# A Survey on Queuing Inventory Models with Server Vacations and Retrials

**Beena P[1], Sindu Mathew P[2]**

[1]Associate Professor, Department of Mathematics, Govt. Engineering College, Thrissur
[2]Assistant Professor, Department of Mathematics, Govt. Engineering College, Thrissur

**Abstract**

This paper presents a comprehensive survey of queueing inventory models that incorporate server vacations and retrials, focusing on their applications, methodologies, and performance metrics. These models address complex real-world systems where service mechanisms are influenced by inventory constraints, server availability, and customer retry behavior. Key features such as vacation policies, retrial queues, and inventory replenishment strategies are examined in detail. The survey highlights the interplay between these components, exploring their impact on system performance, including queue lengths, waiting times, service rates, and inventory costs.

**Keywords:** Queuing Inventory Model, Server Vacations, Retrials

## 1. Introduction

The mathematical study of queues is the focus of queuing theory. Queues are prevalent in numerous places, such as banks, supermarkets, hospitals, gas stations, computer systems, etc. A.K. Erlang's seminal paper on telephone calls marked the beginning of the theory of queueing systems. After that, Kendall (1951 and 1953) created a queuing theory from the stochastic process perspective.

Inventory, sometimes referred to as stocks, is essentially the financially valuable products and raw materials that any corporation would keep on hand and that are either ready for sale or soon will be. A mathematical tool called the inventory model aids businesses in figuring out how much inventory is best to keep on hand while manufacturing is underway. It is also highly helpful for controlling the number of orders, figuring out how much inventory of items or raw materials should be kept on hand, and monitoring the flow so that consumers may receive uninterrupted service.

The queueing system with vacations has been thoroughly researched and has real-world uses in computer network design, data communication systems, and industrial inventory models, among other areas. On a queueing system, the period that the servers are unavailable to deliver the service is known as the server vacation time. Server outages, a lack of work, or any secondary tasks given to the servers may cause them to take vacations. Allowing servers to take vacation time will increase an organization's revenue because the time off can be used for other tasks.

The ability of the server to perform secondary activities during idle time is another benefit of the server vacation queuing mechanism. Because of their usefulness in a variety of domains, queueing inventory models with server vacations have attracted a lot of attention in recent years. Inventory systems with server vacations have only been the subject of a small number of reports thus far.

Levy and Yechiali [1] have investigated the methodology of queueing systems with one or more vacations.

It appears that research on vacation in queuing models started in the early 1970s. Outstanding survey studies on vacation models were proposed by Teghem [2] and Doshi [3]. See the book by Takagi [4] and Tian [5] for a thorough analysis of vacation models. The servers are busy when new clients arrive, so they leave the service area to join orbiting retrial queues, which are groups of disgruntled customers.

In this paper, analytical techniques and solution approaches, including stochastic processes, queueing theory, and simulation, are reviewed, emphasizing challenges in deriving practical insights. This work also identifies gaps in the existing literature, suggesting directions for future research, particularly in optimizing system performance under varying stochastic conditions and integrating emerging technologies like machine learning for dynamic decision-making. The findings provide valuable insights for researchers and practitioners aiming to design efficient and resilient service systems.

## 2. Preliminaries

Advanced queueing models encompass several critical concepts that optimize system performance and resource management. Multi-server queueing models involve multiple parallel servers handling tasks or customer requests, with key performance metrics such as system throughput, waiting times, server utilization, and queue lengths. These models are often framed using structures like M/M/c (Poisson arrivals and exponential service times with cc servers) or M/G/c (general service time distributions with c servers). Inventory models integrate inventory management with queueing systems to address resource constraints like spare parts or consumables. Policies such as (s, S), where restocking occurs when inventory drops below s up to S, and (R,Q), involving periodic review with a fixed order quantity Q, are crucial in influencing throughput, reliability, and waiting times.

Server vacations, designed to optimize server usage or perform maintenance, follow various policies: single vacations (returning immediately if tasks are queued), multiple vacations (taking successive breaks until tasks appear), or working vacations, where servers remain partially active at reduced service rates. Finally, retrials address scenarios where customers retry after finding all servers busy instead of abandoning the system. Retrial customers wait in an orbit before reattempting service, a characteristic feature in applications such as call centers, telecommunications networks, and shared computing systems. This topic centers around multi-server queuing inventory models incorporating server vacations and retrials, which are complex systems commonly encountered in areas like telecommunications, manufacturing, supply chain management, and service industries.

## 3. Literature Review

Both a multiple server vacation with Markovian demand and an inventory system with trial demands were taken into consideration by Sivakumar [6]. In the steady state situation, the joint probability distribution of the number of consumers in the orbit and the inventory level is calculated. The long-term total projected cost rate is computed, together with several system performance metrics in the steady state. Numerous numerical examples that shed light on the system's behaviour are provided.

In Jayaraman et.al. [7] the Poisson process governs the unit demands. To restock, the (s, S) ordering policy is used. The server takes a random vacation when the inventory is totally exhausted. If not restocked at the end of the vacation, the server begins their vacation anew. Customers that visit during server vacations or stock-out periods are given the option to leave the system or join a pool with a limited capacity. Only when the inventory level is higher than s does the server choose each demand in the pool one at a time. The duration between any two consecutive selections is distributed exponentially, with the

parameter based on the number of customers in the pool. The steady-state of the joint probability distribution of the number of customers in the pool and the inventory level is obtained. The long-term total projected cost rate is computed, together with a number of system performance metrics in the steady state.
Padmavathi et al. [8] considered a retrial inventory system with single and modified multiple vacation for the server. They examined an exponentially distributed delivery time and Poisson demand continuous review stochastic (s, S) inventory system. They looked at two models that differed in how servers take their vacations. In the steady state scenario, the joint probability distribution of the server status, the number of demands in the orbit, and the inventory level is determined. The long-term total projected cost rate is computed, together with a number of system performance metrics in the steady state.

Koroliuk et al. [9] suggest a queueing-inventory system model where service times are positive random numbers and there is perishable inventory and server vacations. If the inventory is at zero, the queue is empty, or both occur, the server takes a vacation. If the inventory level is not 0 at the end of the vacation, the server immediately begins answering calls; if not, a new vacation is taken. Restocking adheres to the two-level approach, and calls in the line are impatient. The asymptotic system analysis approach is created, and numerical experiment results are displayed.

In [10], Vijayalakshmi et.al examined a service facility's continuous review (s, S) inventory system, which issues a customer-requested item after servicing it. Primary arrivals can be either positive or negative before entering the orbit, and they follow a Poisson distribution. The distribution of replenishment, vacation, and service times throughout usual busy periods and vacation periods is exponential. For the model's steady-state distribution, a matrix analytic approach is employed.

Koroliuk [11] examined the perishable queueing-inventory system model with server vacations. If there are no customers in the queue when the service is finished, the server takes a vacation; if there are more consumers in the system than a certain threshold, the server begins service at the conclusion of the vacation; if not, it takes another vacation. The parameters of the system inventory level are calculated using both precise and approximative approaches. The conditional distributions of the inventory level are determined for both the server's operational and on-time states, as well as for the server's vacation-related shutdowns. The authors derive a few system performance metrics using the stationary distribution. The impact of the server's vacation on the performance metrics is examined analytically by the writers.

In [12], Kathiresan et.al. examined a continuous review inventory system that had two supply modes for replenishment, one with a shorter lead time, a single server, several vacations, Poisson demand, retrial demand, and exponentially distributed lead time. Gaver's approach is applied to determine the system's stationary distribution. Several cost minimization numerical methods are employed after calculating different system performance metrics.

Senthilkumar [13] examined a continuous review inventory system that has two commodities and a limited number of homogeneous sources that produce demand. A joint reordering policy for placing orders is adopted. It is assumed that the lead time for order delivery follows an exponential distribution and the commodities are interchangeable. It means that a unit of one commodity can satisfy the demand for another when there is no stock of either. Various system performance measures in the steady state are derived

To examine a stock-dependent client arrival process and two distinct server tasks, Sugapriya et. al. [14] examine a continuous review inventory. A maximum of S objects can be stored in this system. There are two modes of server availability: vacation mode and regular mode. The server takes a vacation if the stock level is zero. Only if there is a positive inventory at the end of the vacation will the server return; if not, he will take another vacation. The number of clients in the orbit and the current stock level both affect the retrial procedure for a customer. The steady-state joint distribution with its components, orbit size, server status, and inventory level at any given time is generated using the matrix geometric technique. The nature of the anticipated total cost and other system functions, such as waiting time, are examined in steady-state. A numerical evaluation is also done.

In Zhang et.al. [15] a queueing-inventory system with a random order size policy, lost sales, and ongoing review is examined. A replenishment order is immediately initiated if the inventory is exhausted following a customer's service. A discrete probability distribution may be used to randomly determine the size of the replenishment order. Customers need server service and enter the system based on a Poisson process. After the stock runs out, the server takes several vacations. It is assumed that the lead time, vacation time, and service time are all distributed exponentially. In explicit product form, they obtained the stationary joint distribution of the server status, the on-hand inventory level, and the queue length. Additionally, conditional distributions of the level of on-hand inventory are produced for both when the server is on and functioning and when it is off because of a vacation or depleted inventory. They also compute a few system performance metrics. Analytical research is done on how the server's vacation affects the performance metrics. Lastly, a few numerical findings are shown. The model is simulated and examined in relation to broader arrival processes and service time distributions.

In [16], a multi-server queueing-inventory system's asynchronous server vacation and customer retrial features are examined by Jaganathan et.al. Each server begins a vacation if, following a busy period, they discover that there are not enough clients and objects in the system. The server enters an inactive state when its vacation is ended and it realizes there is no possibility of it becoming busy; if not, it will take another vacation. An estimated overall system cost will be created and numerically integrated with the parameters following the computation of adequate system performance metrics

Yue et al. [17] examined an M/M/1 queuing system with an associated inventory that is subject to a (s, S) control policy. Every time the inventory runs out, the server takes several vacations. Both the lead time and the vacation time are assumed to have exponential distributions. The stability requirement of the system is derived by the authors, who formulate the model as a quasi-birth-and-dearth (QBD) process. The stationary distribution in product form is then obtained for the joint process of the server's status, inventory level, and queue length. Additionally, the conditional distributions of the inventory level are determined for both the server's operational and on-time states, as well as for the server's vacation-related shutdowns. The authors derive a few system performance metrics using the stationary distribution.

## 4. Applications

Queueing models with vacation and retrial mechanisms find widespread applications in various fields that require efficient resource utilization and customer satisfaction. In telecommunications networks, these models manage call centers and data traffic, where servers may take vacations during low-demand periods or handle reduced workloads while idle, ensuring energy efficiency and maintenance without service disruptions. Retrial mechanisms are critical in handling busy signals, as customers are queued in a virtual orbit for retries, reducing call drops and improving accessibility.

In manufacturing and logistics, vacation policies enable machinery maintenance or downtime optimization, while retrials ensure tasks like reprocessing or reattempted deliveries are systematically managed. Similarly, in healthcare systems, these models help manage patient flow, allowing medical staff to take breaks or attend to administrative duties, while patients unable to get immediate service can retry after a delay, improving service equity. Additionally, shared computing systems, such as cloud platforms, leverage these models to manage server workloads dynamically, ensuring efficient handling of retry requests for computing resources during peak demand or maintenance periods. These applications demonstrate how integrating vacations and retrials enhances system resilience, efficiency, and customer experience.

## 5. Challenges and Limitations

Queueing models with vacation and retrial mechanisms, while versatile, face several challenges and limitations. One primary issue is the increased complexity in system analysis and modeling, as the combination of vacations and retrials introduces intricate interactions between server states and customer behavior. Accurately estimating parameters such as retrial rates, vacation durations, and service times can be difficult, particularly in dynamic or real-time environments. Another challenge lies in balancing system efficiency with customer satisfaction, as excessive vacation periods or high retrial delays can lead to longer wait times and decreased throughput. Computational costs for simulating and optimizing these models also grow significantly with larger systems or heterogeneous server configurations.

Additionally, retrial queues can lead to resource contention in high-traffic scenarios, where the orbit size grows uncontrollably, potentially overwhelming the system. Practical implementation challenges include adapting the model assumptions—such as exponentially distributed interarrival or retrial times—to real-world systems, which often exhibit variability and dependencies not captured by classical frameworks. These limitations underscore the need for robust modeling techniques and adaptive strategies to ensure the practical viability of such systems.

## 6. Scope for further research

The scope for further research on queueing models with vacation and retrial mechanisms is vast, driven by the need to address real-world complexities and enhance model applicability. Future work could explore the incorporation of machine learning and data-driven approaches to predict and optimize retrial and vacation parameters dynamically, adapting to changing demand patterns. Developing models that account for more general distributions, customer priorities, and correlated arrivals or retrials could improve the realism of these systems.

Research on energy-efficient vacation policies, particularly in green computing and telecommunications, offers opportunities to minimize energy consumption without compromising service quality. There is also a need to study the impact of retrial behavior in systems with limited or stochastic capacity, such as cloud computing or healthcare networks, where resources fluctuate. Furthermore, hybrid models integrating queueing with other operational paradigms, like inventory systems or transportation networks, could provide comprehensive solutions for complex environments. Lastly, the development of scalable algorithms for analyzing and simulating large-scale systems with heterogeneous servers and multi-class customers remains a critical area for advancing theoretical and practical insights.

## References

1. Y. Levy and U. Yechiali. “An m/m/s queue with servers’ vacations”, INFOR: Information Systems and Operational Research, 14(2):153–163, 1976.
2. J. Teghem Jr. “Control of the service process in a queueing system”. European Journal of Operational Research, 23(2):141–158, 1986.
3. Doshi, B. T. (1986). “Queueing systems with vacations: A survey. Queueing Systems”, 1, 29–66.

4. Takagi, H. (1993). “Queueing analysis—a foundation of performance evaluation” (Vol. 3). Amsterdam: Elsevier
5. Tian, N., & Zhang, Z. G. (2006). “Vacation queueing models—theory and application”. Berlin: Springer.
6. Sivakumar, B. (2011). “An Inventory System with Retrial Demands and Multiple Server Vacation”. Quality Technology & Quantitative Management, 8(2), 125–146. https://doi.org/10.1080/16843703.2011.11673252
7. B. Jayaraman, B. Sivakumar, G. Arivarignan, “A perishable inventory system with postponed demands and multiple server vacations,” in: Modeling and Simulation in Engineering, Hindawi Publishing Corporation, Vol. 2012, Article ID 620960.
8. Padmavathi, I., Sivakumar, B. & Arivarignan, G. “A retrial inventory system with single and modified multiple vacation for server”. Ann Oper Res 233, 335–364 (2015). https://doi.org/10.1007/s10479-013-1417-1
9. Koroliuk, V. S., Melikov, A. Z., Ponomarenko, L. A., & Rustamov, A. M. (2017) , “Asymptotic analysis of the system with server vacation and perishable inventory”, Cybernetics and Systems Analysis, 53, 543-553.
10. Vijaya Laxmi, P., & Soujanya, M. L. (2017). “Retrial inventory model with negative customers and multiple working vacations”. International Journal of Management Science and Engineering Management, 12(4), 237–244. https://doi.org/10.1080/17509653.2016.1233837
11. Koroliuk, V.S., Melikov, A.Z., Ponomarenko, L.A. et al. (2018), “Models of Perishable Queueing-Inventory Systems with Server Vacations”, Cybern Syst Anal 54, 31–44. https://doi.org/10.1007/s10559-018-0005-4
12. J. Kathiresan and N. Anbazhagan (2020), “An inventory system with retrial demands, multiple vacations and two supply modes”, Int. J. of Operational Research, Vol.37,No.4,pp 524-548.
13. P. Senthil Kumar (2021), “A Finite Source Two Commodity Inventory System with Retrial Demands and Multiple Server Vacation”, J. Phys.: Conf. Ser., DOI 10.1088/1742-6596/1850/1/012101.
14. Sugapriya, C., Nithya, M., Jeganathan, K., Anbazhagan, N., Joshi, G. P., Yang, E., & Seo, S. (2022). “Analysis of Stock-Dependent Arrival Process in a Retrial Stochastic Inventory System with Server Vacation. Processes”, 10(1), 176. https://doi.org/10.3390/pr10010176
15. Zhang, Y., Yue, D. & Yue, W. A (2022) “Queueing-inventory system with random order size policy and server vacations”, Ann Oper Res 310, 595–620 . https://doi.org/10.1007/s10479-020-03859-3
16. K. Jeganathan, T. Harikrishnan, K. Prasanna Lakshmi, D. Nagarajan, “A multi-server retrial queueing-inventory system with asynchronous multiple vacations”, Decision Analytics Journal, Volume 9, 2023, 100333, https://doi.org/10.1016/j.dajour.2023.100333
17. Yue, D., Zhang, Y., Xu, X. et al. (2024) “Product Form Solution of a Queuing-Inventory System with Lost Sales and Server Vacation”, J Syst Sci Complex 37, 729–758 (2024). https://doi.org/10.1007/s11424-024-1207-7