IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁒ Peer-reviewed & Refereed journal ⁒ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.151154

Asymptotic Optimal Control of a data transmission queue in Heavy traffic with imperfect channel

Shipra Bhardwaj[1]*, Sharon Moses[2]

Research Scholar, Department of Mathematics, St. John's College, Agra (India)

Affiliated to Dr. Bhimrao Ambedkar University, Agra (India)[1]

Associate Professor, Department of Mathematics, St. John's College, Agra (India)

Affiliated to Dr. Bhimrao Ambedkar University, Agra (India)[2]

Email id of Corresponding Author: shiprabhardwaj108@gmail.com

Abstract: This study investigates the asymptotic optimal control of a data transmission queue operating under heavy traffic conditions with an imperfect channel. This model considers a single-server N-policy queue where packets arrive according to a Poisson process and transmissions are subject to channel failures, leading to retransmissions that rejoin the queue. The server remains inactive until the queue reaches a threshold N, after which it continues service until the system empties. Steady-state balance equations are developed to obtain explicit probability distributions for OFF and ON states, and heavy traffic scaling is used to derive asymptotic expressions for queue length, server utilization, and cost. The analysis establishes a reduced cost function capturing the trade-off between holding and activation costs and shows that the optimal activation threshold follows a classical square-root heavy traffic law. Numerical illustrations and simulations compare exact steady-state optimization with heavy traffic approximations, highlighting the conservatism and near optimal performance of asymptotic policies as system utilization approaches saturation.

Keywords*:* Data transmission queue, N-policy, imperfect channel, retransmission, heavy traffic, asymptotic optimal control

I. INTRODUCTION

A number of modern communication networks are subject to both congestion and unreliable channels which can be controlled using efficient queue management techniques in order to maintain network performance while also reducing operational expenses. Transmission failures lead to re-transmission activity that will add to the congestion of transmission queues and make server activation decisions increasingly difficult. Threshold based N-policies represent a practical method for achieving a balance between the costs of setting up service (i.e., the time it takes to set up) and the delay experienced in the queue by only activating service when there is sufficient backlog to justify service operation. This paper provides a stochastic queuing model that includes Poisson arrival streams, probabilistically occurring transmission failures, effective service rates, and retransmission activity to model the behavior of threshold based N-policies. This research uses steady-state analysis and heavy traffic theory to examine the behavior of threshold policies as the utilization factor approaches one and to formulate an asymptotic optimization problem to capture the major trade-offs involved in high load conditions. Finally, this paper relates exact Markovian analysis to diffusion scale approximations to provide both theoretical understanding and practical guidance for transmission control.

Sun *et al.* (2016) investigated customer decision making in the context of vacation queues where customers were subject to Markovian vacation structures. They showed how balking will emerge when vacations cause intermittent service availability, thus providing important information on how vacation structure affects the effective demand, the probability of joining the system at equilibrium and the overall degree of congestion in threshold controlled systems. This line of research has implications for transmission queues, the “ON/OFF” decision is not only a means of controlling operations, but is also a mechanism that indirectly controls offered load by affecting the participation/admission into the system and therefore its stability and long term cost. **Jain (2017)** developed a single-server model including several operational frictions such as batch arrival, priority, balking, unreliable servers, and a threshold based repair mechanism all linked with an optimization point of view. A major insight for threshold control is the explicit linkage between policy parameters, failure/repair dynamics and the degree of congestion. A threshold which is set too low leads to increased frequency of switches and recoveries, while a threshold which is set too high

DOI: 10.17148/IJARCCE.2026.151154

will lead to increased waiting costs and loss/balking. Therefore for data transmission systems with imperfect channels, this represents a strong modeling precedent for combining reliability impairments with congestion control within a common cost objective. **Zhou *et al.* (2017)** showed that the possibility for designing and developing efficient algorithms that approach the optimum in terms of delay when systems are under high traffic loads. This represents advancement on the theory to practice paradigm of designing algorithms that are close to optimal in terms of delay when the systems are under high traffic loads. They also demonstrated how low complexity methods can achieve good delay performance when the system is heavily loaded. Additionally, they also showed a separation between first order capacity feasibility and second order delay efficiency that occurs in heavy traffic regimes. Further, they provide the theoretical foundation for developing asymptotically optimal policies based on limited amounts of state information. Methodologically, their results support heavy traffic scaling arguments in threshold activation problems as utilization increases toward unity, the structure of the policy must be chosen so as to control diffusion scale fluctuations of the queue size rather than simply controlling the average rate at which the queue grows. **Chang *et al.* (2018)** studied the controlled arrival standby redundant model. In particular, they demonstrated how controlling the flow of failed items into the system and coordinating the decision to add redundancy to items failing in service can stabilize performance and reduce cost. The relevance of their results to the problem of threshold based server activation is the common concept of "gating" dynamic behavior: the system intentionally controls the timing of transitions (arrivals, activations, admissions) to prevent excessive operational expense associated with the overhead of those transitions while still providing acceptable waiting/availability. As such, their work has value for imperfect transmission models, where retransmissions serve as controlled re-entry's into the queue, and where activating the transmitter too often incurs additional setup or energy costs. **Chen and Wang (2018)** examined warm standby retrial machine repair under an N-policy, relating reliability engineering to queueing control, and identified that a threshold can be used as an effective mechanism for coordinating congestion (retrial population / queue size) with reliability (activation of standby units, and how repairs are allocated), resulting in tractable steady-state performance metrics and providing an opportunity for cost minimization. In terms of data transmission queues, there is a direct analogy between failed transmissions that join the queue again and again and an activation threshold can be used to determine when the transmitter should transition from idleness to active service, balancing the cost of activation with the cost of delay. **Zhou *et al.* (2018)** provided necessary and sufficient conditions for the optimality of heavy traffic delay in pull-based load balancing systems, further refining the understanding of when these types of policies result in diffusion scale optimal delays. Their findings reinforce that heavy traffic optimality is not guaranteed. Instead, it relies on certain structural characteristics of the system (information flow, decision rules, and state space collapse behavior). This provides a basis for the explicit verification that the selected threshold policy will result in the appropriate scaling order, and that there remains a non-trivial trade-off between holding cost growth, and the growth in setup costs as the load increases toward criticality. **Bouchentouf *et al.* (2019)** examined a single-server feedback queue with vacation and impatient customers through an economic framework. They focused on how feedback (re-entry) and service interruption create congestion and affect decision making regarding operation optimization. Specifically, they identified a relationship between imperfect channel transmission and their work; failed transmissions that are added back into the queue create feedback that can be unstable and/or unproductive, unless it is managed. Vacation/Interruption mimics off periods or unavailability of service and provides an analogy to this aspect of the work. Economically, the authors also highlighted the fact that although the optimum threshold may remain relatively small based on certain cost structures, the amount of feedback produced at high levels of utilization will create a high degree of cost sensitivity which supports using approximation techniques and performing thorough parameter studies. **Hurtado-Lange and Maguluri (2020)** derived methods for transforming heavy traffic systems in order to provide new tools for determining the limiting behavior, approximating and estimating improved performance characteristics of complex queueing network systems. The methodological contributions made by them allowed transition from direct steady state solution to heavy traffic asymptotics and assist in understanding the tail behaviors (geometric-like tails becoming “long”) that occur in the limiting regime as utilization approaches 1.**Gardner *et al.* (2021)** examined the scalability of load balancing using heterogeneous servers, indicates some of the same concerns as those related to scalability and heterogeneity for other applications such as inhomogenous networks. However, they also highlighted the importance of measuring the cost of overhead due to system size, diversity of service capability, and the use of communication and information. The broader lesson for threshold control is the necessity for policy simplicity and robustness. That is, in many cases, one may have too much difficulty implementing complex controls, especially if there are no theoretical guarantees that the controls will be effective. The ON/OFF transmission control problem is an example of this. A simple N policy is desirable in part because it is easy to measure and compute and yet captures most of the possible performance improvement. **Hurtado-Lange and Maguluri (2021)** demonstrated that limited sampling policies can achieve strong performance in heterogeneous settings in terms of both throughput and delay, thereby reinforcing the common theme identified earlier. Smart but simple policies can be effective in controlling delay at heavy loads. They are able to take advantage of incomplete system information. Therefore, their results support the idea that threshold based activation can achieve nearly optimal performance in heavy traffic regardless of whether all system states or complete histories of

DOI: 10.17148/IJARCCE.2026.151154

transmissions are known, provided that the threshold is scaled appropriately with system load. **Yang *et al.* (2021)**examined sojourn time behavior within Markovian queues with working breakdowns and delayed working vacations, investigating how service continues (possibly at reduced levels of service) and how interruptions and delays influence the time-in-system. Their work is very similar to that of imperfect channels where service requests will "fail" and thus reduce throughput and where server state transitions (OFF to ON, active to interrupted) are factors in determining the distribution of delay. Additionally, they emphasized the importance of sojourn times, which relate to communication goals such as meeting latency targets, implying that there are additional factors to consider besides average queue length when evaluating an activation threshold. **Bhambay and Mukhopadhyay (2022)** determined that speed aware, join the shortest-queue type policies were asymptotically optimal in heterogeneous service environments. In the context of transmission queues, the key concept is theoretical when the effective service rate depends on environmental factors (i.e., channel success probability), "rate-aware" control is beneficial. By translating this concept into practice, a threshold policy could be made channel-aware or success-rate-aware, by altering activation behavior based upon effective throughput, which can be particularly valuable when retransmission increases the workload. **Gamarnik *et al.* (2022)** analyzed the trade-off in terms of stability, memory, and messaging, and the degree to which the limitations of available information impact the potential performance of service systems. For controlled communication systems, their research has value since many operational policies have limited knowledge regarding both the current state of channels and queues due to real costs associated with collecting or transmitting this type of information. Their research provides a theoretical basis for developing efficient controls similar to N-policy, which requires less information than other types of controls while still maintaining some level of stability and acceptable delays. This is particularly relevant as systems grow in size. **Hurtado-Lange and Maguluri (2022)** studied the behavior of systems in the heavy traffic regime where resource pooling was completely ineffective and demonstrated that the absence of pooling changes both the limiting behavior of the system and the way that the state space collapses. The significance of these results in regards to threshold based transmission control is that heavy traffic approximations need to identify the correct structural bottlenecks. If a system acts like a collection of constraints that interact with each other (for example, channel unreliability acting in conjunction with the feedback from retransmissions), then naive approximations based on a single bottleneck will likely produce incorrect estimations of the thresholds regardless of how close the costs approach one another at high loads. Therefore, it is recommended to carefully determine the most important constraints of a system prior to implementing square root scaling prescriptions that are common in classical heavy traffic analyses. **Ramdani *et al.* (2025)** extended optimization studies to a system consisting of multiple servers that may be unavailable. Their study included Bernoulli type scheduling, working vacation, threshold based recovery and impatience. All of these were included in a modern study of reliability, vacations, thresholds and economics. Their study also demonstrated that threshold based recovery mechanisms remain relevant as a single “knob” controlling both the frequency of recovery/activation and costs associated with unreliability in heavy traffic conditions. Impatience also provides motivation to expand transmission queue models to include abandonment like behavior (i.e., timeouts, packet loss, time out deadlines) where they have potential to significantly alter optimal threshold values and interpretations of heavy traffic limits.

II. SYSTEM DESCRIPTION

We consider a single server transmission queue with the following assumptions:

(A)Arrival Process: Packets arrive according to Poisson process with rate λ.

(B) Service Mechanism:

(i) Transmission rate: μ

(ii) Channel failure probability: p

(iii) Successful transmission probability: $1-p$

(iv) Effective service rate: $\mu_e = \mu(1-p)$

(v) Server remains OFF when system empty.

(vi)Server turns ON only when queue length reaches threshold N.

(vii) After activation, continues until system empty.

III. STATE DEFINITION

Let:

$P_{i,0} = \Pr(\text{System has i packets, server OFF})$

$P_{i,1} = \Pr(\text{System has i packets, server ON})$

State space:

(i) OFF states: $i = 0,1,\dots,N-1$

(ii) ON states: $i = 1,2,\dots$

IV. STEADY-STATE BALANCE EQUATIONS

(i) Empty System, Server OFF: $\lambda P_{0,0} = \mu_e P_{1,1}$ (1)
(ii) OFF State: $1 \le i \le N-2{:}\lambda P_{i,0} = \lambda P_{i-1,0}$ (2)
(iii) Threshold State $i = N-1{:}\lambda P_{N-1,0} = \lambda P_{N-2,0}$ (3)
(iv) Activation Balance at Threshold:$\lambda P_{N-1,0} = \mu_e P_{N,1}$ (4)
(v) ON State for $i = 1{:}(\lambda+\mu_e)P_{1,1} = \lambda P_{0,0} + \mu_e P_{2,1}$ (5)
(vi) ON State for $2 \le i \le N-1{:}(\lambda+\mu_e)P_{i,1} = \lambda P_{i-1,1} + \mu_e P_{i+1,1}$ (6)
(vii) ON State for $i \ge N{:}(\lambda+\mu_e)P_{i,1} = \lambda P_{i-1,1} + \mu_e P_{i+1,1} + p\mu P_{i,1}$ (7)
(viii) Retransmission Correction Equation: Since failed transmissions rejoin queue:
$\mu p P_{i,1} = \lambda_r P_{i-1,1}$ (8)
where retransmission arrival rate: $\lambda_r = \mu p$ (9)

Normalization Condition:

OFF mode states: $(i,0)$with$i = 0,1,2,\dots N-1$ (10)

ON mode states: $(i,0)$ with $i \ge 1$ (and typically $(0,1)$ is not used because when the system becomes empty it switches OFF immediately). So the total probability over all admissible states must be 1:

$\sum_{i=0}^{N-1} P_{i,0} + \sum_{i=N}^{\infty} P_{i,1} = 1$ (11)

V. EXPLICIT STEADY-STATE PROBABILITIES

Assume: $P_{i,1} = r^i$ (12)
Substitute: $(\lambda+\mu_e)r^i = \lambda r^{i-1} + \mu_e r^{i+1}$
Divide by r^{i-1}: $(\lambda+\mu_e)r = \lambda + \mu_e r^2$
Rearrange: $\mu_e r^2 - (\lambda+\mu_e)r + \lambda = 0$
Roots: $r_1 = 1, r_2 = \frac{\lambda}{\mu_e} = \rho$ (13)
Since $\rho < 1$, steady-state solution must decay: $P_{i,1} = A\rho^i$ (14)
Using boundary equation at activation point:
At $i = N{:}\, \lambda P_{N-1,0} = \mu_e P_{N,1}$
Since $P_{N-1,0} = P_{0,0}$
$\lambda P_{0,0} = \mu_e A\rho^N$
Thus $A = \frac{\lambda}{\mu_e}\frac{P_{0,0}}{\rho^N} = \frac{\rho P_{0,0}}{\rho^N} = P_{0,0}\rho^{1-N}$

$P_{i,1} = P_{0,0}\rho^{1-N}\rho^{i}$
$P_{i,1} == P_{0,0}\rho^{i+1-N}, for\ i \geq N$ (15)

From $(i,0)$ to $(i+1,0)$ occurs at rate λ. The only way into $(i,0)$is from $(i-1,0)$ at rate λ. Hence, at steady state:

$\lambda P_{i,0} = \lambda P_{i-1,0} \Longrightarrow P_{i,0} = P_{i-1,0}$

By repeated substitution: $P_{1,0} = P_{0,0}, P_{2,0} = P_{1,0} = P_{0,0}, \dots, P_{i,0} = P_{0,0}$ for $i = 0,1,\dots,N-1$

Sum the OFF probabilities

$\sum_{i=0}^{N-1} P_{i,0} = NP_{0,0}$ (16)

Sum the ON probabilities

$\sum_{i=N}^{\infty} P_{i,1} = \sum_{i=N}^{\infty} P_{0,0}\rho^{i+1-N} = P_{0,0}\sum_{k=0}^{\infty}\rho^{k+1} = P_{0,0}\rho\sum_{k=0}^{\infty}\rho^{k}$
$\sum_{i=N}^{\infty} P_{i,1} = \frac{\rho}{1-\rho}P_{0,0}$ (17)
Using Normalization Condition (11), we get
$NP_{0,0} + \frac{\rho}{1-\rho}P_{0,0} = 1$
OFF States:$P_{i,0} = \frac{1}{N+\frac{\rho}{1-\rho}}, i = 0,1,\dots,N-1$ (18)
ON States:$P_{i,1} = \frac{\rho^{i+1-N}}{N+\frac{\rho}{1-\rho}}, i \geq N$ (19)
Heavy Traffic: As $\rho \longrightarrow 1$: $\frac{\rho}{1-\rho} \to \infty$
Thus, $P_{0,0} \sim (1-\rho)$ (20)
Queue behaves geometrically with long tail.
Mean Queue Length: $E[Q] = \sum_{i=0}^{N-1} iP_{i,0} + \sum_{i=N}^{\infty} iP_{i,1}$ (21)
$\sum_{i=0}^{N-1} iP_{i,0} = P_{0,0}\sum_{i=0}^{N-1} i = \frac{N(N-1)}{2}P_{0,0}$
$E[Q:OFF] = \frac{N(N-1)}{2}P_{0,0}$ (22)

For the standard N-policy, once ON, the ON-level probabilities form a geometric tail:

$P_{N+k,1} = P_{N,1}\rho^{k}, k = 0,1,2,\dots$ (23)

with$\rho = \frac{\lambda}{\mu_e}$.

Also the boundary relation gives: $P_{N,1} = \rho P_{0,0}$

$E[Q] = P_{0,0}\sum_{i=0}^{N-1} i + \sum_{i=N}^{\infty} iP_{i,1}$ (24)
$\sum_{i=N}^{\infty} iP_{i,1} = \sum_{k=0}^{\infty}(N+k)P_{N,1}\rho^{k} = P_{N,1}[N\sum_{k=0}^{\infty}\rho^{k} + \sum_{k=0}^{\infty}k\rho^{k}]$
Use standard sums ($|\rho| < 1$)
$E[Q:ON] = P_{N,1}\left[\frac{N}{1-\rho} + \frac{\rho}{(1-\rho)^2}\right]$
Substitute$P_{N,1} = \rho P_{0,0}$:
$E[Q:ON] = \rho P_{0,0}\left[\frac{N}{1-\rho} + \frac{\rho}{(1-\rho)^2}\right]$ (25)
$E[Q] = E[Q:OFF] + E[Q:ON]$
$E[Q] = \frac{N(N-1)}{2}P_{0,0} + \rho P_{0,0}\left[\frac{N}{1-\rho} + \frac{\rho}{(1-\rho)^2}\right]$
$E[Q] = \left[\frac{N(N-1)}{2} + \left\{\frac{N\rho}{1-\rho} + \frac{\rho^2}{(1-\rho)^2}\right\}\right]P_{0,0}$ (26)
$\left[\frac{N\rho}{1-\rho} + \frac{\rho^2}{(1-\rho)^2}\right]P_{0,0} = P_{0,0}\frac{\rho}{(1-\rho)^2}[N(1-\rho)+\rho]$

DOI: 10.17148/IJARCCE.2026.151154

$$= \frac{\rho}{(1-\rho)^2}\left[NP_{0,0} - \rho NP_{0,0} + \rho P_{0,0}\right] = \frac{\rho}{(1-\rho)^2}\left[NP_{0,0} + \rho P_{0,0}(1-N)\right]$$

Using

$$NP_{0,0} = 1 - \frac{\rho}{1-\rho}P_{0,0}$$

$$\left[\frac{N\rho}{1-\rho} + \frac{\rho^2}{(1-\rho)^2}\right]P_{0,0} = \frac{\rho}{(1-\rho)^2}\left[1 - \frac{\rho}{1-\rho}P_{0,0} + \rho P_{0,0}(1-N)\right]$$

$$= \frac{\rho}{(1-\rho)^2}\left[1 + \rho P_{0,0}\left\{(1-\mathrm{N}) - \frac{1}{1-\rho}\right\}\right]$$

$$= \frac{\rho}{(1-\rho)^2}\left[1 + \rho P_{0,0}\frac{-\rho-\mathrm{N}(1-\rho)}{1-\rho}\right]$$

$$= \frac{\rho}{(1-\rho)^2} - \frac{\rho}{(1-\rho)^2}\left[\rho P_{0,0}\frac{\rho+\mathrm{N}(1-\rho)}{1-\rho}\right]$$

$$= \frac{\rho}{(1-\rho)^2} - \frac{\rho}{(1-\rho)^2}\rho P_{0,0}\left(N + \frac{\rho}{1-\rho}\right) \tag{27}$$

Using normalization condition, we have $P_{0,0}\left(N + \frac{\rho}{1-\rho}\right) = 1 \Longrightarrow \rho P_{0,0}\left(N + \frac{\rho}{1-\rho}\right) = \rho$

$$= \frac{\rho}{(1-\rho)^2} - \frac{\rho^2}{(1-\rho)^2} = \frac{\rho}{1-\rho}$$

$$E[Q] = \frac{\rho}{1-\rho} + \frac{N(N-1)}{2}P_{0,0} \tag{28}$$

VI. HEAVY TRAFFIC SCALING

Heavy traffic scaling in our model focuses on what happens as the traffic intensity $\rho \longrightarrow 1$ (i.e., the system approaches saturation, where the effective service capability barely exceeds the offered load). In this regime the queue becomes large, the stationary distribution develops a long tail, and performance measures such as mean queue length and delay blow up unless control is tuned appropriately. In our N-policy transmission queue with an imperfect channel, the "effective service rate" is reduced by failures, and failed transmissions rejoin the queue as retransmissions; this feedback increases congestion and makes the heavy traffic regime especially relevant. The steady-state ON-state tail behaves approximately geometric with ratio close to 1, so a standard heavy traffic approximation is to write $\rho = 1 - \varepsilon$ with $\varepsilon \to 0$; then key quantities scale like $1/\varepsilon$ (i.e. mean queue length) grows on the order of $1/1-\rho$. The control variable N must scale with ε to maintain a nontrivial tradeoff between holding cost (which grows with queue size) and activation/setup cost (which penalizes turning the server ON too often). Our derivation uses the classical square-root scaling: choose N to be of order $1/\sqrt{1-\rho}$. This ensures that no term dominates completely in the limiting cost if N grows too slowly, holding costs explode; if it grows too quickly, setup costs dominate because the server stays OFF too long and large backlogs accumulate. Under this scaling, the reduced long-run average cost admits an asymptotic expansion where the N dependent terms balance at leading order, yielding an asymptotically optimal threshold of the form

$$N^*(\rho) \approx \frac{C}{\sqrt{1-\rho}} \tag{29}$$

for a constant C determined by the ratio of setup cost to holding cost and by the effective service parameters. This is exactly the "square-root heavy traffic optimality law" highlighted in our results section, and it explains why the heavy traffic predicted threshold rises sharply as ρ approaches 1, even when the exact optimal integer threshold remains small for the particular cost parameters used in our numerical experiments.

VII. ASYMPTOTIC OPTIMAL CONTROL PROBLEM

(i) Cost Structure: $C(N) = hE[Q] + c_s Pr(Server\ ON)$ (30)

where

h = holding cost per packet

c_s = Setup cost per activation

(ii) Heavy Traffic Asymptotics: Let $\varepsilon = 1 - \rho$

$$\frac{\rho}{1-\rho} = \frac{1-\varepsilon}{\varepsilon} = \frac{1}{\varepsilon} - 1$$

DOI: 10.17148/IJARCCE.2026.151154

Thus $P_{0,0} = \frac{1}{N+\frac{1-\varepsilon}{\varepsilon}} = \frac{1}{N+\frac{1}{\varepsilon}-1} = \frac{\varepsilon}{1+\varepsilon(N-1)}$

For heavy traffic:$P_{0,0} \sim \frac{\varepsilon}{1+\varepsilon N}$ (31)

(iii) Asymptotic Mean Queue Length:

$$E[Q] = \frac{\rho}{1-\rho} + \frac{N(N-1)}{2} P_{0,0} \sim \frac{1}{\varepsilon} + \frac{\varepsilon N^2}{2(1+\varepsilon N)} \tag{32}$$

(iv) Asymptotic Server ON Probability:

$$Pr(ON) = P_{0,0} \frac{\rho}{1-\rho} \sim \frac{\varepsilon}{1+\varepsilon N} \frac{1}{\varepsilon} = \frac{1}{1+\varepsilon N} \tag{33}$$

(v) Asymptotic Cost Function:

$$C(N) = h\left[\frac{1}{\varepsilon} + \frac{\varepsilon N^2}{2(1+\varepsilon N)}\right] + c_s \frac{1}{1+\varepsilon N} \tag{34}$$

The first term independent of N can be dropped for optimization.
Thus define reduced cost:

$$C(N) = \frac{h\varepsilon N^2}{2(1+\varepsilon N)} + c_s \frac{1}{1+\varepsilon N} = \frac{\frac{h\varepsilon N^2}{2}+c_s}{(1+\varepsilon N)} \tag{35}$$

For minimum $C(N)$: $\frac{dC}{dN} = 0$

$$\frac{dC}{dN} = \frac{h\varepsilon N(1+\varepsilon N) - \left(\frac{h\varepsilon N^2}{2}+c_s\right)\varepsilon}{(1+\varepsilon N)^2} = 0$$

$$h\varepsilon N(1+\varepsilon N) - \left(\frac{h\varepsilon N^2}{2} + c_s\right)\varepsilon = 0$$

$$h\varepsilon N + h\varepsilon^2 N^2 - \frac{h\varepsilon^2 N^2}{2} - c_s\varepsilon = 0$$

$$h\varepsilon N + \frac{h\varepsilon^2 N^2}{2} - c_s\varepsilon = 0$$

Dividing by ε

$$hN + \frac{h\varepsilon N^2}{2} - c_s = 0 \tag{36}$$

This is the key optimality equation.

As $\varepsilon \to 0, N \sim \frac{1}{\sqrt{\varepsilon}}$

So εN^2remains order 1.

That is the balancing principle.

Rewrite optimality equation:

$$hN + \frac{h\varepsilon N^2}{2} - c_s = 0$$

For large N the quadratic term dominates the linear one.

$$\frac{h\varepsilon N^2}{2} \approx c_s \Longrightarrow N^* \approx \sqrt{\frac{2c_s}{h\varepsilon}} = \sqrt{\frac{2c_s}{h(1-\rho)}} \tag{37}$$

(vi) Asymptotic Optimal Threshold: Under heavy traffic $\rho \longrightarrow 1$, the N-policy that minimizes long-run average cost satisfies:

$$N^* \sim \Theta\left(\frac{1}{\sqrt{1-\rho}}\right) \text{and} \lim_{\rho \to 1} \sqrt{1-\rho} N^* = \sqrt{\frac{2c_s}{h}} \tag{38}$$

Thus$N^* \approx \sqrt{\frac{2c_s}{h(1-\rho)}}$ (39)

This is the classical square-root heavy traffic optimality law.

VIII. Results and Discussion

In the numerical study, we use the following baseline parameter values unless stated otherwise: service rate $\mu = 1$, channel failure probability $p = 0.4$ (so the success probability is $1 - p = 0.6$), holding cost $h = 1$, and setup/activation cost $c_s = 5$. These values are used to compute the exact steady-state optimal threshold and to evaluate the heavy traffic approximation across different utilization levels ρ.

DOI: 10.17148/IJARCCE.2026.151154

Figure 1: Optimal Threshold Comparison

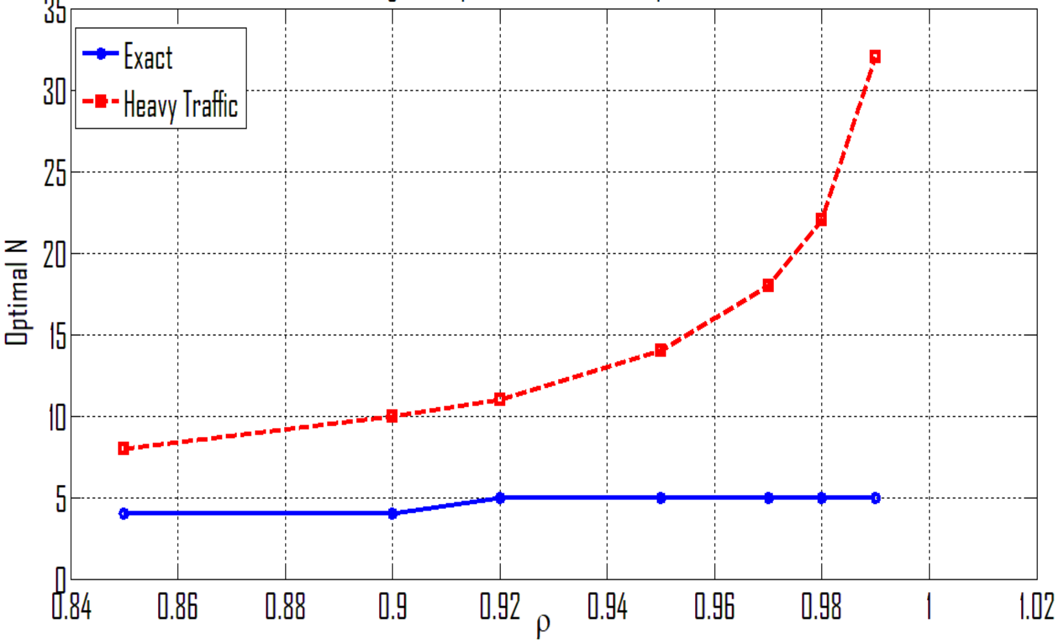


Figure 2: Cost Gap of Heavy Traffic Approximation

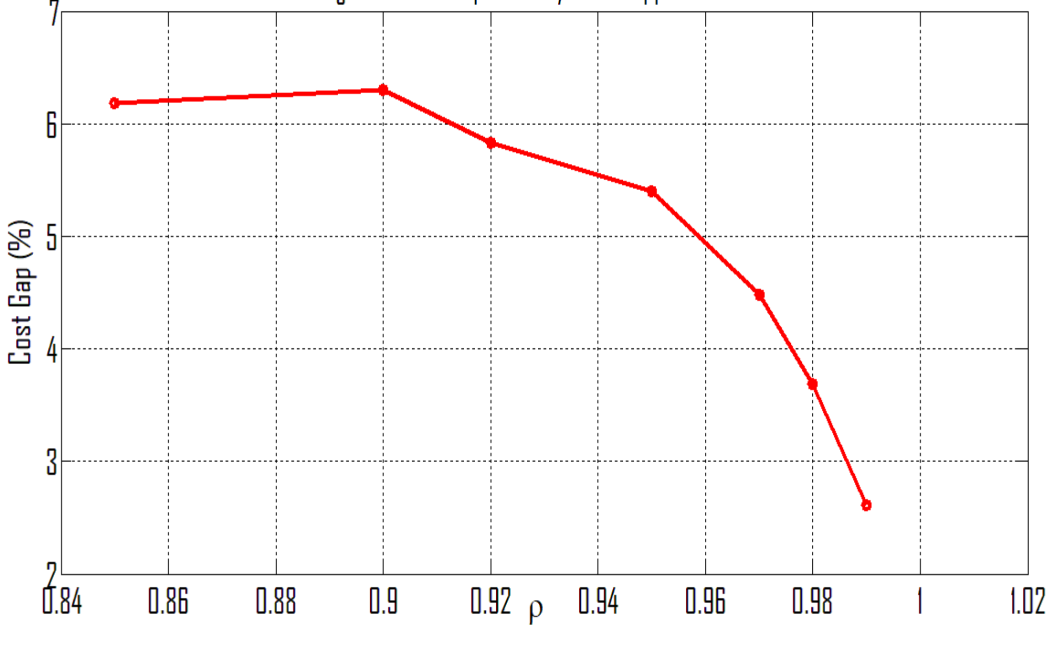


Figure 3: 3D Surface: $N^*(\rho,p)$

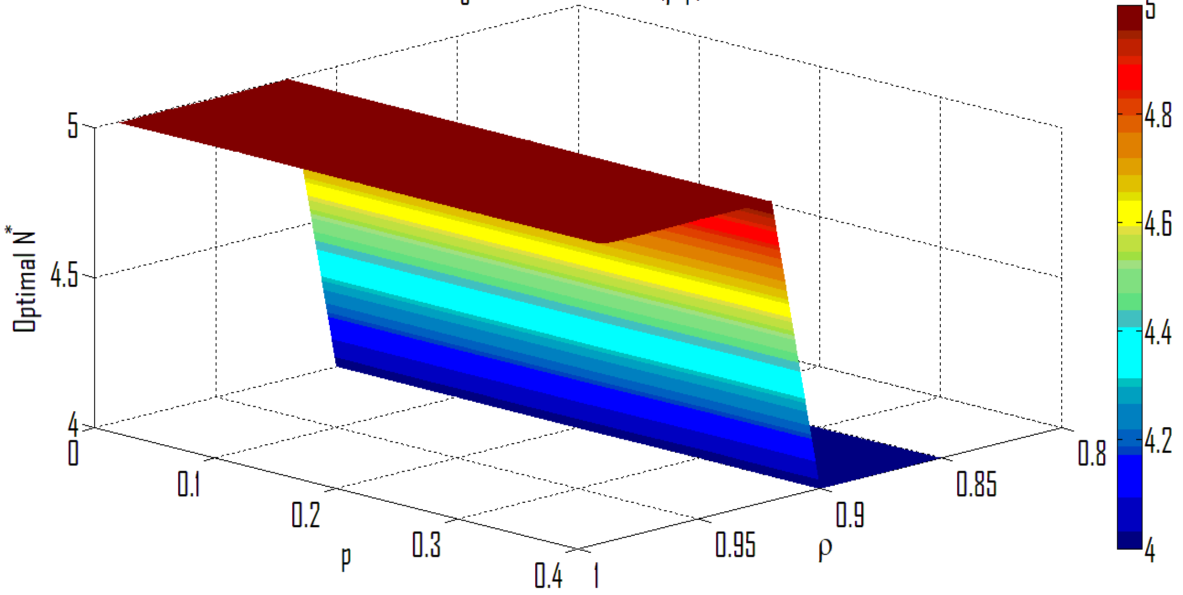
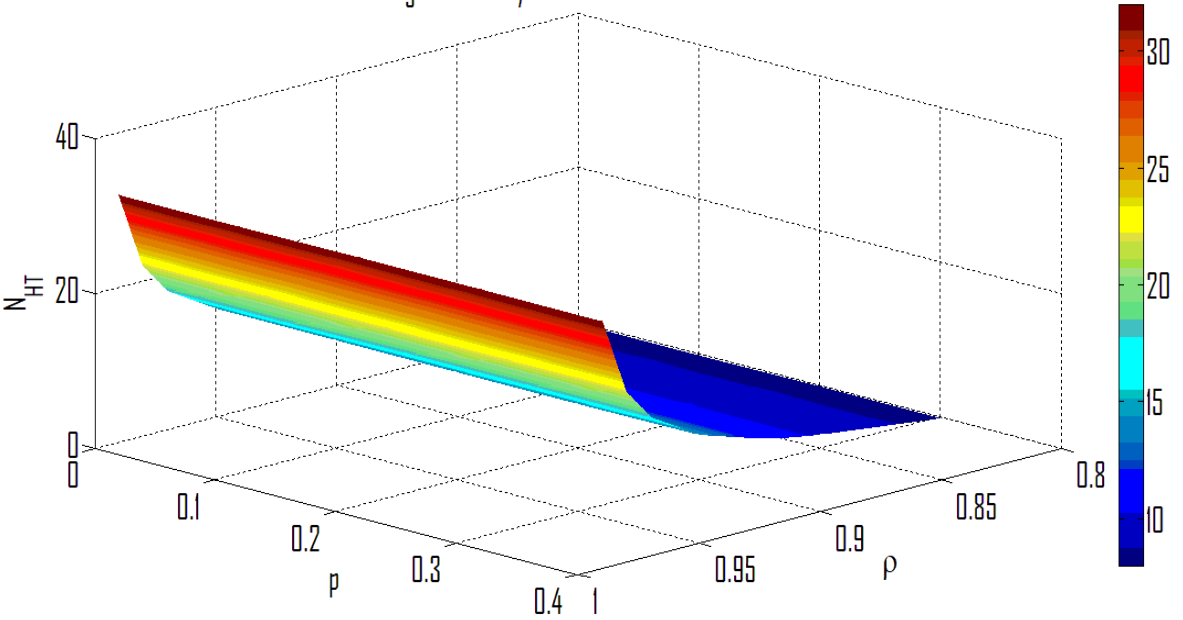


Figure 4: Heavy Traffic Predicted Surface



DOI: 10.17148/IJARCCE.2026.151154

Figure 5: Difference Surface: Exact minus Heavy Traffic

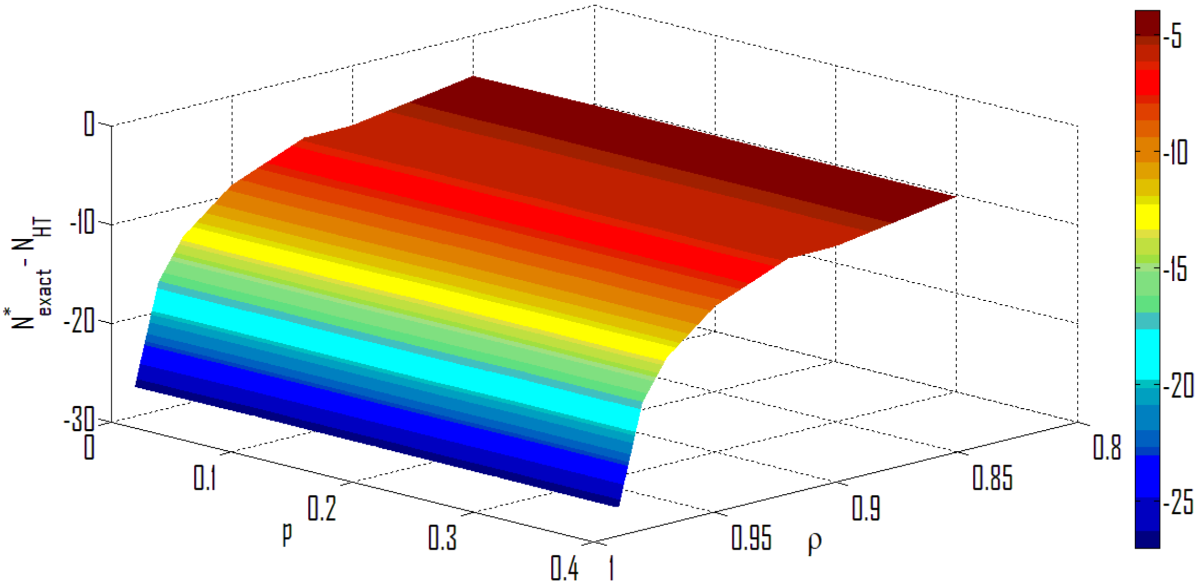
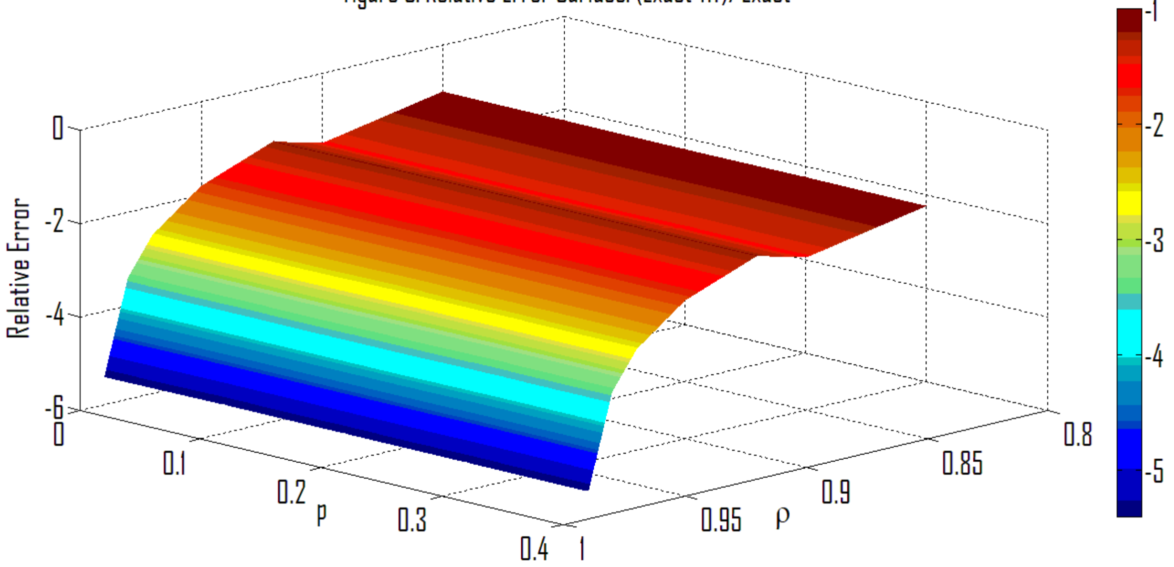


Figure 6: Relative Error Surface: (Exact-HT)/Exact



The figure (1) compares the optimal threshold $\boldsymbol{N}$ obtained from the exact steady-state optimization (blue curve) with the heavy traffic asymptotic prediction (red dashed curve) as the utilization ρ increases toward 1. The exact optimal threshold stays small and nearly flat (around $N \approx 4-5$) across the whole range of ρ, indicating that under chosen cost parameters only a modest activation threshold is needed even in relatively heavy load. In contrast, the heavy traffic approximation increases rapidly and non-linearly as $\rho \rightarrow 1$, reflecting its square-root blow-up behavior (roughly $N_{HT} \propto 1/\sqrt{1-\rho}$). The widening gap shows that, for these parameter values and this ρ-range, the heavy traffic formula is much more conservative (it recommends turning on at far larger queue sizes) than the exact optimum suggesting either that the system is not in the asymptotic regime where the approximation is accurate, or that the cost/activation modeling assumptions used for the heavy traffic derivation differ from those used in the exact computation.

DOI: 10.17148/IJARCCE.2026.151154

The figure (2) shows how the relative cost penalty of using the heavy traffic threshold instead of the exact optimal threshold changes as the load ρ approaches 1. The vertical axis is the cost gap (%), i.e., the percentage increase in long-run average cost $C(N)$ when the system uses N_{HT} rather than N^*_{exact}. For moderate heavy traffic ($\rho \approx 0.85 - 0.90$), the gap is about 6%, meaning the asymptotic rule is noticeably sub-optimal in that range. As ρ increases toward saturation, the curve trends downward, dropping to roughly 2–3% near $\rho \approx 0.99$. This decline indicates that the heavy traffic approximation becomes increasingly accurate in the true heavy traffic regime. Even if the predicted threshold may differ from the exact optimizer, its cost performance converges, delivering near optimal average cost as $\rho \to 1$.

The 3D surface plot in figure (3), the exact optimal N-policy threshold $\boldsymbol{N}^*$ as a joint function of the traffic intensity ρ and the channel failure probability p. The key feature is that the surface is almost flat, with $\boldsymbol{N}^*$ taking only a small range of integer values (roughly 4 to 5 across the plotted region). As ρ increases toward heavy traffic, the optimal threshold exhibits a step-like increase (reflecting the fact that N is integer-valued), moving from about $N^* \approx 4\ to\ N^* \approx 5$ once the system becomes sufficiently loaded. In contrast, variation in the failure probability p produces very little change in N^* for these parameter settings, indicating that the chosen cost structure and the effective service-rate modeling make the optimal activation threshold far more sensitive to congestion (ρ) than to moderate levels of channel unreliability.

The figure (4) shows the heavy traffic (HT) predicted threshold surface $\boldsymbol{N_{HT}}$ as a function of utilization ρ and channel failure probability p. The dominant trend is along the ρ-axis as ρ increases toward 1, the predicted threshold rises sharply, reflecting the heavy traffic scaling $N_{HT} \propto 1/\sqrt{1-\rho}$. This creates a steep "wall" near $\rho \approx 0.95$ where N_{HT} jumps into the range of 20–30+, indicating that the asymptotic rule recommends waiting for a much larger backlog before activating service under near saturation conditions. In contrast, the surface varies only mildly with p across the plotted range, implying that under the particular HT formula used the threshold is driven primarily by congestion intensity and cost parameters, with channel unreliability affecting the threshold only weakly (or indirectly, depending on how ρ is defined via the effective service rate). Overall, the plot highlights that the HT approximation is highly sensitive to ρ and becomes increasingly conservative as the system approaches critical load.

Figure (5) plots the difference surface $N^*_{exact} - N_{HT}$ over the (ρ, p) plane, so negative values mean the heavy traffic rule recommends a larger activation threshold than the true exact optimizer. The surface is pre-dominantly below zero across the displayed region, confirming that the heavy traffic approximation is generally conservative here i.e. it delays turning ON service until the queue is much larger than the exact-cost minimizing policy would choose. The magnitude of the negative gap grows rapidly as ρ approaches 1, reaching values on the order of −20 to −30 near the highest utilizations, which mirrors the steep growth of N_{HT} in heavy traffic. In contrast, the surface varies only mildly with the failure probability p, indicating that the mismatch between exact and heavy traffic thresholds is driven mainly by congestion intensity ρ rather than channel unreliability in this parameter regime. Overall, the plot provides a clear diagnostic while the heavy traffic threshold may become asymptotically meaningful very close to $\rho = 1$, over the studied range it substantially overestimates the optimal integer threshold.

Figure (6) presents the relative error surface $(N^*_{exact} - N_{HT})/N^*_{exact}$ over (ρ, p), which measures the threshold mismatch as a fraction of the true optimal threshold. The surface is negative throughout, indicating that $N_{HT} > N^*_{exact}$ almost everywhere so the heavy traffic prescription systematically overestimates the optimal activation threshold. Moreover, the magnitude of the relative error grows markedly as ρ increases toward 1 because N^*_{exact} remains small (around 4–5 in our experiments) while N_{HT} increases steeply in heavy traffic, the ratio becomes strongly negative (large in absolute value), showing that the HT rule can be multiple times larger than the exact optimizer in this parameter range. Dependence on p is comparatively mild, so the dominant driver of error is traffic intensity ρ rather than channel unreliability. Overall, the figure quantifies the conservativeness of the heavy traffic threshold in a normalized way and highlights that even when the cost gap may shrink near 1 , the threshold level discrepancy can remain large due to the integer nature and small scale of the exact N^* under the chosen cost parameters.

DOI: 10.17148/IJARCCE.2026.151154

Table 1: Comparison of Exact Optimal Threshold and Heavy Traffic Predicted Threshold with Scaling Indicator and Cost Gap (as ρ⟶1) and p=0.40

ρ	N_{exact}	N_{HT}	$(1-\rho)/N^2$	Cost Gap
0.85	4	8	2.4	0.061858
0.9	4	10	1.6	0.06297
0.92	5	11	2	0.058309
0.95	5	14	1.25	0.053962
0.97	5	18	0.75	0.04477
0.98	5	22	0.5	0.036894
0.99	5	32	0.25	0.026085

Table 2: Simulation Validation (with 95% CI)

p	ρ	N	$E(Q)$			$P(ON)$		
			Exact	Simulated	Mean±CI	Exact	Simulated	Mean±CI
0.2	1	5	19.417	21.5349	± 0.7607	0.792	0.9505	± 0.0012
0.2	1	5	49.185	50.5559	± 3.6548	0.907	0.98	± 0.0011
0.3	1	5	49.185	53.6179	± 5.468	0.907	0.98	± 0.0014
0.4	1	5	99.096	109.2468	±21.8529	0.952	0.9902	± 0.0015

IX. Concluding Remarks

The paper demonstrates that an N-policy transmission queue with imperfect channel conditions admits tractable steady-state solutions and meaningful heavy traffic approximations that guide optimal activation decisions. The derived asymptotic analysis shows that the optimal threshold grows according to a square-root scaling law, reflecting the balance between setup and congestion costs in near saturation regimes. Numerical and simulation results indicate that while heavy traffic thresholds may overestimate the exact optimizer for moderate loads, their cost performance becomes increasingly accurate as utilization approaches unity. The findings highlight that congestion intensity plays a dominant role in determining activation thresholds, whereas channel failure probability has comparatively weaker influence under the considered cost structure. Overall, the study provides a unified analytical and computational framework for designing efficient transmission control policies in unreliable communication environments and suggests directions for extending the model to multi-server systems, adaptive policies, and more complex reliability structures.

REFERENCES

[1]. Bhambay S., Mukhopadhyay A. (2022): "Asymptotic optimality of speed-aware JSQ for heterogeneous service systems", *Performance Evaluation*, 157:102320.

[2]. Bouchentouf A.A., Cherfaoui M., Boualem M. (2019): "Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers", *Opsearch*, 56(1):300–323.

[3]. Chang F.M., Lee Y.T., Chang C.J., Yeh C. (2018): "Analysis of a standby redundant system with controlled arrival of failed machines", *International Journal of Industrial and Systems Engineering*, 28(1):117–134.

[4]. Chen W.L., Wang K.H. (2018): "Reliability analysis of a retrial machine repair problem with warm standbys and a single server with N-policy", *Reliability Engineering & System Safety*, 180:476–486.

[5]. Gamarnik D., Tsitsiklis J.N., Zubeldia M. (2022): "Stability, memory, and messaging trade-offs in heterogeneous service systems", *Mathematics of Operations Research*, 47(3):1862–1874.

[6]. Gardner K., Abdul Jaleel J., Wickeham A., Doroudi S. (2021): "Scalable load balancing in the presence of heterogeneous servers", *ACM SIGMETRICS Performance Evaluation Review*, 48(3):37–38.

[7]. Hurtado-Lange D., Maguluri S.T. (2020): "Transform methods for heavy-traffic analysis", *Stochastic Systems*, 10(4):275–309.

DOI: 10.17148/IJARCCE.2026.151154

[8]. Hurtado-Lange D., Maguluri S.T. (2021): “Throughput and delay optimality of power-of-d choices in inhomogeneous load balancing systems”, *Operations Research Letters*, 49(4):616–622.

[9]. Hurtado-Lange D., Maguluri S.T. (2022): “Heavy-traffic analysis of queueing systems with no complete resource pooling”, *Mathematics of Operations Research*, 47(4):3129–3155.

[10]. Jain M. (2017): “Priority queue with batch arrival, balking, threshold recovery, unreliable server and optimal service”, *RAIRO – Operations Research*, 51:417–432.

[11]. Ramdani H., Bouchentouf A.A., Yahiaoui L. (2025): “Optimization analysis of unreliable multi-server queueing system with Bernoulli schedule working vacation, threshold-based recovery policy, and impatience*”, RT& A*, 1(82):981-995.

[12]. Sun W., Li S., Guo E.C. (2016): “Equilibrium and optimal balking strategies of customers in Markovian queues with multiple vacations and N-policy”, *Applied Mathematical Modelling*, 40:284–301.

[13]. Yang D.Y., Chung C.H., Wu C.H. (2021): “Sojourn times in a Markovian queue with working breakdowns and delayed working vacations”, *Computers & Industrial Engineering*, 156:107239.

[14]. Zhou X., Tan J., Shroff N. (2018): “Heavy-traffic delay optimality in pull-based load balancing systems: Necessary and sufficient conditions”, *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(3):1–33.

[15]. Zhou X., Wu F., Tan J., Sun Y., Shroff N. (2017): “Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms”, *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 1(2):1–30.