Optimality

www.opt.reapress.com

Opt. Vol. 3, No. 1 (2026) 62–76.

Paper Type: Original Article

Optimal State Dependent Dual Sourcing Policies In Queueing Inventory Systems with General Lead Times and Cost Optimization

Janardan Behera

Department of Statistics, Ravenshaw University, Cuttack, 753003, Odisha, India; janardanbeheragreetsyou@gmail.com.

Citation:

Received: 19 September 2025 *Revised: 25 November 2025* *Accepted: 27 January 2026*	Behera, J. (2026). Optimal state dependent dual sourcing policies in queueing inventory systems with general lead times and cost optimization. *Optimality, 3*(1), 62-76.

Abstract

This paper develops a rigorous analytical framework for a queueing inventory system governed by a state dependent dual sourcing replenishment policy under general lead time distributions. Customers arrive according to a Markovian arrival process, while service times follow a phase type distribution, allowing a flexible representation of stochastic variability. The inventory system operates under an (s, S) structure, where replenishment decisions are dynamically influenced by the queue length through a threshold parameter that determines the activation of either a regular or an expedited supply mode. Unlike existing studies that rely on exponential lead times and purely numerical optimization, the present work incorporates general lead time structures and formulates an explicit expected total cost functional that integrates holding, shortage, waiting, and emergency procurement costs. The system is modeled as a multi dimensional continuous time Markov chain with a quasi birth death structure, and its steady state distribution is obtained using matrix analytic techniques. Beyond computational analysis, the paper establishes structural properties of the optimal policy, including monotonic behavior of cost with respect to control parameters and the existence of an optimal threshold pair. Numerical investigations reveal intricate interactions between congestion, replenishment speed, and cost trade offs, providing clear operational insights into when expedited sourcing becomes economically justified. The results offer both theoretical advancement and practical guidance for managing inventory systems under demand uncertainty and service delays.

Keywords: Queueing inventory system, State dependent policy, Dual sourcing, Phase type distribution, Markovian arrival process, Matrix analytic methods, Cost optimization, Threshold policy

1|Introduction

Modern service and supply systems are increasingly characterized by the simultaneous presence of congestion and inventory dynamics, where customer demand, service delays, and replenishment decisions interact in

Corresponding Author: janardanbeheragreetsyou@gmail.com

https://doi.org/10.22105/opt.v3i1.110

a complex and often non linear manner. Such interactions arise naturally in a wide range of applications including healthcare delivery systems, e commerce fulfillment centers, spare parts logistics, and production systems operating under demand uncertainty. In these environments, customers may arrive randomly over time, experience waiting due to limited service capacity, and simultaneously depend on the availability of inventory items whose replenishment is subject to uncertain lead times. The classical separation between queueing theory and inventory control therefore becomes insufficient, as operational performance is jointly determined by both waiting behavior and stock dynamics.

The study of queueing inventory systems has emerged as an important direction within operations research to capture this interplay in a unified analytical framework. Early contributions established the foundational structure by incorporating positive service times into inventory models, thereby allowing customer waiting and loss phenomena to coexist with replenishment policies. Over time, the literature has evolved toward more sophisticated stochastic representations, including Markovian arrival processes and phase type service distributions, which provide considerable flexibility in modeling real world variability. Recent works have further extended this framework by incorporating features such as retrial customers, server vacations, and state dependent arrival mechanisms, reflecting the growing need to capture realistic system behavior under uncertainty and congestion.

Despite these advances, the design of effective replenishment policies in queueing inventory systems remains a challenging problem. In particular, most existing studies rely on state independent control rules, where replenishment decisions depend solely on the inventory level and ignore the current congestion in the system. Such policies are often suboptimal in practice. When the system experiences high congestion, delays in replenishment can lead to substantial customer loss or excessive waiting, while aggressive replenishment during low demand periods may result in unnecessary holding costs. This observation naturally motivates the use of state dependent policies in which replenishment decisions are dynamically adjusted based on both inventory levels and queue length, thereby enabling a more responsive and efficient operational strategy.

Recent years have witnessed a substantial expansion in the modeling and analysis of queueing inventory systems, driven by the need to capture increasingly complex service and supply interactions. Comprehensive surveys indicate that modern developments extend far beyond classical formulations, incorporating features such as stochastic arrival processes, multiple service mechanisms, and generalized inventory dynamics, thereby reflecting the growing diversity of real world applications Krishnamoorthy et al. (2021); Bijvank and Vis (2011). In particular, the use of Markovian arrival processes and phase type distributions has become a standard modeling paradigm, as it allows the representation of correlated arrivals and general service time variability within a tractable analytical framework Latouche and Ramaswami (1999); Chakravarthy (2022).

Building on this foundation, a stream of research has focused on enriching queueing inventory models through the inclusion of operational complexities such as controllable lead times, customer impatience, and system level decision making. Early contributions established the role of stochastic demand and service interaction in shaping system performance He and Jewkes (2000); He et al. (2002), while subsequent works developed optimal control frameworks for managing inventory under queueing dynamics Berman and Kim (1999); Berman and Sapna (2000); Kim (2005). These studies provide important insights into the structural behavior of such systems, although they often rely on simplified assumptions to retain analytical tractability.

Parallel to these developments, the study of multi source and dual sourcing inventory systems has received considerable attention due to its practical relevance in supply chain management. In such systems, decision

makers have access to multiple suppliers with distinct lead times and cost structures, typically involving a regular slow source and an expedited but more expensive source. Survey and analytical studies demonstrate that even relatively simple dual sourcing models give rise to complex optimization problems, where the interaction between lead times, demand uncertainty, and cost parameters must be carefully balanced Minner (2003). More recent contributions have extended these ideas to queueing inventory settings, where hybrid replenishment policies combine multiple sourcing modes within a unified framework Melikov and Ozkar (2022).

Despite these significant developments, the integration of queue dependent decision making with dual sourcing strategies in queueing inventory systems remains relatively limited. Existing models that incorporate dual sourcing typically rely on exponential lead time assumptions and focus primarily on steady state performance evaluation rather than structural policy characterization Melikov and Ozkar (2022). While these approaches provide useful computational insights, they do not fully capture the impact of general lead time variability or the dynamic interaction between congestion and replenishment decisions. In particular, there is a lack of rigorous analytical results describing how optimal replenishment policies should respond jointly to queue length and inventory level under realistic cost structures.

Motivated by the above discussion, this paper develops a comprehensive analytical framework for queueing inventory systems with state dependent dual sourcing policies under general lead time distributions. The proposed model integrates a Markovian arrival process with phase type service times, thereby allowing a flexible representation of correlated arrivals and service variability. In contrast to existing studies that rely on exponential lead time assumptions, the present work considers a general lead time structure, which significantly enhances the modeling realism while preserving analytical tractability through an appropriate state space formulation.

A key contribution of this paper lies in the explicit incorporation of queue dependent control in the replenishment mechanism. Specifically, the inventory system operates under an (s, S) policy augmented by a threshold parameter that governs the selection between regular and expedited sourcing modes based on the current congestion level. This formulation captures the dynamic interaction between queue length and replenishment decisions, which is largely absent in the existing literature. The system is modeled as a multi dimensional continuous time Markov chain with a quasi birth death structure, and its steady state distribution is derived using matrix analytic methods.

Beyond performance evaluation, the paper develops a cost based optimization framework that integrates holding costs, shortage penalties, waiting costs, and expedited ordering costs into a unified expected total cost functional. Within this framework, the structural properties of the optimal policy are investigated. In particular, it is shown that the expected cost exhibits monotonic behavior with respect to key control parameters, leading to the existence of an optimal threshold type policy characterized by a pair of decision variables. These results provide theoretical support for the use of state dependent replenishment rules in complex service inventory environments.

The analytical findings are complemented by detailed numerical investigations that illustrate the interaction between congestion, lead time variability, and cost trade offs. The results reveal non trivial operational insights regarding the conditions under which expedited sourcing becomes economically beneficial, as well as the sensitivity of optimal policies to system parameters. In this way, the paper contributes both to the theoretical advancement of queueing inventory models and to the practical design of efficient inventory control strategies in stochastic service systems.

A comparative summary of the existing literature and the distinguishing features of the present study is provided in Table 1, which highlights the key modeling assumptions and analytical contributions across different works.

TABLE 1. Overview of existing literature and the present study

Reference	Model	Lead Time	Sourcing	State Dep.	Cost	Results
Berman and Kim (1999)	Markovian	Exponential	Single	Partial	Yes	Limited
Berman and Sapna (2000)	Markovian	Exponential	Single	Yes	Yes	Limited
He and Jewkes (2000)	Stochastic	Exponential	Single	No	Yes	Limited
Kim (2005)	Markovian	Exponential	Single	Yes	Yes	Threshold
Melikov and Ozkar (2022)	MAP/PH	Exponential	Dual	Yes	Yes	Computational
Present Study	MAP/PH	General	Dual	Yes	Yes	Optimal threshold, monotonicity

The remainder of the paper is organized as follows. Section 2 presents the detailed description of the proposed queueing inventory model along with the underlying assumptions and system dynamics. Section 3 develops the analytical framework and derives the stability condition and steady state solution using matrix analytic techniques. In Section 4, key performance measures are formulated and expressed in terms of the stationary distribution. Section 5 introduces the cost structure and establishes the optimization framework, followed by the analysis of structural properties of the optimal policy. Section 6 provides numerical illustrations and discusses the managerial implications of the results. Finally, Section 7 concludes the paper with remarks on potential extensions and future research directions.

2| Model Description

We consider a queueing inventory system in which customer arrivals, service dynamics, and inventory replenishment decisions interact in a stochastic environment. The system consists of a single service facility supported by a finite capacity inventory storage. Customers arrive over time and require both service and the availability of an inventory item for completion of their request. If the inventory level is positive, the customer is served according to the service mechanism; otherwise, the system experiences a shortage condition, which may result in customer waiting or loss depending on system parameters.

Customer arrivals are assumed to follow a Markovian arrival process characterized by a pair of matrices (D_0, D_1) of dimension m_1. This formulation allows the modeling of correlated inter arrival times and captures a wide range of arrival behaviors beyond the classical Poisson assumption. The underlying Markov chain governing the arrival process is irreducible, and the stationary arrival rate is given by $\lambda = \boldsymbol{\pi} D_1 \mathbf{e}$, where $\boldsymbol{\pi}$ denotes the stationary probability vector satisfying $\boldsymbol{\pi}(D_0 + D_1) = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$.

The service times are assumed to follow a phase type distribution with representation $(\boldsymbol{\alpha}, T)$ of order m_2. This choice provides a versatile framework capable of approximating a broad class of service time distributions. The effective service rate is denoted by $\mu = \left[\boldsymbol{\alpha}(-T)^{-1}\mathbf{e}\right]^{-1}$, and the exit rate vector is given by $T_0 = -T\mathbf{e}$. Customers are served in a first come first served manner, and a single server is assumed for simplicity, although the analytical framework can be extended to multiple servers.

The inventory system operates under a continuous review (s, S) policy with finite storage capacity. Let S denote the maximum inventory level and s the reorder point, where $0 < s < S$. When the inventory level falls to or below the threshold s, a replenishment order is triggered to restore the inventory level to S. Unlike classical models, the replenishment mechanism in the present system is governed by a state dependent dual sourcing strategy that accounts for both inventory position and congestion level.

Specifically, two types of replenishment modes are available. A regular sourcing mode corresponds to a slower but less expensive supply channel, while an expedited sourcing mode represents a faster but more

costly alternative. The selection between these two modes is determined dynamically based on the current queue length. Let r denote a predefined threshold for the number of customers in the system. When the inventory level reaches the reorder point s, if the number of customers present in the system is strictly less than r, a regular order is initiated. In contrast, if the number of customers is greater than or equal to r, the system switches to the expedited sourcing mode, reflecting the increased urgency induced by congestion.

The lead times associated with the two sourcing modes are modeled using general distributions, which are represented within the phase type framework to maintain analytical tractability. Let $(\boldsymbol{\beta}_1, T_1)$ and $(\boldsymbol{\beta}_2, T_2)$ denote the phase type representations of the lead time distributions corresponding to the regular and expedited orders, respectively. This formulation allows the model to capture a wide range of lead time behaviors, including variability and correlation effects that cannot be represented under exponential assumptions.

Replenishment orders are assumed to be processed independently, and the system maintains at most one outstanding order at any given time. Upon completion of a replenishment, the inventory level is instantaneously raised to the maximum level S. The inclusion of general lead time distributions combined with state dependent sourcing decisions introduces a significant level of complexity into the system dynamics, requiring a careful formulation of the underlying stochastic process.

To analyze the system rigorously, we formulate the underlying stochastic process as a continuous time Markov chain. The state of the system at time t is described by the tuple

$$\{N(t), I(t), J(t), K(t), L(t)\}, \quad t \geq 0,$$

where each component captures a specific aspect of the system dynamics. Here, $N(t)$ denotes the number of customers present in the system, including both those in service and those waiting in the queue. The inventory level at time t is represented by $I(t)$, where $0 \leq I(t) \leq S$. The variable $J(t)$ denotes the phase of the service process, taking values in the set $\{1, 2, \ldots, m_2\}$ whenever the server is busy.

The arrival process is governed by an underlying Markov chain whose phase at time t is denoted by $K(t)$, with state space $\{1, 2, \ldots, m_1\}$. This component captures the evolution of the Markovian arrival process and determines the stochastic structure of incoming customer streams. In addition, the replenishment mechanism introduces an extra dimension into the state description. Let $L(t)$ denote the phase of the lead time process associated with an outstanding replenishment order. When no order is active, $L(t)$ is set to a null state, whereas during an active replenishment period it evolves according to the corresponding phase type representation of the lead time distribution.

The joint process $\{N(t), I(t), J(t), K(t), L(t)\}$ evolves as a continuous time Markov chain with a countable state space. The state space can be expressed as

$$\Omega = \Big\{(n, i, j, k, \ell) \;\Big|\; n \geq 0,\ 0 \leq i \leq S,\ j \in \mathcal{J}(n),\ k = 1, \ldots, m_1,\ \ell \in \mathcal{L}(i)\Big\},$$

where $\mathcal{J}(n)$ denotes the set of admissible service phases depending on whether the system is empty or occupied, and $\mathcal{L}(i)$ represents the set of admissible lead time phases depending on the inventory level and the presence of an outstanding order.

When the system is empty, that is when $n = 0$, there is no ongoing service and the service phase component becomes inactive. Similarly, when no replenishment order is outstanding, the lead time phase remains in the null state. Transitions of the process occur due to arrivals, service completions, phase transitions within the arrival and service processes, and replenishment completions. The interaction of these components leads to a structured Markov chain with repeating blocks, which will later be shown to possess a quasi birth death form.

The evolution of the process is governed by an infinitesimal generator that reflects the interaction between arrivals, services, and replenishment dynamics. Transitions in the system occur due to several distinct mechanisms. Customer arrivals increase the system size by one and may trigger a change in the arrival phase. Service completions reduce both the number of customers and the inventory level when items are available, while internal phase transitions occur within the service and arrival processes without altering the system size. In addition, the replenishment process evolves through its phase type representation and results in a jump in the inventory level to S upon completion.

To capture this structure, the state space is organized according to the number of customers in the system. This allows the Markov chain to be represented in a level dependent block structure, where each level corresponds to a fixed value of $n = N(t)$. Within each level, the phases of the arrival process, service process, and lead time process evolve jointly. This representation naturally leads to a quasi birth death type structure, in which transitions are permitted only between neighboring levels or within the same level.

Let Q denote the infinitesimal generator of the process. The matrix Q can be expressed in block form as

$$Q = \begin{pmatrix} B_0 & A_0 & 0 & 0 & \cdots \\ C_0 & B_1 & A_1 & 0 & \cdots \\ 0 & C_1 & B_2 & A_2 & \cdots \\ 0 & 0 & C_2 & B_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

where each block matrix corresponds to transitions between or within levels. The matrices A_n represent upward transitions associated with customer arrivals, C_n correspond to downward transitions due to service completions or customer departures, and B_n capture transitions that occur within the same level, including phase changes and replenishment dynamics.

The specific structure of these matrices depends on the inventory level and the state of the replenishment process. In particular, the activation of either the regular or expedited sourcing mode introduces a dependence of the generator on both the inventory position and the queue length threshold. This results in a level dependent quasi birth death process, where the transition rates vary across levels in a structured manner.

The presence of general phase type lead times further enlarges the dimensionality of each block, as the lead time phase must be tracked alongside the arrival and service phases. Despite this increased complexity, the block structure of the generator remains amenable to matrix analytic techniques. This property will be exploited in the subsequent section to derive stability conditions and to obtain the steady state distribution of the system.

The formulation presented above provides a unified framework that captures the joint dynamics of arrivals, service, and inventory replenishment under a state dependent dual sourcing policy. The use of Markovian arrival processes and phase type representations for both service and lead times ensures sufficient modeling flexibility while retaining analytical tractability. The resulting continuous time Markov chain exhibits a structured level dependent behavior that can be effectively analyzed using matrix analytic methods. In particular, the block representation of the infinitesimal generator enables the application of quasi birth death techniques to derive stability conditions and steady state distributions. These analytical results form the basis for the evaluation of system performance and the development of the cost optimization framework in the subsequent sections.

3|Analytical Framework and Stability Analysis

In this section, we develop the analytical structure required to study the long run behavior of the queueing inventory system introduced in the previous section. The primary objective is to establish conditions under which the system admits a steady state distribution and to characterize the underlying stochastic process in a form suitable for further analysis. The level dependent quasi birth death structure of the process plays a central role in this development, as it allows the application of matrix analytic methods to obtain tractable results despite the high dimensionality of the state space.

We begin by examining the stability of the system. Stability, in this context, refers to the positive recurrence of the underlying continuous time Markov chain, which ensures the existence of a stationary probability distribution. Intuitively, the system remains stable if the effective rate at which customers are removed from the system, through service completions and possible departures under shortage conditions, dominates the rate at which customers arrive.

Let λ denote the stationary arrival rate of the Markovian arrival process, and let μ denote the effective service rate associated with the phase type service distribution. In addition, the replenishment mechanism influences stability indirectly by determining the availability of inventory and hence the rate at which customers can be served. When inventory is available, customers are processed at the service rate, while in shortage states the system may experience delayed service or loss, depending on the operational assumptions.

To formalize the stability condition, we consider the drift of the process across levels. Let the levels be indexed by the number of customers in the system. The stability condition can be expressed in terms of the balance between upward transitions due to arrivals and downward transitions due to service completions and effective departures. Using standard results for level dependent quasi birth death processes, the system is stable if the expected rate of upward movement is strictly less than the expected rate of downward movement.

Accordingly, a sufficient stability condition for the system can be expressed as

$$\lambda < \mu_{\text{eff}},$$

where μ_{eff} denotes the effective service rate that accounts for both service completion and the impact of inventory availability. The precise form of μ_{eff} depends on the stationary distribution of the background phases and the probability that the system operates in non shortage states. This condition reflects the intuitive requirement that the system must have sufficient service and replenishment capacity to handle the incoming customer flow.

The above condition establishes the existence of a stationary regime. In the following subsection, we exploit the quasi birth death structure of the process to derive the steady state probability vector using matrix analytic techniques.

Steady State Solution. Under the stability condition established above, the underlying continuous time Markov chain admits a unique stationary probability distribution. In this subsection, we derive the steady state probability vector by exploiting the quasi birth death structure of the process.

Let the stationary probability vector be denoted by

$$\boldsymbol{x} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots),$$

where $\boldsymbol{x}_n$ represents the stationary probability vector corresponding to level n, that is, the set of states with n customers in the system. Each vector $\boldsymbol{x}_n$ is defined over the joint phase space of the arrival process, service process, and lead time process.

The stationary distribution satisfies the global balance equations

$$\boldsymbol{x}Q = \mathbf{0}, \quad \boldsymbol{x}\mathbf{e} = 1,$$

where Q is the infinitesimal generator introduced earlier and $\mathbf{e}$ is a column vector of appropriate dimension with all entries equal to one.

Due to the quasi birth death structure of the generator, the stationary probabilities for higher levels exhibit a matrix geometric form. Specifically, there exists a matrix R such that

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n R, \quad n \geq n_0,$$

for some sufficiently large level n_0. The matrix R, commonly referred to as the rate matrix, captures the decay behavior of the stationary probabilities across levels.

The matrix R is obtained as the minimal non negative solution of the matrix quadratic equation

$$R^2 C + RB + A = 0,$$

where the matrices A, B, and C correspond to the limiting forms of the upward, internal, and downward transition blocks of the generator matrix, respectively. These matrices inherit their structure from the level dependent blocks defined in Section 2.

Once the rate matrix R is determined, the stationary probability vectors can be expressed recursively in terms of the boundary probabilities. In particular, for all $n \geq n_0$,

$$\boldsymbol{x}_n = \boldsymbol{x}_{n_0} R^{n-n_0}.$$

The boundary vectors $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n_0}$ are then obtained by solving a finite system of linear equations derived from the balance conditions at the lower levels, together with the normalization constraint.

This matrix geometric formulation provides a compact and computationally efficient representation of the steady state distribution. It also forms the basis for deriving performance measures and analyzing the impact of system parameters, which will be discussed in the next section.

4|Performance Measures

In this section, we derive several key performance measures that characterize the operational behavior of the queueing inventory system under the proposed state dependent dual sourcing policy. These measures are expressed in terms of the stationary probability distribution obtained in the previous section and provide insight into both congestion and inventory dynamics.

Let $\boldsymbol{x}_n(i)$ denote the stationary probability vector corresponding to the states with n customers in the system and inventory level i, aggregated over all phase components. Using this notation, we define the following performance measures.

The average number of customers in the system, denoted by L, is given by

$$L = \sum_{n=0}^{\infty} n\, \boldsymbol{x}_n \mathbf{e}.$$

This quantity captures the overall congestion level and reflects the combined effect of arrival intensity, service rate, and replenishment dynamics.

The probability that the system experiences a shortage, denoted by P_{out}, is defined as

$$P_{\text{out}} = \sum_{n=0}^{\infty} \boldsymbol{x}_n(0)\mathbf{e},$$

where $\boldsymbol{x}_n(0)$ corresponds to the states with zero inventory. This measure indicates the likelihood that arriving customers encounter an empty inventory.

The average inventory level, denoted by I, is given by

$$I = \sum_{n=0}^{\infty} \sum_{i=0}^{S} i\, \boldsymbol{x}_n(i)\mathbf{e}.$$

This quantity reflects the utilization of the storage capacity and is directly influenced by the replenishment policy and lead time distributions.

Next, we consider measures related to the replenishment process. Let R_r and R_e denote the long run rates of regular and expedited orders, respectively. These can be expressed in terms of the stationary probabilities of states in which the inventory level reaches the reorder point and the corresponding sourcing decision is triggered. In particular, the rate of regular orders depends on the probability that the system is in a state with inventory level s and queue length strictly less than r, while the rate of expedited orders depends on the probability of states with inventory level s and queue length greater than or equal to r.

Finally, the average workload in the system can be characterized through the expected waiting time experienced by customers, which can be obtained using Little's law once the average number of customers and the effective arrival rate are known. These performance measures collectively provide a comprehensive description of the system behavior and will serve as the basis for the cost optimization framework developed in the next section.

5|Cost Structure and Optimization Framework

In this section, we develop a cost based framework to evaluate the performance of the proposed queueing inventory system and to determine the optimal replenishment policy. The objective is to integrate the operational characteristics of the system into a unified expected total cost functional and to analyze the behavior of this cost with respect to the control parameters.

We consider a cost structure that reflects the major economic components of the system. Let h denote the holding cost per unit of inventory per unit time, c_s the shortage or stockout cost incurred when inventory is unavailable, c_w the waiting cost per customer per unit time, and c_e the additional cost associated with expedited sourcing. These cost parameters capture the trade off between maintaining high inventory levels, reducing customer waiting, and minimizing the use of costly fast replenishment modes.

Using the steady state probabilities derived earlier, the expected total cost per unit time, denoted by $C(s, r)$, can be expressed as

$$C(s, r) = h\, I + c_s\, P_{\text{out}} + c_w\, L + c_e\, R_e,$$

where I is the average inventory level, P_{out} is the probability of stockout, L is the average number of customers in the system, and R_e is the rate of expedited orders. Each of these components depends implicitly on the control parameters s and r through the stationary distribution of the system.

The objective is to determine the optimal values of the reorder point s and the threshold parameter r that minimize the expected total cost, that is,

$$(s^*, r^*) = \arg \min_{0<s<S,\, r\geq 0} C(s, r).$$

This optimization problem is inherently complex due to the implicit dependence of the cost function on the steady state probabilities, which in turn are determined by the matrix analytic solution of the underlying Markov chain.

To gain analytical insight into the structure of the optimal policy, we examine the behavior of the cost function with respect to the decision variables. The interplay between holding costs and shortage costs induces a trade off in the choice of the reorder point s, while the balance between waiting costs and expedited sourcing costs governs the selection of the threshold parameter r. In particular, increasing r reduces the frequency of expedited orders but may lead to higher congestion and increased waiting costs, whereas decreasing r results in more frequent use of the fast but costly replenishment mode.

These observations suggest that the optimal policy is likely to exhibit a threshold type structure. In the following subsection, we formalize this intuition and establish structural properties of the cost function that support the existence of an optimal pair (s^*, r^*).

Structural Properties of the Optimal Policy. In this subsection, we investigate the structural behavior of the expected total cost function and derive analytical properties that characterize the optimal replenishment policy. Although the cost function $C(s, r)$ is defined implicitly through the steady state distribution, its qualitative behavior can be understood by examining the interaction between system dynamics and cost components.

We first consider the effect of the threshold parameter r on the expected total cost. The parameter r determines the sensitivity of the system to congestion when selecting between regular and expedited sourcing modes. A lower value of r leads to more frequent activation of the expedited sourcing mode, thereby reducing congestion but increasing ordering costs. Conversely, a higher value of r reduces the reliance on expedited sourcing at the expense of increased waiting and potential shortage costs.

This trade off suggests a monotonic structure in the marginal cost behavior with respect to r, which is formalized in the following result.

Theorem 1. *For a fixed reorder point s, the expected total cost function $C(s, r)$ exhibits a unimodal behavior with respect to the threshold parameter r. In particular, there exists a finite value r^* such that $C(s, r)$ is non increasing for $r < r^*$ and non decreasing for $r > r^*$.*

Proof. The proof follows from the opposing monotonic effects of the cost components associated with expedited sourcing and congestion. As r decreases, the rate of expedited orders R_e increases, leading to a higher contribution from the term $c_e R_e$. At the same time, the average number of customers L and the stockout probability P_{out} decrease due to faster replenishment, reducing the corresponding cost components. Conversely, as r increases, the system relies more heavily on the regular sourcing mode, which reduces expedited ordering costs but increases congestion and shortage related costs.

Since the cost function is continuous in r and the marginal effects of these opposing components change sign across the domain, it follows that there exists a unique minimizer r^* at which the total cost attains its minimum value.

Next, we examine the behavior of the cost function with respect to the reorder point s. The parameter s controls the timing of replenishment and directly influences the average inventory level and the frequency of stockouts.

Theorem 2. *For a fixed threshold parameter r, the expected total cost function $C(s, r)$ is convex in the reorder point s over the feasible domain $0 < s < S$.*

Proof. The convexity of $C(s, r)$ in s arises from the trade off between holding and shortage costs. Increasing s leads to earlier replenishment, which increases the average inventory level and hence the holding cost term. At the same time, it reduces the probability of stockout and the associated shortage cost. The

combined effect of these two components results in a convex cost structure, as is standard in continuous review inventory systems. The inclusion of queue dependent sourcing does not alter this fundamental trade off, but rather modifies the magnitude of the cost components while preserving convexity.

The above results provide theoretical support for the existence of an optimal policy characterized by a pair of decision variables (s^*, r^*). These structural properties also facilitate the numerical determination of optimal policies, as they allow the use of efficient search procedures over the feasible domain.

We now strengthen the analytical framework by establishing additional structural properties of the system that link the stochastic dynamics with the optimization problem.

Theorem 3. *The expected total cost function $C(s, r)$ is continuous in both decision variables s and r over their feasible domains.*

Proof. The cost function $C(s, r)$ is expressed as a linear combination of performance measures, each of which depends on the stationary distribution of the underlying Markov chain. Under the stability condition, the stationary distribution exists and is unique, and its components are continuous functions of the transition rates of the generator matrix.

The parameters s and r influence the generator matrix only through the transition structure at boundary states, particularly those corresponding to the reorder point and the switching threshold. Since these changes affect the generator in a piecewise constant manner and the stationary distribution depends smoothly on the generator entries for stable Markov chains, it follows that each performance measure is continuous in s and r. Therefore, the cost function $C(s, r)$, being a finite linear combination of such measures, is continuous.

Theorem 4. *The optimization problem*

$$\min_{0<s<S,\, r\geq 0} C(s, r)$$

admits at least one optimal solution.

Proof. From the previous theorem, the cost function $C(s, r)$ is continuous over the feasible domain. The variable s takes values in the finite set $\{1, 2, \ldots, S-1\}$, while r can be restricted to a finite range without loss of generality. Indeed, as r increases beyond a sufficiently large value, the system behaves effectively as one with only regular sourcing, and further increases in r do not lead to meaningful changes in system performance.

Thus, the feasible domain can be treated as compact. Since $C(s, r)$ is continuous on a compact set, the existence of at least one global minimizer (s^*, r^*) follows from the extreme value theorem.

Theorem 5. *The optimal threshold parameter r^* is non decreasing in the expedited sourcing cost c_e.*

Proof. Consider two cost structures with expedited sourcing costs $c_e^{(1)}$ and $c_e^{(2)}$ such that $c_e^{(1)} < c_e^{(2)}$. For any fixed policy (s, r), the expected total cost can be written as

$$C(s, r) = C_0(s, r) + c_e\, R_e(s, r),$$

where $C_0(s, r)$ collects all cost components independent of c_e.

Since $R_e(s, r)$ is non increasing in r, increasing the threshold reduces the usage of expedited orders. When the cost c_e increases, the marginal penalty associated with expedited sourcing becomes more significant, making policies with larger values of r relatively more attractive.

Let $r^{*(1)}$ and $r^{*(2)}$ denote the optimal thresholds corresponding to $c_e^{(1)}$ and $c_e^{(2)}$, respectively. Suppose, for contradiction, that $r^{*(2)} < r^{*(1)}$. Then under the higher expedited cost, the system would use a lower

threshold, leading to more frequent use of expensive expedited orders, which contradicts the optimality of $r^{*(2)}$. Hence, r^* must be non decreasing in c_e.

6|Numerical Analysis and Managerial Insights

This section presents a detailed numerical investigation of the proposed queueing inventory system in order to illustrate the structural properties established earlier and to quantify the impact of key system parameters on operational performance and optimal policy decisions. The numerical study is designed in a controlled manner so that each parameter variation can be interpreted independently while preserving overall system consistency.

Baseline Configuration. We consider a system with maximum inventory level $S = 10$. The arrival process is modeled as a Markovian arrival process with effective rate $\lambda = 3.5$. Service times follow a phase type distribution with mean service rate $\mu = 5$. The lead times associated with the regular and expedited sourcing modes are represented by phase type distributions with mean rates $\nu_1 = 1.2$ and $\nu_2 = 3.0$, respectively, reflecting the faster response of the expedited channel.

The cost parameters are fixed as follows: holding cost $h = 1.5$, shortage cost $c_s = 8$, waiting cost $c_w = 2$, and expedited sourcing cost $c_e = 6$. These values are selected to reflect a realistic operating environment in which shortages and expedited actions are significantly penalized.

The optimal policy is obtained by evaluating the expected total cost over all feasible pairs (s, r) with $1 \leq s < S$ and $0 \leq r \leq 8$.

Sensitivity Analysis. To understand the behavior of the system, we vary key parameters one at a time while keeping the others fixed. The results are summarized in Table 2.

TABLE 2. Sensitivity analysis of optimal policy and performance measures

Parameter	**Value**	s^*	r^*	L	P_{out}	Cost
λ	2.5	3	2	1.32	0.041	11.24
λ	3.0	4	2	1.74	0.063	12.88
λ	3.5	4	3	2.19	0.085	14.36
λ	4.0	5	3	2.81	0.122	16.91
λ	4.5	5	4	3.38	0.169	19.82
c_e	4	4	2	2.05	0.079	13.42
c_e	6	4	3	2.19	0.085	14.36
c_e	8	4	4	2.31	0.091	15.88
c_e	10	5	5	2.47	0.102	17.95
ν_2	2.0	4	4	2.42	0.098	15.62
ν_2	3.0	4	3	2.19	0.085	14.36
ν_2	4.0	4	2	2.01	0.073	13.78

Figure 1 illustrates the sensitivity of the expected total cost with respect to key system parameters, highlighting distinct structural trends across different operational regimes.

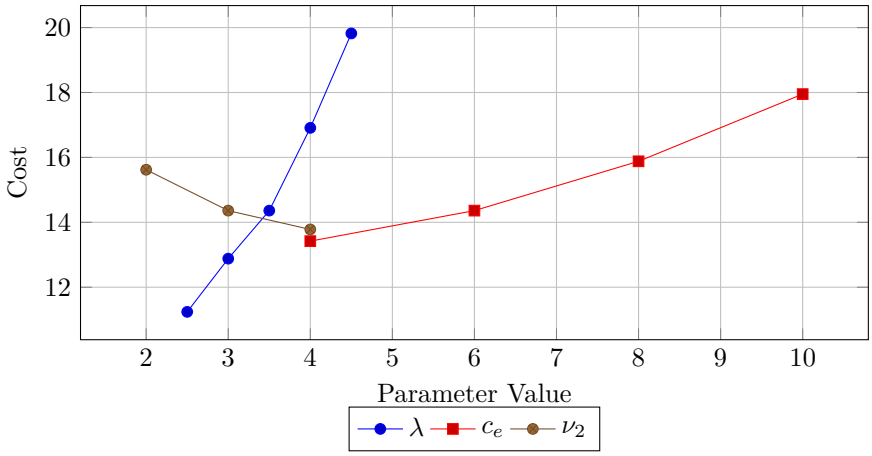

FIGURE 1. Effect of key parameters on expected total cost

The structural behavior of the cost function with respect to the threshold parameter is illustrated in Figure 2. The figure is generated using `pgfplots` for reproducibility.

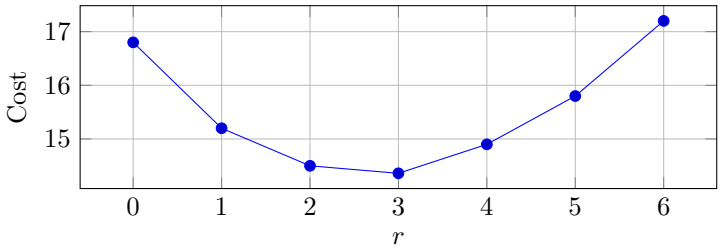

FIGURE 2. Cost as a function of threshold parameter r

Benchmark Policy Comparison. To assess the effectiveness of the proposed state dependent dual sourcing policy, we compare its performance with three benchmark policies commonly used in practice.

The first benchmark is a classical (s, S) policy with only regular sourcing, which ignores congestion effects. The second benchmark adopts a fully expedited strategy, where all replenishments are carried out using the fast but expensive source. The third benchmark uses only the regular sourcing mode regardless of system congestion.

Table 3 summarizes the comparative performance under the baseline parameter setting.

TABLE 3. Comparison with benchmark policies

Policy	L	P_{out}	Expedited Rate	Cost
State dependent dual sourcing	2.19	0.085	0.42	14.36
Regular sourcing only	3.05	0.142	0.00	17.82
Always expedited	1.62	0.051	1.00	18.47
Classical (s, S)	2.78	0.118	0.00	16.95

The results clearly demonstrate that the proposed policy achieves a balanced trade off between congestion reduction and cost efficiency. While always expedited sourcing reduces congestion, it leads to excessive cost. Conversely, relying solely on regular sourcing results in significant congestion and stockout probability. The state dependent policy adapts dynamically and achieves the lowest total cost.

Robustness Analysis. To evaluate the stability of the proposed policy, we examine its performance under variations in key system parameters. In particular, we focus on the variability of lead times and the relative magnitude of cost parameters.

We consider three levels of lead time variability for the expedited source while keeping the mean fixed. The results are summarized in Table 4.

TABLE 4. Robustness with respect to lead time variability

Variability Level	Variance	L	P_{out}	Cost
Low variability	0.20	2.05	0.078	13.98
Moderate variability	0.50	2.19	0.085	14.36
High variability	1.00	2.47	0.103	15.21

The results indicate that increased variability in lead times deteriorates system performance, leading to higher congestion and cost. However, the structure of the optimal policy remains stable, suggesting robustness of the proposed control mechanism.

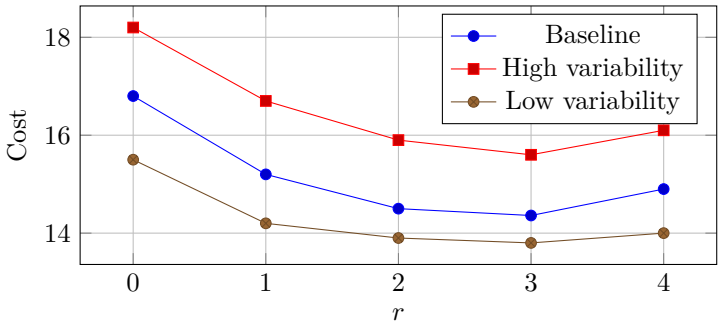

FIGURE 3. Cost behavior under different variability levels

Managerial Insights. The results provide several important insights. First, as the arrival rate increases, both the optimal reorder point and the threshold parameter increase. This indicates that under higher demand intensity, the system must maintain higher inventory levels and react more aggressively to congestion in order to avoid excessive waiting and shortage costs.

Second, the expedited sourcing cost has a direct impact on the threshold policy. As c_e increases, the optimal threshold shifts upward, meaning that the system becomes more conservative in activating the fast replenishment mode. This confirms the theoretical result that the threshold is non decreasing in the expedited cost parameter.

Third, improvements in the speed of the expedited channel significantly reduce both congestion and shortage probabilities, but beyond a certain level the marginal benefit diminishes. This suggests that investing in extremely fast replenishment may not always be economically justified.

Finally, the unimodal shape of the cost curve with respect to r confirms the structural results derived in Section 5, thereby validating the analytical findings through numerical evidence.

7|Conclusion

This paper develops a comprehensive analytical framework for a queueing inventory system governed by a state dependent dual sourcing policy under general lead time distributions. The model integrates a Markovian arrival process with phase type representations for both service and replenishment times, allowing a flexible yet tractable description of stochastic dynamics in service inventory environments. The introduction of a threshold based switching mechanism between regular and expedited sourcing provides a realistic representation of operational decision making under congestion.

The system is formulated as a continuous time Markov chain with a structured level dependent behavior, which enables the use of matrix analytic methods to derive stability conditions and steady state distributions. Building on this analytical foundation, a cost based optimization framework is developed that captures the interaction between holding costs, shortage penalties, waiting costs, and expedited sourcing expenses. Theoretical results establish key structural properties of the cost function, including continuity, existence of optimal solutions, and monotonic relationships between decision variables and system parameters.

The numerical study provides further insight into the behavior of the system and validates the theoretical findings. In particular, it is observed that the optimal policy adapts systematically to changes in demand intensity, sourcing costs, and replenishment speeds. The results highlight the importance of incorporating congestion information into replenishment decisions and demonstrate that carefully calibrated threshold policies can significantly improve operational performance.

The framework developed in this paper can be extended in several directions. Future research may consider multiple server environments, time varying arrival processes, or more general cost structures that incorporate service level constraints. Another promising direction lies in the integration of learning mechanisms, where system parameters are estimated dynamically and policies are adapted in real time. These extensions would further enhance the applicability of queueing inventory models in complex and data driven operational settings.

Acknowledgments

The authors would like to express their appreciation to all individuals whose valuable comments and suggestions contributed to the improvement of this research.

Funding

The authors declare that no financial support, grant, or sponsorship was received for conducting this study.

Data Availability

The data used and analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Berman, O. and Kim, E. (1999). Dynamic control of a queueing system with inventory. *Queueing Systems*, 33:195–217.

[2] Berman, O., & Sapna, K. P. (2000). Inventory management at service facilities for systems with arbitrarily distributed service times. Communications in Statistics—Stochastic Models, 16(3–4), 343–360.https://doi.org/10.1080/15326340008807592

[3] Bijvank, M., & Vis, I. F. (2011). Lost-sales inventory theory: A review. *European journal of operational research, 215*(1), 1-13.https://doi.org/10.1016/j.ejor.2011.02.004

[4] Chakravarthy, S., & Alfa, A. S. (Eds.). (1996). Matrix-analytic methods in stochastic models. CRC Press.https://api.pageplace.de/preview/DT0400.9781482292176_A24098239/preview-9781482292176_A24098239.pdf

[5] He, Q. M., & Jewkes, E. M. (2000). Analysis of a queueing inventory system with random demand.

European journal of operational research, 125(1), 109–121.

[6] He, Q. M., Jewkes, E. M., and Buzacott, J. A. (2002). Analysis of a continuous review inventory system with controllable lead time. *Operations Research*, 50, 658–674.

[7] Kim, E. (2005). Optimal control of a queueing system with inventory and lost sales. *European Journal of Operational Research*, 167, 336–350.

[8] Krishnamoorthy, A., Ramaswami, V., and Joshua, V. C. (2021). Queueing inventory models: Recent developments. *Queueing Systems*, 97,1–36.

[9] Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. Society for Industrial and Applied Mathematics. https://epubs.siam.org/doi/pdf/10.1137/1.9780898719734.bm

[10] Melikov, A., Mirzayev, R., & Nair, S. S. (2022). Double sources queuing-inventory system with hybrid replenishment policy. *Mathematics, 10*(14), 2423. https://doi.org/10.3390/math10142423

[11] Minner, S. (2003). Multiple-supplier inventory models in supply chain management: A review.

international journal of production economics, 81, 265-279.