Queueing Syst (2006) 54:55–78
DOI 10.1007/s11134-006-8710-5

M/M/1 Queueing systems with inventory

Maike Schwarz · Cornelia Sauer · Hans Daduna · Rafal Kulik · Ryszard Szekli

Received: 11 August 2004 / Revised: 6 April 2006

Abstract We derive stationary distributions of joint queue length and inventory processes in explicit product form for various M/M/1-systems with inventory under continuous review and different inventory management policies, and with lost sales. Demand is Poisson, service times and lead times are exponentially distributed. These distributions are used to calculate performance measures of the respective systems. In case of infinite waiting room the key result is that the limiting distributions of the queue length processes are the same as in the classical M/M/1/∞-system.

Keywords Queueing systems · Inventory systems · Performance analysis · Inventory policy · Queue lengths · Service level · Inventory level · Stationary distribution · Convex ordering

1 Introduction

The importance of inventory management for the quality of service (QoS) of today's service systems is generally accepted and optimization of systems in order to maximize QoS is therefore an important topic.

All authors were supported by DAAD/KBN grant number D/02/32206.

M. Schwarz · C. Sauer (✉) · H. Daduna
University of Hamburg, Department of Mathematics, Center of Mathematical Statistics and Stochastic Processes, Bundesstrasse 55, 20146 Hamburg, Germany
e-mail: sauer@math.uni-hamburg.de

R. Kulik · R. Szekli
University of Wroclaw, Mathematical Institute, Pl. Grunwaldzki 2/4, 50-384 Wroclaw, Poland

From the point of view of inventory availability there are many different classical definitions of quality ([22] p. 232), and on the other side from the point of waiting times or queueing in service systems QoS characteristics are well established. But evaluation of these characteristics usually is done in models either from inventory theory or from queueing theory.

Over the last decade research on complex integrated production-inventory systems or service-inventory systems has found much attention, often in connection with the research on integrated supply chain management. Interaction of production/service processes with inventory management for associated inventories is usually described using queueing networks and multi-echelon inventory models. Mathematical methods used in the field are usually aggregation-disaggregation techniques or simulation or hybrid techniques. Analytical models are rare up to now.

Only recently investigations in integrated models appeared concerning the problem of how the classical performance measures (e.g. queue length, waiting time, etc.) are influenced by the management of attached inventory—and vice versa: How inventory management has to react to queueing of demands and customers, which is due to incorporated service facilities.

An early contribution is [17] where approximation procedures are used to find performance descriptions for models, in which the interaction of queueing for service and inventory control is integrated. In a sequence of papers Berman and his coworkers ([2–4, 6]) investigate the behaviour of service systems with an attached inventory. Their approach can be characterized as follows. Define a Markovian system process and then use classical optimization methods to find the optimal control strategy of the inventory. All these models assume that the demand, which arrives during the time the inventory is zero, is backordered. The models vary with

respect to the lead time distribution, the service time distribution, waiting room size, order size and reorder policy. Further a continuous review structure for the inventory is assumed in all these models. The growing use of computer systems to control inventories has boosted the interest in continuous-review inventory models [14].

Our investigation is on continuous-review inventory models with lost sales of customers that arrive during stockout. The lost sales situation arises e.g. in many retail establishments [9], where the intense competition allows customers to choose another brand or to go to another store. This can be considered as a typical situation for being described by a pure inventory model.

But there are other areas of applications, where lost sales models are appropriate as well. E.g. these models apply to cases such as essential spare parts where one must go to the outside of the normal ordering system when a stockout occurs ([8] p. 605). The essential spare part problem is central for many repair procedures, where broken down units arrive at a repair station, queue for repair, and are repaired by substituting a failed part by a spare part from the inventory. A similar problem arises in production processes where rough material items are needed to let the production process run. Both of these latter problems are usually modeled using pure service systems, but these queueing theoretical models neglect the inventory management. Lost sales are in these contexts known as losses of customers. There is a huge amount of literature on loss systems, especially in connection with teletraffic and communication systems, where losses usually occur due to limited server capacity or finite buffer space. But there is another occurrence of losses due to balking or reneging of impatient customers. However, only in the essential spare part problem of repair facilities a sort of inventory at hand is considered.

To summarize these observations: Lost sales in inventory theory and losses of customers in queueing theory are technical terms for similar, even often the same, events in real systems. The difference is set by the appropriate model selection done by the investigators: Either emphasizing the inventory management point of view or emphasizing the service system's point of view, both cases mostly neglect the alternative aspect.

As pointed out above models that incorporate both aspects, queueing of customers and inventory management, and offer closed form solutions are rare in the literature. (Even more, the area of continuous review inventory systems with lost sales remains largely unexplored, the literature on lost sales models is reviewed in the article of Mohebbi and Posner [14].)

The aim of our research is to present explicit performance measures for service systems with an attached inventory under continuous review and lost sales as well as availability measures and service grades for the inventory to directly enable cost optimization in an integrated model. We analyse several single server queueing systems of M/M/1-type with an attached inventory. Customers arrive according to a Poisson process with intensity λ and each customer, who is served, needs exactly one item from the inventory and has an exponentially distributed service time with parameter μ. Consequently, the demand rate of the inventory is equal to λ if there are no customers waiting in queue otherwise the demand rate is equal to the service rate μ. The variable replenishment lead time, which is the time span between ordering of materials and receipt of the goods, is exponentially distributed with parameter ν. The entire order is received into stock at the same time. The type of inventory system is defined to be a continuous review system where the inventory state is inspected after every single demand event and orders are placed every time the inventory on hand reaches a reorder point r. The on-hand stock is the stock that is physically on the shelf. The systems under investigation differ with respect to the size of replenishment orders and the reorder policy. Every system under consideration has the property that no customers are allowed to join the queue as long as the inventory is empty. This corresponds to the lost sales case of inventory management. However, if inventory is at hand, customers are still admitted to enter the waiting room even if the number of customers in the system exceeds the inventory on hand.

The strategy of our investigation in this paper is as follows:
We start from the observation of Berman and Kim [2] who proved that in an exponential system with zero lead times an optimal policy does not place an order unless the inventory is empty and a certain number of customers are waiting. We therefore investigate in Section 2 an $M/M/1/\infty$ queue with inventory, where the set of feasible policies is prescribed by fixing the reorder point 0 and allowing general randomized order sizes that are restricted only by the capacity of the inventory. Coming up with an explicit description of the steady state behaviour for the (queueing/inventory) process we are able to prove that in this class of feasible policies the vector process for (queue length/inventory size) is minimized in the strong stochastic order by using deterministic order size. i.e., in this class of policies the optimal policy is of $(r = 0, Q)$ structure, $Q \in \mathbb{N}_+$.

This observation is the rationale behind the investigation of $M/M/1/\infty$ queues with inventory, where the set of feasible policies is prescribed by fixing the reorder point $r \geq 0$ and deterministic order sizes Q in Section 3. In Section 4 we investigate the respective systems under the classical celebrated (r, S)-policy. Our explicit results enable us to compare the behaviour of systems under (r, Q) and (r, S)-policies. Letting $S = Q + s$, where s is the safety stock in the (r, Q)-system, we find that under (r, Q) the stationary mean inventory position is greater or equal to that in the (r, S)-system,

and all the other performance measures dealt with in both systems are equal.

In Sections 6.1, 6.2 and 6.3 systems similar to those in Sections 2, 3 and 4 are investigated that have only finite waiting rooms. We assume throughout the paper that during the time a replenishment order is outstanding the service place can be used as a waiting place by the customers in the system. This regulation scheme resembles the behaviour of the BLOCKING BEFORE SERVICE—SERVER OCCUPIED strategy of networks with blocking.

For each of our systems we compute the steady state probability distribution and calculate the most important performance measures. It turns out that a special feature of the above models is that the steady state probabilities are of product form. This means, that the asymptotic and stationary distribution of the joint (queue length/inventory size) process factorizes into the stationary queue length and inventory size distribution. Saying it the other way round: In the long run and in equilibrium the queue length process and the inventory process behave as if they are independent. This is a rather strange observation because—as described above—these processes strongly interact, independently of whether being in equilibrium or not.

In case of infinite waiting room the limiting distributions of the queue length processes coincide with that of the M/M/1/∞-system with arrival rate λ and service rate μ. This shows an unexpected and very important invariance property for the queueing systems with inventory management and lost sales investigated in this paper: indeed, we can see that for the effective arrival rate $\lambda_{eff} \neq \lambda$ holds, and that for the effective service rate $\mu_{eff} \neq \mu$ holds.

The unexpected conclusion is: the system by itself regulates the effective service and arrival rates in reaction to the lead time characteristics and the inventory management policy in a way that the service system always experiences a traffic intensity $\rho = \frac{\lambda}{\mu} = \frac{\lambda_{eff}}{\mu_{eff}}$. The only side condition is $\rho = \frac{\lambda}{\mu} < 1$.

In Section 7 we comment on the possibility to perform cost analysis for the systems using the explicit results on inventory size and queue lengths obtained in the previous sections.

Throughout the paper we will assume that unless otherwise specified an underlying probability space $(\Omega, \mathcal{F}, \mathsf{P})$ is given where all random variables are defined on.

2 Single server system with inventory and lost sales

*Definition 2.1 (*The general queueing-inventory system). At a service system with an attached inventory undistinguishable customers arrive one by one and require service. There is a single server with unlimited waiting room under first come, first served (FCFS) regime and an inventory with maximal capacity of M (identical) items. Each customer needs exactly one item from the inventory for service, and the on-hand inventory decreases by one at the moment of service completion. If the server is ready to serve a customer which is at the head of the line and there is no item of inventory this service starts only at the time instant (and then immediately) when the next replenishment arrives at the inventory. Customers arriving during a period when the server waits for the replenishment order are rejected and lost to the system ("lost sales").

A served customer departs from the system at once and the associated item is removed from the inventory at this time instant as well. If there is another customer in the line and at least one further item in the inventory, the next service starts immediately.

There is a policy specified which determines at each decision point whether a replenishment order is placed or not, and how many items are ordered. Admissible decision epochs are arrival and departure epochs. We assume that there is always at most one outstanding order.

There are costs connected with operating the system originating from both, the queueing of the customers and from holding inventory at the system. We have a fixed holding cost h per item and time unit in the inventory, a fixed ordering cost K for each replenishment order, a shortage cost ℓ per unit of lost sales, a cost ω per customer and time unit in the waiting room, and a cost σ per customer and time unit in service. Whenever a customer's service is completed, a revenue R is payed to the system.

Let $X(t)$ denote the number of customers present at the server at time $t \geq 0$, either waiting or in service, and let $Y(t)$ denote the on-hand inventory at time $t \geq 0$. We denote the joint queue length and inventory process by $Z = ((X(t), Y(t)), t \geq 0)$. The state space of Z is $E_Z = \{(n, k) : n \in \mathbb{N}_0, k \in \{0, ..., M\}\}$, where M is the maximal size of the inventory, which depends on the order policy, see Definition 2.3. We shall henceforth refer to Z as the queueing-inventory process.

Definition 2.2 (Assumption on the random behaviour of the system). For the service system with inventory management from Definition 2.1 we assume:

Customers are of stochastically identical behaviour. To the server there is a Poisson-λ-arrival stream, $\lambda > 0$. Customers request an amount of service time which is exponentially distributed with mean 1. Service is provided with intensity $\mu > 0$.

Service times and inter-arrival times constitute an independent family of random variables.

*Definition 2.3 (*Reorder policy). For the service system with inventory management from Definition 2.1 we consider the following policy:

If the inventory is depleted after the service of a customer is completed, a replenishment order is instantaneously triggered. The decision of the order size may be randomized according to a discrete probability density function p on the integers $\{1, 2, \ldots, M\}$, where M is the maximal capacity of the inventory. So the size of a replenishment order is k with probability p_k, where $\sum_{k=1}^{M} p_k = 1$. The corresponding discrete distribution function and its tail distribution function will be denoted by F_p and $\bar{F}_p := 1 - F_p$, respectively. We abbreviate the probability that the size of a replenishment order is at least k units by q_k, i.e. $q_k = \bar{F}_p(k-) = \sum_{h=k}^{M} p_h$. The mean order size is $\bar{p} = \sum_{k=1}^{M} k p_k$.

Fixed (deterministic) order quantities are described by using one-point distributions for the order size distribution.

The replenishment lead time is exponentially distributed with parameter $\nu > 0$. Order size decisions and lead times are independent and independent of the arrival and service times.

The system described above generalizes the lost sales case of classical inventory management where customer demand is not backordered but lost in case there is no inventory on hand (see Tersine [22] p. 207). We postpone a detailed discussion on the modeling assumptions to the end of this subsection.

*Definition 2.4 (*M/M/1/∞ with inventory). A service system with inventory according to Definition 2.1, with the stochastic assumptions of Definition 2.2, and under some prescribed policy, is called an M/M/1/∞ system with inventory or with inventory management under that policy.

Theorem 2.5. *For the M/M/1/∞ system with inventory according to Definition* 2.4 *the stochastic queueing-inventory process Z from Definition* 2.1 *is a homogeneous Markov process. Z is ergodic if and only if $\lambda < \mu$. If Z is ergodic then it has a unique limiting and stationary distribution of product form:*

$$\pi(n,k) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n q_k \ \text{ with } n \in \mathbb{N}_0, 1 \le k \le M, \tag{1}$$

$$\pi(n,0) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n \frac{\lambda}{\nu} \ \text{ with } n \in \mathbb{N}_0, \tag{2}$$

$$\text{and with normalization constant} \quad K = \frac{\mu}{\mu-\lambda}\left(\bar{p} + \frac{\lambda}{\nu}\right). \tag{3}$$

Proof: We first check that for finite K π is a stationary measure which satisfies the global balance equations of Z for all $n \in \mathbb{N}_0$:

$$\begin{aligned}
&\pi(n,M)(\lambda + \mu(1-\delta_{0n})) \\
&\quad = \pi(n-1,M)\lambda(1-\delta_{0n}) + \pi(n,0)\nu p_M, \\
&\pi(n,k)(\lambda + \mu(1-\delta_{0n})) \\
&\quad = \pi(n-1,k)\lambda(1-\delta_{0n}) + \pi(n+1,k+1)\mu \\
&\qquad + \pi(n,0)\nu p_k, \quad M > k \ge 1, \\
&\pi(n,0)\nu = \pi(n+1,1)\mu.
\end{aligned}$$

This is done by insertion, using $q_1 = 1, q_M = p_M$, and $q_k = q_{k+1} + p_k$ for $M > k \ge 1$. Evaluating the normalization constant K yields the ergodicity criterion. □

We note further that the normalization constant K factorizes in the normalization constant for the marginal queue length and for the inventory process as

$$K = K_X \cdot K_Y \ \text{ with } \ K_X := \frac{\mu}{\mu-\lambda} \ \text{ and } \ K_Y := \bar{p} + \frac{\lambda}{\nu}.$$

Remark 2.6. For $\nu = \infty$ the inventory is replenished instantaneously and the inventory position 0 is left immediately. Therefore, the stationary distribution has support $\mathbb{N}_0 \times \{1, 2, \ldots, M\}$ and is given by (1) with $K = \frac{\mu}{\mu-\lambda}\bar{p}$.

Corollary 2.7.

(a) The marginal steady state queue length distribution of $X = (X(t), t \ge 0)$ is equal to the steady state queue length distribution in the classical M/M/1/∞-FCFS system with the same parameters λ and μ. Therefore the mean number of customers in system is

$$\bar{L}_0 = \frac{\lambda}{\mu-\lambda}.$$

(b) The steady state on-hand inventory distribution of $Y = (Y(t), t \ge 0)$ is

$$\mathsf{P}(Y=k) = \begin{cases} K_Y^{-1}\dfrac{\lambda}{\nu}, & \text{for } k = 0 \\ K_Y^{-1} q_k, & \text{for } 1 \le k \le M. \end{cases}$$

Here we have denoted by Y a random variable distributed like the stationary inventory distribution.

Remark 2.8. The result of Corollary 2.7 (a) is remarkable. λ is in effect not the arrival rate of customers to the system

defined in Definition 2.1 and μ is not the actual service rate of customers. However, these rates must have been reduced to the same degree (in equilibrium) since if X has a steady state distribution of the form $\tilde{\pi}(n) = \rho^n(1-\rho)$, then $\rho = \frac{\lambda_{\text{eff}}}{\mu_{\text{eff}}}$ is the quotient of the effective input and service rates. Here $\rho = \frac{\lambda_{\text{eff}}}{\mu_{\text{eff}}} = \frac{\lambda}{\mu}$.

Remark 2.9. We have from Corollary 2.7 (b) that for $M \geq k \geq 1$

$$\mathsf{P}(Y = k | Y > 0) = \frac{q_k}{\sum_{l=1}^{M} q_l} = \frac{\bar{p}^{-1} \sum_{h=k}^{M} p_h}{\sum_{l=1}^{M} \bar{p}^{-1} \sum_{h=l}^{M} p_h}, \quad (4)$$

i.e. the stationary distribution of the inventory content, conditioned on having on-hand inventory greater than 0, is the conditional asymptotic and stationary residual inter-occurence time in a discrete time renewal process with life time distribution $(p_h : 1 \leq h \leq M)$. This resembles the structure of the stationary residual service time for symmetric servers [13][Chapter 3.3] with lattice service time distributions.

A first intuition behind this observation is that (4) describes the conditional state distribution of an alternating renewal process, with life time distributions that are alternating according to $\exp(\nu)$ and a "busy period distribution" of the inventory that consists of random phases, the number of these phases is random according to $p(\cdot)$. Nevertheless even this intuition does not fully meet the real system's behaviour: The aging of the latter life times of this alternating renewal process is randomly interrupted when an associated random environment changes its state: this environment is the queue length process which interrupts aging by entering state 0.

Discussion of the modeling assumptions: We consider the system of Definition 2.1 with the further specifications in Definitions 2.2 and 2.3 to be fundamental. It corresponds to the popular (r, S)- and (r, Q)-policies with reorder point $r = 0$ and random S and Q respectively. The main result is the explicit steady state distribution in (1) and (2) for the joint queueing-inventory process. In the spirit of the usual terms of performance analysis of complex networks the distribution in (1) and (2) is a PRODUCT FORM STEADY STATE. This emphasizes that in the long run and in equilibrium the queue length and the inventory level behave AS IF THEY ARE INDEPENDENT. In the field of queueing networks such product form steady states opened the path to very successful applications of stochastic network models in various fields. It is well known that even in the context of network applications the assumptions which are necessary to obtain these explicit formulas are often rather unrealistic. Nevertheless, these models were usually acknowledged as a fundamental breakthrough in network modeling.

The strong restriction in our present model is that we regulate reordering and admission of the customers only via the inventory level. Customers are only rejected (and lost), when the physical inventory level reaches zero. A more sophisticated policy would include into the decision procedure information on the actual queue length at the feasible decision instant. The gain of posing our restriction on the reorder policy is the result of Theorem 2.5.

The selection of the class of policies determined by Definition 2.3 originates from observations made in connection with various other models in the literature.

- It is shown by Berman and Kim [2], that in case of exponential interarrival and service times and zero lead times a reorder point $r > 0$ is suboptimal if customer demand is backordered and inventory holding costs are involved. An optimal order policy for that system only places an order when the inventory level drops to zero and the number of customers in the system exceeds some threshold value.
- From stochastic dynamic optimization we know that in many situations randomized policies must be considered for optimizing system processes.

  $\Rightarrow$ We therefore will first carry out in detail the calculations for the system with arbitrary random order size out of $\{1, 2, \ldots, M\}$ and reorder point 0 and show how to define the important measures of system performance.
- An outcome of this investigation is the proof that for fixed mean order size deterministic order size is optimal.
- For the case of backordering customers which arrive during a stockout and prescribed fixed order size with non-zero exponential lead times Berman and Kim [3] showed that the optimal policy is of threshold type such that with given inventory level the reorder decision depends on the queue length. The method to prove this is stochastic dynamic optimization (no steady state analysis seems to be possible up to now).

  $\Rightarrow$ This result suggests that it may be profitable to already trigger an order if there is still some inventory on stock in case of stochastic lead times if customer waiting costs are incurred. We therefore turn to policies with deterministic order sizes and reorder point $r \geq 0$.

  Summarized: Our insights into the M/M/1/∞-system with inventory management, reorder point 0 and random size of replenishment orders suggest to extend the system to early replenishments with reorder level $r \geq 0$, without admitting for random order sizes. This results in the M/M/1/∞-system with (r, Q)-policy that will be investigated in Section 3. For comparison we also investigate the M/M/1/∞-system with (r, S)-policy in Section 4.
- The concerns about the restrictions on the order policies and on admitting customers independently of the queue lengths apply in the subsequent models of Sections 3 and 4 as well, but in any case the gain will be a result in parallel to Theorem 2.5.

⇒ We therefore investigate in Section 6 systems under different inventory policies where arriving customers are rejected if a prescribed threshold of the queue length is reached. The results are explicit steady state distributions of product form as well.

2.1 Measures of system performance

We are interested in stationary characteristics of the queueing-inventory system. These are long-run characteristics as well. Note that stationarity is always assumed in the classical inventory theory as well. Having determined the stationary distribution, we can compute several measures of operating characteristics for the system explicitly. We introduce the following measures of system performance for the stationary system: the average inventory position $\bar{I}$, the expected reorder rate λ_R, the expected lost sales per unit time $\overline{LS}$ and per cycle $\overline{LS}_c$, the safety stock s, α- and β-service levels and the average waiting time for a customer $\bar{W}$.

We say that the system goes through one cycle in the time between the placing of two successive orders or equivalently the receipt of two successive procurements with respect to the mean cycle time, i.e. the mean cycle time is λ_R^{-1}.

It will turn out that all except of one performance measure only depend on the mean order size $\bar{p}$ and not on the whole distribution F_p. Only the stationary mean inventory position depends on the second moment of F_p and will be shown to be minimal for fixed order size in Section 2.3.
The stationary average on-hand inventory position is given by

$$\bar{I} = \sum_{k=1}^{M} k \sum_{n=0}^{\infty} \pi(n,k) = K_Y^{-1} \sum_{k=1}^{M} k q_k. \tag{5}$$

Note that

$$\sum_{k=1}^{M} k q_k = \sum_{k=1}^{M} k \sum_{h=k}^{M} p_h = \sum_{k=1}^{M} \left(\sum_{h=1}^{k} h \right) p_k$$
$$= \frac{1}{2} \sum_{k=1}^{M} k(k+1) p_k.$$

Hence, $\bar{I}$ depends on the first and second moment of the distribution F_p.

From (67) of Theorem A.1 in Appendix A the expected number of replenishments per time unit (reorder rate) is

$$\lambda_R = \frac{\lambda}{\bar{p} + \frac{\lambda}{\nu}}. \tag{6}$$

Formula (6) reveals a striking insensitivity property of the systems under consideration: The steady state reorder rate λ_R in the systems with explicit incorporation of service and queueing behaviour is the same as the mean number of replenishment orders given by Hadley and Whitin [11] on page 180 in 4–37 and by Hax and Candea [12] in (4.1.38) for the case of no queueing (i.e. service time = 0). Saying it the other way round: The expected number of replenishments per unit time in the stationary system is independent of the service intensity as long as it is greater than the overall arrival rate λ. Clearly the same observation holds for the reciprocal value, the mean time between two replenishment orders $\frac{\bar{p}}{\lambda} + \frac{1}{\nu}$.

Therefore we have proven that the condition to stabilize the system asymptotically can be decoupled from the inventory management problem as long as only the reorder rate is concerned.

The average number of lost sales incurred per unit of time is given by

$$\overline{LS} = \lambda \mathsf{P}(Y = 0) = \frac{\lambda^2}{\bar{p}\nu + \lambda}.$$

The expected number of lost sales per cycle is given by

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \lambda/\nu.$$

Inventory can be divided into working stock and safety stock. Working stock is inventory acquired and held in advance of requirements so that the expected demand can be satisfied and ordering can be done on a lot size rather than on an as needed basis. Lot sizing is done in order to minimize ordering and holding costs. Safety stock is held in reserve to protect against the uncertainties of supply and demand (Tersine [22] p. 205). The safety stock is usually defined as follows (see Silver/Peterson [18] Chapter 7).

Definition 2.10. The safety stock is the Palm stationary mean value

$$s = \mathsf{E}(\text{net inventory position just before the arrival of a replenishment order}).$$

We define the net inventory position as the inventory on hand

$$Y_{net} = Y.$$

The net inventory position is usually defined as the on-hand inventory minus backorders. In our system customer demand is not backordered. Therefore, the above definition of net inventory is natural for the service facility with inventory and lost sales as defined in Definition 2.1 since the customer who is in service will not use one piece from inventory until he leaves the system. Another approach to evaluate the net inventory position Y_{net} will be investigated in Section 5. In the system under consideration the expected net inventory just before a replenishment order arrives is 0 but we shall need the general form of the safety stock s later in Sections 3 and 4.

Service levels are defined for a specified time period which can be the average replenishment cycle or one unit of time for example. Service levels based on the replenishment cycle simplify computations. To the customers, however, the length of an order cycle is of minor interest, they are interested in the quality of service in every unit of time. The definition of the α-service levels is a combination of the definitions given by [15] and [18]. The definition of the β-service level is standard and can be found in [15] and [18]. Note that the underlying probability measures may be casewise different.

Definition 2.11.

$$\alpha_1 = \mathsf{P}(\text{net inventory position at the end of a cycle } > 0),$$

$$\alpha_2 = \mathsf{P}(\text{net inventory position at the end of a unit of time } > 0),$$

$$\beta = \frac{\mathsf{E}(\text{demand satisfied per cycle})}{\mathsf{E}(\text{total demand per cycle})} \overset{\star}{=} \frac{\mathsf{E}(\text{demand satisfied per unit of time})}{\mathsf{E}(\text{total demand per unit of time})}.$$

Here $\overset{\star}{=}$ is justified for all regenerative processes and therefore is here a conclusion of the ergodicity of Z. α-service levels are event-oriented performance measures. They only value the occurence but not the magnitude of a shortage. The definition of an α-service level depends on the chosen time interval, it represents the fraction of replenishment cycles or time units without a negative net inventory position. In the present case the α_2-service level is just the time stationary probability of a non-negative net inventory position. It is rarely considered in the literature. The β-service level is a quantity-oriented service measure describing the proportion of demands that are met from stock without accounting for the duration of a stockout. β-service levels are widely used in practice [21].

According to Definition 2.10 of the net inventory position Y_{net} the α_1-service level is just 0 in the system defined in Definition 2.1 since the net inventory position is always 0 when an order arrives. (Note that if we use the definition from [15] $\alpha_1 = \mathsf{P}(Y_{net}$ at the end of a cycle $\geq 0)$ then the α_1-service level will be equal to 1 for any M/M/1-system with inventory management and lost sales since with Definition 2.10 the net inventory position Y_{net} cannot become negative.) $\alpha_2 = \mathsf{P}(Y > 0)$. For the β-service level we compute

$$\beta = \frac{\mathsf{E}(\text{satisfied demand per unit of time})}{\mathsf{E}(\text{total demand per unit of time})} = \frac{\lambda - \overline{LS}}{\lambda} = \frac{\bar{p}}{\bar{p} + \overline{LS}_c} = 1 - \frac{\lambda}{\bar{p}\nu + \lambda}.$$

Hence

$$\beta = \mathsf{P}(Y > 0) = \alpha_2 \quad \text{if } Y_{net} = Y,$$

since the proportion of demand which is lost is just controlled by $Y = 0$. Normally, α- and β-service levels are not in a universally valid proportion to each other. Furthermore, since $\mathsf{P}(Y = 0) = \lambda_R/\nu$ we find

$$\lambda_R \bar{\alpha}_1 = \nu \bar{\alpha}_2,$$

where $\bar{\alpha}_i = 1 - \alpha_i$, $i = 1, 2$. In the classical inventory management literature the relation between α_1 and α_2 is $\lambda_R \bar{\alpha}_1 = \bar{\alpha}_2$.

α- and β-service levels do not provide information on the length of the waiting time that a customer may experience. In contrast to these quantity related service measures, a second stream in the literature therefore uses a time criteria to measure inventory control performance. This approach becomes even more important in case of non-zero and stochastic service times, i.e. customer orders are not filled instantaneously from on-hand inventory as assumed in most inventory management systems. For example, Tempelmeier suggests in [20] and [21] to analyse an inventory model with a service constraint on the expected customer waiting time or on the probability that the customer waiting time is larger than a prespecified constant.

From Corollary 2.7 (a) the mean number of customers in the system, $\bar{L}_0$, is the same as in the classical M/M/1/∞-system with parameters λ, μ. For evaluating the mean number of waiting customers, $\bar{L}$, we must take into consideration that the waiting time of a customer formally ends at the moment when he enters the server since there are no more customers waiting in front of him. Nevertheless it may happen that his service does not immediately start then because of a stockout. Taking this into consideration we exclude the replenishment lead time from the customer waiting time. It follows that $\bar{L}$ is the same as in the classical M/M/1/∞-system with parameters λ, μ. However, the mean total time in system and the mean waiting time of the customers are different from the mean total time in system and the expected waiting time in the M/M/1/∞-system since customers are not arriving with overall intensity λ because of the interruption of the arrival stream during the stochastic replenishment lead time. As can be found in Theorem A.1 in Appendix A the mean number of customers arriving per unit time is

$$\lambda_A = \bar{p}\lambda_R = \frac{\bar{p}\lambda\nu}{\bar{p}\nu + \lambda}.$$

Note that for the throughput of the system the following relation holds $\lambda_A = \lambda \mathsf{P}(Y > 0) = \lambda - \overline{LS}$.

From Little's formula the customers' mean sojourn time $\bar{W}_0$ and mean waiting time $\bar{W}$ are

$$\bar{W}_0 = \frac{\bar{L}_0}{\lambda_A} = \frac{\bar{p}\nu + \lambda}{\bar{p}\nu(\mu - \lambda)} = \frac{1}{\mu - \lambda} + \frac{\lambda}{\bar{p}\nu(\mu - \lambda)}, \tag{7}$$

$$\bar{W} = \frac{\bar{L}}{\lambda_A} = \frac{(\bar{p}\nu + \lambda)\lambda}{\bar{p}\nu\mu(\mu - \lambda)} = \frac{\lambda}{\mu(\mu - \lambda)} + \frac{\lambda^2}{\bar{p}\nu\mu(\mu - \lambda)}. \tag{8}$$

$\bar{W}_0$ and $\bar{W}$ are naturally larger than the mean sojourn time and mean waiting time of the classical M/M/1/∞-system respectively, which are just the first summands of the above expressions. The second term is $\lambda/(\bar{p}\nu)$ times the first summand. These characteristics of $\bar{W}_0$ and $\bar{W}$ just result from the mean interarrival time of customers who are admitted to the system, $\lambda_A^{-1} = \lambda^{-1} + (\bar{p}\nu)^{-1}$, which is larger than in the classical M/M/1/∞-system.

Remark 2.12. Consider extremal cases (see Remark 2.6): If $\bar{p} \to \infty$ (and necessarily $M \to \infty$), then asymptotically the system behaves like the classical M/M/1/∞-system with the same $\lambda_A = \lambda$, $\bar{W}_0$ and $\bar{W}$. The inventory size and the cycle length are infinite, no demand is lost and $\beta = 1$. If $\nu \to \infty$ then λ_A, $\bar{W}_0$ and $\bar{W}$ are the same as in the classical M/M/1/∞-system, no demand is lost and $\beta = 1$ as well. However, in this case the cycle length is finite: $\lambda_R = \lim_{\nu\to\infty} \lambda(\bar{p} + \frac{\lambda}{\nu})^{-1} = \lambda/\bar{p}$.

Remark 2.13. Note that only $\bar{W}_0$ and $\bar{W}$ depend on the service rate μ. Some performance measures are not dependent on λ and ν individually but only on their proportion λ/ν, e.g. $\bar{I}$, $\overline{LS}_c$ and β. Concerning the influence of F_p we observe that several performance measures only depend on the first moment of F_p like λ_R, $\overline{LS}$, β and $\bar{W}$ or are completely independent of F_p like $\overline{LS}_c$ and α_1. $\bar{I}$ depends on the first and second moment of F_p.
Hence, for two systems which have the same parameters λ, ν and μ but different order size distributions F_p and $F_{\tilde{p}}$ with the same mean $\bar{p}$ only $\bar{I}$ will be different.

2.2 Examples for specific order size distribution

In this section we investigate two examples for the replenishment order size distribution. We consider the fixed order size Q, which yields an $(0, Q)$-policy and the system with uniformly distributed order sizes on $\{1, ..., Q\}$. In both cases holds $M = Q$. The performance measures for these examples are summarized in Table 1. (It will turn out that several of the entries of the table could be obtained directly from the results of the following sections.)

Deterministic order size. Let us assume that the order size is fixed and equal to $Q \in \mathbb{N}$. Hence, we have $p_k = \delta_{kQ}$ for all $k \in \{1, \ldots, M\}$. Then $q_k = 1$ for all $k \in 1, \ldots, Q$ and $q_k = 0$ otherwise. The mean order size is $\bar{p} = Q$ and the normalization constant is $K = \frac{\mu}{\mu-\lambda}(Q + \frac{\lambda}{\nu})$. Note that as in Berman and Sapna's paper [4] (with backordering) the steady state probabilities for the on-hand inventory to be $k \geq 1$ are equally distributed according to

$$\mathsf{P}(Y = k) = \left(Q + \frac{\lambda}{\nu}\right)^{-1}, \quad 1 \leq k \leq Q.$$

In contrast to their results the probability that a replenishment order is outstanding here has a different probability:

$$\mathsf{P}(Y = 0) = \frac{\lambda}{\nu}\left(Q + \frac{\lambda}{\nu}\right)^{-1}.$$

Uniformly distributed order size. Let the size of a replenishment order be equally distributed on $\{1, \ldots, Q\}$ (Unif($\{1, \ldots, Q\}$)), hence $p_k = 1/Q$, for all $k \in \{1, \ldots, Q\}$ and $p_k = 0$ otherwise. Here $\bar{p} = \frac{Q+1}{2}$ and $q_k = \sum_{h=k}^{Q} p_h = \frac{Q+1-k}{Q}$. The normalization constant is $K = \frac{\mu}{\mu-\lambda}(\frac{Q+1}{2} + \frac{\lambda}{\nu})$.

2.3 Ordering result

Consider two single server systems as in Definition 2.1 which differ only in their probability distribution functions F_p and $F_{\tilde{p}}$ for the size of replenishment orders. The mean order sizes in the two systems are assumed to be the same. Denote by X, $\tilde{X}$ and Y, $\tilde{Y}$ the random variables distributed like the stationary queue length and inventory distribution of the two systems respectively. Then, the following theorem holds (for

Table 1 Performance measures

F_p	Deterministic	Uniform
$\bar{I}$	$\frac{Q}{Q+\frac{\lambda}{\nu}}\frac{Q+1}{2}$	$\frac{(Q+2)[2\lambda+(Q+1)\nu]}{6\lambda}$
λ_R	$\frac{\lambda}{Q+\frac{\lambda}{\nu}}$	$\frac{2\lambda\nu}{(Q+1)\nu+2\lambda}$
$\overline{LS}$	$\frac{\lambda^2}{Q\nu+\lambda}$	$\frac{2\lambda^2}{(Q+1)\nu+2\lambda}$
β	$\frac{Q}{Q+\frac{\lambda}{\nu}}$	$\frac{(Q+1)\nu}{(Q+1)\nu+2\lambda}$
$\bar{W}$	$\frac{(Q\nu+\lambda)\lambda}{Q\nu\mu(\mu-\lambda)}$	$\frac{[(Q+1)\nu+2\lambda]\lambda}{(Q+1)\nu\mu(\mu-\lambda)}$

definition of $<_{cx}$ and $<_{st}$ see e.g. [19] page 6 and 8 respectively):

Theorem 2.14. *If* $F_p <_{cx} F_{\tilde{p}}$ *then* $(X, Y) <_{st} (\tilde{X}, \tilde{Y})$.

Proof: From Condition (iii) ensuing Theorem A in [19] section 1.3, p. 11, $F_p <_{cx} F_{\tilde{p}}$ is equivalent to $\sum_{h=k}^{\infty} \bar{F}_p(k) \leq \sum_{h=k}^{\infty} \bar{F}_{\tilde{p}}(k)$ for all $k \in \mathbb{N}$. This is $\sum_{h=k}^{\infty} q_k \leq \sum_{h=k}^{\infty} \tilde{q}_k$ for all $k \in \mathbb{N}$. Hence, $Y <_{st} \tilde{Y}$ since $\mathsf{P}(Y > k) \leq \mathsf{P}(\tilde{Y} > k)$ holds for all $k \in \mathbb{N}$ and $k = 0$. Moreover, because steady state probabilities are in product form (1) and λ, μ and ν are the same in both systems, the result follows for $k \in \mathbb{N}_0$ ($\mathsf{P}(X = n, Y > k) \leq \mathsf{P}(\tilde{X} = n, \tilde{Y} > k)$ for all $n \in \mathbb{N}_0$ and $k \in \mathbb{N}_0$). Thus, $(X, Y) <_{st} (\tilde{X}, \tilde{Y})$. □

Remark 2.15. The convex ordering of two distribution functions F_p and $F_{\tilde{p}}$, with finite and equal first moments can be ascertained using the Karlin-Novikoff cut-criterion (see for example [19] Section 1.3, Theorem E, p. 17). Suppose that for F_p and $F_{\tilde{p}}$ we have

$$F_p(k) \leq F_{\tilde{p}}(k), \quad \text{for } k \leq \xi,$$
$$F_p(k) \geq F_{\tilde{p}}(k), \quad \text{for } k > \xi,$$

for some $\xi \in \mathbb{R}_+$ then

$$F_p <_{cx} F_{\tilde{p}}.$$

Example 2.16. Let $p \sim \delta_Q$, then $F_p(k) = 0$ for all $k < Q$ and $F_p(k) = 1$ for all $k \geq Q$ and $\bar{p} = Q$. Hence, the Karlin-Novikoff cut criterion can be applied for comparison with all other distribution functions $F_{\tilde{p}}$ with mean Q. An example which has already been studied above is

$$p \sim \delta_{(Q+1)/2} <_{cx} \tilde{p} \sim \text{Unif}(\{1, \ldots, Q\}).$$

Corollary 2.17. *If the mean replenishment size is prescribed, the stationary mean inventory position* $\bar{I}$ *is minimal for the system with fixed order size.*

This can be seen as follows: Consider two systems with different probability functions p and $\tilde{p}$, with $p \sim \delta_Q$ and $\mathsf{E}\tilde{p} = Q$. Then $p <_{cx} \tilde{p}$. If we denote by Y and $\tilde{Y}$ the random variables distributed like the stationary inventory distribution of the two systems it follows from Theorem 2.14 that $Y <_{st} \tilde{Y}$. Hence,

$$\bar{I} = \sum_{k \in \mathbb{N}} \mathsf{P}(Y \geq k) \leq \sum_{k \in \mathbb{N}} \mathsf{P}(\tilde{Y} \geq k) = \bar{\tilde{I}}.$$

Summarized: We have found out that all performance measures except the mean inventory position $\bar{I}$ are the same for two systems that differ only in the order size distribution function F_p but have the same mean $\bar{p}$. $\bar{I}$ is minimal for the fixed order size system from Section 2.2. This is why all other systems can be excluded from further consideration when it comes to an optimization of the costs associated to M/M/1/∞-systems with inventory management and reorder point 0.

For this reason we will now investigate early replenishment systems with reorder point $r \geq 0$ and fixed order size Q. This yields the well-known (r, Q)-system from inventory management theory.

3 M/M/1/∞-system with (r, Q)-policy

In this section we consider an M/M/1/∞-system with inventory management policy that corresponds to the lost sales version of a continuous review, order-point, order-quantity model, typically called (r, Q) model, where a fixed order quantity Q is ordered each time the on-hand inventory reaches the reorder point r (see [22] p. 206 and [18] p. 256). The (r, Q)-policy is applicable if units are demanded one at a time and if transaction reporting is used. We assume throughout the paper that $r < Q$. This avoids degenerate cycles in which no demand occurs.

*Definition 3.1 (*M/M/1/∞-system with (r, Q)-policy). We have a single server with infinite waiting room under FCFS regime and an attached inventory of capacity M as in Definition 2.1 with the assumptions on the stochastic behaviour of the system from Definition 2.2.

If the on-hand inventory reaches a prespecified value $r \geq 0$, a replenishment order is instantaneously triggered. The size of the replenishment order is fixed to $Q < \infty$ units, $Q > r$. We fix $M = r + Q$.

The replenishment lead time is exponentially distributed with parameter ν, $\nu > 0$. During the time the inventory is zero, arriving customers are lost.

Let $X(t)$ denote the number of customers present at the server at time $t \geq 0$, either waiting or in service and let $Y(t)$ denote the on-hand inventory at time $t \geq 0$. Then $Z = ((X(t), Y(t)), t \geq 0)$ is a continuous-time Markov process for the M/M/1/∞-system with (r, Q)-policy. The state space of Z is $E_Z = \{(n, k) : n \in \mathbb{N}_0, 0 \leq k \leq Q + r\}$.

If $\nu = \infty$, early replenishment of inventory with $r > 0$ does not make sense with respect to economical aspects, since r units are never touched by the customers and remain in stock for ever.

Theorem 3.2. *The continuous-time Markov process Z from Definition 3.1 is ergodic if and only if $\lambda < \mu$ and $\nu < \infty$. If Z is ergodic then it has a unique limiting and stationary distribution of product form given by*

$$\pi(n,k) = K^{-1}C(k)\left(\frac{\lambda}{\mu}\right)^n, \quad \text{for } n \in \mathbb{N}_0, 1 \le k \le Q+r, \tag{9}$$

$$\pi(n,0) = K^{-1}\frac{\lambda}{\nu}\left(\frac{\lambda}{\mu}\right)^n, \qquad \text{for } n \in \mathbb{N}_0, \tag{10}$$

with

$$C(k) = \left(\frac{\lambda+\nu}{\lambda}\right)^{k-1}, \quad k = 1,\dots,r,$$
$$C(k) = \left(\frac{\lambda+\nu}{\lambda}\right)^{r}, \qquad k = r+1,\dots,Q,$$
$$C(Q+k) = \left(\frac{\lambda+\nu}{\lambda}\right)^{r} - \left(\frac{\lambda+\nu}{\lambda}\right)^{k-1}, \quad k = 1,\dots,r.$$

The normalization constant is

$$K = \frac{\mu}{\mu-\lambda}\left(Q\left(\frac{\lambda+\nu}{\lambda}\right)^r + \frac{\lambda}{\nu}\right). \tag{11}$$

Proof: The process Z from Definition 3.1 possesses the following global balance equations:

$$\pi(n,k)\big(\lambda + \mu(1-\delta_{0n}) + \nu 1_{\{k\le r\}}\big) = \pi(n-1,k)\lambda(1-\delta_{0n}) + \pi(n+1,k+1)\mu(1-\delta_{Q+r,k}) + \pi(n,k-Q)\nu 1_{\{k\ge Q\}}, \quad 1 \le k \le Q+r, \tag{12}$$

$$\pi(n,0)\nu = \pi(n+1,1)\mu \tag{13}$$

for all $n \in \mathbb{N}_0$. We will show that the distribution from (9) and (10) satisfies Eqs. (12) and (13). Equation (13) is satisfied since $C(1) = 1$. In (12) the term $\pi(n,k)\mu(1-\delta_{0n})$ at the left cancels against the term $\pi(n-1,k)\lambda(1-\delta_{0n})$ at the right side for all $n \in \mathbb{N}_0$ and $1 \le k \le Q+r$. It remains to prove that the term on the left side

$$\pi(n,k)\big(\lambda + \nu 1_{\{k\le r\}}\big) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n C(k)\big(\lambda + \nu 1_{\{k\le r\}}\big)$$

equals the remaining terms at the right side

$$\pi(n+1,k+1)\mu(1-\delta_{Q+r,k}) + \pi(n,k-Q)\nu 1_{\{k\ge Q\}}$$
$$= K^{-1}\left(\frac{\lambda}{\mu}\right)^n \lambda C(k+1)(1-\delta_{Q+r,k}) + K^{-1}\left(\frac{\lambda}{\mu}\right)^n \lambda 1_{\{k=Q\}} + K^{-1}\left(\frac{\lambda}{\mu}\right)^n \nu C(k-Q)1_{\{k>Q\}}$$

for all $n \in \mathbb{N}_0$ and $1 \le k \le Q+r$. Hence, we have to show that for all $1 \le k \le Q+r$

$$C(k+1)(1-\delta_{Q+r,k}) = C(k)\left(1 + \frac{\nu}{\lambda}1_{\{k\le r\}}\right) - 1_{\{k=Q\}} - \frac{\nu}{\lambda}C(k-Q)1_{\{k>Q\}} \tag{14}$$

holds. The equations which have to be satisfied are

$$C(k+1) = C(k)\left(1+\frac{\nu}{\lambda}\right), \quad k = 1,\dots,r,$$
$$C(k+1) = C(k), \quad k = r+1,\dots,Q-1,$$
$$C(Q+1) = C(Q) - 1,$$
$$C(k+1) = C(k) - \frac{\nu}{\lambda}C(k-Q), \quad k = Q+1,\dots,Q+r-1,$$
$$0 = C(Q+r) - \frac{\nu}{\lambda}C(r).$$

Inserting the $C(k)$ inductively into these equations finishes the proof. □

As in Section 2 the normalization constant K factorizes into

$$K_X := \frac{\mu}{\mu-\lambda} \quad \text{and} \quad K_Y := \left(Q\left(\frac{\lambda+\nu}{\lambda}\right)^r + \frac{\lambda}{\nu}\right).$$

The steady state queue length of $X = (X(t), t \ge 0)$ is distributed like the steady state queue length in the classical M/M/1/∞-FCFS system with the same parameters λ and μ.

Remark 3.3. Buchanan and Love [8] investigate an (r, Q) inventory system with lost sales and Erlang-distributed lead times, where customer demands are satisfied without any loss of time. Their result for the special case of exponentially distributed lead times coincides with the marginal distribution of the inventory process of our system.

Theorem 3.4 (Measures of system performance). *The measures of system performance as defined in Section 2.1 of the M/M/1/∞-system with (r, Q)-policy according to*

Definition 3.1 *are given by*

$$\bar{I} = \frac{Q}{Q+\frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}\left(\frac{Q+1}{2}+s\right), \tag{15}$$

$$\lambda_R = \frac{\lambda}{Q+\frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}, \tag{16}$$

$$\lambda_A = \frac{\lambda Q}{Q+\frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}, \tag{17}$$

$$\overline{LS} = \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{\lambda}{Q+\frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}, \tag{18}$$

$$\overline{LS}_c = \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{19}$$

$$s = r - \frac{\lambda}{\nu}\left(1-\left(\frac{\lambda}{\lambda+\nu}\right)^r\right), \tag{20}$$

$$\alpha_1 = 1-\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{21}$$

$$\alpha_2 = \beta = \frac{Q}{Q+\frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}, \tag{22}$$

$$\bar{W}_0 = \frac{1}{\mu-\lambda} + \frac{\lambda}{Q\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{1}{\mu-\lambda}, \tag{23}$$

$$\bar{W} = \frac{\lambda}{\mu(\mu-\lambda)} + \frac{\lambda}{Q\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{\lambda}{\mu(\mu-\lambda)}. \tag{24}$$

Proof: The proof of Theorem 3.4 is presented in Appendix B. □

Some remarks may be in order here. Obviously, $\bar{I} < \frac{Q+1}{2}+s$. The expected number of replenishments per unit time λ_R has the same form as in Section 2.1 (mean demand per unit time/(mean satisfied demand per cycle + mean lost sales per cycle)) and $\lambda_A = \lambda_R \cdot$ (mean satisfied demand per cycle).

The safety stock s is equal to the difference of the reorder point r and the expected lead time demand which can be satisfied. $s = 0$ for $r = 0$ and s is strictly increasing in the reorder point r since for all $r \geq 1$:

$$\begin{aligned} s(r) - s(r-1) &= 1 - \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^{r-1}\left(1-\frac{\lambda}{\lambda+\nu}\right) \\ &= 1-\left(\frac{\lambda}{\lambda+\nu}\right)^r > 0. \end{aligned} \tag{25}$$

α_1 can be interpreted as the proportion of cycles without a stockout or as the proportion of lead time demand which can be satisfied from stock. This explains the relation $s = r - \frac{\lambda}{\nu}\alpha_1$ once again. α_1 is strictly increasing in r.

As before $\beta = \mathsf{P}(Y > 0) = \alpha_2 = \lambda_A/\lambda$. $\bar{W}$ and $\bar{W}_0$ are again larger than in the classical M/M/1/∞-system, since $\bar{W} = \bar{W}_0 \frac{\lambda}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}(1+\frac{\overline{LS}_c}{Q})$.

Remark 3.5. Remark 2.12 carries over to the cases $Q \to \infty$ and $\nu \to \infty$ in the M/M/1/∞-system with (r, Q)-policy as defined in Definition 3.1. Furthermore, if $\nu \to \infty$ the mean inventory on hand is $(Q+1)/2 + r$ and r units remain in stock for ever.

Remark 3.6. As in Section 2 the only performance measures that depend on the service rate μ are $\bar{W}_0$ and $\bar{W}$. The values for $\bar{I}$, $\overline{LS}_c$, s and β depend on λ and ν only through the ratio λ/ν. Concerning the influence of Q and r we observe that all performance measures depend on the reorder point r, but only λ_R, λ_A, $\overline{LS}$, β, $\bar{W}_0$ and $\bar{W}$ are also dependent on Q.

4 M/M/1/∞-system with (r, S)-policy

The M/M/1/∞-system with inventory management studied in this section corresponds to the lost sales version of a continuous review, order-point, order-up-to-level (r, S)-system. Each time the on-hand inventory reaches the reorder point r a variable replenishment quantity is used such that upon replenishment, the on-hand inventory is restocked to level S (see [5] and [17] for an equivalent definition of the (r, S)-policy in the backorder case). In the inventory management literature another definition of the (r, S)-system has appeared as well (see [18] p. 256, [11]), where the order size is determined at the moment the order is placed such that the actual inventory level $x < r$ plus the variable order size is equal to S. In our case, where demands are unit-sized, no overshoot of the order point r is possible and the (r, Q)- and (r, S)-systems are identical when the second definition is used.

*Definition 4.1 (*M/M/1/∞-system with (r, S)-policy). We have a single server with infinite waiting room under FCFS regime and an attached inventory of capacity M as in Definition 2.1 with the assumptions on the stochastic behaviour of the system from Definition 2.2.

If the inventory reaches a prespecified value $r > 0$, a replenishment order is instantaneously triggered. With each replenishment the inventory level is restocked to exactly $S < \infty$ units with $r < S$ no matter how many items are still present in the inventory. We set $M = S$.

The replenishment lead time is exponentially distributed with parameter ν, $\nu > 0$. During the time the inventory is zero, no customers are admitted to join the queue and are lost.

Let $X(t)$ denote the number of customers present at the server at time $t \geq 0$, either waiting or in service and let $Y(t)$ denote the on-hand inventory at time $t \geq 0$. Then $Z = ((X(t), Y(t)), t \geq 0)$ is a continuous-time Markov process for the M/M/1/∞-system with (r, S)-policy. The state space of Z is given by $E_Z = \{(n, k) : n \in \mathbb{N}_0, 0 \leq k \leq S\}$.

As in Sections 2 and 3 an increasing order-up-to-level $S \to \infty$ results in the classical M/M/1/∞-system and an early replenishment of the inventory with $r > 0$ does not make sense if $\nu = \infty$.

Theorem 4.2. *The continuous-time Markov process Z from Definition 4.1 is ergodic if and only if $\lambda < \mu$ and $\nu < \infty$. If Z is ergodic then it has a unique limiting and stationary distribution of product form given by*

$$\pi(n,k) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n C(k), \quad \text{for } n \in \mathbb{N}_0, 1 \leq k \leq S, \tag{26}$$

$$\pi(n,0) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n \frac{\lambda}{\nu}, \quad \text{for } n \in \mathbb{N}_0, \tag{27}$$

with

$$C(k) = \left(\frac{\lambda+\nu}{\lambda}\right)^{k-1}, \quad k = 1, \ldots, r,$$

$$C(k) = \left(\frac{\lambda+\nu}{\lambda}\right)^{r}, \quad k = r+1, \ldots, S$$

and normalization constant

$$K = \frac{\mu}{\mu-\lambda}\left(S - r + \frac{\lambda}{\nu}\right)\left(\frac{\lambda+\nu}{\lambda}\right)^r. \tag{28}$$

Proof: The process Z from Definition 4.1 possesses the following global balance equations:

$$\begin{aligned}\pi(n,k)\big(\lambda + \mu(1-\delta_{0n}) + \nu 1_{\{k\leq r\}}\big) &= \pi(n-1,k)\lambda(1-\delta_{0n}) + \pi(n+1,k+1)\mu(1-\delta_{Sk}) \\ &\quad + \sum_{h=0}^{r}\pi(n,h)\nu 1_{\{k=S\}}, \quad 1 \leq k \leq S, \end{aligned}\tag{29}$$

$$\pi(n,0)\nu = \pi(n+1,1)\mu \tag{30}$$

for all $n \in \mathbb{N}_0$. We will show that the distribution from (26) and (27) satisfies Eqs. (29) and (30). Equation (30) is satisfied since $C(1) = 1$. In (29) the term $\pi(n,k)\mu(1-\delta_{0n})$ at the left cancels against the term $\pi(n-1,k)\lambda(1-\delta_{0n})$ at the right side for all $n \in \mathbb{N}_0$ and $1 \leq k \leq S$. Moreover, we will prove that the term on the left side

$$\pi(n,k)\big(\lambda + \nu 1_{\{k\leq r\}}\big) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n C(k)\big(\lambda + \nu 1_{\{k\leq r\}}\big)$$

corresponds to the following expression at the right side

$$\begin{aligned}&\pi(n+1,k+1)\mu(1-\delta_{Sk}) + \sum_{h=0}^{r}\pi(n,h)\nu 1_{\{k=S\}} \\ &= K^{-1}\left(\frac{\lambda}{\mu}\right)^n \lambda C(k+1)(1-\delta_{Sk}) + K^{-1}\left(\frac{\lambda}{\mu}\right)^n \lambda 1_{\{k=S\}} \\ &\quad + K^{-1}\left(\frac{\lambda}{\mu}\right)^n \nu \sum_{h=1}^{r} C(h) 1_{\{k=S\}}\end{aligned}$$

for all $n \in \mathbb{N}_0$ and $1 \leq k \leq S$. We therefore have to show that for all $1 \leq k \leq S$

$$\begin{aligned}C(k+1)(1-\delta_{S,k}) &= C(k)\left(1 + \frac{\nu}{\lambda}1_{\{k\leq r\}}\right) - 1_{\{k=S\}} \\ &\quad - \frac{\nu}{\lambda}\sum_{h=1}^{r} C(h) 1_{\{k=S\}}\end{aligned}\tag{31}$$

holds. Hence, the equations which have to be satisfied by $C(k)$, $1 \leq k \leq S$ are

$$C(k+1) = C(k)\left(1 + \frac{\nu}{\lambda}\right), \quad k = 1, \ldots, r,$$

$$C(k+1) = C(k), \quad k = r+1, \ldots, S-1,$$

$$0 = C(S) - 1 - \frac{\nu}{\lambda}\sum_{h=1}^{r} C(h).$$

That these equations are solved by the $C(k)$ can be seen by direct insertion. □

The normalization constant K can be split into two parts

$$K_X := \frac{\mu}{\mu-\lambda} \quad \text{and} \quad K_Y := \left(S - r + \frac{\lambda}{\nu}\right)\left(\frac{\lambda+\nu}{\lambda}\right)^r.$$

As in Sections 2 and 3 the steady state queue length distribution of $X = (X(t), t \geq 0)$ is equal to the steady state queue length distribution in the classical M/M/1/∞-FCFS system with the same parameters λ and μ.

Theorem 4.3 (Measures of system performance). *The measures of system performance as defined in Section* 2.1 *of the M/M/1/∞-system with (r, S)-policy according to*

Definition 4.1 *are*

$$\bar{I} = \frac{1}{S-r+\frac{\lambda}{\nu}}\left\{\frac{\lambda}{\nu}s+\frac{(S+1)S-(r+1)r}{2}\right\}, \tag{32}$$

$$\lambda_R = \frac{\lambda}{S-r+\frac{\lambda}{\nu}}, \tag{33}$$

$$\lambda_A = \lambda - \frac{\lambda^2}{(S-r)\nu+\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{34}$$

$$\overline{LS} = \frac{\lambda^2}{(S-r)\nu+\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{35}$$

$$\overline{LS_c} = \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{36}$$

$$s = r-\frac{\lambda}{\nu}\left(1-\left(\frac{\lambda}{\lambda+\nu}\right)^r\right), \tag{37}$$

$$\alpha_1 = 1-\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{38}$$

$$\alpha_2 = \beta = 1-\frac{\lambda}{(S-r)\nu+\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r, \tag{39}$$

$$\bar{W}_0 = \frac{1}{\mu-\lambda}+\frac{\lambda}{(S-s)\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r\frac{1}{\mu-\lambda}, \tag{40}$$

$$\bar{W} = \frac{\lambda}{\mu(\mu-\lambda)}+\frac{\lambda}{(S-s)\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r\frac{\lambda}{\mu(\mu-\lambda)}. \tag{41}$$

Proof: The proof of Theorem 4.3 is presented in Appendix B. □

Remark 4.4. The dependences of the performance measures on the several parameters are the same as in Section 3.

Theorem 4.5 (Comparison of (r, Q)- and (r, S)-policy with the same r). Let $S = Q + s$, where s is the safety stock from (20). Then, the stationary mean inventory position in the M/M/1/∞-system with (r, S)-policy is not larger than the stationary mean inventory position in the M/M/1/∞-system with (r, Q)-policy. All other performance measures are the same.

Proof: The proof is given in Appendix B. □

5 Net inventory revisited

In Definition 2.10 of Section 2.1 we defined the net inventory position as the inventory on hand $Y_{net} = Y$. In the context of service facilities with inventory and lost sales a second common-sense definition of the net inventory position is

$$Y_{net} = \max\{0, Y - X\}. \tag{42}$$

This definition is reasonable since one can argue that for each customer in queue who has been admitted to the system one piece from the inventory is already reserved.

In the following theorems the performance measures dependent on Y_{net}, the safety stock s and the α_1-service level, will be given for all the inventory models considered above using the second definition of Y_{net} from (42).

Theorem 5.1. *In the M/M/1/∞-system with inventory management, reorder point* 0 *and arbitrary size of replenishment orders according to Definition* 2.1*, it follows*

$$s = 0 \quad \textit{and} \quad \alpha_1 = 0,$$

if the net inventory position is defined by (42)*.*

Proof: See Appendix B. □

Theorem 5.2. *In the M/M/1/∞-system with (r, Q)- and (r, S)-policy according to Definition* 3.1 *and* 4.1 *respectively, it follows*

$$s = r-\frac{\lambda}{\nu}\left(1-\left(\frac{\lambda}{\lambda+\nu}\right)^r\right)-\frac{\lambda}{\mu-\lambda}\alpha_1$$

and

$$\alpha_1 = 1-\frac{1}{\mu-(\lambda+\nu)}\left((\mu-\lambda)\left(\frac{\lambda}{\lambda+\nu}\right)^r-\nu\left(\frac{\lambda}{\mu}\right)^r\right),$$

if the net inventory position is defined by (42)*.*

Proof: The proof is given in Appendix B. □

We observe that α_1 and s now depend on the service rate μ. α_1 can be interpreted as the proportion of cycles where at the end of the cycle the number of customers is smaller than the size of the on-hand inventory. $\lambda/(\mu-\lambda)$ is the mean number of customers in the system. Hence, s is equal to the reorder point r minus the lead time demand which can be satisfied from stock minus α_1 times the mean number of customers in the system.

Corollary 5.3. *In the M/M/1/∞-system with (r, Q)- or (r, S)-policy according to Definition 3.1 and 4.1 respectively, α_1 has the following properties if the net inventory position is as in* (42)*:*

1. $0 \le \alpha_1 \le 1 - (\frac{\lambda}{\mu})^r$ *for some fixed* $r \in \mathbb{N}_0$,
2. $\alpha_1(\cdot)$ *is strictly increasing in* $r \in \mathbb{N}_0$.

Proof. See Appendix B. □

Corollary 5.4. *In the M/M/1/*∞*-system with* (r, Q)*- or* (r, S)*-policy according to Definition 3.1 and 4.1 respectively, with net inventory position defined as in* (42) *the following inequality holds*

$$s \le r - \frac{\lambda}{\nu}\left(1 - \left(\frac{\lambda}{\lambda+\nu}\right)^r\right).$$

Proof: As shown in Corollary 5.3 it holds for all $r \in \mathbb{N}_0$ that $\alpha_1 \ge 0$ and α_1 is increasing in r. $\mu > \lambda$ follows from the ergodicity criterion, thus, the second summand of s is non-positive and decreasing in r. This completes the proof. □

6 M/M/1/N-1-systems with inventory and lost sales

In this section we consider the systems from Sections 2, 3 and 4 with a limited number of $N-1$ waiting places, such that only N customers are admitted to the system at the same time. We assume that during the time a replenishment order is outstanding the service place can be used as a waiting place by the customers in the system.

Recall the discussion on page 59 on the restriction of the order policies. The main concern was that a very large queue length does not imply rejection of further arrivals due to having some (or even only a few) items still in the inventory. To a certain extent this drawback of our previous models is now removed because customers that arrive when all waiting places are occupied are immediately rejected although there may be still items in the inventory.

6.1 M/M/1/N-1-system with inventory, reorder point 0, and random size of replenishment orders

*Definition 6.1 (*M/M/1/N-1-system with inventory). We have a single server as described in Definition 2.1 with a limited number of $N-1$ waiting places under FCFS regime. If at the moment of arrival of a customer the number of customers present at the system is less than N and the on-hand inventory is positive, the arriving customer is admitted to the system, otherwise the arriving customer is not admitted to the system. She disappears and never returns.

The assumptions on the stochastic behaviour of the system are as in Definition 2.2.
The attached inventory has capacity of M items.
The randomized order policy is determined according to Definition 2.3.

Let $X(t)$ denote the number of customers present at the server at time $t \ge 0$, either waiting or in service (queue length) and let $Y(t)$ denote the on-hand inventory at time $t \ge 0$. Then $Z = ((X(t), Y(t)), t \ge 0)$ is a continuous-time Markov process for the M/M/1/N-1-system with inventory management, reorder point 0 and random size of replenishment orders. The state space of Z is $E_{Z_N} = \{(n,k) : n \in \{0,1,\ldots,N\}, 1 \le k \le M\} \cup \{(n,0) : n \in \{0,1,\ldots,N-1\}\}$, since the inventory can only be depleted after a customer has been served who took the last item from inventory and no customers join the queue during the replenishment lead time.

Theorem 6.2. *The continuous-time Markov process Z from Definition 6.1 is ergodic and has a unique limiting and stationary distribution of product form given by*

$$\pi(n,k) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n q_k, \quad \textit{with} \quad 0 \le n \le N,\ M \ge k \ge 1, \tag{43}$$

$$\pi(n,0) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n \frac{\lambda}{\nu}, \quad \textit{with} \quad 0 \le n \le N-1, \tag{44}$$

and with normalization constant

$$K = \frac{\mu^{N+1}\left(\bar{p} + \frac{\lambda}{\nu}\right) - \lambda^{N+1}\left(\bar{p} + \frac{\mu}{\nu}\right)}{\mu^N(\mu - \lambda)}. \tag{45}$$

Proof: The process Z from Definition 6.1 possesses the following global balance equations:

$$\begin{aligned}\pi(n,k)&(\lambda(1-\delta_{Nn}) + \mu(1-\delta_{0n})) \\ &= \pi(n-1,k)\lambda(1-\delta_{0n}) + \pi(n+1,k+1)\mu(1-\delta_{Nn}) \\ &\quad \times (1-\delta_{kM}) + \pi(n,0)\nu p_k(1-\delta_{Nn}), \\ &\quad 0 \le n \le N, 1 \le k \le M, \end{aligned} \tag{46}$$

$$\pi(n,0)\nu = \pi(n+1,1)\mu, \quad 0 \le n \le N-1. \tag{47}$$

We have to show that the distribution from (43) and (44) satisfies Eqs. (46) and (47). This is done by insertion using $q_1 = 1$, $q_M = p_M$, and $q_k = q_{k+1} + p_k$. The normalization

constant K is

$$K = \sum_{n=0}^{N}\sum_{k=1}^{M}\left(\frac{\lambda}{\mu}\right)^n q_k + \sum_{n=0}^{N-1}\left(\frac{\lambda}{\mu}\right)^{n+1}\frac{\mu}{\nu}$$
$$= \frac{1-\left(\frac{\lambda}{\mu}\right)^{N+1}}{1-\frac{\lambda}{\mu}}\left(\bar{p}+\frac{\mu}{\nu}\right)-\frac{\mu}{\nu}$$
$$= \frac{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)}{\mu^N(\mu-\lambda)},$$

where $\bar{p}$ denotes the mean size of a replenishment order. □

Remark 6.3. The stationary distribution is the conditional distribution of the open system with infinite waiting room from Definition 2.1 and Theorem 2.5 conditioned on the total population size excluding the state $(N, 0)$. Our pair of systems with finite or infinite waiting room shares this property with standard birth and death queues.

We define

$$K_X = \frac{\mu^N-\lambda^N}{\mu^{N-1}(\mu-\lambda)},$$
$$K_Y := \frac{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)}{\mu(\mu^N-\lambda^N)}$$

such that $K = K_X \cdot K_Y$. Note that in case of a finite waiting room K cannot be splitted up in factors which only depend on queue length and inventory size respectively.

The marginal steady state queue length probabilities are for $n = 0, \ldots, N-1$

$$\mathsf{P}(X=n) = \sum_{k=1}^{M}\pi(n,k)+\pi(n,0) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n\left(\bar{p}+\frac{\lambda}{\nu}\right),$$
$$\mathsf{P}(X=N) = \sum_{k=1}^{M}\pi(N,k) = K^{-1}\left(\frac{\lambda}{\mu}\right)^N\bar{p}.$$

These probabilities do not coincide with the steady state probabilities of the queue length process in the classical M/M/1/N-1-FCFS-system. The steady state probability that the on-hand inventory is $M \geq k \geq 1$ is given by

$$\mathsf{P}(Y=k) = \sum_{n=0}^{N}\pi(n,k) = q_k\frac{\mu^{N+1}-\lambda^{N+1}}{\mu^N(\mu-\lambda)}K^{-1}.$$

The probability that the server is depleted is given by

$$\mathsf{P}(Y=0) = \sum_{n=0}^{N-1}\pi(n,0) = \frac{\lambda}{\nu}\frac{\mu^N-\lambda^N}{\mu^{N-1}(\mu-\lambda)}K^{-1} = \frac{\lambda}{\nu}K_Y^{-1}.$$

Remark 6.4. If $\nu = \infty$ the inventory on hand leaves state 0 immediately and the state space is then given by $E_{Z_N} = \{(n,k) : n \in \{0,1,\ldots,N\}, 1 \leq k \leq M\}$. If in addition the order size is fixed to some $Q \in \mathbb{N}$ then the inventory positions are uniformly distributed as in [4].

Theorem 6.5 (Measures of system performance). *The measures of system performance as defined in Section* 2.1 *of the M/M/1/N-1-system with inventory according to Definition* 6.1 *are*

$$\bar{I} = \frac{\mu^{N+1}-\lambda^{N+1}}{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)}\sum_{k=1}^{M}kq_k$$
$$= \frac{\mu^{N+1}-\lambda^{N+1}}{K_Y\mu(\mu^N-\lambda^N)}\sum_{k=1}^{M}kq_k, \tag{48}$$

$$\lambda_R = \frac{(\mu^N-\lambda^N)\lambda\mu}{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)} = \frac{\lambda}{K_Y}, \tag{49}$$

$$\lambda_A = \frac{(\mu^N-\lambda^N)\lambda\mu\bar{p}}{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)} = \frac{\lambda\bar{p}}{K_Y}, \tag{50}$$

$$\overline{LS} = \lambda\frac{\lambda}{\nu}K_Y^{-1}+\lambda\left(\frac{\lambda}{\mu}\right)^N\bar{p}K^{-1}, \tag{51}$$

$$\overline{LS_c} = \frac{\lambda}{\nu}+\left(\frac{\lambda}{\mu}\right)^N\bar{p}K_X^{-1}, \tag{52}$$

$$s = 0 \quad \textit{for } Y_{net} = Y \textit{ and } Y_{net} = \max\{0, Y-X\}, \tag{53}$$

$$\alpha_1 = 0 \quad \textit{for } Y_{net} = Y \textit{ and } Y_{net} = \max\{0, Y-X\}, \tag{54}$$

$$\alpha_2 = \frac{(\mu^{N+1}-\lambda^{N+1})\bar{p}}{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)}$$
$$= \frac{\bar{p}}{K_Y}\frac{\mu^{N+1}-\lambda^{N+1}}{\mu(\mu^N-\lambda^N)} \textit{ for } Y_{net} = Y, \tag{55}$$

$$\beta = \frac{(\mu^N-\lambda^N)\bar{p}\mu}{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right)-\lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)} = \frac{\bar{p}}{K_Y}, \tag{56}$$

$$\bar{W}_0 = \frac{1}{\mu-\lambda}\left(1+\frac{\lambda}{\nu\bar{p}}\right)-\frac{N\lambda^N}{\mu(\mu^N-\lambda^N)}\left(1+\frac{\mu}{\nu\bar{p}}\right), \tag{57}$$

$$\bar{W} = \frac{\lambda(\mu^{N-1}-\lambda^{N-1})}{(\mu-\lambda)(\mu^N-\lambda^N)}\left(1+\frac{\lambda}{\nu\bar{p}}\right)$$
$$-(N-1)\frac{\lambda^N}{\mu(\mu^N-\lambda^N)}\left(1+\frac{\mu}{\nu\bar{p}}\right). \tag{58}$$

Proof: The proof of Theorem 6.5 is presented in Appendix B. □

Note that the mean time between two orders

$$\frac{1}{\lambda_R} = \frac{\mu^{N+1}\left(\bar{p}+\frac{\lambda}{\nu}\right) - \lambda^{N+1}\left(\bar{p}+\frac{\mu}{\nu}\right)}{(\mu^N-\lambda^N)\lambda\mu}$$
$$= \frac{\mu^{N+1}-\lambda^{N+1}}{(\mu^N-\lambda^N)\lambda}\frac{\bar{p}}{\mu}+\frac{1}{\nu}$$

is larger than in the M/M/1/∞-system as defined in Definition 2.1 if $\mu > \lambda$.
$\overline{LS}$ and $\overline{LS}_c$ are now composed of two parts: the first part represents the lost customers from an empty inventory and the second part represents the lost customers from a full waiting room. As before $\beta = \frac{\lambda_A}{\lambda}$ but $\beta \neq \alpha_2 = \mathsf{P}(Y > 0)$. The mean sojourn time $\bar{W}_{0,s}$ and the mean waiting time $\bar{W}_s$ of the classical M/M/1/N-1-system are given by (see [10] Section 2.4, p. 97)

$$\bar{W}_{0,s} = \frac{1}{\mu-\lambda} - \frac{N\lambda^N}{\mu(\mu^N-\lambda^N)} \quad \text{and}$$
$$\bar{W}_s = \frac{\lambda}{\mu(\mu-\lambda)} - \frac{N\lambda^N}{\mu(\mu^N-\lambda^N)}.$$

Hence,

$$\bar{W}_0 - \bar{W}_{0,s} = \frac{1}{\mu-\lambda}\frac{\lambda}{\nu\bar{p}} - \frac{N\lambda^N}{\mu(\mu^N-\lambda^N)}\frac{\mu}{\nu\bar{p}}$$
$$= \frac{\mu^N - N\lambda^{N-1}\mu + (N-1)\lambda^N}{(\mu-\lambda)(\mu^N-\lambda^N)}\frac{\lambda}{\nu\bar{p}} > 0.$$

A similar computation for the mean waiting times yields

$$\bar{W} - \bar{W}_s = \frac{\mu^{N-1} - (N-1)\lambda^{N-2}\mu + (N-2)\lambda^{N-1}}{(\mu-\lambda)(\mu^N-\lambda^N)}\frac{\lambda^2}{\nu\bar{p}} > 0.$$

Remark 6.6. For $\nu \to \infty$ the following performance measures are the same as in the classical M/M/1/N-1-system: $\lambda_R = \frac{\lambda}{\bar{p}}\mathsf{P}(X < N)$, λ_A, $\overline{LS} = \lambda\mathsf{P}(X = N)$, $\beta = \mathsf{P}(X < N)$, $\bar{W}_0$ and $\bar{W}$.

In case of finite waiting room all performance measures depend on the service rate μ and the parameters λ and ν. The influence of F_p is the same as in the case of an infinite waiting room.

6.2 M/M/1/N-1-system with (r, Q)-policy

We consider the finite version of the M/M/1/∞-system with (r, Q)-policy from Section 3.

Definition 6.7. An M/M/1/N-1-system with (r, Q)-policy is a single server under FCFS regime as described in Definition 3.1 but with the number of waiting places limited to $N-1$. The capacity of the inventory is $M = r + Q$.

Let $X(t)$ denote the number of customers present at the server at time $t \geq 0$, either waiting or in service and let $Y(t)$ denote the on-hand inventory at time $t \geq 0$. Then $Z = ((X(t), Y(t)), t \geq 0)$ is a continuous-time Markov process for the M/M/1/N-1-system with (r, Q)-policy. The state space of Z is $E_{Z_N} = \{(n,k) : n \in \{0,1,\ldots,N\}, 1 \leq k \leq Q+r\} \cup \{(n,0) : n \in \{0,1,\ldots,N-1\}\}$.

Theorem 6.8. *If we assume that a replenishment order is outstanding only if the on-hand inventory is smaller or equal than r and if there are less than N customers present in the system, the continuous-time Markov process Z has a stationary distribution of product form given by*

$$\pi(n,k) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n C(k), \text{ with } 0 \leq n \leq N, 1 \leq k \leq r+Q, \tag{59}$$

$$\pi(n,0) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n \frac{\lambda}{\nu}, \quad \text{with } 0 \leq n \leq N-1. \tag{60}$$

$C(k)$, $1 \leq k \leq Q+r$, are the same as in Theorem 3.2 and the normalization constant is given by

$$K = \frac{\mu^{N+1}-\lambda^{N+1}}{\mu^N(\mu-\lambda)}\left(Q\left(\frac{\lambda+\nu}{\lambda}\right)^r + \frac{\lambda}{\nu}\right) - \frac{\lambda}{\nu}\frac{\lambda^N}{\mu^N}. \tag{61}$$

Proof: The proof of Theorem 3.2 is modified by replacing ν by $\nu(1-\delta_{Nn})$. Furthermore, λ on the left side of (12) and μ on the right side of (12) and (13) are multiplied by $(1-\delta_{Nn})$. □

6.3 M/M/1/N-1-system with (r, S)-policy

We consider the finite version of the M/M/1/∞-system with (r, S)-policy from Section 4.

Definition 6.9. An M/M/1/N-1-system with (r, S)-policy is a single server under FCFS regime as described in Definition 4.1 but now the number of waiting places is limited to $N-1$. The capacity of the inventory is $M = S$.

Let $X(t)$ denote the number of customers present at the server at time $t \geq 0$, either waiting or in service and let $Y(t)$ denote the on-hand inventory at time $t \geq 0$. Then $Z = ((X(t), Y(t)), t \geq 0)$ is a continuous-time Markov process for the M/M/1/N-1-system with (r, S)-policy. The state space of Z is $E_{Z_N} = \{(n,k) : n \in \{0,1,\ldots,N\}, 1 \leq k \leq S\} \cup \{(n,0) : n \in \{0,1,\ldots,N-1\}\}$.

Theorem 6.10. *If we assume that a replenishment order is outstanding only if the inventory on hand is smaller or equal than r and if there are less than N customers present in the system, the continuous-time Markov process Z has a stationary distribution of product form given by*

$$\pi(n,k) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n C(k), \quad \textit{with } 0 \le n \le N, 1 \le k \le S, \tag{62}$$

$$\pi(n,0) = K^{-1}\left(\frac{\lambda}{\mu}\right)^n \frac{\lambda}{\nu}, \quad \textit{with } 0 \le n \le N-1. \tag{63}$$

$C(k), 1 \le k \le S$, are the same as in Theorem 4.2 and the normalization constant is given by

$$K = \frac{\mu^{N+1} - \lambda^{N+1}}{\mu^N(\mu-\lambda)}\left(S - r + \frac{\lambda}{\nu}\right)\left(\frac{\lambda+\nu}{\lambda}\right)^r - \frac{\lambda}{\nu}\frac{\lambda^N}{\mu^N}. \tag{64}$$

Proof: The proof of Theorem 4.2 is modified by replacing ν by $\nu(1-\delta_{Nn})$. Furthermore, λ on the left side of (29) and μ on the right side of (29) and (30) are multiplied by $(1-\delta_{Nn})$. □

7 Cost analysis

From the explicit results obtained in the preceding sections it is an easy task to explicitly write down the usual cost functions and reward functions respectively. Recall the cost-reward structure from Definition 2.1. The mean costs that occur in steady state per time unit are:

- $\lambda_R \cdot K$, the fixed costs associated to replenishment orders that occur with reorder rate λ_R,
- $\bar{I} \cdot h$, the holding costs for inventory of mean size $\bar{I}$,
- $\overline{LS} \cdot \ell$, the shortage costs for the mean number of lost sales $\overline{LS}$,
- $\bar{L} \cdot \omega$, the waiting costs for the mean number $\bar{L}$ of waiting customers,
- $\bar{V} \cdot \sigma$, the costs for the mean number $\bar{V}$ of customers in service.

The revenue obtained by the system's service is per time unit:

- $\lambda_A \cdot R$, the amount of money obtained from the served customers per time unit, which is proportional to the throughput.

A first conclusion for minimization of costs is from Theorem 2.14 and Corollary 2.17: In optimization we can restrict ourselves to deterministic order sizes.

Let us therefore consider the cost structure of the M/M/1/∞ under (r, Q) policy from Section 3 which is

$$F = \lambda_R \cdot K + \bar{I} \cdot h + \overline{LS} \cdot \ell + \bar{L} \cdot \omega + \bar{V} \cdot \sigma. \tag{65}$$

Collecting the results from Theorem 3.4 in Section 3 and recalling that the steady state queue length distribution in this system is the same as in the isolated M/M/1/∞ we obtain:

$$\begin{aligned} F = {} & \frac{\lambda}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r} \cdot K + \frac{Q}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}\left(\frac{Q+1}{2} + s\right) \cdot h \\ & + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{\lambda}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r} \cdot \ell + \frac{\lambda^2}{\mu(\mu-\lambda)} \cdot v_w \\ & + \frac{\lambda}{\mu} \cdot v_s. \end{aligned}$$

This function is to be minimized under side conditions which are prescribed by e.g. bounds for the available service capacity or a maximal acceptable loss rate.

Nevertheless having these explicit functions at hand it is not immediate to prove structural properties of these functions, e.g. convexity or concavity. Even for the much simpler case of having no explicit service times and service constraints in the system's behaviour the application of calculus methods to establish these properties for the cost functions in an analytical sense turns out to be impractical, see [14][p. 268]. But a numerical search procedure similar to the one described there can get started by inserting the formula into a cost function type like (65).

E.g., if in a system with (r, Q)-policy (see Section 3), where no queueing and service occurs (i.e. $\bar{L} = \bar{W} = 0$), we arrive at the optimization problem

$$F = F(r,Q) = \min!, \quad s.t. \quad 0 \le r < Q,$$

as discussed in [14][Problem 1, p. 266].

Problem 2 of [14][p. 266 there] imposes side conditions via stockout risk, i.e. safety stock and service levels. The necessary quantities are computed successively as well.

A similar approach is opened by the formulas obtained in the previous sections for the more complex integrated queueing-inventory systems.

Although it seems hard to perform an analytical sensitivity analysis for the cost functions we have direct access to numerical procedures from our results and do not need long simulations.

8 Conclusions

We have studied M/M/1 queueing systems with attached inventory, exponentially distributed lead times and lost sales for infinite and finite waiting rooms. The discussed inventory management policies are the (r, Q)- and (r, S)-policies and general randomized order policies with reorder point 0. For each of these systems we have determined the stationary distribution. Surprisingly enough all these distributions are of product form. In case of infinite waiting room the limiting distributions of the queue length processes coincide with that of the classical M/M/1/∞-system. To the best of our knowledge these are the first closed form solutions to service facilities with attached inventories and stochastic lead times. Note that state dependent service rates can be incorporated into the different models without any difficulty.

Various definitions for performance measures found in the literature have been quoted and compared in the underlying models.

Directions for future research are investigations of the long-run average cost functions with/without service constraints and to allow for more general lead time and service time distributions.

Appendix A: Customer and event stationary measures and intensities

This section contains results on intensities for the customer arrival/departure processes and the replenishment order arrival process for the systems from Sections 2, 3, 4 and 6. Besides these performance measures the Palm probability associated with the point process of replenishments is computed. These probabilities are needed to calculate the safety stock and α_1-service level as defined in Definitions 2.10 and 2.11.

In addition to these results used in earlier sections we also point out some results related to the arrival theorem for systems with limited and unlimited population size.

We assume that under P Z is a regular, stationary Markov Process with Q-Matrix $Q = \{q(i, j)\}$ and invariant measure π.

Theorem A.1. *In the M/M/1/∞-system with inventory management, reorder point* 0 *and arbitrary size of replenishment orders according to Definition* 2.1 *the intensity of departures λ_D and the intensity of arrivals of customers who are admitted to the system λ_A are*

$$\lambda_D = \lambda_A = \frac{\bar{p}\lambda\nu}{\bar{p}\nu + \lambda}. \tag{66}$$

The intensity of arrivals of replenishment orders λ_R is

$$\lambda_R = \frac{\lambda\nu}{\bar{p}\nu + \lambda}. \tag{67}$$

The Palm probability associated with the point process of replenishments is

$$\mathsf{P}^0_{N_R}(Z_{0-} = (n, 0)) = \frac{\bar{p}\nu + \lambda}{\lambda}\,\pi(n, 0) \quad \text{for } n \in \mathbb{N}_0. \tag{68}$$

Proof: Let H be a subset of $E_Z \times E_Z - \operatorname{diag}(E_Z \times E_Z)$ and let N_H be the point process counting the H-transitions of Z, i.e.

$$N_H(C) = \int_C 1_H(Z_{s-}, Z_s)\, N(ds), \quad C \in \mathbb{B}.$$

Let us consider special sets that describe events of the queueing system under investigation

$$\begin{aligned}
D = {} & \{((n, k), (n-1, k-1)) : n \geq 1, M \geq k \geq 1\} \\
& \text{departure of customers,} \\
A = {} & \{((n, k), (n+1, k)) : n \geq 0, M \geq k \geq 1\} \\
& \text{arrival of customers, who are admitted to the system,} \\
R_k = {} & \{((n, 0), (n, k)) : n \geq 0\} \\
& \text{arrival of replenishment orders of size } M \geq k \geq 1, \\
R = {} & \{((n, 0), (n, k)) : n \geq 0, M \geq k \geq 1\} \\
& \text{arrival of replenishment orders.}
\end{aligned}$$

From formula (1.4.15) on page 38 of [1]

$$\lambda_H = E[N_H((0, 1])] = \sum_{((n,k),(m,l))\in H} \pi(n, k) q((n, k), (m, l)) \tag{69}$$

and therefore

$$\begin{aligned}
\lambda_D &= \sum_{((n,k),(n-1,k-1))\in D} \pi(n, k) q((n, k), (n-1, k-1)) \\
&= \mu \sum_{n\geq 1} \sum_{k=1}^{M} \pi(n, k) = \mu \frac{\lambda}{\mu} \frac{\bar{p}}{\bar{p} + \frac{\lambda}{\nu}} = \frac{\bar{p}\lambda\nu}{\bar{p}\nu + \lambda} = \frac{\lambda\bar{p}}{K_Y}, \\
\lambda_A &= \sum_{((n,k),(n+1,k))\in A} \pi(n, k) q((n, k), (n+1, k)) \\
&= \lambda \mathsf{P}(Y > 0) = \lambda_D, \\
\lambda_{R_k} &= \sum_{((n,0),(n,k))\in R_k} \pi(n, 0) q((n, 0), (n, k)) = \nu p_k \mathsf{P}(Y = 0) \\
&= \frac{\lambda p_k}{K_Y},
\end{aligned}$$

$$\lambda_R = \sum_{k=1}^{M} \lambda_{R_k} = \frac{\lambda}{K_Y}.$$

Since $0 < \lambda_H < \infty$ we can define the Palm probability $P^0_{N_H}$ associated with N_H for all $H \in \{D, A, R_k, R\}$ as in formulas (1.4.18) and (1.4.19).

$$\begin{aligned}
&\mathsf{P}^0_{N_D}(Z_0 = (n,k)) \\
&= \frac{\sum_{(m,l)/((m,l),(n,k))\in D} \pi(m,l) q((m,l),(n,k))}{\lambda_D} \\
&= \frac{\mu(\bar{p}\nu + \lambda)}{\bar{p}\lambda\nu} \pi(n+1, k+1) \\
&= \frac{K_Y}{\bar{p}} \pi(n, k+1), n, 0 \le k \le M \\
&\mathsf{P}^0_{N_D}(Z_{0-} = (n,k)) \\
&= \begin{cases} \dfrac{\mu(\bar{p}\nu + \lambda)}{\bar{p}\lambda\nu} \pi(n,k) & \text{for } 0 < n,\ 0 < k \le M, \\ 0 & \text{otherwise,} \end{cases} \\
&= \begin{cases} \dfrac{K_Y}{\bar{p}} \pi(n-1,k) & \text{for } 0 < n,\ 0 < k \le M, \\ 0 & \text{otherwise,} \end{cases} \\
&\mathsf{P}^0_{N_A}(Z_0 = (n,k)) \\
&= \frac{\sum_{(m,l)/((m,l),(n,k))\in A} \pi(m,l) q((m,l),(n,k))}{\lambda_A} \\
&= \begin{cases} \dfrac{K_Y}{\bar{p}} \pi(n-1,k) & \text{for } 0 < n, 0 < k \le M, \\ 0 & \text{otherwise,} \end{cases} \\
&\mathsf{P}^0_{N_A}(Z_{0-} = (n,k)) \\
&= \pi(n,k) \frac{\sum_{(m,l)/((n,k),(m,l))\in A} q((n,k),(m,l))}{\lambda_A} \\
&= \begin{cases} \dfrac{K_Y}{\bar{p}} \pi(n,k) & \text{for } n \in \mathbb{N}_0, 0 < k \le M, \\ 0 & \text{otherwise,} \end{cases} \\
&\mathsf{P}^0_{N_{R_k}}(Z_0 = (n,k)) \\
&= \frac{\sum_{(m,l)/((m,l),(n,k))\in R_k} \pi(m,l) q((m,l),(n,k))}{\lambda_{R_k}} \\
&= \frac{K_Y \nu}{\lambda} \pi(n,0) \quad \text{for } n \in \mathbb{N}_0, \\
&\mathsf{P}^0_{N_R}(Z_{0-} = (n,0)) \\
&= \pi(n,0) \frac{\sum_{(m,l)/((n,0),(m,l))\in R} q((n,0),(m,l))}{\lambda_R} \\
&= \frac{K_Y \nu}{\lambda} \pi(n,0) \quad \text{for } n \in \mathbb{N}_0. \qquad \square
\end{aligned}$$

For a large class of networks having product-form equilibrium distribution it is known (see [16]) that if a customer belongs to an open subchain the state distribution at arrival points, departure points, and random points are identical. In this system the arrival-instant/departure-instant and the steady-state distributions are not the same. Just before an arrival a customer observes

$$\mathsf{P}^0_{N_A}(Z_{0-} = (n,k)) = \frac{K_Y}{\bar{p}} \pi(n,k), \quad n \in \mathbb{N}_0, 0 < k \le M$$

whereas just after a departure he leaves

$$\mathsf{P}^0_{N_D}(Z_0 = (n,k)) = \frac{K_Y}{\bar{p}} \pi(n,k+1) \quad n \in \mathbb{N}_0, 0 < k \le M.$$

These probabilities are not equal since $\pi(n,k) \neq \pi(n,k+1)$. Note that an arriving customer never sees the state $(n,0)$ without counting himself, whereas a departing customer can leave the state $(n,0)$ if he gets the last item from inventory. The following Palm probabilities are equal.

$$\mathsf{P}^0_{N_A}(Z_0 = (n,k)) = \mathsf{P}^0_{N_D}(Z_{0-} = (n,k)),\ 0 < n, 0 < k \le M.$$

After an arrival a customer observes the state (n,k) with the same probability as a customer just before his departure.

If we look at aggregate states $n \in \mathbb{N}_0$, defined as the queue length, we obtain

$$\begin{aligned}
\mathsf{P}^0_{N_A}(X_{0-} = n) &= \frac{K_Y}{\bar{p}} \sum_{k=1}^{M} \pi(n,k) = K_X \left(\frac{\lambda}{\mu}\right)^n = \mathsf{P}(X = n) \\
&= \frac{K_Y}{\bar{p}} \sum_{k=0}^{M} \pi(n,k+1) = \mathsf{P}^0_{N_D}(X_0 = n).
\end{aligned}$$

This shows that the distributions of the queue length $n \in \mathbb{N}_0$ just before an arrival and just after a departure instant and in the time stationary queue are the same. This is the ESTA (events see time averages) property [7], note that arrivals to the server do not form a Poisson process. For $0 < n$ we compute

$$\begin{aligned}
\mathsf{P}^0_{N_A}(X_0 = n) &= \frac{K_Y}{\bar{p}} \sum_{k=1}^{M} \pi(n-1,k) = \mathsf{P}(X = n-1) \\
&= \mathsf{P}^0_{N_D}(X_{0-} = n).
\end{aligned}$$

Hence, the distributions of n customers in the system just after an arrival and just before a departure instant are the same and they are equal to the time stationary probability with one less customer in the queue. As for every stochastic

process on $\mathbb{N}_0$ with step size one, the steady state probabilities in arrival and departure instants are equal if they exist.

Theorem A.2. *In the M/M/1/∞-system with* (r, Q)*-policy according to Definition* 3.1 *the intensity of departures* λ_D *and the intensity of arrivals of customers who are admitted to the system* λ_A *are*

$$\lambda_D = \lambda_A = \frac{Q\lambda}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}. \tag{70}$$

The intensity of arrivals of replenishment orders λ_R *is*

$$\lambda_R = \frac{\lambda}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}. \tag{71}$$

The Palm probability associated with the point process of replenishments is

$$\mathsf{P}^0_{N_R}(Z_{0-} = (n,h)) = \frac{\nu}{\lambda_R}\pi(n,h) \quad \text{for } 0 \le h \le r, n \in \mathbb{N}_0. \tag{72}$$

Proof: Define special sets D for departure of customers, A for arrival of customers, who are admitted to the system, and R for arrivals of replenishment orders as before. Using Eq. (69) it is straight forward to prove that in the context of the M/M/1/∞-system with (r, Q)-policy the intensities λ_D, λ_A and λ_R are equal to the expressions in (70) and (71). For $0 \le h \le r$ and $n \in \mathbb{N}_0$

$$\begin{aligned}\mathsf{P}^0_{N_R}&(Z_{0-} = (n,h)) \\ &= \pi(n,h)\frac{\sum_{(m,l)/((n,h),(m,l))\in R} q((n,h),(m,l))}{\lambda_R} \\ &= \frac{\nu}{\lambda_R}\pi(n,h).\end{aligned}$$

□

Theorem A.3. *In the M/M/1/∞-system with* (r, S)*-policy according to Definition* 4.1 *the intensity of departures* λ_D *and the intensity of arrivals of customers who are admitted to the system* λ_A *are*

$$\lambda_D = \lambda_A = \lambda - \frac{\lambda^2}{(S-r)\nu + \lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r. \tag{73}$$

The intensity of arrivals of replenishment orders λ_R *is*

$$\lambda_R = \frac{\lambda}{S - r + \frac{\lambda}{\nu}}. \tag{74}$$

The Palm probability associated with the point process of replenishments is

$$\mathsf{P}^0_{N_R}(Z_{0-} = (n,h)) = \frac{\nu}{\lambda_R}\pi(n,h) \quad \text{for } 0 \le h \le r, n \in \mathbb{N}_0. \tag{75}$$

Proof: The proof can be done analogously to the proof of Theorem A.2. □

The observations made above concerning the probabilities in arrival and departure instants carry over to the M/M/1/∞-system with (r, Q)-policy or (r, S)-policy.

Theorem A.4. *In the M/M/1/N-1-system with inventory management, reorder point* 0 *and arbitrary size of replenishment orders according to Definition* 6.1 *the intensity of departures* λ_D *and the intensity of arrivals of customers who are admitted to the system* λ_A *are*

$$\lambda_D = \lambda_A = \frac{\bar{p}\lambda}{K_Y}. \tag{76}$$

The intensity of arrivals of replenishment orders λ_R *is*

$$\lambda_R = \frac{\lambda}{K_Y}. \tag{77}$$

The Palm probability associated with the point process of replenishments is

$$\mathsf{P}^0_{N_R}(Z_{0-} = (n,0)) = \frac{K_Y\nu}{\lambda}\pi(n,0) \quad \text{for } 0 \le n \le N-1. \tag{78}$$

Proof: As before we consider the following special sets that describe events of the queueing system

$$\begin{aligned} D =& \{((n,k),(n-1,k-1)) : 1 \le n \le N, 1 \le k \le M\} \\ &\text{departure of customers,} \\ A =& \{\big((n,k),(n+1,k)\big) : 0 \le n \le N-1, 1 \le k \le M\} \\ &\text{arrival of customers,} \\ R_k =& \{\big((n,0),(n,k)\big) : 0 \le n \le N-1\} \\ &\text{arrival of a replenishment order of size } M \ge k \ge 1, \\ R =& \{\big((n,0),(n,k)\big) : 0 \le n \le N-1, 1 \le k \le M\} \\ &\text{arrival of a replenishment order.} \end{aligned}$$

Computations similar to those in the proof of Theorem A.1 show that λ_D, λ_A and λ_R have the form given in (76) and (77). The calculation of $\mathsf{P}^0_{N_R}(Z_{0-} = (n,0))$ is straightforward.

□

We now study the finite queueing system at the moments when customers arrive at the server or leave the server. In this system the arrival-instant/departure-instant and the steady-state distributions are not the same. We have

$$\mathsf{P}^0_{N_A}(Z_{0-} = (n,k)) \neq \mathsf{P}^0_{N_D}(Z_0 = (n,k)), \qquad 0 < k \leq M, 0 \leq n \leq N-1$$

since $\pi(n,k) \neq \pi(n,k+1)$ and

$$\mathsf{P}^0_{N_A}(Z_0 = (n,k)) = \mathsf{P}^0_{N_D}(Z_{0-} = (n,k)), \qquad 0 < n \leq N, 0 < k \leq M.$$

If we look at aggregate states n for $0 \leq n \leq N-1$, we obtain

$$\mathsf{P}^0_{N_A}(X_{0-} = n) = \mathsf{P}^0_{N_D}(X_0 = n) = \frac{K_Y}{\bar{p}} \sum_{k=1}^{M} \pi(n,k) = K_{ql}^{-1} \left(\frac{\lambda}{\mu}\right)^n \neq \mathsf{P}(X = n).$$

This shows that the distributions of the queue length n just before an arrival and just after a departure instant are the same. The time stationary distribution of the queue length is different. For $0 < n$ we compute

$$\mathsf{P}^0_{N_A}(X_0 = n) = \mathsf{P}^0_{N_D}(X_{0-} = n) = \frac{K_Y}{\bar{p}} \sum_{k=1}^{M} \pi(n-1,k) \neq \mathsf{P}(X = n-1). \tag{79}$$

Hence, the distributions of n customers in the system just after an arrival or just before a departure instant are not equal to the time stationary probability with one less customer in the queue. This is in contrast to the arrival theorem [16] for networks with limited population size.

Appendix B: In the main text omitted proofs

Proof of Theorem 3.4. The average on-hand inventory position is

$$\bar{I} = \sum_{k=1}^{Q+r} k \sum_{n \geq 0} \pi(n,k) = \frac{Q}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r} \left\{ \frac{Q+1}{2} + r - \frac{\lambda}{\nu} + \frac{\lambda}{\nu} \left(\frac{\lambda}{\lambda+\nu}\right)^r \right\},$$

and s from (20) is computed below. This yields (15).

λ_R and λ_A follow from Theorem A.2.

$$\overline{LS} = \lambda \sum_{n \geq 0} \pi(n,0) = \lambda \mathsf{P}(Y = 0) = \lambda \frac{\frac{\lambda}{\nu}}{Q\left(\frac{\lambda+\nu}{\lambda}\right)^r + \frac{\lambda}{\nu}},$$

$$\overline{LS}_c = \frac{\overline{LS}}{\lambda_R} = \frac{\lambda}{\nu} \left(\frac{\lambda}{\lambda+\nu}\right)^r.$$

Involving Definition 2.10 and (72) from Theorem A.2 the safety stock or the expected net inventory position just before a replenishment order arrives is

$$\begin{aligned} s &= \sum_{k \geq 1} k\, \mathsf{P}^0_{N_R}(Y_{net,-} = k) = \sum_{k=1}^{r} k \sum_{n \geq 0} \mathsf{P}^0_{N_R}(Z_{0-} = (n,k)) \\ &= \frac{\nu}{\lambda_R} \sum_{k=1}^{r} k \sum_{n \geq 0} \pi(n,k) = \frac{\nu}{\lambda_R} K_Y^{-1} \sum_{k=1}^{r} kC(k) \\ &= \frac{\nu}{\lambda} \left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{1 - (r+1)\left(\frac{\lambda+\nu}{\lambda}\right)^r + r\left(\frac{\lambda+\nu}{\lambda}\right)^{r+1}}{\left(1 - \frac{\lambda+\nu}{\lambda}\right)^2} \\ &= \frac{\nu}{(\lambda+\nu)^r} \frac{\lambda^{r+1} - (r+1)\lambda(\lambda+\nu)^r + r(\lambda+\nu)^{r+1}}{\nu^2} \\ &= r - \frac{\lambda}{\nu}\left(1 - \left(\frac{\lambda}{\lambda+\nu}\right)^r\right). \end{aligned}$$

Using Definitions 2.10 and 2.11 we have

$$\begin{aligned} \alpha_1 &= \mathsf{P}^0_{N_R}(Y_{0-} > 0) = 1 - \sum_{n \geq 0} \mathsf{P}^0_{N_R}(Z_{0-} = (n,0)) \\ &= 1 - \frac{\nu}{\lambda_R} K_Y^{-1} \frac{\lambda}{\nu} = 1 - \left(\frac{\lambda}{\lambda+\nu}\right)^r. \end{aligned}$$

From Definition 2.11 $\alpha_2 = \mathsf{P}(Y > 0)$ and $\beta = (\lambda - \overline{LS})/\lambda$ hence, the results for α_2 and β are obvious.

The mean number of customers in the system $\bar{L}_0$ and the mean number of waiting customers $\bar{L}$ are the same as in the usual M/M/1/∞-system. Using Little's formula the mean sojourn time $\bar{W}_0$ and the mean waiting time $\bar{W}$ of the customers are

$$\begin{aligned} \bar{W}_0 &= \frac{\bar{L}_0}{\lambda_A} = \frac{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}{\lambda Q} \frac{\lambda}{\mu - \lambda} \\ &= \frac{1}{\mu - \lambda} + \frac{\lambda}{Q\nu} \left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{1}{\mu - \lambda}, \\ \bar{W} &= \frac{\bar{L}}{\lambda_A} = \frac{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}{\lambda Q} \frac{\lambda^2}{\mu(\mu - \lambda)} \\ &= \frac{\lambda}{\mu(\mu - \lambda)} + \frac{\lambda}{Q\nu} \left(\frac{\lambda}{\lambda+\nu}\right)^r \frac{\lambda}{\mu(\mu - \lambda)}. \end{aligned}$$

□

Proof of Theorem 4.3. The average on-hand inventory position is

$$\begin{aligned}\bar{I} &= \sum_{k=1}^{S} k \sum_{n\geq 0} \pi(n,k) \\ &= \frac{1}{S-r+\frac{\lambda}{\nu}} \left\{ \frac{\lambda}{\nu}\left(r - \frac{\lambda}{\nu} + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r\right) \right. \\ &\quad \left. + \frac{(S+1)S-(r+1)r}{2} \right\}.\end{aligned}$$

Using s from (37) the final form of $\bar{I}$ given in (32) follows. λ_R and λ_A are a result of Theorem A.3. The average number of lost sales incurred per unit of time and per cycle are

$$\begin{aligned}\overline{LS} &= \lambda \sum_{n\geq 0} \pi(n,0) = \lambda \mathsf{P}(Y=0) \\ &= \frac{\lambda^2}{(S-r)\nu+\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r, \\ \overline{LS}_c &= \frac{\overline{LS}}{\lambda_R} = \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r.\end{aligned}$$

Since $\lambda_R = \lambda(\frac{\lambda+\nu}{\lambda})^r K_Y^{-1}$ as in the M/M/1/∞-system with (r, Q)-policy (see Theorems A.2 and A.3) and the constants $C(k)$ are the same in both systems for $1 \leq k \leq r$ (see Theorems 3.2 and 4.2) the calculation of s in the proof of Theorem 3.4 carries over to the M/M/1/∞-system with (r, S)-policy. The same reasoning holds for the calculation of α_1. The β-service level is

$$\beta = \frac{\lambda - \overline{LS}}{\lambda} = 1 - \frac{\lambda}{(S-r)\nu+\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r.$$

As before $\beta = \lambda_A/\lambda = \mathsf{P}(Y > 0) = \alpha_2$.
The mean number of customers in the system $\bar{L}_0$ and the mean number of waiting customers $\bar{L}$ are the same as in the usual M/M/1/∞-system. Using Little's formula we obtain the formulas for mean sojourn time $\bar{W}_0$ and mean waiting time $\bar{W}$ in the form given in Theorem 4.3. □

Proof of Theorem 4.5. Let $S = Q + s$, where s is the safety stock in the M/M/1/∞-system with (r, Q)-policy from (20). Because of (37) $s = r - \frac{\lambda}{\nu}(1 - (\frac{\lambda}{\lambda+\nu})^r)$ is also the safety stock of the order-up-to level (r, S)-system. In the (r, S)-system λ_R is

$$\lambda_R = \frac{\lambda}{S - s + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}.$$

If $S - s = Q$ this is equal to λ_R in the (r, Q)-system. Using (37) and $S - s = Q$ once again we obtain from (34)

$$\lambda_A = \lambda - \frac{\lambda^2}{Q\nu + \lambda\left(\frac{\lambda}{\lambda+\nu}\right)^r}\left(\frac{\lambda}{\lambda+\nu}\right)^r = \frac{Q\lambda}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}.$$

This is (17). Since $\overline{LS} = \lambda - \lambda_A$ and $\overline{LS}_c = \overline{LS}\lambda_R^{-1}$ the expected lost sales per unit of time and per cycle are the same in the (r, Q)- and (r, S)-system. α_1 is the same in both systems if the reorder point r is identical. From the definition $\beta = \frac{\lambda - \overline{LS}}{\lambda}$ coincides in the two systems if $\overline{LS}$ is the same. The equality of $\bar{W}_0$ and $\bar{W}$ can be seen from the formulas (23) and (40) and (24) and (41) respectively if $S - s = Q$ holds.

We will now show that if $S = Q + s$ the stationary mean inventory position $\bar{I}_l$ in the M/M/1/∞-system with order-level (r, S)-policy is not larger than the stationary mean inventory position $\bar{I}_q$ in the M/M/1/∞-system with order-quantity (r, Q)-policy. In the (r, Q)-system we have

$$\bar{I}_q = \frac{Q}{Q + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}\left(\frac{Q+1}{2} + s\right).$$

Whereas in the (r, S)-system we have

$$\bar{I}_l = \frac{\frac{\lambda}{\nu}s + \frac{(S+1)S-(r+1)r}{2}}{S - s + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r}.$$

Replacing Q by $S - s$ in the first expression we have

$$\begin{aligned}\bar{I}_l - \bar{I}_q &= \frac{2\frac{\lambda}{\nu}s + (S+1)S - (r+1)r - (S-s)(S+s+1)}{2\left(S - s + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r\right)} \\ &= \frac{2\frac{\lambda}{\nu}s - (r+1)r + (s+1)s}{2\left(S - s + \frac{\lambda}{\nu}\left(\frac{\lambda}{\lambda+\nu}\right)^r\right)}.\end{aligned}$$

Replacing the safety stock s by $r - \frac{\lambda}{\nu}(1 - (\frac{\lambda}{\lambda+\nu})^r)$ (from (20) or (37)) and introducing $a := \frac{\lambda}{\lambda+\nu}$ and $b := \lambda/\nu$ yields with some algebra

$$\begin{aligned}\bar{I}_l - \bar{I}_q &= \frac{2b(r - b(1-a^r)) - (r+1)r + (r - b(1-a^r) + 1)(r - b(1-a^r))}{2(S - r + b)} \\ &= -\frac{b}{2(S-r+b)} \sum_{k=1}^{r} a^{r-k}(1-a^k)^2 < 0.\end{aligned}$$

Hence, $I_l < I_q$ and the proof is complete. □

Proof of Theorem 6.5. The average on-hand inventory is

$$\begin{aligned}\bar{I} &= \sum_{k=1}^{M} k \sum_{n=0}^{N} \pi(n,k) \\ &= \frac{\mu^{N+1} - \lambda^{N+1}}{\mu^{N+1}\left(\bar{p} + \frac{\lambda}{\nu}\right) - \lambda^{N+1}\left(\bar{p} + \frac{\mu}{\nu}\right)} \sum_{k=1}^{M} k q(k).\end{aligned}$$

λ_R and λ_A are given in Theorem A.4. The average number of lost sales incurred per unit of time and per cycle are given by

$$\begin{aligned}\overline{LS} &= \lambda\left(\sum_{n=0}^{N-1}\pi(n,0) + \sum_{k=1}^{M}\pi(N,k)\right) \\ &= \lambda\frac{\lambda}{\nu}K_Y^{-1} + \lambda\left(\frac{\lambda}{\mu}\right)^N \bar{p}K^{-1}, \\ \overline{LS}_c &= \frac{\overline{LS}}{\lambda_R} = \frac{\lambda}{\nu} + \left(\frac{\lambda}{\mu}\right)^N \bar{p}K_X^{-1}.\end{aligned}$$

Just as in the M/M/1/∞-system with inventory, reorder point 0 and arbitrary size of replenishment orders treated in Section 2, the safety stock is 0 for both definitions of net inventory Y_{net}.
For the service levels we compute for the first definition of net inventory 2.10

$$\alpha_1 = \mathsf{P}^0_{N_R}(Y_{net,0-} > 0) = \mathsf{P}^0_{N_R}(Y_{0-} > 0) = 0$$

and the second definition of net inventory from (42)

$$\begin{aligned}\alpha_1 &= \mathsf{P}^0_{N_R}(Y_{net,0-} > 0) = \mathsf{P}^0_{N_R}(\max\{0, Y_{0-} - X_{0-}\} > 0) = 0. \\ \alpha_2 &= \mathsf{P}(Y > 0) = 1 - \frac{\lambda}{\nu K_Y} = \frac{(\mu^{N+1} - \lambda^{N+1})\bar{p}}{\mu(\mu^N - \lambda^N)K_Y}. \\ \beta &= \frac{\lambda - \overline{LS}}{\lambda} = \frac{\bar{p}}{\bar{p} + \overline{LS}_c} = \frac{\bar{p}}{\bar{p}\frac{\mu^{N+1}-\lambda^{N+1}}{\mu(\mu^N-\lambda^N)} + \frac{\lambda}{\nu}} = \frac{\bar{p}}{K_Y}.\end{aligned}$$

The mean number of customers in the system $\bar{L}_0$ and the mean number of waiting customers $\bar{L}$ are not the same as in the usual M/M/1/N-1-system. They are

$$\begin{aligned}\bar{L}_0 &= \sum_{n=1}^{N}\sum_{k=1}^{M} n\pi(n,k) + \sum_{n=1}^{N-1} n\pi(n,0) \\ &= \frac{\lambda(\mu^N - \lambda^N)}{\mu^{N-1}(\mu-\lambda)^2 K}\left(\bar{p} + \frac{\lambda}{\nu}\right) - \frac{N\lambda^{N+1}}{\mu^N(\mu-\lambda)K}\left(\bar{p} + \frac{\mu}{\nu}\right),\end{aligned} \tag{80}$$

$$\begin{aligned}\bar{L} &= \bar{L}_0 - \mathsf{P}(X > 0) = \bar{L}_0 - \left(1 - K^{-1}\left(\bar{p} + \frac{\lambda}{\nu}\right)\right) \\ &= \frac{\lambda^2(\mu^{N-1} - \lambda^{N-1})}{\mu^{N-1}(\mu-\lambda)^2 K}\left(\bar{p} + \frac{\lambda}{\nu}\right) \\ &\quad - \frac{(N-1)\lambda^{N+1}}{\mu^N(\mu-\lambda)K}\left(\bar{p} + \frac{\mu}{\nu}\right).\end{aligned} \tag{81}$$

Using Little's formula the mean sojourn time $\bar{W}_0$ and the mean waiting time $\bar{W}$ of the customers are

$$\begin{aligned}\bar{W}_0 &= \frac{\bar{L}_0}{\lambda_A} \\ &= \frac{1}{\mu-\lambda}\left(1 + \frac{\lambda}{\nu\bar{p}}\right) - \frac{N\lambda^N}{\mu(\mu^N - \lambda^N)}\left(1 + \frac{\mu}{\nu\bar{p}}\right), \\ \bar{W} &= \frac{\bar{L}}{\lambda_A} = \frac{\lambda(\mu^{N-1} - \lambda^{N-1})}{(\mu-\lambda)(\mu^N - \lambda^N)\bar{p}}\left(\bar{p} + \frac{\lambda}{\nu}\right) \\ &\quad - \frac{(N-1)\lambda^N}{\mu(\mu^N - \lambda^N)\bar{p}}\left(\bar{p} + \frac{\mu}{\nu}\right).\end{aligned}$$

□

Proof of Theorem 5.1

$$\begin{aligned}s &= \sum_{k=1}^{M} k\mathsf{P}^0_{N_R}(Y_{net-} = k) = 0, \\ \alpha_1 &= \mathsf{P}^0_{N_R}(\max\{0, Y_{0-} - X_{0-}\} > 0) = 0.\end{aligned}$$

□

Proof of Theorem 5.2. We carry out the proof for the M/M/1/∞-system with (r, Q)- and (r, S)-policy at the same time. With the second definition of the net inventory position (42) we obtain

$$\begin{aligned}s &= \sum_{k=1}^{M} k\mathsf{P}^0_{N_R}(Y_{net,0-} = k) \\ &= \sum_{k=1}^{M} k\mathsf{P}^0_{N_R}(\max\{0, Y_{0-} - X_{0-}\} = k) \\ &= \sum_{h=1}^{\infty}\sum_{k=1}^{h} k\mathsf{P}^0_{N_R}(X_{0-} = h-k, Y_{0-} = h) \\ &= \frac{\nu}{\lambda_R}K^{-1}\sum_{h=1}^{r}\sum_{k=1}^{h} k\left(\frac{\lambda}{\mu}\right)^{h-k} C(h).\end{aligned}$$

For the (r, Q)- and the (r, S)-system $\frac{\nu}{\lambda_R}K^{-1} = \frac{\nu}{\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r K_X^{-1}$ with $K_X = \frac{\mu}{\mu-\lambda}$. Furthermore from Theorems 3.2 and 4.2 the

constants $C(h)$, $1 \le h \le r$ are the same. This yields

$$s = \frac{\nu}{\lambda}\left(\frac{\lambda}{\lambda+\nu}\right)^r K_X^{-1} \sum_{h=1}^{r} C(h)\left(\frac{\lambda}{\mu}\right)^{h-1} \sum_{k=1}^{h} k\left(\frac{\mu}{\lambda}\right)^{k-1}$$
$$= r - \frac{\lambda}{\nu}\left(1 - \left(\frac{\lambda}{\lambda+\nu}\right)^r\right) - \frac{\lambda}{\mu-\lambda}\alpha_1,$$

where α_1 is the α_1-service level for the net inventory position as defined in (42). It is derived next. With the same reasoning as above the α_1-service level of the two systems with the second definition of Y_{net} is

$$\alpha_1 = \mathsf{P}^0_{N_R}(Y_{0-} - X_{0-} > 0) = \sum_{h=1}^{r} \mathsf{P}^0_{N_R}(X_{0-} < h, Y_{0-} = h)$$
$$= \frac{\nu}{\lambda_R} K^{-1} \sum_{h=1}^{r}\sum_{n=0}^{h-1}\left(\frac{\lambda}{\mu}\right)^n C(h)$$
$$= 1 - \frac{1}{\mu-(\lambda+\nu)}\left((\mu-\lambda)\left(\frac{\lambda}{\lambda+\nu}\right)^r - \nu\left(\frac{\lambda}{\mu}\right)^r\right).$$

□

Proof of Corollary 5.3. Define a function g as follows

$$g : \mathbb{N}_0 \to \mathbb{R},$$
$$r \mapsto \frac{1}{\mu-(\lambda+\nu)}\left((\mu-\lambda)\left(\frac{\lambda}{\lambda+\nu}\right)^r - \nu\left(\frac{\lambda}{\mu}\right)^r\right).$$

Then, clearly $\alpha_1(r) = 1 - g(r)$. In order to show the assertion it suffices to proof that $(\frac{\lambda}{\mu})^r \le g(r) \le 1$ and that g is decreasing in r.

The first assertion can be proved by induction.

To show that g is strictly decreasing in r we directly compute

$$g(r-1) - g(r)$$
$$\ge \frac{\nu}{\lambda+\nu}\left(\frac{\lambda}{\mu}\right)^{r-1}\left(1 - \frac{\lambda}{\mu}\right) > 0 \quad \text{for all } r \in \mathbb{N}.$$

□

References

1. F. Baccelli and P. Brémaud, *Elements of queueing theory*. Springer, New York, 2003.
2. O. Berman and E. Kim, Stochastic models for inventory management at service facilities, Stochastic Models 15(4) (1999) 695–718.
3. O. Berman and E. Kim, Dynamic order replenishment policy in internet-based supply chains. Mathematical Models of Operations Research 53 (2001) 371–390.
4. O. Berman and K.P. Sapna, Inventory management at service facilities for systems with arbitrarily distributed service times, Stochastic Models 16 (3, 4) (2000) 343–360.
5. O. Berman and K.P. Sapna, Optimal control of service for facilities holding inventory, Computers and Operations Research 28 (2001) 429–441.
6. O. Berman and K.P. Sapna, Optimal service rates of a service facility with perishable inventory items, Naval Research Logistics, 49 (2002) 464–482.
7. P. Brémaud, van Kannurpatti, and R. Mazumdar, Event and time averages: A review, Advances in Applied Probability 24 (1992) 377–411.
8. D.J. Buchanan and R.F. Love, A (Q,R) Inventory model with lost sales and Erlang-distributed lead times, Naval Research Logistics Quarterly 32 (1985) 605–611.
9. F.M. Cheng and S.P. Sethi, Optimality of state-dependent (s,S) Policies in inventory models with markov-modulated demand and lost sales, Production and Operations Management 8 (1999) 183–192.
10. D. Gross and C.M. Harris, *Fundamentals of Queueing Theory*. 2nd edition. John Wiley and Sons, New York, 1985.
11. G. Hadley and T.M. Whitin, *Analysis of Inventory Systems*. Prentice Hall, Englewood Cliffs, N.J., 1963.
12. A.C. Hax and D. Candea, *Production and Inventory Management*. Prentice Hall, Englewood Cliffs, N.J., 1984.
13. F. P. Kelly, *Reversibility and Stochastic Networks*. John Wiley and Sons, Chichester—New York—Brisbane—Toronto, 1979.
14. E. Mohebbi and M.J.M. Posner, A continuous-review inventory system with lost sales and variable lead time, Naval Research Logistics, 45 (1998) 259–278.
15. H. Schneider, Effect of service-levels on order-points or order-levels in inventory models, International Journal of Production Research 19 (1981) 615–631.
16. K.C. Sevcik and I. Mitrani, The distribution of queueing networks states at input and output instants, Journal of the Association for Computing Machinery 28 (1981) 358–371.
17. K. Sigman and D. Simchi-Levi, Light traffic heuristic for an M/G/1 queue with limited inventory, Annals of Operations Research 40 (1992) 371–380.
18. E.A. Silver and R. Peterson, *Decision Systems for Inventory Management and Production Planning*. John Wiley and Sons, Inc., Chichester—New York—Brisbane—Toronto—Singapore, 1985.
19. R. Szekli, *Stochastic Ordering and Dependence in Applied Probability*. Springer-Verlag, New York—Berlin—Heidelberg, 1995.
20. H. Tempelmeier, Inventory control using a service constraint on the expected customer order waiting time. European Journal of Operational Research 19 (1985) 313–323.
21. H. Tempelmeier, Inventory service-levels in the customer supply chain, Operations Research Spektrum 22 (2000) 361–380.
22. R.J. Tersine, *Principles of Inventory and Materials Management*. 4th Ed. PTR Prentice Hall, Englewood Cliffs, N.J., 1994.