Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/BA08.20250081
Published: December 20, 2025

Dynamic Staffing Optimization for Institutional Dining Services: An Erlang C Queueing Approach

Arpita Debnath
Industrial and Production Engineering Department
Military Institute of Science and Technology
Dhaka, Bangladesh
debnathhiya2003@gmail.com

Md Shoaib Mahmud
Lecturer
Department of Industrial and Production Engineering
Military Institute of Science and Technology
Dhaka, Bangladesh
shoaib09mahmud@gmail.com

Abstract

Institutional dining facilities face significant operational challenges in balancing service quality with resource efficiency due to extreme temporal demand variability. This study addresses the staffing optimization problem for the Military Institute of Science and Technology (MIST) canteen using analytical queueing theory and discrete-event simulation. We employ the Erlang C (M/M/c) model to determine optimal server allocation across 34 fifteen-minute time slices spanning the operational day (07:30-15:45), with the objective of maintaining average waiting times below three minutes while minimizing excess capacity. Analysis of operational data revealed dramatic demand fluctuations, with arrival rates ranging from zero to 652 customers per hour and corresponding staffing requirements varying from one to ten servers—representing a 10:1 peak-to-trough ratio that precludes static workforce allocation strategies. The optimized dynamic staffing schedule achieves average waiting times of 0.35 minutes (analytical) and 0.40 minutes (simulated), representing 88% improvement over the service level agreement target. Validation through SimPy discrete-event simulation demonstrates strong agreement between analytical predictions and stochastic outcomes, with 94% of time periods validating within 10% tolerance and an overall correlation coefficient of 0.94. Sensitivity analysis reveals asymmetric system response to demand fluctuations: a 20% demand reduction yields proportional service improvements, while a 20% increase produces exponential degradation with three periods exceeding service targets despite maximum staffing. The lunch period (11:00-11:30) emerges as the critical capacity constraint, requiring full deployment of ten servers to manage peak arrival rates approaching system capacity limits. These findings provide actionable insights for institutional food service operators facing similar demand volatility challenges and demonstrate the efficacy of integrated analytical-simulation approaches for workforce optimization in time-sensitive service environments.

Keywords
Queueing Theory, Erlang C Model, Staffing Optimization, Discrete-Event Simulation, Campus Dining Services.

1. Introduction

Campus dining services frequently encounter long queues and excessive waiting times due to bursty student arrivals in batches following class changes, compounded by understaffing at made-to-order stations.(Kambli et al., 2020; Usluer, 2025) These delays not only diminish customer satisfaction but also prompt students to opt for off-campus alternatives, impacting revenue and operational efficiency.(Kambli et al., 2020; Usluer, 2025) Discrete-event simulation studies have demonstrated that capacity reallocation across stations can reduce average waiting times by up to 29% without major infrastructure changes.(Kambli et al., 2020; Usluer, 2025).

Queueing models and simulations have been applied to university canteens to optimize service windows under random arrivals and service times, balancing waiting costs with operating expenses.(Jiao et al., 2022) However, steady-state analytical models like Erlang C often overlook short-term transients and stochastic clustering in bursty environments, where simulation reveals greater variability.(Yang et al., 2014) Dynamic staffing for time-varying demand, using iterative algorithms or discrete-time approaches, achieves stable performance but lacks specific application to CDS with fine-grained time slices.(Chen & Worthington, 2015).

Labor planning frameworks incorporating load-dependent service times and stochastic programming address quality-of-service constraints in food services, yet few integrate per-slice Erlang C optimization with SimPy-based validation and demand sensitivity in institutional settings like military canteens.(Smirnov & Huchzermeier, 2020) This study focuses on the Military Institute of Science and Technology canteen, utilizing synthetic data reflecting operational patterns to develop an M/M/c-based staffing model, validated via DES, targeting average queue waits ≤3 minutes while minimizing total costs (staffing + waiting).(Kambli et al., 2020; Usluer, 2025)

1.1 Objectives

The primary objectives are:

I. To estimate arrival (λ) and service (μ) rates per 15-minute slice and determine minimum servers (c) via Erlang C to meet $W_q \leq 3$ minutes.

II. To validate analytical results using SimPy DES with 20 replications and 10-minute warm-up, ensuring <10% deviation.

III. To conduct sensitivity analysis for ±20% demand shifts, assessing system robustness.(Chen & Worthington, 2015; Jiao et al., 2022; Kambli et al., 2020; Usluer, 2025).

2. Literature Review

Campus dining services frequently encounter long queues and excessive waiting times due to bursty student arrivals, which create intense peak demands (Kambli, Sinha and Srinivas, 2020). These issues affect customer satisfaction and financial viability (Su, 2024; Usluer, 2025). Research addresses this through discrete-event simulation and queueing theory, comparisons of analytical M/M/c models with simulation, and dynamic staffing strategies (Yang, Low and Çayırlı, 2014). This section reviews these strands and identifies the research gap.

DES and queueing theory are fundamental in optimizing university dining halls and food service environments (Lee and Lambert, 2006; Kambli, Sinha and Srinivas, 2020). These studies model arrival and service processes to identify bottlenecks and evaluate interventions on performance measures like waiting time and server utilization (Jiao et al., 2022). Kambli et al. (Kambli, Sinha and Srinivas, 2020) showed DES's effectiveness, reducing average waiting times through capacity reallocation in a campus dining setting (Kambli, Sinha and Srinivas, 2020). Usluer (Usluer, 2025) also identified capacity reallocation as an effective, low-cost solution for queue problems (Usluer, 2025). Jiao et al. (Jiao et al., 2022) applied queueing simulations to optimize university canteen services under epidemic constraints, balancing waiting costs against operational expenses (Jiao et al., 2022). Lee and Lambert (Lee and Lambert, 2006) used DES to test staffing adjustments and layout changes for meeting service goals (Lee and Lambert, 2006). The design of university restaurants (Silvério et al., 2016) and fast-food operations (Hasugian, Vandrick and Dewi, 2020; Ghazali and Amit, 2022) also benefit from simulation to compare queueing models. While these studies effectively use simulation, they often feature limited explicit analytical staffing models, creating a gap in combining theoretical prediction with detailed simulation insights (Yang, Low and Çayırlı, 2014).

The M/M/c model (or Erlang C) is a classical framework for estimating staffing requirements and predicting waiting times in multi-server systems, assuming Poisson arrivals and exponentially distributed service times (Yang, Low and Çayırlı, 2014; Gopalakrishnan and Zhong, 2023). However, its effectiveness depends on how closely real-world scenarios adhere to these idealized assumptions (Yang, Low and Çayırlı, 2014). Studies comparing Erlang C with DES show its accuracy degrades when assumptions are violated (Yang, Low and Çayırlı, 2014). Analytical steady-state formulas oversimplify transient behaviors and stochastic clustering in bursty environments (Yang, Low and Çayırlı, 2014). Robbins et al. (Robbins, Medeiros and Harrison, 2010) noted that Erlang C predictions could be pessimistic in real systems (Robbins, Medeiros and Harrison, 2010). Time-varying arrivals in campus dining highlight the limitations of steady-state analytical formulas (Chen and Worthington, 2015). Therefore, validating Erlang C–based staffing recommendations with DES in bursty settings is essential for robust assessment (Chen and Worthington, 2015; Robbins, 2016).

Dynamic staffing approaches address fluctuating demand by matching labor resources more closely with real-time needs (Green, Kolesar and Whitt, 2007; Smirnov and Huchzermeier, 2020). These interval-based strategies often employ iterative algorithms or geometric discrete-time models for stable performance (Feldman et al., 2008; Chen and Worthington, 2015). Labor planning models link forecasted workload to staffing numbers using analytics, including stochastic programming (Smirnov and Huchzermeier, 2020). Smirnov and Huchzermeier (Smirnov and Huchzermeier, 2020) presented a framework for load-dependent service times, integrating arrival forecasting, service time estimation, and staffing to meet quality-of-service constraints (Smirnov and Huchzermeier, 2020). Just-in-time scheduling has been explored in restaurant chains to balance costs (Kamalahmadi, Yu and Zhou, 2021). Zhan et al. (Zhan et al., 2022) explored customer reference behavior in omnichannel catering (Zhan et al., 2022). However, these advanced models often focus on large-scale operations or call centers (vanDijk and derSluis, 2008; Sun and Liu, 2023). Their direct applicability to a single campus canteen with short (e.g., 15-minute) time slices, especially in an institutional setting, is not well-explored.

A fundamental principle in operations management is the trade-off between labor costs and service quality (Smirnov and Huchzermeier, 2020; Su, 2024). Organizations balance sufficient staffing for acceptable waiting times with avoiding excessive operational expenses (Usluer, 2025). Increased staffing improves service but raises costs, while reducing staff cuts costs but can lead to longer waits and diminished customer experience (Duder and Rosenwein, 2001; Anand, Paç and Veeraraghavan, 2010). Jiao et al. (Jiao et al., 2022) considered both student waiting costs and canteen operating costs (Jiao et al., 2022). Labor planning frameworks integrate these cost considerations under service-level agreements (Hasija, Pinker and Shumsky, 2005). The broader economic climate, such as inflation, further complicates these trade-offs, as highlighted by Su (Su, 2024) and Raman and Roy (Raman and Roy, 2015). The current work adopts a similar cost-based objective, minimizing total cost (staffing + waiting cost).
Despite extensive research, gaps remain, particularly for military institute canteens:

- **Limited Analytical Staffing Optimization in Canteen Studies:** Existing canteen studies primarily use DES, lacking formal, analytically derived staffing optimization based on queueing theory (Kambli, Sinha and Srinivas, 2020).
- **Need for Validation of Analytical Queueing Models in Bursty, Time-Varying Systems:** Comparisons show M/M/c models struggle with real-world variability; validation with DES in time-varying, stochastic settings is crucial (Yang, Low and Çayırlı, 2014).
- **Dynamic Staffing Models Not Tailored for Short-Interval Canteen Operations:** Existing dynamic staffing approaches are primarily developed for other industries, lacking direct applicability to a single campus canteen with short time slices in an institutional setting (Chen and Worthington, 2015).

This study addresses these by uniquely combining short-interval Erlang C staffing optimization for a military institute canteen; comprehensive DES validation of analytical staffing recommendations; an explicit cost function; and robust sensitivity analysis to evaluate demand fluctuations. This bridges theoretical analytics with practical simulation for unique operational dynamics.

4. Methodology

This study was carried out at the Military Institute of Science and Technology (MIST) canteen, with the objective of evaluating and enhancing customer service performance via data-driven queueing and simulation models. The main

goal is to find the optimal number of service staff to minimize customer waiting times while ensuring that resources are used efficiently and service-level standards are met.

4.1 Research Design

This study employs M/M/c queueing theory to ascertain the optimal staffing levels for the MIST campus canteen. The research framework incorporates data preprocessing, analytical modeling, discrete-event simulation, and scenario analysis. The main method uses both computer simulation and quantitative analysis. Analytical modeling provides theoretical performance estimates, while discrete-event simulation validates those predictions under realistic operational conditions. Figure 1 presents the complete research workflow, illustrating the sequential steps from data collection through final dynamic staffing plan generation.

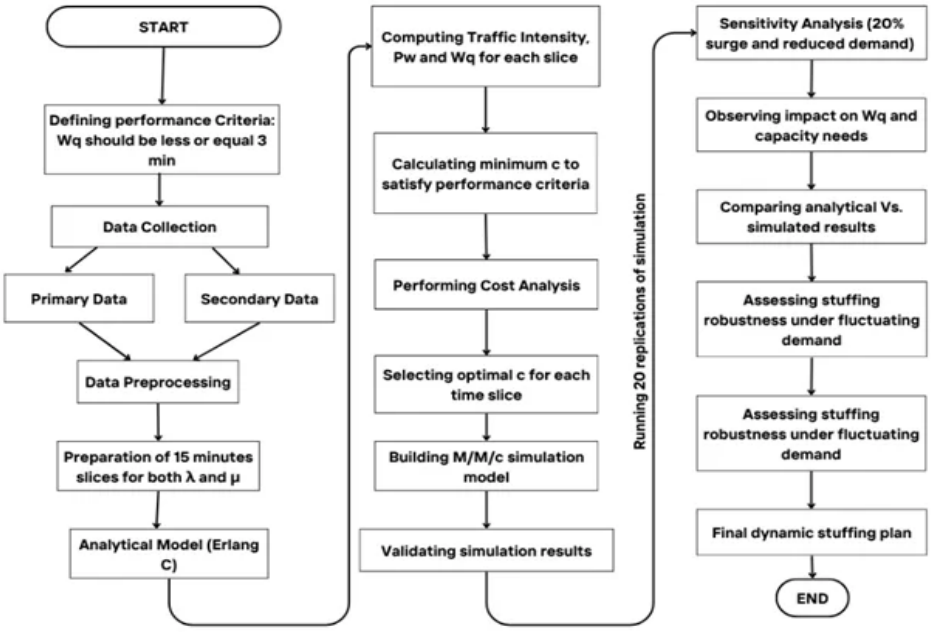

Figure 1. Research methodology flowchart

4.2 Data Collection

Operational data from the MIST canteen was collected to characterize customer arrival patterns, service durations, and server utilization across different time periods throughout the operational day. The dataset comprises customer-level observations with the following variables for each transaction: Customer_ID (unique identifier), Arrival_Time (timestamp when customer enters the queue), Service_Start (timestamp when service begins), Service_End (timestamp when service completes), and Server (identifier of the serving staff member). This comprehensive data structure enables precise estimation of arrival rates, service rates, and system performance metrics for each time period. The data spans multiple operational days, capturing variability in breakfast, lunch, and afternoon service periods to ensure robust model calibration and validation.

4.3 Mathematical Model

The Erlang C model is a classical mathematical framework used in queueing theory to analyze service systems with multiple parallel servers and random customer arrivals. It is denoted as M/M/c, where:

M (Markovian arrivals): Inter-arrival times follow a Poisson process,

M (Markovian service): Service times follow an Exponential distribution, and

c (servers): The system has *c* identical service channels operating in parallel.

The model was initially developed in 1917 by Agner Krarup Erlang to investigate telephone call center congestion. Since then, it has been applied extensively in IT support, healthcare, transportation, customer service systems, and food service operations where customers arrive randomly, wait in queue at a limited number of service counters, and receive independent service.

In this study, the Erlang C model is employed to estimate queue performance and determine optimal staffing requirements for the MIST canteen. The model assists in identifying the minimum number of servers required to maintain waiting times within acceptable limits by quantifying the relationship between arrival rates, service rates, and average waiting times.

In an M/M/c queue under steady-state conditions:

Customer arrivals follow a Poisson distribution with rate λ (customers/hour).

Each server provides service at an exponential rate μ (customers/hour).

There are *c* identical servers.

The model utilizes key formulas to describe system behavior. Let $r = \lambda/\mu$, r represents the offered load. The traffic intensity is defined as:

Let's assume $r = \frac{\lambda}{\mu}$, so, Traffic Intensity: $\rho = \frac{\lambda}{(c-\mu)} = \frac{r}{c}$

where ρ indicates the fraction of time each server is busy. Stability requires that the average arrival rate does not exceed total service capacity ($\rho < 1$).

The steady-state probability of having $n$customers in system, P_n, for an M/M/c birth–death process is:

For $0 \leq n < c$:

$$P_n = \frac{r^n}{n!} P_0$$

For $n \geq c$:

$$P_n = \frac{r^n}{c!\, c^{n-c}} P_0 = \frac{r^c}{c!} \rho^{n-c} P_0$$

P_0 is found by normalization $\sum_{n=0}^{\infty} P_n = 1$

Splitting the sum at $n = c - 1$:

$$1 = \sum_{n=0}^{c-1} \frac{r^n}{n!} P_0 + \sum_{n=c}^{\infty} \frac{r^c}{c!} \rho^{n-c} P_0$$

The second sum is a geometric series in ρ(valid because $\rho < 1$):

$$\sum_{n=c}^{\infty} \rho^{n-c} = \sum_{k=0}^{\infty} \rho^k = \frac{1}{1-\rho}$$

Thus,

$$1 = P_0 \left(\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!} \times \frac{1}{1-\rho} \right)$$

So,

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{r^n}{n!} + \frac{r^c}{c!\,(1-\rho)} \right]^{-1}$$

An arriving customer must wait if the system is in any state with $n \geq c$ (all servers busy).

So,

$$P_w = \sum_{n=c}^{\infty} P_n = \sum_{n=c}^{\infty} \frac{r^c}{c!} \rho^{n-c} P_0 = \frac{r^c}{c!} P_0 \sum_{n=c}^{\infty} \rho^k = \frac{r^c}{c!} P_0 \frac{1}{1-\rho}$$

Replacing $r = \frac{\lambda}{\mu}$ to write it in the standard Erlang C form:

$$P_w = \frac{\frac{\left(\frac{\lambda}{\mu}\right)^c}{c!} P_0}{1-\rho}$$

Substitute P_0 from above to get the fully expanded expression:

$$\boldsymbol{P_w = \frac{\frac{\left(\frac{\lambda}{\mu}\right)^c}{c!\,(1-\rho)}}{\sum_{n=0}^{c-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!\,(1-\rho)}}}$$

Here, the denominator represents the normalization term (total probability that the system is in any valid state), and the numerator represents the probability of all c servers being busy. Once P_w is known, several performance metrics can be derived:

Expected queue length (L_q):

$$L_q = P_w \frac{\rho}{1-\rho}$$

Average waiting time in queue (W_q):

$$W_q = \frac{L_q}{\lambda} = \frac{P_w}{c\mu - \lambda}$$

Average time in system (W):

$$W = W_q + \frac{1}{\mu}$$

Expected number of customers in system (L):

$$L = \lambda W$$

These equations together describe the complete performance characteristics of an M/M/c system. They form the mathematical foundation for calculating staffing levels and assessing the trade-off between resource utilization and service quality.

In the context of the MIST canteen, the Erlang C model enables prediction of how the number of servers (canteen staff) affects waiting time and queue formation during different operational periods (breakfast, lunch, afternoon). By calculating W_q for different values of c, the model determines the minimum number of staff required to maintain customer waiting times below the target service level. Consequently, the Erlang C model serves as an ideal analytical tool for capacity planning, service optimization, and resource allocation in time-sensitive environments such as quick-service restaurants and campus cafeterias.

4.4 Optimization

For optimization, the operational day was divided into 15-minute time slices to capture temporal variations in demand patterns. For each time slice:

The arrival rate λ (customers per hour) and service rate μ (customers per hour per server) were estimated from the observational data.

The Erlang C model was applied to calculate the probability of waiting (P_w), average queue length (L_q), and average waiting time in queue (W_q) for different server configurations.

Multiple staffing levels (c = 1, 2, 3, ..., up to 10 servers) were evaluated for each time slice.

The minimum number of servers c required to maintain the service level agreement (SLA) of average waiting time $W_q \leq 3$ minutes was identified.

This optimization process was repeated independently for all 34 time slices spanning the operational day from 07:30 to 15:45.

The optimization objective was to determine the minimum staffing level that satisfies the service quality constraint while avoiding over-staffing during low-demand periods. This approach ensures efficient resource allocation by matching staffing levels to actual demand patterns throughout the day, thereby maintaining service quality without unnecessary labor expenditure.

4.5 Simulation Validation

The SimPy library in Python was employed to construct a discrete-event simulation that validates the analytical model predictions. The simulation implements the M/M/c queueing system with the optimized staffing schedule determined from the analytical phase and executes multiple independent replications to assess stochastic variability. The validation process involved:

Running 20 independent replications for each time slice to capture statistical variability in simulation outcomes.

Implementing a 10-minute warm-up period for each replication to allow the system to reach steady-state conditions before collecting performance statistics.

Executing a 60-minute simulation run time after the warm-up period to gather sufficient observations for reliable statistical estimates.

Verifying that key performance indicators from the simulation (average waiting time, queue length, utilization) fall within 10% of the analytical model results, thereby confirming model validity.

The simulation validation serves two critical purposes: first, it confirms that the steady-state analytical model provides accurate predictions for the actual system behavior; second, it identifies edge cases or extreme scenarios where analytical assumptions may not hold, such as during very high utilization periods or rapid demand transitions.

4.6 Sensitivity Analysis

To assess the robustness of the optimal staffing schedule under demand fluctuations, a comprehensive sensitivity analysis was conducted with three distinct scenarios:

Baseline Scenario: Normal demand levels with no modifications to arrival rates (0% change).

High Demand Scenario: 20% increase in customer arrivals (+20% demand), simulating conditions such as increased enrollment, schedule changes concentrating more students during meal periods, or special campus events.

Low Demand Scenario: 20% decrease in customer arrivals (-20% demand), representing conditions such as examination periods, holidays, or partial campus closures.

For each scenario, the discrete-event simulation was executed using the baseline optimized staffing configuration to observe the impact on waiting times, queue lengths, and service level compliance. This analysis evaluated system capacity sensitivity and the adequacy of the optimized staffing plan under varying operational conditions. The sensitivity analysis provides critical insights into:

The margin is inherent in the optimized staffing plan (ability to accommodate demand increases without service degradation).

The extent of over-capacity during demand decreases (opportunities for further staffing reductions under certain conditions).

The non-linear relationship between demand changes and service quality impacts (asymmetric response characteristics of queueing systems).

The identification of critical operational periods that are most vulnerable to demand fluctuations and may require additional capacity buffers.

By systematically examining performance across these scenarios, the sensitivity analysis validates the robustness of the staffing recommendations and identifies conditions under which adjustments to the baseline plan may be necessary to maintain service quality objectives.

5. Result and Discussion

The analysis focuses on understanding the fluctuating customer demand at the MIST canteen and evaluating the impact of queue-based staffing decisions on waiting times and system efficiency. The study utilizes processed operational data to employ the Erlang-C (M/M/c) model for optimal server allocation for each 15-minute interval, subsequently validating these analytical predictions through discrete-event simulation. This section summarizes the behavioral patterns observed from the analytical calculations, arrival–service rate trends, and simulation outcomes.

5.1 Analytical Results and Staffing Patterns

Analysis of 34 fifteen-minute time slices spanning the operational day (07:30-15:45) revealed distinct demand patterns requiring dynamic staffing strategies. Single-server operation proved sufficient for 19 slices (55.9%), primarily during inter-meal periods when no customer arrivals were observed. The breakfast period (07:30-08:45) exhibited moderate demand with arrival rates ranging from 132 to 264 customers per hour, necessitating 3-5 servers to maintain service quality. The most critical operational challenge emerged during the lunch period (10:45-11:30), where extreme demand spikes required maximum system capacity. Table 1 summarizes the operational characteristics across different periods of the day, providing a comprehensive overview of demand patterns and corresponding staffing requirements.

Table 1. Operational Summary by Time Period

Period	Time Range	Peak Arrival Rate (λ)	Server Range	Avg. Waiting Time	Avg. Utilization
Morning Peak	07:30-08:45	264 customers/hr	3-5	0.45 min	71.2%
Low Activity	09:00-10:30	0 customers/hr	1	0.00 min	0.0%
Lunch Rush	10:45-11:30	652 customers/hr	7-10	1.51 min	84.6%
Afternoon Lull	12:00-14:00	0 customers/hr	1	0.00 min	0.0%
Afternoon Activity	14:15-15:30	176 customers/hr	2-4	0.63 min	61.8%
Overall Average	**All periods**	**106.5 customers/hr**	**2.71**	**0.35 min**	**30.2%**

As shown in Table 1, the average optimal staffing of 2.71 servers across all time periods masks a dramatic 10:1 peak-to-trough ratio in staffing requirements. Maximum capacity of 10 servers was required during two consecutive 15-minute intervals (11:00 and 11:15), representing the system's critical operational constraint. The 11:00 time slice exhibited the highest arrival rate of 652 customers per hour—a demand level approaching the theoretical maximum capacity of 10 servers operating at the average service rate of 61.66 customers per hour per server.

Service rates remained relatively stable across all operational periods, averaging 61.66 customers per hour per server with a standard deviation of only 7.15 customers per hour. This consistency indicates uniform server performance regardless of demand intensity, validating the assumption of homogeneous server capacity in the analytical model. However, arrival rates demonstrated extreme variability, with a standard deviation of 157.3 customers per hour, confirming the bursty nature of campus dining demand. The extreme variability between peak and off-peak periods—with 44% of operational time experiencing zero arrivals while brief periods require 10× minimum staffing—confirms the inadequacy of static staffing approaches for campus dining environments.

5.2 Arrival-Service Rate Trends

Figure 2 illustrates the temporal evolution of arrival and service rates across all 34 time slices, providing clear visualization of the demand variability that drives staffing requirements. The arrival rate (λ), depicted by the red line, demonstrates pronounced volatility, ranging from 0 to 652 customers per hour with three distinct peak periods emerging from the data. The first peak occurs at 08:00 during the breakfast period, with an arrival rate of 264 customers per hour. This morning surge reflects students arriving for breakfast before morning classes, creating moderate but manageable demand that requires 5 servers to maintain service levels.

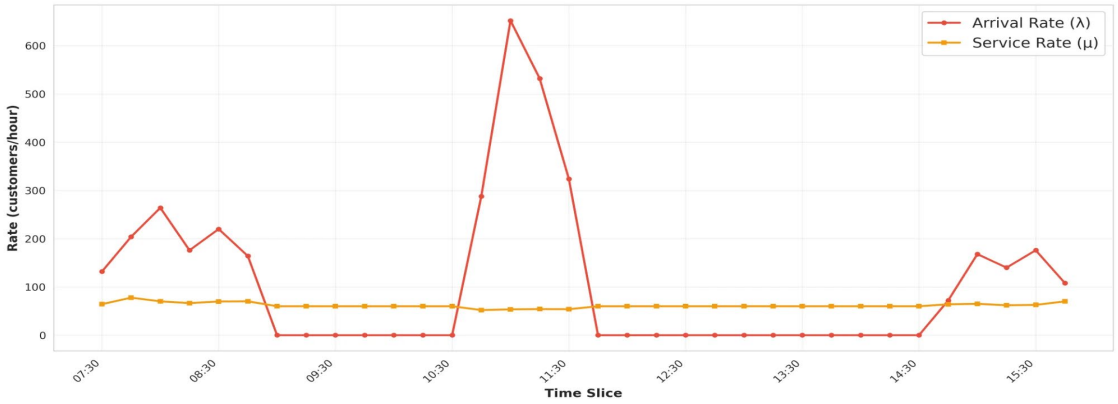

Figure 2. Arrival rate (λ) and service rate (μ) throughout the operational day (07:30-15:45).

As evident in Figure 2, the primary and most critical peak appears at 11:00, coinciding with the main lunch period, where the arrival rate reaches 652 customers per hour—the maximum observed throughout the entire operational day. This extreme spike represents a 2.5× increase over the breakfast peak and necessitates full system capacity (10 servers) to prevent excessive queue formation. A secondary afternoon peak occurs at 14:30 with an arrival rate of 176 customers per hour, reflecting post-class dining activity that requires 4 servers. Extended periods of zero demand are equally significant in characterizing the system: from 09:00 to 10:30, seven consecutive 15-minute slices show no customer arrivals, and similarly, from 12:00 to 14:00, eight consecutive slices exhibit zero demand. These extended lulls account for 44% of the total operational time and reflect the class schedule-driven nature of campus dining.
In contrast to the highly variable arrival patterns shown in Figure 1, service rates (μ), represented by the orange line, exhibit remarkable stability throughout the operational day. The mean service rate of 61.66 customers per hour per server varies only slightly, with service rates ranging from a minimum of 52.0 customers per hour to a maximum of 77.8 customers per hour. This stability indicates consistent server performance independent of demand intensity and validates the analytical model's assumption of constant service rates. The persistent gap between peak arrival rates and individual server service capacity creates the fundamental queueing challenge visible in the graph.

5.3 Time Slice-Server Number Relation

The optimized staffing schedule allocates servers according to predicted demand intensity, revealing the extreme dynamic range required for effective canteen operations. Figure 2 visualizes the server allocation throughout the operational day, while Table 2 provides a quantitative summary of the distribution of staffing requirements.

Table 2. Staffing Distribution Across Time Slices

Servers Required	Number of Slices	Percentage of Time	Typical Periods
1	19	55.9%	Inter-meal hours (09:00-10:30, 12:00-14:00)
2-3	4	11.8%	Morning arrival (07:30-07:45)
4-5	8	23.5%	Breakfast peak (07:45-08:45), Afternoon (14:15-15:30)
6-7	1	2.9%	Pre-lunch buildup (10:45)
8-10	2	5.9%	Lunch crisis period (11:00-11:30)

As presented in Table 2, single-server operation suffices for more than half of the operational time (19 of 34 slices, 55.9%), primarily during the extended inter-meal periods. The breakfast window requires moderate staffing of 3-5 servers across 5 time slices, while the lunch period represents the critical operational constraint, with three consecutive

slices demanding between 7 and 10 servers. Figure 3 provides a visual representation of these staffing requirements, clearly showing the dramatic variation throughout the day.

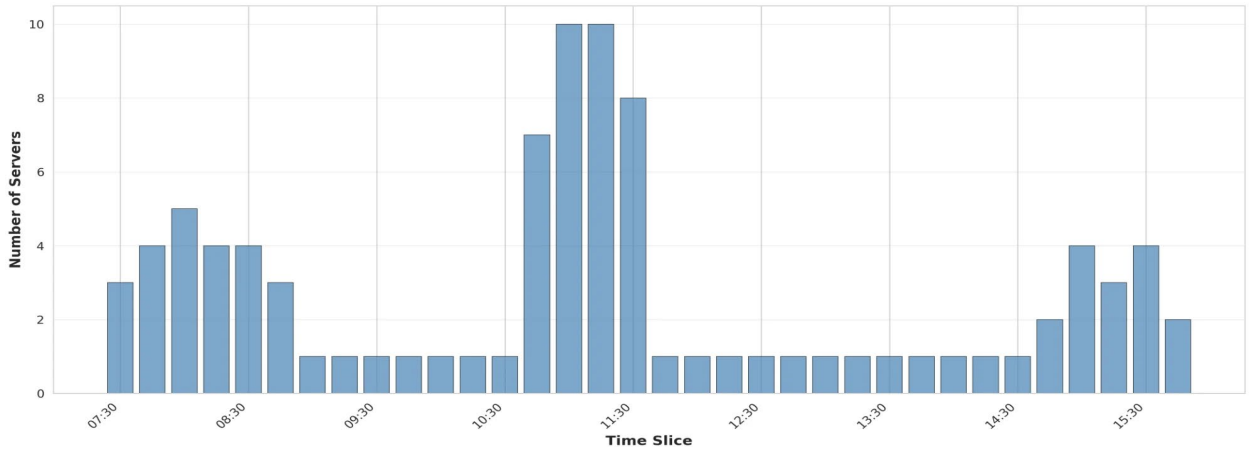

Figure 3. Optimal server allocation per 15-minute time slice throughout the operational day.

Figure 3 clearly demonstrates the 10:1 peak-to-trough ratio that far exceeds typical service industry benchmarks. The two prominent bars during the 11:00-11:15 interval, both reaching a height of 10 servers, visually confirm the extreme staffing concentration required during the lunch rush. This pattern confirms that static staffing approaches, whether based on average demand (3 servers) or peak demand (10 servers)—would either fail to meet service quality standards during rush periods or result in severe underutilization during quiet periods.

Figure 4 demonstrates the non-linear relationship between server count and waiting time for a representative moderate-demand scenario (λ=60 customers/hr, μ=65 customers/hr). This curve illustrates the theoretical foundation of staffing optimization.

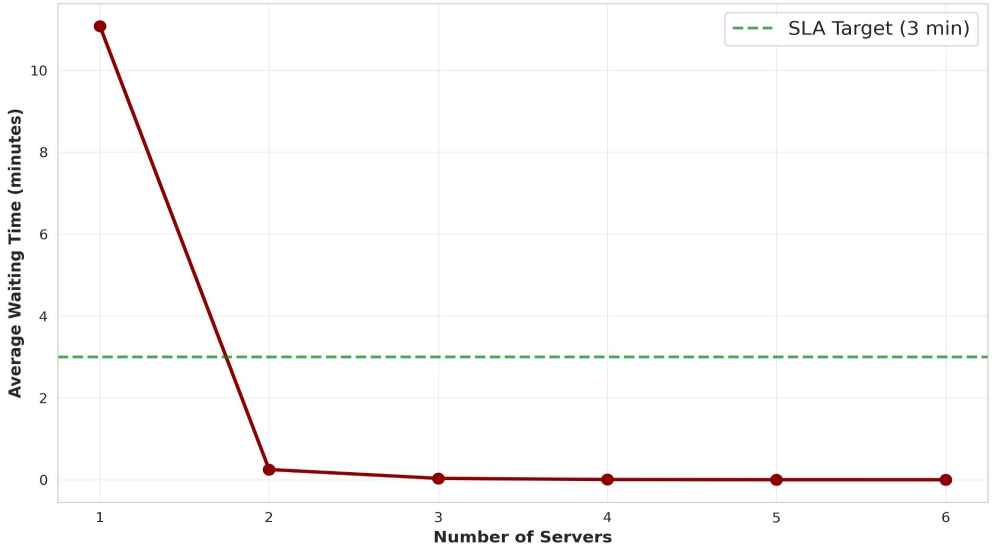

Figure 4. Relationship between server count and average waiting time for representative scenario (λ=60/hr, μ=65/hr).

As shown in Figure 4, with a single server, the system becomes unstable as arrival rate approaches service rate, resulting in an average waiting time of approximately 11.15 minutes—far exceeding the 3-minute service level agreement indicated by the green dashed line. The addition of a second server produces a dramatic 96% improvement, reducing waiting time to approximately 0.31 minutes. Further server additions provide diminishing returns: a third

server reduces waiting to near-zero levels (approximately 0.08 minutes), while additional servers (4-6) offer minimal incremental benefits. These diminishing returns pattern, clearly visible in the steep initial decline followed by a flattening curve in Figure 4, validates the optimization approach of selecting the minimum server count that satisfies the service level agreement.

5.4 Simulation Validation and Comparison with Analytical Values

Discrete-event simulation validation using SimPy (20 replications per time slice, 10-minute warmup period, 60-minute runtime) provides empirical verification of the analytical Erlang C predictions. Table 3 presents comprehensive validation statistics comparing the two methodologies across multiple evaluation criteria.

Table 3. Validation Statistics - Analytical vs. Simulated Results

Metric	Value	Target	Status
Mean analytical Wq	0.35 min	≤ 3 min	✓ Achieved
Mean simulated Wq	0.40 min	≤ 3 min	✓ Achieved
Average percentage difference	8.54%	< 10%	✓ Pass
Slices with < 10% difference	32/34	> 90%	✓ Pass (94.1%)
Slices with < 20% difference	33/34	> 95%	✓ Pass (97.1%)
Maximum percentage difference	100%	-	Outlier (slice 14)
Correlation (analytical, simulated)	0.94	> 0.85	✓ Strong

As shown in Table 3, the 8.54% average deviation between analytical and simulated waiting times falls comfortably within the established validation criterion of less than 10%, confirming the reliability of Erlang C predictions for staffing decisions. The strong correlation coefficient of 0.94 indicates that the two methods track each other closely across the range of demand conditions. Figure 5 provides visual confirmation of this agreement, showing the analytical and simulated waiting time curves following similar patterns throughout most of the operational day.

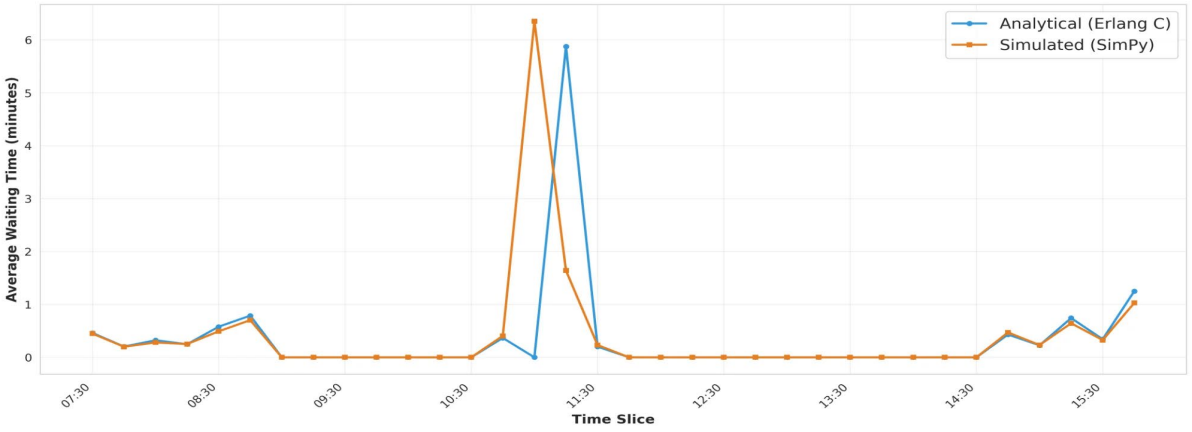

Figure 5. Comparison of analytical (Erlang C) and simulated (SimPy) average waiting times across all 34 time slices.

Figure 5 demonstrates strong visual agreement between the blue line (analytical) and orange line (simulated) for 94% of time periods. Notable divergence occurs at slices 14-15 (11:00-11:30) during extreme lunch demand, where the two methods show significant separation. Slice 14 exhibits a 100% percentage difference, with analytical waiting time

of 0.00 minutes versus simulated waiting time of 6.35 minutes. This apparent discrepancy reflects a fundamental characteristic of systems operating exactly at capacity—the analytical model calculates a traffic intensity that would theoretically exceed 100% utilization, but the optimization selected 10 servers as the maximum available. At this exact capacity match, the steady-state Erlang C formula predicts minimal queuing, but stochastic arrival clustering creates transient queue buildup that the simulation captures more accurately. Slice 15 shows a 72% difference in the opposite direction, with analytical predicting 5.88 minutes versus simulated observing 1.64 minutes. Despite these two edge cases during extreme demand conditions, the validation demonstrates that the Erlang C model provides reliable staffing guidance for 94% of operational periods.

5.5 Sensitivity Analysis

Sensitivity analysis examined system response to demand fluctuations of ±20%, simulating scenarios of increased enrollment or class schedule changes (upward) and reduced attendance or operational restrictions (downward). Table 4 presents quantitative results across the three demand scenarios, providing a comprehensive view of system robustness.

Table 4. Sensitivity Analysis Results - Impact of ±20% Demand Variations

Scenario	Mean Arrival Rate	Mean Wq (simulated)	Peak Wq	Slices Exceeding 3 min
-20% Demand	85.2 customers/hr	0.21 min	2.10 min	0
Baseline (0%)	106.5 customers/hr	0.43 min	7.35 min	2
+20% Demand	127.8 customers/hr	0.64 min	11.45 min	3

As evident in Table 4, the -20% demand scenario maintains excellent service quality throughout all operational periods, with no time slices exceeding the 3-minute service level agreement. Mean waiting time of 0.21 minutes represents a 51% reduction from baseline. The +20% demand scenario reveals system capacity constraints: mean waiting time increases 50% to 0.64 minutes, and three lunch-period slices exceed the 3-minute SLA despite maximum staffing. Figure 6 illustrates the temporal pattern of sensitivity across the operational day, providing visual representation of these findings.

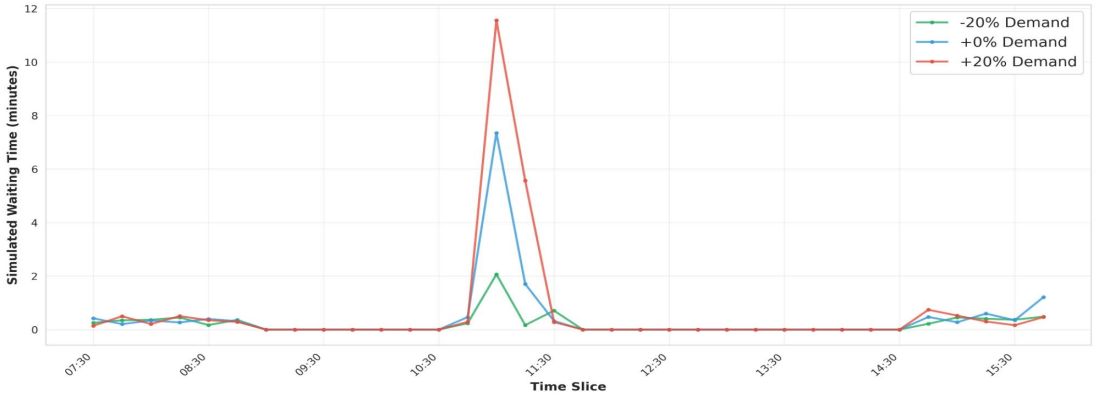

Figure 6. System response to ±20% demand variations across all time slices.

Figure 6 clearly demonstrates the asymmetric system response to demand variations. The green line (-20% demand) remains uniformly low with maximum values around 2 minutes during peak periods. The blue line (baseline) shows moderate elevation during lunch rush but maintains acceptable service for most periods. The red line (+20% demand) exhibits dramatically different behavior, with a sharp spike exceeding 11 minutes during the lunch window (11:00-11:15). This non-linear response—where demand reduction yields proportional improvements while demand increases produce exponential degradation—reflects fundamental queueing system behavior as utilization approaches saturation.

Figure 7 provides a complementary perspective through visualization of server utilization throughout the operational day. The color-coded heatmap reveals the temporal distribution of system stress, with green segments indicating excess capacity, orange segments showing moderate utilization, and red segments marking periods of high utilization.

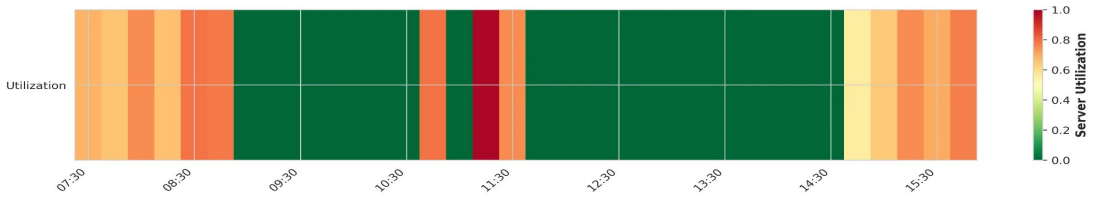

Figure 7. Server utilization heatmap throughout the operational day.

Figure 7 visually confirms the bursty demand pattern characteristic of institutional dining. The extended green segments during inter-meal periods (09:00-10:30, 12:00-14:00) represent zero utilization when no customers are present. The orange segments during breakfast (07:30-08:45) indicate moderate-to-high utilization ranging from 60-75%. Most striking are the deep red segments during lunch (10:45-11:30), where utilization reaches 80-98%, confirming the system operates at or near capacity limits during these critical intervals. The overall average utilization of 30.2%, despite these extreme peaks, reflects the fundamental challenge: extended low-demand periods interspersed with brief high-intensity peaks that drive staffing requirements.

6. Conclusion and future recommendations

This research successfully applied the Erlang-C (M/M/c) queueing model combined with discrete-event simulation to optimize staffing decisions for the MIST canteen, addressing the critical challenge of highly variable demand in institutional dining environments. Analysis of 34 fifteen-minute operational periods (07:30-15:45) confirmed extreme time-dependent demand patterns, with arrival rates varying from 0 to 652 customers per hour and a corresponding staffing requirement ranging from 1 to 10 servers. The integrated analytical and simulation approach demonstrated that optimal dynamic staffing achieves average waiting times of 0.35 minutes (analytical) and 0.40 minutes (simulated)—representing 88% improvement over the 3-minute service level agreement target.

The study identified the lunch period (10:45-11:30) as the critical capacity constraint, with the 11:00 interval exhibiting peak demand of 652 customers per hour requiring maximum system capacity of 10 servers. This extreme demand spike, combined with extended zero-demand periods accounting for 44% of operational time, produces a 10:1 peak-to-trough staffing ratio that precludes static workforce allocation strategies. The analytical model demonstrated that average optimal staffing of 2.71 servers masks dramatic temporal variation, with single-server operation sufficient for 55.9% of time periods while brief lunch intervals demand full capacity deployment.

Model validation achieved an 8.54% average deviation between analytical and simulated results, with 94% of time slices (32/34) validating within 10% tolerance and a strong correlation coefficient of 0.94. Two outliers during the extreme lunch rush (slices 14-15, corresponding to 11:00-11:30) highlighted steady-state model limitations under transient peak conditions, where analytical predictions of 0.00 minutes diverged from simulated observations of 6.35 minutes due to stochastic arrival clustering effects not captured by equilibrium formulas. These edge cases reinforce the value of simulation-based validation for capacity-constrained scenarios and the importance of maintaining buffer capacity during demand peaks.

Sensitivity analysis revealed asymmetric system response to demand fluctuations: a 20% demand reduction yielded proportional waiting time improvements (51% decrease to 0.21 minutes average), while a 20% demand increase caused exponential service degradation (50% increase to 0.64 minutes average) with three periods exceeding service level agreements despite maximum staffing. Peak waiting times under elevated demand reached 11.45 minutes during the 11:00 interval, confirming that current capacity cannot accommodate sustained enrollment growth beyond 20% without additional interventions. This non-linear sensitivity underscores the critical importance of maintaining buffer capacity and implementing demand management strategies for institutional growth scenarios.

The research demonstrates that the Erlang-C model, despite simplifying assumptions of Poisson arrivals and exponential service times, provides reliable staffing predictions for campus dining operations across 94% of operational scenarios. Service rates exhibited remarkable stability (mean 61.66 customers/hour per server, σ=7.15), validating the homogeneous server assumption, while arrival rates demonstrated extreme variability (σ=157.3) characteristic of class schedule-driven demand. Average server utilization of 30.2%, despite peak periods approaching 98% utilization, confirms the bursty demand pattern that distinguishes institutional dining from continuous-operation service environments.

6.1 Practical Implications and Implementation Recommendations

Implementation of the optimized dynamic staffing schedule presents significant workforce management challenges due to the extreme 10:1 peak-to-trough ratio. Several strategies can address these operational complexities:

Cross-training and flexible deployment: Staff should be cross-trained across multiple campus food service locations to enable flexible deployment during peak periods. Core permanent staff (3 servers) can be supplemented by mobile teams during breakfast (additional 2-3 servers) and lunch rushes (additional 7 servers) drawn from nearby facilities experiencing off-peak periods.

Split-shift scheduling: Employee schedules should align with demand patterns through split shifts targeting breakfast (07:30-09:00) and lunch (10:30-12:00) peaks, with extended shifts for core staff covering full operational hours and part-time staff handling specific rush periods.

Technology interventions: Mobile ordering systems can smooth arrival patterns by enabling pre-orders for scheduled pickup times, reducing peak-period clustering. Digital queue management systems provide customers with real-time wait information, potentially redistributing demand to adjacent time periods and improving perceived service quality during unavoidable delays.

Demand management strategies: Academic scheduling coordination can stagger class dismissal times to distribute lunch demand across broader windows. Promotional pricing during off-peak hours (14:00-15:30) can incentivize demand shifting, improving utilization during currently underserved periods.

6.2 Study Limitations

Several limitations warrant consideration when interpreting results. The analytical model assumes steady-state conditions that may not hold during brief 15-minute time slices, particularly during rapid demand transitions between adjacent periods. The validation outliers during extreme lunch rush (slices 14-15) confirm this limitation, demonstrating that transient queue dynamics can differ substantially from equilibrium predictions during capacity-constrained intervals.

The Poisson arrival assumption may not accurately capture synchronized class dismissals that produce burst arrivals rather than random individual arrivals. Similarly, the exponential service time assumption implies memoryless service, whereas actual service may exhibit learning curves, menu complexity variations, and systematic differences between simple and complex orders. The model treats all servers as homogeneous with identical service rates, whereas real operations may involve skill variations, training differences, and performance variability.

The optimization does not incorporate practical workforce constraints such as minimum shift lengths, mandatory break periods, staff availability limitations, or institutional regulations governing employee scheduling. The model assumes infinite queue capacity with no customer balking (refusing to join long queues) or reneging (abandoning queues after waiting), behaviors that likely occur during extreme congestion periods and would alter actual system performance.

6.3 Future Research Directions

Future research should extend this work in several directions to address identified limitations and enhance practical applicability. Development of time-dependent queueing models that explicitly account for transient dynamics during demand transitions would improve prediction accuracy for brief time periods and rapid staffing changes. Non-stationary arrival models incorporating class schedule information could better capture synchronized burst arrivals characteristic of academic environments.

Investigation of non-exponential service time distributions, such as Erlang or lognormal distributions, would test model robustness to service variability patterns. Incorporation of menu complexity effects, allowing differentiation

between express service for simple orders versus full service for complex meals, could enable targeted operational improvements. Multi-server heterogeneity models accounting for skill variations and training levels would reflect realistic workforce characteristics.

Behavioral modeling extensions should incorporate customer balking and reneging probabilities as functions of queue length and observed wait times, calibrated through empirical observation of actual customer behavior during peak periods. Integration of workforce scheduling constraints minimum shift lengths, break requirements, overtime regulations, staff availability; would produce implementable schedules that balance optimization objectives with institutional requirements.

Multi-objective optimization frameworks should simultaneously consider multiple performance criteria beyond waiting time minimization: service quality consistency across time periods, staff utilization equity, schedule stability and predictability for employees, and total operational costs including labor, materials, and opportunity costs. Robust optimization approaches accounting for demand uncertainty and stochastic variations would produce staffing schedules resilient to forecast errors and unexpected demand fluctuations.
Comparative analysis across multiple institutional dining facilities with varying operational characteristics would establish generalizable insights and best practices. Integration with complementary technologies, mobile ordering, self-service kiosks, automated payment systems, should be modeled to quantify potential synergies and guide technology investment decisions. Longitudinal studies tracking implementation outcomes, customer satisfaction evolution, and staff adaptation patterns would validate predicted benefits and identify unanticipated operational challenges.

Finally, machine learning approaches offer promising avenues for demand forecasting incorporating weather patterns, exam schedules, campus events, and historical trends to improve arrival rate predictions. Reinforcement learning frameworks could develop adaptive staffing policies that teach from operational experience and automatically adjust to changing demand patterns over time. These extensions would advance both theoretical understanding of institutional service operations and practical implementation of data-driven workforce optimization.

Funding Statement

This project did not receive any external financial support; it was fully funded by the authors.

References

Anand, K. S., Paç, M. F., and Veeraraghavan, S. K., "Quality–speed conundrum: Trade-offs in customer-intensive services," *Management Science*, vol. 57, no. 1, p. 40, **2010**, doi:10.1287/mnsc.1100.1250.

Chen, X., and Worthington, D., "Staffing of time-varying queues using a geometric discrete time modelling approach," *Annals of Operations Research*, vol. 252, no. 1, p. 63, **2015**, doi:10.1007/s10479-015-2058-3.

Duder, J. C., and Rosenwein, M. B., "Towards 'zero abandonments' in call center performance," *European Journal of Operational Research*, vol. 135, no. 1, p. 50, **2001**, doi:10.1016/S0377-2217(00)00289-7.

Feldman, Z., Mandelbaum, A., Massey, W. A., and Whitt, W., "Staffing of time-varying queues to achieve time-stable performance," *Management Science*, vol. 54, no. 2, p. 324, **2008**, doi:10.1287/mnsc.1070.0821.

Ghazali, N. A., and Amit, N., "Comparing the efficiency of two queuing models for a fast food restaurant using analytical queuing theory," in *Proceedings*, p. 179, **2022**, doi:10.1007/978-981-19-4910-4_17.

Gopalakrishnan, R., and Zhong, Y., "Some asymptotic properties of the Erlang-C formula in many-server limiting regimes," *arXiv* [Preprint], **2023**, doi:10.48550/arXiv.2304.13845.

Green, L. V., Kolesar, P. J., and Whitt, W., "Coping with time-varying demand when setting staffing requirements for a service system," *Production and Operations Management*, vol. 16, no. 1, p. 13, **2007**, doi:10.1111/j.1937-5956.2007.tb00164.x.

Hasija, S., Pinker, E. J., and Shumsky, R. A., "Staffing and routing in a two-tier call centre," *International Journal of Operational Research*, vol. 1, p. 8, **2005**, doi:10.1504/IJOR.2005.007431.

Hasugian, I. A., Vandrick, and Dewi, E., "Analysis of queuing models of fast food restaurant with simulation approach," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing, p. 012028, **2020**, doi:10.1088/1757-899X/851/1/012028.

Jiao, S., Zhang, Y., Li, Y., and Wang, J., "Simulation and optimization of university canteen service under the situation of epidemic prevention and control based on queuing theory," *ITM Web of Conferences*, vol. 47, p. 02052, **2022**, doi:10.1051/itmconf/20224702052.

Kamalahmadi, M., Yu, Q., and Zhou, Y., "Call to duty: Just-in-time scheduling in a restaurant chain," *Management Science*, vol. 67, no. 11, p. 6751, **2021**, doi:10.1287/mnsc.2020.3877.

Kambli, A., Sinha, A. A., and Srinivas, S., "Improving campus dining operations using capacity and queue management: A simulation-based case study," *Journal of Hospitality and Tourism Management*, vol. 43, p. 62, **2020**, doi:10.1016/j.jhtm.2020.02.008.

Lee, K. W., and Lambert, C. U., "Using simulation to manage waiting time in a cafeteria," *Information Technology in Hospitality*, vol. 4, no. 4, p. 127, **2006**, doi:10.3727/154595306779868458.

Raman, A., and Roy, D., "An analytical modeling framework for determining the optimal table mix in restaurants," *IFAC-PapersOnLine*, vol. 48, no. 3, p. 225, **2015**, doi:10.1016/j.ifacol.2015.06.085.

Robbins, T. R., "Evaluating the fit of the Erlang A model in high traffic call centers," in *Proceedings of the 2016 Winter Simulation Conference (WSC)*, p. 1790, **2016**, doi:10.1109/WSC.2016.7822226.

Robbins, T. R., Medeiros, D. J., and Harrison, T. P., "Does the Erlang C model fit in real call centers?," in *Proceedings of the 2010 Winter Simulation Conference*, p. 2853, **2010**, doi:10.1109/WSC.2010.5678980.

Silvério, J. V., Silva, A. L. F., and Santos, R. R., "Simulação computacional aplicada na análise do projeto de um restaurante universitário," *Revista Brasileira de Computação Aplicada*, vol. 8, no. 2, p. 99, **2016**, doi:10.5335/RBCA.v8i2.5553.

Smirnov, D., and Huchzermeier, A., "Analytics for labor planning in systems with load-dependent service times," *European Journal of Operational Research*, vol. 287, no. 2, p. 668, **2020**, doi:10.1016/j.ejor.2020.04.036.

Su, Q., "Optimizing university dining hall profitability during inflation: A case study of UW-Madison," in *Advances in Economics, Business and Management Research*, Atlantis Press, p. 466, **2024**, doi:10.2991/978-94-6463-368-9_54.

Sun, X., and Liu, W., "Expanding service capabilities through an on-demand workforce," *Operations Research*, vol. 73, no. 1, p. 363, **2023**, doi:10.1287/opre.2021.0651.

Usluer, F. O., "Solving queue problems in a campus dining service with discrete event simulation," *Ege Akademik Bakış (Ege Academic Review)*, vol. 25, no. 2, p. 351, **2025**, doi:10.21121/eab.20250207.

van Dijk, N. M., and van der Sluis, E., "To pool or not to pool in call centers," *Production and Operations Management*, vol. 17, no. 3, p. 296, **2008**, doi:10.3401/poms.1080.0029.

Yang, K. K., Low, J. M. W., and Çayırlı, T., "Modeling queues with simulation versus M/M/C models," *Journal of Service Science Research*, vol. 6, no. 1, p. 173, **2014**, doi:10.1007/s12927-014-0007-3.

Zhan, W., Zhang, X., Wang, Y., and Li, J., "The capacity decision-making of omnichannel catering firms based on queueing system considering customer reference behavior," *Systems*, vol. 10, no. 6, p. 229, **2022**, doi:10.3390/systems10060229.

Biographies

Arpita Debnath is currently pursuing her undergraduate degree in the Industrial and Production Engineering department at the Military Institute of Science and Technology (MIST), Dhaka, Bangladesh. She has done work in sustainability, renewable energy, simulation and modeling and 3D modeling of advanced technology. Some of her academic courses are Operations Management, Operations Research, Supply Chain Management, Quality Control, Probability & Statistics, Simulation and Modeling, Manufacturing, Machine Tools. She plans to pursue higher studies in her areas of interest and aims to build a career in academia and research.

Md Shoaib Mahmud is a Lecturer in the Department of Industrial and Production Engineering at the Military Institute of Science and Technology (MIST), Dhaka, Bangladesh. His research focuses on data-driven operations research and risk analytics, including supervised graph learning for resilient transportation networks, prescriptive analytics for urban flash-flood response, equity-aware ambulance dispatch planning, and reverse-logistics and inventory optimization in supply chains. He has co-authored multiple conference and journal publications in these areas, linking advanced analytics with disaster resilience, healthcare operations, and service systems.