

Daduna, Hans

Article — Published Version

On queueing-inventory-location problems

Annals of Operations Research

Provided in Cooperation with:

Springer Nature

Suggested Citation: Daduna, Hans (2022) : On queueing-inventory-location problems, Annals of Operations Research, ISSN 1572-9338, Springer US, New York, NY, Vol. 331, Iss. 2, pp. 679-710, <https://doi.org/10.1007/s10479-022-04839-5>

This Version is available at:

<https://hdl.handle.net/10419/312860>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>



On queueing-inventory-location problems

Hans Daduna¹

Accepted: 14 June 2022 / Published online: 19 July 2022
© The Author(s) 2022

Abstract

A network of queueing-inventory systems is considered where the inventories are replenished by a common central server. Travel times for transport of the items send out from the center and an adaptive dispatching regime to direct transportation towards the locations with queueing-inventory systems are incorporated in the model. The stationary distribution of the system is of product form and is explicitly given. An optimization procedure is developed to find for given locations of the queueing-inventory systems an optimal location for the replenishment center. Optimization is with respect to overall utilization of the resources measured in total throughput of the queueing-inventory systems.

Keywords Queueing networks · Inventory · Base stock policy · State-dependent routing · Location analysis · Throughput

Mathematics Subject Classification 60K25 · 68M20 · 90B22 · 90B05 · 90B06 · 90B85

1 Introduction

In standard OR-literature queueing theory and inventory theory usually are dealt with as different fields of research, see e.g. the volume “Stochastic Models” in the series “Handbooks in Operations Research and Management Science” Heyman and Sobel (1990) with separated sections on “Queueing Theory” Cooper (1990) and “Stochastic Inventory Theory” Porteus (1990). As the editors put it “Queueing theory is more descriptive than normative, inventory is the reverse,...” Heyman and Sobel (1990)[p. ix].

On the other side, planning and analysing complex systems in connection with nowadays large supply chains calls for models which encompass integration of production processes and inventory holding. Increasing complexity of supply chains requires integration of other areas of OR, e.g. reliability, logistics (transportation), and location analysis, see the discussion in Heckmann and Nickel (2019).

Realizing the need for models which integrate congestion phenomena (queueing) and inventory control, over the last 30 years queueing-inventory models of various structures and different level of complexity have been developed, see the review of selected literature below.

✉ Hans Daduna
hans.daduna@uni-hamburg.de

¹ Department of Mathematics, Universität Hamburg, Bundesstrasse 55, T17, 20146 Hamburg, Germany

But observing the complexity of supply chains or production plants of today, it is obvious that integration of queueing and inventory aspects only will not meet all the features inherent in these systems.

The aim of this article is to integrate into queueing-inventory systems spatial aspects of distributed queueing-inventory systems. Starting this project it is immediately visible that aspects of transport have to be integrated as well. Integration of location decisions, logistics activities, production, and inventory management generates problems with decision making of different levels: Strategic, tactical, and operational decisions are mixed up. Heckmann and Nickel emphasize that "...making location decisions ignoring primary logistics activities like production or distribution may result in excessive costs incurred throughout the lifetime of the facilities supporting the logistics system." Heckmann and Nickel (2019)[p. 456].

In modelling supply chains the areas of queues, inventories, location analysis constitute important sub-problems which from the very definition of a supply chain are strongly interconnected. Nevertheless, due to the system's overall complexity often the related parts of production, storage, transportation, and decision for location of facilities are separated to obtain feasible (optimization-) problems where classical solution procedures from the respective fields could be applied to solve the isolated sub-problems. This separation is often justified by considering the location, production, inventory, and transport control problems as being problems that occur on different levels of decision making: Strategic, tactical, and operational scales of management.

On the other side, for several of the mentioned composite problems it has been shown recently that separation of the partial problems leads to sub-optimal solutions of the relevant global optimization problems, see for separated location-transportation control problems Salhi and Rand (1989).

Our aim is to construct a network model which integrates aspects of production with associated inventories, transport, and location of a common replenishment system (for to satisfy reorders of inventories) under random influences. Incorporating randomness excludes to consider many details which are part of a deterministic modeling process as it is presented in Cordeau et al. (2006) and Song and Wu (2022).

Main contributions of the paper. For the sketched network model

- we compute the steady state distribution in explicit product form, i.e. the state process is separable,
- we find an optimal location for the central replenishment server (supporting strategic decisions),
- we determine the needed inventory to satisfy on a make-to-order basis the arriving demand (supporting tactical decisions),
- we construct an adaptive dispatching regime for to send out items from the replenishment station (supporting operational decisions).

Structure of the paper. In Sect. 2 the general problem and its components are described in more detail. In Sect. 3 a short review of related literature is given. The formal model with all details is presented in Sect. 4. Section 4.1 is devoted to the analysis of the system in steady state with short sections on cost analysis in Sect. 4.2 and generalizations in Sect. 4.3. The location problem for the central unit is described in Sect. 5. A reduced model is developed in Sect. 5.1 which enables to determine the optimal location of the center and the needed inventory under optimal location in Sect. 5.2. Conclusions are summarized in Sect. 6. In Appendix 1 technical necessities are reviewed.

Notations and conventions. $(x)_+ := \max(x, 0)$ for $x \in \mathbb{R}$.

Empty sums are 0, empty products are 1. $\mathbb{N}_0 = \{0, 1, 2, \dots\}$, $\mathbb{N}_+ = \{1, 2, \dots\}$.

$\bar{J} = \{1, \dots, J\}$, and $\bar{J}^+ = \{1, \dots, J, J + 1\}$ for $J \in \mathbb{N}_+$.

Increasing \equiv non-decreasing, decreasing \equiv non-increasing.

All random variables and processes which occur are defined on a common probability space (Ω, \mathcal{F}, P) . Expectations under P are expressed as $\mathbb{E}(\cdot)$.

For a probability measures π on a discrete space we write shortly $\mathbb{E}_\pi(\cdot)$ for expectations under π .

2 Problem statement

We investigate a system which consists of a set of production facilities, each with an associated inventory (\equiv queueing-inventory system). The positions of the locations with production-inventory system are known and fixed in the plane. We seek a (nearly) optimal position for a common replenishment facility according to a suitable optimization criterion which incorporates the effects of transportation, varying demand, and holding inventories.

To be more specific: We are given $J \geq 1$ queueing-inventory systems, numbered $1, \dots, J$ at positions $a_j = (a_{j1}, a_{j2})$ in the plane. At each location the production unit is represented by a single server with ample waiting room under First-Come-First-Serve (FCFS) regime. External demand (represented by customers) at location j follows a Poisson- λ_j stream. Production is according to customers' local demand on a make-to-order basis and to satisfy a customer's demand the production server needs one (single) item of raw material from the associated inventory. For the inventory a base-stock level is prescribed which is location specific.

If a customer's service is completed the customer departs from the system and the consumed item from the inventory is removed formally from the inventory at this instant. Concurrently an order for a new item of raw material for this location is placed which arrives at the replenishment center without time delay. So replenishment orders from the inventories follow local continuous review base-stock policies.

The items to be stored in the inventories are manufactured by a central replenishment unit which is represented by a single server with ample waiting room under FCFS. Whenever an item of raw material is produced it is send out to a location according to an adaptive dispatching regime which will be specified below. We assume that items are exchangeable. This implies that we can decide about the target location when the item is produced.

Customers' behaviour on arrival at the locations can be described as following a combined backordering/lost-sales principle: Customers who find on arrival an empty queueing system and an item on stock in the local inventory enter service immediately. If on arrival there is already a customer in service and the inventory is not empty these new customers enter the system and queue up (backordering). If the inventory is depleted every new arrival is rejected by the system (lost sales).

The local production-inventory systems with central supplier constitute a standard (integrated) queueing-inventory problem as investigated e.g. in Otten et al. (2016). We add to this model a location-transportation problem by integrating the transport of the manufactured items from the replenishment center to the locations and a decision for locating the center.

Transportation is by trucks which transport single items, starting when an item is send out. For modelling this feature we use infinite server systems which are standard devices of queueing theory to model transportation systems. We assume that there is ample transportation capacity, i.e. whenever an item is manufactured and ready to be send out a truck is available.

Such an integrated system is termed “logistics network” in Cordeau et al. (2006)[p. 60]: “A logistics network is a set of suppliers, manufacturing plants and warehouses organized to manage the procurement of raw materials, their transformation into finished products, and the distribution of finished products to customers.” In Cordeau et al. (2006) a deterministic one-period optimization problem is formulated and two approaches for solving the problem are described. Our approach is completely different. The main goals of our investigation are

- to find specifications of the system which allow to determine explicitly the asymptotic and stationary behaviour of the system which is driven by stochastic processes, especially random arrival processes of demand, stochastic service and production times, random travel times for delivering the replenishment items, and
- to find an optimal location for the central replenishment center which supplies the predetermined locations with items to be held on stock for the ongoing production processes, and
- to include an adaptive routing regime that decides automatically which orders for replenishment by the locations should be satisfied next, and
- to include (different from e.g. Cordeau et al. (2006)) the time dimension in our optimization procedure in the sense of Tapiero (1971). We evaluate the time average of local total production (\equiv satisfied external demand) over a long time horizon which due to stabilization of the system is computed as steady state throughput.

Summarizing, our model integrates strategic (facility location) and tactical/operational (allocation, scheduling, service, stock holding) aspects of decision making. These integrated logistics-location-service models occur e.g. in supply chain planning and operation, see e.g. Melo et al. (2009) and Heckmann and Nickel (2019). Clearly, there arise many difficulties with such integration procedures in combined location-routing-service problems. A typical example is discussed in Min et al. (1998)[p.10].

Our integrated queueing-inventory-location problems (QILPs) are related to location-allocation problems, queueing-inventory problems (QIPs), location-routing problems (LRPs), location-inventory problems (LIPs), location-inventory-routing problems (LIRPs), and transportation-location-allocation problems under random influences.

3 Literature review

The interplay of location analysis and decisions with aspects and problems from other areas of OR has been investigated and described in many research articles. We restrict our description to integrated models under random influences, i.e. the abundant literature around (one-period) deterministic optimization models will not be considered here.

Queueing-location problems, i.e., investigation and optimization of combined queueing and location models is a well established part of Operations Research. Larson (1974) initiated research on location problems within the scope of queueing systems, follow up papers in the context of discrete location problems are e.g. by Larson, Berman and coauthors (e.g. Berman et al. (1985, 1987)). Relevant survey chapters in collections are Drezner and Hamacher (2004)[Chapter 11], Mirchandani and Francis (1990)[Chapter 13], (vehicle routing problems under stochastic side constraints). A recent very detailed survey is Berman and Krass (2019). Two main research directions are:

(1) Clients are fixed, servers move. In Drezner et al. (1990) and Scott et al. (1999) the authors consider how to locate mobile servers in a plane to serve demands which occur as Poisson processes at fixed locations. The mobile servers are modelled as standard queueing

systems. Travel times of the mobile servers to and from the clients are incorporated in the service times, a survey is Berman and Krass (2004). An investigation of determining locations in case of emerging queueing phenomena is Dan and Marcotte (2019).

(2) Servers are fixed, clients move. See e.g. Berman and Drezner (2007), Aboolian et al. (2008, 2009) with additional references. Demand is generated locally according to renewal processes, the service is provided by $M/M/k/\infty$ systems. A detailed review is Berman and Krass (2019).

Location-inventory problems (LIPs) are often summarized under the heading “location analysis”, which means to find optimal locations for warehouses or central suppliers. LIPs aim “to integrate strategic supply chain decisions with tactical and operational inventory management decisions”, see Farahani et al. (2015) with a survey of research on basic LIPs and more evolved variants. A basic LIP is described in Farahani et al. (2015)[Section2].

Location routing problems (LRPs) combine location analysis and vehicle routing decision. First principles of vehicle routing are described in Laporte (1988). A recent survey mainly with locations on networks is Albareda-Sambola and Rodriguez-Pereira (2019). Closer to location in the plane are Salhi and Nagy (2009) and Manzour-al-Ajdad et al. (2012).

Location-inventory-routing problems (LIRPs) as defined in Song and Wu (2022) constitute an integrated approach to model parts of a supply chain, including location of distribution centers, inventory holding, allocation decisions, and routing schedules for transportation. Predecessors of that investigation are summarized in Song and Wu (2022) [p. 3].

Integration of strategic and tactical/operational aspects of planning is a common topic of almost all of the mentioned work. In Salhi and Rand (1989) for LRPs it is shown that separating decisions on location and routing can lead to sub-optimal decisions. Under the heading “Why logistics matters in location modelling” this problem is discussed indepth in Heckmann and Nickel (2019) [Section 6.2], stating as “main conclusion ...that making location decisions ignoring primary logistics activities ...may result in excessive costs”

Literature on queueing-inventory problems is overwhelming as can be seen from the recent review Krishnamoorthy et al. (2021). Earlier reviews with additional comments on specific aspects of queueing-inventory systems are Krishnamoorthy et al. (2011) and the “Short review of known results” in Melikov et al. (2016). This points out that the integrated model which was introduced to the research community only 30 years ago independently of another in Melikov and Molchanov (1992) and Sigman and Simchi-Levi (1992) is of interest to theoretical research and practical applications. Instead of repeating the review in Krishnamoorthy et al. (2021) I refer only to fundamental methodological approaches here.

A first branch exploits *approximation procedures*, Melikov and Molchanov (1992) and Sigman and Simchi-Levi (1992) are typical examples. Realizing that the global balance equations for the dedicated composite models do not have an easy to find explicit solution the authors turn to suitable procedures to obtain approximate solutions for the steady state distribution. This technique reoccurs in many other articles, e.g. by Melikov, Molchanov, Ponomarenko, Koroliuk, Bagirova, et al. as demonstrated in Krishnamoorthy et al. (2021).

A second branch of methods to attack queueing-inventory systems are Markov decision processes and stochastic dynamic optimization which are standard methods to obtain optimal policies in classical inventory theory, see Puterman (1990). Especially, Berman and coworkers applied these techniques to obtain structural information on optimal policies to control queueing-inventory systems, see e.g. Berman and Kim (1999), Berman and Sapna (2001).

A third branch utilizes the state space structure of single queues with inventory as being a product $\mathbb{N}_0 \times F$ with queue length in \mathbb{N}_0 and with F counting for the inventory size (possibly plus some external environment). Any Markov process which fits into this class can be

described as *Quasi-birth-death process (QBD)* and the steady state distribution (if it exists) can be obtained numerically with matrix-geometrical methods developed and described for this class of processes in Neuts (1981), Latouche and Ramaswami (1999). This applies especially if the arrival process of demands is not Poisson. Seemingly, most of the results published on queueing-inventory systems are obtained via matrix-geometrical methods, see Krishnamoorthy et al. (2021), to name a few authors we refer to the work of the group around Krishnamoorthy, Chakravarthy, Shajin, Narayanan, Lakshmy, et al.—for details see Krishnamoorthy et al. (2021).

A fourth branch of articles exploits internal structures of the describing Markov processes for the queueing-inventory systems. These processes exhibit in the internal structures of the state processes local or partial balance properties leading to surprising simple stationary distributions which are called *product form equilibrium*, which means that asymptotically and in equilibrium the joint distribution of the queueing and the inventory component decouple (\equiv being stochastically independent). According to Krishnamoorthy et al. (2021) the number of articles where this characterization applies is minor compared to branches one and three. Contributions to this field started with Schwarz et al. (2006, 2007), and are continued among others with Krishnamoorthy and Narayanan (2013), Shajin et al. (2018) and other members of the group around Krishnamoorthy, and with Saffari et al. (2013) and some further articles by Saffari, Haji, and their coworkers.

Remarkably, almost all work is on single servers with an attached inventory. The investigation of networks of queues with attached inventory is just in the beginning. The research presented in this article is related to investigation of network models in Otten et al. (2016, 2020), Otten (2017)[Section 3.4].

4 The queueing-inventory-location model

We have a set of locations $\bar{J} := \{1, \dots, J\}$ located in the plain at $a_j = (a_{j1}, a_{j2})$, each equipped with an exponential single server as production server with infinite waiting room under First-Come-First-Served (FCFS) and an attached inventory of maximal size $b_j \geq 1$, $j = 1, \dots, J$. The service rates $\mu_j(\cdot)$ of the production servers depend on the (local) queue length. If at location j are $n_j > 0$ customers present (in service or waiting) then $\mu_j(n_j) > 0$. If necessary we set $\mu_j(0) = 0$. We assume throughout that the $\mu_j(n_j)$ are non-decreasing in $n_j \in \mathbb{N}_0$. Recall that service of a customer (production of a demanded unit) at j needs one item from inventory which is consumed formally at the instant when service of this customer is finished. This customer leaves the system immediately and at the same time instant an order for manufacturing a new item arrives at the replenishment station (base stock policy).

The replenishment station (identified by number $J + 1$) is located at $x = (x_1, x_2)$ and is an exponential single server with infinite waiting room under FCFS with service rate $\nu > 0$. An item produced at $J + 1$ is dispatched to one of the stations, say j , which are not satisfied with items in the local inventory (k_j items on stock at j) or on the way to j (m_j items on transport to j) with $m_j + k_j < b_j$. This restriction is due to the maximal capacity b_j ($=$ base-stock level) at j . We incorporate in the dispatching rules the condition that the total stock (\equiv inventory position or system stock in standard inventory systems with backordering Porteus (1990)[p. 605]) at location j includes the items already on transport to j . (We do not include items which are ordered but not produced and on transport to the respective location because this would in case of base-stock policy under lost sales always sum up to b_j .)

Given an item is on leave from station $J + 1$, station j is selected with branching probability $r_{J+1,j}(z)$ (to be defined below) if the global state of the queueing-inventory-transportation system is $z \in E$ (\equiv state space).

Transportation of an item by a truck from center $J + 1$ at x to location j at a_j is modelled by serving the truck (as a customer) at an exponential infinite server (ample service capacity). For simplicity of presentation we assume that trucks travel with unit speed, i.e. if the distance between $J + 1$ and j is $d_j(x) := d(x, a_j)$, the mean travel (service) time is $d_j(x)$. Here $d(\cdot)$ is any suitable distance function (metric). So, we assume that travel times are exponentially distributed with rate $d_j(x)^{-1}$. Modeling traffic on a lane by infinite server queueing system is a standard device from queueing theory in transportation theory, see e.g. Newell (1982)[Section 6]. (In the algorithm to determine necessary inventory capacity to fulfill predetermined demand realistic speeds will be incorporated, see Algorithm 5.13.)

We construct a multivariate Markov process $Z = (Z(t) : t \geq 0)$ which describes the evolution over time of this system. The state space of Z is

$$E := \{(m_1, k_1, n_1; \dots m_j, k_j, n_j; \dots; m_J, k_J, n_J) : m_j, k_j, n_j \in \mathbb{N}_0, 0 \leq m_j + k_j \leq b_j, j = 1, \dots, J\} \quad (4.1)$$

A typical local state of location j at time $t \geq 0$ is

$$(V_j(t), Y_j(t), X_j(t)) = (m_j, k_j, n_j). \quad (4.2)$$

The meaning of (4.2) is: $V_j(t) = m_j$ items are on transport with destination j , the size of the onhand inventory at j is $Y_j(t) = k_j$, and the queue length is $X_j(t) = n_j$ at the production server at j . The queue length at the central replenishment server $J + 1$ is implicitly determined as $\sum_{j=1}^J (b_j - m_j - k_j)$.

Imposing the usual (conditional) independence assumptions on the arrival streams, service and transport times, and routing (dispatching) decisions, it follows by standard arguments that

$$Z = (Z(t) : t \geq 0), \text{ with } Z(t) = ((V_j(t), Y_j(t), X_j(t)), j = 1, \dots, J), \quad (4.3)$$

is a strong Markov process with countable state space E from (4.1).

Summarizing: The model encompasses a set of J exponential single server queues representing production which are coupled with a closed network representing support of the production. There are $b = b_1 + \dots + b_J$ customers cycling in the closed network according to problem specific rules. These customers represent throughout one cycle different objects (have alternating identities): Orders from an inventory waiting at the replenishment station, trucks carrying new items, and items stored in an inventory.

State dependent routing with blocking (here due to $V_j(t) + Y_j(t) \leq b_j$) poses specific difficulties when separability is an aim of the modelling process. The following definition follows ideas of Towsley (1980) who constructed queueing networks with state dependent routing protocols which exhibit detailed balance and product form steady state in the sense of Gordon-Newell and Jackson networks.

Definition 4.1 The routing probabilities $R = (r_{J+1,j}(\cdot) : j = 1, \dots, J)$ for a truck (with an item) on leave from the replenishment center are

$$r_{J+1,j}(m_1, k_1; \dots m_j, k_j; \dots; m_J, k_J) := h_j(m_j + k_j) \cdot h(m_1 + k_1 + \dots + m_j + k_j + \dots + m_J + k_J), \quad (4.4)$$

for destination $j \in \{1, \dots, J\}$, with

$$h_j(m) = (b_j - m)_+, \quad m \in \mathbb{N}_0 \quad \text{for } j = 1, \dots, J, \quad (4.5)$$

$$h(m) = \left[\left(\sum_{j=1}^J b_j \right) - m \right]_+^{-1}, \quad m \in \mathbb{N}_0, \quad (4.6)$$

for strictly positive integers b_1, \dots, b_J .

Remark 4.2 The notation $h_j(\cdot)$ is used for easier reading. In fact these functions are functions of the destination branch j and the base stock level b_j :

$$h_j(n) = h_{j,b_j}(n) = (b_j - n)_+, \quad n \geq 0.$$

Similarly, $h(\cdot) = h_b(b - \cdot)$ with $b = b_1 + \dots + b_J$. The extended notion will be used if necessary.

Towsley (1980)[Theorem 3] showed that the above functions can be taken as $h_j(n) = C \cdot n + d_j$, $j = 1, \dots, J$, and $1h(n) = C \cdot n + d_1 + \dots + d_J$ for positive constants C and d_j . In Daduna (1985) it was shown that in a simple star-like network $C \in \mathbb{R}$ may be allowed, which introduced a blocking scheme as needed for the more general network in the present investigation.

Definition 4.3 The infinitesimal generator (intensity matrix) of Z is denoted by $\mathbf{Q} = (q(z, z') : z, z' \in E)$ with the following positive rates. (For readability we only depict the relevant local states for the respective transitions, e.g. we write $(\dots m_j, k_j, n_j; \dots)$ instead of $(m_1, k_1, n_1; \dots m_j, k_j, n_j; \dots m_J, k_J, n_J)$.)

For $j = 1, \dots, J$ we have

$$\begin{aligned} q((\dots m_j, k_j, n_j; \dots), (\dots m_j + 1, k_j, n_j; \dots)) \\ &= 1_{(m_j + k_j < b_j)} \cdot v \cdot r_{J+1,j}(\dots m_j, k_j; \dots), \\ q((\dots m_j, k_j, n_j; \dots), (\dots m_j - 1, k_j + 1, n_j; \dots)) &= 1_{(m_j > 0)} \cdot d_j(x)^{-1} \cdot m_j, \\ q((\dots m_j, k_j, n_j; \dots), (\dots m_j, k_j, n_j + 1; \dots)) &= 1_{(k_j > 0)} \cdot \lambda_j, \\ q((\dots m_j, k_j, n_j; \dots), (\dots m_j, k_j - 1, n_j - 1; \dots)) &= 1_{(n_j > 0, k_j > 0)} \cdot \mu_j(n_j). \end{aligned}$$

The diagonal elements $q(z, z)$ of \mathbf{Q} are selected in a way that row sums are zero. All other rates $q(z, z')$, $z \neq z'$, are zero.

4.1 Steady-state analysis

Performance evaluation for the queueing-inventory-location model is either based on long time evaluation with averaging over time or based on observing the system in steady state (if the system is stable). For a system described by a Markov process these methods yield the same performance metrics if the process is ergodic.

Theorem 4.4 The Markov process Z is ergodic if and only if for all $j \in \bar{J} = \{1, \dots, J\}$ holds

$$\sum_{n_j=0}^{\infty} \prod_{i=1}^{n_j} \frac{\lambda_j}{\mu_j(i)} < \infty. \quad (4.7)$$

If Z is ergodic its stationary distribution π is with normalization constant $G(b_1, \dots, b_J; \bar{J})$

$$\begin{aligned} & \pi(m_1, k_1, n_1; \dots m_j, k_j, n_j; \dots m_J, k_J, n_J) \\ &= \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v d_\ell(x)}{i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\} \\ & \cdot \left(\prod_{i=0}^{m_1+k_1+\dots m_J+k_J-1} h(i) \right) \cdot G(b_1, \dots, b_J; \bar{J})^{-1}. \end{aligned} \quad (4.8)$$

If Z is ergodic the normalization constant is

$$\begin{aligned} G(b_1, \dots, b_J; \bar{J}) &= \sum_{n_1=0}^{\infty} \dots \sum_{n_J=0}^{\infty} \left(\prod_{\ell=1}^J \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right) \cdot H(b_1, \dots, b_J; \bar{J}), \\ \text{with } H(b_1, \dots, b_J; \bar{J}) &= \sum_{(m_1, k_1; \dots m_J, k_J): 0 \leq m_j + k_j \leq b_j, j \in \bar{J}} \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v d_\ell(x)}{i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \right\} \\ & \cdot \left(\prod_{i=0}^{m_1+k_1+\dots m_J+k_J-1} h(i) \right). \end{aligned} \quad (4.9)$$

Proof Abbreviating $b := b_1 + \dots + b_J$ and $d_\ell(x) =: \eta_\ell^{-1}$ the global balance equations for Z are with unknowns $x(z)$, $z \in E$,

$$\begin{aligned} & x(m_1, k_1, n_1; \dots m_j, k_j, n_j; \dots m_J, k_J, n_J) \left[\underbrace{v 1_{(m_1+k_1+\dots+m_J+k_J < b)}}_{=:(a)} \right. \\ & \left. + \sum_{j=1}^J \left(\underbrace{1_{(m_j > 0)} \eta_j m_j}_{=:(b)} + \underbrace{1_{(k_j > 0)} \lambda_j}_{=:(c)} + \underbrace{1_{(n_j > 0, k_j > 0)} \mu_j(n_j)}_{=:(d)} \right) \right] \end{aligned} \quad (4.10)$$

$$\begin{aligned} &= \sum_{j=1}^J 1_{(m_j > 0)} x(m_1, k_1, n_1; \dots m_j - 1, k_j, n_j; \dots m_J, k_J, n_J) \\ & \quad v \cdot r_{J+1, j}(m_1, k_1; \dots m_j - 1, k_j; \dots m_J, k_J) \end{aligned} \quad (4.11)$$

$$\begin{aligned} &+ \sum_{j=1}^J 1_{(k_j > 0)} x(m_1, k_1, n_1; \dots m_j + 1, k_j - 1, n_j; \dots m_J, k_J, n_J) \\ & \quad \cdot \eta_j \cdot (m_j + 1) \end{aligned} \quad (4.12)$$

$$\begin{aligned} &+ \sum_{j=1}^J 1_{(n_j > 0, k_j > 0)} x(m_1, k_1, n_1; \dots m_j, k_j, n_j - 1; \dots m_J, k_J, n_J) \cdot \lambda_j \end{aligned} \quad (4.13)$$

$$\begin{aligned} &+ \sum_{j=1}^J 1_{(k_j < b_j)} x(m_1, k_1, n_1; \dots m_j, k_j + 1, n_j + 1; \dots \\ & \quad \dots m_J, k_J, n_J) \cdot \mu_j(n_j + 1) \end{aligned} \quad (4.14)$$

In a first step we show that a non-normalized solution of this system of global equilibrium equations is

$$\begin{aligned}
 & x(m_1, k_1, n_1; \dots m_j, k_j, n_j; \dots m_J, k_J, n_J) \\
 &= \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\} \\
 & \cdot \left(\prod_{i=0}^{m_1+k_1+\dots m_J+k_J-1} h(i) \right), \quad (m_1, k_1, n_1; \dots m_J, k_J, n_J) \in E \quad (4.15)
 \end{aligned}$$

For $j \in \{1, \dots, J\}$ we obtain in (4.11) (inserting the definition of the routing probabilities)

$$\begin{aligned}
 & 1_{(m_j>0)} x(m_1, k_1, n_1; \dots m_j - 1, k_j, n_j; \dots m_J, k_J, n_J) \\
 & \cdot v \cdot h_j(m_j - 1 + k_j) h \left(\left(\sum_{i=1}^J m_i + k_i \right) - 1 \right) \\
 &= 1_{(m_j>0)} \prod_{\ell=1, \ell \neq j}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\} \\
 & \cdot \left\{ \left(\prod_{i=0}^{m_j-1+k_j-1} h_j(i) \right) \left(\prod_{i=1}^{m_j-1} \frac{v}{\eta_j \cdot i} \right) \left(\frac{v}{\lambda_j} \right)^{k_j} \left(\prod_{i=1}^{n_j} \frac{\lambda_j}{\mu_j(i)} \right) \right\} \\
 & \cdot \left(\prod_{i=0}^{m_1+k_1+\dots m_j-1+\dots-1} h(i) \right) \cdot v \cdot h_j(m_j - 1 + k_j) h \left(\left(\sum_{i=1}^J m_i + k_i \right) - 1 \right)
 \end{aligned}$$

which after simple manipulation is (b).

For $j \in \{1, \dots, J\}$ we obtain in (4.12)

$$\begin{aligned}
 & 1_{(k_j>0)} x(m_1, k_1, n_1; \dots m_j + 1, k_j - 1, n_j; \dots m_J, k_J, n_J) \eta_j \cdot (m_j + 1) \\
 &= 1_{(k_j>0)} \prod_{\ell=1, \ell \neq j}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\} \\
 & \cdot \left\{ \left(\prod_{i=0}^{m_j+1+k_j-1-1} h_j(i) \right) \left(\prod_{i=1}^{m_j+1} \frac{v}{\eta_j \cdot i} \right) \left(\frac{v}{\lambda_j} \right)^{k_j-1} \left(\prod_{i=1}^{n_j} \frac{\lambda_j}{\mu_j(i)} \right) \right\} \\
 & \cdot \left(\prod_{i=0}^{m_1+k_1+\dots m_j+1+k_j-1+\dots-1} h(i) \right) \cdot \eta_j (m_j + 1)
 \end{aligned}$$

which after simple manipulation is (c).

For $j \in \{1, \dots, J\}$ we obtain in (4.13)

$$\begin{aligned}
 & 1_{(n_j>0, k_j>0)} x(m_1, k_1, n_1; \dots m_j, k_j, n_j - 1; \dots m_J, k_J, n_J) \cdot \lambda_j \\
 &= 1_{(n_j>0, k_j>0)} \prod_{\ell=1, \ell \neq j}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\}
 \end{aligned}$$

$$\cdot \left\{ \left(\prod_{i=0}^{m_j+k_j-1} h_j(i) \right) \left(\prod_{i=1}^{m_j} \frac{v}{\eta_j \cdot i} \right) \left(\frac{v}{\lambda_j} \right)^{k_j} \left(\prod_{i=1}^{n_j-1} \frac{\lambda_j}{\mu_j(i)} \right) \right\} \\ \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_j+k_j-1} h(i) \right) \cdot \lambda_j$$

which after simple manipulation is (d).

Writing (4.14) in detail yields

$$\sum_{j=1}^J 1_{(k_j < b_j)} x(m_1, k_1, n_1; \dots, m_j, k_j + 1, n_j + 1; \dots, m_J, k_J, n_J) \cdot \mu_j(n_j + 1) \\ = \sum_{j=1}^J 1_{(k_j < b_j)} \prod_{\ell=1, \ell \neq j}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\} \\ \cdot \left\{ \left(\prod_{i=0}^{m_j+k_j+1-1} h_j(i) \right) \left(\prod_{i=1}^{m_j} \frac{v}{\eta_j \cdot i} \right) \left(\frac{v}{\lambda_j} \right)^{k_j+1} \left(\prod_{i=1}^{n_j+1} \frac{\lambda_j}{\mu_j(i)} \right) \right\} \\ \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_j+k_j+1+\dots+m_J+k_J-1} h(i) \right) \cdot \mu_j(n_j + 1) \\ = \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \left(\prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \right) \right\} \\ \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_j+k_j+\dots+m_J+k_J-1} h(i) \right) \cdot v \\ \cdot \underbrace{\sum_{j=1}^J 1_{(k_j < b_j)} h_j(m_j + k_j) h \left(\sum_{i=1}^J (m_i + k_i) \right)}_{(\star)}.$$

If $(m_j + k_j = b_j)$ holds for all $j = 1, \dots, J$, the sum (\star) equals 0 because $h_j(m_j + k_j) = (b_j - m_j - k_j)_+$, and with $m_1 + k_1 + \dots + m_J + k_J = 0$ this coincides with (a) in this case.

If $(m_j + k_j < b_j)$ for some $j \in \{1, \dots, J\}$ then it holds $1_{(m_1+k_1+\dots+m_J+k_J < b)} = 1$ and it follows with $(m_j + k_j < b_j) \Rightarrow (k_j < b_j)$

$$(\star) = \sum_{j=1}^J 1_{(k_j < b_j)} (b_j - (m_j + k_j))_+ \cdot \left(\sum_{i=1}^J (b_i - (m_i + k_i)) \right)_+^{-1} = 1,$$

which coincides with (a) in this case. This finishes the first part of the proof.

Because irreducibility of Z follows directly from the definition of E and routing R in (4.4) summability of $(x(z) : z \in E)$ guarantees ergodicity of Z . Direct computation separates terms associated to queue lengths, i.e. $\prod_{j=1}^J \left(\prod_{i=1}^{n_j} \frac{\lambda_j}{\mu_j(i)} \right)$ in (4.9). Because $H(b_1, \dots, b_J; \bar{J})$ is a finite sum this finishes the second part of the proof. \square

Consider the system from Theorem 4.4 with transportation times = 0. This is a classical starlike system of queues with additional attached inventories at the branches. For this system Otten (2017)[Theorem 3.4.4] has found the stationary distribution in product form under the branching regime (termed “weak priorities”) from Daduna (1985).

Corollary 4.5 Denote by

$$\pi_\ell(n_\ell) := \prod_{i=1}^{n_\ell} \frac{\lambda_\ell}{\mu_\ell(i)} \cdot G_\ell^{-1}, \quad n_\ell \in \mathbb{N}_0, \quad (4.16)$$

the stationary distribution of an exponential single server queue with Poisson- λ_ℓ arrival stream and state dependent service rates $\mu_\ell(\cdot)$. Then the result of Theorem 4.4 for $(m_1, k_1, n_1; \dots, m_J, k_J, n_J) \in E$ can be expressed as

$$\pi(m_1, k_1, n_1; \dots, m_J, k_J, n_J) = \left(\prod_{\ell=1}^J \pi_\ell(n_\ell) \right) \cdot \theta(m_1, k_1; \dots, m_J, k_J) \quad (4.17)$$

where

$$\begin{aligned} & \theta(m_1, k_1; \dots, m_J, k_J) \\ &= \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v d_\ell(x)}{i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \right\} \\ & \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_J+k_J-1} h(i) \right) \cdot H(b_1, \dots, b_J; \{1, \dots, J\})^{-1} \end{aligned} \quad (4.18)$$

is a probability measure on state space

$$E_o = \{(m_1, k_1; \dots, m_J, k_J) : m_j, k_j \in \mathbb{N}_0, 0 \leq m_j + k_j \leq b_j, j = 1, \dots, J\}.$$

Corollary 4.5 says that the stationary distribution of Z is separable with factors π_ℓ , $\ell = 1, \dots, J$, and θ . Separability opens the path for successful investigation of a system because several performance metrics are then directly accessible. We shall henceforth denote by $(V_1, Y_1, X_1; \dots, V_J, Y_J, X_J)$ a random vector which is distributed according to the stationary distribution (4.8), respectively (4.17).

Possibly, the most important performance metric for production systems is the throughput which quantifies the total output of the system. It is defined either as a time average over a long time horizon or as a steady state measure per time unit. Roughly, the throughput quantifies the utilization of the system’s resources. More formally, we define

$$D_j(b_1, \dots, b_J; x)(T) = \int_0^T \mu_j(X_j(t)) 1_{(X_j(t)>0, Y_j(t)>0)} dt$$

as the total amount of served demand at location j within time horizon $[0, T]$ with base stock levels b_1, \dots, b_J and the center being located at x . This definition incorporates (contrasting to the standard one-period optimization procedures in deterministic location-inventory problems) a time dimension in the sense of Tapiero (1971) which is adequate for strategic decisions which are concerned e.g. with decision making for the replenishment center.

The relevant performance criterion for the composite system is then the asymptotic time average of total satisfied demand

$$D(b_1, \dots, b_J; x) := \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{j=1}^J D_j(b_1, \dots, b_J; x)(T).$$

Because we consider a stable (ergodic) system described by a Markov process the asymptotic time average is the stationary amount of served demand over all locations per time unit which is

$$TH(b_1, \dots, b_J; x) := \mathbf{E}_\pi \left(\sum_{j=1}^J \mu_j(X_j) \cdot 1_{(X_j > 0, Y_j > 0)} \right). \quad (4.19)$$

For readability we shall often abbreviate $TH := TH(b_1, \dots, b_J; x)$ and similarly other terms.

Theorem 4.6 *The steady state mean number of served customers (demand) per time unit at location $j \in \{1, \dots, J\}$ is the local throughput at j :*

$TH_j = \mathbf{E}_\pi (\mu_j(X_j) \cdot 1_{(X_j > 0, Y_j > 0)})$. *The total throughput is $TH = \sum_{j=1}^J TH_j$. It holds*

$$TH_j = v \cdot \frac{b_j}{b_1 + \dots + b_J} \frac{H(b_1, \dots, b_j - 1, \dots, b_J, \{1, \dots, J\})}{H(b_1, \dots, b_j, \dots, b_J, \{1, \dots, J\})} \quad (4.20)$$

Proof To simplify notation we consider the case $j = J$ and use the abbreviation $d_\ell(x) =: \eta_\ell^{-1}$ again. The product form of the steady state distribution applies in (\star) below to cancel the queue lengths related terms together with $1_{(n_J > 0)} \cdot \mu_J(n_J)$ and yields

$$\begin{aligned} TH_J &= \sum_{z=(\dots, m_J, k_J, n_J) \in E} \pi(z) \cdot 1_{(n_J > 0, k_J > 0)} \cdot \mu_J(n_J) \\ &\stackrel{(\star)}{=} \sum_{(m_1, k_1; \dots, m_J, k_J): 0 \leq m_j + k_j \leq b_j, j \in \bar{J}} \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell + k_\ell - 1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \right\} \\ &\quad \cdot \left(\prod_{i=0}^{m_1 + k_1 + \dots + m_J + k_J - 1} h(i) \right) \cdot 1_{(k_J > 0)} \cdot \lambda_J \cdot H(b_1, \dots, b_J; \{1, \dots, J\})^{-1}. \quad (4.21) \end{aligned}$$

Using the extended notion for the routing probabilities from Definition 4.1 we have

$$\begin{aligned} &\prod_{i=0}^{m_J + k_J - 1} h_{J, b_J}(i) \cdot \prod_{i=0}^{m_1 + k_1 + \dots + m_J + k_J - 1} h_{b_1 + \dots + b_J}(i) \cdot 1_{(k_J > 0)} \\ &= b_J \cdot \prod_{i=0}^{m_J + (k_J - 1) - 1} h_{J, b_{J-1}}(i) \\ &\quad \cdot (b_1 + \dots + b_J) \cdot \prod_{i=0}^{m_1 + k_1 + \dots + m_J + (k_J - 1) - 1} h_{b_1 + \dots + b_{J-1}}(i) \cdot 1_{(k_J > 0)}. \end{aligned}$$

We insert this into the sum (and common factor) which is relevant for location J and obtain

$$\sum_{(m_J, k_J): 0 \leq m_J + k_J \leq b_J} \left\{ \left(\prod_{i=0}^{m_J + k_J - 1} h_J(i) \right) \left(\prod_{i=1}^{m_J} \frac{v}{\eta_J \cdot i} \right) \left(\frac{v}{\lambda_J} \right)^{k_J} \right\}$$

$$\begin{aligned}
& \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_J+k_J-1} h(i) \right) \cdot 1_{(k_J>0)} \cdot \lambda_J \\
& = \sum_{(m_J, k_J): 0 \leq m_J+k_J \leq b_J-1} \left\{ \left(\prod_{i=0}^{m_J+k_J-1} h_J(i) \right) \left(\prod_{i=1}^{m_J} \frac{v}{\eta_J \cdot i} \right) \left(\frac{v}{\lambda_J} \right)^{k_J} \right\} \\
& \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_J+k_J-1} h(i) \right) \cdot \frac{b_J}{b_1 + \dots + b_J} \cdot v,
\end{aligned}$$

where we reintroduced the shorthand notation $h_{J, b_J-1}(\cdot) \rightarrow h_J(\cdot)$ and $h_{b-1}(\cdot) \rightarrow h(\cdot)$. This finishes the proof. \square

TH is not the total throughput (= overall expected number of departures in the network per time unit) of the closed system consisting of the replenishment server, the transportation and inventory nodes with stationary distribution θ . Nevertheless, the TH_j resemble the expressions of local throughputs at specified nodes in a Gordon-Newell network with these exponential nodes.

The state dependent routing probabilities from Definition 4.1 determine an adaptive routing control for sending out produced items from the replenishment server.

$r_{J+1,(\cdot)}(m_1, k_1; \dots m_J, k_J)$ selects the destination for a finished item according to the free stock capacity, including items on transport. Conditioned on state $(m_1, k_1, n_1; \dots m_J, k_J, n_J) \in E$ the probability for sending an item to location j is proportional to $(b_j - (m_j + k_j))_+$. Being interested in global performance of the queueing-inventory-location system it is natural to ask whether it is possible to compute the steady state expected routing probabilities. By ergodicity these expectations would be the long time averages of the routing process as well. We determine the “expected virtual routing probabilities” in the sense of distinguishing virtual waiting times from actual waiting times in a queueing system: We compute in steady state the probability that at time $t \geq 0$ an item would be send out to location j if a departure from node $J+1$ would occur at time t , $j = 1, \dots, J$, see e.g. Wolff (1989)[Section 5.13].

Proposition 4.7 *The expected routing probability to location j , $j = 1, \dots, J$, is*

$$\tilde{r}(J+1, j) = \frac{b_j}{b_1 + \dots + b_J} \frac{H(b_1, \dots, b_j - 1, \dots, b_J, \{1, \dots, J\})}{H(b_1, \dots, b_j, \dots, b_J, \{1, \dots, J\})} \quad (4.22)$$

Proof We demonstrate the case $j \rightarrow J$ and abbreviate again $d_\ell(x) =: \eta_\ell^{-1}$. Using separability at (\star) below we obtain

$$\begin{aligned}
\tilde{r}(J+1, J) &= \sum_{(m_1, k_1, n_1; \dots m_J, k_J, n_J): 0 \leq m_J+k_J \leq b_J, j \in \bar{J}} \pi(m_1, k_1, n_1; \dots m_J, k_J, n_J) \\
&\cdot r_{J+1, J}(m_1, k_1; \dots m_J, k_J) \\
&\stackrel{(\star)}{=} \sum_{(m_1, k_1; \dots m_J, k_J): 0 \leq m_J+k_J \leq b_J, j \in \bar{J}} \theta(m_1, k_1; \dots m_J, k_J) \\
&\cdot r_{J+1, J}(m_1, k_1; \dots m_J, k_J) \\
&= \sum_{(m_1, k_1; \dots m_J, k_J): 0 \leq m_J+k_J \leq b_J, j \in \bar{J}} \prod_{\ell=1}^J \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{v}{\eta_\ell \cdot i} \right) \left(\frac{v}{\lambda_\ell} \right)^{k_\ell} \right\}
\end{aligned}$$

$$\begin{aligned}
& \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_J+k_J-1} h(i) \right) H(b_1, \dots, b_J; \{1, \dots, J\})^{-1} \\
& \cdot h_{J,b_J}(m_J + k_J) \cdot h_{b_1+\dots+b_J}(m_1 + k_1 + \dots + m_J + k_J) \\
& = \sum_{(m_1, k_1; \dots, m_{J-1}, k_{J-1}): 0 \leq m_J + k_J \leq b_J} \prod_{\ell=1}^{J-1} \left\{ \left(\prod_{i=0}^{m_\ell+k_\ell-1} h_\ell(i) \right) \left(\prod_{i=1}^{m_\ell} \frac{\nu}{\eta_\ell \cdot i} \right) \left(\frac{\nu}{\lambda_J} \right)^{k_J} \right\} \\
& \cdot \sum_{(m_J, k_J): 0 \leq m_J + k_J \leq b_J} \left\{ \left(\prod_{i=0}^{m_J+k_J-1} h_J(i) \right) \left(\prod_{i=1}^{m_J} \frac{\nu}{\eta_J \cdot i} \right) \left(\frac{\nu}{\lambda_J} \right)^{k_J} \right\} \\
& \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+m_J+k_J-1} h(i) \right) H(b_1, \dots, b_J; \{1, \dots, J\})^{-1} \\
& \cdot h_{J,b_J}(m_J + k_J) \cdot h_{b_1+\dots+b_J}(m_1 + k_1 + \dots + m_J + k_J)
\end{aligned}$$

For the terms referring to J we obtain for fixed $(m_1, k_1; \dots, m_{J-1}, k_{J-1})$ (using the extended notation for the routing probabilities and noticing $h_{J,b_J}(b_J) = 0$)

$$\begin{aligned}
& \sum_{g=0}^{b_J-1} \sum_{m_J+k_J=g} \left(\prod_{i=0}^{g-1} h_{J,b_J}(i) \right) \left\{ \left(\prod_{i=1}^{m_J} \frac{\nu}{\eta_J \cdot i} \right) \left(\frac{\nu}{\lambda_J} \right)^{k_J} \right\} \\
& \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+g-1} h_{b_1+\dots+b_J}(i) \right) H(b_1, \dots, b_J; \{1, \dots, J\})^{-1} \\
& \cdot h_{J,b_J}(g) \cdot h_{b_1+\dots+b_J}(m_1 + k_1 + \dots + g) \\
& = \sum_{g=0}^{b_J-1} \left(\prod_{i=0}^g h_{J,b_J}(i) \right) \sum_{m_J+k_J=g} \left\{ \left(\prod_{i=1}^{m_J} \frac{\nu}{\eta_J \cdot i} \right) \left(\frac{\nu}{\lambda_J} \right)^{k_J} \right\} \\
& \cdot \left(\prod_{i=0}^{m_1+k_1+\dots+g} h_{b_1+\dots+b_J}(i) \right) H(b_1, \dots, b_J; \{1, \dots, J\})^{-1}
\end{aligned}$$

Now

$$\left(\prod_{i=0}^g h_{J,b_J}(i) \right) = b_J \left(\prod_{i=0}^{g-1} h_{J,b_J-1}(i) \right)$$

and

$$\begin{aligned}
& \left(\prod_{i=0}^{m_1+k_1+\dots+m_{J-1}+k_{J-1}+g} h_{b_1+\dots+b_J}(i) \right) \\
& = (b_1 + \dots + b_J)^{-1} \left(\prod_{i=0}^{m_1+k_1+\dots+m_{J-1}+k_{J-1}+g-1} h_{b_1+\dots+(b_J-1)}(i) \right)
\end{aligned}$$

finishes the proof. \square

A bit surprising with the result of Proposition 4.7 is that $\tilde{r}(J+1, j)$ is proportional to TH_j . From the hindsight, a little reflection indicates that this should be the case.

Remark 4.8 The state-dependent routing protocol of Definition 4.1 due to Towsley (1980) has found successors in the literature which often requires additional structural properties of the networks, e.g. local balance properties or quasi-reversibility of the nodes considered in isolation, for reviews see e.g. Huisman and Boucherie (2011) or Balsamo (2000).

Although the stationary distribution (4.8) has an appealing explicit form it usually poses difficulties for computational evaluation of, e.g., throughputs. Similarly to standard closed queueing networks computational algorithms can be developed, see Sauer (1983) for a Convolution Algorithm to compute normalization constants and Mean Value Analysis for determining more detailed information on expected queue lengths, expected passage times, etc. Explicit formulas are provided there for branches of length 1, i.e. consisting of a single queue.

4.2 Cost analysis, revenue

The explicit form of the stationary distribution (4.8) allows to compute cost for running the system over a long time horizon, respectively in stationary state per time unit, under the prescribed system layout. Because revenue is obtained from successfully served demand we have immediately from Theorem 4.6 the

Proposition 4.9 *Let $e_j > 0$ be the profit obtained from serving one unit of demand at location j . Then the total expected revenue R per time unit is*

$$R = \sum_{i=1}^J TH_j \cdot e_j.$$

The costs for running the system include (assumed all to be > 0)

- $w(j)$: waiting cost for a customer (demand) at the server of location j (waiting or in service),
- $i(j)$: holding cost for an item on stock in the inventory at location j ,
- $t(j)$: transport cost for an item (on a truck) on the way from replenishment server $J+1$ to location j ,
- $s(j)$: shortage cost for each demand rejected and lost at location j ,
- $c(j)$: capacity cost per time unit for providing the inventory at location j (e.g. rent, insurance),
- $w(J+1)$: waiting cost for an order at the replenishment server at location $J+1$ (waiting or in service).

The cost function is

$$\begin{aligned} f(b_1, \dots, b_J)(m_1, k_1, n_1; \dots, m_J, k_J, n_J) \\ := \sum_{j=1}^J (b_j \cdot c(j) + w(j) \cdot n_j + t(j) \cdot m_j + i(j) \cdot k_j + 1_{(k_j=0)} \lambda_j \cdot s(j)) \\ + w(J+1) \left(b - \sum_{j=1}^J (m_j + k_j) \right) \end{aligned}$$

The time average costs over long time horizon are (asymptotically) determined by ergodicity of Z .

Proposition 4.10 Denote by $(V_1, Y_1, X_1; \dots, V_J, Y_J, X_J)$ a vector which is distributed according to the stationary distribution π of the queueing-inventory-location system. The stationary costs per time unit of the system are

$$\begin{aligned} & \mathbb{E}\left(f_{(b_1, \dots, b_J)}(V_1, Y_1, X_1; \dots, V_J, Y_J, X_J)\right) \\ &= \sum_{j=1}^J b_j(w(J+1) + c(j)) + \sum_{j=1}^J w(j) \mathbb{E}_{\pi_j}(n_j) \\ &+ \sum_{j=1}^J (t(j) - w(J+1)) \mathbb{E}_{\theta}(m_j) \\ &+ \sum_{j=1}^J (i(j) - w(J+1)) \mathbb{E}_{\theta}(k_j) + \sum_{j=1}^J s(j) (\lambda_j - TH_j). \end{aligned} \quad (4.23)$$

The expectations $\mathbb{E}_{\pi_j}(n_j)$ can be directly evaluated, while the $\mathbb{E}_{\theta}(m_j)$ and $\mathbb{E}_{\theta}(k_j)$ can be computed using mean value analysis, and $\lambda_j - TH_j$ can be computed using the convolution algorithm following Sauer (1983)[Section 3].

Proof The expectation (4.23) is

$$\begin{aligned} & \sum_{j=1}^J b_j(w(J+1) + c(j)) + \sum_{j=1}^J w(j) \cdot \mathbb{E}(X_j) \\ &+ \mathbb{E} \left(\sum_{j=1}^J \left(V_j(t(j) - w(J+1)) + Y_j(i(j) - w(J+1)) + 1_{(Y_j=0)} s(j) \cdot \lambda_j \right) \right) \\ &\stackrel{(\star)}{=} \sum_{j=1}^J b_j(w(J+1) + c(j)) + \sum_{j=1}^J w(j) \cdot \mathbb{E}_{\pi_j}(n_j) \\ &+ \mathbb{E}_{\theta} \left(\sum_{j=1}^J \left(m_j(t(j) - w(J+1)) + k_j(i(j) - w(J+1)) + 1_{(k_j=0)} s(j) \cdot \lambda_j \right) \right), \end{aligned}$$

where (\star) follows from separability of the system. To transform the last expression $\sum_{j=1}^J s(j) \lambda_j \mathbb{E}_{\theta}(1_{(k_j=0)})$ we note that (4.21) in the proof of Theorem 4.6 can for any j be expressed as $TH_j = \mathbb{E}(1_{[Y_j>0]}) \cdot \lambda_j$, which yields $P(Y_j = 0) = 1 - \frac{TH_j}{\lambda_j}$. \square

4.3 Generalizations and modifications

We have presented our composite model in a simple form which to a certain extend included assumptions which oversimplify the system with the aim to provide easier access to our results. In this section we first sketch some easy generalizations. Thereafter we discuss a class of related problems which seemingly need much more effort to approach explicit solutions.

Replenishment server The single exponential server with state-independent service rate ν can be substituted

- by a node with general service discipline Daduna (2001)[Definition 9.1] with exponential service time request or a node with symmetric service discipline Daduna (2001)[Definition 9.5] according to Kelly (1979)[Section 3.3], or
- by a replenishment network as in Otten et al. (2020). A necessary restriction is that departure from the replenishment network into the transportation subsystem is from a unique dedicated “departure node”.

The proofs of both of these generalizations follow from Towsley (1980).

Modeling transport times For simplicity of presentation we assumed that for given distance $d_j(x) = d(x, a_j)$ the mean travel time (with travelling speed 1) is $d_j(x)$, and we applied an exponential- $d_j^{-1}(x)$ distribution for the service time at the infinite server which models travelling the lane. As a result of the so-called “insensitivity property of infinite servers” we can apply other travel time distributions with the same mean and obtain an explicit stationary distribution for a “supplemented Markov process” with a state space where the infinite server components (here the $Y_j(t)$) carry as additional information the residual service time (travel time) or the obtained service time as supplementary variable, for more details see Towsley (1980) and Chandy et al. (1977). The most relevant consequence for our project is obvious: We are in a position to apply our findings to more realistic situations, especially to non-deterministic travel times with means $d_j(x)$ and a small variance which allows some overtaking.

Control policies for inventories The inventory control policies for the inventories in the queueing-inventory systems in this article are base-stock policies. These policies have been investigated in many previous research papers related to the subject of our investigation. Other inventory control policies of interest are of type (r, Q) and (r, S) . We sketch here only research on queueing-inventory systems which provided as an outcome the stationary distribution in explicit expressions. Research which is relevant for networks of queueing-inventory systems with a common replenishment system and base-stock control of the inventories is communicated in Otten (2017), Otten et al. (2016, 2020).

Employing the popular (r, Q) and (r, S) policies seemingly poses severe difficulties in case of networks of queueing-inventory systems with a common replenishment system. In Otten (2017)[Section 7] for different classes of networks of queueing-inventory systems with stochastic non-zero lead times the local inventory control is by location specific (r_j, S_j) policies which means that whenever at location j the stock size decreases to the reorder point r_j a replenishment order is sent out. When the replenishment arrives at location j the inventory is filled up to maximal stock size S_j . For special network structures (with transportation times $= 0$) for $r_j \in \{0, 1\}$ an explicit product form of the stationary distribution is obtained. Seemingly all other research articles which provided explicit expressions of the stationary distributions under (r, Q) and (r, S) policies are on single queueing-inventory systems. Examples are Krishnamoorthy and Narayanan (2013), Baek and Moon (2014), Saffari et al. (2013), Schwarz et al. (2006). More details are provided in the survey Krishnamoorthy et al. (2021) [Section 5.2.1.1].

Results aside from these streams of research are provided in Schwarz et al. (2007). In a classical Jackson network, considered as a Flexible Manufacturing System (FMS), to some servers an inventory is attached and manufacturing at this server needs exactly one item from the associated inventory. The inventory control is by either (r_j, Q_j) or (r_j, S_j) policy with node-specific control limits.

5 Location analysis

The overall productivity of a system is usually measured by its throughput, i.e. the steady state mean number of produced items per time unit. Similarly, for the queueing-inventory-location system the most important single performance measure is the satisfied overall long time external demand (\equiv total production over all locations). This is for systems running under stability conditions the expected number of produced units (demanded from the exterior), i.e. the overall throughput as defined in (4.19) in Sect. 4.1. Theorem 4.6 enables us to compute this directly for a given system layout as dealt with in Theorem 4.4, i.e. with fixed locations, base stock levels, and local production functions (the $\mu_j(\cdot)$) and an adaptive routing function $r_{J+1,j}(\cdot)$.

In this section we discuss the possibility to find for given locations $a_j = (a_{j1}, a_{j2})$, $j = 1, \dots, J$, a location $x = (x_1, x_2)$ for the replenishment center such that throughput is maximized. We therefore indicate in the overall throughput $TH(x)$ the location $x = (x_1, x_2)$ as decision variable.

$$TH(x) := \mathbf{E}_{(x)} \left(\sum_{j=1}^J \mu_j(X_j) \cdot 1_{(X_j > 0, Y_j > 0)} \right), \quad (5.1)$$

where $\mathbf{E}_{(x)}(\cdot)$ refers to the expectation under process distribution determined by the central location placed at x . We write $TH_j(x)$ for local throughput of the server at location j with center located at x .

Typically this optimization problem is considered as a strategic decision problem. The classical problem is to determine for given a_j and some distance measure $d(\cdot, \cdot)$ a location x such that for certain problem specific weights $v_j \geq 0$, $j = 1, \dots, J$, the total weighted distance

$$\sum_{j=1}^J v_j \cdot d(x, a_j) \quad (5.2)$$

is minimized. This is the well known “Generalized Weber Problem”. To find suitable weights v_j usually is a non trivial problem. Information on the Weber problem can be found in Wesolowsky (1993) or Drezner et al. (2004).

The problem we are faced with is that we are not only interested in minimizing (weighted total) distances, but in maximizing the overall satisfied demand which is a global revenue. Additionally, the (strategic) decision for an optimal location has to pay attention for long-time development of the system. Especially one needs a prediction of the future demand. Consequently, we start our investigation with setting in force the

Assumption 5.1 The long time expected demand $\lambda_j > 0$ per time unit at the locations $j = 1, \dots, J$, is (approximately) known.

This assumption is reasonable if the amount of satisfied demand is a decision criterion. This assumption imposes a necessary condition on the system’s specification if we are interested in running a system which stabilizes over time. The proof is a direct consequence of Theorem 4.4.

Corollary 5.2 For stabilizing the queueing-inventory system the service rates $\mu_j(\cdot)$ (\equiv capacity functions) of the production servers at the locations must be (asymptotically) larger than

the arrival rates λ_j , $j = 1, \dots, J$, in the sense that

$$\sum_{n=1}^{\infty} \prod_{i=1}^n \frac{\lambda_j}{\mu_j(i)} < \infty \quad (5.3)$$

holds for all $j = 1, \dots, J$. For state independent service rates μ_j this means $\lambda_j < \mu_j$.

Consequently, we set in force

Assumption 5.3 The service rates can (and will) be selected in a way that (5.3) is fulfilled for all $j = 1, \dots, J$. (But the $\mu_j(\cdot)$ need not be fixed in advance according to the next corollary.)

A somewhat surprising but important consequence for the global optimization problem to find the location for the center is the following corollary, the proof of which is direct from Theorem 4.6 under Assumption 5.3.

Corollary 5.4 The production rates $\mu_j(\cdot)$, $j = 1, \dots, J$, at the locations are not relevant for placing the central location such that the overall throughput $TH(x)$ is maximized.

So, we have already proved, although in a stylized (but rather detailed) model, that the strategic decision for location of the center and the tactical decision for installing service capacities at the locations can indeed be separated. This is different from other integrated problem settings, e.g. with locating depots and transportation decisions in Salhi and Rand (1989).

Nevertheless, the rates $\mu_j(\cdot)$, $j = 1, \dots, J$, will reoccur below.

The main problem is now to maximize with respect to location x in the plane

$$\sum_{j=1}^J TH_j(x) = \sum_{j=1}^J v \cdot \frac{b_j}{b_1 + \dots + b_J} \frac{H(b_1, \dots, b_j - 1, \dots, b_J, \{1, \dots, J\})}{H(b_1, \dots, b_j, \dots, b_J, \{1, \dots, J\})}$$

with $H(b_1, \dots, b_j, \dots, b_J, \{1, \dots, J\})$ given in Theorem 4.4.

5.1 The reduced model

We did not succeed in solving this maximization problem for the detailed queueing-inventory-location model investigated so far. The main difficulties are to incorporate the specified base stock levels b_j and the adaptive routing for departures from the replenishment server $J + 1$ which evaluates online the local total stock sizes (onhand inventory + items on transport). We therefore reduce the complexity of the model to obtain tractable processes. The main components of the reduced model are

- the central replenishment server where items are produced in a make-to-order regime,
- transportation of items from the center to the locations,
- the locations with inventories where items are delivered with a rate that enables the server to meet with the location's throughput the external demand of the original model (in the reduced model), and
- reorders which are send without time delay from the inventories to the replenishment servers when an item is delivered, and
- an overall limit on the total stock capacity.

The restrictions imposed by the local base-stock levels are weakened. Instead of prescribing local restrictions b_j , we fix the total number of "orders send out + items on transport

+ items as onhand-inventory” $= b_1 + \dots + b_J =: b$. Because we assumed for the local base-stock levels that $b_j \geq 1$, $j = 1, \dots, J$, holds, $b \geq J$ holds for the original system. For easier reading we allow in this section $b \geq 1$. In the final Algorithm 5.16 for determining the base-stock levels the restrictions $b_j \geq 1$ will be re-invented.

This reduced model is a closed starlike network of Gordon-Newell type with $2 \cdot J + 1$ nodes and b customers cycling (having different interpretation at the different nodes of the network). The central node is the exponential- ν replenishment single server and there are J branches numbered $1, \dots, J$. Each branch consists of a tandem with (i) an exponential- $d_j(x)^{-1}$ infinite server (which represents transport), and (ii) a location with state dependent exponential- $\mu_j(\cdot)$ single server, $j = 1, \dots, J$. An item which is served (delivered at a location) departs immediately from that location and is sent to the center, occurring there as an order.

The most difficult problem is to select suitably the routing probabilities for items on leave from the central server. Recall that the predicted loads for the locations are λ_j , $j = 1, \dots, J$ (Assumption 5.1).

A simple policy is to distribute the total amount of items produced by the replenishment server (\equiv the network’s total throughput!) proportional to these predicted loads. Therefore the local throughputs should be proportional to the local demands, which implies fair sharing of the available capacity according to the locations’ demand. Selection of this dispatching policy is supported by the following observation: From Proposition 4.7 we know that in the detailed model from Sect. 4 the expected routing probabilities are proportional to the local throughputs.

This suggests to apply in the reduced model a routing scheme $\hat{R} = (\hat{r}(J+1, j), j = 1, \dots, J)$ according to

$$\hat{r}(J+1, j) = \frac{\lambda_j}{\lambda_1 + \dots + \lambda_J}, \quad j = 1, \dots, J. \quad (5.4)$$

Summarizing this construction, we describe the evolution over time of this reduced model by

Definition 5.5 Let for $j = 1, \dots, J$, denote $\hat{V}_j(t)$ the number of items on the way to location j at time $t \geq 0$, $\hat{Y}_j(t)$ the size of the inventory at j at time $t \geq 0$, and $\hat{Y}_{J+1}(t)$ the number of orders waiting or in service at the central replenishment server $J+1$ at time $t \geq 0$. The network process $\hat{Z} = (\hat{Z}(t) : t \geq 0)$ with $\hat{Z}(t) = (\hat{Y}_{J+1}(t), \hat{V}_1(t), \hat{Y}_1(t), \dots, \hat{V}_J(t), \hat{Y}_J(t))$ is an ergodic Markov process with state space

$$\hat{E} := \{(k_{J+1}; m_1, k_1; \dots, m_J, k_J) \in \mathbb{N}_0^{2J+1} : k_{J+1} + \sum_{j=1}^J (m_j + k_j) = b\}.$$

The intensity matrix (infinitesimal generator) $\hat{Q} = (\hat{q}(z, z') : z, z' \in \hat{E})$ of \hat{Z} has the following positive rates. For readability we only depict the relevant local states for the respective transitions, e.g. we write $(k_{J+1}; \dots, m_j, k_j; \dots)$ instead of $(k_{J+1}; m_1, k_1; \dots, m_j, k_j; \dots, m_J, k_J)$. For $j = 1, \dots, J$ we have

$$\begin{aligned} \hat{q}((k_{J+1}; \dots, m_j, k_j; \dots), (k_{J+1} - 1; \dots, m_j + 1, k_j; \dots)) &= 1_{(k_{J+1} > 0)} \cdot \nu \cdot \hat{r}(J+1, j) \\ \hat{q}((k_{J+1}; \dots, m_j, k_j; \dots), (k_{J+1}; \dots, m_j - 1, k_j + 1; \dots)) &= 1_{(m_j > 0)} \cdot d_j(x)^{-1} \cdot m_j \\ \hat{q}((k_{J+1}; \dots, m_j, k_j; \dots), (k_{J+1} + 1; \dots, m_j, k_j - 1; \dots)) &= 1_{(k_j > 0)} \cdot \mu_j(k_j). \end{aligned}$$

The diagonal elements $\widehat{q}(z, z)$ of $\widehat{\mathbf{Q}}$ are selected in a way that row sums are zero. All other rates $\widehat{q}(z, z'), z \neq z'$, are zero.

The intensity matrix $\widehat{\mathbf{Q}}$ determines the steady state distribution of the Gordon-Newell network process $\widehat{\mathbf{Z}}$ and the asymptotic throughput (respectively the steady state throughput) of the network. We summarize without proofs the facts necessary to evaluate the performance of the reduced model. These are standard results in Gordon-Newell network theory, see e.g. Daduna (2001).

Proposition 5.6 *The Markov process $\widehat{\mathbf{Z}}$ is ergodic. If the center $J + 1$ is located at $x = (x_1, x_2)$, the stationary distribution $\psi = \psi(b; x)$ is with $\widehat{G}(b, \{1, \dots, J, J + 1\}; x)$ as normalization constant*

$$\begin{aligned} & \psi(k_{J+1}; m_1, k_1; \dots m_J, k_J) \\ &= \left(\frac{\alpha_{J+1}}{\nu} \right)^{k_{J+1}} \cdot \prod_{j=1}^J \left\{ (\alpha_j \cdot d_j(x))^{m_j} \frac{1}{m_j!} \prod_{i=1}^{k_j} \frac{\alpha_j}{\mu_j(i)} \right\} \\ & \cdot \widehat{G}(b, \{1, \dots, J, J + 1\}; x)^{-1}, \\ & \text{for } (k_{J+1}; m_1, k_1; \dots m_J, k_J) \in \widehat{E}. \end{aligned} \quad (5.5)$$

Here $\alpha_{J+1} = 1/3$ and $\alpha_j = \lambda_j / (3(\lambda_1 + \dots + \lambda_J))$, $j = 1, \dots, J$, are obtained as the probability solution of the standard traffic equation. Note that α_j occurs twice in (5.5) for each $j = 1, \dots, J$.

The network's total throughput, i.e. the overall departure rate from all nodes is

$$TH(b; x) = \frac{\widehat{G}(b - 1, \{1, \dots, J, J + 1\}; x)}{\widehat{G}(b, \{1, \dots, J, J + 1\}; x)}. \quad (5.6)$$

The throughput from location j , i.e. the departure rate from the single server node at location j is

$$TH_{loc,j}(b; x) = \alpha_j \cdot TH(b; x), \quad j = 1, \dots, J. \quad (5.7)$$

The throughput from all locations, i.e. the sum of the departure rates from all single server nodes at the locations $1, \dots, J$ is

$$TH_{loc}(b; x) = TH(b; x) \sum_{j=1}^J \alpha_j = \frac{1}{3} TH(b; x). \quad (5.8)$$

Before proceeding, a short discussion is necessary. A first important observation about the predicted demand rates λ_j occurring implicitly in (5.5) and (5.6) is that within the factors $(\alpha_j \cdot d_j(x))$ and $(\alpha_j / \mu_j(\cdot))$ of the stationary distribution only the relative demands $\lambda_j / (\lambda_1 + \dots + \lambda_J)$ occur and that $(\alpha_{J+1} / \nu) = 1 / (3 \cdot \nu)$ is independent of demands.

The second observation is that in taking instead of (b_1, \dots, b_J) the sum $b := b_1 + \dots + b_J$, we have “averaged” the base stock levels.

This leads to the third observation that the stationary distribution (5.5) and the throughput (5.8) depend on the decision variable $x \in \mathbb{R}^2$ for locating the center and on the potential decision variable for providing the overall inventory capacity b . Both variables will turn out to be of fundamental importance for optimal responding to the total predicted demands because

- x will be the result of a strategic decision for locating the replenishment center,

- b will be a result of tactical decisions if the sum of the demands $\lambda_1 + \dots + \lambda_J$ are given.

The main problem is that these decision variables are intertwined via (5.6) and optimization of $TH(b; x)$ has at a first glance to be concurrently in b and x according to

Optimization Problem 5.7 *Determine*

$$\min_{b \in \mathbb{N}_+} \left(\max_{x \in \mathbb{R}^2} \{TH_{loc}(b; x)\} \right) \quad \text{and} \quad \arg \left\langle \min_{b \in \mathbb{N}_+} \left(\max_{x \in \mathbb{R}^2} \{TH_{loc}(b; x)\} \right) \right\rangle$$

$$\text{subject to } TH_{loc,j}(b; x) \geq \lambda_j, j = 1, \dots, J.$$

Note that the side constraints guarantee that the total produced and sent-out items are delivered in such a way that the local demands are satisfied. It will turn out that similar to the solution procedure in Lange and Daduna (2022) the main effort in solving Optimization Problem 5.7 is to solve the sequence of maximization sub-problems

Optimization Problem 5.8 *Determine for each total inventory size $b \geq 1$*

$$\max_{x \in \mathbb{R}^2} \{TH_{loc}(b; x)\} \quad \text{and} \quad \arg \left\langle \max_{x \in \mathbb{R}^2} \{TH_{loc}(b; x)\} \right\rangle.$$

If the manufacturing capacity ν of the replenishment server $J + 1$ is too small compared with the demand that has to be delivered there might be no solution of the Optimization Problem 5.7 because there emerge bottlenecks in the network. Nevertheless, all optimization sub-problems in Optimization Problem 5.8 have solutions. We assume (if necessary) that the capacities guarantee that a feasible solution of the Optimization Problem 5.7 exists, i.e. the side constraints in the Optimization Problem 5.7 can be satisfied.

The astonishing fact is that the solutions (in fact there will be only one common solution for all b) to all sub-problems of Optimization Problem 5.8 satisfy automatically the side constraints in Optimization Problem 5.7, whenever the ν and the $\mu_j(\cdot)$ guarantee the existence of a feasible solution.

The Optimization Problems 5.7 and 5.8 will be solved in a way similar to the location problem in “logistics and services networks under congestion” in Lange and Daduna (2022). It turns out that the present problem can be attacked almost in the same way. We therefore will only sketch proofs and refer for details to Lange and Daduna (2022). In our solution procedure we will exploit the standard “Weber Problem”. For an introduction with general remarks on extensions and on the history of the problem see Wesolowsky (1993) and Drezner et al. (2004). The next theorem is a modification of Theorem 4.4 in Lange and Daduna (2022) [Section 4.2].

Theorem 5.9 *Consider locations $a_j = (a_{j1}, a_{j2}) \in \mathbb{R}^2, j = 1, \dots, J$, in the plane and associated weights $v_j := \lambda_j / (\lambda_1 + \dots + \lambda_J)$. Let $x^* \in \mathbb{R}^2$ be a solution of the standard Weber problem with weighted distances:*

$$\text{Find } \min_{x \in \mathbb{R}^2} \left\{ \sum_{j=1}^J v_j d_j(x) \right\} \quad \text{and} \quad x^* = \arg \left\langle \min_{x \in \mathbb{R}^2} \left\{ \sum_{j=1}^J v_j d_j(x) \right\} \right\rangle. \quad (5.9)$$

Then x^ is a solution of the Optimization Problem 5.8 for any $b \geq 1$ as well.*

The proof relies on the observation that the reduced model for the queueing-inventory-location network can be described in terms of a closed queueing network of Gordon-Newell type as sketched above. The proof of the next theorem will be given implicitly by proving correctness of Algorithm 5.13 below.

Theorem 5.10 *If a solution of the Optimization Problem 5.7 exists for the capacity v , the service intensities $\mu_j(\cdot)$, and the demands λ_j , $j = 1, \dots, J$, the solution for the optimal value of $b \geq 1$ is uniquely determined if x^* is given.*

The results of Theorems 5.9 and 5.10 are striking and comments are necessary. (i) If the Optimization Problem 5.7 has a solution, i.e., the side constraints are satisfied with capacities v and $\mu_j(\cdot)$, $j = 1, \dots, J$, these capacities do not matter for optimizing the overall throughput with respect to the location of the center. Similarly, the number b of total inventory is not relevant for finding the optimal location x^* . The relevant information for the location decision only comprises for $j = 1, \dots, J$,

- distances $d_j(x) := d(a_j, x)$, which determine travel times, and
- proportions $v_j = \lambda_j / (\lambda_1 + \dots + \lambda_J)$ of items to be dispatched to location j .

(ii) It is intuitive that increasing the loading capacity at the center increases the throughput at any location. However, the following is less intuitive: Fix the capacity at the replenishment center and the service rates at **all but one dedicated location** and increase the service rate at the **dedicated location**, then the throughput at **all** locations increases. Both facts are consequences of Theorem 14.B.13 of Shaked and Shanthikumar (1994).

Theorem 5.9 guarantees that under changes of capacities at the locations the decision for the optimal location has not to be renewed, as long as the solution of the Optimization Problem 5.7 exists.

(iii) Theorem 5.9 does not propose that throughput at the locations is independent of local properties. Details about functional dependencies will be given below. Moreover, it is not clear in advance whether a prescribed overall throughput can be met with sufficiently high total inventory capacity b by a given set of parameter values. If this is the case, Theorem 5.10 (implicitly in connection with the main result of van der Wal (1989)) guarantees that by successively adding inventory we can increase the throughput until the total requirements from the locations can be satisfied. Otherwise, if capacities do not suffice, bottlenecks occur. The relevant proofs show, that we can increase the throughput by increasing the speed of production v at the replenishment center (capacity). Theorem 5.9 states that in any case the selected position for the center remains optimal.

The next lemma indicates how the “Weber Problem” enters our solutions.

Lemma 5.11 *Denote with $\bar{J}^+ := \{1, \dots, J, J+1\}$ for $g = 0, \dots, b$,*

$$C_g(b, \bar{J}^+) := \sum_{k_{J+1}+k_1+\dots+k_J=b-g} \left(\frac{\alpha_{J+1}}{v} \right)^{k_{J+1}} \cdot \prod_{j=1}^J \left(\prod_{i=1}^{k_j} \frac{\alpha_j}{\mu_j(i)} \right). \quad (5.10)$$

Then for the normalization constant in (5.5) holds

$$\widehat{G}(b, \{1, \dots, J, J+1\}; x) = \sum_{g=0}^b \frac{\left(\sum_{i=1}^J \alpha_i \cdot d_i(x) \right)^g}{g!} \cdot C_g(b, \bar{J}^+). \quad (5.11)$$

Proof The proof is along the lines of the proof of Theorem 4.7 in Lange and Daduna (2022). We sketch some details to fix ideas.

$$\begin{aligned} & \widehat{G}(b, \{1, \dots, J, J+1\}; x) \\ &= \sum_{k_{J+1}+m_1+k_1+\dots+m_J+k_J=b} \prod_{j=1}^J \left(\frac{(\alpha_j \cdot d_j(x))^{m_j}}{m_j!} \right) \left(\frac{\alpha_{J+1}}{v} \right)^{k_{J+1}} \prod_{j=1}^J \left(\prod_{i=1}^{k_j} \frac{\alpha_j}{\mu_j(i)} \right) \end{aligned}$$

$$\begin{aligned}
&= \sum_{g=0}^b \underbrace{\left\{ \sum_{m_1+\dots+m_J=g} \prod_{j=1}^J \left(\frac{(\alpha_j \cdot d_j(x))^{m_j}}{m_j!} \right) \right\}}_{:= (\star)} \\
&\cdot \underbrace{\left\{ \sum_{k_{J+1}+k_1+\dots+k_J=b-g} \left(\frac{\alpha_{J+1}}{v} \right)^{k_{J+1}} \cdot \prod_{j=1}^J \left(\prod_{i=1}^{k_j} \frac{\alpha_j}{\mu_j(i)} \right) \right\}}_{=: C_g(b, \bar{J}^+)}.
\end{aligned}$$

We finish the proof by observing that for $g \in \{0, \dots, b\}$ holds

$$\begin{aligned}
\star &= \left(\sum_{m_1+\dots+m_J=g} \prod_{j=1}^J \left(\frac{\alpha_j \cdot d_j(x)}{\underbrace{\sum_{i=1}^J \alpha_i \cdot d_i(x)}_{\text{density of multinomial distribution}}} \right)^{m_j} \frac{g!}{m_j!} \right) \\
&\cdot \frac{(\sum_{i=1}^J \alpha_i \cdot d_i(x))^g}{g!}.
\end{aligned}$$

□

The representation (5.11) of $\widehat{G}(b, \{1, \dots, J, J+1\}; x)$ shows that it is the same normalization constant as that of a Gordon-Newell network with b customers, a single server with rate v , J stations with the service rates $\mu_j(\cdot)$, and a *single infinite server* with visit ratio $1/3$ and exponentially distributed service time with mean $(\sum_{j=1}^J (\lambda_j / (\lambda_1 + \dots + \lambda_J)) \cdot d_j(x))$.

The consequences of Lemma 5.11 are manifold. The main benefit is helping to prove Theorem 5.9. It separates the sum of the weighted travel times from the other ingredients of the network characteristics. This weighted sum is exactly what the Weber problem is about: For any solution x^* of the Weber problem (5.2) with $v_j = \lambda_j / (\lambda_1 + \dots + \lambda_J)$ we know that

$$\sum_{j=1}^J \frac{\lambda_j}{\lambda_1 + \dots + \lambda_J} \cdot d(a_j, x^*)$$

minimizes this weighted sum over $x \in \mathbb{R}^2$.

The proof of Theorem 5.9 follows the lines of the proof of Theorem 4.4 in Lange and Daduna (2022), which is rather technical and tedious. A short path to the main idea is as follows.

We have to investigate throughputs i.e. quotients of normalization constants. According to Lemma 5.11 these can be split into a term which can be interpreted as coming from a single infinite server node, and a second term which reminds normalization constants of a single server Gordon-Newell network. Realizing that throughput of a closed network with only one infinite server node increases with the speed of service (\equiv speed of transport which is maximized by the solution of the Weber problem) we have to show that the attached Gordon-Newell parts, represented by the $C_g(b, \bar{J}^+)$, do not disturb this isotonicity. A main technical ingredient is the monotonicity property of throughputs in Gordon-Newell networks with non decreasing service rates of van der Wal (1989) plus some combinatorial identities.

A second benefit of Lemma 5.11 is that it enables us to construct a rather simple algorithm to evaluate the optimized throughputs successively in the number $b \geq 1$ of customers (items) in the system. This enables us to find the minimum total size of needed inventory to realize the required throughputs.

Remark 5.12 It is possible that the location of the center coincides with one of the locations. If the center's location is, say $x = a_j$, this results in distance to this node $d_j(x) = 0$ and the service times at the infinite server (travel times) in branch j are zero. In this case we always have $m_j = 0$.

5.2 Determining needed inventory sizes

We demonstrate the power of Lemma 5.11 by showing how to determine efficiently (i) the total inventory needed to fulfill the overall demand, and (ii) the size of the local inventories. We assume that the center's location x and capacities $\mu_j(\cdot)$ and v are fixed and sufficiently high to satisfy demands λ_j eventually. We assumed up to now that trucks travel from replenishment station to the locations with unit speed (1 km/hour), which implies that $d_j(x)$ is exactly the mean time to travel distance $d_j(x)$. For the present demonstration we allow general speed $S > 0$ for trucks. The mean time for travelling distance $d_j(x)$ is then the mean service time $d_j(x)/S$ at the infinite server j , $j = 1, \dots, J$.

With notation from Proposition 5.6 and Lemma 5.11, we apply Buzen's Algorithm (Algorithm 7.1 in the Appendix) in a first step to a Gordon-Newell network consisting of stations $1, 2, \dots, J, J+1$.

Algorithm 5.13 [Determine the total inventory needed.] Let $S \in (0, \infty)$ denote the speed of the trucks and recall that $\bar{J}^+ := \{1, \dots, J, J+1\}$.

Initialization: STORE

$$G(0, \{1, \dots, j\}) := 1, j = 1, \dots, J+1, \quad \lambda := \sum_{j=1}^J \lambda_j, \quad \alpha_{J+1} = 1/3,$$

$$\alpha_j = \lambda_j / (3\lambda) \quad j = 1, \dots, J, \quad \kappa := \sum_{j=1}^J \alpha_j \cdot d_j(x) / S,$$

$$\hat{F}(1, \bar{J}^+; x) := 1.$$

SET $b \leftarrow 1$.

Iterate (*) FOR b DO

STORE $G(b, \{1\}) := g_1(b)$ FROM (7.1).

COMPUTE WITH $\mu_j(k)$ FROM ASSUMPTION 5.3, $j = 1, \dots, J$,

AND REPLENISHMENT RATE v

FROM (7.2) : $G(0, \bar{J}^+), G(1, \bar{J}^+), \dots, G(b, \bar{J}^+)$.

$$\text{COMPUTE } \hat{G}(b, \bar{J}^+; x) = \sum_{n=0}^b \left[G(b-n, \bar{J}^+) \cdot \frac{\kappa^n}{n!} \right].$$

IF $\frac{1}{3} \frac{\hat{F}(b, \bar{J}^+; x)}{\hat{G}(b, \bar{J}^+; x)} \geq \lambda$: IF $b < j$: THEN $b \leftarrow b + j$; OUTPUT b , STOP.

ELSE STORE

$\hat{F}(b+1, \bar{J}^+; x) \leftarrow \hat{G}(b, \bar{J}^+; x)$ AND SET $b \leftarrow b + 1$. GO TO (*)

OUTPUT is the minimal total inventory needed to guarantee

the required total stock size for satisfying demand $\lambda = \lambda_1 + \dots + \lambda_J$.

Proof Following the discussion before Definition 5.5, $\hat{G}(b, \bar{J}^+; x)$ from Proposition 5.6 can be interpreted as normalization constant in a Gordon-Newell network with nodes $1, \dots, J+1$ having visit ratios α_j and service rates $\mu_j(n_j)$ at stations $j = 1, \dots, J$, v at station $J+1$, and an additional infinite server node with visiting ratio $1/3$ and exponential- $(\sum_{j=1}^J \lambda_j / (\lambda_1 +$

$\cdots + \lambda_J) \cdot d_j(x))$ service time distribution. Because the $\mu_j(n_j)$, v and $n \cdot (\sum_{j=1}^J \lambda_j / (\lambda_1 + \cdots + \lambda_J) \cdot d_j(x))$ are nondecreasing in n_j and n , from van der Wal's theorem in van der Wal (1989) it follows that the throughput of this artificial Gordon-Newell network is nondecreasing in b . As can be seen from the proof in van der Wal (1989) the throughput is strictly increasing in b . This guarantees that the algorithm stops after a finite number of iterations, because we assumed that the capacities are high enough to satisfy all the demands eventually.

Taking $b \leftarrow b + J$ guarantees that minimal base stock level 1 can be guaranteed for all locations. \square

A direct consequence is that with Algorithm 5.13 we obtain automatically that the total throughput, which is $\geq \lambda$, is distributed to the locations as desired.

Corollary 5.14 *Assume that for the reduced model (see Proposition 5.6) the optimal location for the center at x^* is found and the necessary total inventory to satisfy the overall demand $\lambda = \lambda_1 + \cdots + \lambda_J$ is determined as b . Then the local demands for the individual locations are satisfied as well.*

Proof Algorithm 5.13 provided the number b of total inventory such that the sum of the departure rates from the locations (see (5.8)) fulfills $TH_{loc}(b; x^*) = (1/3)TH(b; x^*) \geq \lambda$. Then the throughput at location $j \in \{1, \dots, J\}$ fulfills $(1/3)\alpha_j \cdot TH(b; x^*) = \alpha_j \cdot TH_{loc}(b; x^*) \geq \alpha_j \cdot \lambda = \lambda_j$. \square

It remains to determine the distribution of the total inventory over the locations which is needed to re-transform the results from the reduced model to the model from Sect. 4. Because we did not restrict the number of items present at, say location j , in the reduced model all b items can “queue up” at j . The problem is to restrict the inventory places positioned at j in a way that in the long run (\equiv in steady state) the mean overflow at j is minimized for all $j \in \{1, \dots, J\}$. We need the following conditional probability of the vector of joint inventory positions. The proof is immediate from Proposition 5.6.

Lemma 5.15 *Let $\widehat{Z} = (\widehat{Y}_{J+1}, \widehat{V}_1, \widehat{Y}_1, \dots, \widehat{V}_J, \widehat{Y}_J)$ denote a vector which is distributed according to the stationary distribution $\psi = \psi(b; x^*)$ from Proposition 5.6 (see (5.5)) with b from Algorithm 5.13. The conditional distribution of $(\widehat{Y}_1, \dots, \widehat{Y}_J)$ given $(\widehat{Y}_{J+1} = 0, \widehat{V}_1 = 0, \dots, \widehat{V}_J = 0)$ is with $\beta_j = \lambda_j / (\lambda_1 + \cdots + \lambda_J)$*

$$P(\widehat{Y}_1 = k_1, \dots, \widehat{Y}_J = k_J | \widehat{Y}_{J+1} = 0, \widehat{V}_1 = 0, \dots, \widehat{V}_J = 0) \\ = \frac{\prod_{j=1}^J \left(\prod_{i=1}^{k_j} \frac{\beta_j}{\mu_j(i)} \right)}{\sum_{\ell_1 + \dots + \ell_J = b} \prod_{j=1}^J \left(\prod_{i=1}^{\ell_j} \frac{\beta_j}{\mu_j(i)} \right)} \quad \forall k_1, \dots, k_J \text{ such that } k_1 + \dots + k_J \leq b. \quad (5.12)$$

The result of the Lemma 5.15 is surprising because

$P(\widehat{Y}_1 = k_1, \dots, \widehat{Y}_J = k_J | \widehat{Y}_{J+1} = 0, \widehat{V}_1 = 0, \dots, \widehat{V}_J = 0)$ is independent of x , respectively x^* . This is a consequence of the product form stationary distribution. Therefore a tempting heuristic to distribute the total inventory is as follows.

Algorithm 5.16 [Determine the local inventories.] Let $b \geq J$ denote the output from Algorithm 5.13.

FOR $j \in \{1, \dots, J\}$ COMPUTE $\tilde{b}_j := \mathbb{E}(\widehat{Y}_j | \widehat{Y}_{J+1} = 0, \widehat{V}_1 = 0, \dots, \widehat{V}_J = 0)$.
DETERMINE $b_1, \dots, b_J \in \mathbb{N}_+$ SUCH THAT

$$\begin{aligned}
 &b_1 + \dots + b_J = b \\
 &\sum_{j=1}^J |b_j - \tilde{b}_j| = \min ! \\
 &b_j \geq 1, \quad \forall j = 1, \dots, J. \\
 &\text{OUTPUT } (b_1, \dots, b_J).
 \end{aligned}$$

Remark. The generalizations described in Sect. 4.3 apply to the reduced model directly as a consequence of insensitivity theory for queueing networks, see Daduna (2001)[Section 9].

6 Conclusion

We have developed a model for to integrate in a single Markovian system the features of manufacturing at various locations, inventory holding and control, replenishment production, transportation with dispatching according to system state dependent adaptive regimes. This enabled an optimization procedure with respect to an overall performance measure for the system. We developed two models, which together enabled us to explicitly compute the stationary distribution of the respective system in product form and to solve the integrated location problem. Possibly the most important benefit is to single out the decision for the location of the central manufacturing unit in terms of a generalized Weber problem.

According to the classification in Min et al. (1998) or Nagy and Salhi (2007) the problem tackled in this article is related to location-allocation problem because when the location of the replenishment center is fixed it is assumed that there are only radial trips from the center to the locations. The problem is related to location-routing problems because it deals with “location planning with tour planning aspects taken into account, see Nagy and Salhi (2007)[p.1]”, and because we have to decide about subsequent radial tours of the trucks to visit different locations by applying the congestion dependent scheduling regimes. The relation to transportation-location problems in the sense of Cooper (1972, 1976) is obvious because we consider both of these problems, especially that the center has limitations on its capacity to manufacture and ship the items Cooper (1972)[p. 94]. Addressing all these sub-problems, the investigation in this article is related to location-inventory-routing problems as considered in Song and Wu (2022). Our optimization criterion encompasses furthermore a time dimension as it is considered in Tapiero (1971) for the classical setting of Cooper. Aspects considered in the present article, which are not included in the above standard problem classes, are:

- The effect of limited manufacturing capacities at the locations and the central production facility. These limitations result in congestion, queueing problems, and blocking, respectively rejection of demand.
- The effect of including “time dimension of transportation-location-allocation problems” in the optimization criterion as discussed in Tapiero (1971)[p. 383]. In Tapiero (1971)[Section 5] questions concerning delivery time lags are sketched.

Acknowledgements I thank the managing editor and three reviewers for their critical reading of the paper and their suggestions which helped to improve the article.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence,

and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix: Buzen's Algorithm

We present the version of Buzen's algorithm (which is often termed "convolution algorithm") using notation which is directly applicable to prove the Algorithm for determining the minimal number of overall inventory. We have introduced the stationary distribution of the reduced model as the stationary distribution of a Gordon-Newell network with probability solution of the traffic equation. Buzen's algorithm, see Bruell and Balbo (1980)[Section 2.2.1], applies without assuming this restriction. Because our application is to compute the $C_g(b, \bar{J}^+)$ from Lemma 5.11 we have $\sum_{j=1}^{J+1} \alpha_j = 2/3$.

Algorithm 7.1 Buzen's Algorithm.

INPUT: $\alpha_j, j = 1, \dots, J, v, \mu_j(n_j), n_j \in \mathbb{N}_+, j = 1, \dots, J$.

FOR $j = 1, \dots, J, J + 1$ SET $g_j(0) := 1$,

FOR $j = 1, \dots, J$ SET $g_j(n_j) := \prod_{k=1}^{n_j} \frac{\alpha_j}{\mu_j(k)}, n_j \geq 1$.

FOR $j = J + 1$ SET $g_{J+1}(n_{J+1}) := \left(\frac{\alpha_{J+1}}{v}\right)^{n_{J+1}}$

SET BOUNDARY VALUES

$G(0, \{1, \dots, j\}) := 1, j = 1, \dots, J + 1, G(m, \{1\}) := g_1(m), m = 1, 2, \dots$

DENOTE FOR $j \geq 1, m \geq 1$,

$$G(m, \{1, \dots, j\}) := \sum_{n_1 + \dots + n_j = m} \prod_{j=1}^j g_j(n_j). \quad (7.1)$$

COMPUTE ITERATIVELY NORMING CONSTANTS FOR $j \geq 1$ AND $m \geq 1$:

$$G(m, \{1, \dots, j\}) = \sum_{\ell=0}^m G(\ell, \{1, \dots, j-1\}) \cdot g_j(m-\ell). \quad (7.2)$$

References

- Aboolian, R., Berman, O., & Drezner, Z. (2008). Location and allocation of service units on a congested network. *IIE Transactions*, 40, 422–433.
- Aboolian, R., Berman, O., & Drezner, Z. (2009). The multiple server center location problem. *Annals of Operations Research*, 167, 337–352.
- Albareda-Sambola, M., & Rodriguez-Pereira, J. (2019). Location-routing and location-arc routing. In G. Laporte, S. Nickel, & F. Saldanha da Gama (Eds.), *Facility location, chapter 15* (2nd ed., pp. 431–451). Cham: Springer.
- Baek, J. W., & Moon, S. K. (2014). The M/M/1 queue with a production-inventory system and lost sales. *Applied Mathematics and Computation*, 233, 534–54.

- Balsamo, S. (2000). Product form queueing networks. Lecture notes in computer science. In C. Lindemann, G. Haring, & M. Reiser (Eds.), *Performance evaluation white book* (Vol. 1769, pp. 377–401). New York: Springer.
- Berman, O., & Drezner, Z. (2007). The multiple server location problem. *Journal of the Operational Research Society*, 58(1), 91–99.
- Berman, O., & Kim, E. (1999). Stochastic models for inventory management at service facilities. *Communications in Statistics. Stochastic Models*, 15(4), 695–718.
- Berman, O., & Krass, D. (2004). Facility location problems with stochastic demands and congestion. In Z. Drezner & H. W. Hamacher (Eds.), *Facility location: Applications and theory, chapter 11* (1st ed., pp. 329–371). Berlin: Springer (**2. Printing**).
- Berman, O., & Krass, D. (2019). Stochastic location models with congestion. In G. Laporte, S. Nickel, & F. Saldanha da Gama (Eds.), *Facility location, chapter 17* (2nd ed., pp. 477–535). Cham: Springer.
- Berman, O., Larson, R. C., & Chiu, S. S. (1985). Optimal server allocation on a network operating as an M/G/1 queue. *Operations Research*, 33, 746–771.
- Berman, O., Larson, R. C., & Parkan, C. (1987). The stochastic queue p-median problem. *Transportation Sciences*, 21, 207–216.
- Berman, O., & Sapna, K. P. (2001). Optimal control of service for facilities holding inventory. *Computers & Operations Research*, 28, 429–441.
- Bruell, S. C., & Balbo, G. (1980). *Computational algorithms for closed queueing networks*. New York: North-Holland.
- Chandy, K. M., Howard, H., Jr., & Towsley, D. F. (1977). Product form and local balance inn queueing networks. *Journal of the Association for Computing Machinery*, 24(2), 250–263.
- Cooper, L. (1972). The transportation-location problem. *Operations Research*, 20, 94–108.
- Cooper, L. (1976). An efficient heuristic algorithm for the transportation-location problem. *Journal of Regional Sciences*, 16(3), 309–315.
- Cooper, R. B. (1990). Queueing theory. In D. P. Heyman & M. J. Sobel (Eds.), *Stochastic models, volume 2 of Handbooks in operations research and management science, chapter 10* (pp. 469–518). Amsterdam: North-Holland.
- Cordeau, J. F., Pasin, F., & Solomon, M. M. (2006). An integrated model for logistics network design. *Annals of Operations Research*, 144, 59–82.
- Daduna, H. (1985). The cycle time distribution in a central server network with state-dependent branching. *Optimization*, 16(4), 617–626.
- Daduna, H. (2001). Stochastic networks with product form equilibrium. In D. N. Shanbhag & C. R. Rao (Eds.), *Stochastic processes: Theory and methods, volume 19 of Handbook of statistics, chapter 11* (pp. 309–364). Amsterdam: Elsevier.
- Dan, T., & Marcotte, P. (2019). Competitive facility location with selfish users and queues. *Operations Research*, 67(2), 479–497.
- Drezner, Z., & Hamacher, H. (Eds.). (2004). *Facility location, applications and theory* (1st ed.). Berlin: Springer (**2. Printing**).
- Drezner, Z., Klamroth, K., Schöbel, A., & Wesolowsky, G. O. (2004). The Weber problem. In Z. Drezner & H. W. Hamacher (Eds.), *Facility location: Applications and theory* (1st ed., pp. 1–36). Berlin: Springer (**2. Printing**).
- Drezner, Z., Schaible, S., & Simchi-Levi, D. (1990). Queueing-location problems on the plane. *Naval Research Logistics*, 37, 929–935.
- Farahani, R. Z., Bajgan, H. R., Fahimnia, B., & Kaviani, M. (2015). Location-inventory problem in supply chains: A modelling review. *International Journal of Production Research*, 53(12), 3769–3788.
- Heckmann, I., & Nickel, S. (2019). Location logistics in supply chain management. In G. Laporte, S. Nickel, & F. SaldanhadaGama (Eds.), *Facility location, chapter 16* (2nd ed., pp. 453–476). Cham: Springer.
- Heyman, D. P., & Sobel, M. J. (1990). *Stochastic models, volume 2 of handbooks in operations research and management science*. Amsterdam: North Holland (**Editors: Nemhauser, G.L. and Rinnoy Kan, A.H.G.**).
- Huisman, T., & Boucherie, R. J. (2011). Decomposition and aggregation in queueing networks. In R. J. Boucherie & N. M. van Dijk (Eds.), *Queueing networks: A fundamental approach, volume 154 of International series in operations research and management science, chapter 7* (pp. 313–344). New York: Springer.
- Kelly, F. P. (1979). *Reversibility and stochastic networks*. Chichester: Wiley.
- Krishnamoorthy, A., Lakshmy, B., & Manikandan, R. (2011). A survey on inventory models with positive service time. *OPSEARCH*, 48, 153–169.
- Krishnamoorthy, A., & Narayanan, Viswanath C. (2013). Stochastic decomposition in production inventory with service time. *European Journal of Operational Research*, 228, 358–366.

- Krishnamoorthy, A., Shajin, D., & Viswanath, C. N. (2021). Inventory with positive service time: A survey. In V. Anisimov & N. Limnios (Eds.), *Queueing theory 2—advanced trends in queueing theory, mathematics and statistics series, sciences, chapter 6* (pp. 201–238). London: Wiley.
- Lange, V., & Daduna, H. (2022). *The Weber problem in logistic and services networks under congestion*. Preprint, [arXiv.org](https://arxiv.org/abs/2205.12345) (Submitted).
- Laporte, G. (1988). Location-routing problems. In B. L. Golden & A. A. Assad (Eds.), *Vehicle routing: Methods and studies* (pp. 163–198). Amsterdam: North-Holland.
- Larson, R. C. (1974). A hypercube queueing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, 1, 67–95.
- Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling. ASA-SIAM series on statistics and applied probability*. Philadelphia: ASA-SIAM.
- Manzour-al-Ajjad, S. M. H., Torabi, S. A., & Salhi, S. (2012). A hierarchical algorithm for the planar single-facility location routing problem. *Computers & Operations Research*, 39(2), 461–470.
- Melikov, A. Z., & Molchanov, A. A. (1992). Stock optimization in transportation/storage systems. *Cybernetics and Systems Analysis*, 28(3), 484–487.
- Melikov, A. Z., Ponomarenko, L. A., & Bagirova, S. A. (2016). Analysis of queueing-inventory systems with impatient customers. *Journal of Automation and Information Sciences*, 48(1), 53–68.
- Melo, M. T., Nickel, S., & Saldanha-da-Gama, F. (2009). Facility location and supply chain management—A review. *European Journal of Operational Research*, 196, 401–412.
- Min, H., Jayaraman, V., & Srivastava, R. (1998). Combined location-routing problems: A synthesis and future research directions. *European Journal of Operational Research*, 108, 1–15.
- Mirchandani, P. B., & Francis, R. L. (Eds.). (1990). *Discrete location theory*. New York: Wiley.
- Nagy, G., & Salhi, S. (2007). Location-routing: Issues, models and methods. *European Journal of Operational Research*, 177(2), 649–742.
- Neuts, M. F. (1981). *Matrix geometric solutions in stochastic models—An algorithmic approach*. Baltimore, MD: John Hopkins University Press.
- Newell, G. F. (1982). *Applications of queueing theory. Monographs on statistics and applied probability*. (2nd ed.). London: Chapman and Hall.
- Otten, S. (2017). *Integrated models for performance analysis and optimization of queueing-inventory systems in logistics networks*. Ph.D. thesis, Universität Hamburg, Department of Mathematics, Hamburg, Germany.
- Otten, S., Krenzler, R., & Daduna, H. (2016). Models for integrated production-inventory systems: Steady state and cost analysis. *International Journal of Production Research*, 54(20), 6174–6191.
- Otten, S., Krenzler, R., & Daduna, H. (2020). Separable models for interconnected production-inventory systems. *Stochastic Models*, 36(1), 48–93.
- Porteus, E. L. (1990). Stochastic inventory theory. In D. P. Heyman & M. J. Sobel (Eds.), *Stochastic models, volume 2 of Handbooks in operations research and management science, chapter 12* (pp. 605–652). Amsterdam: North-Holland.
- Puterman, M. L. (1990). Markov decision processes. In D. P. Heyman & M. J. Sobel (Eds.), *Stochastic models, volume 2 of Handbooks in operations research and management science, chapter 8* (pp. 331–434). Amsterdam: North-Holland.
- Saffari, M., Asmussen, S., & Haji, R. (2013). The M/M/1 queue with inventory, lost sale and general lead times. *Queueing Systems*, 75, 65–77.
- Salhi, S., & Nagy, G. (2009). Local improvement in planar facility location using vehicle routing. *Annals of Operations Research*, 167, 287–296.
- Salhi, S., & Rand, G. K. (1989). The effect of ignoring routes when locating depots. *European Journal of Operational Research*, 39(2), 150–156.
- Sauer, C. H. (1983). Computational algorithms for state-dependent queueing networks. *ACM Transactions on Computer Systems*, 1(1), 67–92.
- Schwarz, M., Sauer, C., Daduna, H., Kulik, R., & Szekli, R. (2006). M/M/1 queueing systems with inventory. *Queueing Systems*, 54, 55–78.
- Schwarz, M., Wichelhaus, C., & Daduna, H. (2007). Product form models for queueing networks with an inventory. *Communications in Statistics. Stochastic Models*, 23, 627–663.
- Scott, C., Jefferson, T., & Drezner, Z. (1999). Various objectives for the queueing-location problem on the plane. *Asia-Pacific Journal of Operational Research*, 16, 203–214.
- Shajin, D., Benny, B., Razumchik, R. V., & Krishnamoorthy, A. (2018). Discrete product inventory control with positive service time and two operation modes. *Automation and Remote Control*, 79(9), 1593–1608.
- Shaked, M., & Shanthikumar, J. G. (1994). *Stochastic orders and their applications. Probability and mathematical statistics*. Boston: Academic Press.
- Sigman, K., & Simchi-Levi, D. (1992). Light traffic heuristic for an M/G/1 queue with limited inventory. *Annals of Operations Research*, 40, 371–380.

- Song, L., & Wu, Z. (2022). An integrated approach for optimizing location-inventory and location-inventory-routing problem for perishable products. *International Journal of Transportation Science and Technology*.
- Tapiero, C. S. (1971). Transportation-location-allocation problems over time. *Journal of Regional Sciences*, 11(3), 377–384.
- Towsley, D. (1980). Queueing network models with state-dependent routing. *Journal of the Association for Computing Machinery*, 27(2), 323–337.
- van der Wal, J. (1989). Monotonicity of the throughput of a closed exponential queueing network in the number of jobs. *ORSpektrum*, 11, 97–100.
- Wesolowsky, G. O. (1993). The Weber problem: History and perspectives. *Location Science*, 1(1), 5–23.
- Wolff, R. W. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs: Prentice-Hall International Editions.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.