SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Queueing-Based Latency Minimization in Video Streaming Systems with Adaptive Service Control

S. P. Niranjan[1], K. Aswini [2*]

[1]Associate professor. Department of Mathematics, Vel Tech Rangarajan Dr. Sagunthala R&D institute of science and technology, Avadi, Chennai, Tamil Nadu, India. E-mail: niran.jayan092@gmail.com

[2] Research scholar, Department of Mathematics, Vel Tech Rangarajan Dr. Sagunthala R&D institute of science and technology, Avadi, Chennai, Tamil Nadu, India. E-mail: vtd1428@veltech.edu.in

Abstract
Maintaining low latency and uninterrupted playback is a major challenge in video streaming systems operating under varying traffic loads and server failures. A bulk service queueing model is developed in which requests arrive in batches according to a Poisson process and are served under a general service-time distribution. The server dynamically switches among regular, slow, and fast service modes according to queue congestion and failure conditions, while vacation and dormant states represent periods of low demand. The supplementary variable technique is used to derive the probability generating function of the queue length and obtain key performance measures, including mean queue length, waiting time, busy period, idle period, and server utilization. A threshold-based cost analysis is employed to examine the trade-off between system performance and resource utilization. Numerical results demonstrate that adaptive service control reduces latency and improves system responsiveness under changing workloads. The proposed framework provides useful guidance for the design of reliable and latency-sensitive video streaming systems.

Keywords: Video streaming latency, server failure, bulk queueing model, service rate fluctuation, supplementary variable technique, adaptive service rate, real-time buffering reduction.

1. Introduction

The rapid growth of video streaming services has increased the demand for efficient server systems capable of delivering uninterrupted content with minimal latency and buffering. In practical streaming environments, fluctuating traffic intensity and server failures can significantly affect service quality and overall system performance. Queueing theory provides an effective analytical framework for studying such stochastic service systems under uncertain operating conditions.

Bulk service queueing models have been extensively studied due to their applications in communication networks, cloud platforms, and multimedia service systems. Niranjan et al. [1] introduced a bulk retrial queueing model with state-dependent arrivals, Bernoulli feedback, and multiple vacations. Later, Niranjan et al. [2] extended this framework by incorporating server failures and vacation policies. Upadhyaya [3] presented a comprehensive overview of queueing systems with vacations, while Rani and Indhira [4] reviewed bulk queueing models with server breakdowns. Haridass and Nithya [5] analyzed the impact of vacation interruption due to server failures, and Chandrasekaran et al. [6] studied working vacation queueing models relevant to systems operating at a reduced service rate. Ahmed et al. [7] developed mathematical models for dynamic systems under uncertainty, Sharma and Khanna [8] presented theoretical results using frame theory, while Wang and Meng [9] analyzed stochastic dynamics with distributed delay. These studies demonstrate the effectiveness of advanced mathematical techniques for analyzing complex dynamic systems. Further contributions include Sasikala et al. [10], who analyzed bulk service queues with setup time, multiple vacations, and server breakdowns, and Jeyakumar and Senthilnathan [12,13], who examined working vacation queueing systems with breakdown mechanisms. Pradhan and Karan [15], Sultan et al. [16], and Arumuganathan and Ramaswami [22] further explored adaptive service structures under varying operational conditions. Agrawal et al. [27] studied an M/M/1 queue with minor and major server breakdowns during regular and working vacation periods, while Seenivasan and Chandralekha [28], Seeniraj et al. [29], and Gautam et al. [30] analyzed vacation-based stochastic queueing systems using supplementary variable and generating function approaches.

However, most existing studies assume fixed or limited-service adaptation and do not adequately represent systems with dynamic transitions among multiple operational states under congestion and failure conditions. The integration of regular, slow, fast, vacation, and dormant service modes within a unified bulk

queueing framework remains relatively unexplored, particularly for latency-sensitive streaming applications.

Motivated by this gap, the present study develops a stochastic bulk service queueing model for video streaming systems with dynamically varying service rates governed by queue conditions and server reliability. Using the supplementary variable technique, steady-state performance measures and a threshold-based cost optimization framework are derived to support efficient latency-aware streaming system design.

2. MODEL MOTIVATION, ASSUMPTIONS AND APPLICATION

2.1 Research Motivation and Gap

Most existing queueing models assume fixed service rates and do not account for real-time server behavior changes due to failures or demand fluctuations. They lack the ability to model partial server failures where the system continues to operate at reduced capacity. Additionally, previous studies do not incorporate dynamic switching between multiple service states based on real-time conditions. There is also little attention given to cost-performance trade-offs in latency-sensitive applications like video streaming. This paper addresses these gaps by introducing a comprehensive, adaptive queueing model that captures fluctuating service rates, server failure modes and performance cost balancing making it highly relevant for modern streaming platforms like Netflix**.**

2.2 Model Assumptions

The proposed model assumes that video streaming requests arrive in batches following a Poisson process, and service begins only when a minimum number of requests are accumulated, following the general bulk service rule. The server can operate in five different modes: regular, slow, fast, vacation and dormant. The service rate dynamically changes based on queue length and server condition. If the queue is long, the server speeds up (fast mode)'if there's a partial failure, it slows down but continues operating (slow mode). During very low demand, the server either takes a break(vacation) or shuts down temporarily(dormant). This dynamic switching allows the server to maintain performance under varying traffic and failure conditions. The model evaluates key performance metrics like queue length, waiting time, throughput and cost using mathematical tools like probability generating functions.

2.3 Practical Application to Video Streaming Systems

2.3.1 Fluctuating Service Rate

Fluctuation in service rate refers to variations in the speed at which the server processes requests based on system Conditions.

The server dynamically switches between regular, slow and fast service rates depending on queue length, server failures.

2.3.2 Example: Netflix video streaming platform.

In Netflix video streaming platform, if the queue length exceeds a threshold(N), the server automatically switches to fast service quickly to reduce the queue length. Even if the server partially fails, it slows down but continues to operate.

The proposed model is well-suited for real-time video streaming platforms such as Netflix, where maintaining smooth playback and minimizing latency are critical. For instance, under normal conditions, the server processes about 20 video requests per minute using regular service. During peak hours, when the queue length exceeds a threshold (eg.50 requests) the server switches to a fast mode and processes 40 requests every 30 seconds to quickly reduce congestion. In the case of a partial server failure, the system does not stop but continues operating in slow mode, processing only 10 requests per minute to maintain service continuity. When demand is very low (e.g., fewer than 5 requests), the server enters vacation or dormant mode to conserve energy. If a small number of requests arrive-say 5, below the minimum batch threshold-it waits until enough requests accumulate before resuming service. This adaptive mechanism helps reduce user witing time, avoids buffering, maintains continuous streaming even during failures, and optimizes energy usage while preserving the desired quality of service (QoS).

3. MODEL DESCRIPTION

This model examines batch size queueing system where server failure cause various service rates. Arrivals follow a Poisson process and service follows general bulk service rules.

Depending on the queue length and server's circumstances, the server alternates between regular service, slow service, fast service, vacation and dormant period.

One important aspect of this approach is its capacity to optimize performance and reduce waiting time by switching to slower service during failures and fast service during periods of high demand. Using supplementary variable technique, steady-state analysis and various performance metrics including expected queue length and busy period are obtained. Because it minimizes waiting time of customers, this

technology is especially useful in real-time service contexts, such as a cloud platforms and telephone networks.

3.1 GENERAL BULK SERVICE RULE

The general bulk service rule was first introduced by Neuts. "The general bulk service rule states that the server will start to provide service only when at least" customers are present in the queue, and the maximum service capacity is 'b' (b > a**)".**

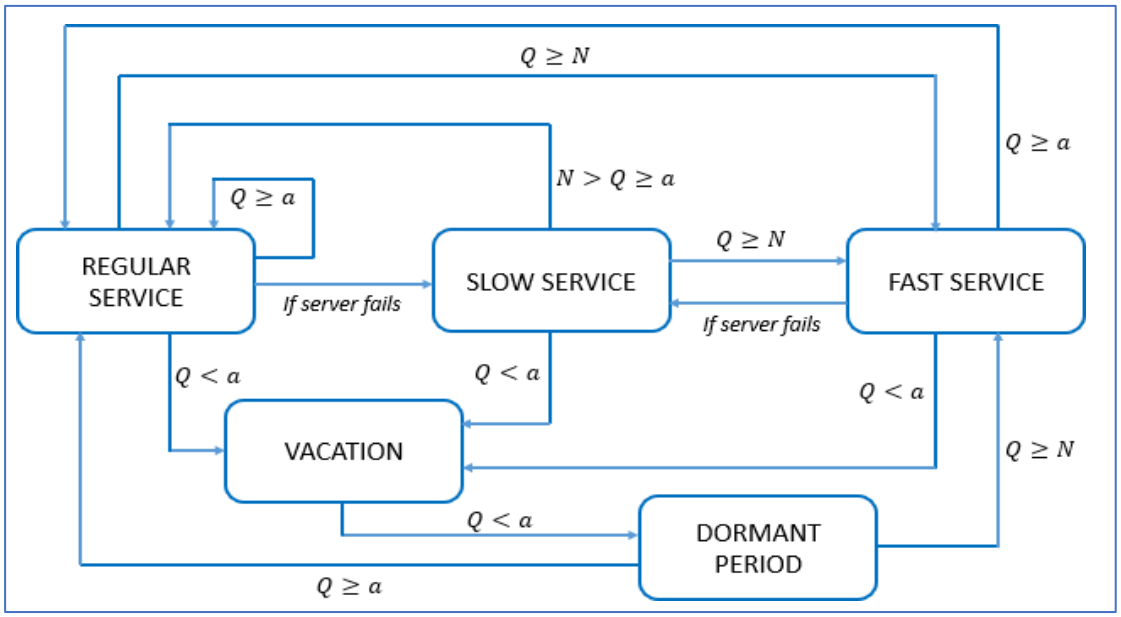

Fig2. Schematic representation of proposed model

The model describes a queueing system where arrivals follow a Poisson distribution. The server begins service only when at least requests accumulate, with maximum batch size of 'n'
The server operates at different service rates depending on queue conditions:

- Regular service for moderate queues.
- Fast service to quickly reduce long queues.
- Slow service during failures or congestion.

If the queue size falls below 'm' the server enters Vacation mode and pauses temporarily. For extended low traffic, the server transitions to Dormant mode, shutting down completely until demand increases. This dynamic system optimizes performance and resource usage based on queue length and server status.

3.2 NOTATIONS

- ω-Poisson arrival rate
- α-group size random variable of the arrival
- γ_a-probability that 'a' customers arrive in a batch
- $\alpha(k)$- probability generating function of α
- Z_s(h)- number of units under the service at time h
- Z_q(h)- number of units waiting for service at time h

3.3 SERVER STATES

X(H)=a/b/c/d/e

- a – if the server is busy with regular service
- b – if the server is busy with slow service
- c – if the server is busy with fast service
- d – if the server is on vacation
- e –if the server is on dormant period

The state probabilities are defined as follows.

(I_{mn}(y,h),dh = Pr{Z_s(h)=m, Z_q(h)=n, y\le R_{10}(h)\le y+dh, X(h)=a}), (u\le m\le v,; n\ge0), is the probability that, at time (h), the server is busy with (m) customers under regular batch service and (n) customers in the queue, and the remaining service time of a customer under service is between (y) and (y+dh). (J_{mn}(y,h),dh = Pr{Z_s(h)=m, Z_q(h)=n, y\le R_{20}(h)\le y+dh, X(h)=b}) is the probability that, at time (h),

the server is busy with (m) customers under slow batch service and (n) customers in the queue, and the remaining service time of a customer under service is between (y) and (y+dh).
(f_{mn}(y,h),dh = Pr{Z_s(h)=m, Z_q(h)=n, y\le R_{30}(h)\le y+dh, X(h)=d}) is the probability that, at time (h), the server is busy with (m) customers under fast batch service and (n) customers in the queue, and the remaining service time of a customer under service is between (y) and (y+dh).
(L_n(y,h),dh = Pr{Z_q(h)=n, y\le R_{40}(h)\le y+dh, X(h)=d}) is the probability that, at time (h), the server is on vacation, the queue length is (n), and the remaining vacation time of a customer at an arbitrary time is between (z) and (z+dt).
(T_n(h)=Pr{Z_q(h)=n, X(h)=e},;0\le n\le a-1), is the probability that, at time (h), the server is in the dormant state, the queue length is (n), and the remaining time of a customer at an arbitrary time is between (z) and (z+dt).
The steady-state analysis is as follows.
Equation (1) represents the following:
During regular service completion, if (m) customers are in the queue and (i) customers are in service, then the entire customers will be taken for service.
During slow service completion, if (m) customers are in the queue and (i) customers are in service, then the entire customers will be taken for service.
During fast service, if (m) customers are in the queue and (i) customers are in service, then the entire customers will be taken for service.
During vacation, if (m) customers are in the queue, then the entire customers will be taken for service.
During the dormant state, if (i) customers are in the queue, then (m-i) customers can arrive.
Similarly, it follows for the below equations.
The steady-state balance equations obtained for the regular, slow, fast, vacation, and dormant server states are first transformed by applying the Laplace–Stieltjes transform with respect to the supplementary variable. The resulting transformed equations are then multiplied by appropriate powers of the generating variable and summed over all admissible queue lengths. By eliminating the intermediate state generating functions through algebraic manipulation and incorporating the corresponding boundary conditions, a closed-form expression for the probability generating function of the queue length at an arbitrary epoch is obtained.
Accordingly, the probability generating function is given by the following equation.
Where (d_m=I_m+J_m+F_m), (q_m=T_n+L_n).
The probability generating function (P(K)) has to satisfy (P(1)=1). In order to satisfy this condition, applying Hospital's rule and evaluating (\lim_{K\to1}P(K)) and equating the expression to 1, it is derived that (\rho<1) is the condition to be satisfied for the existence of steady state for the model under consideration, where
[
\rho=\frac{\omega[\alpha]\left(E[R_1]+E[R_2]+E[R_3]\right)}{V}.
]
The unknown constants (q_m) are expressed in terms of (d_m) as
[
q_m=\sum_{i=0}^{m}d_{m-i}\beta_i,\qquad n=0,1,2,\ldots,u-1,
]
where (\beta_i) is the probability that (i) customers arrive during the vacation.
Proof:
From the equation we have
[
L(K,0)=\widetilde{R}_4(\omega-\omega\alpha(K))
]
and
\sum_{n=0}^{u-1}
\left(
\sum_{i=0}^{n}d_{m-i}\beta_i
\right)K^n.
]
Equating the coefficient of (K^n), (n=0,1,2,3,\ldots,u-1), on both sides of the equation, we get
[
q_m=\sum_{i=0}^{m}d_{m-i}\beta_i.
]
Expected Length of Busy Period
Let (B) be the busy random variable. Then, the expected length of the busy period is
[
E(B)=E(T),

]
where
[
E(T)=E(R_1)+E(R_2)+E(R_3).
]
Here, (E(R_1)) is the expected regular service time, (E(R_2)) is the expected slow service time, and (E(R_3)) is the expected fast service time.
Therefore,
[
E(B)=E(R_1)+E(R_2)+E(R_3).
]
Expected Length of Idle Period
[
E(I)=
\frac{E(V)}
{1-\displaystyle\sum_{i=0}^{u-1}\sum_{n=0}^{\infty}\gamma_j I_{n-1}}.
]
Where (\gamma_i) is the probability that (i) customers arrive during vacation, and (n) is the probability of (n) customers being in the queue at a departure epoch.
Expected Waiting Time in the Queue
The mean waiting time of the customers in the queue, (E(W)), can be easily obtained using Little's formula.
Optimization techniques are widely used to reduce the total average cost associated with queueing systems in many practical applications. The cost analysis process takes into account several factors, such as start-up cost, operating cost, holding cost, set-up cost, and reward cost (if applicable). The fundamental goal of system management is to achieve the minimum possible total average cost. In this section, the cost structure of the proposed queueing system is analyzed under the following assumptions.
(Z_h): Holding cost per customer per unit time.
(Z_0): Operating cost per unit time.
(Z_s): Startup cost per cycle.
(Z_r): Reward cost per cycle due to vacation.
Since the length of the cycle is the sum of the idle period and busy period,
[
E(T_C)=E(\text{length of idle period})+E(\text{length of the busy period}).
]
The Total Average Cost is given by
\left[
\frac{Z_s-Z_rE(I)}{E(T_C)}
\right]
+Z_hE(Q)+Z_0\rho.
]
Where
[
\rho=
\frac{\omega E[\alpha]\left(E[R_1]+E[R_2]+E[R_3]\right)}{V}.
]
The simple value direct search method to find the optimal policy for a threshold (N^*) to minimize the total average cost is defined as follows.
Step 1: Fix the value of maximum batch size (v).
Step 2: Select the value (N) which satisfies the following relation:
[
TACS(N^*)\le TA(N),\qquad 1\le N\le v.
]
Step 3: The value (N^*) is optimum, since it gives the minimum total average cost.
The above procedure gives the optimum value of (u), which minimizes the total average cost function. To justify the above solution, a numerical illustration will be presented in the next section.

NUMERICAL ILLUSTRATIONS

Arrival rate follows poisson distribution with parameter	ω
Regular service time distribution is 4-Erlang with parameter	η_1

Slow service time distribution is 4-Erlang with parameter	η_2
Fast service time distribution is 4-Erlang with parameter	η_3
Batch size distribution of the arrival is geometric with mean	2
Vacation time is exponential with parameter	ε_v
Minimum server capacity	u=2
Maximum server capacity	v= 4

3.4 IMPACT OF SYSTEM PARAMETERS ON KEY PERFORMANCE METRICS

In this paper, the performance of a batch service queueing system with fluctuating service rates caused by server failure was analysed. numerical results were obtained by using MATLAB software and presented in Table 1 to 6.

The following key observation are made:

9.1.1. Sensitivity of performance measures to arrival rate (ω)

From Table 1 and fig 3 it is observed that as the arrival rate increases, Expected Queue Length increases because more customers arrive in a shorter time, Expected Busy period increases as the server stays busy longer to serve more arrivals. Expected Idle period decreases because there are fewer gaps between customer arrivals. Expected Waiting time increases due to longer queues.

9.1.2. Sensitivity of performance measures to Regular service Rate

From Table 2 and fig 4 it is observed that as the regular service rate increases, Expected Queue length, Expected busy period, expected Waiting time Decrease because the server can handle more customers per unit time. Expected Idle period increases as the server clears queues faster leading to more idle time.

9.1.3. Sensitivity of performance measures to Slow service Rate

From Table 3 and fig 5 it is observed that as the slow service rate increases (i.e., Failure impact reduces), Expected Queue length, Expected Busy period and Expected Waiting time Decreases and Expected Idle period Increases because Improved slow service reduces queue build-up.

9.1.4. Sensitivity of performance measures to Fast service Rate

From Table 4 and Fig 6 it is observed that as the fast service rate increases, Expected Queue Length and Expected Waiting time decreases significantly as the server quickly clears larger queues. Expected Busy period decreases as faster service completes jobs quicker. Expected Idle time increases as fast service creates more idle periods.

9.1.5. Effect of Arrival Rate on Server State Probabilities

From Table 5 it is observed that as arrival rate increases the probability of the server being busy in regular, slow, fast states increases. The probability of the server being on vacation or in dormant state decreases because of the continuous arrival of customers keeping the server occupied.

9.1.6. Effect of Threshold Value (u) on Total Average Cost (TAC)

As the threshold value increases, Expected Queue length, Expected Busy period and expected idle time increases because the server waits longer for batch accumulation before starting service. The total average cost initially decreases due to improved batching efficiency but later increases due to longer delays, indicating there is an optimal threshold level for minimizing TAC.

The results highlight the importance of dynamically adjusting the server's service rate and threshold settings in response to real-time arrival rates and queue conditions. Proper tuning of parameters like arrival rate, service rates and threshold values can minimize queue length, waiting time, and system cost, making the model suitable for Real-time video streaming platforms like Netflix, where maintaining low latency and high server utilization is critical even under failure conditions.

TABLE 1: Arrival rate Vs Performance measures (u=2,v=4)

Arrival rate (ω)	E(Q)	E(B)	E(I)	E(W)
4.0	6.9205	4.1265	1.2397	3.7261
4.3	10.2415	6.4387	1.0951	4.9237
4.6	12.3467	9.5432	0.8327	6.5387

4.9	13.9345	11.2653	0.6391	8.2314
5.5	15.5235	13.3432	0.5183	10.2121
5.9	19.2651	16.7623	0.2381	12.2314
6.3	23.17643	20.2352	0.0972	14.2364

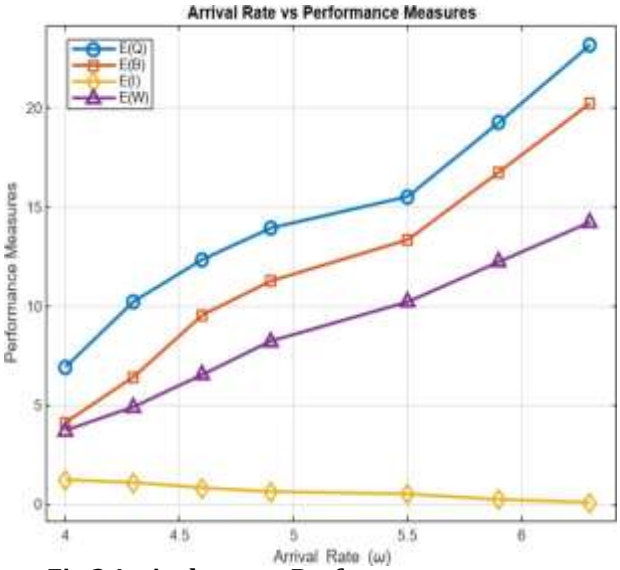

Fig.3 Arrival rate vs Performance measures

Table 2: **Regular service rate Vs Performance measures. ($\eta_1 = 3.2$, $u = 2, v = 4$)**

Regular Service rate	E(Q)	E(B)	E(I)	E(W)
2.3	5.9452	4.2412	1.2341	3.9241
2.6	4.5267	3.4982	2.1123	3.2253
2.9	3.5234	2.7712	2.2192	2.5793
3.2	3.0323	2.1341	3.5321	1.9287
3.5	2.1543	1.5411	4.1374	1.2341
4.2	1.5734	1.2179	4.1379	0.7271
4.5	0.9321	0.8253	5.1982	0.2397

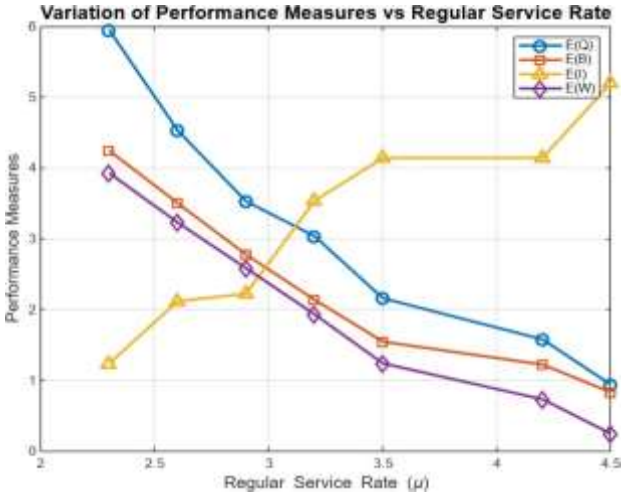

Fig.4 Regular service rate vs Performance measures

Table 3: Slow service rate Vs Performance measures. ($\eta_2 = 2.6, u = 2, v = 4$)

Slow service rate	E(Q)	E(B)	E(I)	E(W)
2.0	4.3273	4.1311	0.9921	3.5421
2.2	3.5421	3.8726	1.2341	3.1789
2.4	2.9354	3.3176	1.5377	2.5183
2.6	2.4197	2.9065	2.1329	2.0211
2.8	1.9363	2.4192	3.4789	1.4371
3.0	1.2171	1.9391	4.1762	0.9862
3.2	0.8321	1.5521	5.2198	0.2198

Fig.5 Slow service rate vs Performance measures

Table 4: Fast service rate Vs Performance measures. ($\eta_4 = 4, u = 2, v = 4$)

Fast service rate	E(Q)	E(B)	E(I)	E(W)
2.5	6.1142	3.9261	2.3271	3.1172
3.0	5.4397	3.2343	2.9364	2.9546
3.5	4.8271	2.7721	3.2372	2.1271
4.0	3.9673	2.1372	4.5462	1.8321
4.5	3.1281	1.5432	5.2391	1.5463
5.0	2.9323	1.1123	6.7481	1.2362
5.5	1.5421	0.8392	7.5232	0.8279

Fig.6. Fast service rate vs Performance measures

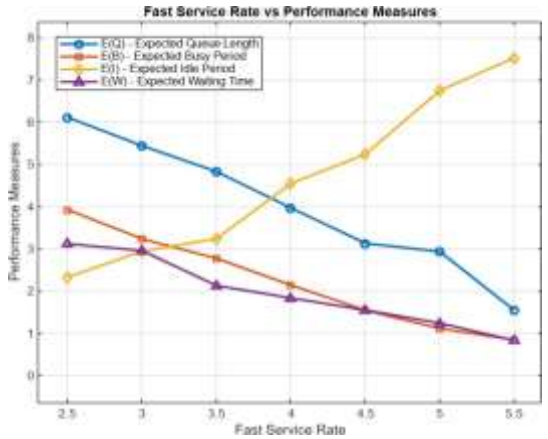

Table 5: Arrival rate Vs Probability that the server busy with Regular service, slow service, fast service and vacation.

Arrival rate	$P(R_1)$	$P(R_2)$	$P(R_3)$	$P(R_4)$	$P(D)$
4.0	0.2432	0.2175	0.1843	0.1542	0.2008
4.7	0.2641	0.2312	0.1875	0.1736	0.1436
5.6	0.2853	0.2435	0.1901	0.1924	0.0879
6.8	0.3056	0.2547	0.1910	0.2124	0.0363
7.6	0.3257	0.2652	0.1915	0.2312	0.0264
8.3	0.3442	0.2754	0.1917	0.2498	0.0389
9.5	0.3621	0.2847	0.1918	0.2683	0.0331

Table 6: Threshold Value (Vs) Performance Measures and Total Average Cost.

u	E(Q)	E(B)	E(I)	TAC
2	0.4211	2.1841	1.2321	0.9172
3	0.5623	3.4132	2.2200	0.7621
4	**0.5921**	**4.2351**	**3.1341**	**0.6342**
5	0.6382	10.2321	4.5321	0.6211
6	0.8211	12.4392	5.6372	0.7233
7	0.9345	17.1321	7.8321	0.7831
8	1.0271	18.2321	8.2121	0.8231
9	1.2343	20.6891	9.7621	0.9211
10	1.3521	22.4321	10.7421	0.9325

Minimum total average cost occurs when u=4, v=5, N=9

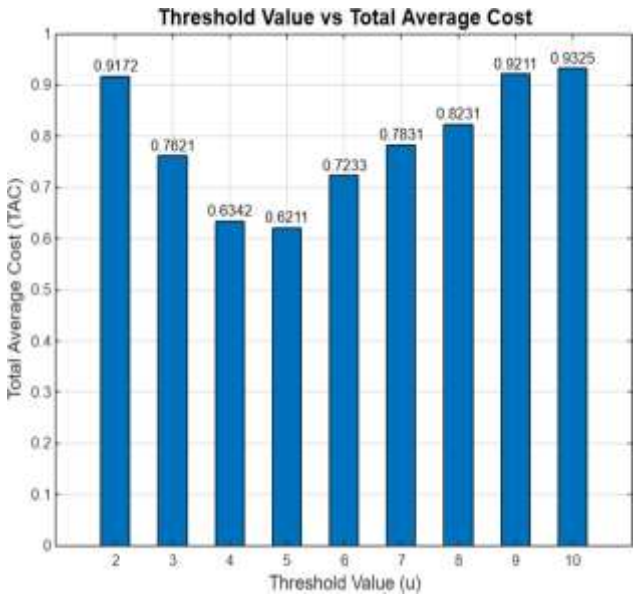

Fig.8 Threshold Value (Vs) E(Q) and Total Average Cost

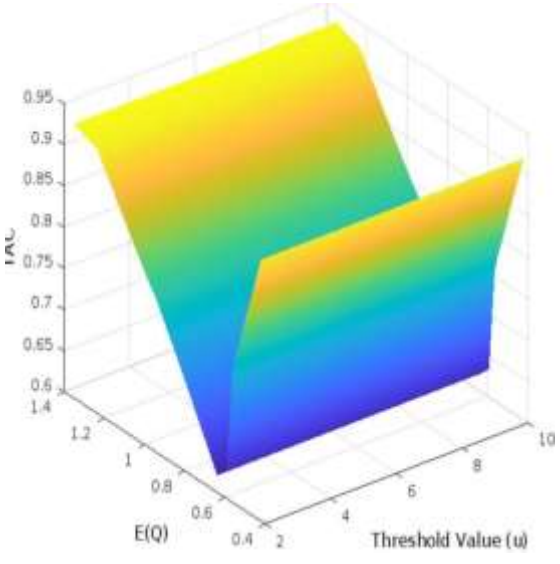

Fig.7 Threshold value vs TAC

Based on the numerical results and total average cost analyzis, it is concluded that for minimizing total average cost while ensuring low latency in video streaming platforms like Netflix, the management should select optimal Threshold values based on Maximum server capacity. Specifically, for N=4 or N=6 or N=9, the fixed threshold values are u=2, v=3; for u=3, v=4; for u=4, v=5.These threshold values help balance customer waiting time, server utilization, and operational cost. The model thus provides an effective cost-performance for real-time streaming platforms under varying traffic and server conditions.

4. Conclusion

In this paper, a queueing model is developed to reflect the dynamic and fault-prone nature of real-time streaming servers. By incorporating batch arrivals, general service rules, and five distinct server states-regular, slow, fast, vacation and dormant -the model captures the fluctuating service rates caused by varying queue lengths and server failures. The steady-state analysis using supplementary variable techniques and generating functions provides key insights into system behavior. Performance metrics such as expected queue length, waiting time and server utilization are systematically derived. The model proves especially effective in minimizing user latency and maintaining service continuity, even during server degradation. Through numerical results, the model demonstrates its practical applicability in optimizing performance and energy consumption in video streaming platforms. This work exte traditional queueing theory by introducing an adaptive frame work suitable for modern, high-demand service environments.

Reference

[1]. Niranjan, S. P., V. M. Chandrasekaran, and K. Indhira. "State dependent arrival in bulk retrial queueing system with immediate Bernoulli feedback, multiple vacations and threshold." *IOP conference series: materials science and engineering*. Vol. 263. No. 4. IOP Publishing, 2017.

[2]. Niranjan, S. P., V. M. Chandrasekaran, and K. Indhira. "Performance characteristics of a batch service queueing system with functioning server failure and multiple vacations." *Journal of physics: conference series*. Vol. 1000. No. 1. IOP Publishing, 2018.

[3]. Upadhyaya, Shweta. "Queueing systems with vacation: an overview." *International journal of mathematics in operational research* 9.2 (2016): 167-213.

[4]. Rani, R., and K. Indhira. "A REVIEW ON BULK QUEUE WITH SERVER BREAKDOWN MODELS." *Reliability: Theory & Applications* 19.2 (78) (2024): 25-34.

[5]. Haridass, M., and R. P. Nithya. "Analysis of a bulk queueing system with server breakdown and vacation interruption." *International Journal of Operations Research* 12.3 (2015): 069-090.

[6]. Chandrasekaran, V. M., et al. "A survey on working vacation queueing models." *Int. J. Pure Appl. Math* 106.6 (2016): 33-41.

[7]. Ahmed, Hamdy M., et al. "Models for COVID-19 daily confirmed cases in different countries." *Mathematics* 9.6 (2021): 659.

[8]. Sharma, Raksha, and Nikhil Khanna. "Naimark-Type Results Using Frames." *Journal of Function Spaces* 2024.1 (2024): 8588361..

[9]. Wang, Sheng, and Jiarui Meng. "Stochastic dynamics and density function of a food chain model with infinite distributed delay." *Filomat* 40.6 (2026): 2041-2066.

[10]. Sasikala, S., K. Indhira, and V. M. Chandrasekaran. "General bulk service queueing system with N-policy, multiple vacations, setup time and server breakdown without interruption." *IOP Conference Series: Materials Science and Engineering*. Vol. 263. No. 4. IOP Publishing, 2017.

[11]. Niranjan, S. P., and K. Indhira. "A review on classical bulk arrival and batch service queueing models." *Int. J. Pure Appl. Math* 106.8 (2016): 45-51.

[12]. Jeyakumar, S., and B. Senthilnathan. "Steady state analysis of bulk arrival and bulk service queueing model with multiple working vacations." *International Journal of Mathematics in Operational Research* 9.3 (2016): 375-394.

[13]. Jayaraman, D., R. Nadarajan, and M. R. Sitrarasu. "A general bulk service queue with arrival rate dependent on server breakdowns." *Applied mathematical modelling* 18.3 (1994): 156-160.

[14]. Jeyakumar, S., and B. Senthilnathan. "Modelling and analysis of a bulk service queueing model with multiple working vacations and server breakdown." *RAIRO-Operations Research-Recherche Operationnelle* 51.2 (2017): 485-508.

[15]. Pradhan, Sourav, and Prasenjit Karan. "Performance analysis of an infinite-buffer batch-size-dependent bulk service queue with server breakdown and multiple vacation." *Journal of Industrial & Management Optimization* 19.6 (2023).

[16]. Sultan, A. M., N. A. Hassan, and N. M. Elhamy. "Computational analysis of a multi-server bulk arrival with two modes server breakdown." *Mathematical and Computational Applications* 10.2 (2005): 249-259.

[17]. Jeyakumar, S., and B. Senthilnathan. "A study on the behaviour of the server breakdown without interruption in a MX/G (a, b)/1 queueing system with multiple vacations and closedown time." *Applied Mathematics and Computation* 219.5 (2012): 2618-2633.

[18]. Haridass, M., and R. Arumuganathan. "A batch service queueing system with multiple vacations, setup time and server's choice of admitting reservice." International Journal of Operational Research 14.2 (2012): 156-186.

[19]. Lee, Ho Woo, et al. "On a batch service queue with single vacation." Applied mathematical modelling

16.1 (1992): 36-42.
[20]. Choudhury, Gautam. "A batch arrival queue with a vacation time under single vacation policy." Computers & Operations Research 29.14 (2002): 1941-1955.
[21]. Reddy, GV Krishna, R. Nadarajan, and R. Arumuganathan. "Analysis of a bulk queue with N-policy multiple vacations and setup times." Computers & operations research 25.11 (1998): 957-967.
[22]. Arumuganathan, R., and K. S. Ramaswami. "Analysis of a bulk queue with fast and slow service rates and multiple vacations." Asia-Pacific Journal of Operational Research 22.02 (2005): 239-260.
[23]. Gray, William J., P. Patrick Wang, and Meckinley Scott. "A queueing model with multiple types of server breakdowns." *Quality Technology & Quantitative Management* 1.2 (2004): 245-255.
[24]. Doshi, Bharat T. "Queueing systems with vacations—a survey." *Queueing systems* 1 (1986): 29-66.
[25]. Ayyappan, G., and S. Nithya. "Analysis of batch arrival bulk service priority queue with multiple vacation, close down, setup, differentiate breakdown, and repair." *International Journal of Mathematics in Operational Research* 30.1 (2025): 1-21.
[26]. Tamrakar, G. K., and A. Banerjee. "On state dependent batch service queue with single and multiple vacation under Markovian arrival process." *International Journal of Operational Research* 52.4 (2025): 531-557.
[27]. Agrawal, Praveen Kumar, Anamika Jain, and Madhu Jain. "M/M/1 queueing model with working vacation and two type of server breakdown." *Journal of Physics: Conference Series*. Vol. 1849. No. 1. IOP Publishing, 2021.
[28]. Seenivasan, M., and S. Chandralekha. "Single server queueing model with multiple working vacation and with breakdown." *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. IEEE, 2022.
[29]. Seeniraj, G., R. Jayaraman, and D. Mogan raj. "Steady state analysis of queueing system with random vacation subject to supplementary variable method." *J. Math. Comput. Sci.* 11.2 (2021): 1838-1848.
[30]. Gautam, Choudhury, Kalita Priyanka, and Selvamuthu Dharmaraja. "Analysis of a model of batch arrival single server queue with random vacation policy." *Communications in Statistics-Theory and Methods* 50.22 (2021): 5314-5357.