alternate routing

overflow from
——— 1: primary to secondary server group
- - - - 2: primary storage to secondary server group
-·-·- 3: primary storage to secondary storage

In Fig. 3, three different overflow strategies are shown which can be applied to alternate routing in data networks [40-42]. The strategies yield different results with respect to carried traffics on the primary and secondary routes, means and pdf's of waiting times, and loss probabilities, as well. Taking the costs for trunking and storage equipment as well as waiting times into consideration, a suitable strategy can be chosen which minimizes the total costs similar as in telephone networks [37-38].

2.2.4 Multiprogrammed computer systems

An efficient use of central processing units (CPU) and input/output-channels (I/O) in a computer system requires simultaneous work of the CPU and I/O-units and competition of several programs or parts of programs (segments, pages) in main memory with respect to processing. The traffic flow through a

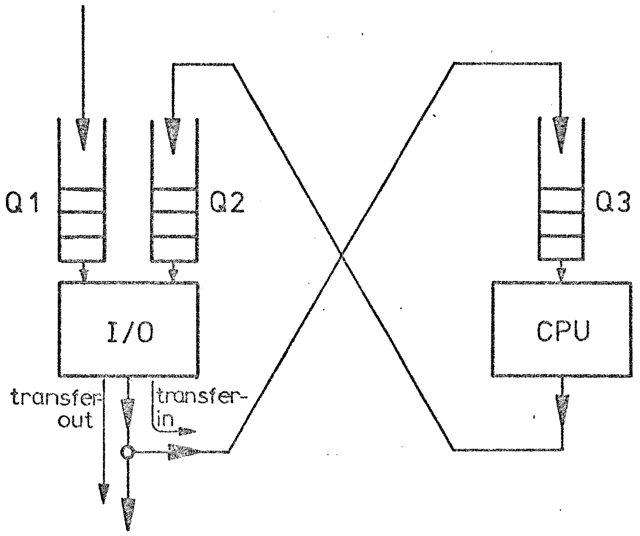

Fig. 4. Queuing model of a multiprogrammed computer system

multiprogrammed computer system can be described by a queuing model acc. to Fig. 4 [49-50]: New programs enter queue Q1 (fast background memory) and wait for a transfer into queue Q3 (main memory) via the I/O-unit. The CPU serves a certain program until completion or interruption by a "page-fault" and generates a request (queue Q2) for the I/O-channel concerning a transfer-out and/or a transfer-in of pages, cf. also [20, 43-50].

The throughput of the system depends on many criteria as page size, main memory capacity, transmission speed of the I/O-channel, CPU-speed, page-replacement strategy, system-overhead phases, distribution of program lengths and computing times, as well. For an optimum working of such a system, the lay-out of components and strategies must be synchronized carefully to the program behavior. Solutions have been derived by mathematical analysis as well as simulations [43-50]. Up to now, simulation studies of such systems are superior since there are only few methods for the analysis of such complicated system structures, strategies, or service time characteristics [50].

3. ANALYSIS AND OPTIMIZATION OF SERVICE SYSTEMS

3.1 General aspects

The analysis of a queuing problem can be carried out either analytically or by simulation. Analytical methods are usually more adequate for general statements and numerical results; simulations are used either when there is no analytical theory or no method to evaluate an analytical theory, and, furthermore, for checking results of approximate analytical methods.

Queuing theory has developed a number of analytical methods, cf. [1-10]. For practical applications, however, only those methods are important which can be suitably evaluated. Numerical evaluation causes often difficulties as, e.g., solutions of systems of linear or differential equations of extremely high order, partial difference equations, and inverse transformations of generating functions or Laplace transforms. If the numerical evaluation can not be carried out often approximate methods have to be derived which yield in most cases results sufficiently accurate with respect to the model assumptions (simplified structures and operating modes, unreliable traffic parameters etc.).

In the following sections, three examples will be given with different optimization aims originating from system design. The analysis methods are also referred to briefly.

3.2 Examples for system analysis and optimization

3.2.1 Study on optimum grading structures for multi-queue delay systems [14].

3.2.1.1 System structures

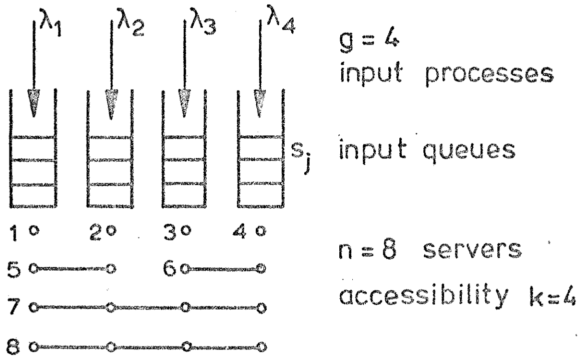

Fig. 5. Multi-queue delay system with limited accessibility

The multi-queue delay system consists of g input queues (grading group queues) of capacity s_j, j = 1,2,...,g, 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). The interconnection scheme is also called as grading. In Fig. 5 an example is given having n = 8 servers, accessibility k = 4, and g = 4 grading groups.

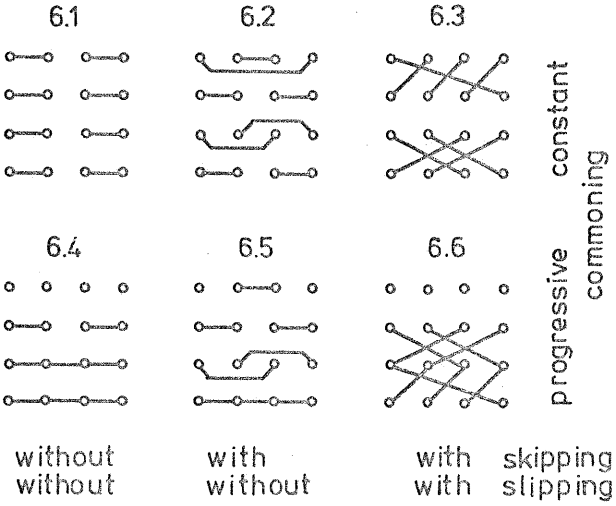

Fig. 6. Types of gradings with various wiring methods

The special interconnection scheme (wiring) has an important influence on the efficiency of a grading and has been intensively studied for loss systems. 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.6.

Besides these structural criteria, the efficiency of gradings is furthermore reflected by the mean interconnecting number M = gk/n and the matrix for the distribution of busies, cf. [14].

3.2.1.2 Operating rules

The operating rules are given by the hunting, interqueue, and queue disciplines: The servers are hunted sequentially; queues are selected for service at random, and the queue discipline may be arbitrary (as far as no pdf of waiting time is considered).

3.2.1.3 Input and service processes

Calls are assumed to arrive acc. to Poisson pdf's with arrival rate $\lambda_j=\lambda/_g$, j = 1,2,...,g. The service times are negative exponentially distributed with termination rate ε.

3.2.1.4 Analysis

The analysis is carried out by means of state equations in the stationary case. A system state ξ may be defined by a (n+g)-dimensional vector

$$\xi = (\ldots,x_i,\ldots;\ldots,z_j,\ldots), \quad \xi \in \Xi \quad , \qquad (1)$$

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

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

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 $q_\xi = \sum_{\pi \neq \xi} q_{\xi\pi}$. The equations of state can be generated by a computer program for arbitrary system structures, interqueue disciplines, arrival and termination rates. They are solved by the method of successive overrelaxation.

From the probabilities of state, further values can be derived as probabilities for waiting and loss, carried traffics, mean queue lengths, and mean waiting times.

For larger gradings or storage capacities, the number of unknowns grows too large so that efficient approximate methods have to be applied. Methods for approximate calculations were developed on the basis of macrostate-descriptions and use of blocking probabilities for gradings of various types which yield results in close accordance with simulations [11, 13, 14]. The pdf of waiting time have also been calculated exactly as well as approximately by means of birth and death processes and higher moments, respectively, cf. [13, 14, 40].

3.2.1.5 Optimum grading structures

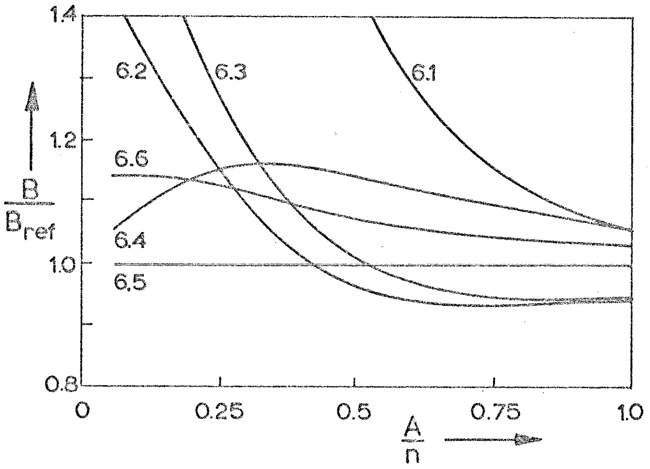

Fig. 7. Efficiency of wiring methods acc. to Fig. 6 with respect to loss probability B ($B_{ref} \hat{=}$ Fig. 6.5)

The above analysis method will be applied to various grading structures shown in Fig. 6 with $s_j = 1$ storage place in front of each grading group. The efficiency of the grading structures is shown by means of loss probability B versus the occupancy A/n ($A = \lambda/\varepsilon$ offered traffic).

As shown by Fig. 7, for small occupancies ($\frac{A}{n} < 0.4$) the straight inhomogeneous grading with progressive commoning and skipping, Fig. 6.5, is best, whereas for higher occupancies ($\frac{A}{n} > 0.4$) the straight homogeneous grading with skipping, Fig. 6.2, is best. Similar effects are already known from loss systems. For delay systems with small occupancies, the optimum grading for a loss system will be the best, too. For higher occupancies, calls queue up and the termination process of all servers determines more and 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. 6.2 and 6.3 shows that for sequential hunting slipping is worse than skipping. The optimum grading for a delay system, irrespective of the offered traffic, should therefore be a grading with certain progression and a considerable homogeneous part with skipping. Grading 6.5 forms a good compromise.

Similar results were also obtained for large gradings by simulations and approximate calculations [14], i.e., the results of exact calculations for small systems can be generalized and applied to optimum grading design.

3.2.2 Study on optimum routing strategies and storage capacities for data networks [40, 42]

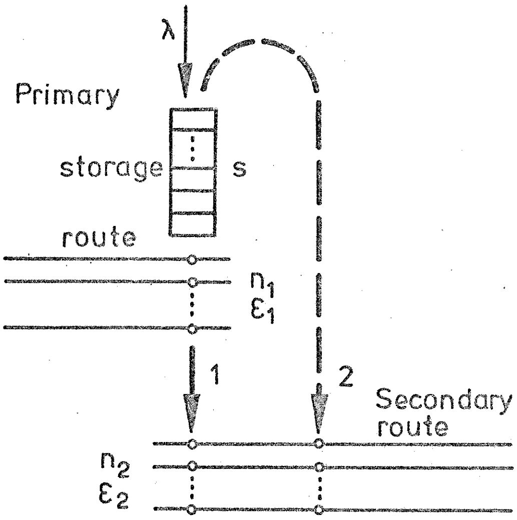

Fig. 8. Basic overflow system with two overflow strategies acc. to Section 2.2.3

3.2.2.1 Model

Fig. 8 shows a basic overflow system having n_1 servers in the primary route, n_2 servers in the secondary route, and a storage with limited capacity s. Two different overflow strategies are applied: (1) overflow from primary to secondary server group and common storage; (2) overflow from primary storage to secondary server group. The queue discipline may be arbitrary as far as no pdf of waiting time is considered. Calls are generated by a Poisson pdf with arrival rate λ, the service times are negative exponentially distributed, generally with different mean termination rates ε_i for servers of the i-th route, i = 1,2.

3.2.2.2 Analysis for overflow strategy 1

The analysis is carried out by means of state equations either based on microstates, as outlined in the preceeding example [13] or based on macrostates $(x_1,x_2;z)$, where x_i the number of busy servers of route i, i = 1,2, and z the number of waiting calls [42]. Clearly, $z = 0$ for $x_1 + x_2 < n_1 + n_2$, and $z \geq 0$ for $x_1 + x_2 = n_1 + n_2$. In the second method, the state equations are not solved directly, but indirectly by introduction of the generating function

$$F(x_2|x_1,t) = \sum_{x_2=0}^{n_2} p(x_1,x_2;0)(1+t)^{x_2} = \sum_{r=0}^{\infty} M_r(x_2|x_1)\frac{t^r}{r!}, \qquad (3)$$

where

$$M_r(x_2|x_1) = \sum_{x_2=0}^{n_2} r!\binom{x_2}{r}\, p(x_1,x_2;0) \qquad (4)$$

the conditional factorial moment of r-th order. The application of Eqs.(3) and (4) to the state equations results in a differential-difference equation for the generating function and corresponding difference equations for the factorial moments. The latter can be solved recursively yielding the state probabilities $p(n_1,n_2;z)$, $z \geq 0$, the carried traffics Y_1 and Y_2, the variance coefficient D_2, the probabilities of waiting and loss, W and B, the mean queue length Ω, the mean waiting time t_W, and, in case of the FIFO-queue discipline, the pdf of waiting times $W(>t)$, too.

3.2.2.3 Analysis for overflow strategy 2

The analysis can be carried out again by numerical solution of the corresponding equations of state. A general solution based on the famous idea of "substitute primary arrangements" describing the overflow traffic by its first and second moment has been reported in [40]. This solution yields all the characteristic values mentioned in 3.2.2.2.

3.2.2.4 Optimum routing strategy and storage capacity

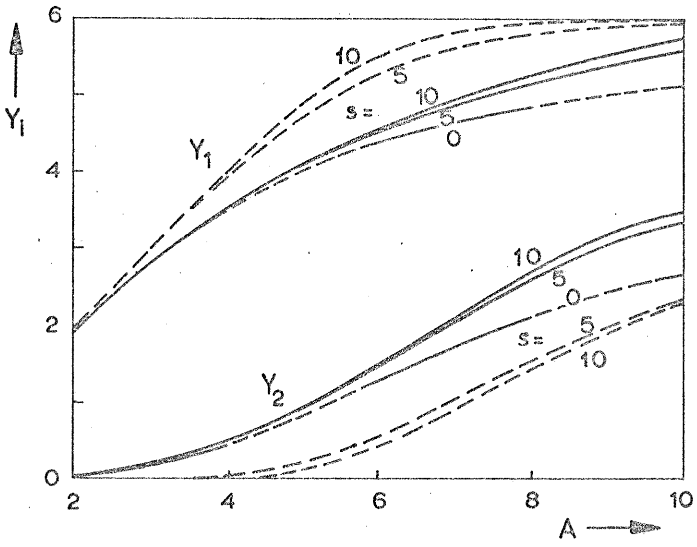

Fig. 9a. Carried traffics Y_1 and Y_2 versus offered traffic A
———— Overflow strategy 1
– – – – Overflow strategy 2

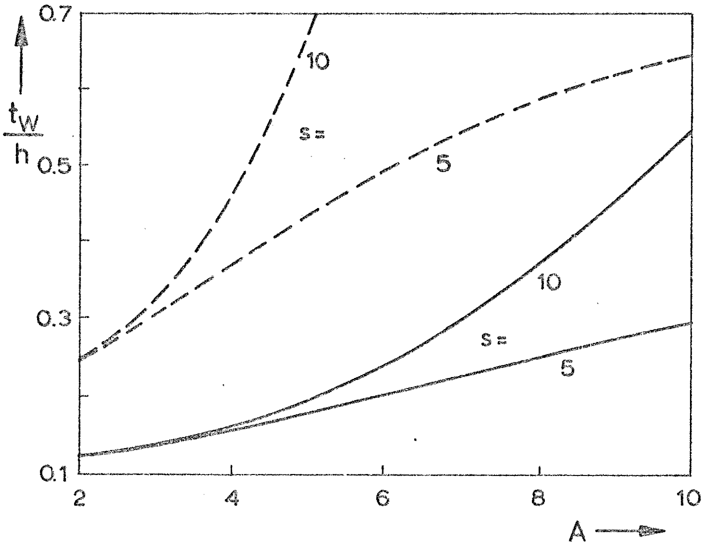

Fig. 9b. Mean waiting times t_W of waiting calls versus offered traffic A

As example, $n_1 = 6$ servers of the primary and $n_2 = 4$ servers of the secondary route are considered with $\varepsilon_1 = \varepsilon_2 = \frac{1}{h}$ for the storage capacities s = 0,5, and 10. In Figs. 9a,b the carried traffics of the primary and secondary route, Y_1 and Y_2, and the mean waiting times t_W of waiting calls are shown versus the offered traffic $A = \lambda/\varepsilon$.

As shown by the results, overflow strategy 2 yields a much better utilization of the primary route than overflow strategy 1. This utilization is paid by an increase of the mean waiting times. The results are further dependent on the capacity of the storage s. Moreover, both strategies and the capacity of the storage influence the variance coefficient of the traffic on the secondary route significantly [42]; this effect has also to be taken into consideration when the secondary route carries various overflow traffics additionally to a direct traffic in more complicated systems.

An optimum lay-out for given traffic amount is found by minimizing the function for total costs considering the costs of primary route, secondary route, storage, and (if possible) waiting time, too (This procedure can be carried out similarly as for telephone networks with alternate routing, cf. [35-38]). The minimization procedure has to be applied to various routing strategies to find out that routing strategy yielding minimal total costs. For purposes of field engineering, the numerical results must be given by tables or curves for a sufficiently large number of parameter combinations.

Further system structures and strategies were also dealt with, cf. [40,41].

3.2.3 Study on optimum number of interruptions in a real-time computer system with background programs [20]

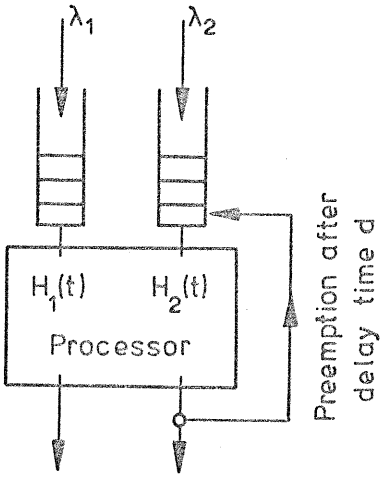

Fig. 10. Real-time processor under preemption-delay strategy

3.2.3.1 Model

In Fig. 10, a real-time processor is shown to which two types of programs are offered: (1) high-priority (foreground) real-time programs; (2) low-priority (background) programs for a reasonable utilization of the processor. The real-time programs require a fast response which is usually realized by interruption (preemption) of a background program in service. Preemption, however, needs certain overhead and reduces therefore the maximum throughput. To reduce the number of preemptions, a "preemption-delay" strategy has been suggested by which a real-time program interrupts a background program not before a certain delay time d [23,20]. For d = 0 and $d \longrightarrow \infty$ the usual cases of preemptive and non-preemptive priority are obtained, respectively. Interrupted programs continue their service at that point they had reached until preemption (preemption-resume). The queue disciplines may be FIFO.

Both real-time and background programs arrive according to Poisson pdf's with arrival rates λ_1 and λ_2, respectively. Real-time programs have a general service time T_{H1} with pdf $H_1(t)$ and mean $h_1 = 1/\varepsilon_1$; service times T_{H2} of background programs are negative exponentially distributed according to $H_2(t) = 1-\exp(-\varepsilon_2 t)$ with mean $h_2 = 1/\varepsilon_2$. For reasons of stationarity, the offered traffics $A_j = \lambda_j/\varepsilon_j$, j = 1,2, are bounded by $A = A_1 + A_2 < 1$.

3.2.3.2 Analysis

The analysis can be carried out by the method of imbedded Markov chain and has been reported for saturated background [23] as well as unsaturated background [20]. In the following, the more general results for unsaturated background will be outlined.

The preemption-delay strategy may be defined by the pdf of that delay time T_V a real-time program has to undergo by a background program in service:

$$V(t) = \begin{cases} 1-\exp(-\varepsilon_2 t) \; , & 0 \leqq t < d \\ 1 \; , & t \geqq d \end{cases} . \qquad (5)$$

The regeneration points of the imbedded Markov chain are those points immediately after service of a real-time program. For adequate description, a fictive service time is introduced for those real-time programs which meet at their arrival a background program in service and no waiting real-time program with pdf

$$F(t) = \left[1-P_V(d)\right] \cdot H_1(t) + P_V(d) \cdot V(t) * H_1(t) \; , \qquad (6)$$

where $P_V(d)$ the probability of delay for those real-time programs.

By the aid of Eqs.(5) and (6), the transition probabilities for the transition between the system states i (i = number of real-time programs in the system, i ≧ 0) can be determined easily. The application of the generating function

$$G(s) = \sum_{i=0}^{\infty} p(i)s^{i} \quad , \quad |s| \leqq 1 \, , \tag{7}$$

of the state probabilities p(i) on the Chapman-Kolmogorov-equations yields

$$G(s) = p(0)\Psi_1(\lambda_1-\lambda_1 s)\cdot\frac{P_V(d)s\phi(\lambda_1-\lambda_1 s) + [1-P_V(d)]\,s - 1}{s - \Psi_1(\lambda_1-\lambda_1 s)} \quad , \tag{8}$$

where $\Psi_1(s)$ and $\phi(s)$ the Laplace-Stieltjes transforms of $H_1(t)$ and $V(t)$, respectively, and $p(0) = (1-A_1)/[1+P_V(d)\lambda_1 E[T_V]]$.

From $G'(1) = \lambda_1 t_{R1}$, the mean rsponse time of real-time programs t_{R1} results

$$t_{R1} = h_1 + \frac{\lambda_1}{2}\cdot\frac{E[T_{H1}^2]}{1 - A_1} + P_V(d)\cdot\frac{E[T_V] + \frac{1}{2}\lambda_1 E[T_V^2]}{1 + P_V(d)\lambda_1 E[T_V]} \; . \tag{9}$$

A further analysis considering only mean values yields the mean response time of background programs t_{R2}

$$t_{R2} = h_2 + \frac{1}{1-A}\left[A_1 P_V(d)\frac{E[T_V] + \frac{\lambda_1}{2}E[T_V^2]}{1+P_V(d)\lambda_1 E[T_V]} + \frac{\lambda_1}{2}\cdot\frac{E[T_{H1}^2]}{1 - A_1} + \frac{\lambda_2}{2}E[T_{H2}^2] + \frac{A_1}{\varepsilon_2}\exp(-\varepsilon_2 d)\right]. \tag{10}$$

The delay probability $P_V(d)$ was found to be

$$P_V(d) = A_2/\left\{1-A_1 + \frac{\lambda_1}{\varepsilon_2}(1-A)[1 - \exp(-\varepsilon_2 d)]\right\} \; . \tag{11}$$

3.2.3.3 Optimum number of interruptions

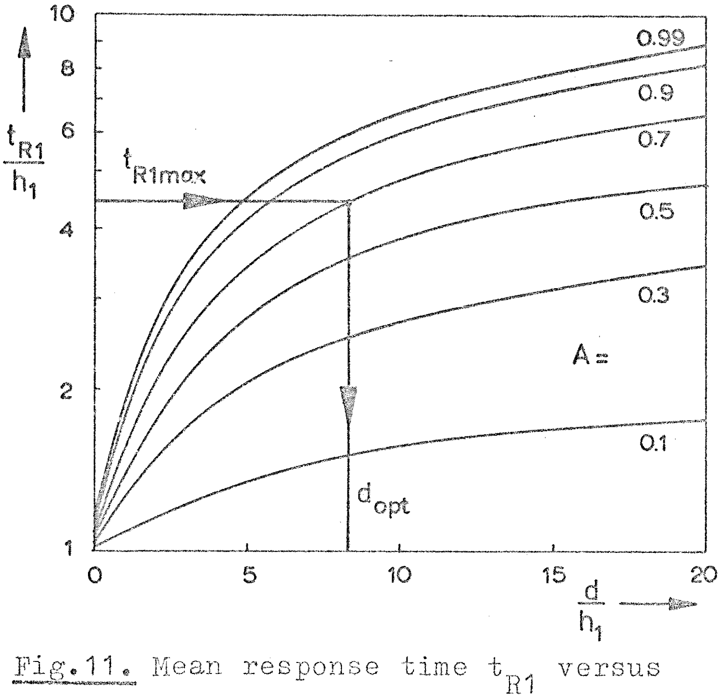

Fig.11. Mean response time t_{R1} versus delay time d, parameter A

Fig.11 shows an example with parameters $A_2=10A_1$, $\lambda_1=\lambda_2$, $\varepsilon_1=10\varepsilon_2$, $H_1(t)$ hyperexponential distributed with $E[T_{H1}^2] = 3/\varepsilon_1^2$.

As indicated in Fig. 11, for a prescribed upper value of the response time t_{R1max}, an optimum value d_{opt} can be chosen yielding a mean number of interruptions I referred to one background program

$$I = \frac{\lambda_1}{\varepsilon_2}\exp(-\varepsilon_2 d_{opt}) . \tag{12}$$

Compared with the usual preemptive priority (d = 0),

$$\Delta I = \frac{\lambda_1}{\varepsilon_2}[1-\exp(-\varepsilon_2 d_{opt})] \tag{13}$$

interruptions referred to one background program are saved.

Final remarks

The given examples were chosen such that in each case an individual analysis was necessary. This seems to be typical in many applications since the actual problems lead often to models not yet being investigated. Furthermore, the time available during a project phase is usually too short to carry out extensive analytical investigations so that simulation methods have to be applied

In some application cases, however, the queuing problem results in a standard queuing model or can be simplified such yielding a standard queuing model as, e.g., the M/M/n-, M/D/n-, and M/G/1- infinite-source queue, M/M/n-finite-source queue, or the M/M/n- infinite-source queue with limited accessibility. The characteristics of such standard queuing models were given in tables and charts for a large number of parameter combinations, cf. [51-60]. From the viewpoint of applications, however, these means have to be completed with respect to more complicated system structures, operational strategies, input and service characteristics, as well.

The optimization criteria can be quite different. For the usual cases, the problem is to dimension system components as, e.g., numbers of crosspoints, trunks, registers, etc., with respect to congestion as well as economic considerations. This problem is inherently related to management sciences, cf., e.g., Moe's principle [1,51], or methods applied to telephone networks with alternate routing[37,38]. Besides these quantitative aspects, this paper intends to draw the attention also to more qualitative aspects of optimization as studies on optimum system structures and operating strategies.

4. CONCLUSION

The paper gives a short review of typical queuing problems in communications and computer systems emphasizing those questions which are most important from the viewpoint of applications. Three different examples have been reported which refer to various types of optimization criteria characterized by quantitative aspects (optimum numbers of servers or storage places) as well as qualitative aspects (optimum structures and operating strategies). In connection with these examples, some typical analysis methods of queuing theory have been outlined to demonstrate their application to practical queuing problems and their impact on the optimization of communications and computer systems if the results can be evaluated suitably. Vice versa, the tasks of actual system design cause a number of new questions to queuing theory to be answered.

REFERENCES

[1] Syski, R.: Introduction to congestion theory in telephone systems. Oliver and Boyd, Edinburgh and London, 1960.

[2] Syski, R.: Markovian queues. Symp. on Congestion Theory. The Univ. of North Carolina Press, Chapel Hill, 1965, 170-227.

[3] Feller, W.: An introduction to probability theory and its applications. John Wiley and Sons, Inc., New York/London/Sidney, 1957.

[4] Khintchine, A.Y.: Mathematical methods in the theory of queuing. Griffin, London, 1960.

[5] Saaty, T.L.: Elements of queuing theory. McGraw-Hill Book Company, Inc., New York/Toronto/London, 1961.

[6] Riordan, J.: Stochastic service systems. John Wiley and Sons, Inc., New York and London, 1962.

[7] Takács, L.: Introduction to the theory of queues. Oxford University Press, New York, 1962.

[8] Cohen, J.W.: The single server queue. North-Holland Publish. Comp., Amsterdam and London, 1969.

[9] Kendall, D.G.: Stochastic processes occuring in the theory of queues and their analysis by the method of the imbedded Markov chain. Ann. Math. Statist., Vol. 24 (1953), 338-354.

[10] Pollaczek, F.: Concerning an analytical method for the treatment of queuing problems. Symp. on Congestion Theory. The Univ. of North Carolina Press, Chapel Hill, 1965, 1-42.

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

[12] Hieber, L.: The calculation of the probability of delay and the mean waiting time of link systems with unlimited queuing. 11th Report on Studies in Congestion Theory. Inst. f. Switching and Data Technics, Univ. of Stuttgart, 1970.

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

[14] Kühn, P.: Waiting time distributions in multi-queue delay systems with gradings. 7th ITC, Stockholm, 1973.

[15] Wagner, W.: On a combined delay and loss system with priorities. 6th Report on Studies in Congestion Theory. Inst. f. Switching and Data Technics, Univ. of Stuttgart, 1968.

[16] Brandt, G.J.: The pre-emptive delay and loss system. 14th Report on Studies in Congestion Theory. Inst. f. Switching and Data Technics, Univ. of Stuttgart, 1971.

[17] Segal, M.: A preemptive priority model with two classes of customers. ACM/IEEE Second Symp. on Problems in the Optimization of Data Comm. Systems. Palo Alto, 1971, 168-174.

[18] Coffman, E.G. and Kleinrock, L.: Some feedback queuing models for time-shared systems. 5th ITC, New York 1967, Prebook 288-304.

[19] Adiri, I.: A note on some mathematical models of time-sharing systems. J.ACM, Vol. 18 (1971), 611-615.

[20] Herzog, U., Kühn, P. and Zeh, A.: Classification and analysis of traffic models for the performance activities in computer systems. Nachrichtentechn. Fachber., Vol. 44 (1972), 181-198.

[21] Gaver, D.P.: On priority type disciplines in queuing. Symp. on Congestion Theory. The Univ. of North Carolina Press, Chapel Hill, 1965, 228-252.

[22] Marte, G.: Optimal time scheduling for time-shared computer systems with piecewise exponential computing time distribution. Proc. Intern. Comp. Symp. (ACM), Bonn, 1970, 184-206.

[23] Coffman, E.G.: On the tradeoff between response and preemption costs in a foreground-background computer service discipline. IEEE Transact. Comp., Vol. C-18 (1969), 942-947.

[24] Herzog, U.: Preemption-distance priorities in real-time computer systems. NTZ, Vol. 25 (1972), 201-203.

[25] Langenbach-Belz, M.: Sampled queuing systems. Symp. on Comp. Comm. Networks and Teletraffic. Polytechnic Press of the Polytechnic Institute of Brooklyn, New York, 1972.

[26] Jackson, R.R.P.: Random queuing processes with phase-type service. J.R.S.S. Ser.B., Vol. 18 (1956), 129-132.

[27] Hunt, G.C.: Sequential arrays of waiting lines. J. Op. Res., Vol. 4 (1956), 674-683.

[28] Reich, E.: Departure processes. Symp. on Congestion Theory. The Univ. of North Carolina Press, Chapel Hill, 1965, 439-457.

[29] Avi-Itzhak, B. and Yadin, M.: A sequence of two servers with no intermediate queue. Management Science, Vol. 11 (1965), 553-564.

[30] Burke, P.J.: The output process of a stationary M/M/s queuing system. Ann. Math. Stat., Vol. 39 (1968), 1144-1152.

[31] Neuts, M.F.: Two queues in series with a finite, intermediate waitingroom. J. Appl. Prob., Vol. 5 (1968), 123-142.

[32] Krämer, W.: Total waiting time distribution function and the fate of a customer in a system with two queues in series. 7th ITC, Stockholm, 1973.

[33] Langenbach-Belz, M.: Two-stage queuing model with sampled parallel input queues. 7th ITC, Stockholm, 1973.

[34] Wilkinson, R.I.: Theories for toll traffic engineering in the USA. BSTJ, Vol. 35 (1956), 421-514.

[35] Lotze, A.: Problems of traffic theory in the design of international direct distance dialling networks. NTZ-Comm. J., Vol. 7 (1968), 41-46.

[36] Herzog, U. and Lotze, A.: The RDA-method, a method regarding the variance coefficient for limited access trunk groups. NTZ-Comm. J., Vol. 7 (1968), 47-52.

[37] Lotze, A. and Schehrer, R.: The design of alternate routing systems with regard to the variance coefficient. NTZ-Comm. J., Vol. 7 (1968), 52-56.

[38] Lotze, A.: Field engineering methods for economic network planning with or without alternate routing. TIMS XX, XX International Meeting. The Institute of Management Sciences, Tel Aviv, 1973.

[39] Boehm, B.W. and Mobley, R.L.: Adaptive routing techniques for distributed communications systems. IEEE Transact. Comm. Techn., Vol. CT-17 (1969), 340-349.

[40] Herzog, U. and Kühn, P.: Comparison of some multi-queue models with overflow and load-sharing strategies for data transmission and computer systems. Symp. on Comp. Comm. Networks and Teletraffic. Polytechnic Press of the Polytechnic Institute of Brooklyn, New York, 1972.

[41] Herzog, U.: Message-switching networks with alternate routing. 7th ITC, Stockholm, 1973.

[42] Krämer, W., Kühn, P. and Wörn, H.-W.: On the calculation of overflow systems with fully accessible primary and secondary trunk groups and common storage. Inst. f. Switching and Data Technics, Univ. of Stuttgart, Monograph No. 390, 1973.

[43] Gaver, D.P.: Probability models for multiprogrammed computer systems. J. ACM, Vol. 14 (1967), 423-438.

[44] Chen, Y.C. and Shedler, G.S.: A cyclic queue network model for demand paging computer systems. IBM Res. Rep. RC-2398, 1969.

[45] Shedler, G.S.: A cyclic queue model of a paging machine. IBM Res. Rep. RC-2814, 1970.

[46] Spies, P.P.: A queuing model analysis of the multiplexed use of a central processor unit and an I/O-channel. Proc. Intern. Comp. Symp. (ACM), Bonn, 1970, 282-299.

[47] Avi-Itzhak, B. and Heyman, D.P.: Approximate queuing models for multiprogramming computer systems. Techn. Mem. Bell Teleph. Lab., MM-7k-1713-15, 1971.

[48] Buzen, J.P.: Queuing network models of multiprogramming. Ph.D. Thesis, Harvard University, 1971.

[49] Adiri, I., Hofri, M. and Yadin, M.: A multiprogramming queue. IBM Res. Rep. RC-3566, 1971.

[50] Böttinger, R., Herzog, U., Krämer, W. and Kühn, P.: Classification and simulation of traffic models for multiprogrammed computer systems. Inst. f. Switching and Data Technics, Univ. of Stuttgart, Monograph No. 385, 1972.

Tables and charts:

[51] Jensen, A.: Moe's principle. Copenhagen Telephone Co., Copenhagen, 1950.

[52] Peck, L.G. and Hazelwood, R.N.: Finite queuing tables. John Wiley and Sons, Inc., New York, 1958.

[53] Zimmermann, G.O. and Störmer, H.: Wartezeiten in Nachrichtenvermittlungen mit Speichern. R. Oldenbourg Verlag, München, 1961.

[54] Descloux, A.: Delay tables for finite- and infinite-source systems. McGraw-Hill Book Company, Inc., New York/Toronto/London, 1962.

[55] — :Dimensioning data for planning of communication systems. Telefonbau und Normalzeit, Frankfurt/Main, 1966.

[56] — :Projektierungsunterlagen für Vermittlungssysteme. Standard Elektrik Lorenz AG, Stuttgart, 1966.

[57] Thierer, M.: Delay-tables for limited and full availability according to the Interconnection Delay Formula (IDF).Stuttgart, 1968, cf. [11].

[58] — :Telephone traffic theory. Tables and charts, Part 1. Siemens AG, Berlin/München, 1970.

[59] Rouault, J.M.: Teletraffic. Éditions Eyrolles, Paris, 1970.

[60] Everling, W.: Exercices in computer systems analysis. Springer-Verlag, Berlin/Heidelberg/New York, 1972.