A Birth-Death Approximation for a Fluid Source in a Token Bucket Model[*]

ANDREAS BRANDT
Institut für Operations Research, Humboldt-Universität zu Berlin,
Spandauer Str. 1, D-10178 Berlin, Germany, Email: brandt@wiwi.hu-berlin.de

MANFRED BRANDT
Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB),
Takustr. 7, D-14195 Berlin, Germany, Email: brandt@zib.de

Berlin, August 2004

1 Introduction and model description

In the third generation mobile networks (UMTS) which base on IP protocols, real time services like speech and moving pictures are important applications. For ensuring an appropriate quality of service for such services, there is an admission control necessary, rejecting service requests of users if their acceptance probably would imply an overload congestion for the network. Several models for packet streams, data handling systems and admission control strategies have been investigated. In particular, the superposition of many on-off sources, leading to bursty packet streams, is often modeled and approximated by certain stochastic processes, in particular by Markov-modulated rate processes (fluid approximation), cf. e.g. [AMS], [SE], [K], [BB], [EM], [R] and the references therein. Admission control strategies often base on monitoring the actual packet arrival rate at the system: if too "many" packets arrive, then a control mechanism regulates the arrival rate. Token bucket algorithms are used for preventing overload for the processors handling the packets. Various kinds of stochastic processes and queueing models including fluid models are extensively used for modeling and analyzing packet handling mechanisms, cf. e.g. [IKKM], [ADRS], [AR], [ARK], [AS], [BBS], [EM], [LP], [PVL] and the references therein.

In this paper we consider a node with a processor where arriving requests are rejected if the actual packet arrival rate exceeds a certain level, and overload for the processor is prevented by a token bucket algorithm. The

[*] This work was partially supported by a grant from the Siemens AG.

model is as follows: At a node of a network there arrives a Poisson process of requests (session activation requests) of intensity λ from outside. Each request accepted by the node – the admission strategy will be described below – generates a geometrically with parameter $q \in [0, 1)$ distributed positive number C of on-off cycles, cf. Figure 1.1. The on and off periods are exponentially distributed with mean $1/\alpha$ and $1/\beta$, respectively. A request generates in mean $1/(1 - q)$ on periods, and during the on periods packets are generated at constant rate r. The cumulative packet arrival process, generated by the requests which are in an on period, i.e. active, has to be served by a processor.

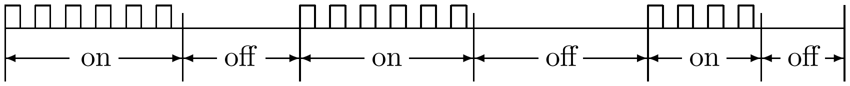

Figure 1.1: *Packet stream generated by a request consisting of $C = 3$ on-off cycles, i.e., the request induces three on periods.*

Since the on and off periods of accepted requests change dynamically in time and independent of each other the actual packet arrival rate varies considerably, and hence it may induce congestion of the system. For preventing session breakdowns – due to limited capacity – the following easy to implement *admission control strategy for the requests* is considered: An arriving new request (session activation request) is accepted iff the actual packet arrival rate, generated by the accepted requests which are in an on period, i.e. active, is not larger than a given threshold r^*. However, although a reasonable choice of r^* will limit overflow in some sense, the admission control based on the number of active requests cannot prevent overload for the system. Note that the number of active requests is dynamically, and it may happen that new requests are accepted because only few requests are active, but some times later more requests in the node may become active leading to a packet rate exceeding the system capacity. The following *token bucket algorithm for the packets*, cf. Figure 1.2, prevents overload of the system: At constant rate μ tokens arrive at a token bucket of capacity b. Arriving tokens which find the bucket full get lost. An arriving packet is paired with a token from the token bucket – if there is any – and then served by the system. An arriving packet which cannot be paired with a token is discarded, i.e. gets lost. For this system the probability that an arriving new request will not be accepted by the node, the mean rate of packets generated by requests accepted by the node and the packet loss probability due to an empty token bucket are of interest.

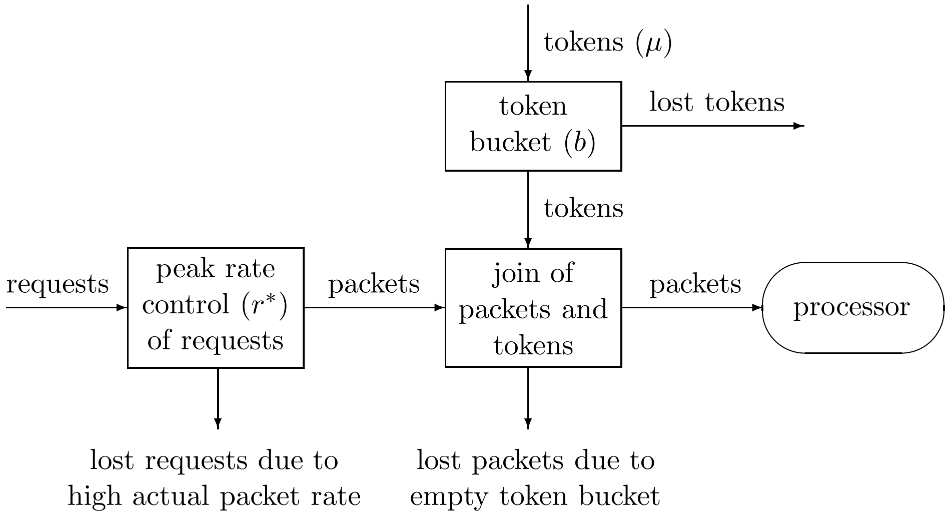
tokens (μ)
token bucket (b)
lost tokens
tokens
requests
peak rate control (r^*) of requests
packets
join of packets and tokens
packets
processor
lost requests due to high actual packet rate
lost packets due to empty token bucket

Figure 1.2: *Node with processor: the requests are controlled by a peak rate mechanism and the packets by a token bucket algorithm.*

Remark 1.1 *The last off period generated by a request does not affect the dynamics of the system and hence can be neglected. In other words, an accepted request is equivalently characterized by a geometrically with parameter q distributed positive number C of on periods and $C-1$ off periods, where the requests start and end with an on period, cf. Figure 1.3. In particular, a request is determined by the number and durations of the on periods and the durations of the off periods between the on periods.*

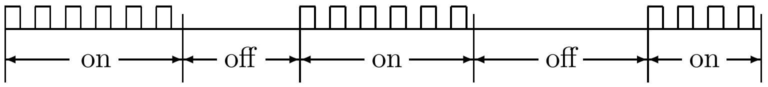

Figure 1.3: *Packet stream generated by a request consisting of $C = 3$ on periods and $C - 1 = 2$ off periods.*

The paper is organized as follows. For obtaining approximations of reasonable complexity for the performance measures of interest we introduce the fluid flow approximation in Section 2.1. The fluid flow source is modeled via two different two-node networks with state-dependent arrival rate, but for the parameters of interest the balance equations for the stationary occupancy distribution in these networks cannot be solved numerically due to their complexity. Thus in Section 2.2 we consider some limiting cases, which are the basis for an approximation of the dynamics of active requests by a birth-death process constructed in Section 3. Theorem 3.1 tells us that the induced approximation for the stationary distribution of the number of active requests is exact in these limiting cases. Within the birth-death process approximation the request loss probability, the mean packet arrival rate and the packet loss probability in case of $b = 0$ can be computed efficiently. In Section 4.1 we propose $E[(A - b)_+]/EF$ as an approximation for the packet loss probability in the general case, where A is the amount of arriving fluid during two successive time instants where the packet arrival rate hits μ from below, which has to be paired from the token buffer, and F is the total amount of arriving fluid in this time interval. For applying this approximation it remains to evaluate $E[\min(A, b)]$. In Section 4.2 we approximate $E[\min(A, b)]$ by linear combinations of the LST $A^*(s)$ of A at some $s \in \mathbb{R}_+$, and $A^*(s)$ is given as a continued fraction in Section 4.3.

2 Approximation of the packet arrival stream in the fluid flow model

2.1 Fluid flow sources: modeling via two-node networks

According to the fluid flow approximation, arriving packets during an on period are approximated by a fluid with rate r, i.e., the discrete nature of the packets is ignored. There are several reasons for the attraction of fluid models, cf. [IKKM] p. 87: the small and uniform packet size; the constant inter-arrival time between packets in an on period (burst) fits naturally in the fluid framework, and it is difficult to handle in the queueing context; the complexity of numerically solving and of simulating fluid models is considerably less than for similar queueing models. Further, the fluid approximation presumes separation of time scales – note that the inter-arrival time of packets is small with respect to on periods (bursts). For comparing performance measures of packet models obtained by simulation and solution of fluid models, cf. e.g. [EM].

Let $N_1(t)$ and $N_2(t)$ be the number of requests in the system at time $t+0$ and which are in an on and in an off period, respectively. The admission control strategy for the requests in the fluid flow model approximation then reads as: an arriving new request at time t is accepted iff $N_1(t-0) \leq n^*$, where $n^* := \lfloor r^*/r \rfloor$. In view of the Poisson arrival assumption for new requests, the exponentially distributed on and off times, the geometrically distributed number of on periods and the independence assumptions, the process $(N_1(t), N_2(t))$, $t \in \mathbb{R}$, is a Markov process with state space $\mathbb{Z}_+^2$. The dynamics of $(N_1(t), N_2(t))$ correspond to the dynamics of a two-node network with state-dependent arrival rate at the first node, given in Figure 2.1.

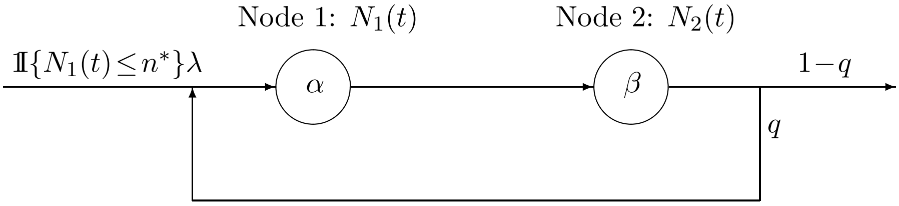

Figure 2.1: *Markov-modulated fluid flow source as a two-node network with state-dependent arrival rate.*

The two nodes in Figure 2.1 are infinite server systems with exponential service times with parameter α in the first and parameter β in the second node. The service times represent the durations of the on and off periods, respectively. At Node 1 requests arrive from outside according to the state-dependent arrival intensity $1\!\!1\{n \leq n^*\}\lambda$ if there are n requests in Node 1, modeling the admission control strategy for arriving new requests. After leaving the second node, with probability q the request is transferred to the first node again and with probability $1-q$ it leaves the network. The mean sojourn time EV of an accepted request in this two-node network is

$$EV = \frac{\alpha+\beta}{\alpha\beta(1-q)}\,. \tag{2.1}$$

The fluid arrival rate $R(t) := rN_1(t)$ approximates the actual packet arrival rate of the system. Note that $R(t)$ corresponds to a Markov-modulated fluid flow source. The two-node network is of non-product type. Hence numerical algorithms and approximations for relevant performance measures will be developed in the following.

The assumptions of the model imply that $(N_1(t), N_2(t))$ is an irreducible Markov process with state space $\mathbb{Z}_+^2$ whose stationary distribution

$$p(n_1, n_2) := \lim_{t\to\infty} P(N_1(t)=n_1, N_2(t)=n_2)\,, \quad (n_1, n_2) \in \mathbb{Z}_+^2\,,$$

exists for an arbitrary set of parameters λ, α, $\beta > 0$, $q \in [0,1)$ and $n^* \in \mathbb{Z}_+$. In the following we assume that $(N_1(t), N_2(t))$ is a stationary process. The balance equations for $p(n_1,n_2)$ read

$$\begin{aligned}
&(\mathbb{1}\{n_1 \le n^*\}\lambda + n_1\alpha + n_2\beta)p(n_1,n_2) \\
&\quad = \mathbb{1}\{n_1{-}1 \le n^*\}\lambda p(n_1{-}1,n_2) + (n_1{+}1)\alpha p(n_1{+}1,n_2{-}1) \\
&\qquad + (n_2{+}1)\beta q p(n_1{-}1,n_2{+}1) + (n_2{+}1)\beta(1{-}q)p(n_1,n_2{+}1)\,, \\
&\qquad\qquad (n_1,n_2) \in \mathbb{Z}_+^2\,, \qquad (2.2)
\end{aligned}$$

where $p(n_1,n_2) := 0$ for $(n_1,n_2) \in \mathbb{Z}^2 \setminus \mathbb{Z}_+^2$. The normalizing condition reads

$$\sum_{(n_1,n_2)\in\mathbb{Z}_+^2} p(n_1,n_2) = 1\,. \qquad (2.3)$$

In the limiting case of $n^* = \infty$, i.e., if there is no admission control for arriving new requests, the resulting two-node network has the product form solution

$$p(n_1,n_2) = p_1(n_1)\,p_2(n_2)\,, \quad (n_1,n_2) \in \mathbb{Z}_+^2\,, \qquad (2.4)$$

where

$$p_i(n) := e^{-\varrho_i}\,\frac{\varrho_i^n}{n!}\,, \quad n \in \mathbb{Z}_+\,, \quad i \in \{1,2\}\,, \qquad (2.5)$$

$$\varrho_1 := EN_1(t) = \frac{\lambda}{\alpha(1{-}q)}\,, \quad \varrho_2 := EN_2(t) = \frac{\lambda}{\beta(1{-}q)}\,. \qquad (2.6)$$

Note that ϱ_i can be interpreted as the offered traffic intensity for the i-th node. Eqs. (2.4), (2.5) imply that $N_1(t)$ and $N_2(t)$ are stochastically independent and Poisson-distributed with parameters ϱ_1 and ϱ_2, respectively.

Denote by

$$p(n) := P(N_1(t){=}n) = \sum_{n_2=0}^{\infty} p(n,n_2)\,, \quad n \in \mathbb{Z}_+\,, \qquad (2.7)$$

the distribution of the number $N_1(t)$ of active requests, i.e., which are in an on period. Taking into account the PASTA property, the probability $p_{\ell,r}$ that an arriving new request will not be accepted, i.e. gets lost, is given by

$$p_{\ell,r} = \sum_{n=n^*+1}^{\infty} p(n)\,. \qquad (2.8)$$

The conservation principle applied to Node 1 yields immediately

$$\frac{\lambda(1-p_{\ell,r})}{1-q} = \alpha E N_1(t)\,, \tag{2.9}$$

where we used the fact that the mean number of on periods of a request is $1/(1-q)$. Thus the mean rate $ER := rEN_1(t)$ of packets generated by the requests accepted by the system is

$$ER = r\sum_{n=1}^{\infty} np(n) = \frac{r\lambda(1-p_{\ell,r})}{\alpha(1-q)}\,. \tag{2.10}$$

The conservation principle applied to $\{N_1(t)=n\}$ and to the flow between Node 1 and Node 2 yields the following two identities

$$(\mathbb{1}\{n \le n^*\}\lambda + \beta q E[N_2(t) \,|\, N_1(t)=n])p(n) = (n+1)\alpha p(n+1)\,, \qquad n \in \mathbb{Z}_+\,, \tag{2.11}$$

$$\alpha E N_1(t) = \beta E N_2(t)\,. \tag{2.12}$$

Note that (2.9) and (2.12) provide

$$\lambda(1-p_{\ell,r}) = \beta(1-q)EN_2(t)\,, \tag{2.13}$$

which is the conservation law that the intensity of requests accepted by the system is equal to the intensity of requests leaving the system.

In the fluid model, in case of $b > 0$, the buffer content of the token bucket is a continuous random variable governed by the Markov-modulated process $N_1(t)$, cf. Section 4 for details. In general it is a very difficult task – analytically as well as numerically – to determine the buffer content distribution and hence the packet loss probability, cf. e.g. [AS], [IKKM], [DS]. However, in case of $b = 0$, i.e., if there is no buffer for the tokens, in the fluid model the probability $p_{\ell,p}$ that an arriving packet gets lost is the fraction of fluid which exceeds μ and the total arriving fluid*, i.e.,

$$p_{\ell,p} = \lim_{t\to\infty} \frac{\int\limits_0^t (rN_1(t')-\mu)_+ \,\mathrm{d}t'}{\int\limits_0^t rN_1(t')\,\mathrm{d}t'}\,.$$

*In the fluid model the arriving tokens are modeled as a fluid with rate μ.

Dividing the numerator and denominator by t and applying the individual ergodic theorem it follows

$$p_{\ell,p} = \frac{E[(rN_1(t)-\mu)_+]}{E[rN_1(t)]} = \frac{E[(N_1(t)-\tau)_+]}{EN_1(t)}\,, \tag{2.14}$$

where

$$\tau := \mu/r\,. \tag{2.15}$$

From (2.9) and (2.14) we obtain

$$p_{\ell,p} = \frac{\alpha(1-q)}{\lambda(1-p_{\ell,r})}\,E[(N_1(t)-\tau)_+]\,. \tag{2.16}$$

The following theorem gives monotonicity results for the request loss probability and for the mean packet rate.

Theorem 2.1 *The request loss probability $p_{\ell,r}$ is a monotonically decreasing function, the mean packet rate ER a monotonically increasing function of the admission control parameter r^*.*

Proof. Note that n^* is a monotonically increasing function of r^*. We consider the two-node network given in Figure 2.1 and the modified model where n^* is replaced with n^{*+} $(> n^*)$ and correspondingly $p_{\ell,r}$ with $p^+_{\ell,r}$, ER with ER^+ and $(N_1(t), N_2(t))$ with $(N^+_1(t), N^+_2(t))$. Moreover, we consider the modified model with the following modified admission and service discipline as well as with marked and non marked requests: Let $\tilde{N}_1(t)$ and $\tilde{N}_2(t)$ be the number of non marked requests in Node 1 and Node 2 as well as $\tilde{N}^+_1(t)$ and $\tilde{N}^+_2(t)$ be the number of all requests in Node 1 and Node 2 of the latter model at time $t+0$, respectively. A request arriving from outside at time t in this model will be accepted and remains non marked iff $\tilde{N}_1(t-0) \le n^*$, but in case of $\tilde{N}^+_1(t-0) > n^{*+}$ a chosen at random marked request from Node 1 leaves the system immediately while the service of the request arrived from outside at Node 1 starts (preemptive priority for the non marked requests). A request arriving from outside at time t will be accepted and marked iff $\tilde{N}_1(t-0) > n^*$ and $\tilde{N}^+_1(t-0) \le n^{*+}$.

In view of this admission and service discipline, the system dynamics of the non marked requests are the same as in the original model of Figure 2.1 and not influenced by the numbers of marked requests in Node 1 and Node 2. Thus the distribution of $(\tilde{N}_1(t), \tilde{N}_2(t))$ is equal to the distribution of $(N_1(t), N_2(t))$. Moreover, due to the exponential service time in Node 1 and the Bernoulli feedback, the distribution of $(\tilde{N}_1^+(t), \tilde{N}_2^+(t))$ is equal to the distribution of $(N_1^+(t), N_2^+(t))$. Therefore it follows

$$N_1^+(t) \overset{\mathcal{D}}{=} \tilde{N}_1^+(t) \overset{\mathcal{D}}{\geq} \tilde{N}_1(t) \overset{\mathcal{D}}{=} N_1(t)\,, \tag{2.17}$$

and we obtain $ER^+ \geq ER$. Because of

$$\lambda(1-p_{\ell,r}^+) = \alpha(1-q)EN_1^+(t) \geq \alpha(1-q)EN_1(t) = \lambda(1-p_{\ell,r})\,,$$

cf. (2.9) and (2.17), we find $p_{\ell,r}^+ \leq p_{\ell,r}$.

□

Taking into account Remark 1.1, the modified two-node network given in Figure 2.2, cf. also Figure 1.3, describes the dynamics of requests consisting of a positive number (geometrically distributed with parameter q) of on periods (exponentially distributed with parameter α) and off periods (exponentially distributed with parameter β) between them.

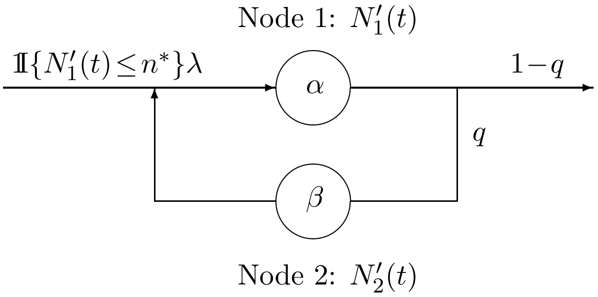

Figure 2.2: *Modified two-node network, where the requests start and end with an on period and $N_1'(t) = N_1(t)$.*

Denoting by $N_1'(t)$ and $N_2'(t)$ the number of requests in the first and second node at time $t+0$, respectively, it holds

$$N_1'(t) = N_1(t)\,.^\dagger \tag{2.18}$$

The stationary distribution $p'(n_1, n_2)$, $(n_1, n_2) \in \mathbb{Z}_+^2$, of the modified network, which is uniquely determined for λ, α, $\beta > 0$, $q \in [0,1)$, $n^* \in \mathbb{Z}_+$,

[†] The corresponding processes can be constructed appropriately.

satisfies the balance equations

$$\begin{aligned}
&(\mathbb{1}\{n_1 \le n^*\}\lambda + n_1\alpha + n_2\beta)p'(n_1,n_2) \\
&\qquad = \mathbb{1}\{n_1{-}1 \le n^*\}\lambda p'(n_1{-}1,n_2) + (n_1{+}1)\alpha q p'(n_1{+}1,n_2{-}1) \\
&\qquad\qquad + (n_1{+}1)\alpha(1{-}q)p'(n_1{+}1,n_2) + (n_2{+}1)\beta p'(n_1{-}1,n_2{+}1)\,, \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (n_1,n_2) \in \mathbb{Z}_+^2\,, \qquad (2.19)
\end{aligned}$$

where $p'(n_1,n_2) := 0$ for $(n_1,n_2) \in \mathbb{Z}^2 \setminus \mathbb{Z}_+^2$, and the normalizing condition

$$\sum_{(n_1,n_2)\in\mathbb{Z}_+^2} p'(n_1,n_2) = 1\,. \qquad (2.20)$$

In view of (2.18), (2.7), it holds

$$p(n) = \sum_{n_2=0}^{\infty} p'(n,n_2)\,, \quad n \in \mathbb{Z}_+\,. \qquad (2.21)$$

The advantage of the modified network is that the solution of (2.19), (2.20) can be computed a bit more efficiently than that of (2.2), (2.3), and hence the performance measures $p_{\ell,r}$ and ER, too, cf. (2.8), (2.10) and (2.21).

Having in mind real life applications where $p(n_1,n_2)$ and $p'(n_1,n_2)$ would be concentrated mainly on $n_1,n_2 \le 10^4$, a linear system of equations with approximately 10^8 variables has to be solved, which is too huge. Thus we are interested in approximations for $p(n)$ which can be computed much more efficiently. The proposed approximations base on limiting cases, which will be considered in the next section.

2.2 Limiting cases

In this section we use the two-node network given in Figure 2.1.

Limiting case $q \to 1$: Binomial model

Let λ, $\alpha(1-q)$, $\beta(1-q)$ and $n^* \in \mathbb{Z}_+$ be fixed. In case of $q \to 1$ (or equivalently $\alpha \to \infty$ or $\beta \to \infty$, respectively) each of the requests in the two-node network will be – when looking at an arbitrary time instant at the system – with probability

$$p = \frac{\beta(1{-}q)}{\alpha(1{-}q) + \beta(1{-}q)} \qquad (2.22)$$

in Node 1, i.e. active, and with probability $1-p$ in Node 2, i.e. passive. Further, for the Laplace-Stieltjes transform of the sojourn time V of an accepted request in the system we obtain

$$\begin{aligned} Ee^{-sV} &= E\Big[E\Big[e^{-sV}\Big|C\Big]\Big] = E\Big[\Big(\frac{\alpha}{\alpha+s}\,\frac{\beta}{\beta+s}\Big)^{C}\,\Big] \\ &= \frac{\alpha\,\beta\,(1-q)}{(\alpha+s)(\beta+s)-q\alpha\beta} \longrightarrow \frac{\gamma}{\gamma+s} \end{aligned}$$

as $q \to 1$, where

$$\gamma := 1/EV = \frac{\alpha(1-q)\beta(1-q)}{\alpha(1-q)+\beta(1-q)} \tag{2.23}$$

is fixed. In the limiting case $q \to 1$ hence the sojourn times are exponentially distributed with parameter γ. Thus for $q \to 1$ the dynamics of the number $J(t) := N_1(t) + N_2(t)$ of requests in the system correspond to a birth-death process with birth rates

$$\lambda_j^{(1)} := \lambda \sum_{n=0}^{n^*} \binom{j}{n} p^n (1-p)^{j-n}\,, \quad j \in \mathbb{Z}_+\,, \tag{2.24}$$

and death rates $j\gamma$, $j \in \mathbb{N} := \mathbb{Z}_+ \setminus \{0\}$. Therefore its stationary distribution $\pi^{(1)}(j) := P(J(t) = j)$ is given by

$$\pi^{(1)}(j) = \pi^{(1)}(0) \prod_{\ell=1}^{j} \frac{\lambda_{\ell-1}^{(1)}}{\ell\gamma}\,, \quad j \in \mathbb{Z}_+\,, \tag{2.25}$$

$$\pi^{(1)}(0) = \Big(\sum_{j=0}^{\infty} \prod_{\ell=1}^{j} \frac{\lambda_{\ell-1}^{(1)}}{\ell\gamma} \Big)^{-1}\,. \tag{2.26}$$

For the stationary distribution of the number of active requests we find

$$p^{(1)}(n) = \sum_{\ell=n}^{\infty} \pi^{(1)}(\ell) \binom{\ell}{n} p^n (1-p)^{\ell-n}\,, \quad n \in \mathbb{Z}_+\,, \tag{2.27}$$

cf. (2.22). Thus taking into account (2.8), (2.10), we obtain

$$p_{\ell,r} = \sum_{n=n^*+1}^{\infty} p^{(1)}(n)\,, \quad ER = r \sum_{n=1}^{\infty} np^{(1)}(n) = \frac{r\lambda(1-p_{\ell,r})}{\alpha(1-q)}\,. \tag{2.28}$$

In view of (2.22), (2.24) and (2.27), for a given number $J(t) = j$ of requests in the system the number of active requests is binomially distributed with parameters j and p. We refer to the birth-death model (2.22)–(2.26) and the binomial sampling (2.27) of on periods as the *binomial model* of the distribution of the number $N_1(t)$ of on periods. Although the distribution $p^{(1)}(n)$, $n \in \mathbb{Z}_+$, in the binomial model corresponds to the distribution of $N_1(t)$ in the limiting case $q \to 1$ where $\alpha(1-q)$ and $\beta(1-q)$ are fixed, the binomial model is well defined for the general model, too. In Section 3 we will use the binomial model as a first approximation for the general model. In the binomial model a conservation law analogous to (2.9) holds.

Lemma 2.1 *For given λ, $\alpha(1-q)$, $\beta(1-q)$ and n^* in the binomial model it holds the conservation law*

$$\lambda \sum_{n=0}^{n^*} p^{(1)}(n) = \alpha(1-q) \sum_{n=1}^{\infty} np^{(1)}(n)\,. \tag{2.29}$$

Proof. From (2.22)–(2.27) we obtain

$$\begin{aligned}
\alpha(1-q) \sum_{n=1}^{\infty} np^{(1)}(n) &= \frac{\gamma}{p} \sum_{n=1}^{\infty} n \sum_{\ell=n}^{\infty} \pi^{(1)}(\ell) \binom{\ell}{n} p^n (1-p)^{\ell-n} \\
&= \sum_{n=1}^{\infty} \sum_{\ell=n}^{\infty} \lambda_{\ell-1}^{(1)} \pi^{(1)}(\ell-1) \binom{\ell-1}{n-1} p^{n-1}(1-p)^{\ell-n} \\
&= \lambda \sum_{m=0}^{n^*} \sum_{\ell=0}^{\infty} \binom{\ell}{m} p^m (1-p)^{\ell-m} \pi^{(1)}(\ell) \sum_{n=0}^{\ell} \binom{\ell}{n} p^n (1-p)^{\ell-n} \\
&= \lambda \sum_{m=0}^{n^*} \sum_{\ell=m}^{\infty} \binom{\ell}{m} p^m (1-p)^{\ell-m} \pi^{(1)}(\ell) = \lambda \sum_{m=0}^{n^*} p^{(1)}(m)\,.
\end{aligned}$$

□

Limiting case $q \to 0$: $M/M/n^* + 1/0$ system

Let λ, α, β, n^* be fixed. For $q \to 0$ the dynamics of the two-node network converge to the dynamics of a two-node series network, cf. Figure 2.3.

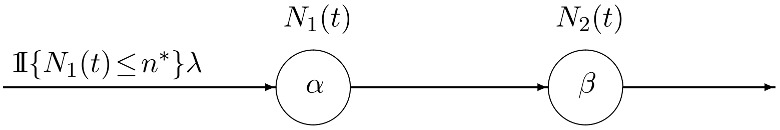

Figure 2.3: *Limiting case $q \to 0$: Two-node series network.*

The dynamics of $N_1(t)$ correspond to the dynamics of a $M/M/n^* + 1/0$ system with arrival rate λ and service rate α. The stationary distribution $p(n) = P(N_1(t) = n)$ is given by

$$p(n) = p(0)\,\frac{1}{n!}\Big(\frac{\lambda}{\alpha}\Big)^n, \quad n = 0, \ldots, n^*{+}1\,, \tag{2.30}$$

$$p(0) = \Big(\sum_{n=0}^{n^*+1} \frac{1}{n!}\Big(\frac{\lambda}{\alpha}\Big)^n\Big)^{-1}, \tag{2.31}$$

and it holds

$$p_{\ell,r} = p(n^*{+}1)\,. \tag{2.32}$$

Limiting case $\beta \to 0$: Single-node model with two arrival streams

Let λ, α, q, n^* be fixed. In case of $\beta \to 0$ the dynamics of active requests converge to the dynamics of a single-node with two arrival streams, cf. Figure 2.4.

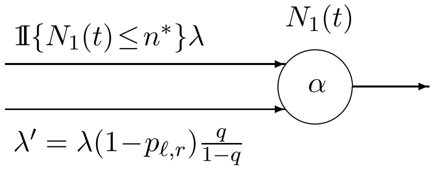
$1\!\!1\{N_1(t)\le n^*\}\lambda$
$N_1(t)$
α
$\lambda' = \lambda(1-p_{\ell,r})\frac{q}{1-q}$

Figure 2.4: *Limiting case $\beta \to 0$: Single-node model with two arrival streams. The first stream has the state-dependent arrival intensity $1\!\!1\{N_1(t) \le n^*\}\lambda$, and the second one is a Poisson process of intensity $\lambda' = \lambda\,(1 - p_{\ell,r})\,q/(1 - q)$.*

The first one is, as in the two-node model, the process of accepted new requests from outside and hence has the state-dependent arrival intensity $1\!\!1\{N_1(t) \le n^*\}\lambda$. As for $\beta \to 0$ the sojourn times in Node 2 of Figure 2.1 converge to infinity and since they are independent of each other, the process of requests routed from Node 2 to Node 1 in Figure 2.1 converges – heuristically – to a Poisson process of some intensity λ'. Since the mean number of revisits of Node 1 by an accepted request is $EC - 1 = q/(1 - q)$, cf. Section 1, and the intensity of the point process of accepted requests from outside is $\lambda(1 - p_{\ell,r})$, we have

$$\lambda' = \lambda(1{-}p_{\ell,r})\,\frac{q}{1{-}q}\,, \tag{2.33}$$

where the probability $p_{\ell,r}$ that an arriving request from outside gets lost is given by

$$p_{\ell,r} = \sum_{n=n^*+1}^{\infty} p(n) \tag{2.34}$$

and $p(n) = P(N_1(t) = n)$, $n \in \mathbb{Z}_+$, denotes the stationary distribution of the number $N_1(t)$ of requests in the single-node model given in Figure 2.4. Since $N_1(t)$ is a birth-death process with birth rates $\mathbb{1}\{n \leq n^*\}\lambda + \lambda'$ and death rates $n\alpha$ the $p(n)$ are given by

$$p(n) = p(0)\,\frac{1}{n!}\Big(\frac{\lambda(1-q\,p_{\ell,r})}{\alpha(1-q)}\Big)^{\min(n,n^*+1)}\Big(\frac{\lambda(1-p_{\ell,r})q}{\alpha(1-q)}\Big)^{(n-n^*-1)_+}, \qquad n \in \mathbb{Z}_+\,, \tag{2.35}$$

$$p(0) = \Big(\sum_{n=0}^{\infty}\frac{1}{n!}\Big(\frac{\lambda(1-q\,p_{\ell,r})}{\alpha(1-q)}\Big)^{\min(n,n^*+1)}\Big(\frac{\lambda(1-p_{\ell,r})q}{\alpha(1-q)}\Big)^{(n-n^*-1)_+}\Big)^{-1}. \tag{2.36}$$

Note that (2.34)–(2.36) yield a fixed point equation for $p_{\ell,r}$.

Limiting case $\beta \to \infty$: $M/M/n^* + 1/0$ system

Let λ, α, q, n^* be fixed. In case of $\beta \to \infty$ the requests leaving Node 1 are fed back with probability q immediately to Node 1, and with probability $1 - q$ they leave the system. The limiting model, cf. Figure 2.5, is a single-node feedback model with state-dependent arrival intensity $\mathbb{1}\{N_1(t) \leq n^*\}\lambda$ of arriving new requests.

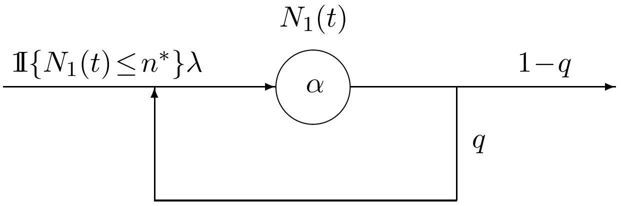

Figure 2.5: *Limiting case $\beta \to \infty$: Single-node feedback model with state-dependent arrival intensity $\mathbb{1}\{N_1(t) \leq n^*\}\lambda$.*

The number $N_1(t)$ of requests in the node is a birth-death process with birth rates $\mathbb{I}\{n \leq n^*\}\lambda$ and death rates $n\alpha(1-q)$. The stationary distribution $p(n)$, $n \in \mathbb{Z}_+$, of $N_1(t)$ is given by

$$p(n) = p(0)\,\frac{1}{n!}\Big(\frac{\lambda}{\alpha(1-q)}\Big)^n, \quad n = 0, \dots, n^*+1, \tag{2.37}$$

$$p(0) = \Big(\sum_{n=0}^{n^*+1} \frac{1}{n!}\Big(\frac{\lambda}{\alpha(1-q)}\Big)^n\Big)^{-1}, \tag{2.38}$$

and it holds

$$p_{\ell,r} = p(n^*+1)\,. \tag{2.39}$$

Note that the dynamics correspond to the dynamics of a $M/M/n^* + 1/0$ system with arrival rate λ and service rate $\alpha(1-q)$.

The time scaling $t \Rightarrow (1-q)t$ provides the birth-death process $\hat{N}_1(t)$ with birth rates $\mathbb{I}\{n \leq n^*\}\lambda/(1-q)$ and death rates $n\alpha$. Note that the stationary occupancy distribution is invariant with respect to time scaling. Further, since the superposition of two independent Poisson processes of intensities λ and $\lambda q/(1-q)$, respectively, is a Poisson process of intensity $\lambda/(1-q)$, the dynamics of $\hat{N}_1(t)$ are equivalent to the dynamics of the single-node model with two state-dependent arrival streams given in Figure 2.6. The splitting into two streams will be used later.

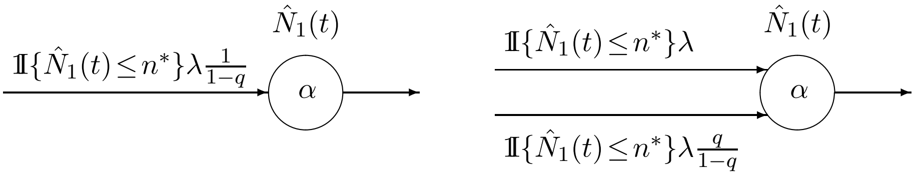
$\hat{N}_1(t)$
$\mathbb{I}\{\hat{N}_1(t) \leq n^*\}\lambda\frac{1}{1-q}$
α
$\hat{N}_1(t)$
$\mathbb{I}\{\hat{N}_1(t) \leq n^*\}\lambda$
α
$\mathbb{I}\{\hat{N}_1(t) \leq n^*\}\lambda\frac{q}{1-q}$

Figure 2.6: *Limiting case $\beta \to \infty$: Single-node model after the time scaling $t \Rightarrow (1-q)t$. The two systems are equivalent.*

Limiting case $n^* \to \infty$: $M/M/\infty$ system

In this case from the product form solution (2.4)–(2.6) for the stationary distribution $p(n) = P(N_1(t) = n)$ we obtain

$$p(n) = e^{-\frac{\lambda}{\alpha(1-q)}}\,\frac{1}{n!}\Big(\frac{\lambda}{\alpha(1-q)}\Big)^n, \quad n \in \mathbb{Z}_+\,. \tag{2.40}$$

Note that $N_1(t)$ is a birth-death process with rates $\lambda/(1-q)$ and $n\alpha$, respectively. The corresponding node is given in Figure 2.7, where the same superposition result is used as in case of $\beta \to \infty$.

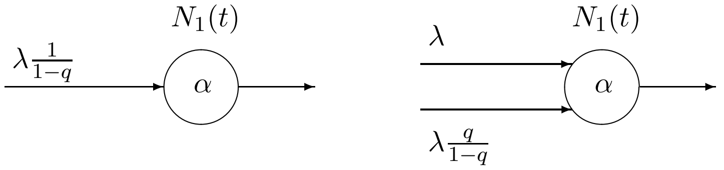

Figure 2.7: *Limiting case $n^* \to \infty$: Two equivalent systems.*

3 Approximation of the dynamics of the number of on periods by a birth-death process

Remember, the dynamics of the number $N_1(t)$ of on periods, i.e. of the number of active requests, in the two-node networks given in Figure 2.1 and Figure 2.2 are Markov-modulated processes. The corresponding underlying Markov processes are two-dimensional and numerically intractable for parameter regions where we are interested in, cf. the end of Section 2.1. In this section we derive an approximation for the dynamics of $N_1(t)$ by a birth-death process, which can be computed efficiently. We proceed in two steps. In the following let λ, α, $\beta > 0$, $q \in [0,1)$ and $n^* \in \mathbb{Z}_+$ be fixed.

Step 1: Fitting of a birth-death process to the binomial approximate model

First let us approximate the stationary distribution $p(n)$, $n \in \mathbb{Z}_+$, of the number $N_1(t)$ of on periods, i.e. of active requests, by the binomial model obtained via the limiting case $q \to 1$: The distribution $p^{(1)}(n)$, $n \in \mathbb{Z}_+$, defined by (2.22)–(2.27) is – in some sense – an approximation of $p(n)$, $n \in \mathbb{Z}_+$, which we call the *binomial approximate model.* Note that for $q \to 1$, if λ, $\alpha(1-q)$, $\beta(1-q)$ and n^* are fixed, the distribution $p(n)$, $n \in \mathbb{Z}_+$, in the original model converges to the distribution $p^{(1)}(n)$, $n \in \mathbb{Z}_+$.

Next we define a birth-death process $N_1^*(t)$ with birth rates λ_n^*, which have to be determined, and death rates $\mu_n := n\alpha$ [‡] such that the stationary distribution $p^*(n)$, $n \in \mathbb{Z}_+$, of $N_1^*(t)$ fits $p^{(1)}(n)$, $n \in \mathbb{Z}_+$, i.e.,

$$p^*(n) = p^{(1)}(n)\,, \quad n \in \mathbb{Z}_+\,. \tag{3.1}$$

From

$$p^*(n) = p^*(0) \prod_{\ell=1}^{n} \frac{\lambda_{\ell-1}^*}{\ell\alpha}\,, \quad n \in \mathbb{Z}_+\,, \quad p^*(0) = \Big(\sum_{n=0}^{\infty} \prod_{\ell=1}^{n} \frac{\lambda_{\ell-1}^*}{\ell\alpha} \Big)^{-1}$$

[‡] If n requests are active then $n\alpha$ is the rate that an on period will end.

and (3.1) it follows

$$\lambda_n^* = (n+1)\alpha\,\frac{p^*(n+1)}{p^*(n)} = (n+1)\alpha\,\frac{p^{(1)}(n+1)}{p^{(1)}(n)}\,, \quad n \in \mathbb{Z}_+\,, \tag{3.2}$$

and thus taking into account (2.27), (2.25), (2.23), (2.22), we obtain

$$\begin{aligned}\lambda_n^* &= (n+1)\alpha\,\frac{\sum\limits_{\ell=n+1}^{\infty}\Big(\prod\limits_{j=1}^{\ell}\frac{\lambda_{j-1}^{(1)}}{j\gamma}\Big)\binom{\ell}{n+1}p^{n+1}(1-p)^{\ell-n-1}}{\sum\limits_{\ell=n}^{\infty}\Big(\prod\limits_{j=1}^{\ell}\frac{\lambda_{j-1}^{(1)}}{j\gamma}\Big)\binom{\ell}{n}p^{n}(1-p)^{\ell-n}} \\ &= \frac{\lambda}{1-q}\,\frac{\sum\limits_{\ell=0}^{\infty}\Big(\prod\limits_{j=0}^{\ell+n}\frac{\lambda_j^{(1)}}{\lambda}\Big)\frac{1}{\ell!}\Big(\frac{\lambda}{\beta(1-q)}\Big)^{\ell}}{\sum\limits_{\ell=0}^{\infty}\Big(\prod\limits_{j=0}^{\ell+n-1}\frac{\lambda_j^{(1)}}{\lambda}\Big)\frac{1}{\ell!}\Big(\frac{\lambda}{\beta(1-q)}\Big)^{\ell}}\,, \quad n \in \mathbb{Z}_+\,. \end{aligned} \tag{3.3}$$

Moreover, shifting the index in the numerator on the r.h.s. of (3.3) provides the representation

$$\lambda_n^* = \frac{\sum\limits_{\ell=0}^{\infty}\beta\ell\Big(\prod\limits_{j=0}^{\ell+n-1}\frac{\lambda_j^{(1)}}{\lambda}\Big)\frac{1}{\ell!}\Big(\frac{\lambda}{\beta(1-q)}\Big)^{\ell}}{\sum\limits_{\ell=0}^{\infty}\Big(\prod\limits_{j=0}^{\ell+n-1}\frac{\lambda_j^{(1)}}{\lambda}\Big)\frac{1}{\ell!}\Big(\frac{\lambda}{\beta(1-q)}\Big)^{\ell}}\,, \quad n \in \mathbb{Z}_+\,, \tag{3.4}$$

where, cf. (2.24), (2.22),

$$\frac{\lambda_j^{(1)}}{\lambda} = \sum_{n=0}^{n^*}\binom{j}{n}p^n(1-p)^{j-n}\,, \quad j \in \mathbb{Z}_+\,, \quad p = \frac{\beta}{\alpha+\beta}\,. \tag{3.5}$$

From (3.5) it follows $0 < \lambda_j^{(1)}/\lambda \le 1$, $j \in \mathbb{Z}_+$, and in view of (3.3) hence it holds

$$0 < \lambda_n^* \le \lambda/(1-q)\,, \quad n \in \mathbb{Z}_+\,. \tag{3.6}$$

Step 2: Modification of the birth-death process $N_1^*(t)$

Note that the birth-death process $N_1^*(t)$ only depends on the parameters $\lambda/(1-q)$, α, β and n^* while the Markov-modulated process $N_1(t)$ additionally depends on q. Thus modifying the birth rates λ_n^* of $N_1^*(t)$ by an additional parameter could improve the approximation. In the limiting cases

$\beta \to 0$ and $\beta \to \infty$ the distribution of $N_1(t)$ corresponds to the distribution of the number of requests in an infinite server system with two arrival streams, cf. Figure 2.4 and Figure 2.6, where the first arrival stream in both cases is a state-dependent arrival process of intensity $\mathbb{1}\{n \le n^*\}\lambda$ and the second process is a Poisson process of intensity $\lambda\,(1-p_{\ell,r})q/(1-q)$ and a state-dependent arrival process of intensity $\mathbb{1}\{n \le n^*\}\lambda q/(1-q)$, respectively. These observations suggest to replace the birth rates λ_n^* of $N_1^*(t)$ with

$$\lambda_n := \mathbb{1}\{n \le n^*\}\lambda + c\lambda_n^*\,, \quad n \in \mathbb{Z}_+\,, \tag{3.7}$$

where $c \in \mathbb{R}_+$ is a parameter which has to be determined. Let $X(t)$ be the corresponding stationary birth-death process with birth rates (3.7) and death rates $\mu_n = n\alpha$. Its distribution $p^{(a)}(n) := P(X(t) = n)$, $n \in \mathbb{Z}_+$, is given by

$$p^{(a)}(n) = p^{(a)}(0) \prod_{\ell=1}^{n} \frac{\lambda_{\ell-1}}{\ell\alpha}\,, \quad n \in \mathbb{Z}_+\,, \quad p^{(a)}(0) = \Big(\sum_{n=0}^{\infty} \prod_{\ell=1}^{n} \frac{\lambda_{\ell-1}}{\ell\alpha}\Big)^{-1}\,. \tag{3.8}$$

Since also for an approximation of $N_1(t)$ the rate of accepted requests should be equal to the rate of requests leaving the system, cf. the conservation law (2.9), we claim for $X(t)$

$$\lambda \sum_{n=0}^{n^*} p^{(a)}(n) = \alpha(1-q) \sum_{n=1}^{\infty} np^{(a)}(n)\,, \tag{3.9}$$

which determines the unknown parameter $c \in \mathbb{R}_+$ in (3.7) uniquely.

Lemma 3.1 *Let λ, α, $\beta > 0$, $q \in [0,\,1)$ and $n^* \in \mathbb{Z}_+$ be given. Then* (3.9) *has a uniquely determined solution $c \in \mathbb{R}_+$.*

Proof. 1. Using (3.8), (3.7), it follows easily that (3.9) is equivalent to

$$\frac{\lambda}{1-q} \sum_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod_{j=0}^{n-1} (\lambda + c\lambda_j^*) = \sum_{n=0}^{\infty} \frac{1}{n!\,\alpha^n} \prod_{j=0}^{n} (\mathbb{1}\{j \le n^*\}\lambda + c\lambda_j^*)\,. \tag{3.10}$$

For fixed λ, $\alpha > 0$, $q \in [0,\,1)$ and λ_j^*, $j \in \mathbb{Z}_+$, where the λ_j^* are bounded because of (3.6), with respect to $c \in \mathbb{R}_+$ the l.h.s. and r.h.s. of (3.10) defines a continuous function $f(c)$ and $g(c)$, respectively, and (3.10) is equivalent to $f(c) = g(c)$. For $c = 0$ we find

$$f(0) = \frac{\lambda}{1-q} \sum_{n=0}^{n^*} \frac{\lambda^n}{n!\,\alpha^n} \ge \sum_{n=0}^{n^*} \frac{\lambda^{n+1}}{n!\,\alpha^n} = g(0)\,. \tag{3.11}$$

On the other hand, in view of $\lambda_n^* > 0$ for $n \in \mathbb{Z}_+$, if c is sufficiently large such that $\lambda/(1-q) \le \lambda + c\lambda_n^*$ for $0 \le n \le n^*$ then it holds $f(c) \le g(c)$. In view of the continuity of $f(c)$ and $g(c)$, thus there exists $c^* \in \mathbb{R}_+$ such that $f(c^*) = g(c^*)$, i.e., c^* is a solution of (3.9).

2. In the following it will be shown that $g(c)/f(c)$ is strictly monotonically increasing for $c \in \mathbb{R}_+$, which provides the uniqueness. It holds

$$\frac{\lambda}{1-q}\,\frac{g(c)}{f(c)} = \alpha\,\frac{\sum\limits_{n=0}^{n^*} \frac{n}{n!\,\alpha^n} \prod\limits_{j=0}^{n-1} (\lambda + c\lambda_j^*)}{\sum\limits_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod\limits_{j=0}^{n-1} (\lambda + c\lambda_j^*)} + \frac{\sum\limits_{n=n^*}^{\infty} \frac{1}{n!\,\alpha^n} \prod\limits_{j=n^*+1}^{n} (c\lambda_j^*)}{\sum\limits_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod\limits_{j=n}^{n^*} (\lambda + c\lambda_j^*)^{-1}}\,.$$

As the second summand on the r.h.s. obviously is strictly monotonically increasing in $c \in \mathbb{R}_+$, it is sufficient to prove that the derivative of

$$h(c) := \frac{\sum\limits_{n=0}^{n^*} \frac{n}{n!\,\alpha^n} \prod\limits_{j=0}^{n-1} (\lambda + c\lambda_j^*)}{\sum\limits_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod\limits_{j=0}^{n-1} (\lambda + c\lambda_j^*)}\,, \quad c \in \mathbb{R}_+\,,$$

is nonnegative. Using the abbreviation $\xi_j := \lambda + c\lambda_j^*$, it follows

$$\begin{aligned}
&h'(c)\Big(\sum_{n=0}^{n^*} \frac{1}{n!\,\alpha^n} \prod_{j=0}^{n-1} \xi_j\Big)^2 \\
&= \Big(\sum_{n=0}^{n^*} \frac{n}{n!\,\alpha^n} \Big(\prod_{j=0}^{n-1} \xi_j\Big)\Big(\sum_{j=0}^{n-1} \frac{\lambda_j^*}{\xi_j}\Big)\Big)\Big(\sum_{m=0}^{n^*} \frac{1}{m!\,\alpha^m} \Big(\prod_{j=0}^{m-1} \xi_j\Big)\Big) \\
&\quad - \Big(\sum_{n=0}^{n^*} \frac{n}{n!\,\alpha^n} \Big(\prod_{j=0}^{n-1} \xi_j\Big)\Big)\Big(\sum_{m=0}^{n^*} \frac{1}{m!\,\alpha^m} \Big(\prod_{j=0}^{m-1} \xi_j\Big)\Big(\sum_{j=0}^{m-1} \frac{\lambda_j^*}{\xi_j}\Big)\Big) \\
&= \sum_{n,m=0}^{n^*} \frac{n}{n!\,\alpha^n\,m!\,\alpha^m} \Big(\prod_{j=0}^{n-1} \xi_j\Big)\Big(\prod_{j=0}^{m-1} \xi_j\Big)\Big(\sum_{j=0}^{n-1} \frac{\lambda_j^*}{\xi_j} - \sum_{j=0}^{m-1} \frac{\lambda_j^*}{\xi_j}\Big) \\
&= \sum_{0 \le m < n \le n^*} \frac{1}{n!\,\alpha^n\,m!\,\alpha^m} \Big(\prod_{j=0}^{n-1} \xi_j\Big)\Big(\prod_{j=0}^{m-1} \xi_j\Big)(n-m)\Big(\sum_{j=m}^{n-1} \frac{\lambda_j^*}{\xi_j}\Big) \ge 0\,.
\end{aligned}$$

□

Summarizing the results up to now, for given λ, α, $\beta > 0$, $q \in [0,1)$ and $n^* \in \mathbb{Z}_+$ we propose to approximate the stationary Markov-modulated process $N_1(t)$ of the number of active requests by a stationary birth-death process $X(t)$ with birth rates λ_n given by (3.7), death rates $\mu_n = n\alpha$, and where c has to be determined by solving (3.9) or (3.10). The parameters λ_n^* are given by (3.3)–(3.5). Note that the birth rates λ_n are bounded, cf. (3.6), (3.7), and that $EX(t)$ is finite, cf. (3.9).

Having in mind (2.8), (2.10), we suggest to approximate $p_{\ell,r}$ and ER by $p_{\ell,r}^{(a)} := P(X(t) \geq n^* + 1)$ and $ER^{(a)} := rEX(t)$, respectively:

$$p_{\ell,r} \approx p_{\ell,r}^{(a)} = \sum_{n=n^*+1}^{\infty} p^{(a)}(n)\,, \tag{3.12}$$

$$ER \approx ER^{(a)} = \frac{r\lambda(1-p_{\ell,r}^{(a)})}{\alpha(1-q)}\,, \tag{3.13}$$

where we used the conservation law (3.9) and (3.12) in (3.13). In case of $b = 0$, i.e., if there is no buffer for the tokens, analogously to (2.14)–(2.16) we suggest to approximate $p_{\ell,p}$ by $p_{\ell,p}^{(a)} := E[(X(t) - \tau)_+]/EX(t)$:

$$p_{\ell,p} \approx p_{\ell,p}^{(a)} = \frac{\alpha(1-q)}{\lambda(1-p_{\ell,r}^{(a)})} \sum_{n=\lceil\tau\rceil}^{\infty} (n-\tau)\, p^{(a)}(n)\,, \tag{3.14}$$

where $\tau = \mu/r$ and $p_{\ell,r}^{(a)}$ is given by (3.12).

The following theorem tells us that in several limiting cases the proposed approximation $p^{(a)}(n)$, $n \in \mathbb{Z}_+$, of $p(n)$, $n \in \mathbb{Z}_+$, is exact. This gives some evidence that $p^{(a)}(n)$, $n \in \mathbb{Z}_+$, $p_{\ell,r}^{(a)}$ and $ER^{(a)}$ will provide good approximations for $p(n)$, $n \in \mathbb{Z}_+$, $p_{\ell,r}$ and ER, respectively.

Theorem 3.1 *In the limiting cases*

(i) $q \to 1$, *where* λ, $\alpha(1-q)$, $\beta(1-q)$ *and* n^* *are fixed;*

(ii) $q \to 0$; *(iii)* $\beta \to 0$; *(iv)* $\beta \to \infty$; *(v)* $n^* \to \infty$

it holds

$$p(n) = p^{(a)}(n)\,, \quad n \in \mathbb{Z}_+\,, \quad p_{\ell,r} = p_{\ell,r}^{(a)}\,, \quad ER = ER^{(a)}\,. \tag{3.15}$$

Proof. (*i*) Remember that for given λ, α, $\beta > 0$, $q \in [0,\, 1)$ and $n^* \in \mathbb{Z}_+$ the stationary distribution $p^{(1)}(n)$, $n \in \mathbb{Z}_+$, of the binomial model given in Section 2.2 is well defined. In case of $q \to 1$, where $\alpha(1-q)$ and $\beta(1-q)$ are fixed, it follows $\alpha \to \infty$, and thus using (3.7), (3.2) we obtain

$$\begin{aligned}\lim_{\alpha\to\infty}\prod_{\ell=1}^{n}\frac{\lambda_{\ell-1}}{\ell\alpha} &= \lim_{\alpha\to\infty}\prod_{\ell=1}^{n}\Big(\frac{1\!\!\mathrm{I}\{\ell-1\le n^*\}\lambda}{\ell\alpha} + c\,\frac{p^{(1)}(\ell)}{p^{(1)}(\ell-1)}\Big)\\ &= \lim_{\alpha\to\infty}\prod_{\ell=1}^{n} c\,\frac{p^{(1)}(\ell)}{p^{(1)}(\ell-1)} = c^n\,\frac{p^{(1)}(n)}{p^{(1)}(0)}\,, \quad n \in \mathbb{Z}_+\,. \qquad (3.16)\end{aligned}$$

From (3.8) we conclude that the conservation law (3.9) is equivalent to

$$\lambda\sum_{n=0}^{n^*}\prod_{\ell=1}^{n}\frac{\lambda_{\ell-1}}{\ell\alpha} = \alpha(1-q)\sum_{n=1}^{\infty} n\prod_{\ell=1}^{n}\frac{\lambda_{\ell-1}}{\ell\alpha}\,.$$

For $q \to 1$ and fixed $\alpha(1-q)$, $\beta(1-q)$ thus from (3.16) we obtain

$$\lambda\sum_{n=0}^{n^*} c^n p^{(1)}(n) = \alpha(1-q)\sum_{n=1}^{\infty} nc^n p^{(1)}(n)\,. \qquad (3.17)$$

From Lemma 2.1 we find that $c = 1$ is a solution of (3.17). Analogously to the second part of the proof of Lemma 3.1 it follows that $c = 1$ is the only nonnegative solution of (3.17). Thus from (3.16) we obtain

$$\frac{p^{(1)}(n)}{p^{(1)}(0)} = \lim_{\alpha\to\infty}\prod_{\ell=1}^{n}\frac{\lambda_{\ell-1}}{\ell\alpha}\,, \quad n \in \mathbb{Z}_+\,. \qquad (3.18)$$

Summing up the r.h.s. of (3.18) over $n \in \mathbb{Z}_+$ and (3.8) yield $p^{(a)}(0) = p^{(1)}(0)$ and hence

$$p^{(a)}(n) = p^{(1)}(n) = p(n)\,, \quad n \in \mathbb{Z}_+\,,$$

as $q \to 1$ where $\alpha(1-q)$, $\beta(1-q)$ are fixed. In view of (2.8), (2.10), (3.12) and (3.13), thus it holds (3.15).

(*ii*) In case of $q \to 0$ using (3.8) and (3.7) from the conservation equation (3.9) we obtain

$$\begin{aligned}\lambda\sum_{n=0}^{n^*} p^{(a)}(n) &= \alpha\sum_{n=0}^{\infty} np^{(a)}(n) = \sum_{n=0}^{\infty}(1\!\!\mathrm{I}\{n\le n^*\}\lambda + c\lambda_n^*)p^{(a)}(n)\\ &= \lambda\sum_{n=0}^{n^*} p^{(a)}(n) + c\sum_{n=0}^{\infty}\lambda_n^*\,p^{(a)}(n)\,,\end{aligned}$$

yielding $c = 0$. Hence it holds $\lambda_n = 1\!\!1\{n \leq n^*\}\lambda$, $n \in \mathbb{Z}_+$, and thus in case of $q \to 0$ the dynamics correspond to those of the original model, cf. Figure 2.3, in this case.

(iii) Let $\ell(\beta)$ be the largest $\ell \in \mathbb{Z}_+$ such that $\beta\ell \leq \lambda_\ell^{(1)}/(1-q)$. In case of $\beta \to 0$ from (3.5) it follows that $\beta\ell(\beta)$ converges to the uniquely determined positive solution λ^* of the fixed point equation

$$x = \frac{\lambda}{1-q}\, e^{-\frac{x}{\alpha}} \sum_{n=0}^{n^*} \frac{1}{n!}\left(\frac{x}{\alpha}\right)^n .$$

As for any $\varepsilon > 0$ and any fixed $n \in \mathbb{Z}_+$ the summands of both series on the r.h.s. of (3.4) become maximal in $U_\varepsilon(\beta) := \mathbb{Z}_+ \cap ((1-\varepsilon)\ell(\beta), (1+\varepsilon)\ell(\beta))$ and the portion of the sum over $U_\varepsilon(\beta)$ is larger than $1-\varepsilon$ for both series if $\beta > 0$ is sufficiently small, it follows that λ_n^* converges to λ^* in case of $\beta \to 0$ for any fixed $n \in \mathbb{Z}_+$. Hence (3.7) reads $\lambda_n = 1\!\!1\{n \leq n^*\}\lambda + c\lambda^*$ in case of $\beta \to 0$, and we obtain the single-node model given in Figure 2.4 where λ' is replaced with $c\lambda^*$. As $c \in \mathbb{R}_+$ is uniquely determined by the conservation law (3.9), cf. the second part of the proof of Lemma 3.1, it follows $c = \lambda'/\lambda^*$, and thus in case of $\beta \to 0$ the dynamics of the proposed approximation correspond to the dynamics of the original model given in Figure 2.4.

(iv) In case of $\beta \to \infty$ from (3.5) we find $p = 1$ and $\lambda_n^{(1)} = 1\!\!1\{n \leq n^*\}\lambda$. Hence (3.3) provides $\lambda_n^* = 1\!\!1\{n \leq n^*\}\lambda/(1-q)$, and thus from (3.7) it follows $\lambda_n = 1\!\!1\{n \leq n^*\}\lambda(1 + c/(1-q))$. From (3.8) and (3.9) we obtain $c = q$ and hence $\lambda_n = 1\!\!1\{n \leq n^*\}\lambda/(1-q)$. Thus in case of $\beta \to \infty$ the dynamics of the proposed approximation correspond to the dynamics of the time-scaled original system, cf. Figure 2.6, and as the stationary distribution $p(n)$, $n \in \mathbb{Z}_+$, is invariant with respect to time scaling it holds (3.15).

(v) In case of $n^* \to \infty$ from (3.5) it follows $\lambda_n^{(1)} = \lambda$. Hence (3.3) provides $\lambda_n^* = \lambda/(1-q)$, and thus from (3.7) we obtain $\lambda_n = \lambda(1 + c/(1-q))$. From (3.8) and (3.9) we find $c = q$ and hence $\lambda_n = \lambda/(1-q)$. Thus in case of $n^* \to \infty$ the dynamics of the proposed approximation correspond to the dynamics of the original model given in Figure 2.7.

□

4 Approximation of the packet loss probability for finite token buckets

In accordance to the fluid flow model for the packet arrival stream described in Section 2.1, we model the tokens arriving with rate μ at the token bucket as fluid, too. Remember that the case of a token bucket with capacity $b = 0$ has already been discussed in Section 2.1 and Section 3. Thus in this section we consider a token bucket of capacity b, where b is a given positive real number, cf. Figure 1.2. The buffer content $B_0(t)$ of the tokens at time t corresponds to the amount of fluid at time t in a reservoir of capacity b where at rate μ fluid is filled in the reservoir and depleted at rate $R(t) = rN_1(t)$. The process $B_0(t)$ is driven by $(N_1(t), N_2(t))$, $t \in \mathbb{R}$, such that

$$\frac{\mathrm{d}B_0(t)}{\mathrm{d}t} = \begin{cases} \max(r(N_1(t), N_2(t)), 0)\,, & B_0(t) = 0\,, \\ r(N_1(t), N_2(t))\,, & 0 < B_0(t) < b\,, \\ \min(r(N_1(t), N_2(t)), 0)\,, & B_0(t) = b\,, \end{cases}$$

where the input rates are given by

$$r(n_1, n_2) := \mu - rn_1\,, \quad (n_1, n_2) \in \mathbb{Z}_+^2\,.$$

The three-dimensional Markov process $(N_1(t), N_2(t), B_0(t))$, $t \in \mathbb{R}$, is the fluid flow model approximation of the token bucket model. There is a lot of references in the literature dealing with analytical as well as numerical aspects for solving fluid flow models, cf. e.g. [AS], [IKKM], [DS], [FV] and the references therein. However, there seems not to be an appropriate solution technique available which is of reasonable complexity and stability for our parameters of interest.

Since the dynamics of $B_0(t)$ are driven only by the number $N_1(t)$ of active requests, we propose to approximate the process $N_1(t)$ by the birth-death process $X(t)$ constructed in Section 3. Denoting the resulting buffer content of the fluid reservoir by $B(t)$, the Markov process $(X(t), B(t))$, $t \in \mathbb{R}$, is a birth-death fluid queue. Birth-death fluid queues with infinite buffer capacity were studied and surveyed in [DS] Section 5. Having in mind that the birth rates λ_n, $n \in \mathbb{Z}_+$, for $X(t)$, cf. (3.7), are of a complex structure, analytical methods of moderate complexity for the zeros of the characteristic polynomial seem not to be available. A numerical computation of them would lead to stability problems in view of complexity. Furthermore, a linear system of equations has to be solved for determining coefficients related to boundary conditions. There is a lot of stability problems in a numerical solution of this type of equations, too, cf. e.g. [FV]. In view of these difficulties, which arise already in case of an infinite buffer, and since in case of a finite buffer the analysis would become yet more complicated, we are interested in an approximation for the packet loss probability $p_{\ell,p}$ in case of a finite positive buffer size b.

4.1 Approximation of the packet loss probability

Consider the stationary birth-death process $X(t)$, $t \in \mathbb{R}$, with rates λ_n and $\mu_n = n\alpha$, cf. Section 3, approximating the number of active requests. Let $\tau \in \mathbb{R}_+ \setminus \{0\}$ be fixed and $X^{(\tau)}(t)$, $t \in \mathbb{R}$, be the birth-death process with distribution $P(X(t) \in (\cdot) \,|\, X(0-0) = \lceil\tau\rceil - 1, X(0) = \lceil\tau\rceil)$, i.e., $X^{(\tau)}(t)$, $t \in \mathbb{R}$, has the Palm distribution of $P(X(t) \in (\cdot))$ with respect to the jump epochs from $\lceil\tau\rceil - 1$ to $\lceil\tau\rceil$. Further let $T_\ell^{(\tau)}$, $\ell \in \mathbb{Z}$, be the epochs where $X^{(\tau)}(t)$ jumps from $\lceil\tau\rceil - 1$ to $\lceil\tau\rceil$, i.e. $X^{(\tau)}(T_\ell^{(\tau)} - 0) = \lceil\tau\rceil - 1$, $X^{(\tau)}(T_\ell^{(\tau)}) = \lceil\tau\rceil$, and assume $\ldots < T_{-1}^{(\tau)} < T_0^{(\tau)} = 0 < T_1^{(\tau)} < \ldots$. Note that $T_\ell^{(\tau)}$ are the hitting times of τ from below. The intensity $\lambda^{(\tau)}$ of jumps of $X(t)$ from $\lceil\tau\rceil - 1$ to $\lceil\tau\rceil$, i.e. the intensity of the time stationary version of the point process $\{T_\ell^{(\tau)}\}_{\ell=-\infty}^{\infty}$, is given by

$$\lambda^{(\tau)} = \lambda_{\lceil\tau\rceil-1} p^{(a)}(\lceil\tau\rceil - 1) = \lceil\tau\rceil \alpha p^{(a)}(\lceil\tau\rceil)\,, \tag{4.1}$$

cf. (3.8). Since $X^{(\tau)}(t)$, $t \in \mathbb{R}$, is a Markov process, it is a regenerative process with respect to the embedded regeneration points $T_\ell^{(\tau)}$, $\ell \in \mathbb{Z}$, too. Because of $T_0^{(\tau)} = 0$, the process $X^{(\tau)}(t)$, $t \in [0, T_1^{(\tau)})$, is a regeneration cycle of the process $X^{(\tau)}(t)$, $t \in \mathbb{R}$. Let

$$D := \min\{t \in \mathbb{R}_+ \,:\, X^{(\tau)}(t) < \tau\}\,, \tag{4.2}$$

i.e., $X^{(\tau)}(t) \geq \tau$ for $t \in [0, D)$ and $X^{(\tau)}(t) < \tau$ for $t \in [D, T_1^{(\tau)})$ in view of the fact that a birth-death process jumps only by values ±1. Note that during the interval $[0, D)$ the content of the token buffer will be constant or emptied since the rate of new arriving fluid of packets is greater or equal to $r\tau = \mu$, cf. (2.15). Hence

$$A := \int_0^D r(X^{(\tau)}(t) - \tau)\,\mathrm{d}t = \int_0^{T_1^{(\tau)}} r(X^{(\tau)}(t) - \tau)_+\,\mathrm{d}t \tag{4.3}$$

is the random amount of arriving fluid during $[0, T_1^{(\tau)})$ which has to be paired from the token buffer, cf. Figure 4.1.

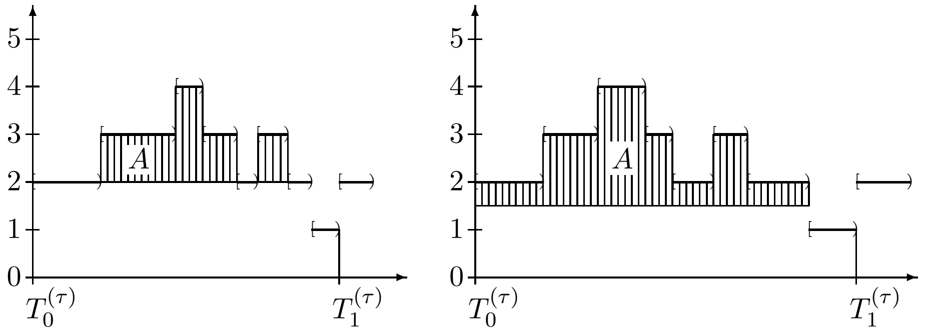

Figure 4.1: *Sample path of the r.v. A (hatched area) in case of $r = 1$ and $\tau = 2$, $\tau = 1.5$, respectively.*

From (4.3), using the cycle formula for regenerative processes with respect to $r(X^{(\tau)}(t) - \tau)_+$, cf. e.g. [A] p. 126, we obtain that

$$EA = E\,T_1^{(\tau)}\,E[r(X(t) - \tau)_+]\,, \tag{4.4}$$

where

$$E\,T_1^{(\tau)} = \frac{1}{\lambda^{(\tau)}} = \frac{1}{\lceil\tau\rceil\alpha p^{(a)}(\lceil\tau\rceil)}\,, \tag{4.5}$$

cf. (4.1). From (4.4) and (4.5) it follows

$$EA = \frac{r}{\lceil\tau\rceil\alpha p^{(a)}(\lceil\tau\rceil)} \sum_{n=\lceil\tau\rceil}^{\infty} (n - \tau)p^{(a)}(n)\,. \tag{4.6}$$

Let B be the random buffer content at $T_0^{(\tau)} = 0$. Then $E[(A - B)_+]$ is the mean amount of packet fluid during $[0, T_1^{(\tau)})$ which gets lost. The mean total fluid EF of arriving packets during $[0, T_1^{(\tau)})$ is given by

$$EF = E\Bigg[\int_0^{T_1^{(\tau)}} rX^{(\tau)}(t)\,\mathrm{d}t\Bigg] = E\,T_1^{(\tau)}\,E[rX(t)]\,, \tag{4.7}$$

where the last equation follows from the cycle formula for regenerative processes again, and it holds

$$E[rX(t)] = r\sum_{n=1}^{\infty} np^{(a)}(n) = \frac{r\lambda(1 - p_{\ell,r}^{(a)})}{\alpha(1 - q)} \tag{4.8}$$

in view of (3.13). The fraction $E[(A-B)_+]/EF$ is the packet loss probability $p_{\ell,p}$ within the birth-death process approximation of the fluid flow model. Assuming that at the beginning of the cycle at $T_0^{(\tau)}=0$ the buffer is full, we obtain the approximation $p_{\ell,p}^{(a)} := E[(A-b)_+]/EF$ for the packet loss probability $p_{\ell,p}$ in case of a finite positive buffer size b:

$$\begin{aligned} p_{\ell,p} \approx p_{\ell,p}^{(a)} &= \frac{\alpha(1-q)}{r\lambda(1-p_{\ell,r}^{(a)})}\lceil\tau\rceil\alpha p^{(a)}(\lceil\tau\rceil)\,E[(A-b)_+] \\ &= \frac{\alpha(1-q)}{\lambda(1-p_{\ell,r}^{(a)})}\sum_{n=\lceil\tau\rceil}^{\infty}(n-\tau)p^{(a)}(n) \\ &\qquad - \frac{\alpha(1-q)}{r\lambda(1-p_{\ell,r}^{(a)})}\lceil\tau\rceil\alpha p^{(a)}(\lceil\tau\rceil)\,E[\min(A,b)] \end{aligned} \tag{4.9}$$

because of (4.5)–(4.8) and $(x-b)_+ = x-\min(x,b)$.

Remark 4.1 *(i) From* (4.9) *it follows*

$$\lim_{b\downarrow 0} p_{\ell,p}^{(a)} = \frac{\alpha(1-q)}{\lambda(1-p_{\ell,r}^{(a)})}\sum_{n=\lceil\tau\rceil}^{\infty}(n-\tau)\,p^{(a)}(n)\,.$$

Comparing this with (3.14) *we find that $p_{\ell,p}^{(a)}$ is continuous at $b=0$, and* (4.9) *is valid in case of $b=0$, too.*
(ii) Note that within the birth-death process approximation it holds

$$p_{\ell,p}^{(a)} \le p_{\ell,p} \le p_{\ell,p}^{(a)}\Big|_{b=0}\,.$$

In case of $b=0$ the approximation $p_{\ell,p}^{(a)}$ can be computed directly from the stationary distribution $p^{(a)}(n)$ via (3.14), (3.12). In case of $b>0$ the approximation depends also on the dynamics of the birth-death process $X(t)$, $t\in\mathbb{R}$, and we have to compute additionally $E[\min(A,b)]$. In the next two sections we will derive an efficient algorithm for computing this quantity approximately. First in Section 4.2 we show that $E[\min(A,b)]$ can be approximated – in principle with arbitrary accuracy – by linear combinations of the LST $A^*(s)=E[e^{-sA}]$ of A at $s=k/b$ for some $k\in\mathbb{Z}_+$. In Section 4.3 we derive a representation of $A^*(s)$, $s\in\mathbb{R}_+$, as a fast convergent continued fraction.

4.2 Approximation of $E[\min(A,b)]$ by means of the LST $A^*(s)$

First we will derive polynomial approximations of the function $(\ln(y)+1)_+$, $y \in (0,1]$, which will yield approximations of $E[\min(A,b)]$ later. Consider the polynomials

$$\varphi_n(y) := y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Bigl(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Bigr)_+ (1-y)^{j-n}\,, \quad n \in \mathbb{N}\,. \tag{4.10}$$

Note that the $\varphi_n(y)$ are positive on $(0,1]$. Defining

$$m(n) := \min\Bigl\{ j \in \{n, n+1, n+2, \dots\} \,:\, \sum_{i=n}^{j} \frac{1}{i} \ge 1 \Bigr\} \tag{4.11}$$

it follows that $\varphi_n(y)$ is a polynomial of degree $m(n)$ and that

$$\varphi_n(y) = y^n \sum_{j=n}^{m(n)} \binom{j-1}{n-1} \Bigl(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Bigr) (1-y)^{j-n}\,, \quad n \in \mathbb{N}\,. \tag{4.12}$$

Hence there is a representation

$$\varphi_n(y) = \sum_{k=n}^{m(n)} c_k^{(n)}\, y^k\,, \quad n \in \mathbb{N}\,. \tag{4.13}$$

Note that $m(n) < en$, $n \in \mathbb{N}$, because of

$$1 > \sum_{i=n}^{m(n)-1} \frac{1}{i} \ge \sum_{i=n}^{m(n)-1} (\ln(i+1) - \ln(i)) = \ln(m(n)) - \ln(n)\,.$$

From (4.12) we find

$$\begin{aligned}
\varphi_n(y) &= y^n \sum_{j=n}^{m(n)} \binom{j-1}{n-1} \Bigl(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Bigr) \sum_{k=0}^{j-n} \binom{j-n}{k} (-y)^k \\
&= y^n \sum_{k=0}^{m(n)-n} (-y)^k \sum_{j=n+k}^{m(n)} \binom{j-n}{k} \binom{j-1}{n-1} \Bigl(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Bigr) \\
&= y^n \sum_{k=0}^{m(n)-n} (-y)^k \sum_{j=n+k}^{m(n)} \binom{n+k-1}{n-1} \binom{j-1}{n+k-1} \Bigl(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Bigr) \\
&= \sum_{k=n}^{m(n)} (-1)^{k-n} \binom{k-1}{n-1} \sum_{j=k}^{m(n)} \binom{j-1}{k-1} \Bigl(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Bigr) y^k\,.
\end{aligned} \tag{4.14}$$

Thus for the coefficients $c_k^{(n)}$ of the polynomial $\varphi_n(y)$, cf. (4.13), we obtain

$$c_k^{(n)} = (-1)^{k-n} \binom{k-1}{n-1} \sum_{j=k}^{m(n)} \binom{j-1}{k-1} \Big(1 - \sum_{i=n}^{j-1} \frac{1}{i} \Big) ,$$

$$k = n, \ldots, m(n) , \quad n \in \mathbb{N} . \qquad (4.15)$$

Lemma 4.1 *For the polynomials* $\varphi_n(y)$ *defined by* (4.10) *it holds*

$$0 \le \varphi_n(y) - (\ln(y)+1)_+ < \frac{0.4}{\sqrt{n-0.5}} , \quad y \in (0,1] , \quad n \in \mathbb{N} . \qquad (4.16)$$

Proof. 1. For the proof we need some elementary preliminaries: For $y \in (0,1]$ it follows

$$\begin{aligned} & y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} (1-y)^{j-n} \\ & \quad = y^n \sum_{j=0}^{\infty} \binom{-n}{j} (y-1)^j = y^n \,(1+(y-1))^{-n} = 1 , \end{aligned} \qquad (4.17)$$

$$\begin{aligned} & y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big(\sum_{i=n}^{j-1} \frac{1}{i} \Big) (1-y)^{j-n} \\ & \quad = -y^n \sum_{j=0}^{\infty} \binom{-n}{j} \Big(- \sum_{i=n}^{n+j-1} \frac{1}{i} \Big) (y-1)^j \\ & \quad = -y^n \lim_{\varepsilon \to 0} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \sum_{j=0}^{\infty} \binom{-n+\varepsilon}{j} (y-1)^j \\ & \quad = -y^n \lim_{\varepsilon \to 0} \frac{\mathrm{d}}{\mathrm{d}\varepsilon} \,(1+(y-1))^{-n+\varepsilon} = -\ln(y) , \end{aligned} \qquad (4.18)$$

$$\begin{aligned} & y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big(\Big(\sum_{i=n}^{j-1} \frac{1}{i} \Big)^2 - \sum_{i=n}^{j-1} \frac{1}{i^2} \Big) (1-y)^{j-n} \\ & \quad = y^n \sum_{j=0}^{\infty} \binom{-n}{j} \Big(\Big(\sum_{i=n}^{n+j-1} \frac{1}{i} \Big)^2 - \sum_{i=n}^{n+j-1} \frac{1}{i^2} \Big) (y-1)^j \\ & \quad = y^n \lim_{\varepsilon \to 0} \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} \sum_{j=0}^{\infty} \binom{-n+\varepsilon}{j} (y-1)^j \\ & \quad = y^n \lim_{\varepsilon \to 0} \frac{\mathrm{d}^2}{\mathrm{d}\varepsilon^2} \,(1+(y-1))^{-n+\varepsilon} = (\ln(y))^2 . \end{aligned} \qquad (4.19)$$

2. Now let us prove the inequality on the l.h.s. of (4.16). For $y \in (0,1]$ from (4.10), $x_+ = x + (-x)_+$, (4.17) and (4.18) we obtain that

$$\begin{aligned}\varphi_n(y) &= y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big(1 - \sum_{i=n}^{j-1} \frac{1}{i}\Big) (1-y)^{j-n} \\ &\qquad + y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big(\sum_{i=n}^{j-1} \frac{1}{i} - 1\Big)_+ (1-y)^{j-n} \\ &= (\ln(y)+1) + y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big(\sum_{i=n}^{j-1} \frac{1}{i} - 1\Big)_+ (1-y)^{j-n}, \qquad (4.20)\end{aligned}$$

which implies immediately $\varphi_n(y) \geq \ln(y)+1$ for $y \in (0,1]$. Remembering $\varphi_n(y) > 0$ for $y \in (0,1]$, cf. (4.10), we conclude that

$$\varphi_n(y) \geq (\ln(y)+1)_+\,, \quad y \in (0,1]\,.$$

3. Now we prove the r.h.s. inequality in (4.16). Using for $\varphi_n(y) - (\ln(y)+1)$ the expression given by (4.20) and for $\varphi_n(y)$ the r.h.s. of (4.10) by taking into account $x_+ + (-x)_+ = |x|$ we obtain that

$$2\varphi_n(y) - (\ln(y)+1) = y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big| \sum_{i=n}^{j-1} \frac{1}{i} - 1 \Big| (1-y)^{j-n}$$

for $y \in (0,1]$. The Cauchy-Schwarz inequality provides

$$\begin{aligned}(2\varphi_n(y) - (\ln(y)+1))^2 &\leq \Big(y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} (1-y)^{j-n}\Big) \\ &\quad \times \Big(y^n \sum_{j=n}^{\infty} \binom{j-1}{n-1} \Big(\sum_{i=n}^{j-1} \frac{1}{i} - 1\Big)^2 (1-y)^{j-n}\Big),\end{aligned}$$

and in view of (4.17)–(4.19) it follows

$$\begin{aligned}&(2\varphi_n(y) - (\ln(y)+1))^2 \\ &\qquad \leq (\ln(y)+1)^2 + y^n \sum_{j=n+1}^{\infty} \binom{j-1}{n-1} \Big(\sum_{i=n}^{j-1} \frac{1}{i^2}\Big) (1-y)^{j-n}\,. \qquad (4.21)\end{aligned}$$

Since for $n \in \mathbb{N}$ and $j = n+1, n+2, \ldots$ it holds

$$\sum_{i=n}^{j-1} \frac{1}{i^2} < \sum_{i=n}^{j-1} \Big(\frac{2}{2i-1} - \frac{2}{2i+1}\Big) = \frac{2}{2n-1} - \frac{2}{2j-1} < \frac{2}{2n-1}\, \frac{j-n}{j-1}\,,$$

from (4.21), (4.17) we conclude

$$(2\varphi_n(y) - (\ln(y)+1))^2 \le (\ln(y)+1)^2 + \frac{2(1-y)}{2n-1}\,, \quad y \in (0,1]\,,$$

which implies

$$2\varphi_n(y) - (\ln(y)+1) \le |\ln(y)+1| + \sqrt{\frac{2(1-y)}{2n-1}}\,, \quad y \in (0,1]\,.$$

In view of $x + |x| = 2x_+$ hence it holds

$$\varphi_n(y) - (\ln(y)+1)_+ \le \sqrt{\frac{1-y}{4n-2}}\,, \quad y \in (0,1]\,. \tag{4.22}$$

From (4.12) and $m(n) < en$ for $y \in (0, e^{-1}]$ it follows

$$\varphi_n'(y) = y^{n-1} \sum_{j=n}^{m(n)} \binom{j-1}{n-1} \left(1 - \sum_{i=n}^{j-1} \frac{1}{i}\right)(n-jy)(1-y)^{j-n-1} > 0\,,$$

and hence $\varphi_n(y) - (\ln(y)+1)_+$ is monotonically increasing for $y \in (0, e^{-1}]$. Thus from (4.22) we obtain

$$\varphi_n(y) - (\ln(y)+1)_+ \le \sqrt{\frac{1-e^{-1}}{4n-2}} < \frac{0.4}{\sqrt{n-0.5}}\,, \quad y \in (0,1]\,.$$

□

Remark 4.2 *(i) Lemma 4.1 implies that the sequence of polynomials $\varphi_n(y)$, $n \in \mathbb{N}$, converges uniformly on $(0,1]$ to the function $(\ln(y)+1)_+$ from above.*
(ii) By tedious asymptotic analysis one can show that

$$\lim_{n\to\infty} \sqrt{n} \max_{y\in(0,1]} (\varphi_n(y) - (\ln(y)+1)_+) = \sqrt{\frac{1-e^{-1}}{2\pi}} \approx 0.3172\,.$$

(iii) Note that

$$\varphi_n(y) - (\ln(y)+1)_+ = \mathcal{O}(y^n(1-y)^{m(n)+1-n})$$

for any $y \in (0,1]$ but fixed $n \in \mathbb{N}$ because of (4.10), (4.11), (4.20) and that $\varphi_n(y)$ is the only polynomial of degree at most $m(n)$ having this property.

Example 4.1 *From* (4.11), (4.13), (4.15) *by elementary algebra one finds*

$$\varphi_1(y) = y\,, \qquad 2\,\varphi_2(y) = 5\,y^2 - 4\,y^3 + y^4$$

and by computer algebra the more interesting representation

$$\begin{aligned}
27\,720\,\varphi_8(y) = {} & 419\,436\,855\,y^8 - 3\,956\,676\,320\,y^9 + 17\,536\,973\,256\,y^{10} \\
& - 48\,063\,232\,800\,y^{11} + 90\,400\,696\,650\,y^{12} - 122\,605\,306\,560\,y^{13} \\
& + 122\,692\,101\,840\,y^{14} - 91\,129\,997\,376\,y^{15} + 49\,795\,472\,925\,y^{16} \\
& - 19\,500\,386\,400\,y^{17} + 5\,190\,375\,080\,y^{18} - 842\,447\,520\,y^{19} \\
& + 63\,018\,090\,y^{20}\,. \qquad (4.23)
\end{aligned}$$

Inequality (4.16) *provides*

$$0 \le \varphi_8(y) - (\ln(y)+1)_+ < 0.15\,, \quad y \in (0,1]\,.$$

In view of $\min(x,1) = 1 - (1-x)_+$, choosing $y = e^{-x}$ in Lemma 4.1 yields an approximation for the function $\min(x,1)$.

Corollary 4.1 *For $n \in \mathbb{N}$ it holds*

$$0 \le \min(x,1) - (1-\varphi_n(e^{-x})) < \frac{0.4}{\sqrt{n-0.5}}\,, \quad x \in \mathbb{R}_+\,,$$

where the $\varphi_n(y)$ are defined by (4.10).

Now we are capable of giving the announced approximation for $E[\min(A,b)]$ in terms of the LST $A^*(s) = E[e^{-sA}]$ of A.

Theorem 4.1 *Let A be defined by* (4.3) *and $b \in \mathbb{R}_+ \setminus \{0\}$. Then for $n \in \mathbb{N}$ it holds*

$$0 \le E[\min(A,b)] - b\left(1 - \sum_{k=n}^{m(n)} c_k^{(n)} A^*(k/b)\right) < \frac{0.4}{\sqrt{n-0.5}}\,b\,, \qquad (4.24)$$

where $m(n)$ and the coefficients $c_k^{(n)}$ are given by (4.11), (4.15), *respectively.*

Proof. From Corollary 4.1 for $x = A/b$ and (4.13) we obtain

$$0 \le \min(A/b,1) - \left(1 - \sum_{k=n}^{m(n)} c_k^{(n)}\, e^{-kA/b}\right) < \frac{0.4}{\sqrt{n-0.5}} \quad \text{a.s.}$$

for $n \in \mathbb{N}$. Multiplying by b and taking expectation yields the assertion (4.24) in view of $E[e^{-kA/b}] = A^*(k/b)$.

□

4.3 Representation of $A^*(s)$ as continued fraction

As in Section 4.1 let $\tau \in \mathbb{R}_+ \setminus \{0\}$ be fixed and consider the stationary birth-death process $X(t)$, $t \in \mathbb{R}$, with birth rates λ_n and death rates μ_n, cf. Section 3, approximating the number of active requests. For fixed $i \in \{\lceil\tau\rceil, \lceil\tau\rceil + 1, \ldots\}$ let $X^{(i)}(t)$, $t \in \mathbb{R}$, be the birth-death process with distribution $P(X(t) \in (\cdot) \,|\, X(0-0) = i-1, X(0) = i)$, i.e., $X^{(i)}(t)$, $t \in \mathbb{R}$, has the Palm distribution of $P(X(t) \in (\cdot))$ with respect to the jump epochs from $i-1$ to i. Further let

$$D_i := \min\{t \in \mathbb{R}_+ \,:\, X^{(i)}(t) < \tau\}\,, \quad i \geq \lceil\tau\rceil\,, \tag{4.25}$$

which is the first hitting time of state $\lceil\tau\rceil - 1$ after time zero. The integrals

$$A_i := \int\limits_0^{D_i} r(X^{(i)}(t) - \tau)\,\mathrm{d}t\,, \quad i \geq \lceil\tau\rceil\,, \tag{4.26}$$

correspond to the random amount of arriving fluid during $[0, D_i)$ which has to be paired from the token buffer. Note that $D_{\lceil\tau\rceil} = D$, $A_{\lceil\tau\rceil} = A$, cf. (4.2), (4.3). Let $A_{\lceil\tau\rceil-1} := 0$ for notational convenience and

$$A_i(x) := P(A_i \leq x)\,, \quad x \in \mathbb{R}\,, \quad i \geq \lceil\tau\rceil - 1\,,$$

be the distribution function of A_i. Applying the law of total probability with respect to the first jump epoch of $X^{(i)}(t)$ after time zero and the properties of birth-death processes provide the renewal type equation

$$A_i(x) = \int\limits_{\mathbb{R}_+} \left(\frac{\lambda_i}{\lambda_i + \mu_i} A_{i+1}\left(x - (i-\tau)rt\right) + \frac{\mu_i}{\lambda_i + \mu_i} A_{i-1}\left(x - (i-\tau)rt\right) \right) (\lambda_i + \mu_i)\, e^{-(\lambda_i + \mu_i)t} \mathrm{d}t\,, \quad x \in \mathbb{R}\,, \quad i \geq \lceil\tau\rceil\,. \tag{4.27}$$

From (4.27) for the LST $A_i^*(s) = E[e^{-sA_i}]$ we obtain

$$\begin{aligned} A_i^*(s) \;&=\; \int\limits_{\mathbb{R}_+} \left(\frac{\lambda_i}{\lambda_i + \mu_i} A_{i+1}^*(s) \;+\; \frac{\mu_i}{\lambda_i + \mu_i} A_{i-1}^*(s) \right) (\lambda_i + \mu_i)\, e^{-(\lambda_i + \mu_i)t}\, e^{-s(i-\tau)rt} \mathrm{d}t \\ &=\; \frac{\lambda_i}{\lambda_i + \mu_i + (i-\tau)rs} A_{i+1}^*(s) + \frac{\mu_i}{\lambda_i + \mu_i + (i-\tau)rs} A_{i-1}^*(s)\,, \\ &\qquad s \in \mathbb{R}_+\,, \quad i \geq \lceil\tau\rceil\,, \end{aligned} \tag{4.28}$$

and $A_{\lceil\tau\rceil-1}(x) = \mathbb{I}\{x \geq 0\}$ yields

$$A_{\lceil\tau\rceil-1}^*(s) = 1\,, \quad s \in \mathbb{R}_+\,. \tag{4.29}$$

Remark 4.3 *From* (4.28), (4.29) *one can derive explicit formulae for* EA_i *and* EA_i^2*, respectively.*

In the following let $\tau \in \mathbb{R}_+ \setminus \{0\}$, $s \in \mathbb{R}_+$ be fixed and

$$x_i := \frac{A_i^*(s)}{A_{i-1}^*(s)}\,, \quad i \geq \lceil \tau \rceil\,. \tag{4.30}$$

Since A_i, $i \geq \lceil \tau \rceil - 1$, is a stochastically monotonically increasing sequence of r.v.'s and $s \in \mathbb{R}_+$, it follows that $E[e^{-sA_i}]$ is a decreasing sequence of positive real numbers, and hence it holds

$$0 < x_i \leq 1\,, \quad i \geq \lceil \tau \rceil\,. \tag{4.31}$$

Moreover, (4.30), (4.29) provide

$$x_{\lceil \tau \rceil} = A_{\lceil \tau \rceil}^*(s) = A^*(s)\,. \tag{4.32}$$

Replacing in (4.28) $A_{i+1}^*(s)$ and $A_{i-1}^*(s)$ with $A_i^*(s)x_{i+1}$ and $A_i^*(s)/x_i$, respectively, cf. (4.30), we obtain

$$x_i = \frac{\mu_i}{\mu_i + (i-\tau)rs + \lambda_i(1-x_{i+1})}\,, \quad i \geq \lceil \tau \rceil\,. \tag{4.33}$$

Thus the sequence $\{x_i\}_{i \geq \lceil \tau \rceil}$ satisfies a continued fraction equation, and $A^*(s)$ is represented as a continued fraction because of (4.32). Since the birth rates λ_i are bounded, cf. (3.6), (3.7), in view of $\mu_i = i\alpha$, $i \in \mathbb{N}$, and (4.31), from (4.33) it follows

$$\lim_{i \to \infty} x_i = \frac{\alpha}{\alpha + rs}\,. \tag{4.34}$$

Let

$$\varphi_i(x) := \frac{\mu_i}{\mu_i + (i-\tau)rs + \lambda_i(1-x)}\,, \quad x \in [0,1]\,, \quad i \geq \lceil \tau \rceil\,. \tag{4.35}$$

Note that $x \in [0,1]$ implies $\varphi_i(x) \in [0,1]$,

$$0 < \varphi_i'(x) \leq \frac{\lambda_i}{\mu_i}\,, \quad x \in [0,1]\,, \quad i \geq \lceil \tau \rceil\,, \tag{4.36}$$

and

$$\varphi_i(x_{i+1}) = x_i\,, \quad i \geq \lceil \tau \rceil\,, \tag{4.37}$$

cf. (4.33). Further let

$$\psi_n(x) := \varphi_{\lceil\tau\rceil}(\varphi_{\lceil\tau\rceil+1}(\varphi_{\lceil\tau\rceil+2}(\ldots(\varphi_{n-1}(\varphi_n(x)))\ldots)))\,, \qquad x\in[0,1]\,,\quad n\geq\lceil\tau\rceil\,. \tag{4.38}$$

In view of $\mu_i = i\alpha$, $i\in\mathbb{N}$, and (3.8), from (4.38), (4.36) we obtain

$$0<\psi_n'(x)\leq\prod_{i=\lceil\tau\rceil}^{n}\frac{\lambda_i}{\mu_i}=\frac{\mu_{n+1}}{\mu_{\lceil\tau\rceil}}\prod_{i=\lceil\tau\rceil+1}^{n+1}\frac{\lambda_{i-1}}{\mu_i}=\frac{(n{+}1)p^{(a)}(n{+}1)}{\lceil\tau\rceil p^{(a)}(\lceil\tau\rceil)}\,, \qquad x\in[0,1]\,,\quad n\geq\lceil\tau\rceil\,, \tag{4.39}$$

and (4.32), (4.37), (4.38) provide

$$\psi_n(x_{n+1}) = A^*(s)\,,\quad n\geq\lceil\tau\rceil\,. \tag{4.40}$$

From (4.31), (4.39), (4.40) it follows

$$\psi_n(0)<A^*(s)\leq\psi_n(1)\,,\quad n\geq\lceil\tau\rceil\,, \tag{4.41}$$

as well as

$$\begin{aligned}|\psi_n(x){-}A^*(s)| \;=\; |\psi_n(x){-}\psi_n(x_{n+1})| &\leq \frac{(n{+}1)p^{(a)}(n{+}1)}{\lceil\tau\rceil p^{(a)}(\lceil\tau\rceil)}\,|x{-}x_{n+1}|\\ &\leq \frac{(n{+}1)p^{(a)}(n{+}1)}{\lceil\tau\rceil p^{(a)}(\lceil\tau\rceil)}\,,\quad x\in[0,1]\,,\quad n\geq\lceil\tau\rceil\,.\end{aligned} \tag{4.42}$$

As $EX(t)$ is finite, cf. (3.9), from (4.42) we obtain

$$A^*(s)=\lim_{n\to\infty}\psi_n(x)\,,\quad x\in[0,1]\,. \tag{4.43}$$

Note that the bound on the r.h.s. of (4.42) does not depend on $s\in\mathbb{R}_+$. In view of (4.42) and (4.34), in particular $\psi_n(\alpha/(\alpha+rs))$ should converge fast to $A^*(s)$ for $n\to\infty$.

Remark 4.4 *A continued fraction approach in the context of $M/M/1$ driven fluid queues has been used in [PVL] for corresponding Laplace transforms.*

References

[ADRS] Adan, I.J.B.F., van Doorn, E.A., Resing, J.A.C., Scheinhardt, W.R.W., Analysis of a single-server queue interacting with a fluid reservoir. Queueing Systems 29 (1998) 313–336.

[AR] Adan, I.J.B.F., Resing, J.A.C., A two-level traffic shaper for an on-off source. Performance Evaluation 42 (2000) 279–298.

[ARK] Adan, I.J.B.F., Resing, J.A.C., Kulkarni, V.G., Stochastic discretization for the long-run average reward in fluid models. Probability in the Engineering and Informational Sciences 17 (2003) 251–265.

[AS] Akar, N., Sohraby, K., Algorithmic solution of finite Markov fluid queues. Proceedings of the 18th Int. Teletraffic Congress, Berlin, Germany, 2003. Eds. J. Charzinski, R. Lehnert, P. Tran-Gia, *Providing Quality of Service in Heterogeneous Environments*, Elsevier Science, 621–630.

[AMS] Anick, D., Mitra, D., Sondhi, M.M., Stochastic theory of a data-handling system with multiple sources. Bell System Technical Journal 61 (1982) 1871–1894.

[A] Asmussen, S., *Applied Probability and Queues.* Wiley, Chichester, 1987.

[BB] Brandt, A., Brandt, M., On the distribution of the number of packets in the fluid flow approximation of packet arrival streams. Queueing Systems 17 (1994) 275–315.

[BBS] Brandt, A., Brandt, M., Sulanke, H., A single server model for packetwise transmission of messages. Queueing Systems 6 (1990) 287–310.

[DS] van Doorn, E.A., Scheinhardt, W.R.W., Analysis of birth-death fluid queues. Proceedings of Applied Mathematics Workshop. Ed. B.D. Choi, Korea Advanced Institute of Science and Technology, Taejon (1996) 13–29.

[EM] Elwalid, A.I., Mitra, D., Analysis and design of rate-based congestion control of high speed networks, I: Stochastic fluid models, access regulation. Queueing Systems 9 (1991) 29–63.

[FV] Fiedler, M., Voos, H., *Fluid flow-Modellierung von ATM-Multiplexern.* Mathematische Grundlagen und numerische Lösungsmethoden. Herbert Utz Verlag Wissenschaft, Munich, 1997.

[IKKM] Igelnik, B., Kogan, Y., Kriman, V., Mitra, D., A new computational approach for stochastic fluid models of multiplexers with heterogeneous sources. Queueing Systems 20 (1995) 85–116.

[K] Kosten, L., Stochastic theory of data-handling systems with groups of multiple sources. Eds. H. Rudin, W. Bux, *Performance of Computer-Communication Systems.* Elsevier, Amsterdam 1984, 321–331.

[LP] Lenin, R.B., Parthasarathy, P.R., A computational approach for fluid queues driven by truncated birth-death processes. Methodology in Computing and Appl. Probab. 2 (2000) 373–392.

[PVL] Parthasarathy, P.R., Vijayashree, K.V., Lenin, R.B., An M/M/1 driven fluid queue – continued fraction approach. Queueing Systems 42 (2002) 189–199.

[R] Roberts, J.W. (ed.), *Performance Evaluation and Design of Multiservice Networks.* Final Report of the COST 224 Project, Commission of the European Communities.

[SE] Stern, T.E., Elwalid, A.I., Analysis of separable Markov-modulated rate models for information-handling systems. Adv. Appl. Probab. 23 (1991) 105–139.