

DISCRETE-EVENT SIMULATION AND QUEUING ANALYSIS OF LANDSIDE TAXI DISPATCH LOGISTICS AT KLIA TERMINAL 1

Aimi Zulliyana binti Rosli^{1*}

Noor Athirah binti Zakaria²

W Nurfahizul Ifwah binti W Alias³

Nurul Suhada binti Aziz⁴

Wan Nurul Husna binti Wan Nordin⁵

Suziana Aida binti Othman⁶

Rahaidah binti Muhammad⁷

¹²³⁴⁵⁶⁷ Faculty of Computer and Mathematical Sciences (FSKM), University Teknologi MARA, Kelantan Branch

*Corresponding author: Aimi Zulliyana binti Rosli (Email : zullia068@uitm.edu.my)

Article history

Received date : 10-9-2026

Revised date : 10-9-2026

Accepted date : 17-10-2026

Published date : 25-10-2026

To cite this document:

Rosli, A. Z., Zakaria, N. A., W Alias, W. N. I., Aziz, N. S., Wan Nordin, W. N. H., Othman, S. A., & Muhammad, R. (2026). Discrete-event simulation and queuing analysis of landside taxi dispatch logistics at KLIA Terminal 1. *Journal of Islamic, Social, Economics and Development (JISED)*, 11 (87), 1000-1017.

Abstract: Ground transportation systems are critical for managing fluctuating passenger volumes at major international airports such as Kuala Lumpur International Airport (KLIA) Terminal 1. During peak flight periods, sudden surges in passenger arrivals may create bottlenecks at taxi loading bays, resulting in long passenger waiting times and queue growth. This study evaluates taxi queueing dynamics at KLIA Terminal 1 under assumed demand and service conditions and examines potential operational strategies for reducing congestion. A hybrid approach combining an analytical M/M/12 queueing model and discrete-event simulation was employed. The simulation model was developed in Python using the SimPy framework. For the simulated peak-demand scenario, the assumed arrival rate was 300 passenger groups per hour, with 12 loading bays and an average service rate of 20 groups per hour per bay. This produced a traffic intensity of $\rho = 1.25$, indicating that demand exceeded service capacity under these assumptions. In the two-hour baseline simulation, the maximum queue length reached 75 passenger groups, while the maximum waiting time was 17.37 minutes. When the assumed service time was reduced to 2.5 minutes per group, equivalent to $\mu = 24$ groups per hour per bay, the simulated maximum waiting time fell to 2.75 minutes and the maximum queue length fell to 15 groups. However, traffic intensity remained above one ($\rho = 1.0417$), so this scenario did not achieve steady-state stability. These findings suggest that faster dispatch and loading processes may reduce congestion during a finite peak period, but further evaluation using observed operational data is needed before drawing conclusions about long-term queue stability or recommending a specific intervention.

Keywords: discrete-event simulation, M/M/12 queueing model, taxi dispatch, airport operations, KLIA Terminal 1

Introduction

Airport taxi services are a crucial component of ground transportation, connecting arriving passengers to their final destinations after baggage collection. Unlike conventional public transport services, airport taxi platforms operate as multi-server queuing systems that must accommodate highly variable and concentrated passenger demand. Landside taxi staging is therefore an important component of operations at Kuala Lumpur International Airport (KLIA) Terminal 1, where efficient passenger-to-taxi dispatching is necessary to prevent congestion from extending into other landside passenger-processing areas.

In the peak-demand scenario examined in this study, passenger-group arrivals may exceed the combined service capacity of the taxi loading bays, causing queues and waiting times to increase. When several large aircraft arrive within a short period, a large number of passengers may enter the landside transportation area almost simultaneously. This sudden increase in passenger demand can exceed the service capacity of the available taxi loading bays. Bottlenecks may develop because of manual coupon-processing procedures, delays in loading luggage. Consequently, passenger queues may grow when the arrival rate exceeds the loading bays' service capacity, increasing waiting times during the modelled peak-demand period. Because observed arrival times, service durations and queue lengths for the KLIA Terminal 1 taxi platform were unavailable, this study examines assumed demand and service conditions rather than documenting an observed congestion event.

To address these operational challenges, this study integrates classical steady-state queuing theory with discrete-event simulation (DES) to model and evaluate the landside taxi dispatch system at KLIA Terminal 1. The analytical component uses a multi-server M/M/c queuing model in which passenger arrivals follow a Poisson process with arrival rate λ and service times follow an exponential distribution with service rate μ ; the baseline scenario assumes $c = 12$ parallel taxi loading bays. This model provides baseline performance measures, including the expected passenger waiting time in the queue (W_q), expected queue length (L_q) and overall loading-bay utilisation (ρ).

Nevertheless, when the passenger arrival rate exceeds the total service capacity and traffic intensity reaches or exceeds unity ($\rho \geq 1$), the system does not have a finite steady-state solution. Therefore, although the M/M/12 model identifies when no finite steady-state solution exists, it does not show how the queue develops during a temporary demand surge. A discrete-event simulation model was developed using Python and SimPy to examine these changes over a finite period. The simulation tracks passenger arrivals and taxi-service events chronologically over a finite peak-demand period, allowing queue formation and dissipation to be examined over time. It also enables alternative operational strategies, including the addition of taxi loading bays and the introduction of a digital dispatch system, to be evaluated in a virtual environment without disrupting ongoing terminal operations or requiring premature capital expenditure.

Literature Review

Changes in airport passenger demand make landside queue management an important consideration in ground transportation planning. At taxi loading bays, short periods of high passenger-group arrivals may increase waiting times when demand exceeds the assumed service capacity. This section reviews three areas relevant to the present model: multi-server M/M/c queuing models, discrete-event simulation, and operational bottlenecks in airport taxi dispatch systems.

Multi-Server M/M/c Queuing Models

Multi-server queuing models have been widely applied to represent service systems with several parallel service channels (Shortle et al., 2018; Wang et al., 2022). In airport taxi operations, an M/M/c model can be used when passengers are served through multiple taxi loading bays. The first “M” indicates that passenger arrivals follow a Poisson process, or equivalently that interarrival times are exponentially distributed. The second “M” indicates that service times are exponentially distributed, while c represents the number of identical parallel servers.

The M/M/c model provides closed-form expressions for performance measures such as the expected queue length (L_q), expected waiting time in the queue (W_q), and server utilisation (ρ). For a system with c servers, traffic intensity is expressed as:

$$\rho = \frac{\lambda}{c\mu}$$

where λ is the passenger arrival rate and μ is the service rate of each server. A finite steady-state distribution exists only when $\rho < 1$. When $\rho \geq 1$, the arrival rate is equal to or greater than the total service capacity, and the system does not converge to a finite steady-state distribution. Although the M/M/c model is useful for evaluating baseline system performance, it assumes stationary arrival and service processes over a sufficiently long period. These assumptions may not adequately represent airport operations because passenger demand can vary significantly according to flight schedules, aircraft capacity, immigration processing and baggage-retrieval times (Balakrishnan & Simaiakis, 2014; Oprea et al., 2024). During temporary peak-demand periods, the model cannot show the minute-by-minute accumulation and subsequent dissipation of passenger queues. Therefore, a dynamic modelling approach is required to complement the analytical results.

Discrete-Event Simulation and the SimPy Framework

Discrete-event simulation (DES) represents a system as a chronological sequence of events. It has been applied to airport processes such as security screening and passenger-flow analysis (Dorton, 2011; Oprea et al., 2024). Unlike a steady-state analytical model, DES can represent a specified operating period, stochastic service times and operational rules. These features make it useful for examining how a queue develops under the assumed demand conditions in this study.

Related simulation research has examined how changes in passenger demand affect airport terminal capacity. For example, Inturri et al. (2026) used agent-based simulation to assess passenger flows and capacity at Comiso Airport. Although their method and terminal setting differ from the present taxi-dispatch model, the study illustrates the value of testing airport capacity under alternative demand scenarios. Lee et al. (2025) combined dynamic rescheduling with discrete-event simulation to examine passenger congestion at airport security screening areas. Although their study focused on security screening rather than taxi dispatch, it demonstrates how simulation can support the evaluation of operational changes under varying passenger demand.

This approach is particularly useful for examining temporary periods in which $\rho \geq 1$. Although such a system has no finite steady-state solution, DES can still estimate queue growth and

passenger waiting times over a specified simulation period. It can also show whether the queue decreases after passenger demand returns below the system's service capacity.

SimPy is an open-source, process-based DES framework implemented in Python. It allows researchers to represent passengers, taxis and loading bays as processes, entities and shared resources. Researchers can use SimPy to implement customised dispatch rules, stochastic arrival and service-time distributions, and changes in resource capacity. Consequently, alternative operational strategies can be evaluated virtually before they are implemented in the actual system. Nevertheless, the reliability of simulation results depends on appropriate input data, verification, validation, repeated simulation runs and statistical analysis of the outputs.

Airport Taxi Queuing Dynamics and Operational Bottlenecks

Airport taxi platforms generally experience fluctuating passenger demand and operate under physical space and capacity constraints (Jia et al., 2022; Qie et al., 2021; Wang et al., 2022). Their performance depends not only on the number of loading bays but also on taxi availability, passenger-processing procedures, luggage-loading times and dispatch coordination. A bottleneck may develop when one or more of these processes cannot accommodate the passenger arrival rate. In the broader context of airport terminal planning, Pivac and Bartulović (2026) assessed facility capacity using peak-hour demand and spatial requirements, illustrating why demand and available space should be considered together. Jenčová et al. (2025) used simulation to identify bottlenecks in airport passenger-handling processes. Their study concerned terminal processes rather than taxi loading, but it highlights the importance of considering service capacity when examining passenger-flow constraints.

Process improvements may reduce the time required for passenger registration, taxi assignment and luggage loading. In this study, shorter service durations are examined as hypothetical simulation scenarios rather than measured effects of a specific intervention. Queueing analysis and discrete-event simulation are used to examine how those assumed durations affect waiting times and queue lengths.

Analytical queueing models have also been applied to taxi and transportation systems (Jia et al., 2022; Qie et al., 2021; Wang et al., 2022). For example, single-server models such as $M/E_k/1$ may be appropriate when customers are processed through one service channel and service times follow an Erlang distribution. However, a single-server model does not directly represent an airport taxi platform in which several loading bays serve passengers simultaneously. A multi-server model is therefore more appropriate when the loading bays operate as parallel service channels.

Nevertheless, the application of a multi-server model depends on how a server is defined. If each loading bay independently completes the entire passenger-loading process, each bay may be treated as a server. Conversely, if passenger registration, taxi assignment and luggage loading occur at separate sequential stations, the system may need to be represented as a multi-stage queueing network rather than a single $M/M/c$ queue.

Synthesis of Previous Studies and Research Gap

Existing studies demonstrate that analytical queueing models provide useful baseline measures of system stability and long-run performance, whereas discrete-event simulation can represent transient demand, finite operating periods and more complex dispatching rules. However, both approaches have limitations when used independently. An analytical $M/M/c$ model may

not adequately represent temporary passenger surges, while a simulation model requires appropriate input data, verification and validation to produce credible results.

Based on the literature reviewed in this study, the combined application of a multi-server queuing model and finite-period discrete-event simulation appears to be relevant for evaluating landside taxi dispatch operations. Therefore, this study applies an M/M/12 analytical model together with a SimPy-based discrete-event simulation to examine baseline and peak-demand conditions at KLIA Terminal 1. The combined approach is used to evaluate system performance and compare alternative operational strategies, including the addition of taxi loading bays and improvement of the service rate through a digital dispatch system.

Methodology

This study employed a quantitative hybrid approach combining secondary operational data, an analytical multi-server queuing model and discrete-event simulation (DES). An M/M/12 model was used to evaluate the steady-state performance and stability of the landside taxi dispatch system at KLIA Terminal 1. A finite-period simulation was subsequently developed using Python and the SimPy framework to evaluate transient queue behaviour during peak passenger-arrival periods and compare alternative operational strategies.

Data Sources and Parameter Estimation

Secondary data were obtained from administrative records, aviation reports and publicly available transportation sources. According to the APAD Airport Taxi Registry (Agensi Pengangkutan Awam Darat, 2020), 857 active vehicles were registered for operations associated with KLIA Terminal 1, compared with 1,946 active airport-taxi licences recorded nationally. The registry indicated that 767 vehicles, representing 89.50% of the identified KLIA fleet, were standard four-seat vehicles. Other recorded vehicle categories included six-, eight- and nine-seat vehicles. Therefore, the modal registered vehicle capacity was four passengers.

The APAD vehicle registry provides information on registered airport taxis but does not contain time-stamped observations of passenger-group arrivals, service durations or queue lengths at the KLIA Terminal 1 taxi platform. Therefore, taxi-queue arrival rates could not be estimated separately for peak hours, off-peak hours or full-day operations. The demand and service rates examined in this study were specified as modelling assumptions.

The arrival rate of $\lambda = 300$ passenger groups per hour was specified as a hypothetical extreme-demand scenario and applied over the 120-minute simulation period. The baseline configuration of 12 loading bays was also assumed for modelling purposes; the APAD vehicle registry does not establish the number of taxi loading bays at KLIA Terminal 1. Any exceedance of loading-bay service capacity under this configuration is a result of the scenario assumptions, rather than an observed capacity violation at KLIA Terminal 1. The simulation baseline service rate was set at $\mu = 20$ passenger groups per hour per bay, equivalent to an assumed mean service duration of three minutes per group. These parameters were treated as modelling assumptions rather than empirically estimated KLIA operational parameters. Table 1 summarises the parameters used in the analytical and simulation models.

Table 1: The parameters used in the analytical and simulation models

Parameter	Symbol	Value	Unit	Basis
Assumed baseline number of loading bays	c	12	bays	Modelling assumption
Peak arrival rate	λ_p	300	passenger groups/hour	Assumed extreme-demand scenario
Baseline service rate	μ_0	20	passenger groups/hour/bay	Assumed 3.0-minute service duration
Strategy 1 capacity	c_1	14	bays	Simulated scenario
Strategy 2 mean service durations	-	2.9, 2.8 & 2.5	minutes/group	Simulated scenarios
Simulation horizon	TT	120	minutes	Study design

The values assigned to the improvement strategies were treated as simulation assumptions unless supported by observed operational data.

System Definition and Modelling Assumptions

The model represents the process from the arrival of a passenger group at the taxi queue until the group completes boarding and leaves the taxi loading bay. Each passenger group is treated as one customer because a single taxi-loading transaction may involve more than one individual travelling together. Each loading bay is treated as one server, and the 12 bays are assumed to operate independently and in parallel.

The model adopts the following assumptions:

1. Passenger-group arrivals follow a Poisson process.
2. Interarrival times are exponentially distributed.
3. Service times are exponentially distributed.
4. The 12 loading bays have identical service rates.
5. Passenger groups are served according to a first-come, first-served discipline.
6. Queue capacity and the potential customer population are assumed to be unlimited for modelling purposes; physical space constraints at the taxi platform are not represented.
7. No passenger group leaves the queue before receiving service.
8. A taxi is available whenever a passenger group reaches an available loading bay.
9. Bay closures, vehicle breakdowns and road traffic conditions are not modelled.

The assumption of continuous taxi availability means that the model evaluates the capacity of the loading and dispatch process rather than taxi-supply shortages. If taxi availability is modelled separately, taxis must be represented as an additional entity or resource.

Analytical M/M/12 Queuing Model

The landside taxi dispatch system was represented as an M/M/12 queue with a first-come, first-served discipline. Passenger-group arrivals were modelled as a Poisson process with

arrival rate λ , while service times were assumed to be exponentially distributed with service rate μ at each of the 12 parallel loading bays.

The traffic intensity is calculated as:

$$\rho = \frac{\lambda}{c\mu} \quad (1)$$

where $c=12$. A finite steady-state distribution exists only when:

$$\rho < 1$$

The probability that no customers are present in the system is calculated as:

$$P_0 = \left[\sum_{k=0}^{c-1} \frac{(\lambda/\mu)^k}{k!} + \frac{(\lambda/\mu)^c}{c!(1-\rho)} \right]^{-1} \quad (2)$$

The expected number of passenger groups waiting in the queue is:

$$L_q = P_0 \frac{(\lambda/\mu)^c \rho}{c!(1-\rho)^2} \quad (3)$$

Using Little's Law, the expected waiting time in the queue is:

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

When λ is measured in passenger groups per hour, W_q is initially obtained in hours. It is converted into minutes using:

$$W_q(\text{minutes}) = \frac{L_q}{\lambda} \times 60 \quad (5)$$

Equations (2)–(5) were applied only to scenarios satisfying $\rho < 1$. When $\rho \geq 1$, the system has no finite steady-state solution; therefore, its finite-period behaviour was evaluated using DES.

Discrete-Event Simulation Development

The DES model was implemented in Python using the SimPy framework. A `simpy.Environment` instance controlled the simulation clock and chronological execution of events. The 12 taxi loading bays were represented as a shared `simpy.Resource` with a capacity of 12.

Passenger-group interarrival times were generated from an exponential distribution using:

$$T_a \sim \text{Exponential}(\lambda_{\min})$$

where:

$$\lambda_{min} = \frac{\lambda}{60}$$

Service times were generated using:

$$T_s \sim \text{Exponential}(\mu_{min})$$

where:

$$\mu_{min} = \frac{\mu}{60}$$

For each passenger group, the following sequence of events was simulated:

1. The next interarrival time was generated.
2. The passenger group entered the system, and its arrival time was recorded.
3. The passenger group requested access to a taxi loading bay.
4. If all loading bays were occupied, the passenger group joined the first-come, first-served queue.
5. When a loading bay became available, the service-start time was recorded, and the waiting time was calculated.
6. The boarding and luggage-loading duration was generated.
7. The loading bay was released upon service completion.
8. For passenger groups completing service within the simulation horizon, the departure time and relevant performance measures were recorded.

Each scenario was simulated over a fixed 120-minute period, and the simulation was terminated at the end of this horizon. Consequently, the reported results represent finite-period system performance and should not be interpreted as complete processing of all passenger groups arriving during the simulation. The experimental scenarios evaluated in this study are presented in Table 2.

Table 2: Experimental Scenarios

Scenario	Arrival rate	Bays	Mean service duration
Baseline peak	300 groups/hour	12	3.0 minutes/group
Strategy 1: bay expansion	300 groups/hour	14	3.0 minutes/group
Strategy 2A: shorter service duration	300 groups/hour	12	2.9 minutes/group
Strategy 2B: shorter service duration	300 groups/hour	12	2.8 minutes/group
Strategy 2C: shorter service duration	300 groups/hour	12	2.5 minutes/group

Strategy 1 represented the addition of two loading bays, increasing capacity from 12 to 14 bays. Strategies 2A, 2B and 2C represented hypothetical reductions in mean service duration from the baseline of 3.0 minutes per passenger group to 2.9, 2.8 and 2.5 minutes, respectively, while retaining 12 bays. These reductions were modelling assumptions rather than measured effects of a digital dispatch system.

The performance of each scenario was evaluated using:

- Mean and maximum passenger-group waiting times
- Time-weighted mean and maximum queue lengths
- Loading-bay utilisation
- Throughput
- Number of passenger groups remaining at the end of the arrival period

Each scenario should be run for several independent replications using different random seeds. Results should be reported using the sample mean, standard deviation and 95% confidence interval. Reporting results from only one stochastic simulation run would not be sufficient.

Verification and Validation

The simulation code was verified by checking event sequences, resource allocation, queue discipline and the conservation of passenger groups entering and leaving the system. For stable scenarios, simulation results were compared with the analytical M/M/12 results.

Model validation should be conducted by comparing the simulated waiting times, queue lengths or throughput with observed operational data. If no observed KLIA queue data or expert assessment is available, the study should describe this process as analytical or conceptual verification rather than empirical validation.

Results and Discussion

This section presents the results obtained from the analytical M/M/12 queuing model and the finite-period discrete-event simulation. The analytical model was used to evaluate system stability under different passenger-arrival rates, whereas the simulation was used to examine queue development over a specified 120-minute peak-demand period. Because the analytical and simulation components used different service-rate assumptions, their results are presented as separate analyses rather than as direct numerical comparisons.

Analytical Evaluation across Demand Scenarios

The M/M/12 equations were applied to three passenger-demand scenarios to examine the relationship between passenger arrival rates, total service capacity and system stability. The analytical results are presented in Table 3.

Table 3: Analytical M/M/12 results across passenger-arrival scenarios

Operational metric	Scenario A: Off-peak	Scenario B: Baseline peak	Scenario C: Extreme surge
Arrival rate, λ	20 groups/hour	140 groups/hour	300 groups/hour
Service rate per bay, μ	15 groups/hour	15 groups/hour	15 groups/hour
Number of loading bays, c	12	12	12
Traffic intensity, ρ	0.1111	0.7778	1.6667
Empty-system	0.2636	0.0001	Not applicable

Under Scenario A, the traffic intensity was $\rho=0.1111$, indicating that the system operated substantially below its total service capacity. The expected queue length and waiting time were both close to zero. However, these values represent very small theoretical averages and should not be interpreted as a guarantee that every passenger group receives immediate service.

Under Scenario B, the traffic intensity increased to $\rho=0.7778$. Because this value remained below unity, the system had a finite steady-state solution. The expected queue length was approximately 1.82 passenger groups, while the expected waiting time was approximately 0.78 minutes.

Under Scenario C, the arrival rate increased to 300 passenger groups per hour, exceeding the total analytical service capacity of:

$$c\mu=12(15)=180 \text{ groups per hour.}$$

This resulted in:

$$\rho = \frac{300}{180} = 1.6667$$

Because $\rho>1$, the system did not have a finite steady-state distribution. Therefore, P_0 , L_q and W_q could not be calculated as finite steady-state measures for this scenario. This result does not represent a failure of queuing theory; instead, it indicates that the assumed demand exceeded the system's long-run service capacity.

Finite-Period Simulation Results

A discrete-event simulation was conducted to examine queue behaviour over a finite 120-minute peak-demand period. The simulation used an arrival rate of 300 passenger groups per hour and an initial configuration of 12 loading bays. Unlike the analytical analysis in Table 3, the simulation assumed a baseline service rate of 20 passenger groups per hour per bay, equivalent to an average service duration of 3.0 minutes.

The baseline configuration was compared with Strategy 1, which increased the number of loading bays from 12 to 14, and three process-acceleration scenarios that reduced the assumed average service duration. The preliminary single-run results obtained using RANDOM_SEED = 42 are presented in Table 4.

Table 4: Preliminary single-run results

Performance measure	Baseline: 12 bays, 3.0 min	Strategy 1: 14 bays, 3.0 min	Strategy 2A: 12 bays, 2.9 min	Strategy 2B: 12 bays, 2.8 min	Strategy 2C: 12 bays, 2.5 min
Mean waiting time	7.43 min	8.36 min	4.82 min	2.65 min	0.62 min
Maximum waiting time	17.37 min	18.62 min	12.15 min	7.40 min	2.75 min
Maximum queue length	75 groups	78 groups	52 groups	34 groups	15 groups
Traffic intensity, ρ	1.2500	1.0714	1.2083	1.1667	1.0417
Analytical stability status	No finite steady state	No finite steady state	No finite steady state	No finite steady state	No finite steady state

The results in Table 4 are based on one stochastic simulation run for each scenario. They describe queue performance during the specified 120-minute period and should be interpreted as preliminary findings. Multiple independent simulation runs and uncertainty analysis are needed before the results can support a definitive comparison between strategies.

Baseline Simulation Performance

Under the baseline simulation configuration, the mean waiting time was 7.43 minutes, while the maximum recorded waiting time was 17.37 minutes. The maximum queue length reached 75 passenger groups during the 120-minute simulation period.

For this configuration, the total service capacity was:

$$12(20)=240 \text{ groups per hour,}$$

while the arrival rate was 300 groups per hour. Therefore:

$$\rho = \frac{300}{240} = 1.25$$

This value indicates that the assumed arrival rate exceeded the total service capacity. Although the simulation produced finite waiting-time and queue-length values, these values resulted from the finite 120-minute simulation horizon. They should not be interpreted as evidence that the system was stable.

Strategy 1: Increasing the Number of Loading Bays

Strategy 1 increased the number of loading bays from 12 to 14 while maintaining an average service duration of 3.0 minutes. The resulting total service capacity was:

$$14(20)=280 \text{ groups per hour.}$$

The corresponding traffic intensity was:

$$\rho = \frac{300}{280} = 1.0714$$

Although Strategy 1 increased the total service capacity, it remained below the assumed arrival rate of 300 groups per hour. Therefore, this configuration did not satisfy the steady-state stability condition.

In the reported simulation run, Strategy 1 produced a mean waiting time of 8.36 minutes and a maximum queue length of 78 passenger groups. These values were unexpectedly higher than the baseline results. Because the model did not explicitly include vehicle-merging conflicts, lane-crossover delays or physical interference between loading bays, the increase cannot be attributed to these operational mechanisms.

The unexpected result may reflect stochastic variation, inconsistent random-number streams, the simulation termination rule or an implementation issue. Therefore, additional model verification and repeated simulation runs are required before drawing conclusions about the effectiveness of Strategy 1.

Strategy 2: Reducing Service Duration

The Strategy 2 scenarios examined the effect of progressively reducing the average service duration while retaining 12 loading bays. In the reported simulation run, shorter service durations were associated with lower waiting times and shorter maximum queues.

Reducing the average service duration from 3.0 to 2.9 minutes decreased the reported mean waiting time from 7.43 to 4.82 minutes. The maximum queue length also decreased from 75 to 52 passenger groups. When the service duration was reduced to 2.8 minutes, the reported mean waiting time decreased to 2.65 minutes, while the maximum queue length decreased to 34 groups.

Strategy 2C, which used an average service duration of 2.5 minutes, produced the lowest values in the reported run. The mean waiting time was 0.62 minutes, the maximum waiting time was 2.75 minutes and the maximum queue length was 15 passenger groups.

Relative to the baseline run, the reported reduction in mean waiting time was:

$$\frac{7.43 - 0.62}{7.43} \times 100 = 91.66\%$$

The reported reduction in maximum queue length was:

$$\frac{75 - 15}{75} \times 100 = 80.00\%$$

Nevertheless, these percentages describe only the difference between the reported single-run results. They should not be interpreted as generalisable performance improvements without repeated simulations and uncertainty analysis.

Stability Interpretation of Strategy 2C

An average service duration of 2.5 minutes corresponds to:

$$\mu = \frac{60}{2.5} = 24 \text{ groups per hour per bay.}$$

Therefore, the traffic intensity for Strategy 2C was:

$$\rho = \frac{300}{12(24)} = 1.0417$$

Although Strategy 2C produced substantially lower finite-period queue measures, its traffic intensity remained above unity. It should therefore not be described as fully stable or steady-state stable. A more appropriate interpretation is that the assumed reduction in service duration slowed queue accumulation and improved performance during the specified 120-minute period.

To obtain $\rho=0.8929$, the required service rate would be 28 passenger groups per hour per bay:

$$\rho = \frac{300}{12(28)} = 0.8929$$

This is equivalent to an average service duration of:

$$\frac{60}{28} = 2.14 \text{ minutes}$$

Therefore, an additional scenario using an average service duration of approximately 2.14 minutes would need to be simulated before concluding that Strategy 2 achieves steady-state stability.

Overall Discussion

The analytical results demonstrate that system stability depends on the relationship between passenger demand and total service capacity. When $\rho < 1$, the M/M/12 model provides finite steady-state performance measures. When $\rho \geq 1$, DES remains useful for examining queue behaviour over a finite period; however, finite simulation results do not indicate long-run system stability.

The preliminary simulation results suggest that reducing service duration may improve waiting times and queue lengths during a finite peak-demand period. However, the unexpected result obtained for Strategy 1 and the use of only one random seed prevent a definitive comparison between physical capacity expansion and process acceleration.

Consequently, the present results are insufficient to establish that process acceleration is more effective than infrastructure expansion. Multiple independent simulation replications, consistent random-number streams, model verification and statistical analysis using 95% confidence intervals are required before a definitive operational recommendation can be made.

Conclusion and Recommendations

This study combined a classical steady-state M/M/12 queuing model with discrete-event simulation to evaluate landside passenger-to-taxi operations at Kuala Lumpur International Airport (KLIA) Terminal 1. Secondary administrative and aviation-related data were used to estimate selected model parameters, while the simulation provided a virtual representation of queue development during a finite peak-demand period.

The principal mathematical finding is that queue stability depends on the ratio of passenger-group arrivals to the combined service capacity of the loading bays, $\rho = \lambda / (c\mu)$. In the analytical extreme-surge scenario, $\lambda = 300$ groups per hour, $c = 12$ and $\mu = 15$ groups per hour per bay give $\rho = 1.6667$; hence, no finite steady-state queue measures exist. In the separate 120-minute simulation, the baseline configuration produced a maximum queue of 75 groups and a maximum waiting time of 17.37 minutes. These simulation results describe one finite-period run and require further verification before they can support operational decisions.

Key Findings

The analytical M/M/12 model provided a useful reference for examining the relationship between passenger demand and total service capacity. Under the analytical extreme-surge scenario, an arrival rate of 300 passenger groups per hour and a service rate of 15 passenger groups per hour per bay produced:

$$\rho = \frac{300}{12(15)} = 1.6667$$

Because $\rho > 1$, the system did not have a finite steady-state solution under these assumptions. This finding indicates that the assumed arrival rate exceeded the total analytical service capacity.

The discrete-event simulation was used to examine queue development over a finite 120-minute peak-demand period. Under the baseline simulation configuration, the mean waiting time was 7.43 minutes, the maximum waiting time was 17.37 minutes and the maximum queue length reached 75 passenger groups. These results describe the system's behaviour within the specified simulation period and should not be interpreted as long-run steady-state values.

In the reported simulation run, Strategy 1, which increased the number of loading bays from 12 to 14, produced a mean waiting time of 8.36 minutes. This result was unexpected because increasing the number of otherwise identical servers would generally be expected to improve queuing performance. Since vehicle-merging friction and lane-weaving interactions were not explicitly represented in the model, the observed increase cannot be attributed to these factors. Model verification and multiple simulation replications are required before conclusions can be drawn regarding the effectiveness of this strategy.

The process-acceleration scenarios produced progressively lower waiting times and queue lengths as the assumed service duration was reduced. Strategy 2C, which used an average service duration of 2.5 minutes, recorded a mean waiting time of 0.62 minutes, a maximum waiting time of 2.75 minutes and a maximum queue length of 15 passenger groups. Compared with the baseline run, the reported mean waiting time decreased by approximately 91.65%.

Nevertheless, a 2.5-minute service duration corresponds to a service rate of 24 passenger groups per hour per bay and a traffic intensity of:

$$\rho = \frac{300}{12(24)} = 1.0417$$

Therefore, Strategy 2C cannot be classified as steady-state stable. Its results suggest improved queue performance during the specified 120-minute period, but the queue may continue to increase if the arrival rate of 300 passenger groups per hour is sustained.

Overall, the preliminary findings suggest that reducing service duration may improve finite-period queue performance. However, the current results do not establish that process acceleration is conclusively superior to physical capacity expansion. Further model verification, repeated simulation experiments and empirical validation are required before a definitive operational recommendation can be made.

Operational Recommendations

1. Conduct a Pilot Study on Process Improvement

KLIA Terminal 1 management may consider conducting a pilot study to examine whether passenger-processing and taxi-loading times can be reduced. Potential interventions may include digital ticketing, improved passenger-to-taxi assignment and better coordination between dispatch personnel and available taxis.

The average service duration should be measured before and after the pilot implementation. This would establish whether the proposed interventions can realistically produce the service-rate improvement assumed in the simulation.

For an arrival rate of 300 passenger groups per hour with 12 loading bays, each bay must serve more than 25 passenger groups per hour to satisfy $\rho < 1$. Therefore, the average service duration must be less than 2.4 minutes. For example, a service rate of 28 passenger groups per hour per bay, equivalent to approximately 2.14 minutes per group, would produce:

$$\rho = \frac{300}{12(28)} = 0.8929.$$

This additional configuration should be evaluated through simulation and verified using observed operational data.

2. Re-evaluate Physical Capacity Expansion

The effect of increasing the number of loading bays should be examined further. At a service rate of 20 passenger groups per hour per bay, 14 bays provide a total service capacity of:

$$14(20) = 280$$

passenger groups per hour. This remains below the assumed arrival rate of 300 passenger groups per hour and produces:

$$\rho = \frac{300}{14(20)} = 1.0714$$

Therefore, 14 bays would not theoretically provide a stable system under the assumed peak demand. At least 16 bays would be required to reduce the traffic intensity below unity:

$$\rho = \frac{300}{16(20)} = 0.9375$$

Before physical expansion is considered, a feasibility study should examine space availability, construction requirements, operating costs, staffing needs and vehicle circulation. Lane-weaving and merging effects should not be included as conclusions unless they are explicitly represented in the simulation or supported by field observations.

3. Consider Demand-Responsive Staff Deployment

Airport management may consider aligning staff deployment with expected flight-arrival waves. Additional personnel could be assigned during identified peak periods to assist with passenger guidance, dispatch coordination and luggage handling.

However, the present simulation did not explicitly model staffing levels. Therefore, the effect of additional staff on service duration and queue performance should be evaluated through field measurements or an extended simulation model before implementation.

4. Use DES as a Decision-Support Method

Discrete-event simulation may be used as a decision-support method for evaluating alternative operating configurations, temporary demand surges and loading-bay closures without disrupting live terminal operations. However, the use of open-source software should not be described as entirely “zero-cost”, as model development, data collection, validation, staff training and maintenance may still require resources.

Before the model is used for operational decision-making, it should be validated against observed passenger waiting times, queue lengths, service durations and throughput at KLIA Terminal 1. A practical next step is to measure passenger-group arrivals during flight-arrival waves, loading times and the available waiting space at the taxi platform. These observations could then be used to test whether changes to dispatch procedures reduce service times and whether additional loading bays are feasible. Future simulations should incorporate time-varying arrivals and explicit physical queue-capacity constraints.

Recommendations for Future Research

Future studies should conduct multiple independent simulation replications, report 95% confidence intervals, validate the model using observed operational data from KLIA Terminal 1, and incorporate time-varying passenger arrivals, taxi availability, and vehicle-movement interactions. A cost-benefit analysis should also be conducted to compare physical capacity expansion with process-improvement strategies.

This study provides a preliminary quantitative framework for examining landside taxi-queue congestion at KLIA Terminal 1. Subject to further model verification and empirical validation, the combined analytical and simulation approach may support the evaluation of strategies for improving passenger flow and taxi-dispatch efficiency.

Acknowledgement

The authors would like to acknowledge Universiti Teknologi MARA (UiTM) and the Faculty of Computer and Mathematical Sciences (FSKM) for the academic support provided through the FSKM Writing and Publication Workshop 2026. This paper was developed based on a final-year student project conducted under the supervision of the corresponding author. The authors also acknowledge Agensi Pengangkutan Awam Darat (APAD) for making the vehicle-registration data used in this study publicly available.

References

- Agensi Pengangkutan Awam Darat. (2020). *Teksi lapangan terbang di Semenanjung Malaysia* [Data set]. Retrieved from <https://archive.data.gov.my/data/msMY/dataset/teksi-lapangan-terbang-di-semenanjung-malaysia>
- Balakrishnan, H., & Simaiakis, I. (2014). *A queuing model of the airport departure process* (Technical Report No. LIDS-P-3021). Cambridge, MA: Massachusetts Institute of Technology.
- Banks, J., Carson, J. S., II, Nelson, B. L., & Nicol, D. M. (2010). *Discrete-event system simulation* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Dorton, S. L. (2011). *Analysis of airport security screening checkpoints using queuing networks and discrete event simulation* (Unpublished master's thesis). Embry-Riddle Aeronautical University, Daytona Beach, FL.
- Dorton, S. L., & Liu, D. (2016). Effects of baggage volume and alarm rate on airport security screening checkpoint efficiency using queuing networks and discrete event simulation. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 26(1), 95–109. doi:10.1002/hfm.20616
- Heizer, J., Render, B., & Munson, C. (2017). *Operations management: Sustainability and supply chain management* (12th ed.). Boston, MA: Pearson.
- Inturri, G., Pluchino, A., & Di Grande, C. M. (2026). Capacity assessment of small airport passenger terminals using agent-based simulation: A case study of Comiso Airport. *Case Studies on Transport Policy*, 25, 101896. <https://doi.org/10.1016/j.cstp.2026.101896>
- Jenčová, E., Gajdušková, T., Jezný, M., & Hudák, P. (2025). Identification of bottlenecks in passenger handling processes using data-driven tools. *Applied Sciences*, 15(15), 8760. <https://doi.org/10.3390/app15158760>
- Jia, W., Huang, Y.-L., Zhao, Q., & Qi, Y. (2022). Modeling taxi drivers' decisions at airport based on queueing theory. *Research in Transportation Economics*, 92, 101093. doi:10.1016/j.retrec.2021.101093
- Kierzkowski, A., & Kisiel, T. (2017). Simulation model of security control system functioning: A case study of the Wrocław Airport terminal. *Journal of Air Transport Management*, 64, 173–185.
- Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). New York, NY: McGraw-Hill Education.
- Lee, Y., Choi, S. C., Lee, K., & Cho, S. W. (2025). Dynamic rescheduling strategy for passenger congestion balancing in airport passenger terminals. *Mathematics*, 13(13), 2208. <https://doi.org/10.3390/math13132208>
- Manataki, I. E., & Zografos, K. G. (2010). Assessing airport terminal performance using a system dynamics model. *Journal of Air Transport Management*, 16(2), 86–93. doi:10.1016/j.jairtraman.2009.10.007
- Mujica Mota, M. (2015). Check-in allocation improvements through the use of a simulation–optimization approach. *Transportation Research Part A: Policy and Practice*, 77, 320–335. doi:10.1016/j.tra.2015.04.016
- Oprea, C., Rosca, M., Rosca, E., Costea, I., Ilie, A., Dinu, O., & Ruscă, A. (2024). Analyzing passenger flows in an airport terminal: A discrete simulation model. *Computation*, 12(11), 223. doi:10.3390/computation12110223
- Pivac, J., & Bartulović, D. (2026). Airport terminal facilities software for low-cost carriers: Development and evaluation at a case-study airport. *Applied Sciences*, 16(2), 852. <https://doi.org/10.3390/app16020852>
- Qie, N., Zhang, Y., & Li, H. (2021). Analysis of airport taxi problem based on M/Ek/1 queuing model. In *Proceedings of the 18th International Conference on Informatics in Control*,

Automation and Robotics (pp. 408–415). Setúbal, Portugal: SCITEPRESS.
doi:10.5220/0010623304080415

Shortle, J. F., Thompson, J. M., Gross, D., & Harris, C. M. (2018). *Fundamentals of queueing theory* (5th ed.). Hoboken, NJ: John Wiley & Sons. doi:10.1002/9781119453761

Wang, Y., Zhang, L., & Chen, J. (2022). The optimization of airport management based on queueing theory. *Journal of Advanced Transportation*, 2022, 1–15.
doi:10.1155/2022/9482310