2025. Volume 9, Number 1 Pages 19-25

UOT: 519.872
DOI: https://doi.org/10.30546/09090.2025.210.019

OPTIMAL COST ANALYSIS IN A DOUBLE-SOURCE QUEUING-INVENTORY SYSTEM UNDER HYBRID REPLENISHMENT POLICY

RAMIL MIRZAYEV
Baku Engineering University,
mirzayev.ramil@gmail.com
Baku, Azerbaijan

ARTICLE INFO

Article history:
Received:2025-07-07
Received in revised form:2025-07-07
Accepted:2025-07-10
Available online

Keywords:
Queuing-inventory systems;
Hybrid replenishment policy;
Stochastic modelling;
Cost optimization;
Markov chains
2010 Mathematics Subject Classifications: 60K25, 90B22, 90B05, 90C40

ABSTRACT

This paper investigates the optimization of total expected cost (TEC) in a finite queuing-inventory system (QIS) managed under a hybrid replenishment policy with two distinct supply sources. The system incorporates finite buffer and inventory capacity, stochastic arrivals of both service-demanding and destructive customers, and probabilistic customer behavior. The primary challenge addressed is balancing service quality and inventory efficiency under uncertainty and capacity constraints. To model the system, a two-dimensional continuous-time Markov chain is developed, capturing the dynamics of queue length and inventory levels. Performance measures such as inventory level, customer loss probability, and replenishment rates are derived using matrix-analytic methods. A cost function combining replenishment, holding, penalty, damage, and waiting costs is formulated, and the optimal reorder thresholds are identified through numerical experiments. The results demonstrate that the hybrid policy can significantly reduce operational costs and improve service efficiency when the thresholds are properly selected. This study provides both theoretical insights and practical tools for managing inventory-service systems with dual sourcing strategies.

1. Introduction

Managing service systems that rely on inventory presents a core challenge in operations research, particularly when these systems have finite capacity, stochastic arrivals, and non-identical supply sources. Traditional queuing models often oversimplify by assuming infinite inventory or single-source resupply, which is inadequate for applications such as spare-parts logistics, emergency medical stock management, or just-in-time production [1,2].

This research study a finite-buffer queuing-inventory system (QIS) with a single server and two distinct supply sources. Customers are categorized as consumer customers, who require inventory items for service, and destructive customers, who deplete stock without receiving service. Replenishment follows a hybrid, threshold-based policy: when stock hits the reorder level s, a regular order is placed to the slow, low-cost supplier; if inventory further falls to the critical level $r < s$, that order is cancelled and an emergency order is placed with the fast, high-cost supplier. This policy mirrors systems analyzed in prior dual-sourcing models [3–5].

We model the system using a two-dimensional continuous-time Markov chain (2DMC), capturing both queue length and inventory level. Transition rates are explicitly derived, and steady-state probabilities are calculated using matrix-analytic techniques. This extends analyses from similar dual-source QIS frameworks [1,2].

The study then formulates a control problem to identify optimal threshold pairs (s, r) that minimize total expected cost (TEC). This cost function includes replenishment, holding, penalty, and cancellation components and follows optimization methodologies from earlier QIS studies [1,5–7].

Finally, numerical experiments assess how changes in system parameters – such as arrival rate, buffer size, and thresholds – affect performance metrics and costs. The findings not only demonstrate the benefits of hybrid replenishment but also offer practical guidance for system design under realistic uncertainties.

2. Model Description and Analytical Framework

This study investigates a double-source, single-server, finite QIS with restricted buffer capacity $N,\ N < \infty$ and maximal store capacity $S, S < \infty$. The model is designed to jointly manage stochastic customer arrivals and inventory replenishment dynamics under capacity and responsiveness constraints. The inventory is essential for service provision and is replenished via a hybrid policy involving two heterogeneous suppliers, see [8].

The system supports two types of customers:

- Consumer customers (c-customers), arriving according to a Poisson process with rate λ, each requesting one item of inventory for service.
- Destructive customers (d-customers), arriving independently with rate κ, each removing one inventory item upon arrival.

If inventory is available and the server is idle, a c-customer is immediately started for service. If the server is busy but inventory is still available, the c-customer joins a finite buffer. If no inventory is present upon arrival, the customer joins the queue with probability φ_1, or leaves with probability $\varphi_2 = 1 - \varphi_1$. Customers in the buffer may abandon the system after an exponentially distributed patience time τ^{-1} if inventory remains unavailable.

After service, a c-customer may accept the item (with probability σ_2) or don't accept it (with probability $\sigma_1 = 1 - \sigma_2$). The corresponding service times are exponentially distributed with rates μ_2 and μ_1, respectively.

Inventory replenishment follows a hybrid policy:

- When inventory drops to a threshold $s, s < S/2$, an order of size $Q = S - s$ is placed to Source-1 (slow, cheaper).
- If inventory reaches a more critical level $r, r < s$, the order from Source-1 is canceled (with penalty), and a new "Up-to-S" order is placed to Source-2 (fast, expensive).
- Both Source-1 and Source-2 exhibit exponentially distributed lead times, with average delivery delays of ν_1^{-1} and ν_2^{-1}, respectively. Since $\nu_2 > \nu_1$, Source-2 provides quicker deliveries. Nevertheless, this increased speed comes at a higher cost, creating a balance that must be considered between cost-effectiveness and responsiveness.

This behavior creates a Markovian system modeled as a 2DMC with states (n, m), where n is the number of c-customers in the system, and m is the current inventory level. The state space is

$E = \{0,1,\dots,N\} \times \{0,1,\dots,S\}$. The transition rates from a given state $(n_1, m_1) \in E$ to another state $(n_2, m_2) \in E$ are represented by $q\big((n_1, m_1); (n_2, m_2)\big)$. The specific transition intensities for state (n_1, m_1) are given below, see [8]:

$$q_1\big((n_1, m_1); (n_1 + 1, m_1)\big) = \lambda\varphi_1 \cdot I(n_1 < N) \cdot I(m_1 = 0);$$
$$q_1\big((n_1, m_1); (n_1 + 1, m_1)\big) = \lambda \cdot I(n_1 < N) \cdot I(m_1 > 0);$$
$$q_1\big((n_1, m_1); (n_1 - 1, m_1)\big) = \mu_1\sigma_1 \cdot I(m_1 > 0);$$
$$q_1\big((n_1, m_1); (n_1 - 1, m_1 - 1)\big) = \mu_2\sigma_2 \cdot I(m_1 > 0);$$
$$q_1\big((n_1, m_1); (n_1, m_1 - 1)\big) = \kappa \cdot I(m_1 > 0);$$
$$q_1\big((n_1, m_1); (n_1 - 1, m_1)\big) = \tau \cdot I(n_1 > 0) \cdot I(m_1 = 0);$$
$$q_1\big((n_1, m_1); (n_1, m_1 + S - s)\big) = \nu_1 \cdot I(r < m_1 \le s);$$
$$q_1\big((n_1, m_1); (n_1, S)\big) = \nu_2 \cdot I(0 \le m_1 \le r).$$

Here, the notation $I(A)$ refers to the indicator function, which takes the value 1 when the condition A holds true, and 0 otherwise.

The steady-state probabilities $p(n,m)$ are satisfied the system of balance equations (SBE). Both exact and approximate methods to solve the SBE are developed in [8]. In that paper the following explicit formulas are proposed to calculate the key performance measures:

- Average inventory level (S_{av})

$$S_{av} = \sum_{m=1}^{S} m \sum_{n=0}^{N} p(n,m) \approx \sum_{m=1}^{S} m\pi(< m >)\,; \; (1)$$

Average supply from Source-i, $i = 1, 2$, $\big(V_{av}(i)\big)$

$$V_{av}(1) = (S - s) \sum_{m=r+1}^{s} \sum_{n=0}^{N} p(n,m) \approx (S - s) \sum_{m=r+1}^{s} \pi(< m >)\,; \; (2)$$

$$V_{av}(2) = \sum_{m=0}^{r} (S - m) \sum_{n=0}^{N} p(n,m) \approx \sum_{m=0}^{r} (S - m)\pi(< m >)\,; \; (3)$$

- Average number of c-customers in system (L_{av})

$$L_{av} = \sum_{n=1}^{N} n \sum_{m=0}^{S} p(n,m) \approx \sum_{n=1}^{N} n \sum_{m=0}^{S} \rho_m(n)\pi(< m >)\,; \; (4)$$

- Average damaging rate of stocks (DRS):

$$DRS = \kappa\left(1 - \sum_{n=0}^{N} p(n,0)\right) \approx \kappa\big(1 - \pi(< 0 >)\big); \; (5)$$

- Average reorder rate from Source-1 (RR_1):

$$RR_1 = \kappa p(0, s+1) + (\mu_2\sigma_2 + \kappa) \sum_{n=1}^{N} p(n, s+1) \approx \pi(s+1)\Big(\kappa + \mu_2\sigma_2\big(1 - \rho(0)\big)\Big)\,; \; (6)$$

- Average reorder rate from Source-2 (RR_2):

$$RR_2 = \kappa p(0, r+1) + (\mu_2\sigma_2 + \kappa)\sum_{n=1}^{N} p(n, r+1) \approx \pi(<r+1>)\left(\kappa + \mu_2\sigma_2\big(1-\rho(0)\big)\right); (7)$$

- Loss probability of *c*-customers (PL):

$$PL = \varphi_2 \sum_{n=0}^{N} p(n,0) + \sum_{m=0}^{S} p(N,m) + \frac{\tau}{\tau + \lambda\varphi_1 + \nu_2}\sum_{n=1}^{N-1} p(n,0) + \frac{\tau}{\tau+\nu_2} p(N,0) \approx$$

$$\approx \varphi_2\pi(<0>) + \sum_{m=0}^{S} \rho_m(n)\pi(<m>) + \frac{\tau}{\tau + \lambda\varphi_1 + \nu_2}\pi(<0>)\big(1-\rho(0)\big) +$$

$$+\frac{\tau}{\tau+\nu_2}\rho_0(N)\pi(<0>). (8)$$

This allows solve the minimization of the TEC associated with operating the QIS.

3. Optimization of Total Expected Cost

As a natural extension of the stochastic modeling and performance analysis discussed in [8], we now focus on optimizing the TEC. The cost function incorporates both service and inventory considerations and is formulated as follows:

$$\text{TEC}(s,r) = \sum_{i=1}^{2}\big(K_i + c_r(i)V_{av}(i)\big)RR_i + c_c RR_2 + c_h S_{av} + c_d DRS + c_l\lambda PL + c_w L_{av}, (9)$$

where:

- K_i is the fixed cost of placing an order from Source-i ($i = 1,2$),
- $c_r(i)$ is the unit inventory cost from Source-i ($i = 1,2$),
- $V_{av}(i)$ is the average ordered volume from Source-i ($i = 1,2$),
- RR_i is the average number of replenishment requests from Source-i ($i = 1,2$) per unit time,
- c_c is the penalty cost for canceling an order from Source-1,
- c_h is the holding cost per unit inventory per unit time,
- c_d is the damage cost per unit inventory,
- DRS is the average number of inventory items destroyed by d-customers per unit time,
- c_l is the cost of losing a c-customer,
- PL is the probability of customer loss due to inventory shortage,
- c_w is the cost of customer waiting per unit time,
- L_{av} is the average number of waiting customers.

We assume that both warehouse and buffer capacity as well as loading parameters of the QIS are fixed parameters. So, the decision variables in this cost minimization problem are the thresholds s and r. The optimal pair (s^*, r^*) is defined as the solution to the following minimization problem:

$$(s^*, r^*) = arg \min_{(s,r)\in X} TC(s,r) . (10)$$

where X is the set of all admissible threshold pairs (s, r) satisfying $0 \leq r < s < S/2$.

This optimization framework enables practitioners to tune replenishment strategies based on a cost-performance trade-off, balancing inventory responsiveness, service reliability, and operational cost under uncertainty.

4. Numerical Experiments

A series of numerical experiments were conducted to evaluate the performance of the proposed model and to investigate the impact of various parameters on the system's behavior. A primary objective of these experiments was the minimization of the TEC, a crucial metric for evaluating the efficiency of the supply chain.

For these experiments, we focus on optimizing the controllable parameters, specifically s and r. The core of the optimization problem is to identify the optimal pair of values (s, r) that minimizes the functional presented in (10).

Table 1 and Table 2 detail the results of these numerical experiments, providing insights into the optimal operational strategies for minimizing TEC. Here, the values of the system parameters were taken as follows:

$$S = 24, N = 50, s = 10, \lambda = 20, \mu_1 = 35, \mu_2 = 15, \sigma_1 = 0.4, \sigma_2 = 0.6,$$
$$s = 10, \varphi_1 = 0.7, \varphi_2 = 0.3, \upsilon_1 = 0.5, \upsilon_2 = 1, \tau = 2, \kappa = 1.$$

Experiment 1. Let's take the cost parameters as follows:

$$K_1 = 100, K_2 = 200, \ c_r(1) = 50, c_r(2) = 100, c_c = 50, c_h = 35, c_d = 75, c_l = 200, c_w = 50.$$

Using the given parameter values, the total cost values $\text{TEC}(s, r)$ are computed according to formula (9). The results are presented in Table 1 for different values of the reorder level r and the review level s.

Table 1. Results of Formula (9)

		s									
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
r	*0*	4300.11	4273.44	4249.78	4229.05	4211.16	4196.05	4183.67	4174.02	4167.07	4162.85
	1		4227.37	4203.75	4182.99	4165.00	4149.73	4137.12	4127.13	4119.76	4115.00
	2			4160.20	4139.50	4121.51	4106.15	4093.38	4083.13	4075.41	4070.19
	3				4098.35	4080.45	4065.11	4052.25	4041.84	4033.84	4028.24
	4					4041.57	4026.35	4013.52	4003.03	3994.86	3988.98
	5						3989.61	3976.94	3966.51	3958.27	3952.20
	6							3942.23	3931.99	3923.83	3917.71
	7								3899.20	3891.27	3885.24
	8									3860.28	3854.54
	9										**3825.26**

From Table 1, it is evident that the TEC decreases monotonically as both s and r increase within the admissible range. The minimal TEC value in the table is:

$$min\ \text{TEC} = 3825.262 \text{ achieved at } (s^*, r^*) = (10, 9).$$

This indicates that the optimal reorder thresholds s^* and r^* which minimize the TEC are $s^* = 10$ and $r^* = 9$. The minimal value of the TEC is made bold in Table 1.

Experiment 2. For this experiment, to illustrate the effect of parameter variation, the cost and system parameters were modified as follows:

$$\lambda = 8, \kappa = 2, \mu_1 = 45, \mu_2 = 15, \tau = 1.5, \nu_1 = 2, \nu_2 = 8,$$

$$K_1 = 150, K_2 = 250,\ c_r(1) = 60, c_r(2) = 120, c_c = 60, c_h = 40, c_d = 90, c_l = 220, c_w = 55.$$

The results of the formula (9) are showed in Table 2 for $s = \overline{1,10}$ and $r = \overline{0, s-1}$.

Table 2. Results of Formula (9)

		s									
		1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***
r	***0***	871.26	864.94	863.51	866.49	873.41	883.87	897.49	913.96	933.04	954.57
	1		837.22	844.44	853.09	863.79	876.78	892.10	909.73	929.61	951.69
	2			**834.09**	847.71	861.18	875.65	891.70	909.62	929.54	951.53
	3				844.86	861.73	877.71	894.25	912.10	931.67	953.20
	4					861.97	880.73	898.24	916.14	935.28	956.17
	5						882.06	902.14	920.80	939.81	960.11
	6							903.64	924.87	944.57	964.64
	7								926.09	948.45	969.22
	8									949.17	972.73
	9										972.80

According to the values presented in Table 2, the lowest total cost recorded in the table is:

$$min\text{TEC} = 834.09$$

which occurs at the reorder and review levels $(s^*, r^*) = (3, 2)$. This optimal value of the TEC is bolded in the table for clarity.

5. Conclusion

This study presents a comprehensive framework for optimizing the TEC in finite QISs operating under a hybrid replenishment policy with dual supply sources. By integrating a 2DMC model with a detailed cost function, we analyze both the performance and economic efficiency of systems where customer demand, inventory depletion, and replenishment dynamics interact under uncertainty and capacity constraints. The proposed model successfully captures key operational characteristics, such as inventory-dependent service capability, probabilistic customer behavior, and the trade-offs between lead time and replenishment cost.

Overall, this work not only advances the theoretical understanding of hybrid QIS models but also provides actionable insights for decision-makers aiming to balance cost efficiency with service quality in complex inventory-service systems. Future research may extend this model to incorporate more dynamic replenishment strategies, partial deliveries, or real-time data-driven control mechanisms.

REFERENCE LIST

1. Melikov, A., Mirzayev, R., & Sztrik, J. (2023). Double sources queuing inventory systems with finite waiting room and destructible stocks. *Mathematics*, *11*(1), 226.
2. Melikov, A., Mirzayev, R., & Nair, S. (2022). Double sources queuing inventory system with hybrid replenishment policy. *Mathematics*, *10*(14), 2423.
3. Lawrence Arputham, S., Melikov, A., & Sivakumar, B. (2023). Analysis and optimization of hybrid replenishment policy in a double sources queueing-inventory system with MAP arrivals. *Annals of Operations Research*, *331*(2), 1249–1267.
4. Melikov, A, Shahmaliyev, M., & Nair, S. (2021). Matrix-geometric method to study queuing system with perishable inventory. *Automation and Remote Control*, *82*, 2168–2181.
5. Xin, L., & Goldberg, D. A. (2017). Asymptotic optimality of tailored base-surge policies in dual-sourcing inventory systems. *Management Science*, *64*(1), 437–452. https://doi.org/10.1287/mnsc.2016.2607
6. Xin, L., & Van Mieghem, J. A. (2021). Dual-sourcing, dual-mode dynamic stochastic inventory models: A review. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3885147
7. Song, J.-S., Xiao, L., Zhang, H., & Zipkin, P. (2017). Optimal policies for a dual-sourcing inventory problem with endogenous stochastic lead times. *Operations Research*, *65*(2), 379–395.
8. Mirzayev, R., & Aliyeva, S. (2025). Modeling and optimization of finite buffer queuing-inventory systems with hybrid replenishment. In *Proceedings of the International Conference on Management and Control in Solving Engineering Problems (MaCoSEP* 25*)*. Baku, Azerbaijan.