OVERFLOW- AND LOAD-SHARING STRATEGIES

Comparison of Some Multi-Queue Models with Overflow and Load-Sharing Strategies for Data Transmission and Computer Systems *

ULRICH HERZOG and PAUL KUHN
Institute for Switching and Data Techniques
University of Stuttgart
Stuttgart, FRG

ABSTRACT

This paper deals with different routing strategies for data transmission as well as different load-sharing and reliability configurations for computer systems.

Starting from two separate service systems without mutual overflow, a review of different overflow systems is given considering three types of overflow strategies applicable to data transmission. Furthermore, systems of two and more computers are inspected with various configurations and operating strategies. For reasons of simplicity, all multi-queue models are demonstrated by example of systems with two (limited) queues.

A comparison between the models of both application areas shows a close similarity with respect to system structure and operating strategies. Therefore, the models of both areas can be treated by the same mathematical methods.

The analysis of the different systems with respect to the service quality is carried out on the basis of the state equations under Markovian assumptions. Finally, the most important models will be compared with each other with respect to various traffic criteria.

1. INTRODUCTION

During the last years, many investigations have been published about single-queue models for data transmission and computer systems under various operating strategies and traffic properties.

In modern communication networks, special routing strategies are used, such as alternative or adaptive routing [1-3]. Real-time computers are duplicated or operate in a load-sharing mode [4]. For such systems single-queue models are often not applicable in order to investigate their traffic behaviour.

Data communication networks with different routing strategies as well as different reliability configurations are rather described by multi-queue models.

In the second chapter, various configurations of many-server systems with two queues are discussed under three modes of overflow strategies:

1. overflow from primary to a secondary server group
2. overflow from a storage in front of a primary server group
to a secondary server group
3. overflow from primary storage to a secondary storage directly to a secondary server group.

All three overflow strategies can be applied to data transmission networks as well as to load-shared or breakdown-reliable computer systems.

The third chapter gives an outline of the mathematical analysis of such overflow systems to investigate the capability of the different configurations and strategies. For the analysis the Markovian assumptions are assumed. Solutions are given either by exact evaluation of systems of equations or by approaches using two moments of the overflow traffic.

In the final chapter, comparisons are made for different system configurations and strategies. Numerical results are presented for probabilities of waiting and loss, mean waiting times, carried traffic (throughput) as well as the distribution of waiting times.

List of Symbols

- $g$: number of incoming groups or queues
- $n$: total number of servers (trunks, lines, computers)
- $n_r$: total number of servers of route $r$
- $k_j$: accessibility within incoming group $j$, $j = 1(1)g$
- $e_{n_j}^r$: accessibility matrix
- $e_{r_j}$: number of busy servers with respect to route $r$
- $x_r$: number of occupied waiting places within queue $j$
- $z_j$: total number of calls (units, demands, requests, jobs) in the system with respect to route $r$
- $w_j$: interarrival time for calls of incoming group $j$
- $b_r$: service (holding) time with respect to route $r$
- $j$: waiting time with respect to queue $j$
- $\lambda_j$: arrival rate for calls of incoming group $j$
- $\sigma_r$: service (termination) rate of an occupied server of route $r$
- $e_{r_j}$: service rate of queue $j$ with respect to route $r$
- $[Fr_j]$: interqueue discipline matrix
- $p_{g_j}^r$: stationary probability of state $g$
- $W_j$: probability of waiting for calls of incoming group $j$
- $B_j$: probability of loss for calls of incoming group $j$
- $k_j$: offered traffic to incoming group $j$
- $y_r$: carried traffic within route $r$
- $R_j$: overflow traffic of group $j$
- $V_j$: variance of overflow traffic of group $j$
- $D_j$: variance coefficient of overflow traffic of group $j$
- $\Omega_j$: mean queue length of queue $j$
- $t_{w_j}$: mean waiting time of calls waiting in queue $j$, referred to all waiting calls within this queue
- $P^j_u (t | g_j)$: complementary conditional distribution function (d.f.) of waiting time for calls waiting in queue $j$ under condition, that state $g_j$ had been met on arrival
- $P^j_0 (t)$: total d.f. of waiting time for calls waiting in queue $j$
- $m_k(u_{j\mu_2})$: conditional ordinary moment of $k$-th order
- $m_k$: ordinary moment of $k$-th order
- $M_k(u_{j\mu_2})$: conditional factorial moment of $k$-th order
- $M_k$: factorial moment of $k$-th order
2. MULTI-QUEUE MODELS FOR DATA TRANSMISSION AND COMPUTER SYSTEMS

In Fig. 1 and Fig. 2 some configurations of many-server systems with two queues are reviewed which are able to operate under various overflow and load-sharing strategies. Models of Fig. 1 are suited for data transmission systems, models of Fig. 2 are applicable to systems of computers.

2.1 Multi-queue server configurations

The model Fig. 1.1 shows two (separate) queuing systems for the traffic to direct routes 1 and 2 with \( n_j \) servers (trunks, lines) for route \( r_j, r=1,2 \). A call of an incoming group \( j \) is allowed to wait in a storage with \( s_j \) waiting places when the servers of the corresponding direct route are blocked and if there is at least one waiting place available, \( j = 1,2 \). In Fig. 1.2 and 1.3 two different configurations of many-server systems are shown with two queues and "fully accessible" servers. Assuming a sequential hunting mode, route no.1 of Fig. 1.2 represents a high usage route which carries most of the traffic while route no.2 takes only the "overflow" traffic. In Fig. 1.3 both routes no.1 and no.2 are working as high usage routes for their own offered traffic; in addition, the overflow traffic of the other route is offered to servers of each route. In model Fig. 1.4 a part of each high usage route is reserved for the own offered traffic, the residual servers are allowed to carry also overflow traffic. Finally, model Fig. 1.5 shows a configuration where the traffic for two direct routes can overflow to a third ("final") route no.3. Both models Fig. 1.4 and 1.5 incorporate "limited accessible" servers.

The configurations of Fig. 2.1 to 2.5 are very similar to the models of Fig. 1. The servers may be considered as computers serving calls (real-time requests or batch jobs) stored within the queues, respectively. All the different configurations with mutual aid allow better utilization, load-sharing and reliability with respect to breakdown.

2.2 Overflow strategies

The multi-queue server configurations discussed in section 2.1 (Fig.1 and 2) may operate under different overflow strategies. Three main modes will be considered:

- **Overflow strategy 1 (S1):**
  
  Overflow from primary (direct) to a secondary (final) server group.
  
  Waiting is only allowed if all accessible servers of the primary and secondary server group are occupied.

- **Overflow strategy 2 (S2):**
  
  Overflow from a storage in front of a primary server group to a secondary server group.
  
  The accessible servers of the secondary server group are only hunted if both the primary server group and the primary storage are fully occupied.

- **Overflow strategy 3 (S3):**
  
  Overflow from primary storage to a secondary storage or directly to a secondary server group.
  
  The call, which finds the own server group and all accessible servers of other server groups occupied, queues in the storage of its incoming group. However, if there is no free waiting place, the call is diverted to the other incoming group and is treated there like an original call of this group.

These different overflow strategies have a significant influence on the traffic criteria, in particular the probabilities of loss and waiting, as well as the mean waiting times. In the following chapters, the analysis of these queuing systems is outlined and the most important configurations are compared with each other with respect to the overflow strategies in question.
3. MATHEMATICAL ANALYSIS

This chapter gives an outline of the mathematical analysis of multi-queue models under different overflow strategies. In section 3.1 the exact calculation of multi-queue models is shown for overflow strategy 1. Sections 3.2 and 3.3 handle multi-queue models for overflow strategies 2 and 3 based on a method regarding two moments of the overflow traffic, respectively. For the mathematical analysis, the Markovian properties are assumed throughout the paper.

3.1 Analysis of multi-queue models for overflow strategy 1

3.1.1 Model

The general multi-queue model with overflow from primary servers (e.g. trunks of high usage routes, highly used computers etc.) to secondary servers (e.g. trunks of final routes, remote computers etc.) can generally be considered as a service system with full or limited accessibility. The special assignment of servers to incoming groups and outgoing routes is determined by a "grading matrix". In the most general case, each server represents a different route; by combination of various servers to outgoing groups all possible configurations can be obtained from the general case.

The structure of the multi-queue overflow model is laid down by the following parameters:

\[ s_{nj} \]
number of incoming groups or input queues

\[ n \]
number of servers

\[ k_j \]
accessibility within incoming group \( j \), \( j=1(1)g \)

\[ a_{rij} \]
accessibility matrix, where \( a_{rij} = 1(0) \) if incoming group \( j \) has (has no) access to server number \( r \), \( r=1(1)n, j=1(1)g \).

\[ b_{rj} \]
interqueue discipline matrix, where \( b_{rj} \) the probability that queue \( j \) will be served when server \( r \) finishes service, \( r=1(1)n, j=1(1)g \).

The operation mode is characterized by the following criteria:

- sequential hunting of all accessible servers within each incoming group (overflow strategy S1)
- first-in, first-out (FIFO)-service within each queue (queue discipline)
- arbitrary probability law for service between the queues (interqueue discipline).

The interqueue discipline is fully determined by

The interqueue discipline matrix is in general state-dependent. By choice of special matrix elements, however, special cases are obtained; for instance, the case of nonpre-emptive priorities between the queues.

Input and termination processes are Markovian, i.e. the inter-arrival times \( a \) and service times \( b \) are exponential:

\[ A_j(t) = \text{Prob}\{a_j \leq t\} = 1 - \exp(-\lambda_j t), \quad j=1(1)g \] \hspace{2cm} (1.1)

\[ B_{rj}(t) = \text{Prob}\{b_{rj} \leq t\} = 1 - \exp(-\varphi_{rj} t), \quad r=1(1)n \] \hspace{2cm} (1.2)

For illustration, Fig. 3 shows a simple 3-server system with two queues.
3.1.2 Principle of solution

At first, the occupations of servers and waiting places will be described by a multi-dimensional state. For the probabilities of state, the linear system of equations (Kolmogorov-forward-equation) is derived in section 3.1.3. By means of the state probabilities, characteristic traffic values (mean values) are defined in section 3.1.4. For the treatment of waiting time distributions, a waiting process is considered which is constructed from the process of system states. In section 3.1.5 the linear system of differential equations (Kolmogorov-backward-equation) for the conditional distribution functions (d.f.) of waiting time is discussed. The total d.f. of waiting time is found by regarding the probabilities of initial states (where the waiting process starts) combined with the corresponding conditional d.f. of waiting time, c.f. section 3.1.6.

3.1.3 Stationary probabilities of state

The system state \( \xi \) may be defined as

\[
\xi = (x_1, \ldots, x_p, \ldots, x_n; z_1, \ldots, z_j, \ldots, z_g), \quad \xi \in \Xi, \quad (1.3)
\]

where

\[
x_r = \begin{cases} 
0 & \text{if server } r \text{ is idle} \\
1 & \text{if server } r \text{ is occupied, } r = 1(1)n,
\end{cases}
\]

\[
z_j = \text{number of occupied waiting places within queue } j, \quad z_j \in [0, a_j], \quad j = 1(1)g.
\]

In Eq. (1.3) not all possible patterns are physically realizable; a queue \( j \) can only be built up if all accessible \( k_j \) servers are blocked, \( j = 1(1)g \).

The stationary probabilities of states \( p(\xi) \) can be determined by the Kolmogorov-forward-equations

\[
q_\xi \cdot p(\xi) - \sum_{\kappa \neq \xi} q_{\xi \kappa} \cdot p(\kappa) = 0, \quad \xi \in \Xi, \quad (1.4)
\]

which are found by considering all states \( \xi \) in statistical equilibrium with their neighbour states \( [5, 6] \). In Eq. (1.4) \( q_{\xi \kappa} \) means the transition coefficient for transition from state \( \xi \) to state \( \kappa \), and

\[
q_\xi = \sum_{\kappa \neq \xi} q_{\xi \kappa}.
\]

Application of the statistical equilibrium to the general state Eq. (1.3) results in the following equation:

\[
\sum_{\xi_1=1}^{2^g} (\xi - \xi' \xi'_{j_1}, \ldots, \xi'_{j_m}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
= \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot (\sum_{\xi_{j_1}=1}^{2^g} \xi_{j_1} \xi'_{j_1}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot (\sum_{\xi_{j_1}=1}^{2^g} \xi_{j_1} \xi'_{j_1}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot (\sum_{\xi_{j_1}=1}^{2^g} \xi_{j_1} \xi'_{j_1}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

\[
+ \sum_{\xi_1=1}^{2^g} (\xi - x_1 \xi'_{j_1}, \ldots, x_r \xi'_{j_m}) \cdot (\sum_{\xi_{j_1}=1}^{2^g} \xi_{j_1} \xi'_{j_1}) \cdot p(\ldots, x_r, \ldots, z_j, \ldots)
\]

where \( \delta_{i,j} \) the Kronecker symbol, and \( \kappa_j = \sum_{i=1}^{n} a_{j,i} \cdot x_i \) the number of occupied servers in group \( j \). In Eq. (1.5) all probabilities of physically not possible states are zero.

To illustrate Eq. (1.5), in Fig. 4 the 5-dimensional state space with transition coefficients is shown for the 3-server system Fig. 3. In Fig. 4 \( \xi_{j} = \xi_{j} \cdot p_{j} \) is used as abbreviation for the service rate of queue \( j \) with respect to server \( r \).
The linear system of equations (1.5) is generally solved by an iterative method (successive overrelaxation) and normalized by the condition, that the sum of all state probabilities equals unity. In some special cases of fully accessible servers, an explicit solution or recursion algorithms can be derived [7].

3.1.4 Characteristic traffic values

The most important mean values are the probabilities of waiting and loss, the carried traffic, the mean queue lengths and mean waiting times. All these values can be derived from the probabilities of state.

a) Probability of waiting $W_j$

$$W_j = \sum_{x_0} \sum_{x_1=0}^{\lambda_1} \sum_{x_2=0}^{\lambda_2} ... \sum_{x_f=0}^{\lambda_f} \left( \prod_{i=1}^f (x_i) \right) \cdot p(y).$$  \hspace{1cm} (1.6)

b) Probability of loss $B_j$

$$B_j = \sum_{x_0} \sum_{x_1=0}^{\lambda_1} \sum_{x_2=0}^{\lambda_2} ... \sum_{x_f=0}^{\lambda_f} \sum_{x_f+1}^{\lambda_f} \left( \prod_{i=1}^f (x_i) \right) \cdot p(y).$$  \hspace{1cm} (1.7)

c) Carried traffic $Y_r$

$$Y_r = \sum_{x_0} \sum_{x_1=0}^{\lambda_1} \sum_{x_2=0}^{\lambda_2} ... \sum_{x_f=0}^{\lambda_f} x_r \cdot p(y).$$  \hspace{1cm} (1.8)

d) Mean queue length $\Omega_j$

$$\Omega_j = \sum_{x_0} \sum_{x_1=0}^{\lambda_1} \sum_{x_2=0}^{\lambda_2} ... \sum_{x_f=0}^{\lambda_f} x_j \cdot p(y).$$  \hspace{1cm} (1.9)

e) Mean waiting time $t_{wj}$ referred to all waiting $j$-calls

$$t_{wj} = \frac{\Omega_j}{\lambda_j W_j}.$$  \hspace{1cm} (1.10)

3.1.5 Conditional distribution functions of waiting time

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

To describe the waiting process of $j$-calls, a waiting state $\xi_j$ is introduced. The waiting state $\xi_j$ is built from the occupation states $x_r$ of all those servers, which have no access to group $j$, and the states $i_y$ of all queues. When the interqueue discipline does not depend on the actual length of the queues, the waiting time of the $j$-test call is not influenced by subsequent arriving $j$-calls (FIPO). In such cases, the waiting state $\xi_j$ can be defined by a $(n-k_j+1)$-dimensional variable:

$$\xi_j = (..., x_{r1}, ..., x_{ry}, ...) \Leftrightarrow \xi_j = (i_1, ..., i_y),$$  \hspace{1cm} (1.11)

where

$$x_r = \begin{cases} 0 & \text{if server } r \text{ is idle} \\ 1 & \text{if server } r \text{ is occupied} \end{cases}$$

$$i_r = \text{number of waiting calls within queue } r, \; \nu = j,$$

$$i_j = \text{number of calls waiting in front of the } j\text{-test call}.$$  

For the waiting time of a $j$-test call, which met an arbitrary state $\xi_j$ at arrival, a conditional (complementary) d.f. is defined:

$$P^C_j(t | \xi_j) = \text{Prob} \{ w_j > t | \xi_j \}, \quad \xi_j \in \mathbb{Z}_j.$$  \hspace{1cm} (1.12)
For the conditional d.f. of waiting time a linear system of differential equations (Kolmogorov-backward-equation) holds [7,8,9]:

\[
\frac{dP^s_j(t|\xi_j)}{dt} = -q^s_j \cdot P^s_j(t|\xi_j) \cdot \sum_{\xi_j \rightarrow \eta_j} q^s_{j \rightarrow \eta_j} \cdot P^s_j(t|\eta_j), \quad \xi_j, \eta_j \in Z^s_j (1.13)
\]

with initial conditions \(P^s_j(0|\xi_j) = 1, \xi_j \in Z^s_j\). In Eq. (1.13) \(q^s_{j \rightarrow \eta_j}\) means the transition coefficient for transition of waiting state \(\xi_j\) to waiting state \(\eta_j\); \(q^s_{j \rightarrow \eta_j}\) is the coefficient for leaving the state \(\xi_j\) inclusing

To illustrate Eq. (1.13), the state space of the waiting process is given in Fig. 5 for the 3-server system of Fig. 3 with respect to 1-calls.

For the example of Fig. 5, the differential equation for the initial state \(\xi_j^* = (x^*_2,i_1,i_2), \quad i_1 > 0, \quad 0 \leq i_2 \leq s_2\), reads as follows:

\[
\frac{dP^s_j(t|\xi_j^*)}{dt} = -(2 \cdot \xi_j^* \cdot \lambda^s_j \cdot P^s_j(t|\xi_j^*) + \lambda^s_j \cdot P^s_j(t|\xi_j^*) + \lambda^s_j \cdot P^s_j(t|\xi_j^*) + \lambda^s_j \cdot P^s_j(t|\xi_j^*))
\]

At the bottom of Fig. 5, the coefficients for termination ("death") of the waiting process are also shown.

A detailed discussion of waiting processes for various queue and interqueue disciplines has been reported by Kühn [7]. The linear systems of differential equations can be suitable solved by a method of successive power series expansions even for a high order of the differential equation system and prescribed accuracy [7].

By integration of Eq. (1.13), the corresponding linear equations for the conditional mean waiting times and higher moments can be obtained as well [7,8].

3.1.6 Total distribution function of waiting time

The total d.f. of waiting time can be obtained by averaging over all conditional d.f. combined with the probabilities of initial states (condition):

\[
P^s_j(t) = \int \int \int \int P^s_j(t|\xi_j) \cdot p(\xi_j) \cdot d\xi_j
\]

The probabilities of initial states \(p(\xi_j)\) are identical with the corresponding probabilities of state \(p(\xi_j)\); the difference between the states \(\xi_j\) and \(\xi_j^*\) originates from the \(k_j\) (occupied) servers accessible from incoming group \(j\).

3.2 Analysis of multi-queue models for overflow strategy 2

3.2.1 Model

Be given a structure as shown in Fig. 6: at first, the traffic \(A_j (i = 1,2)\) is offered to a primary arrangement consisting of \(n_j\) servers and \(s_j\) waiting places. If there is blocking, i.e. all primary servers and all waiting places are occupied, calls will be diverted to a secondary group (overflow group) with \(n_j\) servers. Calls waiting for service in a queue are served in the order of their arrival (FIFO). The service rates in the different server groups may be different.

3.2.2 Principle of solution

In principle, an exact solution for the characteristic traffic values is possible; introducing three-dimensional state probabilities, the equations of state can be found relatively easy and the evaluation may be done by a relaxation method [10]. However, for arrangements with a large number of trunks and waiting places as well as for structures with more than two primary arrangements, this evaluation is not possible, even on the largest digital computers. Therefore, a handy approximate method is suggested, using the fundamental idea of so-called "substitute primary arrangements" [11,12,13,14,15]:

a) All overflow traffics are characterized by their first and second moment (mean value \(R\) and variance \(V\) or variance coefficient \(D = V - R\), respectively). Moments of third and higher order are neglected. The exact calculation of overflow traffic moments is outlined later on in section 3.2.3.
b) Because all traffic overflowings from different primary arrangements are independent of each other, the total overflow traffic, offered to the common secondary group, can be prescribed by the sum of all mean values \( R_j \) and the sum of all variance coefficients \( D_j \). Therefore, one gets in case of two primary arrangements as shown in Fig. 6:

\[
\begin{align*}
\bar{R} &= R_1 + R_2 \\
\bar{D} &= D_1 + D_2
\end{align*}
\]  

(2.1)

c) In order to calculate the traffic characteristics of the secondary group, a "substitute primary arrangement" and a "generating traffic" \( A^* \) are determined such that an overflow traffic is generated with mean value \( \bar{R} \) and variance coefficient \( \bar{D} \) (c.f. Fig. 7). In other words: all actual traffic overflowings from the various primary arrangements are prescribed approximately by a substitute overflow traffic with the same mean \( \bar{R} \) and variance \( \bar{D} \). Then all characteristic values of the actual secondary group are calculated, taking into account the structure of the substitute primary arrangement as well as the generating traffic \( A^* \).

Artificial traffic trials performed on a digital computer have shown that for practical applications it is unimportant whether the substitute arrangement is a full available group with or without waiting places. Therefore, by reason of simple evaluations, a full available group with \( n^* \) servers, offered traffic \( A^* \), and term.ination rate \( \varepsilon^* = \varepsilon_3 \) is chosen as a substitute primary arrangement (\( n^* \) and \( A^* \) have to be determined by iteration such that the overflow traffic \( \bar{R}, \bar{D} \) is generated).

Now, obviously, the traffic \( (\bar{R}_3, \bar{D}_3) \) rejected also by the secondary group is given by \[11,12,13,15]:

\[
\begin{align*}
\bar{R}_3 &= A^* \cdot E_3 \cdot \mu_3 n_3 (A^*) \\
\bar{D}_3 &= \bar{R}_3^2 \cdot \left[ E_3 \cdot A^* + n_3 (A^*) \left( \frac{A^*}{n_3} + \frac{1}{\lambda_3} + 1 - A^* + \bar{R}_3 \right) - 1 \right],
\end{align*}
\]  

(2.2)

where

\[
E_3(A) = \sum_{i=0}^{n_3} \frac{A^i}{i!}
\]

3.2.3 Calculation of the moments of overflow traffic

Be given a full available primary arrangement with \( n_j \) servers and \( s_j \) waiting places \((j = 1, 2)\) and an infinite secondary arrangement \((n_j \rightarrow \infty)\). If there is offered random traffic (Poisson input and negative exponential service times, c.f. Eqs. (1.1, 1.2)) to the primary arrangement, the mean value \( R_j \) of the traffic overflowing to the secondary group with termination rate \( \varepsilon_j \) is given by the well-known formula [16]:

\[
R_j = \frac{\lambda_j}{\varepsilon_j} \cdot \frac{A_j}{\bar{R}_j} \cdot \left[ \frac{\lambda_j^m A^m_j}{\bar{R}_j^{m+1}} \right] = \frac{\lambda_j}{\varepsilon_j} \cdot \frac{A_j}{\bar{R}_j} \cdot \left[ \frac{\lambda_j^m A^m_j}{\bar{R}_j^{m+1}} \right]
\]

where

\[
A_j = \frac{\lambda_j}{\varepsilon_j}
\]

Basharin [17] seems to be the first dealing with the variance of overflow traffic in delay-loss systems: recurrent formulae are presented for the computation of the moments of overflow traffic if there are \( s_j \geq 1 \) waiting places and uniform termination rate \( \varepsilon_j = \frac{\bar{R}_j}{E_j} = \varepsilon \) for both primary and (infinite) secondary arrangement.

In the following it is concisely outlined how to calculate recursively all overflow traffic moments even if there are different termination rates \( \varepsilon_j \) and \( \varepsilon_j \). For uniform termination rates also an explicit solution is presented for the variance \( V_j \) or variance coefficient \( D_j \), respectively.

The calculation has been performed following a way of solution which is similar to that one successfully applied for overflow systems without waiting places [18,19]:

a) Defining with \[u_j, w_j\] the state that there are simultaneously \( u_j \) demands being served or waiting for service in the primary arrangement and \( w_j \) demands in the secondary system, it is possible to find the following equations of state for the state probabilities \( p(u_j, w_j)\):

\[
\begin{align*}
\mu_j < \mu_j :
\lambda_j = & \mu_j \cdot p(u_j, w_j) + \mu_j + \lambda_j \cdot p(u_j, w_j) + \\
= & \mu_j \cdot p(u_j + 1, w_j) + \mu_j \cdot p(u_j, w_j)
\end{align*}
\]  

(2.4a)

b) Because all traffics overflowing from different primary arrangements are independent of each other, the total overflow traffic, offered to the common secondary group, can be prescribed by the sum of all mean values \( R_j \) and the sum of all variance coefficients \( D_j \). Therefore, one gets in case of two primary arrangements as shown in Fig. 6:

\[
\begin{align*}
\bar{R} &= R_1 + R_2 \\
\bar{D} &= D_1 + D_2
\end{align*}
\]  

(2.1)

c) In order to calculate the traffic characteristics of the secondary group, a "substitute primary arrangement" and a "generating traffic" \( A^* \) are determined such that an overflow traffic is generated with mean value \( \bar{R} \) and variance coefficient \( \bar{D} \) (c.f. Fig. 7). In other words: all actual traffic overflowings from the various primary arrangements are prescribed approximately by a substitute overflow traffic with the same mean \( \bar{R} \) and variance \( \bar{D} \). Then all characteristic values of the actual secondary group are calculated, taking into account the structure of the substitute primary arrangement as well as the generating traffic \( A^* \).

Artificial traffic trials performed on a digital computer have shown that for practical applications it is unimportant whether the substitute arrangement is a full available group with or without waiting places. Therefore, by reason of simple evaluations, a full available group with \( n^* \) servers, offered traffic \( A^* \), and termination rate \( \varepsilon^* = \varepsilon_3 \) is chosen as substitue primary arrangement (\( n^* \) and \( A^* \) have to be determined by iteration such that the overflow traffic \( \bar{R}, \bar{D} \) is generated).

Now, obviously, the traffic \( (\bar{R}_3, \bar{D}_3) \) rejected also by the secondary group is given by \[11,12,13,15]:

\[
\begin{align*}
\bar{R}_3 &= A^* \cdot E_3 \cdot \mu_3 n_3 (A^*) \\
\bar{D}_3 &= \bar{R}_3^2 \cdot \left[ E_3 \cdot \mu_3 n_3 (A^*) + n_3 \cdot \frac{1}{\lambda_3} + 1 - A^* + \bar{R}_3 \right] - 1
\end{align*}
\]  

(2.2)

where

\[
E_3(A) = \sum_{i=0}^{n_3} \frac{A^i}{i!}
\]

3.2.3 Calculation of the moments of overflow traffic

Be given a full available primary arrangement with \( n_j \) servers and \( s_j \) waiting places \((j = 1, 2)\) and an infinite secondary arrangement \((n_j \rightarrow \infty)\). If there is offered random traffic (Poisson input and negative exponential service times, c.f. Eqs. (1.1, 1.2)) to the primary arrangement, the mean value \( R_j \) of the traffic overflowing to the secondary group with termination rate \( \varepsilon_j \) is given by the well-known formula [16]:

\[
R_j = \frac{\lambda_j}{\varepsilon_j} \cdot \frac{A_j}{\bar{R}_j} \cdot \left[ \frac{\lambda_j^m A^m_j}{\bar{R}_j^{m+1}} \right]
\]

(2.3)

where

\[
A_j = \frac{\lambda_j}{\varepsilon_j}
\]

Basharin [17] seems to be the first dealing with the variance of overflow traffic in delay-loss systems: recurrent formulae are presented for the computation of the moments of overflow traffic if there are \( s_j \geq 1 \) waiting places and uniform termination rate \( \varepsilon_j = \frac{\bar{R}_j}{E_j} = \varepsilon \) for both primary and (infinite) secondary arrangement.

In the following it is concisely outlined how to calculate recursively all overflow traffic moments even if there are different termination rates \( \varepsilon_j \) and \( \varepsilon_j \). For uniform termination rates also an explicit solution is presented for the variance \( V_j \) or variance coefficient \( D_j \), respectively.

The calculation has been performed following a way of solution which is similar to that one successfully applied for overflow systems without waiting places [18,19]:

a) Defining with \[u_j, w_j\] the state that there are simultaneously \( u_j \) demands being served or waiting for service in the primary arrangement and \( w_j \) demands in the secondary system, it is possible to find the following equations of state for the state probabilities \( p(u_j, w_j)\):

\[
\begin{align*}
\mu_j < \mu_j :
\lambda_j = & \mu_j \cdot p(u_j, w_j) + \mu_j + \lambda_j \cdot p(u_j, w_j) + \\
= & \mu_j \cdot p(u_j + 1, w_j) + \mu_j \cdot p(u_j, w_j)
\end{align*}
\]  

(2.4a)
\( \eta_j < u_j < \eta_j + s_j : \)
\[
\lambda_j \cdot p(u_j-1, u_j) + (u_j+1) \cdot \varepsilon_3 \cdot p(u_j, u_j+1) + \\
+ \eta_j \cdot \varepsilon_3 \cdot p(u_j+1, u_j) = (\eta_j \cdot \varepsilon_3 \cdot \lambda_j) \cdot p(u_j, u_j) 
\]
\( \mu_j = \eta_j + s_j : \)
\[
\lambda_j \cdot p(u_j, u_j+1) + \lambda_j \cdot p(u_j+1, u_j+1) + \\
+ (\eta_j \cdot \varepsilon_3 \cdot p(u_j+1, u_j+1) = (\eta_j \cdot \varepsilon_3 \cdot \lambda_j) \cdot p(u_j+1, u_j+1) 
\]
\( b) \) Solving this system of linear equations, the basic idea is to introduce factorial moments \( M_k(u_j) \), conditional factorial moments \( M_k(u_j | u_j) \), and conditional moments generating function \( P(u_j | u_j, t) \) for the overflow traffic:
\[
M_k(u_j) = \sum_{u_j=0}^{\eta_j+s_j} M_k(u_j | u_j) = \sum_{u_j=0}^{\eta_j+s_j} \sum_{u_{j-1}=0}^{u_j-1} \frac{1}{k!} \cdot p(u_j, u_j) \quad (2.5)
\]
\[
P(u_j | u_j, t) = \sum_{k=0}^{\infty} M_k(u_j | u_j) \frac{t^k}{k!} = \sum_{k=0}^{\infty} (t+1)^{u_j} \cdot p(u_j, u_j) \quad (2.6)
\]
All normal moments \( m_k(u_j) \) of the overflow traffic can be expressed by factorial moments, especially:
\[
k = 0 : \quad m_0(u_j) = M_0(u_j) = 1 \\
k = 1 : \quad m_1(u_j) = M_1(u_j) = R_j \\
k = 2 : \quad m_2(u_j) = M_2(u_j) + m_1(u_j) 
\]
and the variance coefficient
\[
D_j = M_2(u_j) - R_j^2 \quad (2.8)
\]
Therefore, in order to determine the variance coefficient \( D_j \), we have to calculate the factorial moment of second order.
\( c) \) By analogy to the Laplace transform [20] it is possible to transform Eqs. (2,4a,b,c) in corresponding equations for the generating function. Then using Eq. (2.6) it is possible to find, by comparing coefficients, the following equations for the conditional factorial moments:
\[
\eta_j < u_j < \eta_j + s_j:
- \lambda_j \cdot M_k(u_j | u_j) + (u_j+1) \cdot \varepsilon_3 \cdot M_k(u_j | u_j+1) - \\
- (u_j+1) \cdot \varepsilon_3 \cdot M_k(u_j | u_j+1) = 0 \quad (2.9a)
\]
\[
\eta_j < u_j < \eta_j + s_j:
- \lambda_j \cdot M_k(u_j | u_j) + (u_j+1) \cdot \varepsilon_3 \cdot M_k(u_j | u_j+1) - \\
- (u_j+1) \cdot \varepsilon_3 \cdot M_k(u_j | u_j+1) = 0 \quad (2.9b)
\]
\[
\eta_j < u_j < \eta_j + s_j:
- \lambda_j \cdot M_k(u_j | u_j) + (u_j+1) \cdot \varepsilon_3 \cdot M_k(u_j | u_j+1) = 0 \quad (2.9c)
\]
Summing up Eqs. (2.9a,b,c) over all possible values of \( u_j \) we find directly the following equation:
\[
\varepsilon_3 \cdot \sum_{u_j=0}^{\eta_j+s_j} M_k(u_j | u_j) = \varepsilon_3 \cdot M_k(u_j) = \lambda_j \cdot M_{k-1}(u_j | \eta_j + s_j) \quad (2.10)
\]
Eq. (2.10) shows that factorial moments of arbitrary order \( k \) may be obtained if only the conditional factorial moment of order \( k-1 \) is known.
Therefore, all factorial and conditional factorial moments of arbitrary order can be determined by the following procedure:
- By means of Eqs. (2.9a,b) with \( k = 1 \) and the relation
\[
\sum_{u_j=0}^{\eta_j+s_j} M_j(u_j | u_j) = M_j(u_j) = R_j \quad (2.11)
\]
it is possible to determine all conditional factorial moments \( M_j(u_j | u_j) \) (iteration of the values of the moments such that Eq. (2.11) is fulfilled).

Then, Eq. (2.10) allows with \( k = 2 \) to determine the factorial moment \( M_2(u_j) \) and hence Eq. (2.8) the variance coefficient \( D_j \).
For $k = 2$ a second iteration allows to determine all conditional factorial moments $M_{2}(u_{j}|u_{j})$ and hence the calculation of the third factorial moment $M_{3}(u_{j})$ by means of Eq.(2.10) with $k=3$.

Adequate procedures allow to determine factorial and normal moments of arbitrary order.

It should be mentioned at this point that, if there exists the same termination rate for both primary and secondary arrangement, an explicit solution is also available for all conditional factorial moments $M_{j}(u_{j}|u_{j})$ and therefore an explicit solution for the variance coefficient $D_{j}$. This explicit solution can be found when determining all moments $M_{j}(u_{j}|u_{j})$ as a function of $M_{j}(u_{j}|0)$ by means of Eqs.(2.9a,b) (linear homogeneous difference equations of second order). Then, taking into account Eq.(2.11), the explicit solution for the variance coefficient $D_{j}$ is given by:

$$D_{j} = \mathcal{R}_{j}^{2} \left[ \frac{c_{j} \cdot A_{j}}{c_{j} \cdot R_{j}} - 1 \right], \quad (2.12)$$

where

$$C_{j} = \frac{1}{k_{j}} \left[ 1 + \frac{A_{j}}{n_{j}} \cdot C_{j} - (k_{j}+k_{j}) \right] \left( k_{j}+k_{j} \right) S_{j}^{j+1} - \left[ 1 + \frac{A_{j}}{n_{j}} \cdot C_{j} - (k_{j}+k_{j}) \right] \left( k_{j}+k_{j} \right) S_{j}^{j+1} \right]$$

$$C_{2} = n_{j} - A_{j} + A_{j} \cdot C_{j} + \frac{A_{j}}{2 \cdot k_{j}} \left[ 1 + \frac{A_{j}}{n_{j}} \cdot C_{j} - (k_{j}+k_{j}) \right] \sum_{j=0}^{n_{j}+1} \left( k_{j}+k_{j} \right) S_{j}^{j+1} - \left[ 1 + \frac{A_{j}}{n_{j}} \cdot C_{j} - (k_{j}+k_{j}) \right] \sum_{j=0}^{n_{j}+1} \left( k_{j}+k_{j} \right) S_{j}^{j+1} \right]$$

$$C_{3} = \frac{\mathcal{R}_{j}^{n_{j}+1}}{(n_{j}+1)!} = \mathcal{E}_{j, n_{j}+1}(A_{j})$$

$$k_{1} = \frac{n_{j} - 1 + A_{j}}{2 \cdot n_{j}}$$

$$k_{2} = \sqrt{\left( \frac{n_{j} + A_{j}}{2 \cdot n_{j}} \right)^{2} - \frac{A_{j}}{n_{j}}}$$

3.2.4 Probability of waiting, mean waiting time and waiting time distribution

Remember that waiting is allowed only in the primary arrangements (c.f.3.2.1) and not for calls overflowing to the secondary route. Therefore, the probability of waiting, the mean waiting time, and the waiting time distribution can be obtained immediately when using Störmer's [16] well-known formulas for delay-loss systems.

3.2.5 Comparison with simulation results

As shown before, all moments of the overflow traffic distribution (section 3.2.3) and all waiting characteristics (section 3.2.4) can be determined exactly.

Using the method of "substitute primary arrangements" (see section 3.2.2) also the moments of its lost traffic ($\bar{R}, \bar{B}$) and, therefore, the probability of loss $B = \bar{R}/\bar{B}$ can be determined in close approximation. This approximate method has been checked by a large number of simulation runs using a digital computer [21]. Figures 8 and 9 present some typical examples. Comparison shows the very good accordance between simulated and calculated values.

3.3 Analysis of multi-queue models for overflow strategy 3

3.3.1 Model

Be given two delay-loss systems as shown in Fig.10. New calls of class one and two are offered at first to the corresponding arrangement number one (primary) or two (secondary). If the primary arrangement is blocked, i.e. the server and all waiting places $n_{1}$ are occupied, new arriving demands of class one are overflowing to the secondary arrangement (one server and $n_{2}$ waiting places). This secondary arrangement accepts all arriving demands of class one or two if there is at least one free waiting place. Calls of both class one and two are served without priorities in the order of their arrival (PIFO). If there is blocking, arriving calls are rejected.
3.3.2 Principle of solution

Calculating all characteristic traffic values, the basic idea is the same as shown in section 3.2. However, the investigation of the secondary arrangement with offered random plus overflow traffic is more complicated:

a) Traffic, overflown from the primary to the secondary arrangement, is described by the first and second overflow traffic moment (mean value $R_1$ and variance $V_1$ or variance coefficient $D_1 = V_1 - R_1$, respectively). Exact formulas for these moments are also included in the solutions presented in section 3.2.3.

b) The total traffic offered to the secondary group is described by the sum of both overflow traffic ($R_1, D_1$) and direct traffic ($R_2, D_2$)

$$\bar{R} = R_1 + R_2$$
$$\bar{D} = D_1 + D_2$$

c) In order to investigate the traffic characteristics of the secondary arrangement, a "substitute" primary arrangement (one server and an "waiting places") and a "generating" random traffic $\lambda$ are determined such that an overflow traffic is generated with the exact mean value $\bar{R}$ and the exact variance coefficient $\bar{D}$ (c.f. Fig. 11. On the contrary to section 3.2.2, the substitute primary arrangement has also waiting places; this generalization allows to investigate special structures and service strategies to be published later on).

Describing the traffic flow in the substitute primary and the actual secondary arrangement by a two-dimensional Markovian process, characteristic traffic values for the secondary arrangement are determined in the following sections.

3.3.2 Calculation of the state probabilities $p(u_1, u_2)$

Defining with $\{u_1, u_2\}$ the state that there are at the same time $u_1$ calls in the primary arrangement and $u_2$ in the secondary system, the following equations of state are found for the state probabilities:

$$\mu_1 < \lambda + S_1, \mu_2 = \lambda + S_2$$

$$p(u_1, S_1) = p(\lambda + S_1) \cdot p(o, 1 + S_2)$$
$$p(u_1 + 1, 1 + S_2) = p(\lambda + 1 + S_1) \cdot p(u_1 + 1, S_2) = 0$$

This equation is a system of $(n_1 + 1)$ linear homogeneous difference equations of second order with constant coefficients. Using the Z-transformation [20] it is possible to express all probabilities $p(u_1, 1 + S_2)$ as a function of $p(0, 1 + S_2)$:

$$p(u_1, 1 + S_2) = p(0, 1 + S_2) \cdot \left\{ \left( \frac{1 + A_1}{A_1} \right) \cdot \beta_{\mu_1} - A_1 \cdot \beta_{\mu_2} \right\}$$

with

$$\beta_{\mu_1} = \frac{\mu_1^\mu_1 - \mu_2^\mu_2}{\mu_1 - \mu_2}; \quad \frac{A_1}{A_1} = \frac{\xi}{\xi}; \quad A_1 = \frac{\lambda}{\xi}$$

$$\alpha_1, \alpha_2 = \frac{(1 + \xi + A_1)^2 - \sqrt{(1 + \xi + A_1)^2 - 4A_1}}{2}$$

By analogy, we also find:

$$\mu_1 \leq S_1, \mu_2 \leq S_2$$

$$p(u_1, 0) = p(0) \cdot \left\{ \left( \frac{1 + A_1}{A_1} \right) \cdot \beta_{\mu_1} - A_1 \cdot \beta_{\mu_2} \right\} - \sum_{\xi = 0}^{\frac{\lambda}{\xi}} p(\xi, \lambda - 1) \cdot \beta_{\mu_2}$$

$$\mu_1 \leq S_1, \mu_2 = 0$$

$$p(u_1, 0) = p(0) \cdot \left\{ \left( \frac{1 + A_1}{A_1} \right) \cdot \beta_{\mu_1} - A_1 \cdot \beta_{\mu_2} \right\} - \sum_{\xi = 0}^{\frac{\lambda}{\xi}} p(\xi, 1) \cdot \beta_{\mu_2}$$

with

$$\beta_{\mu_1} = \frac{(\mu_1^\mu_1 - \mu_2^\mu_2)/\mu_1}{\mu_1^\mu_1 - \mu_2^\mu_2}; \quad \alpha_1, \alpha_2 = \frac{(1 + A_1)^2 - \sqrt{(1 + A_1)^2 - 4A_1}}{2}$$

$$\mu_1 \leq 1 + S_1, \mu_2 \leq S_2$$

$$\mu_1 + 1, 1 + S_2 = 0$$

$$p(u_1, 0) = p(0 + S_1) \cdot p(1 + 1 + S_2) \cdot p(u_1 + 1, 1 + S_2)$$

$$+ A_1 \cdot p(u_1 + 1, S_2) = - A_1 \cdot p(u_1 + 1, S_2)$$
Eqs. (3.2a, b, c) allow to describe all probabilities \( p(\mu_1, \mu_2) \) as a function of the "basic state" \( p(0,\nu_2) \) and "higher" values \( p(\xi,\nu_2+1) \) \( [\xi = 0,1, \ldots, \nu_2+1] \). Therefore, when starting with the highest possible value \( \nu_2 = s_2+1 \) and systematic replacement of unknown values it is possible to express all state probabilities \( p(\mu_1,\mu_2) \) by the marginal values \( p(0,\nu_2), \ldots, p(0,s_2+1) \), i.e. the two-dimensional state relations can be reduced to a one-dimensional system [10]. Finally, equation (3.2d) and the normalizing condition allow to find an explicit solution for all state probabilities.

\[
p(\mu_1, \mu_2) = \frac{A_{\mu_1}}{1 + A_{\mu_1}} \frac{\sum_{\lambda_1=0}^{s_1} C_{\lambda_1, \mu_2+1}^\lambda \beta_{\mu_1, \mu_2+1}^\lambda}{\sum_{\lambda_2=0}^{s_2} C_{\lambda_2, \mu_2+1}^\lambda \beta_{\mu_1, \mu_2+1}^\lambda}
\]

Carried traffic

\[
Y_2 = \sum_{\mu_1=0}^{s_1} \sum_{\mu_2=1}^{s_2} p(\mu_1, \mu_2) = \frac{\lambda_1}{E_2} \left[ p(S_1, S_2) - p(S_1+1, S_2+1) \right]
\]

Lost traffic

\[
R_2 = \frac{\lambda_1}{E_2} \cdot p(1+S_1, 1+S_2)
\]

Probability of waiting (probability that an arriving call has to wait under the condition that it is offered to the secondary arrangement)

\[
W_2 = \frac{\sum_{\lambda_2=0}^{s_2} p(1+S_1, \lambda_2)}{\sum_{\lambda_2=0}^{s_2} p(1+S_1, \lambda_2)}
\]

Mean number of calls waiting for service

\[
\Omega_2 = \sum_{\lambda_1=0}^{s_1} (\mu_1-1) \cdot p(\mu_1, \mu_2)
\]

Mean waiting time for waiting calls

\[
t_{w2} = \frac{\Omega_2}{\lambda_1 \sum_{\lambda_2=0}^{s_2} p(1+S_1, \lambda_2)}
\]

In addition, it is possible to determine explicitly the waiting time distribution for the total system as well as for the secondary arrangement only. These results will be published at some future time.
3.1.5 Comparison with simulation results

As shown before, all characteristic traffic values referring to the primary arrangement (such as waiting times and overflow traffic) can be determined exactly.

All characteristic traffic values referring to the secondary arrangement are approximate values because the actually offered overflow traffics are replaced by a fictitious traffic generating exactly only the first two moments of the actual traffics.

Therefore, a large number of simulation runs using a digital computer have been performed [22]. Comparison of both simulated and calculated values shows the accuracy of this approach (c.f. Fig. 12 and 13).

4. COMPARISON OF DIFFERENT MULTI-QUEUE MODELS WITH RESPECT TO UNBALANCED LOAD

In this chapter, three numerical examples are given for demonstration of the effect of unbalanced load, different server configurations, as well as different overflow strategies on the characteristic traffic values (grade of service). By means of such numerical evaluations, suitable server configurations and operation strategies can be found, which meet the requirements of overflow, load-sharing, and reliability with respect to breakdowns in data transmission networks and computer systems.

4.1 Comparison of single-server systems with overflow strategies S1 and S3

In the upper left of Fig. 14a, three different models are shown each having two servers and two queues. Model (a) represents two separated single-server, single-queue systems without any mutual aid; model (b) works under overflow strategy S1, and model (c) under overflow strategy S3. The arrival rate to the second input remains constant \( \lambda_2 = 0.6 \text{ sec}^{-1} \), while \( \lambda_1 \) varies. The service rates of the servers are identical \( \mu_1 = \mu_2 = 1 \text{ sec}^{-1} \). The second server of model (b) serves the second queue according to a nonpre-emptive priority mode.

Fig. 14a shows the probabilities of loss \( B_1 \) and \( B_2 \), Fig. 14b shows the mean waiting times \( t_{w1} \) and \( t_{w2} \) referred to waiting calls, dependent on the arrival rate \( \lambda_1 \). In comparison with model (a) models (b) and (c) result in lower loss probabilities \( B_1 \) on the expense of \( B_2 \).

Model (b) reduces the mean waiting times \( t_{w1} \) drastically, while \( t_{w2} \) is not influenced. Model (c) guarantees a maximum throughput of 1-calls by the expense of increasing mean waiting times \( t_{w2} \).

In Fig. 14c the d.f. of waiting times are shown for the case of \( \lambda_1 = 1.0 \text{ sec}^{-1}, \lambda_2 = 0.6 \text{ sec}^{-1} \).

4.2 Comparison of many-server systems with overflow strategies S1 and S2

In the lower right of Fig. 15a, two overflow models are shown each having two primary routes and a common secondary route. Model (a) works under overflow strategy S1, model (b) operates according to overflow strategy S2. Again, \( \lambda_2 \) remains constant \( \lambda_2 = 2 \text{ sec}^{-1} \), while \( \lambda_1 \) varies. The service rates of the different servers are identical \( \mu_1 = 1 \text{ sec}^{-1} \).

Fig. 15a and 15b show that the probabilities of loss \( B_1 \) and \( B_2 \) as well as the waiting times \( t_{w1} \) and \( t_{w2} \) are greater for model (b) compared with model (a). However, model (b) yields a greater utilization for the primary routes by decreasing the load of the common secondary route simultaneously, cf. Fig. 15c.

4.3 Comparison of different overflow server configurations with overflow strategy S1

In Fig. 16 four different models are shown with two routes each having different service rates (transmission speeds) \( \mu_1 = 2 \text{ sec}^{-1}, \mu_2 = 1 \text{ sec}^{-1} \) as previously discussed in section 2.1. As before, the arrival rate \( \lambda_2 \) remains constant \( \lambda_2 = 2 \text{ sec}^{-1} \), while \( \lambda_1 \) assumes different values (cf. Table 1). Common servers serve their own queue according to a nonpre-emptive priority rule.

The characteristic traffic values \( B_j, W_j, t_{wj}, \) and \( Y_j, j = 1,2, \) are given in Table 1 for comparison of the efficiency of the different models 16.1 to 16.4.
References


List of Figure Captions

Fig. 1: Configurations of many-server systems with two queues and overflow capability for data transmission systems

Fig. 2: Configurations of many-server systems with two queues and load-sharing capability for computer systems

Fig. 3: Example of a 3-server overflow system with two queues

Fig. 4: State space and transitions for a 3-server overflow system with two queues according to Fig. 3. System state \( (x_1, y_1, z_1, i_2) \)

Fig. 5: State space and transitions for the waiting process of 1-calls for a 3-server system with two queues according to Fig. 3. Waiting state \( s(t) = (x_2, i_1, i_2) \). Initial state (example): \( s_0 = (1, i_1, i_2) \)

Fig. 6: The model

Fig. 7: Replacement of the real traffic by an equivalent traffic

Fig. 8: Probability of loss for the secondary arrangement. Comparison between calculation and simulation (simulation results with 95% confidence intervals within the circles; assuming Poisson behaviour for the overflow traffic the dashed line is obtained).

Fig. 9: Comparison between calculation and simulation (simulation results with 95% confidence interval; all termination rates \( \lambda_1^2 = \lambda_2^2 = \lambda_4 = 1/\text{sec} \)

Fig. 10: The model

Fig. 11: Replacement of the real traffic by an equivalent traffic

Fig. 12: Probability of waiting (directly offered and overflowing) calls waiting in the secondary arrangement for service. Comparison between calculation and simulation (simulation results with 95% confidence intervals within the circles; assuming Poisson behaviour for the overflow traffic the dashed line is obtained).

Fig. 13: Mean waiting time for calls waiting in the secondary arrangement. Comparison between calculation and simulation (simulation results with 95% confidence intervals within the circles; assuming Poisson behaviour for the overflow traffic the dashed line is obtained).

Fig. 14 Comparison of single-server systems, overflow strategies S1 and S3.
14a: Probabilities of loss versus arrival rate \( \lambda_1 \)
14b: Mean waiting times versus arrival rate \( \lambda_1 \)
14c: Distribution functions of waiting time

Fig. 15 Comparison of many-server systems, overflow strategies S1 and S2.
15a: Probabilities of loss versus arrival rate \( \lambda_1 \)
15b: Mean waiting times versus arrival rate \( \lambda_1 \)
15c: Carried traffic versus arrival rate \( \lambda_1 \)

Fig. 16 Comparison of different overflow server configurations, overflow strategy S1, c.f. Table 1.

Table 1 Comparison of different overflow server configurations, overflow strategy S1 (server configurations, c.f. Fig. 16).

<table>
<thead>
<tr>
<th>Interqueue discipline: non-preemptive priority</th>
<th>Parameters:</th>
<th>n_1=n_2=3, s_1=s_2=4, ( \lambda_1=2.1/\text{sec} ), ( \lambda_2=1.1/\text{sec} ).</th>
</tr>
</thead>
<tbody>
<tr>
<td>I/sec</td>
<td>$1/\eta_2$</td>
<td>Partial Rates</td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16 2 10 2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16 2 10 2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16 2 10 2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16 2 10 2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>16 2 10 2</td>
</tr>
</tbody>
</table>