Analysis and Application of an MMPP/PH/n/m Multi-Server Model

Udo R. Krieger Dietmar Wagner ‡

‡ Informatik IV, Technische Hochschule Darmstadt, D-6100 Darmstadt, F.R.G.

Abstract

Modeling a link in a B-ISDN at the connection level by Markovian techniques, we derive an MMPP/PH/n/m multi-server system as simplified, generic model. First, we calculate the steady-state vector of the underlying Markov chain of the model by advanced numerical methods. Then we determine the time and call congestion of the delay-loss system and the actual waiting-time distribution of a customer. Finally, the properties of some MMPP/PH/n/n loss systems are studied.

1 Introduction

Nowadays, modeling and analysis of complex distributed, technical systems such as telecommunication networks within a broadband ISDN (B-ISDN) environment have become important issues of performance analysis. Normally, such physical systems are described by queueing networks employing Markovian modeling techniques. Usually, either a discrete- or continuous-time modeling approach is used. In the following, we restrict our attention to continuous-time modeling techniques.

As the derived queueing networks arising from the continuous-time modeling of modern communication networks with state-dependent routing or advanced congestion-control mechanisms violate the restrictions of classical product-form networks of BCMP or Kelly type (cf. [11]), only simulation or computational techniques based on numerical solution methods for Markov chains are available as analysis methods (cf. [13], [29], [32]).

Regarding, however, the difficulties of simulation techniques to study rare events, for instance, cell loss probabilities in the range of 10^{-7} arising from the investigation of ATM networks (cf. [26]), numerical analysis methods seem to be the only feasible approach. The main drawback of such an approach stems from the huge, untractable number of states in Markovian models derived from actual networks. Therefore, suitable modeling techniques including special decomposition methods have to be employed to analyze large networks by parts.

Regarding modeling of packet-switched networks, a classical decomposition approach proposed by Kühn [15], Whitt [34] and Marie [17] among others uses GI/GI/1 models as generic elements of the queueing networks. It is well known, however, that the assumption that all streams within the network are renewal streams is crucial, apart from

limiting regimes such as networks in a heavy traffic environment. Usually the traffic streams in a B-ISDN are bursty point processes since they are generated in the sources by sampling and packetizing procedures employing variable bit rate coding techniques. They have to be modeled by versatile non-renewal streams chosen such that the resulting generic queueing models are tractable by analytical or numerical analysis methods. Therefore, it is reasonable to use queueing systems with special Semi-Markovian arrival streams (SMPs) that yield tractable Markovian models as building blocks of the model world. An important class of streams is provided by Markov-modulated Poisson processes (MMPPs) (cf. [13], [21], [16], [26], [9], [18]).
The holding times of network resources in a B-ISDN environment such as the occupation of links can be modeled by deterministic or arbitrary general distributions on $[0, \infty)$. It is known that they may be approximated with suitable accuracy by Coxian or, more generally, phase-type (PH) distributions (cf. [20]). Therefore, the service processes associated with the derived queueing models can be described by PH-distributions.
It is our main objective to analyze parts of an integrated broadband network that employs advanced routing techniques to set up virtual connections (cf. [32], [14], [27, §12]). Considering a classical circuit-switched network, Erlang's loss system M/G/n/n is the generic queueing model of a link between two exchanges. If we model a broadband network at the connection level applying the standard decomposition approach, the derived generic models include the generalized loss system MMPP/PH/n/n with one traffic class and combined delay-loss systems $\sum_i \text{MMPP}_i / \overrightarrow{\text{PH}} / n/m$ with two or more traffic classes of different characteristics (multi-service traffic models - cf. [27, §12-3-2, p. 670ff.], [14], [32]). Taking into account the different bandwidth requirements of traffic streams in a B-ISDN environment, the generic models are multi-class delay-loss systems $\sum_i \text{MMPP}_i^{X_i} / \overrightarrow{\text{PH}} / n/m$ with batch arrivals and bulk service as optional service discipline.
From a practical point of view, it is important that modeling and analysis are supported by convenient software tools. They should offer advanced user interfaces based on modern window and menu techniques. Following these considerations, a convenient software tool called MACOM (Markovian Analysis of Communication Systems) has been developed (cf. [13], [29]). Its model world provides multi-server queueing systems of the types PH/PH/n/m and MMPP/PH/n/m as generic elements. Furthermore, special variants of the multi-class delay-loss system $\sum_i \text{MMPP}_i^{X_i} / \overrightarrow{\text{PH}} / n/m$ with fixed batch size can be handled, too. MACOM implements a computational approach for modeling and analysis of communication systems based on Markovian techniques. It employs modern software design techniques and numerical solution methods for finite Markov chains.
In this paper, we restrict our attention to the simplest generic ISDN model of type MMPP/PH/n/m. We present some efficient algorithms for its analysis based on iterative solution techniques for Markov chains. They may be used to calculate the steady-state characteristics of this delay-loss system. These algorithms are considered as a supplement of the standard analysis techniques provided by MACOM that are based on Grassmann's algorithm and point iterative schemes such as the Gauss-Seidel and SOR procedure with optional aggregation-disaggregation steps (cf. [13]). They can be employed to study MMPP/PH/n/m models in isolation, as basic block iterative

solution methods in MACOM or as building blocks of a decomposition approach based on Semi-Markovian techniques. Extending the previous work of Stewart and Marie [31] and Seelen [30], we prove the convergence of the proposed iterative algorithms.
The paper is organized as follows: Section 2 describes the features of an MMPP/PH/n/m multi-server model. In section 3 the generator matrix associated with the basic Markov chain of this model is constructed and its properties are studied. In section 4 we present some efficient algorithms based on iterative solution techniques. They may be used to calculate the steady-state distribution corresponding to the basic Markov chain of the model. Section 5 is devoted to the calculation of the performance measures associated with the considered delay-loss system. Finally, we investigate the performance characteristics of some MMPP/PH/n/n loss systems.

2 Description of the MMPP/PH/n/m model

Variants of the MMPP/PH/n/m multi-server model are basic elements of the described model world of MACOM. Their features and analysis will be discussed subsequently. The related MMPP/PH/1/m single-server model and its generalized version N/G/1/m with Neuts' versatile Markovian point process as arrival stream have been investigated by Heffes and Lucantoni [9] and Blondia [3]. Interested readers are referred to their articles and the references therein.
In the following section we assume the reader to be familiar with the theory of homogeneous discrete- and continuous-time Markov chains with finite state spaces, abbreviated DTMC and CTMC, to the extent of the books of Heyman and Sobel [10, Chap. 7, 8] and Kemeny and Snell [12]. Furthermore, we shall adopt the terminology of Heyman and Sobel and we use the following notation of Berman and Plemmons [2, Chap. 2, p. 26] w.r.t. vector and matrix orderings: let $x \in \mathbb{R}^n$, then $x \gg 0 \Leftrightarrow x_i > 0$ for each $i \in \{1,\ldots,n\}$, $x > 0 \Leftrightarrow x_i \geq 0$ for each $i \in \{1,\ldots,n\}$ and $x_j > 0$ for some $j \in \{1,\ldots,n\}$, $x \geq 0 \Leftrightarrow x_i \geq 0$ for each $i \in \{1,\ldots,n\}$.
Let us now consider a multi-server model of type MMPP/PH/n/m. It has the following properties:

- The arrival stream is a Markov-modulated Poisson process (MMPP) with s states, an irreducible generator matrix $Q \in \mathbb{R}^{s\times s}$ and the arrival rate vector $\lambda = (\lambda_1 \ldots, \lambda_s)^t > 0$ (cf. [21, p. 269]). Let $Y(t), t \geq 0$, denote the corresponding irreducible CTMC on the state space $\{1,\ldots,s\}$ describing the phase of the MMPP.

- The service facility consists of $n > 1$ identical, parallel servers.

- The capacity of the system comprising the number of servers and the number of waiting places is $m = n + l$, i.e., there are $l \geq 0$ waiting positions and n parallel servers in the system.

- The service discipline is 'delay-loss with FIFO'. If there are less than n customers in the system, an arriving customer selects a free server at random and occupies it for a random service period. If all servers are busy at his arrival instant, the

customer joins the waiting line. If no waiting positions are available, he is lost and has no further impact on the system.

- The service times of the customers are independent, identically distributed random variables governed by a phase-type distribution F of order k with irreducible representation (β, T). We assume $\beta^t e = 1$, where e is the vector of all ones, and set $\mu_i = -T_{ii} > 0$, $1 \leq i \leq k$.
  Furthermore, the service times are assumed to be independent of the arrival process.

According to its definition (cf. [20]), the generic service-time distribution F evaluates the time to absorption in the state $k+1$ of a CTMC with finite state space $\{1, \ldots, k, k+1\}$ and transient states $E = \{1, \ldots, k\}$ provided that the chain is started in the transient set E according to the probability vector β. The corresponding generator matrix is given by $G = \begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix}$ with a regular M-Matrix $-T \in \mathbb{R}^{k \times k}$ and a vector $0 < T^0 = (T_{1k+1}, \ldots, T_{kk+1})^t \in \mathbb{R}^k$ satisfying $Te + T^0 = 0$. Then $F(u) = 1 - \beta^t \cdot e^{Tu} \cdot e$, $u \geq 0$, holds.
In analogy to the uniformization of ergodic Markov chains, we may proceed to an embedded DTMC associated with the absorbing CTMC defined above (cf. [4], [10]). Let $D = -\mathrm{diag}(T_{11}, \ldots, T_{kk}, 0) > 0$. Here, $D = \mathrm{diag}(x)$ denotes a diagonal matrix associated with a vector x which is defined by $D_{ii} = x_i$. Then we set $\alpha' = D^\sharp G + I$ where I denotes the identity matrix and $D^\sharp = -\mathrm{diag}(T_{11}^{-1}, \ldots, T_{kk}^{-1}, 0)$ is the group inverse of D. Hence,

$$\alpha'_{ij} = \begin{cases} -T_{ij}/T_{ii} & \text{for } i \neq j, 1 \leq i \leq k, 1 \leq j \leq k+1 \\ 1 & \text{for } i = j = k+1 \\ 0 & \text{otherwise} \end{cases}$$

holds and α' is a stochastic matrix. It follows $G = D(\alpha' - I)$. Obviously, α'_{ij} is the probability to proceed to state j after departure from state i. The sojourn time in a transient state $i \in E = \{1, \ldots, k\}$ is governed by an exponential distribution with parameter $\mu_i = D_{ii} = -T_{ii} > 0$ (cf. [4]).
Let us denote the $k \times k$ principal submatrix of α' by α. Hence, $\alpha = [-\mathrm{diag}(T_{11}, \ldots, T_{kk})]^{-1} T + I$ is a strictly substochastic matrix, i.e., $\alpha \geq 0, \alpha e < e$.
The behavior of the MMPP/PH/n/m queueing model may be described by a stochastic process $Z(t) = (R(t), H(t), Y(t)), t \geq 0$. Here, $R(t) \in \{0, \ldots, m\}$ denotes the number of customers in the system at time t. Given $n > 1$, $H(t) = (h_1(t), \ldots, h_k(t))$ with $h_j(t) \in \{0, \ldots, n\}$, $1 \leq j \leq k$, is the phase vector of the numbers of customers just served in the different phases j of the service process at time t. $Y(t) \in \{1, \ldots, s\}$ is the phase of the CTMC controlling the arrival stream at time t. The number of busy servers is given by $N(t) = H(t) \cdot e = \sum_{j=1}^k h_j(t) = \min(R(t), n)$. According to the assumptions, the vector process $Z(t), t \geq 0$, is a CTMC on the finite state space $S = \{z = (r, h_1, \ldots, h_k, y) \in \mathbb{N}_0^{k+2} \mid 0 \leq r \leq m, 1 \leq y \leq s, 0 \leq h_j \leq n \text{ for } 1 \leq j \leq k \text{ subject to } \sum_{j=1}^k h_j = \min(r, n)\}$. Here we imbed, of course, all admissible vectors $H = (h_1, \ldots, h_k) \in \mathbb{N}_0^k$ in S by identifying the vectors (r, H, y) and $(r, h_1, \ldots, h_k, y) \in S$.

3 Construction of the generator matrix

Important steady-state performance characteristics of the MMPP/PH/n/m model such as the time and call congestion are defined in terms of the steady-state distribution π of the CTMC $Z(t)$. In order to calculate π, we have to construct the generator matrix $\tilde{Q}$ of $Z(t)$. For this purpose, we have to fix an ordering of the states of the Markov chain first. Then we shall enumerate the states and determine the rates of all transitions between these states.

3.1 Ordering of states

We use the convenient ordering of states '$\prec$' defined by Stewart and Marie [31] for the related M/PH/n/m model. First, we divide the state space into macrostates $[r] = \{(r, H, y) \mid \forall H \geq 0, y \in \{1, \ldots s\} : (r, H, y) \in S\}, r = 0, \ldots, m$, called levels or R-lumps (cf. [20, p. 5]). They are defined by fixing the number of customers in the system, for instance, $R(t) = r$. These macrostates are ordered according to the lexicographical ordering. The microstates $[(r, H)] = \{(r, H, y) \mid \forall y \in \{1, \ldots s\} : (r, H, y) \in S\}$ within a macrostate determined by fixing both $R(t) = r$ and $H(t) = (h_1(t), \ldots, h_k(t)) = H = (h_1, \ldots, h_k)$ are called H_r-lumps. We order the vectors H of all H_r-lumps within each R-lump $[r]$ according to the reverse lexicographical ordering of their components. The last component determined by $Y(t) = y$ is lexicographically ordered again, i.e., $(r, r, \ldots, 0, 1) \prec \ldots \prec (r, r, \ldots, 0, s) \prec (r, r-1, 1, \ldots, 0, 1) \prec \ldots \prec (r, r-1, 1, \ldots, 0, s) \prec \ldots (r, 0, 0, \ldots, r, 1) \prec \ldots \prec (r, 0, 0, \ldots, r, s)$.

To construct the generator matrix $\tilde{Q}$, we have at first to identify its zero structure. Therefore, it is necessary to enumerate the states based on the prescribed ordering, i.e., we have to define a position mapping $p : S \to \mathbb{N}, \quad p : z \mapsto p(z) = p_z$. Suppose there are r customers in the system. Then one has to distribute these r customers among n servers and $l = m - n$ waiting places. Each busy server stays in one of the k phases of the service process. Based on these observations, it is easy to see that each R-lump $[r]$ comprises $\binom{r+k-1}{r}s$ states if $r < n$ and $\binom{n+k-1}{n}s$ states if $r \geq n$ holds.

In the following, we exploit for $\nu \leq r$ the identities

$$\sum_{j=0}^{r-\nu} \binom{j+k-t}{j} = \sum_{i=\nu}^{r} \binom{r-i+k-t}{r-i} = \binom{r-\nu+k-t+1}{r-\nu} = \binom{r-\nu+k-t+1}{k-t+1}$$

and the standard boundary conventions $\binom{n}{m} = 0$ for all integers $0 \leq n < m$ and $\binom{n}{-m} = 0$ for $n = 0, \pm 1, \pm 2, \ldots$ and all integers $m \geq 1$ (cf. [24, p. 1]). Then it follows by some algebraic manipulations that the position index p_z of an arbitrary state $z = (r, h_1, \ldots, h_k, y) \in S$ is determined by

$$\begin{aligned} p_z \;=\; & \left[\binom{\min(r,n)+k-1}{k} + (\max(r,n)-n)\binom{n+k-1}{n}\right. \\ & \left. + \sum_{i=1}^{k-1} \binom{\min(r,n)-(h_1+\ldots+h_i+1)+k-i}{k-i}\right] s + y \end{aligned}$$

given $\sum_{i=1}^{0} \equiv 0$ (cf. [33]).

3.2 Determination of the transition rates

To determine the transition rates between the states of the CTMC $Z(t)$, we have to analyze the transition behavior of the Markov chain. Suppose $Z(t) = (R(t), H(t), Y(t)) = (r, h_1, \ldots, h_k, y)$. We must consider only those events occuring in an interval $[t, t+\delta)$ of infinitesimally small length δ whose transition probabilities exceed $o(\delta)$. Obviously, the following four distinct events cause such transitions:

1. an arrival of a new customer (who can enter the system)
   This birth event changes the components $R(t)$ and $H(t)$ where the R-level is incremented by one if the customer can enter the system. If he is lost there is no change at all.
   If $r < n$ and the customer occupies a free server starting its service in phase i, then $Z(t+\delta) = (r+1, \ldots, h_i+1, \ldots, y)$ results, whereas $Z(t+\delta) = (r+1, h_1, \ldots, h_k, y)$ holds for $n \leq r < m$. In the latter case, all servers are busy and the customer can enter the waiting room of the system. The transition rate of the first event is given by $\lambda_y \beta_i$, that of the second event by λ_y.

2. a phase shift of one customer's service process without service completion
   This internal phase shift of a service process changes only $H(t)$. But it can happen only if $R(t) = r \geq 1$ holds. $Z(t+\delta) = (r, \ldots, h_j - 1, \ldots, h_i + 1, \ldots, y)$ results, for instance, if a transition from phase j to i occurs. The corresponding rate of this event is $h_j \mu_j \alpha_{ji}$.

3. a phase shift of the CTMC controlling the arrival stream
   It changes only $Y(t)$. $Z(t+\delta) = (r, h_1, \ldots, h_k, u)$, for instance, corresponds to a transition from y to u. The rate of this event is Q_{yu}.

4. a service completion of a customer
   This death event changes the components $R(t)$ and $H(t)$ simultaneously decrementing the R-level by one. To determine the transition rates, we have to distinguish the events, that there are no customers waiting or that some are waiting. In the first case $1 \leq r \leq n$, $Z(t+\delta) = (r-1, \ldots, h_j - 1, \ldots, y)$ holds if the customers leaves its service process in phase j. The transition rate of this event is given by $h_j \mu_j \alpha_{jk+1}$. In the second case $n < r \leq m$, a waiting customer will occupy the free server in phase i of the service process immediately after the departure of the served customer from phase j of the service process. If $i \neq j$ holds, then $Z(t+\delta) = (r-1, \ldots, h_j - 1, \ldots h_i + 1, \ldots, y)$ occurs with transition rate $h_j \mu_j \alpha_{jk+1} \beta_i$. If $i = j$ holds, then $Z(t+\delta) = (r-1, h_1, \ldots, h_k, y)$ occurs with transition rate $\sum_{i=1}^{k} h_i \mu_i \alpha_{ik+1} \beta_i$.

In the last subsection we have shown that each R-lump $[r]$ has $d_r = \binom{\min(r,n)+k-1}{\min(r,n)} s$ states. Thus the generator matrix $\tilde{Q} \in \mathbb{R}^{d \times d}$ associated with $Z(t)$ has the order

$$d = \sum_{i=0}^{m} d_i = \left[\binom{n+k}{n} + (m-n) \binom{n+k-1}{n} \right] s \,.$$

Obviously, d is linear in s and m, but it grows exponentially fast for $n > 1$ and $k > 1$. Some examples are provided by Table 1. In comparison to that, for $k = 1$ $d = (m+1)s$ holds. For $n = 1$ it follows $d = (m \cdot k + 1)s$ since in this case $H(t) \in \{0, \ldots, k\}$ only records the phase of the service process at time t, where $H(t) = 0$ indicates an idle server.

Regarding the lexicographical ordering of R-lumps, the generator matrix evidently possesses a block tridiagonal structure determined by the levels $R(t) = r$:

$$\tilde{Q} = \begin{pmatrix} \tilde{Q}_{00} & \tilde{Q}_{01} & 0 & \cdots & \cdots & \cdots & 0 \\ \tilde{Q}_{10} & \tilde{Q}_{11} & \tilde{Q}_{12} & \ddots & \cdots & \cdots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \ddots & \tilde{Q}_{nn-1} & \tilde{Q}_{nn} & \tilde{Q}_{nn+1} & \ddots & \vdots \\ \vdots & \cdots & \ddots & \tilde{Q}_{n+1n} & \tilde{Q}_{n+1n+1} & \ddots & 0 \\ \vdots & \cdots & \cdots & \ddots & \ddots & \ddots & \tilde{Q}_{m-1m} \\ 0 & \cdots & \cdots & \cdots & 0 & \tilde{Q}_{mm-1} & \tilde{Q}_{mm} \end{pmatrix} \tag{1}$$

Each block matrix $\tilde{Q}_{ij}$ is a $d_i \times d_j$ matrix. The upper diagonal blocks $\tilde{Q}_{ii+1}$ correspond to arrival events of type 1. The lower diagonal blocks $\tilde{Q}_{ii-1}$ are associated with departure events of type 4. The diagonal blocks $\tilde{Q}_{ii}$ correspond to events of types 2 and 3.

Provided that $\tilde{Q}$ is irreducible, all off-diagonal blocks are nonzero, nonnegative matrices, whereas the diagonal blocks are regular Metzler-Leontief matrices; i.e., $-\tilde{Q}_{ii}$ is a regular M-matrix since it is a proper principal submatrix of an irreducible singular M-matrix after an appropriate permutation (cf. [2]).

The construction of the generator matrix $\tilde{Q}$ may be performed by special modeling tools such as MACOM (cf. [29], [13]) or by the direct generation of the block matrices employing Lucantoni's and Ramaswami's algorithms (cf. [23], [22]). Subsequently, we adopt Ramaswami's and Lucantoni's notation and apply their algorithms. Setting $\tilde{G} = \begin{pmatrix} 0 & 0 \\ T^0 & T \end{pmatrix}$, the construction algorithm reads as follows:

$$\begin{array}{lcll} \tilde{Q}_{ii+1} & = & P_i(\beta^t) \otimes \Lambda & \text{for} \quad i = 0, \ldots, n-1 \\ \tilde{Q}_{ii+1} & = & I_{\tilde{d}_i} \otimes \Lambda & \text{for} \quad i = n, \ldots, m-1 \\ \tilde{Q}_{ii-1} & = & L_{n-i}(n, \tilde{G}) \otimes I_s & \text{for} \quad i = 1, \ldots, n \\ \tilde{Q}_{ii-1} & = & Q(n, T^0\beta^t) \otimes I_s & \text{for} \quad i = n+1, \ldots, m \\ \tilde{Q}_{00} & = & Q - \Lambda & \\ \tilde{Q}_{ii} & = & A(\min(i,n), T) \otimes I_s + I_{\tilde{d}_i} \otimes Q - \tilde{\Delta}_i & \text{for} \quad i = 1, \ldots, m \end{array} \tag{2}$$

Here $\tilde{\Delta}_i \geq 0$ are diagonal matrices associated with $\tilde{Q}_{ii}$ that guarantee $\tilde{Q}e = 0$ and $\tilde{d}_i = d_i/s$. I_l denotes the identity matrix of order l and $\Lambda = \mathrm{diag}(\lambda)$ is the matrix of the arrival rates. The diagonal elements of $Q(n, T^0\beta^t)$ are calculated by means of [22, Theorem 2, p. 424]. Obviously, apart from the diagonal elements, all diagonal blocks coincide for $i = n, \ldots, m$. Moreover, we have used the relation $\tilde{Q}_{00} = Q - \tilde{\Delta}_0$ with $\tilde{\Delta}_0 = \mathrm{diag}((P_0(\beta^t) \otimes \Lambda) \cdot e) = \mathrm{diag}(\beta^t \cdot e \otimes \Lambda \cdot e) = \mathrm{diag}(\lambda) = \Lambda$.

3.3 Irreducibility properties of the generator matrix

Regarding the calculation of the steady-state distribution π of $Z(t)$, it is necessary to characterize the irreducibility of the generator matrix $\tilde{Q}$. The following Proposition provides a necessary and sufficient condition in terms of the zero structure of the service process. It can be proved in a straightforward manner (cf. [33]).

Proposition 1
The generator matrix $\tilde{Q}$ is irreducible iff for each phase $j \in \{1,\ldots,k\}$ of the service-time distribution the following property (P) holds:

(P) There exist an index i and a set $\{\alpha_{l_0 l_1}, \alpha_{l_1 l_2}, \ldots, \alpha_{l_s l_{s+1}}, \ldots, \alpha_{l_{s+r} l_{s+r+1}}\}$ such that $\beta_i > 0$, $l_0 = i, l_{s+1} = j, l_{s+r} = j_r, l_{s+r+1} = k+1$ and $\alpha_{l_m l_{m+1}} > 0$ for $m = 0,\ldots,s+r$ hold.

□

Note, that the assumptions $\beta \geq 0, \beta^t e = 1$ and $T^0 > 0$ imply the existence of indices i, j_r with $\beta_i > 0$ and $\alpha_{j_r,k+1} > 0$. Now we are able to determine the irreducibility of $\tilde{Q}$ analyzing only the structure of the PH-type service-time distribution.

Proposition 2
The generator matrix $\tilde{Q}$ is irreducible iff the service-time distribution with PH-representation (β, T) is irreducible, i.e., if the matrix $T + T^0\beta^t$ is irreducible. □

As this assumption on the PH-representation is satisfied in the given context, the constructed generator matrix $\tilde{Q}$ of the MMPP/PH/n/m model is irreducible. Thus, the existence and uniqueness of the steady-state distribution π of the system is guaranteed. π is the positive solution of

$$A \cdot \pi = -\tilde{Q}^t \cdot \pi = 0 \tag{3}$$

subject to the normalization condition $\pi^t \cdot e = 1$.
Regarding the numerical solution of the homogeneous system (3) of linear equations, the block tridiagonal structure and sparsity should be exploited. Some iterative solution procedures require the irreducibility of the diagonal blocks A_{ii} of the singular M-matrix $A = -\tilde{Q}^t$. Therefore, it is necessary to characterize this property by equivalent conditions. They are provided by the following Proposition (cf. [33]).

Proposition 3
The following conditions are equivalent:

(i) All diagonal blocks $\tilde{Q}_{ii}$ of the generator matrix $\tilde{Q}$ are irreducible.

(ii) The strictly substochastic matrix $\alpha = -[diag(T_{11},\ldots,T_{kk})]^{-1}T + I \in \mathbb{R}^{k\times k}$ associated with the service-time distribution of PH-type (β, T) is irreducible.

(iii) The regular Metzler-Leontief matrix $T \in \mathbb{R}^{k\times k}$ is irreducible.

□

4 Computation of the steady-state distribution by numerical solution methods

Obviously, the special structure (1), (2) of the generator matrix $\tilde{Q}$ implies that the steady-state distribution π of the CTMC $Z(t)$ has a generalized matrix-geometric form (cf. [8]). In this section, we present an alternative approach for calculating the steady-state vector π. It is based on the numerical solution of the homogeneous system (3) and well suited for an efficient implementation on a parallel or vector computer. Regarding the block tridiagonal structure and sparsity of the system (3), it is advantageous to employ iterative solution methods based on matrix splittings for singular M-matrices such as the block Gauss-Seidel or block SOR procedure (cf. [13]). To guarantee, however, the convergence of these procedures, some structural requirements on the system matrix A have to be fulfilled.
Let us first consider an MMPP/PH/n/m model with a non-exponential service-time distribution. Regarding the convergence of block iterative procedures based on regular matrix splittings $A = M - N$, a unifying framework is provided by R-regular splittings $A = (D - L) - (L(N) + U(N) + D(N))$ introduced by Rose [25].

Definition 1
Let $A \in \mathbb{R}^{n \times n}$ be a (singular) M-matrix with block partition $A = (A_{ij})_{1 \le i,j \le p}$, provided that $p > 1$. Assume the block splitting $A = (D - D(N)) - (L + L(N)) - U(N)$ has the following properties:

(1) $D = diag(D_{ii})_{1 \le i \le p}$ and $D(N)$ are block diagonal matrices with $D(N) \ge 0$. L and $L(N)$ are strictly lower block triangular matrices such that $L \ge 0, L(N) \ge 0$ hold. $U(N)$ is a strictly upper block triangular matrix with $U(N) \ge 0$.

(2) $D_{ii}^{-1} \gg 0$ for $1 \le i \le p$.

(3) $M = D - L$ is a lower block triangular matrix.

(4) $N = L(N) + U(N) + D(N) \ge 0$.

(5) $A_0 = D - L - U(N)$ is irreducible.

(6) The block matrix graph $\Gamma(A_0) = (V, E)$ has a monotone decreasing cycle, this is a sequence $c = (i_1, i_2, \ldots, i_l, i_1)$ of adjacent nodes with the property $i_l \ne i_1$ and $i_j \ge i_{j+1}, 1 \le j \le l-1$.
Recall that the block matrix graph $\Gamma(A_0) = (V, E)$ is a directed matrix graph with nodes $V = \{V_i \mid 1 \le i \le p\}$ and directed edges $(V_i, V_j) \in E$. V_i results from the partition of the index set $\{1, \ldots, n\}$ according to the block partition. $(V_i, V_j) \in E$ iff $A_{ij} \ne 0$, that means, there are indices $l \in V_i, m \in V_j$ such that $(l, m) \in E_{\Gamma(A)}$ is an edge in the matrix graph of A.

A block splitting $A = M - N$ with the properties (1) to (6) is called R-regular (block) splitting.

Such schemes include the block Gauss-Seidel procedure, defined by M = D - L, D(N) = L(N) = 0, N = U(N), and its modifications that exploit the sparsity structure of A. By that means it is possible to employ incomplete LU-factorization techniques in the solution process.

We want to exploit the natural partition $A = (A_{ij})_{0\le i,j\le m}$ defined by the block tridiagonal structure (1). Therefore, we assume that an R-regular splitting based on this natural partition is given. A sufficient condition that guarantees property (2) in Definition 1 and, hence, the convergence of the iterative scheme arising from the R-regular splitting, e.g. the block Gauss-Seidel procedure, requires, however, that the diagonal blocks $D_{00},\dots,D_{mm}$ are regular, irreducible M-matrices (cf. [25], [13]). Hence, the convergence of the block Gauss-Seidel procedure can only be guaranteed if the matrix T associated with the service-time distribution of PH-type is irreducible (cf. [25, Cor. 2, p. 98]). In this case, the resulting iterative scheme may be accelerated by the use of relaxation techniques or the insertion of some aggregation-disaggregation steps (cf. [13], [28]).

Regarding the generator matrix $\tilde{Q}$, we note that the diagonal blocks $\tilde{Q}_{nn}$, $\tilde{Q}_{n+1n+1},\dots,\tilde{Q}_{mm}$ have the same off-diagonal elements. Furthermore, the row sums of all lower diagonal blocks $\tilde{Q}_{ii-1}$, $i = 1,\dots,m$, are equal to $\sum_{j=1}^{k} h_j\mu_j\alpha_{jk+1}$. As all upper diagonal matrices $\tilde{Q}_{nn+1},\dots,\tilde{Q}_{m-1m}$ coincide with $I\otimes\Lambda$, all diagonal elements of $\tilde{Q}_{nn},\dots,\tilde{Q}_{m-1m-1}$ are equal and the elementwise maximum of all positive diagonals of $-\tilde{Q}_{nn},-\tilde{Q}_{n+1n+1},\dots,-\tilde{Q}_{mm}$ is given by that of the nth block $-\tilde{Q}_{nn}$. Thus, the difference of the diagonals of $\tilde{Q}_{ii}$, $i = n,\dots,m-1$, and that one of $\tilde{Q}_{mm}$ is equal to $I\otimes\Lambda$ provided that a matrix notation is used.

If T is irreducible, the following R-regular splitting $A = -\tilde{Q}^t = M - N$ exploits this structure:

$$M = \begin{pmatrix} A_{00} & 0 & \dots & \dots & \dots & \dots & 0 \\ A_{10} & A_{11} & 0 & \dots & \dots & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \dots & \dots & \vdots \\ \vdots & \ddots & A_{nn-1} & A_{nn} & 0 & \dots & \vdots \\ \vdots & \dots & \ddots & A_{n+1n} & A_{nn} & \ddots & \vdots \\ \vdots & \dots & \dots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \dots & \dots & 0 & A_{mm-1} & A_{nn} \end{pmatrix} = D - L \qquad (4)$$

$$N = \begin{pmatrix} 0 & -A_{01} & 0 & \dots & \dots & \dots & 0 \\ 0 & \ddots & -A_{12} & \ddots & \dots & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \dots & \vdots \\ \vdots & \dots & 0 & 0 & -A_{nn+1} & \ddots & \vdots \\ \vdots & \dots & \dots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & \dots & \dots & \ddots & 0 & -A_{m-1m} \\ 0 & \dots & \dots & \dots & \dots & 0 & I\otimes\Lambda \end{pmatrix}$$

$$= L(N) + U(N) + D(N) \qquad (5)$$

It is particularly recommended if $m \gg n$ holds.
According to the construction algorithm (2), $A_{ii-1} = (-\tilde{Q}^t)_{ii-1} = -P^t_{i-1}(\beta^t) \otimes \Lambda \leq 0$, $i = 1, \ldots, n$, and $A_{ii-1} = -I \otimes \Lambda \leq 0$, $i = n+1, \ldots, m$ hold. Furthermore, $A_{ii} = -\tilde{Q}^t_{ii}$, $i = 0, \ldots, n$, are irreducible regular M-matrices, hence, $D_{ii}^{-1} = A^{-1}_{\min(i,n),\min(i,n)} \gg 0$ follows. Obviously, M is a regular M-matrix and $L(N) = 0$, $N = U(N) + D(N) \geq 0$ hold. Moreover, $A_0 = D - L - U(N)$ has the same zero structure as A implying the irreducibility of the block tridiagonal matrix A_0. Hence, the corresponding block matrix graph $\Gamma(A_0)$ possesses a monotone decreasing cycle, too. Thus the proposed splitting (4), (5) is both an R-regular and an M-splitting.
Regarding the generator matrix

$$G = \begin{pmatrix} -\mu_1 & \mu_1 & 0 & \ldots & 0 \\ 0 & -\mu_2 & \mu_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & -\mu_k & \mu_k \\ 0 & \ldots & \ldots & 0 & 0 \end{pmatrix}$$

of a generalized Erlang distribution ($F \in GE_k \subset PH$), we see, however, that the corresponding submatrix T is reducible. Hence, D_{ii} are reducible regular M-matrices implying $D_{ii}^{-1} > 0$. In such cases, a modification of the R-regular splitting (4), (5) must be used. It is based on the following result (cf. [1, Theorem 3.9], [13]) that employs the well-known relaxation technique.

Theorem 1
Let $A \in \mathbb{R}^{n \times n}$ be a singular irreducible M-matrix with block partition $A = (A_{ij})_{i,j=1,\ldots,p}$. Consider a block splitting $A = (D - D(N)) - (L + L(N)) - U(N)$ which satisfies the properties (1), (3), (4) of an R-regular splitting according to Definition 1 and, in addition, (2') $D_{ii}^{-1} > 0$ for $1 \leq i \leq p$. Define the splitting $M_\omega = \frac{1}{\omega}(D - \omega L)$, $N_\omega = \frac{1}{\omega}((1-\omega)D + \omega N)$, $N = L(N) + U(N) + D(N)$ for some $0 < \omega < 1$.
Then $A = M_\omega - N_\omega$ is a weak regular splitting and the iteration matrix $T_\omega = M_\omega^{-1} N_\omega$ is semiconvergent. The scheme $x^{(k+1)} = T_\omega \cdot x^{(k)}$ converges to a nonnegative, nontrivial solution of $Ax = 0$ provided that the initial vector $x^{(0)}$ is positive. □

Regarding the MMPP/M/n/m variant of the model with exponentially distributed service times, a similar convergent, iterative procedure may be applied (cf. [25], [13]). It is also based on an R-regular splitting and has been suggested by Meier-Hellstern [18, §3], but without a rigorous mathematical proof of the convergence of the resulting iterative scheme. The algorithm exploits the uniform structure of the diagonal blocks $A_{ii} = -Q^t + \min(i,n)\mu I + \Lambda(1 - \delta_{im})$, $i = 0, \ldots, m$, that are of the same size $d_i = s$ now. Here, $\Lambda = \mathrm{diag}(\lambda)$ is again the matrix of the arrival rates.
As the diagonal blocks A_{ii} are irreducible regular M-matrices in this case, it follows from [25, Cor. 2, p.139] that the block Gauss-Seidel procedure and, hence, the corresponding block SOR variant are convergent, too. Of course, these algorithms may be accelerated by inserting several aggregation-disaggregation steps during the iteration according to

the IAD scheme (cf. [13], [28]).
In comparison with the block Gauss-Seidel or SOR scheme, the proposed R-regular splitting procedure has the advantage that all diagonal blocks of the matrix M are identical with $-Q^t + n\mu I + \Lambda$. Therefore, it is necessary to decompose only this small matrix and to store its inverse during the iteration process. The resulting algorithm is well suited for an implementation on a vector processor. Experimental results concerning the performance of the scheme have been provided by Meier-Hellstern [18].

5 Performance measures of the model

Regarding the application of an MMPP/PH/n/m model in teletraffic theory, it is our objective to calculate the steady-state performance characteristics of this delay-loss system, namely, the time- and arrival-stationary distribution of the number of customers in the system and the actual waiting-time distribution of the customers. For this purpose, we have to use the steady-state distribution π of the CTMC $Z(t)$.

5.1 Time-stationary distribution of the number of customers

Given a feasible state $z = (r, H, y) \in S$, the time-stationary probabilities $\pi(r, H, y) = \pi_{(r,H,y)} = \lim_{t\to\infty} \text{Prob}\{R(t) = r, H(t) = H = (h_1, \ldots, h_k), Y(t) = y\}$ may be computed by iterative procedures as normalized solution of the balance equations (3). Obviously, the time-stationary distribution P of the number of customers in the system is given by the marginal probabilities

$$P_r = (\pi_{[r]})^t e = [(\pi(r, H, y))_{(r,H,y)\in[r]}]^t e = \sum_{\forall(r,H,y)\in[r]} \pi(r, H, y)$$

of each R-lump $[r] = \{(r, H, y) \mid \forall H, y : (r, H, y) \in S\}, r = 0, \ldots, m$. Hence, the time congestion of the model is determined by $P_m = (\pi_{[m]})^t e$.

5.2 Arrival-stationary distribution of the number of customers

To calculate the arrival-stationary distribution $P^{(0)}$ of the number of customers in the system, i.e., the probabilities $P_r^{(0)}$ that an arriving customer finds r customers in the system at his arrival instant, we may employ Melamed's approach [19]. It uses a level crossing argument for Markov chains that is also known as stochastic intensity principle in the general setting of marked point processes (cf. [6]).
As we have to count all transitions caused by an arrival including the overflow events that do not change the state of the underlying Markov chain $Z(t)$, we use the flip-flop marking technique described by Melamed [19, p. 126]. By this means, we mark all transitions of the Markov chain $Z(t)$ corresponding to arrival instants. Following Melamed's approach and using his notation, we see that the arrival-stationary distribution of the number of customers in the system is given by the term

$$P_r^{(0)} = \Psi^-([r]) = \frac{(\pi_{[r]})^t \; \tilde{Q}_{rr+1} \; e}{\sum_{j=0}^{m} (\pi_{[j]})^t \; \tilde{Q}_{jj+1} \; e}, \tag{6}$$

where we set $\tilde{Q}_{mm+1} = I \otimes \Lambda$. Relation (6) states that the arrival-stationary probability $P_r^{(0)}$ coincides with the ratio of the stochastic intensity of those arrival instants leaving R-lump $[r]$ and the total intensity of all arrivals in the system. In the denominator we have to count all arrivals including those that find the system occupied and overflow. Hence, the call congestion of the model is determined by

$$P_m^{(0)} = \frac{(\pi_{[m]})^t \, \tilde{Q}_{mm+1} \, e}{\sum_{j=0}^{m} (\pi_{[j]})^t \, \tilde{Q}_{jj+1} \, e} .$$

Furthermore, it is evident that the stream of lost calls is also an MMPP.

5.3 Actual waiting-time distribution of the customers

The actual waiting-time distribution of the customers is an important performance measure of the MMPP/PH/n/m delay-loss system. Subsequently, we assume that the system is in steady state and note that the actual waiting times of the customers are identically distributed random variables.

Let $W^{(0)}$ denote the waiting time observed by an arriving customer. Let $R^{(0)}$ be the number of customers and $H^{(0)}$ the phase vector of the service process, both seen in the system at the arrival instant of a customer. $W(x)$ denotes the conditional probability that an arriving customer has to wait at most x time units until he is served, provided that he can enter the system and has to wait, i.e., $W(x) = \text{Prob}\{W^{(0)} \leq x \mid n \leq R^{(0)} < m\}$.

Let us now consider an arriving customer who finds $R^{(0)} = r \in [n, m)$ customers in the system and the service process in state $H^{(0)} = H = (h_1, \ldots, h_k)$. Before his service can start, he has to wait until the remaining service time of the fastest of the n customers in service has elapsed, all other $r - n$ customers in front of him in the waiting line have entered the service facility and one server becomes idle again. We note that all n parallel servers of the system are governed by a service-time distribution of PH-type (β, T) and further arrivals are not taken into account. Hence, the actual waiting time S_{r-n} of the tagged customer coincides with the time to serve $r - n + 1$ customers in the system provided that the service process was started in state $H = (h_1, \ldots, h_k)$. Therefore, the distribution of S_{r-n} can be computed as time until absorption in a system with n independent, parallel Markovian service processes and $r - n$ customers in the queue applying Ramaswami's algorithms (cf. [23, p. 399], [22]) since only the numbers of servers in each phase of the service process have to be recorded. Hence, for $r > n$ S_{r-n} follows a PH-distribution $(\phi_{(r,H)}, L_{(r,H)})$ of order $(r - n + 1)\binom{n+k-1}{n}$. It comprises the probability vector $\phi^t_{(r,H)} = (e^t_{p(H)}, 0)$ and the regular Metzler-Leontief matrix

$$L_{(r,H)} = \begin{pmatrix} Q(n,T) & Q(n,T^0\beta^t) & 0 & \ldots & 0 \\ 0 & Q(n,T) & Q(n,T^0\beta^t) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ldots & \ddots & Q(n,T) & Q(n,T^0\beta^t) \\ 0 & \ldots & \ldots & 0 & Q(n,T) \end{pmatrix}$$

with $r-n+1$ blocks $Q(n,T)$ along the diagonal. Here, $e_{p(H)} \in \mathbb{R}^{\tilde{d}_n}$ with $\tilde{d}_n = \binom{n+k-1}{n}$ denotes the $p(H)$th unit vector and $p(H)$ is the position index of the vector H for the reverse lexicographical ordering, i.e.,

$$(e_{p(H)})_l = \begin{cases} 1 & \text{for } l = p(H) = \sum_{i=1}^{k-1} \binom{n-(h_1+\ldots+h_i+1)+k-i}{n-(h_1+\ldots+h_i+1)} + 1 \\ 0 & \text{otherwise} \end{cases} .$$

The matrices $Q(n,T), Q(n,T^0\beta^t) \in \mathbb{R}^{\tilde{d}_n \times \tilde{d}_n}$ are computed by Ramaswami's algorithms applying [22, Theorem 2] to calculate their diagonal elements.
For $r = n$ no customer is waiting in the line and the tagged customer has to wait until the fastest customer in the service facility leaves system. Hence, S_0 follows a PH-distribution $(\phi_{(n,H)}, L_{(n,H)})$ of order $\tilde{d}_n$ with $\phi_{(n,H)} = e_{p(H)}$ and $L_{(n,H)} = Q(n,T)$.
Obviously, the actual waiting-time distribution of a customer is a mixture of the PH-distributions in the family

$$\begin{aligned} \mathcal{L} &= \{(\phi_{(r,H)}, L_{(r,H)}) \mid \forall r, H : n \le r < m, 0 \le H = (h_1, \ldots, h_k) \text{ s.t. } \sum_{i=1}^{k} h_i = n\} \\ &\equiv \{(\phi_l, L_l) \mid \exists l : 1 \le l \le w = \binom{n+k-1}{n} (m-n) \text{ s.t. } (\phi_l, L_l) = (\phi_{(r,H)}, L_{(r,H)})\} . \end{aligned}$$

Here, we order the $w = \binom{n+k-1}{n} \cdot (m-n)$ different tuples (r, H) in $\mathcal{L}$ in such a way that the first components r are arranged in a descending order and the second components H follow the ordering specified in S. Then we enumerate the PH-distributions in $\mathcal{L}$ according to this sequence of indices, i.e., $((m-1, H_1), \ldots, (m-1, H_{\tilde{d}_n}), \ldots (n, H_1), \ldots, (n, H_{\tilde{d}_n})) \equiv (1, \ldots, w)$.
The probability vector $m = (m_{(r,H)})_{(r,H)}$ of this mixture has the components $m_{(r,H)} = K \cdot P^{(0)}_{(r,H)}$. Here, K is a normalization constant and

$$P^{(0)}_{(r,H)} = \Psi^-([(r,H)]) = \frac{(\pi_{[(r,H)]})^t \, \tilde{Q}_{[(r,H)],[r+1]} \, e}{\sum_{\forall (j,H)} (\pi_{[(j,H)]})^t \, \tilde{Q}_{[(j,H)],[j+1]} \, e} \tag{7}$$

is the arrival-stationary probability that an arriving customers finds r customers in the system and the service process in state $H = (h_1, \ldots, h_k)$. It is calculated according to Melamed's approach [19]. In (7) $\tilde{Q}_{[(j,H)],[j+1]}$ denotes the submatrix of the generator $\tilde{Q}$ on the H_j-lump $[(j,H)]$ and the R-lump $[j+1]$. $\pi_{[(j,H)]}$ is the vector of time-stationary probabilities corresponding to this H_j-lump. Obviously, $K^{-1} = \sum_{r=n}^{m-1} \sum_{\forall H : H_r \in [r]} P^{(0)}_{(r,H)}$ holds. Regarding $\tilde{Q}_{jj+1} = I \otimes \Lambda$ for $j \ge n$, we conclude that

$$m_{(r,H)} = K P^{(0)}_{(r,H)} = \frac{\sum_{y=1}^{s} \pi_{(r,H,y)} \, \lambda_y}{\sum_{j=n}^{m-1} \sum_{\forall H : H_j \in [j]} \sum_{y=1}^{s} \pi_{(j,H,y)} \, \lambda_y}$$

holds for the probabilities of the mixture.
As the finite mixture of PH-distributions is again a PH-distribution (cf. [20, Theorem 2.2.4, p. 53]), the actual waiting-time distribution W coincides with a PH-distribution $(\alpha, \widetilde{W})$. The vector $\alpha = (\alpha_l)_{1 \le l \le w}$ is defined by

$$\alpha_l^t = \alpha_{(r,H)}^t = m_{(r,H)} \phi_{(r,H)}^t = K P^{(0)}_{(r,H)} \left(e^t_{p(H)} \cdot \delta_{rn} + (1 - \delta_{rn}) \cdot (e^t_{p(H)}, 0) \right)$$

where $l = p((r,H)) = \binom{n+k-1}{n} \cdot (m-1-r) + \sum_{i=1}^{k-1} \binom{n-(h_1+\ldots+h_i+1)+k-i}{n-(h_1+\ldots+h_i+1)} + 1$ is the position index of the tuple (r,H) in the specified ordering and δ_{rn} is 1 for $r = n$ and 0 otherwise. The representation matrix is given by

$$\widetilde{W} = \begin{pmatrix} L_1 & 0 & \ldots & 0 \\ 0 & L_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & L_w \end{pmatrix}.$$

Obviously, all $L_{(r,H)}$ are submatrices of the matrix $\widehat{W} = L_{(m-1,H)}$ with the largest order $\hat{w} = (m-n) \cdot \binom{n+k-1}{n}$. To describe this feature, we define a tailoring operator τ in the following way: for $1 \leq j \leq m-n-1$ let $\tau^j(\widehat{W})$ be the matrix obtained from $\widehat{W}$ by deleting the first j block rows and columns and let τ^0 coincide with the identity. As all matrices $L_{(r,H)}$ of the mixture have an upper triangular block structure, it is possible to improve the representation of the distribution W by exploiting the relationship

$$(0, \phi^t_{(r,H)}) \cdot e^{L_{(m-1,H)}\, x} = \phi^t_{(r,H)} \cdot e^{L_{(r,H)}\, x}$$

for $r < m-1$.
Hence, the actual waiting-time distribution W coincides with the PH-distribution $(\hat{\alpha}, \widehat{W})$ defined by $\widehat{W} = L_{(m-1,H)}$ and

$$\hat{\alpha}^t = (m_{(m-1,H_1)}, \ldots, m_{(m-1,H_{d_n})}, \ldots, m_{(n,H_1)}, \ldots, m_{(n,H_{d_n})}) .$$

It may be represented in the form

$$\begin{aligned} W(x) &= 1 - \sum_{r=n}^{m-1} \sum_{i=1}^{d_n} \alpha^t_{(r,H_i)} \cdot e^{\tau^{m-1-r}(\widehat{W})\, x} \cdot e \\ &\stackrel{!}{=} 1 - \hat{\alpha}^t \cdot e^{\widehat{W} x} \cdot e . \end{aligned} \tag{8}$$

If the service times are exponentially distributed, it can be shown that the actual waiting-time distribution is a mixture of Erlang distributions with 1 to $m-n$ phases (cf. [18]).

6 Investigation of some MMPP/PH/n/m models

In the tool MACOM PH/PH/n/m and MMPP/PH/n/m systems are used as generic elements of the queueing networks to model parts of a B-ISDN at the connection level or policing algorithms in these networks such as the rectangle sliding window technique (cf. [32], [5]). Hereby, the question arises whether arrival and service processes with a small number of phases are sufficient to represent the behavior of basic elements in broadband communication networks. As the complexity of a model dramatically increases with the number of phases, it is necessary to use processes with a small number to limit the efforts of analysis. On the other hand, there are some insensitivity results (cf. [35], [7]) which give rise to conjecture that the major performance measures of an MMPP/PH/n/m model such as the time and call congestion are not very sensitive to deviations in the structures of the service and arrival processes. To investigate the sensitivity of the model characteristics, it is necessary to compute the performance measures of MMPP/PH/n/m systems with varying parameters. This task can be performed very efficiently by means of MACOM.

In the following, we restrict our attention to special cases of the SMP+M/PH/n/m model, namely, variants of an SMP+M/M/n/n loss system. They represent simplified versions of a digital transmission link for data and packetized voice traffic that disregard the different holding time characteristics, bandwidth requirements and traffic handling (cf. [27, p. 661ff]). The SMP+M/M/n/n loss system was investigated by Willie [35] who derived a very interesting insensitivity property of the model. Willie proved that the call congestion of the Semi-Markovian point process coincides with the congestion rate of an associated renewal process in the GI+M/M/n/n loss systems if $n = 1$ holds and the interarrival-time distribution of the renewal stream is given by the generic interarrival-time distribution of the SMP (cf. [35]).

We consider two variants of this loss system, the MMPP+M/C_2/n/n and PH+M/C_2/n/n system, respectively. In the first case, we suppose that the MMPP arrival stream is a superposition of two independent point processes, namely, a Poisson process with rate λ_M and a Markov-modulated Poisson process with generator matrix $Q \in \mathbb{R}^{s\times s}$ and rate vector λ_{MMPP}. They may be considered as simplified descriptions of independent voice and data traffic streams. Then the generic interarrival-time distribution associated with the MMPP is a PH-renewal distribution with representation $(r, Q - \Lambda)$ where r is the steady-state vector corresponding to $(\Lambda - Q)^{-1}\Lambda$ and $\Lambda = \mathrm{diag}(\lambda_{MMPP})$ is the arrival rate matrix.

The service process is governed by a Coxian distribution with two phases (C_2) having the same parameter. The mean service time $1/\mu$ is set to 1. The coefficient of variation of the service time may be varied to study its influence on the call-congestion rates.

Let us consider a loss system with $n = 5$ parallel servers and set $\lambda_M = 5$, $\lambda_{MMPP} = (1.0, 2.0)^t$, $Q = \begin{pmatrix} -0.5 & 0.5 \\ 2.0 & -2.0 \end{pmatrix}$. Then $r^t = (2/3, 1/3)$ follows. The generic interarrival time is governed by a PH-distribution with representation matrix $T = Q - \Lambda = \begin{pmatrix} -1.5 & 0.5 \\ 2.0 & -4.0 \end{pmatrix}$. In Table 2 some results are shown for the related PH+M/C_2/5/5 and MMPP+M/C_2/5/5 models. They illustrate the weak sensitivity of the time- and call-

congestion rates of this model if the coefficient of variation of the service process is modified.
The next example (see Table 3) illustrates that there may be large differences between the time and call congestion. We recall that different MMPP streams offered to a common link observe different call-congestion rates although their superposition is again an MMPP stream. By the way, both examples emphasize the necessity to calculate the individual and average call-congestion rates of different traffic streams in B-ISDN models. Furthermore, they show that the approximation of these rates by the time congestion is impossible.

7 Conclusion

In this paper we have discussed modeling and analysis of a B-ISDN by Markovian queueing networks employing a decomposition approach and numerical solution methods for Markov chains. First, we have briefly sketched the concepts of a computational approach for modeling and analysis of such connection-oriented communication systems with adaptive routing based on advanced Markovian techniques. They have been implemented by the software tool MACOM. Variants of the PH/PH/n/m and MMPP/PH/n/m delay-loss systems constitute the generic elements of its model world.

Then we have investigated the simplest generic model of a communication link in B-ISDN, namely, the MMPP/PH/n/m model. After describing the features of this multi-server system, the generator matrix associated with the underlying Markov chain was constructed and its properties were studied. We have developed new, convergent block iterative schemes based on R-regular splittings of the generator matrix. They can be used to calculate the steady-state distribution corresponding to the basic Markov chain of the model. Based on these steady-state probabilities, formulas for the time- and arrival-stationary distributions of the number of customers in the system have been derived including the time and call congestion. Furthermore, the actual waiting-time distribution of a customer who has to wait after entering the system was stated, too.
The presented block iterative schemes may be incorporated in MACOM as new block iterative solution methods for MMPP/PH/n/m models. They can also be used as building blocks of a network analysis method that implements a decomposition approach based on Semi-Markovian techniques. It is worthwhile to mention that the proposed algorithms are most suitable for an implementation on a vector processor.
Finally, we have investigated some variants of the MMPP/PH/n/n loss system and pointed out the relevance of the corresponding results in B-ISDN modeling.

Acknowledgment
The project MACOM has considerably benefitted from the contributions of M. Sczittnick, S. Zäske and B. Müller-Clostermann. The authors wish to express their appreciation to these colleagues working at the department of computer science in Dortmund.

References

[1] G. P. Barker and S. J. Yang. Semi-iterative and iterative methods for singular M-matrices. *SIAM J. Matrix Anal. Appl.*, 9(2), 168–180, 1988.

[2] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Academic Press, New York, 1979.

[3] C. Blondia. The N/G/1 finite capacity queue. *Comm. Statist.-Stochastic Models*, 5(2), 273–294, 1989.

[4] P. P. Bocharov. Queueing system of limited capacity with state dependent distributions of phase type. *Autom. Remote Control*, 46(10), 1229–1236, 1985.

[5] L. Dittmann, S. Jacobsen, and K. Moth. Flow enforcement algorithms for ATM networks. *IEEE Journal on Selected Areas in Communications*, 9(3), 343–350, 1991.

[6] P. Franken, D. König, U. Arndt, and V. Schmidt. *Queues and Point Processes.* John Wiley, New York, 1982.

[7] B. W. Gnedenko and D. König. *Handbuch der Bedienungstheorie II.* Akademie-Verlag, Berlin, 1984.

[8] B. Hajek. Birth-and-death processes on the integers with phases and general boundaries. *Journal of Applied Probability*, 19, 488–499, 1982.

[9] H. Heffes and D. M. Lucantoni. A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance. *IEEE Journal on Selected Areas in Communications*, 4(6), 856–868, 1986.

[10] D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research.* Volume I: Stochastic Processes and Operating Characteristics, McGraw-Hill, New York, 1982.

[11] F. P. Kelly. *Reversibility and Stochastic Networks.* John Wiley, New York, 1979.

[12] J. G. Kemeny and L. J. Snell. *Finite Markov Chains.* Springer, New York, 1976.

[13] U. Krieger, B. Müller-Clostermann, and M. Sczittnick. Modeling and analysis of communication systems based on computational methods for Markov chains. *IEEE Journal on Selected Areas in Communications*, 8(9), 1630–1648, 1990.

[14] U. Krieger and M. Sczittnick. A Markovian approach for modelling and analysis of advanced telecommunication networks. In A. Jensen and V.B. Iversen, editors, *Teletraffic and Datatraffic in a Period of Change, Proc. ITC 13*, pp. 717–722, North-Holland, Amsterdam, 1991.

[15] P. J. Kühn. Approximate analysis of general queuing networks by decomposition. *IEEE Trans. on Communications*, 27(1), 113–126, 1979.

[16] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Comm. Statist.-Stochastic Models*, 7(1), 1–46, 1991.

[17] R. Marie. Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queues. *Performance Evaluation Review*, 9, 117–125, 1980.

n	m	k	d/s
5	10	5	882
5	10	10	13013
10	10	5	3003
10	10	10	184756
20	200	5	1965810

Table 1: Order d of the generator matrix

Service process	PH-model			MMPP-model		
Coefficient of variation	Congestion rates of stream			Congestion rates of stream		
	M	PH	PH+M	M	MMPP	MMPP+M
0.80	0.373672	0.378163	0.374542	0.373571	0.378927	0.374608
1.00	0.373680	0.377989	0.374515	0.373586	0.378726	0.374581
1.20	0.373685	0.377873	0.374495	0.373596	0.378578	0.374560
1.40	0.373687	0.377794	0.374482	0.373603	0.378470	0.374545
1.60	0.373689	0.377738	0.374473	0.373608	0.378390	0.374532

Table 2: Comparison of the congestion rates of an $MMPP+M/C_2/5/5$ and its related $PH+M/C_2/5/5$ model with generic interarrival-time distribution of the MMPP. The mean service time is $1/\mu = 1$. The MMPP arrival rate is 1.2, the arrival rate of the Poisson process $\lambda_M = 5$.

Service process - Coeff. of variation	Congestion rates	
	Time congestion	Call congestion
0.8	4.26109e-01	6.05973e-01
1.0	4.28444e-01	6.04779e-01
1.2	4.30154e-01	6.03855e-01
1.4	4.31403e-01	6.03153e-01
1.6	4.32325e-01	6.02619e-01
1.8	4.33016e-01	6.02209e-01
2.0	4.33543e-01	6.01891e-01

Table 3: Comparison of the time and call congestion associated with an $\mathrm{MMPP}/C_2/5/5$ loss system varying the coefficient of variation of the service time. The service-time distribution is a Coxian distribution with 2 phases, equal rates and mean $1/\mu = 1$. The MMPP possesses the arrival rate vector $\lambda^t = (5.0, 30.0)$ and its generator matrix $Q \in \mathbb{R}^{2\times 2}$ is determined by the elements $Q_{12} = 0.5, Q_{21} = 2.0$.