International Journal of Multidisciplinary Research and Growth Evaluation.

Optimizing Service Throughput in Heterogeneous Traffic Streams: A Discrete-Event Simulation of Subsidized Energy Distribution

Sriyanto [1*], **Yudha Arif Budiyani** [2], **M Mujiya Ulkhaq** [3]

[1-3] Department of Industrial Engineering, Diponegoro University, Indonesia

* Corresponding Author: **Sriyanto**

Article Info

ISSN (Online): 2582-7138
Impact Factor (RSIF): 7.98
Volume: 06
Issue: 06
November - December 2025
Received: 01-10-2025
Accepted: 30-10-2025
Published: 03-11-2025
Page No: 883-886

Abstract

In the context of developing urban logistics, the distribution of subsidized energy is often constrained by stochastic demand and significantly heterogeneous vehicle streams—specifically motorcycles versus automobiles—which frequently results in unbalanced server utilization and high customer balking rates. This study aims to optimize service throughput by evaluating the operational performance of segregated versus pooled queuing configurations through a discrete-event simulation (DES) approach. Validated against empirical arrival data (N = 2,329 entities) from a high-volume refueling node modeled as a general queuing network (approx. M/G/2), the research contrasts a rigid dedicated-lane model against a proposed flexible resource-pooling model. Simulation results demonstrate that the resource-pooling configuration significantly outperforms traditional segregated designs in high-density scenarios, achieving a balanced server utilization rate and successfully eliminating customer balking probability. Furthermore, the pooled strategy reduced average waiting times for large vehicles by 52.8%, providing empirical evidence that facility managers in mixed-traffic environments should prioritize dynamic server allocation over rigid physical segregation to maximize service accessibility and operational efficiency.

Keywords: Discrete-Event Simulation, Heterogeneous Traffic Streams, Resource Pooling, Customer Balking, Subsidized Energy Distribution

Introduction

Operational characteristics of service systems in emerging economies differ fundamentally from established Western models, particularly in the domain of transportation and logistics infrastructure. While traditional operations management literature often assumes homogeneous traffic flows dominated by automobiles, urban environments in Southeast Asia are characterized by highly heterogeneous traffic streams where two-wheeled vehicles (motorcycles) constitute a significant proportion of the volume [1, 2]. This heterogeneity presents unique challenges for facility planning, as the service time variability and spatial requirements between motorcycles and automobiles create complex queuing dynamics that standard Erlang models fail to capture accurately.

The distribution of subsidized energy (fuel) represents a critical infrastructure challenge within this context. In many developing nations, fuel subsidies create inelastic demand patterns characterized by extreme stochastic surges, particularly during peak commuting hours [3]. When service nodes reach saturation, the resulting congestion leads to significant operational inefficiencies. A critical behavioral consequence of this saturation is “customer balking”—the decision of an arriving customer not to join the queue due to perceived excessive wait times [4]. In the context of essential public services, high balking rates represent not only lost throughput but also a failure in equitable service delivery.

Despite the prevalence of this issue, there is a paucity of research addressing queuing optimization specifically within mixed-vehicle environments. Existing literature on discrete-event simulation (DES) largely focuses on homogeneous entities or distinct manufacturing processes [5, 6]. Few studies have empirically quantified the operational trade-offs between "segregated service" (dedicated lanes for specific vehicle types) and “resource pooling” (flexible lanes serving all types) in a space-constrained energy distribution node.

The prevailing assumption often favors segregation, yet this rigid approach frequently results in localized bottlenecks where one server is overwhelmed while others sit idle.

This study addresses this gap by employing DES to evaluate optimal queuing configurations for heterogeneous traffic streams. Using empirical data from a high-volume distribution node in Indonesia—a representative example of a motorcycle-dominant economy—we modeled the system as a general queuing network (M/G/2 approximation). The primary objective is to determine whether a flexible resource-pooling strategy can mitigate customer balking and optimize server utilization compared to the traditional segregated approach. By validating the simulation against real-world arrival patterns, this research offers a scalable framework for infrastructure managers to balance efficiency and service equity.

2. Materials and Methods

Materials and Case Context

The research object is a high-volume public refueling station (SPBU 44.502.15) located in a metropolitan area of Indonesia, selected to represent a typical distribution node constrained by heterogeneous traffic streams. The facility operates dedicated islands for subsidized fuel (RON 90), serving two distinct entity classes: motorcycles (two-wheelers) and automobiles (four-wheelers).

The primary material for this study consists of empirical time-series data collected during peak operational windows (15:00–18:00) on weekdays. The dataset comprises N = 2,329 entities, capturing three critical variables: (1) inter-arrival times per vehicle class, (2) service times, and (3) the frequency of balking behavior.

For data processing and modeling, two specific software tools were utilized:

1. **SPSS:** Used to determine the statistical probability distributions of the raw data.
2. **ExtendSim 10:** Used as the Discrete-Event Simulation (DES) environment to construct the baseline and improvement models due to its capability to handle complex logic blocks and attribute-based routing [7].

Simulation methods and experimental design

The study followed a three-stage methodological framework: Input Modeling, Model Construction, and Verification/ Validation [8].

Input Modeling: The raw data underwent Goodness-of-Fit tests using the Kolmogorov-Smirnov (K-S) statistic to identify the best-fitting theoretical distributions. A significance level of $\alpha = 0.05$ was set as the acceptance criterion. These validated distributions (detailed in the Results section) served as the stochastic generators for the simulation [9].

Model Construction (DES): The system was modeled as a general queuing network (approx. M/G/2) with a specific logic block for Balking Behavior [10]. The simulation logic dictates that an entity assesses the queue length (Lq) upon arrival; if Lq exceeds a tolerable threshold (derived from empirical observation), the entity reneges, recorded as a "Lost Customer."

Validation and Scenario Testing: Validation was performed by comparing the simulation output (O_{sim}) against the real-system baseline (O_{real}) using Independent Sample t-Tests. Upon confirming model validity ($p > 0.05$), two experimental scenarios were stress-tested:

- **Model 1 (Baseline):** Rigid segregation (Dedicated lanes for motorcycles vs. cars).
- **Model 2 (Proposed):** Resource pooling (Flexible lanes allowing cross-utilization based on server availability).

Queue performance measures (M/G/2)

The M/G/2 queue model describes a system in which customer arrivals follow a Markovian (Poisson) distribution, meaning that the interarrival times follow an exponential distribution [9]. The service times in this system follow a general distribution, indicating that service times can follow any probability distribution and are not limited to the exponential distribution. Additionally, the system has only two servers that handle all incoming customers.

Customer behavior varies among individuals some choose to remain in line even when the queue is long, while others decide to leave under the same circumstances [10]. However, from the perspective of queuing models, such behavior can only be considered if it can be measured quantitatively and incorporated into mathematical models. Since queuing systems assume uniform behavior among all customers, individual differences cannot be explicitly modeled. A logical way to account for habitual tendencies is by adjusting the service time per customer. In general, there are three types of arrival behaviors in queuing systems: balking, when a customer refuses to enter the queue; reneging, when a customer leaves the queue after waiting; and jockeying, when a customer switches from one queue to another in hopes of receiving faster service. These behaviors help explain variations in queue dynamics and can be represented through specific mathematical formulations [11].

Probability of no customers in the system

$$P_0 = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1}$$

$$P_0 = \left(\frac{1}{1-\rho}\right)^{-1}$$

$$P_0 = 1 - \rho \qquad (1)$$

Average number of customers in the queue

$$L_q = \frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)} \qquad (2)$$

Average number of customers in the system

$$L = \rho + L_q \qquad (3)$$

Average waiting time in the queue

$$W_q = \frac{L_q}{\lambda} \qquad (4)$$

Average time in the system

$$W = W_q + \frac{1}{\mu} \qquad (5)$$

Average time in the system

$$\lambda_{\text{ef}} = \lambda(1-\beta)$$

$$\beta = \frac{\text{Number of customers who left the system}}{\text{Number of arriving customers}} \quad (6)$$

Legend

P_0 = probability there are no customers in the system
ρ = utilization
n = number of customers
σ = standard deviation of service time
λ = average arrival rate
λ_{ef} = average effective arrival rate
μ = average service rate
L_q = average number of customers in the queue
W_q = average waiting time in the queue
L = average number of customers in the system
W = average time in the system

3. Results

Input analysis and traffic characteristics

The heterogeneity of the traffic stream is confirmed through the distribution fitting analysis. As presented in Table 1, while the arrival patterns for both vehicle classes follow a standard Exponential distribution (consistent with Poisson arrival processes), the service times exhibit distinct characteristics. Motorcycles follow a Beta distribution, reflecting the variability in tank sizes and payment readiness, whereas automobiles follow a Log-Logistic distribution. These validated parameters served as the stochastic inputs for the simulation model.

Model verification and validation

Before evaluating alternative scenarios, the baseline simulation model was subjected to rigorous validation to ensure it accurately represented the real-world system. A "warm-up period" was implemented to eliminate initial bias, followed by 30 replications. The validation process compared the simulation output (O_{sim}) against empirical field data (O_{real}) for the critical metric: Average Number of Customers in the System (L_s).

An Independent Sample *t*-test (Table 2) revealed no statistically significant difference between the simulation and real-world data (p-value > 0.05). The deviation margin was recorded at less than 5%, confirming that the DES model is a valid proxy for predicting system behavior under new configurations.

Comparative Analysis: Segregated vs. Resource Pooling

The core analysis contrasts two distinct operational configurations:

- **Scenario A (Segregated/Status Quo):** Dedicated servers for motorcycles and automobiles.
- **Scenario B (Resource Pooling):** A flexible server configuration where specific islands can serve both vehicle types dynamically based on availability.

Table 3 presents the performance comparison. In the Segregated scenario, the system exhibits severe imbalance; the motorcycle server is over-utilized ($p \approx 0.99$), leading to high congestion, while the automobile server remains under-utilized. The introduction of Resource Pooling (Scenario B) drastically alters this dynamic. The most significant impact is observed in the automobile stream, where average waiting time (W_q) plummeted by 52.8% (from 273.2s to 128.9s).

Table 1: Stochastic Parameters for Simulation Input

Entity Type	Variable	Distribution	Parameters	K-S Rank
Motorcycles	Inter-arrival Time	Exponential	$\lambda = 0.0045$	1
	Service Time	Beta	$\alpha_1 = 1.5$, $\alpha_2 = 2.2$, min = 15, max = 120	1
Automobiles	Inter-arrival Time	Exponential	$\lambda = 0.0305$	1
	Service Time	Log-Logistic	$\alpha = 45.2$, $\beta = 3.1$	1

Table 2: Validation Results (Real System vs. Simulation)

Metric	Real System (Oreal)	Simulation Output (Osim)	Difference (%)	Conclusion
Avg. Queue Length	9.04 vehicles	9.12 vehicles	+0.88%	Valid
Avg. Waiting Time	294.58 sec	291.40 sec	-1.08%	Valid

Table 3: Performance Comparison of Queuing Configurations

Performance Metric	Scenario A (Segregated)	Scenario B (Resource Pooling)	Improvement / Change
Automobile Waiting Time (W_q)	273.2 sec	128.9 sec	+52.8% (Improved)
Motorcycle Waiting Time (W_q)	294.6 sec	301.2 sec	-2.2% (Trade-off)
Server Utilization Balance	Imbalanced (0.99 vs 0.49)	Balanced (0.74 vs 0.74)	Optimal Load Leveling
Balking Probability (P_b)	4.0% (High Loss)	0.0% (Zero Loss)	Eliminated

4. Discussion

The Mechanism of Resource Pooling

The substantial reduction in automobile waiting times confirms the theoretical advantage of resource pooling in stochastic systems. By allowing servers to act as a shared resource, the system absorbs the variability of arrival bursts more effectively. In Scenario A, an idle car server could not assist a queued motorcycle, creating "artificial scarcity." Scenario B eliminates this inefficiency, ensuring that no server sits idle while demand exists. Applying decision intelligence principles allows managers to interpret simulation outcomes for strategic interventions, improving service efficiency and equity[11].

The Efficiency-Equity Trade-Off

An interesting phenomenon was observed regarding motorcycle service times, which saw a marginal increase (approx. 2%). This represents a classic "Efficiency vs. Equity" trade-off in operations management. While the pooled system is vastly more efficient for the aggregate system (total throughput increases), smaller entities (motorcycles) may experience slightly longer service times due to the presence of larger entities (cars) in the shared stream. However, this trade-off is justifiable given the elimination of the "bottleneck" effect for cars.

Mitigating Balking Behavior

Perhaps the most critical finding for revenue retention is the reduction of Balking Probability (P_b) to zero. In the segregated model, visual congestion in the motorcycle lane triggered a 4% balking rate. The pooled model dissipates these long queues by distributing entities across multiple lanes. Psychologically, this faster-moving (or physically shorter) queue reduces the anxiety that triggers balking, ensuring the facility captures 100% of the potential demand.

5. Conclusion

The simulation results provide critical insights into the operational dynamics of heterogeneous traffic systems, confirming the theoretical advantage of resource pooling over rigid segregation. By allowing servers to act as a shared resource, the proposed model eliminates the "artificial scarcity" observed in the baseline scenario, where idle car servers frequently could not assist queued motorcycles. Although this reconfiguration introduces a marginal 2% increase in motorcycle service times, reflecting a classic trade-off between aggregate efficiency and service equity, this cost is operationally justifiable given the substantial reduction in automobile delays and the elimination of system-wide bottlenecks.

Crucially, the study identifies a significant behavioral impact regarding revenue retention: the reduction of Balking Probability (P_b) to zero. In the segregated model, visual congestion in single-use lanes triggered a 4% balking rate due to perceived excessive waiting times. The pooled model dissipates these long queues by distributing entities across multiple lanes, effectively reducing the visual anxiety that drives customers away. This confirms that flexible server allocation not only optimizes physical throughput but also mitigates the psychological triggers of lost demand, ensuring the facility captures 100% of its potential market.

6. References

1. Alnowibet KA, Khireldin A, Abdelawwad M, Mohamed AW. Airport terminal building capacity evaluation using queuing system. Alexandria Eng J. 2022;61(12):10109–18. doi: 10.1016/j.aej.2022.03.055.
2. Varshosaz F, Moazzami M, Fani B, Siano P. Day-ahead capacity estimation and power management of a charging station based on queuing theory. IEEE Trans Ind Inform. 2019;15(10):5561–74. doi: 10.1109/TII.2019.2906650.
3. Xu G, Xu M, Wang Y, Liu Y, Assogba K. Optimization of energy supply system under information variations based on gas stations queuing analyses. Syst Sci Control Eng. 2018;6(2):10–23. doi: 10.1080/21642583.2018.1480434.
4. Economou A, Logothetis D, Manou A. The value of reneging for strategic customers in queueing systems with server vacations/failures. Eur J Oper Res. 2022;299(3):960–76. doi: 10.1016/j.ejor.2022.01.010.
5. Zhao C, Wang Z. The impact of line-sitting on a two-server queueing system. Eur J Oper Res. 2023;308(2):782–800. doi: 10.1016/j.ejor.2022.12.016.
6. Kim B, Kim J. The waiting time distribution for a correlated queue with exponential interarrival and service times. Oper Res Lett. 2018;46(2):268–71. doi: 10.1016/j.orl.2018.02.001.
7. Diamond B, Lamperti S, Nastasi A, Tag P. Extendsim user reference. San Jose: Imagine That Inc.; 2022.
8. Sargent RG. Verification and validation of simulation models. In: Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME, editors. Proceedings of the 2013 Winter Simulation Conference. Piscataway: IEEE; 2013. p. 321–7.
9. Frederick SH, Gerald JL. Introduction to operations research. 10th ed. New York: McGraw-Hill Education; 2016.
10. Wang Y, Guo J, Ceder AA, Currie G, Dong W, Yuan H. Waiting for public transport services: queueing analysis with balking and reneging behaviors of impatient passengers. Transp Res Part B Methodol. 2014;63:53–76. doi: 10.1016/j.trb.2014.02.004.
11. Tasleem N, Raghav RS, Ansari MN. A decision intelligence framework: integrating human intuition with AI models. J Artif Intell Gen Sci. 2024; (in press or provide volume/issue if available).

How to Cite This Article

Sriyanto, Budiyani YA, Ulkhaq MM. Optimizing service throughput in heterogeneous traffic streams: a discrete-event simulation of subsidized energy distribution. Int J Multidiscip Res Growth Eval. 2025;6(6):883-6.

Creative Commons (CC) License