

**Lectures at the Instytut Matematyczny, Uniwersytet Wrocławski,
on**

Discrete Time Stochastic Networks

2 State Dependent Bernoulli Servers

In this section¹ we consider the most elementary queueing system in discrete time. It is the analogue of the state dependent exponential single server queue in continuous time under First-Come-First-Served (FCFS) regime. We begin with systems where all customers are indistinguishable and summarize in section 2.1 the steady state performance measures. Most of those problems can be interpreted as special questions about the behaviour of random walks in discrete time. This does not hold in case of customers of different types arriving at the service node which we consider in section 2.2. In Section 2.3 we reconsider both models and allow immediate feedback of departed customers to the server.

Notations and conventions:

For ease of notation throughout the lectures we fix a common probability space (Ω, \mathcal{F}, P) and assume all random variables which occur to be defined on (Ω, \mathcal{F}, P) , unless otherwise specified.

If a probability measure P on (Ω, \mathcal{F}) that governs some stochastic process

$$X = ((X_n : (\Omega, \mathcal{F}, P) \longrightarrow (E, \mathcal{S})) : n \in \mathbb{N})$$

has to be specified further, often for guaranteeing that X is stationary with stationary distribution (= steady state = equilibrium) π , this will be indicated by writing P_π . An expectation e.g. of X_n , with respect to such P_π is written $E_\pi X_n$. Similar expressions will be selfexplaining.

\mathbb{R} denotes the real numbers, $\mathbb{R}_+ := [0, \infty)$.

The natural numbers are $\mathbb{N} := \{0, 1, 2, \dots\}$, the strict positive natural numbers are $\mathbb{N}_+ := \{1, 2, 3, \dots\}$, and we denote $\mathbb{Z} := \{\dots, -2, -1, 0, 1, 2, \dots\}$.

We denote by

$$\delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$$

the Kronecker delta, and by

$$\eta(a, b) = \begin{cases} 1 & \text{if } a \neq b \\ 0 & \text{if } a = b \end{cases}$$

the complementary Kronecker delta.

¹wr2.tex

2.1 Indistinguishable Customers

The time scale for our systems is $\mathcal{N} = \{0, 1, 2, \dots\}$, sometimes we use $\mathcal{Z} = \mathcal{N} \cup -\mathcal{N} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. There is a single service facility where at each time instant at most one customer may be served. Customers are indistinguishable and arrive one by one randomly at the server. If at the arrival instant the server is free the service of the arriving customer immediately commences. Otherwise an arriving customer enters the waiting room which is organized on a FCFS basis (sometimes called FIFO: First-In-First-Out). If a customer has obtained his total service request he departs immediately from the system. If a customer departs and there is at least one further customer present then the customer at the head of the waiting line enters the server, his service commences immediately, and all other waiting customers are shifted one place up in the line. The time needed for reorganizing the queue is assumed to be neglectible (zero time).

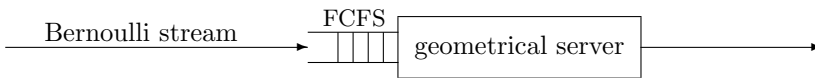


Figure 1: Bernoulli Server

The system's development over time will be described by a discrete time stochastic process $X = (X_t : t \in \mathcal{N})$ with state space \mathcal{N} . X_t denotes the number of customers present at time t , either in service or waiting, shortly the *queue length* at time t .

The randomness of the system is due to the following assumptions:

If at time t a customer is in service and if there are $n - 1 \geq 0$ other customers present then this service ends in the time segment $[t, t + 1)$ with probability $p(n) \in (0, 1)$ and the customer will depart at the end of this time slot; with probability $q(n) = 1 - p(n)$ this customer will stay at least one further time quantum. The decision for a customer whether to stay or to leave is made independently of anything else other than the queue length at time t . If at time $t \in \mathcal{N}$ there are $n \geq 0$ customers present then at the end of time slot $[t, t + 1)$ a new customer will arrive with probability $b(n)$; with probability $c(n) = 1 - b(n)$ there will be no arrival. The decision whether an arrival will occur or not is made independently of anything else other than the queue length at time t . Such an arrival stream will be termed henceforth *state dependent Bernoulli arrival process*. (For some special models we shall allow $p(n) = 1$ and/or $b(n) = 1$; see e.g. corollary 2.8, example 2.9.)

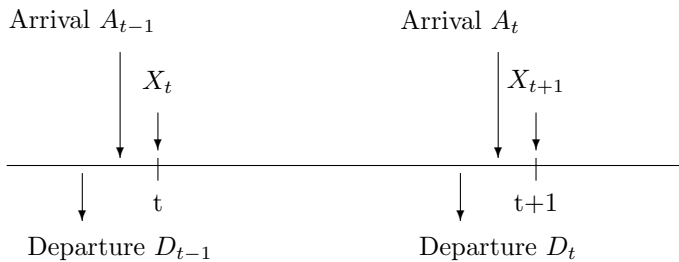


Figure 2: Regulation of arrivals and departures

Unless otherwise specified, we shall assume throughout the notes the following rules for the regulation of simultaneous events: All arrivals (and departures) occur at the end of the respective time slots (*LA-rule*; late arrivals). If at the same epoch an arrival and a departure occur we always assume that the departure event takes place first. (*D/A-rule*; departure before arrival, see [GH92].) The state of the system is recorded at times $t \in \mathbb{N}$ just after possible departures and arrivals have happened. (Figure 2.)

A pathwise analysis of the system considers sequences of successive arrivals and departures which result in the queue length process observed at the time epochs at slot boundaries $t \in \mathbb{N}$.

Definition 2.1 (Departure and arrival process) We denote by

$$D = (D_t : t \in \mathbb{N}) \quad \text{and} \quad A = (A_t : t \in \mathbb{N})$$

the sequences of numbers of departing, resp., arriving customers in time slot $[t, t + 1)$, and assume

$$P(D_{t+1} = k | X_s, D_s, A_s : s \leq t) = P(D_{t+1} = k | X_t), \quad t \geq 0,$$

and

$$P(A_{t+1} = k | X_s, D_s, A_s : s \leq t) = P(A_{t+1} = k | X_t), \quad t \geq 0$$

to hold.

Corollary 2.2 (Queue length process) The queue length process is pathwise defined by

$$X_{t+1} = X_t - D_t + A_t, \quad t \in \mathbb{N}, \quad X_0 \text{ prescribed.} \quad (1)$$

$X = (X_t : t \in \mathbb{N})$ is a time homogeneous Markov chain.

In the literature there is a great variety of arrival and departure regimes defined. These regimes usually reflect physical behaviour of the systems under consideration and specification of protocols which govern the interaction of different processes running concurrently in the system. For the single server queue considered in this section an overview and a comparison of different regulation schemes is given by Hunter [Hun83a], more recent are [GH92], and paying more attention to the effect of different regulation schemes for simultaneous events in networks of queues on the performance analysis of customer oriented quality of service, [DD99], [Des97].

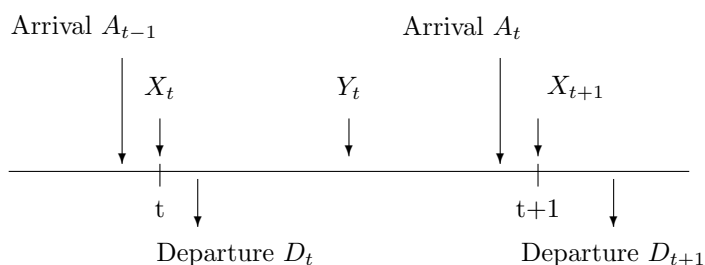


Figure 3: Regulation of arrivals and departures in [CMP99]

In [CMP99], p.345, the regulation of arrivals and departures is pointed out to be essential for proving quasi-reversibility of the process. It is assumed that early departures occur at times $t+$ and arrivals enter

the system at $(t+1)-$ in time slot $[t, t+1)$, see figure 3. The state process X is observed at slot boundaries t , and additionally a second state process Y is observed at times $t+1/2$, as is done, e.g., in [GH92] as well.

It should be noticed that the pathwise construction of the queue length process in corollary 2.2 coincides with the queue length process constructed and observed at the boundaries of the time slots in [CMP99], p.346, but the structure of the transition probabilities in definition 2.1 differ from the the internal structure (intermediate jump probabilities) of their transition mechanism which requires

$$P(A_{t+1} = k | X_s, D_s, A_s : s \leq t, X_{t+1}, D_{t+1}) = P(A_{t+1} = k | X_{t+1} - D_{t+1}), \quad t \geq 0$$

to hold. But nevertheless in many cases the overall transition probability for the state process X can be represented in either way.

The discrete time queueing system described so far is called a *state dependent Bernoulli server*. The term *Bernoulli server* stems from the case of state independent service and arrival probabilities (see corollary 2.7 below) and is generalized to our setting. The fundamental characteristics of the system are well known from the discrete time birth and death chain theory, see remark 2.4.

Theorem 2.3 (Steady state) *Let $X = (X_t : t \in \mathbb{N})$ with state space \mathbb{N} denote the queue length process of the state dependent Bernoulli server described above.*

X is a homogeneous Markov chain.

If $p(n), b(n) \in (0, 1), \forall n \in \mathbb{N}$, then X is irreducible and aperiodic on \mathbb{N} .

If X is irreducible and aperiodic on \mathbb{N} , then X is positive recurrent, hence ergodic, if and only if

$$H := \sum_{n=0}^{\infty} \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)} < \infty. \quad (2)$$

If X is ergodic then its unique stationary and limiting distribution $\pi = (\pi(n) : n \in \mathbb{N})$ is

$$\pi(n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)} \cdot H^{-1}, \quad n \in \mathbb{N}. \quad (3)$$

Proof : From the independence assumptions and the specification of the arrival and departure probabilities via queue lengths only it follows that X is a homogeneous Markov chain. From $p(n), b(n) \in (0, 1), n \in \mathbb{N}$, we have $P(X_{t+1} = n | X_t = n) > 0$ and $P(X_{t+|m-n|} = m | X_t = n) > 0$, hence X is aperiodic and irreducible.

The equilibrium equations $\mathbf{x} = \mathbf{x} \cdot \mathbf{p}$, where \mathbf{p} is the one-step transition matrix of X and $\mathbf{x} = (x(n) : n \in \mathbb{N})$ is a nonnegative row vector, are solved by

$$x(n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)}, \quad n \in \mathbb{N}.$$

Normalizing \mathbf{x} yields the probability solution (3) if (2) holds. ⊙

Remark 2.4 (Random walks in discrete time) *The queue length of the Bernoulli server, i.e., the Markov chain X , moves at most one step up or down in a time slot, i.e., it is a general random walk on \mathbb{N} (with reflection at 0) in discrete time in the sense of [KSK76], p. 84, or a birth and death chain in discrete time, see [Hun83b], (Vol. I), p. 178. Theorem 2.3 and the corollaries below are therefore simple consequences of the limiting and stationary behaviour of birth and death chains, see [Hun83a], (Vol. II), Example 7.2.2, p. 107.*

Remark 2.5 (Computational problems) *Due to allowing the general form of arrival and service probabilities it is in general not possible to give closed form solutions for steady state probabilities (even if these exist). This is in parallel to the continuous time birth and death process theory. Even determining recurrence and transience has to refer to the data of the specific system under consideration. For details see: Hunter [Hun83a], Vol. II, example 7.2.2, p. 107.*

An example with explicitly given norming constant is provided by systems with state independent arrival and service probabilities as described in the corollary 2.7 below. This yields a (homogeneous) random walk with reflection at 0.

Remark 2.6 (Reversibility) *For the case of state independent arrival probabilities, i.e., if the arrival process is a Bernoulli-(b) stream, Hsu and Burke [HB76] proved that in steady state the queue length process X is reversible in time. This implies that the departure process in equilibrium is a Bernoulli-(b) process. Further, in equilibrium, the departure process up to t and the state at t are independent. This lead Hsu and Burke to apply separability properties to tandem queues.*

For random walks on \mathbb{N} in discrete time the Bernoulli property of the departure process means, that even with state dependent death rates the downstep process is homogeneous if the birth probabilities are constant. In [CMP99], example 12.10, and the remark below on p.354, it is shown that this queue is quasi-reversible according to the definition 12.6 there. This again leads to the statement of Hsu and Burke on separably coupling such queues in a tandem.

The system dealt with in theorem 2.3 is neither reversible nor quasi-reversible.

Corollary 2.7 (State independent Bernoulli server) *Assume that in the setting of theorem 2 we have $p(n) = p \in (0, 1)$, and $b(n) = b \in (0, 1), n \in \mathbb{N}$.*

Then X is ergodic if and only if $b < p$, and if this holds the stationary distribution of X is

$$\pi(n) = \left(1 - \frac{b}{p}\right) \left(\frac{bq}{cp}\right)^n \left(\frac{1}{q}\right)^{\eta(0,n)}, \quad n \in \mathbb{N}. \quad (4)$$

The arrival process is a Bernoulli process with probability b for successes and the service times are geometrically distributed on $\mathbb{N}_+ = \{1, 2, 3, \dots\}$ with parameter p . The interarrival times are geometrically distributed, and the service process can be thought to be regulated by a Bernoulli-(p) process.

Weakening the assumptions on the arrival and service probabilities leads to some interesting special cases.

Corollary 2.8 (Loss systems) *Assume that in the setting of Theorem 2 we have $b(n) \in (0, 1), n \leq L-1 > 0, b(n) = 0, n \geq L$.*

Then X is ergodic on $E = \{0, 1, \dots, L\}$, and the stationary distribution of X is $\pi = (\pi(n) : n \in E)$ given by

$$\pi(n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)} \cdot H^{-1}, \quad n \in \mathbb{N}, \quad (5)$$

where

$$H := \sum_{n=0}^L \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)}$$

is the norming constant.

Example 2.9 (Deterministic service times) *The case of deterministic service times is of specific interest, because in modelling transmission lines where cells with constant length are to be transmitted this constant cell length τ constitutes a generic discrete time quantum such that $\mathbb{N} \cdot \tau = \{0, \tau, 2\tau, 3\tau, \dots\}$ is a suitable time scale for a process reflecting the development of the system over time. Assuming $\tau = 1$ and $p(n) = 1, n \in \mathbb{N}_+$, and starting the system with $P(X_0 = n) > 0$ for all $n \in \mathbb{N}$, we find that X is not irreducible on \mathbb{N} .*

Indeed, if $b(n) \in (0, 1)$ for all $n \in \mathbb{N}$, then X is irreducible and ergodic on $E = \{0, 1\}$. The stationary distribution π is given by

$$\pi(0) = \frac{1}{c(0)} \cdot H^{-1}, \quad \pi(1) = \frac{b(0)}{c(0)c(1)} \cdot H^{-1},$$

which formally fits into (3).

If $b = 1$ and $P(X_0 = n) > 0, \forall n \in \mathbb{N}_+, P(X_0 = 0) = 0$, then X is stationary for any such initial distribution.

The discrete time counterpart to the classical $M/M/s/\infty$ queue with s identical service channels under FCFS poses surprisingly many problems with respect to analytical evaluation. No simple closed form expressions for the steady state seem to be at hand. Usually numerical procedures with root solving for multidimensional boundary equations are applied. Models related to that problem are dealt with e.g. in [BSDP92], [DT92] [SZ94], [BK93], section 4.1.2, where the service time is deterministic 1 while the arrival streams are more general. The case of no waiting room is e.g. considered in [CG96]. The complexity of problems and the numerical difficulties which arise in using these multiserver queues in discrete time for modeling controlled ATM switches are described e.g. in [RMW94], where a leaky-bucket control is investigated in detail. Pestien and Ramakrishnan [PR94], section 3, proved that including these servers into a closed cycle of geometrical queues destroys the product form equilibrium if $s < \infty$.

It is well known that in continuous time for the $M/M/s/\infty$ queue there exists an equivalent state dependent single server queue with the same steady state distribution. From the discussion above and the explicit steady state (3) it follows that such an equivalence is not possible in discrete time. However it is for a subclass of $M/M/s/\infty$ queues possible to construct a single server approximation in a rather direct way.

Example 2.10 (Multiserver queue approximation in low traffic)

Consider an $M/M/s/\infty$ queue in discrete time with Bernoulli arrival stream of intensity $b \in (0, 1)$ and service probability $p \in (0, 1)$ such that $b < p \cdot s$ holds. The conditional service intensity for that system (mean number of departures per time unit given the number of customer in system at the begin of a time slot) is the same as that of a single server state dependent queue according to theorem 2.3 with service probabilities

$$p(n) := p \cdot \min(n, s).$$

Studying the behaviour of individual customers with respect to their delay behaviour (waiting time and sojourn or passage time distribution) in equilibrium is a main topic in performance analysis. A first step towards this is determining under equilibrium conditions the state distribution at arrival instants, i.e., what an arriving customer observes just before he enters the system. To be more specific: An arriving customer observes at his arrival at $t-$ exactly n other customers present, if at time t $X_t = n + 1$ holds and an arrival occurs. A customer who is to be observed during his stay in the node for determining (probabilistic) characteristics of his behaviour is usually attributed to be a test customer. To deal with this notion and then to derive from a test customer's behaviour conclusions about the behaviour of a typical customer

needs care because e.g. knowing that a test customer commences his sojourn in the system prevents in general the system from being in equilibrium. The theoretical framework for dealing with this problem is Palm calculus of point process theory (see [BB94]) which fortunately enough can be proved to reduce to investigate elementary conditional probabilities in discrete time models (see [BB94], section 7.4). We formulate this as arrival theorems for customers, usually starting from a system in equilibrium.

Theorem 2.11 (Arrival theorem) *Let the queue length process X of a state dependent Bernoulli server be in steady state π according to (3). Consider for $t > 0$ the event*

$A(t) = \{ \text{At time } t \text{ a new customer arrives at the system} \}$. Then for $n \in \mathbb{N}$

$$\pi_1(n) := P(X_t = n + 1 | A(t)) = \frac{\prod_{m=0}^n b(m)}{\prod_{m=0}^{n+1} c(m)} \cdot \frac{\prod_{m=1}^n q(m)}{\prod_{m=1}^n p(m)} \cdot H_1^{-1}, \quad (6)$$

is the probability (with normalization constant H_1) that conditioned on an arrival at time t the arriving customer finds n other customers before him in service or waiting.

We denote by $\pi_1 = (\pi_1(n) : n \in \mathbb{N})$ the arrival probability. The arrival state and arrival probability refer to the disposition of the other customers present, the new arrival is not counted.

Proof : For $t > 0$ we compute elementary conditional probabilities $P(X_t = n + 1 | A(t)) = \sum_{r=0}^{\infty} P(X_t = n + 1, X_{t-1} = r, A(t)) P(A(t))^{-1}$ ◉

The common interpretation of π_1 is that it describes the distribution of the other customers' disposition in an arrival instant at the node under equilibrium conditions, the jumping customer not counted. Remarkable is that this arrival distribution has not the form of the equilibrium of the system even if the arrival process is a state independent Bernoulli process.. In continuous time it is true for the case of state independent arrivals that the time stationary and customer arrival stationary distribution coincide. For open systems with external Poisson arrivals this is the celebrated PASTA property (Poisson Arrivals See Time Averages) [Wol82].

From a general point of view the PASTA theorem and its successors and relatives determine the stationary and asymptotic distribution of systems when the observation points are prescribed by an associated (embedded) point process. Comparing the stationary distribution of the embedded state process with the time stationary distribution of the systems (seen by an outside observer) is the topic of many research activities in queueing theory. Since Wolff's PASTA theorem [Wol82] appeared the research in this field yielded many generalizations of that property. These have been proved and popularized under names like ASTA, EPSTA, MUSTA. For a review see [BB94], Chapter 4, Section 3, or [KS89], and [ETS99], Chapters 3,4. Using point process terminology, EPSTA and its relatives are concerned with properties of special Palm measures of stationary point processes: Papangelou's formula, which connects the point process intensity and time stationary expectations with Palm stationary expectations, allows the derivation of a variety of formulas summarized under the title *job observer properties* in single node systems as well as in networks of queues. For continuous time, see e.g. [BB94], Example 3.2.2, [MW90], Example 3, [ETS99], Section 4.3, and [DS99]. Starting from Palm theory in discrete time ([BB94], Chapter 1, Section 7.4) a similar development is possible. Palm measures in discrete time are expressed as elementary conditional probabilities. This makes the theory more elementary although explicit computations are tedious, see sections 3.4 and 3.5.

In discrete time, however, a PASTA analogue usually does not hold, although exceptions can be found under certain conditions. An early result was proved by Halfin in [Hal83]. Characterisation theorems of

the PASTA type (thereby strengthening the BASTA–results (Bernoulli Arrivals See Time Averages) from [MMW89]) were proved by El-Taha and Stidham [ETS92], (see also [ETS99], Section 2, Theorem 3.18 and Corollary 3.19. Miyazawa and Takahashi [MT92] proved ASTA in a discrete time point process setting by using a rate conservation principle. They also observed that for some natural systems this property does not hold. In this notes we prove “arrival theorems” for the different cases, where individual customers’ behaviour when entering the system is of interest. These provide us with the necessary basic information to compute customer oriented performance measures. The proofs can be performed either by applying the general theory or by directly evaluating elementary conditional probabilities, i.e. Palm probabilities in discrete time processes. The latter approach is generally chosen here. We assume in any case that the Markov process which describes the time evolution of the system is stationary.

Note: If we rescale time, approaching in the limit a continuous time scale, our systems transform to exponential queueing system: Then in the limit the distinction between arrival distribution and steady state of the system disappears for state independent arrival probabilities.

Because customers are scheduled according to FCFS regime the arrival theorem provides us with information about the number of services which will be performed before a test customer, who enters the system in equilibrium, will be served. This enables us in principle to write down the end–to–end–delay time distribution. But even for this simple single node, single server system explicit results seem to be not known in full generality. We consider the case of state independent service probabilities:

Theorem 2.12 (End–to–end–delay) *Consider the Bernoulli server with state dependent arrival rates $b(n) \in (0, 1)$ and state independent service rates $p(n) = p \in (0, 1)$ in equilibrium with a test customer arriving at time 0 finding the other customers distributed according to π_1 , see (6). We denote by P_{π_1} a probability measure on (Ω, \mathcal{F}) which governs the development of X under this conditions and by $E_{\pi_1}[\cdot]$ expectations under P_{π_1} .*

Denote by S this test customer’s total sojourn time in the system and by

$$E_{\pi_1} \theta^S = \sum_{s=0}^{\infty} P_{\pi_1}(S = s) \theta^s, \quad |\theta| \leq 1,$$

the generating function (of the distribution) of S . If the generating function of the arrival factors in π , see (3), is

$$\alpha(\theta) = \sum_{n=0}^{\infty} \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \theta^n, \quad |\theta| \leq q/p, \quad (7)$$

then

$$E_{\pi_1} \theta^S = \frac{\alpha\left(\frac{q\theta}{1-q\theta}\right) - \alpha(0)}{\alpha\left(\frac{q}{p}\right) - \alpha(0)}, \quad |\theta| \leq 1. \quad (8)$$

Proof : We have by conditioning

$$\begin{aligned} E_{\pi_1} \theta^S &= \sum_{n=0}^{\infty} P_{\pi_1}(X_0 = n) E_{\pi_1} [\theta^S | X_0 = n] \\ &\stackrel{(1)}{=} \sum_{n=0}^{\infty} \frac{\prod_{m=0}^n b(m)}{\prod_{m=0}^{n+1} c(m)} \cdot \left(\frac{q}{p}\right)^n \cdot H_1^{-1} \cdot \left(\frac{p\theta}{1-q\theta}\right)^{n+1}. \end{aligned}$$

In ⁽¹⁾ we used that the service times of the customers are geometrically distributed with parameter p on \mathbb{N}_+

and independent. From $H_1 = (p/q)(\alpha(q/p) - \alpha(0))$ the result follows immediately. \odot

2.2 Customers of Different Types

In this section we consider the service system of section 2.1 and assume that customers arriving at the system may be of different types. All customers are served according to the same rules (FCFS) and have the same distribution for their requested amount of service. Later on the type description will determine a customer's behaviour and service in networks of state dependent nodes. The details:

The node characteristics remain the same as in the previous section. Customers of different types arrive according to a state dependent Bernoulli process at the node, are served according to FCFS, and thereafter depart from the system.

Joint arrivals and departures are scheduled according to LA-D/A regime (*late arrivals and departure before arrivals*), see [GH92].

We assume that there is a single chain of customers, all customers share the same type set M . The entrance type of an arriving customer is chosen according to the following rules: The external arrival probabilities depend on the history of the system only through the actual total population size of the system and on the type of the arrival, i.e., if at time $t \in \mathcal{I}N$ there are n customers present, then a new arrival of type m appears in $(t, t + 1]$ with probability $b(n) \cdot a(m) \in (0, 1)$.

Departure and arrival decisions are conditionally independent given the actual vector of queue lengths.

A typical state of the system is described by a type sequence $x = (x_1, \dots, x_n) \in M^n$, where for $n > 0$ x_1 is the type of the customer in service, x_2 is the type of the customer at the head of the queue, \dots , x_n is the type of the customer who arrived most recently. The empty system is denoted by $x = e$. (For definiteness we set for the empty system the queue length $n = 0$.) These states are sufficient for constructing the state space of the system where X is living on.

Let $X(t)$ denote the state of the node at time $t, t \in \mathcal{I}N$. $X = (X(t) : t \in \mathcal{I}N)$ is a discrete time Markov chain with state space $\tilde{S} := \{e\} \cup \bigcup_{n=0}^{\infty} M^n$ and transition matrix $p = (p(x, y) : x, y \in \tilde{S})$. X is irreducible on \tilde{S} . The problem of stabilisation for this tandem system is solved by the following theorem.

Theorem 2.13 (Steady state) *The Markov chain X is ergodic if and only if*

$$H = \sum_{(n \in \mathcal{I}N)} \left(\frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \right) \left(\frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)} \right) < \infty.$$

If this ergodicity condition is fulfilled, then the unique equilibrium distribution of X is $\pi = (\pi(x) : x \in \tilde{S})$ given by

$$\begin{aligned} \pi(x) &= \pi(x_1, \dots, x_n) \\ &= \left(\frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \right) \left(\prod_{k=1}^n a(x_k) \right) \left(\frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^n p(m)} \right) \cdot H^{-1}, \\ &x = (x_1, \dots, x_n) \in \tilde{S}. \end{aligned} \tag{9}$$

Proof : By inserting formula (9) into the equilibrium equation. We shall consider the case of networks of such nodes in detail below. \odot

Remark 2.14 (Steady state decomposition) *Similarly to theorems 2.3, 2.11, we observe a decomposition (separation) of the steady states into factors concerning the arrival probabilities, the service probabilities, and in addition the type selection probabilities. This separability is common to almost all product form steady states in continuous time and will occur later on in the framework of discrete time queueing networks as well. The separability opens the way to search successfully for explicit performance measures.*

Our aim is to investigate an individual customer's delay behaviour. We must therefore incorporate the types into an arrival theorem similar to theorem 2.11.

Theorem 2.15 (Arrival Theorem) *Consider the state process X of the node in equilibrium and denote by*

$A(m, t) = \{ \text{at time } t \text{ a customer of type } m \text{ arrives at the node} \}$
the arrival event of interest. Then

$$\begin{aligned} \pi_{1,m}(x) &= P(X(t) = ((x_1, \dots, x_n), m) | A(m, t)) \\ &= \left(\frac{\prod_{m=0}^n b(m)}{\prod_{m=0}^{n+1} c(m)} \right) \left(\prod_{k=1}^n a(x_k) \right) \left(\prod_{m=1}^n \frac{q(m)}{p(m)} \right) H_1^{-1}, \\ &x = (x_1, \dots, x_n) \in \tilde{S}, \end{aligned} \quad (10)$$

H_1 is the norming constant, which does not depend on the type of the arriving customer.

For the interpretation of $\pi_{1,m}$ see the remark following theorem 2.11. Note that the arrival distribution does not depend on the type of the arriving customer. This is due to the fact that we consider an "open" system. For "closed" systems this property does not hold.

Corollary 2.16 (End-to-end-delay) *Consider the Bernoulli server with state dependent arrival rates $b(n) \in (0, 1)$ and state independent service rates $p(n) = p \in (0, 1)$ in equilibrium with a test customer of type m arriving at time 0 finding the other customers distributed according to $\pi_{1,m}$, see (10). We denote by $P_{\pi_{1,m}}$ a probability measure on (Ω, \mathcal{F}) which governs the development of X under this conditions and by $E_{\pi_{1,m}}[\cdot]$ expectations under $P_{\pi_{1,m}}$.*

Denote by S this test customer's total sojourn time in the system. Then the generating function of the distribution of S is

$$E_{\pi_{1,m}} \theta^S = \sum_{s=0}^{\infty} P_{\pi_{1,m}}(S = s) \theta^s = \frac{\alpha\left(\frac{q\theta}{1-q\theta}\right) - \alpha(0)}{\alpha\left(\frac{q}{p}\right) - \alpha(0)}, \quad |\theta| \leq 1. \quad (11)$$

Here the $\alpha(\cdot)$ is the generating function of the arrival factors, given in (7).

Proof : Because the arrival distribution does not depend on the individual type of the test customer, the factors concerning types cancel and we end with computations similar to those in the proof of theorem 2.11. \odot

2.3 Bernoulli Servers with Immediate Feedback

Feedback queues are standard models in continuous time systems to model repeated visits of an item to a production or service facility. Another situation with typical feedback structure is rework which occurs due to production control at the exit point of a production stage. A class of network models which encompass these features are re-emtrant lines which are surveyed in [Kum93] in the continuous time setting, and which are to

be described for the discrete time case later on. Applications in the realm of ATM transmission systems are described in [STH98] where in a node with service time deterministic–(1) the feedback mechanism models the successive transmission of cells of a message, the length of which is geometrically distributed. An important consequence is that introducing the feedback destroys the FCFS structure of the systems, which reflects real systems’ protocol behaviour. In fact, the queueing regime used by the authors is what is called *Round–Robin* queueing discipline, and which is described in [LB96], see [Lae96] and the references there. We shall describe here the single feedback station in isolation, which will be used later on as the nodes of tandems.

We consider a state dependent Bernoulli server under FCFS as described in section 2.1 with indistinguishable customers and introduce a Bernoulli switch at the departure point of the node. A customer departing from the queue leaving behind $m - 1$ customers is fed back into the waiting room (to the tail of the queue) with probability $r(m)$. If he was the only customer present he will obtain immediately a further service, otherwise he will join the tail of the queue. With probability $1 - r(m)$ he will leave the system. The decision whether to leave or to reenter the node is made independently of anything else. The regulation of customer movements in case of multiple events is: Departure before arrival for a joint arrival and departure (D/A) and feedback before arrival for a joint feedback and arrival (F/A).

The queue length process $X = (X_t : t \in \mathbb{N})$ with state space \mathbb{N} is Markovian. If $0 < b(\cdot), p(\cdot) < 1$, then X is irreducible and aperiodic on \mathbb{N} . A short reflection shows:

Observing the queue length process at times $t \in \mathbb{N}$ only is not sufficient to decide whether a feedback happened or no arrival *and* no departure happened or one arrival *and* one departure. We can determine from observing a path of X only single external arrivals and single departures to the external. The transition probabilities of the feedback queue are therefore

$$\begin{aligned}
p(0, 0) &= c(0), & (12) \\
p(0, 1) &= b(0), \\
p(0, m) &= 0, \quad m \neq 0, 1, \\
p(n, n - 1) &= c(n)p(n)(1 - r(n)), \quad n \geq 1, \\
p(n, n) &= c(n)(q(n) + p(n)r(n)) + b(n)p(n)(1 - r(n)), \quad n \geq 1, \\
p(n, n + 1) &= b(n)(q(n) + p(n)r(n)), \quad n \geq 1, \\
p(n, m) &= 0, \quad n \geq 1, m \neq n, m - 1, m + 1.
\end{aligned}$$

These transition probabilities are identically to those of a state dependent Bernoulli server without feedback, with arrival probabilities $b(\cdot)$, and with service probabilities $p'(n) = p(n)(1 - r(n), q'(n) = 1 - p(n)(1 - r(n)), n \in \mathbb{N}$. Therefore the Markovian queue length processes of these different systems are stochastically indistinguishable, and we conclude that theorem 2.3 applies for computing the steady state of the feedback queue.

Corollary 2.17 (Steady state of the feedback queue) *Let $X = (X_t : t \in \mathbb{N})$ with state space \mathbb{N} denote the queue length process of the state dependent Bernoulli server with immediate feedback.*

If X is irreducible and aperiodic on \mathbb{N} , then X is positive recurrent, hence ergodic, if and only if

$$H := \sum_{n=0}^{\infty} \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} (q(m) + p(m)r(m))}{\prod_{m=1}^n p(m)(1 - r(m))} < \infty.$$

If X is ergodic then its unique stationary and limiting distribution $\pi = (\pi(n) : n \in \mathbb{N})$ is

$$\pi(n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} (q(m) + p(m)r(m))}{\prod_{m=1}^n p(m)(1-r(m))} \cdot H^{-1}, \quad n \in \mathbb{N}. \quad (13)$$

Note that even in the case of state independent arrival, departure, and feedback probabilities the effective arrival stream at the waiting room, i.e., the superposition of the external arrival stream and the feedback stream, is neither geometrical nor a state dependent Bernoulli arrival process as defined in section 2.1.

We consider now a feedback node where customers of different types from the set M of possible types are served, i.e., the model of section 2.2 with a queue length dependent Bernoulli feedback as described before. The state space for a Markovian description is $\tilde{S} = \{e\} \cup \bigcup_{n=1}^{\infty} M^n$, as defined before theorem 2.13. We notice that we usually can realize the occurrence of a feedback by a resulting permutation of the customers. We even can decide when a batch arrival due to a simultaneous feedback and external arrival appears. The only exception: All customers present are of the same type.

Consequently the transition probabilities of the feedback node have no longer an interpretation in terms of a Bernoulli node without feedback and suitably adjusted service probabilities. In the light of this the following result is somewhat surprising. The proof is straight forward but tedious checking the balance equations.

Theorem 2.18 (Feedback queue with different customer types) *Let $X = (X_t : t \in \mathbb{N})$ with state space \tilde{S} denote the Markov chain describing the evolution of the state dependent feedback queue with different customer types. If $p(n), b(n) \in (0, 1)$, and $r(n) < 1$ for all $n \in \mathbb{N}$, then X is irreducible and aperiodic on \tilde{S} . If X is irreducible and aperiodic, then X is positive recurrent, hence ergodic, if and only if*

$$H := \sum_{n=0}^{\infty} \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \frac{\prod_{m=1}^{n-1} (q(m) + p(m)r(m))}{\prod_{m=1}^n p(m)(1-r(m))} < \infty.$$

If X is ergodic then its unique stationary and limiting distribution $\pi = (\pi(x) : x \in \tilde{S})$ is

$$\begin{aligned} & \pi(x_1, \dots, x_n) && (x_1, \dots, x_n) \in \tilde{S}, \\ & = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^n c(m)} \cdot \prod_{k=1}^n a(x_k) \cdot \frac{\prod_{m=1}^{n-1} (q(m) + p(m)r(m))}{\prod_{m=1}^n p(m)(1-r(m))} \cdot H^{-1}. \end{aligned} \quad (14)$$

We obtain the round-robin discipline dealt with in [STH98], where service time is deterministic (1) and the total system time being geometrically distributed, by putting $p(n) = 1$ for all $n \geq 1$ and the feedback probability being state independent as well. Nevertheless, the model in [STH98] includes several further modeling features not represented in the simplified feedback Bernoulli server here. In [LB96], [Lae96] it is assumed that the number of packets arriving per slot is determined by a sequence of identically distributed random variables (batch Bernoulli arrival process), the service time is of (discrete) phase type, and the service mechanism is round-robin. Due to the general arrival process the steady state of the queue length process is no longer of product form, as it is in case of state dependent single arrivals of different customer classes [DS81].

References

[BB94] F. Baccelli and P. Bremaud. *Elements of Queueing Theory*. Springer, New York, 1994.

- [BK93] H. Bruneel and Byung G. Kim. *Discrete-Time Models for Communication Systems including ATM*. Kluwer Academic Publications, Boston, 1993.
- [BSDP92] H. Bruneel, B. Steyaert, E. Desmet, and G.H. Petit. An analytical technique for the derivation of the delay performance of ATM switches with multivserver output queues. *Intern. Journ. of Digital and Analog Communication Systems*, 5:193–201, 1992.
- [CG96] M.L. Chaudhry and U.C. Gupta. Transient behaviour of the discrete time Geom/Geom/m/m Erlang loss model. In A.C. Borthakur and M.L. Choudhry, editors, *Probability Models and Statistics, A.J. Medhi Festschrift*, pages 133 – 145, New Delhi, 1996. New Age International Limited, Publishers.
- [CMP99] X. Chao, M. Miyazawa, and M. Pinedo. *Queueing Networks – Customers, Signals, and Product Form Solutions*. Wiley, Chichester, 1999.
- [DD99] B. Desert and H. Daduna. Discrete time tandem networks of state dependent queues: The effect of different regulation schemes for simultaneous events on customer oriented performance measures. Preprint 99-07, Institut für Mathematische Stochastik der Universität Hamburg, 1999. submitted.
- [Des97] B. Desert. *Lineare stochastische Netzwerke in diskreter Zeit: Gleichgewichtsverhalten und Durchlaufzeitverteilungen*, 1997. Diploma thesis.
- [DS81] H. Daduna and R. Schassberger. A discrete-time round-robin queue with bernoulli input and general arithmetic service time distributions. *Acta Informatica*, 15:251 –263, 1981.
- [DS99] H. Daduna and R. Szekli. Conditional job observer properties in multitype closed queueing networks. Preprint, Mathematical Institute of the University of Wroclaw, 1999. to appear: Journal of Applied Probability.
- [DT92] J. N. Daigle and St. C. Tang. The queue length distribution for multiserver discrete time queues with batch Markovian arrivals. *Comm.Statist.–Stochastic Models*, 8:665–683, 1992.
- [ETS92] M. El-Taha and S. Jr. Stidham. A filtered ASTA property. *Queueing Systems and Their Applications*, 11:211–222, 1992.
- [ETS99] M. El-Taha and S. Jr. Stidham. *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publisher, Boston, 1999.
- [GH92] A. Gravey and G. Hebuterne. Simultaneity in discrete-time single server queues with Bernoulli inputs. *Performance Evaluation*, 14:123–131, 1992.
- [Hal83] S. Halfin. Batch delays versus customer delays. *The Bell System Technical Journal*, 62:2011–2015, 1983.
- [HB76] J. Hsu and P.J. Burke. Behaviour of tandem buffers with geometric input and markovian output. *IEEE Transactions on Communications*, 24:358 – 361, 1976.
- [Hun83a] J. J. Hunter. *Mathematical Techniques of Applied Probability*, volume II: *Discrete Time Models: Techniques and Applications*. Academic Press, New York, 1983.

- [Hun83b] J. J. Hunter. *Mathematical Techniques of Applied Probability*, volume I: *Discrete Time Models: Basic Theory*. Academic Press, New York, 1983.
- [KS89] D. Koenig and V. Schmidt. EPSTA: The coincidence of time-stationary and customer-stationary distributions. *Queueing Systems and Their Applications*, 5:247–264, 1989.
- [KSK76] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov Chains*. Springer-Verlag, New York – Heidelberg – Berlin, 1976. Reprint of the book published in 1966 by Van Nostrand, Princeton.
- [Kum93] P. R. Kumar. Re-entrant lines. *Queueing Systems and Their Applications*, 13:87–110, 1993.
- [Lae96] K. Laevens. The round-robin service discipline in discrete time for phase-type distributed packet-lengths. Preprint, SMAC Research Group, University of Ghent, 1996.
- [LB96] K. Laevens and H. Bruneel. Discrete-time queueing models with feedback for input-buffered ATM switches. *Performance Evaluation*, 27,28:71–87, 1996.
- [MMW89] A. Makowski, B. Melamed, and W. Whitt. On averages seen by arrivals in discrete time. In *IEEE Conference on Decision and Control, Vol. 28*, pages 1084–1086, Tampa, FL., 1989.
- [MT92] M. Miyazawa and Y. Takahashi. Rate conservation principle for discrete-time queues. *Queueing Systems and Their Applications*, 12:215–230, 1992.
- [MW90] Benjamin Melamed and Wy Whitt. On arrivals that see time averages. *Operations Research*, 38:156–172, 1990.
- [PR94] V. Pestien and S. Ramakrishnan. Features of some discrete-time cyclic queueing networks. *Queueing Systems and Their Applications*, 18:117 – 132, 1994.
- [RMW94] J.-F. Ren, J. W. Mark, and J.W. Wong. Performance analysis of a leaky-bucket controlled ATM multiplexer. *Performance Evaluation*, 19:73–101, 1994.
- [STH98] Y. Sakai, Y. Takahashi, and T. Hasegawa. Discrete time multi-class feedback queue with vacations and close time under random order of service discipline. *Journal of the Operations Research Society of Japan*, 41:589–609, 1998.
- [SZ94] K. Sohraby and J. Zhang. Spectral decomposition approach for transient analysis of multi-server discrete-time queues. *Performance Evaluation*, 21:131–150, 1994.
- [Wol82] R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30:223–231, 1982.