PHILIPPINE ENGINEERING JOURNAL
PEJ 2025; Vol. 46, No. 1: 101-115
Copyright 2025

Received: 31 March 2025
Received in Revised Form: 26 May 2025
Accepted: 9 June 2025

A Simulation-Optimization Approach for Optimizing Service Provision of Multi-service Queues

***Simon Anthony D. Lorenzo*[1*]*, Pierre Allan C. Villena*[1]*, Angelo C. Ani*[2]*, Ven-Rem Bill A. Pasion*[2]
[1]Department of Industrial Engineering and Operations Research, College of Engineering, University of the Philippines Diliman, Quezon City 1101, Philippines
[2]Department of Industrial Engineering, College of Engineering and Agro-Industrial Technology, University of the Philippines Los Baños, Los Baños 4031, Laguna, Philippines
**Corresponding author: sdlorenzo@up.edu.ph*

Abstract – *Queueing systems in the real world can involve multiple types of services provided, such as doctors with different specializations in hospitals, highway toll booths handling cash or RFID payment, and the provision of several fuels in various dispensers in gasoline stations. These types of queues diverge from the common queue types in queueing theory, where it is assumed that only one service type is provided. This study investigates the scenario where a queueing system is to be designed to optimize the system performance with respect to relevant metrics, in particular, the average sojourn time of all customers that entered the system. Specifically, the study tackles the problem of determining which services to offer in a queueing system with a fixed number of servers and a fixed service capacity (i.e. number of services provided) per server. In order to provide a mathematically tractable solution, the combinatorial optimization problem is formulated as an integer program that is solved using the Particle Swarm metaheuristic. Results show improvements of up to 6.9342% in the identified performance upon the implementation of the optimal configuration of the queueing system. Sensitivity analysis shows the robustness of the methodology for varying mean values of the arrival distribution, allowing for a wider range of applicability in the real world.*

Keywords: *queueing optimization, multi-service queues, discrete-events simulation*

I. INTRODUCTION

Basic queueing models involve assumptions for system performance metrics, such as average queue time, average sojourn time and average queue length (Kendall, 1953). These assumptions include the provision of only a single service type within the system, which is not the case in certain instances of real-world queueing systems. For instance, banks often offer a variety of services such as deposits, withdrawals, and account opening. These tasks are often delegated to different servers (i.e., employees) due to various reasons and objectives, such as the minimization of the sojourn time of customers seeking quick transactions.

In this type of queueing system, henceforth called *multi-service queueing systems*, optimizing system performance is not just a matter of determining the number of servers to provide (as is with traditional queueing optimization problems), but also which services should be provided by each server. As an example, gasoline stations typically have a few dispensers, with each one containing pumps for a certain number of fuel types. Determining which fuel types to allocate to each pump, within each dispenser, would be a critical decision to make when trying to minimize the station's customer service time.

Relative to general queueing theory, there are only a few papers that deal with the topic of multi-service queueing systems. Within this already limited literature, most deal with the configuration of the queueing system and the number of servers, rather than the assignment of services to servers.

Two very early studies, Gumbel (1960) and Ancker and Cafarian (1963), initially conceptualized and modelled queueing systems with heterogeneous servers, initially focusing on variance in service time distributions, then also incorporating multiple service types. The concept of queueing systems with lane selection was introduced in Schwartz (1974). In this type of queueing system, customers are classified by type, and each one can only be served by designated servers of their respective type. This characterization of queueing systems is similar to the one being investigated by this study, but with some key differences. The related study only investigated two services types, and the said study tackled the problem through the derivation of certain quantities rather than the assignment of offered services to optimize a system. Green (1985) extended the study of this type of queueing model, but it narrowed down the classification of servers to general-use (i.e., can serve all types of customers) and limited-use (i.e., can serve only one type of customer). The same approach of system characterization rather than optimization was taken in the study. Gans and Van Ryzin (1997) considered a job shop queueing system which handles several jobs that exhibit routing in accordance with the needs of certain jobs. A linear program was developed and solved to determine the queueing system configuration for a certain set of jobs. Whitt (1999) took a different approach to investigating multi-service queueing systems, focusing on the classification and routing of customers of different types to existing servers. Wallace and Whitt (2005) tackled skill-based routing in multi-service queueing systems, specifically in call centers, where each server has a set of skills that corresponds to certain customer types that they can handle. The study focused on determining the number of agents and their respective skills in order to optimize the system.

Looking at more recent studies, Kim et al. (2011) did a study on heterogeneous (i.e., dissimilar) servers, but with the focus of varying service times and distributions instead of service types. Li and Stanford (2016) looked into a multi-class, multi-server queueing model with heterogenous servers, where different service types are present. The focus of the study was on service prioritization, as the queueing system that the study investigated was in the call center industry. Galankashi et al. (2016) examined the design of petrol stations as queueing systems, focusing specifically on two separate services (fueling and payment) and the number of servers that provide them. Simulation was used to model the queueing system, and design of experiments was used to analyze the results. Similar to Galankashi et al. (2016), Dwijendra et al. (2022) focused on fueling and payment as two separate services in petrol stations. The same results were obtained, showing that fuel dispenser and cashier count significantly affected queueing length. Hillas et al. (2024) investigated heavy-traffic multi-class, multi-server bipartite queueing systems wherein customers can only be served by a subset of the servers. The type of queueing system investigated is also similar to that being studied by the system, but the scope of the study is substantially different. This study did not look into service provision per server, and focused on server assignment instead. Additionally, the inclusion of the heavy-traffic condition further differentiates this study.

From here, we see that the problem of selecting which services to provide in a queueing system with multiple multi-service servers has not been tackled, even though it has been established that this type of queueing system is relevant in the real world. As such, this study addresses this gap through the mathematical formulation of this type of queueing system, as well as the modelling through simulation of a real-world queueing system that exhibits this behavior. The remainder of the paper is as follows: Section 2 discusses the mathematical formulation of the problem, Section 3 deals with the methodology for finding an optimal solution for this problem, Section 4 discusses the results produced, Section 5 presents relevant hypothetical scenarios through sensitivity analysis, and finally, Section 6 concludes the paper.

II. METHODOLOGY

2.1 System Description

To more concretely illustrate the operations of multi-service queues, the study investigates a concrete real-world example of a queueing system of this type. Specifically, we look at a petrol station that has multiple dispensers. Without loss of generality, this study assumes that each dispenser can hold 3 pumps as observed in the fuel station visited. Note though that dispensers with 2 or 4 pumps are available in other petrol stations. Each pump delivers a certain fuel type. The following characterize the system of interest in more detail:

1. The station sells I fuel types.
2. The station has J fuel dispensers. Each fuel dispenser accommodates two queues each (i.e., one queue on each side of the dispenser). Hence, the number of queues in the system is $2J$, each having its own server.
3. While a dispenser is connected to 6 color-coded pumps – three on each side – the physical design of the pumps dictates that both sides should discharge the same set of fuel types, since the pairs of pumps are situated on a single dispenser. Thus, only three fuel types can be assigned to a fuel dispenser.
4. Vehicle arrivals have been observed to follow a Poisson process with mean arrival rate λ.
5. The vehicles are classified into K types, $k \in \{A, B\}$. Type A vehicles are those with four or more wheels, while type B are those with two or three wheels. Their respective proportions are denoted as p_k. Respective arrival distributions are still Poisson distributed, with respective mean arrival rates λ_A and λ_B, $\lambda_A + \lambda_B = \lambda$.
6. Different vehicle types require different fuel types. This results in differing probabilities for the required fuel type by each vehicle type. The probability that type k vehicle will require fuel j is denoted by p_{kj}.
7. Service times have also been observed to be different for the two vehicle types. Their probability distributions are characterized by the cumulative distribution functions $G_A(y)$ and $G_B(y)$.

A sample layout for an instance of the problem is illustrated in Figure 1. In this example, there are $J = 3$ fuel dispensers, producing 6 queueing lines as shown.

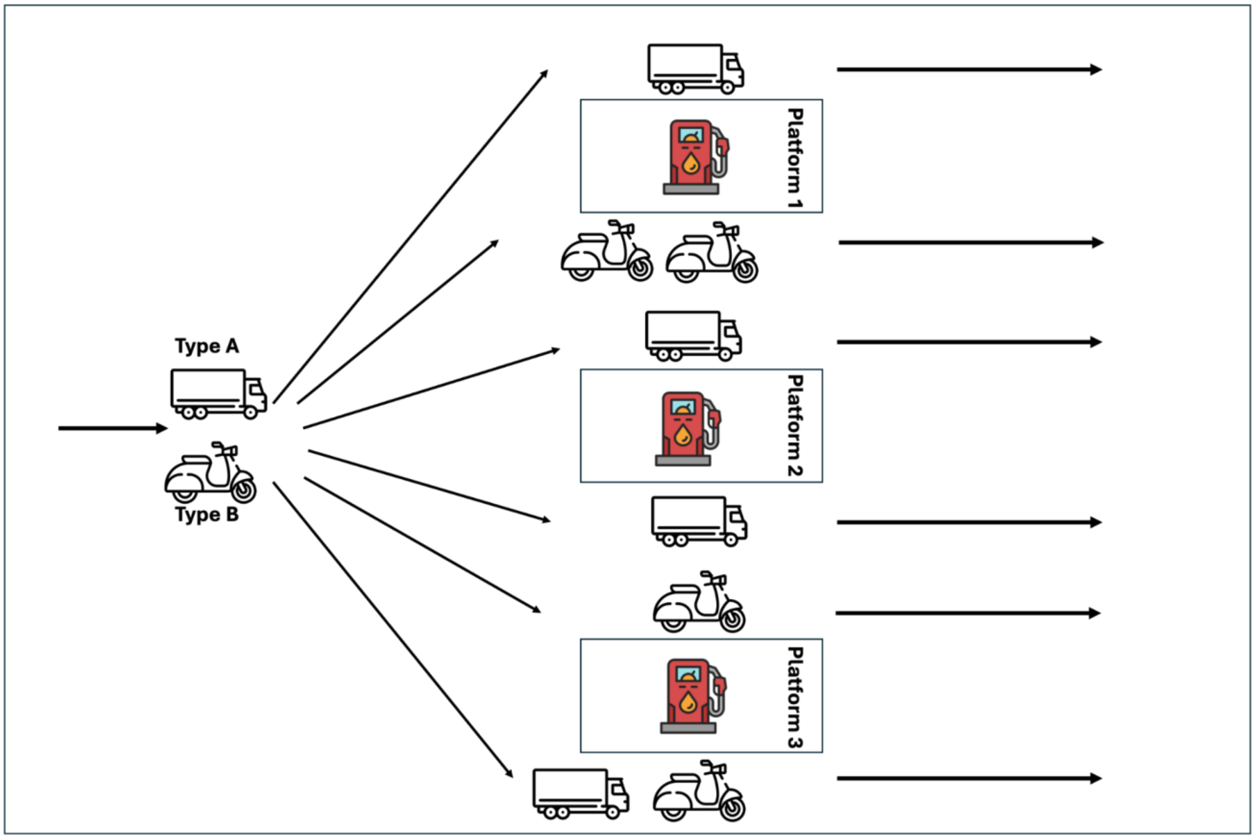

Figure 1. Petrol station layout with 3 dispensers and 6 queues.

This queueing system is fairly straightforward. As described in Figure 2, it begins with the vehicles arriving at the station, where customers first identify the dispensers that offer their desired fuel type. Once the appropriate dispensers are determined, customers check which of these dispensers has the least number of customers (i.e., shortest queue length including those in queue and in server). The vehicle is then refueled when it arrives at the chosen dispenser. While this is ongoing, the customer may choose to avail of additional services, such as windshield cleaning. Refueling, additional services, and payment are lumped into one service in the model by aggregating their service times and defining it as a single distribution. Following refueling, the customer makes the payment and leaves the petrol station.

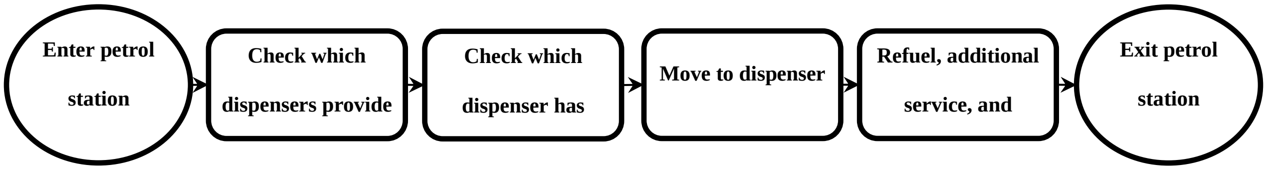
Enter petrol station
Check which dispensers provide
Check which dispenser has
Move to dispenser
Refuel, additional service, and
Exit petrol station

Figure 2. Process Mapping Diagram for vehicles entering the petrol station.

2.2 Mathematical Model

In the real world, it is most likely the case that the mathematical model describing this queueing process is one that maximizes profit. Another objective often used in queueing systems is customer satisfaction, which is affected by the experience of the customer within the system. For this study, we focus on the objective of the average customer *sojourn time*, which is the total time a customer spends in the system. This is the sum of the time spent waiting for service (if there is any) and the time spent being served. This metric, which is represented by the variable W in queuing theory, directly influences a customer's positive experience (i.e., customer satisfaction) (Liang, 2016). That is, minimizing W would roughly be equivalent to maximizing customer satisfaction. We thus use the system performance metric W as our objective function value to be minimized.

Based on the characteristics of the problem described in the previous subsection, we present the following mathematical formulation. The indices, decision variables, and parameters are summarized in Table 1, Table 2, and Table 3, respectively.

Table 1. Indices used in mathematical model.

Index	Description
i	Fuel type
j	Fuel dispenser
k	Vehicle type

Table 2. Decision variables used in mathematical model.

Decision Variable	Description
X_{ij}	Binary variable with value 1 indicating fuel type i is allocated to fuel dispenser j, 0 otherwise
W	Average customer sojourn time, also commonly known as average customer waiting time in the system

Table 3. Parameters used in mathematical model.

Parameter	Description
λ	Mean of Poisson-distributed Vehicle Arrival Rate
p_k	Proportion of vehicle type k
p_{kj}	Proportion of fuel type j requirement for vehicle k
$G_k(y)$	Cumulative distribution function of service time y of vehicle k

$$Min\ Z = W \tag{1}$$

s.t.

$$\sum_{j} X_{ij} \geq 1, \forall\, i \tag{2}$$

$$\sum_{i} X_{ij} = 3, \forall\, j \tag{3}$$

$$W \ = f\left(X_{ij}, \lambda, p_k, p_{kj}, G_k(y)\right) \tag{4}$$

$$X_{ij} \in \{0,1\}, \forall\, i, j \tag{5}$$

As previously discussed, the objective function (1) is the minimization of the average customer sojourn time across the system, W. Constraint (2) ensures that each service type is provided by at least one dispenser in the system. Constraint (3) limits the capacity of each fuel dispenser to exactly three fuel types. Constraint (4) indicates that W is a function of various system parameters and decision variables. Finally, Constraint (5) ensures that the assignment decision variables are binary.

III. SOLUTION METHODOLOGY

With the model formulation provided in the previous section, Constraint (4) denotes that the objective function *W* is a continuous variable that is a function of system parameters and decision variables. Due to the complexity of the customer queue joining process (i.e., determining valid servers, selecting the shortest server) and model assumptions (e.g., non-Markovian service times, heterogeneous service), traditional ways of expressing *W* through closed-form mathematical expressions are not applicable. As such, the study makes use of simulation to model the complex behavior of customers as they traverse the queueing system.

In order to improve the reward returned by the simulation (i.e., objective function value, *W*), we incorporate an optimization step that considers this returned value of each simulation, resulting in a *simulation-optimization* approach. Since this problem of selecting which services to provide results in a combinatorial optimization problem that is NP-hard (Toth, 2000), this study makes use of a metaheuristic, specifically Particle Swarm Optimization (PSO), to produce optimal (or locally optimal) solutions.

3.1 Particle Swarm Algorithm

The particle swarm optimization (PSO) is a population-based metaheuristic that resembles group dynamics of birds and fish (Talbi, 2009). PSO is able to provide a close-to-optimal solution by efficiently iterating through the solution space, for computationally-intensive problems that cannot be solved optimally through solution exhaustion. The use of

PSO in this study stemmed from its simplicity. It relies on basic mathematical operations and is computationally efficient, requiring minimal memory and processing power (Kennedy and Eberhart, 1995). This is especially critical considering that each PSO iteration involves hundreds of replications, with each simulating hundreds of vehicles.

Binary encoding is used to represent a candidate solution to the problem, through the use of a Sigmoid function (Nezamabadi-pour et al., (2008). A vector **x** of binary variables of size *IJ* reflects the solution. The first *I* elements indicate whether the corresponding fuel type should be dispensed by the first dispenser, the second *I* elements by the second dispenser, and so on.

A pseudocode of the metaheuristic is found in Figure 3. The algorithm stops when the set number of iterations are performed.

```
Set maximum iterations, number of particles, dimension, $w$, $c_1$, $c_2$

Generate Swarm Population
    For all particles p
        Initialize particle position vector $\mathbf{x_p}^1$
        Initialize particle velocity, $\mathbf{v_p}^1$
        Evaluate fitness function $f(\mathbf{x_p}^1)$, using queueing simulation
        Update $Pbest_p^1$, $Gbest^1$
For all iterations t
    For all particles p
        Update velocity vector $\mathbf{v_p}^t = w\mathbf{v_p}^{t-1} + c_1\mathbf{r_1}(Pbest_p - \mathbf{x_p}^{t-1}) + c_2\mathbf{r_2}(Gbest - \mathbf{x_p}^{t-1})$, $\mathbf{r_1}$ and $\mathbf{r_2}$
            are vectors of random decimals
         For all dimensions d
            If $r_3 < S(v_{pd}^t)$, then $x_{pd}^t = 1$, else $x_{pd}^t = 0$, $r_3$ is a random decimal and $S$ is the sigmoid
                function
        Evaluate fitness function $f(\mathbf{x_p}^t)$, using queueing simulation
        Update $Pbest_p^t$, $Gbest^t$
Return $Gbest^t$, corresponding solution $\mathbf{x_p}^t$
```

Figure 3. Pseudocode of PSO as applied in this study.

3.2 Simulation Model

Given the difficulty in evaluating the fitness function of a candidate solution as no analytical formula relates the average system waiting time and the other queueing model parameters, this study employs simulation to evaluate system performance. When the value of the fitness function is sought by the PSO algorithm, the queueing simulation is triggered.

In this simulation, the number of replications is set by the analyst. The details of the simulation are reflected in Figure 4. Aside from the details provided in the description of the system, the following additional assumptions are made in the simulation.

1. Upon arrival, a vehicle identifies the fuel dispensers that offer their desired fuel type. From these fuel dispensers, the vehicle selects the queue with the least number of vehicles, considering both those being served and waiting in line.

2. If there are multiple queues that meet the criteria (must offer desired fuel type, shortest queue length), the vehicle will randomly select from these queues.

3. Balking, reneging, and jockeying behaviors are negligible and are not considered in the study. This means that no vehicle is barred from entering the queueing system. Once a customer enters the queue, they remain in the system until they are fully serviced (refueling, additional services, payment). They also do not switch queues.

The simulation of the queueing system was implemented using Python, through the use of the SimPy (v4.1.1) simulation framework. The pseudocode of the model is presented in Figure 4.

```
Set number of replicates, simulation time

For all replicates r
    Repeat
        Generate vehicle
        Simulate vehicle inter-arrival time based on Poisson arrival
        Compute vehicle arrival time
        Simulate vehicle type based on Bernoulli distribution
        Simulate vehicle fuel type based on probability distribution dependent on vehicle
        type
        Simulate vehicle service time based on probability distribution appropriate for
        the vehicle type
    While vehicle arrival time is within simulation time

For each vehicle that arrives within the simulation time
    Select queue with the least vehicle count that offers the desired fuel type. Break ties
    randomly.
    If server is occupied, wait in queue until server becomes available.
    Complete service time.

    Evaluate average vehicle waiting time in the system for the replicate, computed when
    simulation time ends

Return average vehicle waiting time in the system considering all replicates
```

Figure 4. Pseudocode of the queueing system simulation.

IV. NUMERICAL RESULTS

4.1 Estimation of System Parameters

In order to get realistic estimates of the relevant parameters, actual data was gathered on the arrival times, service times, and service types required by customers in a real-world queuing service system that exhibits the characteristics considered in this study. The observation was conducted during a time window when vehicle arrival is relatively at its peak within the day.

The fuel station observed offers $I = 5$ fuel types among its $J = 3$ fuel dispensers. There are $K = 2$ vehicle types as described earlier. Other system parameters are presented in Table 4.

Table 4. Queueing system parameter values.

Parameter	Value
Mean of Poisson-distributed vehicle arrival rate, λ	62.52 vehicles / hr
Proportions of vehicle types, p_k	$p_A = 0.47$, $p_B = 0.53$
Proportions of fuel requirement of a vehicle type, p_{kj}	$p_{A1} = 0.0417$, $p_{A2} = 0.2917$, $p_{A3} = 0.0417$, $p_{A4} = 0.2917$, $p_{A5} = 0.3333$, $p_{B1} = 0.0000$, $p_{B2} = 0.0000$, $p_{B3} = 0.0000$, $p_{B4} = 0.5000$, $p_{B5} = 0.5000$
Service time distribution per vehicle type, in seconds	$G_A(y)$: gamma(shape = 5, scale = 49.59) $G_B(y)$: gamma(shape = 3, scale = 36.46)

4.2 Parameters of PSO algorithm and queueing simulation.

The hyperparameters of the PSO metaheuristic include inertia weight (w), cognitive coefficient (c_1), social coefficient (c_2), and population size. For this study, the values used were w = 0.7, $c_1 = 1.5$, and $c_2 = 1.5$, with 50 particles in each iteration. These values are within the commonly-used values for PSO hyperparameters, slightly adjusted to follow a slower, more stable convergence (Bigdeli, 2015; Eberhart and Shi, 2000). After 500 iterations, the algorithm is terminated with the identified near-optimal solution.

Regarding the queueing simulation, 100 replicates were performed per run. This results in a standard error of 0.012689 for the base (current) model, suggesting fairly consistent results. The simulation run time is set at 2 hours, reflecting the estimated duration of a peak period in the fuel station. Warm up period is deemed unnecessary since peak period is usually preceded by a lull in vehicle arrival, at least for the station observed. However, the addition of warm up time may be considered in other models depending on how the initial situation affects the results.

4.3 Results

The simulation is initially run to illustrate the base case of the queuing system. Using the data obtained through observation as values to the simulation parameters, we run the simulation-optimization methodology to come up with the best solution to our mathematical program. We arrive at the following solution to the model, summarized through the following vector of decision variables.

From the equation of vectors, we identify that solution for the optimal configuration, *O*, the optimal configuration in Table 5 is providing fuel types 3, 4 and 5 to dispenser 1, fuel types 1, 4 and 5 to dispenser 2 and fuel types 1, 2, and 4 to dispenser 3, resulting in a W^O value of 3.2050 minutes.

Table 5. Optimal Solution.

Solution Type	Variable Values
Optimal Solution	$[X_{11}^O\ X_{21}^O\ X_{31}^O\ X_{41}^O\ X_{51}^O\ X_{12}^O\ X_{22}^O\ X_{32}^O\ X_{42}^O\ X_{52}^O\ X_{13}^O\ X_{23}^O\ X_{33}^O\ X_{43}^O\ X_{53}^O]$ = [0 0 1 1 1 1 0 0 1 1 1 1 0 1 0]

In order to assess whether the *simulation-optimization* methodology of the study has produced a better solution than the current one, we compare the base case W^C, which was obtained using the current configuration shown in Table 6.

Table 6. Current Configuration.

Solution Type	Variable Values
Current Configuration	$[X_{11}^C\ X_{21}^C\ X_{31}^C\ X_{41}^C\ X_{51}^C\ X_{12}^C\ X_{22}^C\ X_{32}^C\ X_{42}^C\ X_{52}^C\ X_{13}^C\ X_{23}^C\ X_{33}^C\ X_{43}^C\ X_{53}^C]$ = [1 1 1 0 0 0 1 0 1 1 0 1 0 1 1]

The current configuration has an evaluated W^C = 3.3821 minutes. Compared to the current configuration, the identified optimal solution with W^O = 3.2050 minutes comes up to a 5.2364% decrease in average sojourn time for the customer, simply by reconfiguring the services provided by the servers. The 95% confidence intervals of the current configuration and the optimal solution are shown in Table 7 and Table 8 respectively. This set of results was minimally different from the result reported earlier as these figures were identified by running the relevant configuration in the queueing simulation with 1,000 replications.

Table 7. Confidence Interval of the Current Configuration.

Metric	Mean	Std Dev	± 95% CI
Total Time in the System	3.3888	0.3511	0.0218

Table 8. Confidence Interval of the Optimal Solution.

Metric	Mean	Std Dev	± 95% CI
Total Time in the System	3.1540	0.2613	0.0162

These show that the difference between the distribution means is statistically significant (p-value = 0.000), numerically proving that the improvement in *W* is statistically significant as well. The proposed design, shown in Figure 5, aims to minimize the average sojourn time at the petrol station. To achieve this, it is recommended to make the following adjustments: assign products 4 and 5 to dispenser 1 (replacing products 1 and 2), move product 1 to dispenser 3 (replacing product 5), and increase the number of pumps allocated for product 1 on dispenser 2 (replacing product 2).

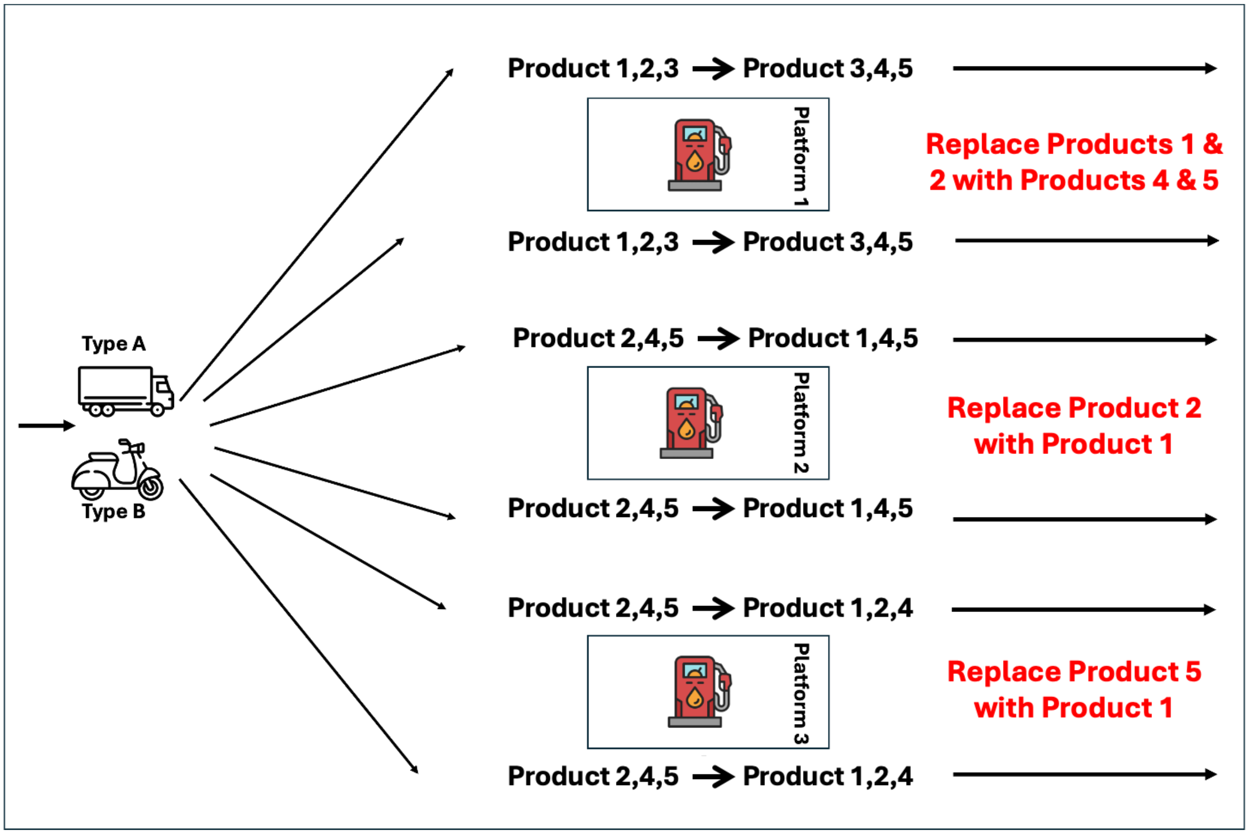
Type A
Type B
Product 1,2,3 → Product 3,4,5
Platform 1
Replace Products 1 & 2 with Products 4 & 5
Product 1,2,3 → Product 3,4,5
Product 2,4,5 → Product 1,4,5
Platform 2
Replace Product 2 with Product 1
Product 2,4,5 → Product 1,4,5
Product 2,4,5 → Product 1,2,4
Platform 3
Replace Product 5 with Product 1
Product 2,4,5 → Product 1,2,4

Figure 5. Proposed Design for the Petrol Station.

Further analysis was conducted to evaluate the proposed design of the PSO algorithm. As shown in Table 4, the vehicles requiring product 1 have one of the lowest arrival rates. However, the PSO algorithm recommended increasing the number of pumps assigned to these vehicles. To investigate this counterintuitive recommendation, two modifications were made: first, the changes in dispenser 2 were reversed by not replacing product 2 with product 1; second, the changes in dispenser 3 were reversed by not replacing product 5 with product 1. The results of these adjustments are presented below.

Table 9. Results of Reversing the PSO Recommendation.

Changes	Configuration	Mean	Standard Deviation	P-Value (vs. recommended)
Do not replace 2 with 1	[0 0 1 1 1 0 1 0 1 1 1 1 0 1 0]	3.1393	0.2577	0.2054
Do not replace 5 with 1	[0 0 1 1 1 1 0 0 1 1 0 1 0 1 1]	3.1481	0.2507	0.6064

The results above indicate that reversing the recommendation of the PSO algorithm does not lead to a statistically-significant change in performance. The notable improvement comes from increasing the number of pumps offering product 4 – specifically, when all dispensers are equipped with pumps for product 4. This configuration provides greater flexibility, allowing vehicles to switch to alternative dispensers when those serving product 5 are occupied. This is particularly beneficial given the high arrival rate of cars requiring product 5.

V. SENSITIVITY ANALYSIS

In order to test the robustness of the study's methodology, we run additional scenarios where the mean arrival rate, λ, is adjusted to values lower and higher than the base case. Though customer arrival is stochastic, there may be instances where the actual distribution parameters change, such as the increase of the mean arrival rate due to residential establishments increasing the customer base of the fuel station. We look into these situations to test whether or not the study's methodology still holds for different values of λ and not just the one observed. The same objective function value *W* is compared for the current configuration and for the optimal configuration as determined by the *simulation-optimization* methodology.

Our sensitivity analysis tests 5 new scenarios where the new mean customer arrival rate λ' is obtained through the multiplication of a constant factor. Specifically, we test the scenarios where $\lambda' = 0.5\lambda$, 0.75λ, 1.5λ, 2λ, 2.5λ. This tests the robustness of the methodology of the study. In this phase, the *simulation-optimization* methodology of simulating the queueing system then utilizing PSO to come up with the optimal solution is run for each new λ. Table 10 shows the summary of the results:

Table 10. Sensitivity analysis results.

Adjusted Arrival Rate, λ'	Optimal Configuration	W, mins		% Change in W	p-value of difference
		Current config	Optimal Config		
0.50λ	[1 1 1 0 0 1 0 0 1 1 1 0 0 1 1]	2.958	**2.943**	0.5071%	0.0016
0.75λ	[0 1 1 1 0 1 0 0 1 1 1 1 0 0 1]	3.129	**2.983**	4.6660%	0.0000
1.00λ	[1 1 1 0 0 0 1 0 1 1 0 1 0 1 1]	3.389	**3.154**	6.9342%	0.0000

1.50λ	[1 0 0 1 1 0 1 1 0 1 0 1 1 1 0]	5.020	**3.933**	21.6534%	0.0000
2.00λ	[0 0 1 1 1 1 1 0 0 1 0 1 1 1 0]	11.041	**7.402**	32.9590%	0.0000
2.50λ	[1 0 1 0 1 1 0 1 0 1 1 1 0 1 0]	18.253	**15.728**	13.8333%	0.0000

We can see from the table that W has been seen to decrease upon utilization of the study's methodology to determine the optimal configuration. We see a consistent decreasing trend in W, though varying in magnitude. This tells us that the study's methodology is consistent in producing better-than-current solutions even if distribution parameters change.

To prove the significance of the differences between the current and optimal configurations, the p-value of corresponding t-test comparisons are also computed. The p-values are all close to 0, further solidifying the significance of the improvement of the sojourn time.

On the cases where $\lambda' = 2.50\lambda$ *and* $\lambda' = 2.00\lambda$, we see that the magnitude decrease has lowered from the previous value of 13.8333% and 32.9590%, respectively. This is due to the % decrease in W being a relative metric, and the absolute reduction of W has been sustained. However, this tells us that in higher values of λ, other solutions may be needed to further reduce W, as required.

In order to further illustrate the benefits of the methodology, other metrics are computed for each new value of λ as well, as shown in 11.

Table 11. Sensitivity analysis supporting metrics.

Arrival Rate λ'	**W_s, mins**		**Type A, num**		**Type B, num**	
	Current config	**Optimal Config**	**Current config**	**Optimal Config**	**Current config**	**Optimal Config**
0.50λ	2.9116	2.8978	29	30	33	33
0.75λ	2.9134	2.9055	44	43	50	50
1.00λ	2.9082	2.9014	58	60	66	65
1.50λ	2.9177	2.9152	87	88	97	98
2.00λ	2.9345	2.9084	106	109	112	124
2.50λ	2.9727	2.8691	116	116	116	130

From Table 11, we see that W_s does not change significantly at all. This aligns with the theoretical expectation that the service time W_s is not affected by changes in the system (only changes in the service distribution would change W_s). We also compute the actual count of customers served and see that there is no significant difference between the current configuration and the optimal configuration. This tells us that while the methodology is effective in reducing the main metric *W*, it is not effective in increasing other metrics such as the system throughput. Provided that all fuel types are available and system capacity can accommodate the demand, all customers are eventually served.

VI. CONCLUSION AND AREAS FOR FURTHER STUDY

6.1 Conclusion

Muti-service queueing systems have several applications in real life, and determining the service types to provide is a real problem that can increase system performance when solved. This study utilized a simulate-and-optimize approach to determine if improvements can be made to a real-world multi-service queue. With the aid of PSO, a close-to-optimal solution that performs better than the current configuration has been produced within a reasonable amount of time, especially in the context of the strategic-to-tactical decision of which service to provide. The customer sojourn time has been shown to decrease by 6.9342% by transitioning from the current setup to the close-to-optimal one, and this is done without any modification to the queuing system other than reconfiguring the services provided.

6.2 Recommendations for Future Studies

While this study closely approximates actual fuel stations, future studies may incorporate additional elements in the model to further enhance its validity. For instance, the inclusion of pump attendants in the simulation may be considered, especially if their availability is a limiting resource. Balking, reneging, and jockeying may also influence model results when they become prevalent. For fuel stations along two-way roads, direction of vehicles and the location of vehicle fuel tank can considerably affect the driver's choice of preferred queue. The impact of this may be explored by future studies. It is also common to see fuel stations with two sequential fuel dispensers per platform (i.e., the physical structure where dispensers are placed), representing multi-server setup. If a station is configured this way, blocking may be substantial, which merits additional analysis.

VII. ACKNOWLEDGEMENTS

The authors are grateful to the fuel station owner, who allowed the team to collect data relevant for this study. We also wish to thank the reviewers who provided valuable insights that improved this paper.

References:

[1] Ancker CJ, Cafarian AV. 1963. Queuing with reneging and multiple heterogeneous servers. Naval Research Logistics. 10:125-149. https://doi.org/10.1002/nav.3800100112

[2] Bigdeli N. 2015. Optimal management of hybrid PV/fuel cell/battery power system: A comparison of optimal hybrid approaches. Renewable and Sustainable Energy Reviews. 42:377-393. https://doi.org/10.1016/j.rser.2014.10.032

[3] Dwijendra N, Vaslavskaya I, Skvortsov NV, Rakhlis TP, RahardjaU, Ali MH, Iswanto AH, Thangavelu L, Kadhim MM. 2022. Application of experimental design in optimizing fuel station queuing system. Industrial Engineering and Management Systems. 21:381-389. https://doi.org/10.7232/iems.2022.21.2.381

[4] Eberhart RC, Shi Y. 2000. Comparing inertia weights and constriction factors in particle swarm optimization. 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512); La Jolla, CA, USA. IEEE. p. 84-88. https://doi.org/10.1109/CEC.2000.870279

[5] Galankashi MR, Fallahiarezoudar E, Moazzami A, Yusof NM, Helmi SA. 2016. Performance evaluation of a petrol station queuing system: A simulation-based design of experiments study. Advances in Engineering Software. 92:15-26. https://doi.org/10.1016/j.advengsoft.2015.10.004

[6] Gans N, Van Ryzin G. 1997. Optimal control of a multiclass, flexible queueing system. Operations Research. 45:677-693. https://doi.org/10.1287/opre.45.5.677

[7] Green L. 1985. A Queueing system with general-use and limited-use servers. Operations Research. 33:168-182. https://doi.org/10.1287/opre.33.1.168

[8] Gumbel H. 1960. Waiting lines with heterogeneous servers. Operations Research. 8:504-511. https://doi.org/10.1287/opre.8.4.504

[9] Hillas LA, Caldentey R, Gupta V. 2024. Heavy traffic analysis of multi-class bipartite queueing systems under FCFS. Queueing Syst. 106:239–284. https://doi.org/10.1007/s11134-024-09903-4

[10] Kendall DG. 1953. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov Chain. The Annals of Mathematical Statistics. 24:338-354.

[11] Kennedy J, Eberhart R. 1995. Particle Swarm Optimization. International Conference on Neural Networks; Perth, WA, Australia. IEEE. 4:1942-1948. doi: 10.1109/ICNN.1995.488968

[12] Kim JH, Ahn HS, Righter R. 2011. Managing queues with heterogenous servers. Journal of Applied Probability 48(02):435-452. DOI: 10.1017/S002190020000797X

[13] Li N, Stanford DA. 2016. Multi-server accumulating priority queues with heterogeneous servers. European Journal of Operational Research. 252:866-878. https://doi.org/10.1016/j.ejor.2016.02.010

[14] Liang CC. 2016. Queueing management and improving customer experience: empirical evidence regarding enjoyable queues. Journal of Consumer Marketing. 33:257-268. https://doi.org/10.1108/JCM-07-2014-1073

[15] Nezamabadi-pour H, Rostami M, Farsangi M. 2008. Binary particle swarm optimization: Challenges and new solutions. The Journal of Computer Society of Iran On Computer Science and Engineering 6.

[16] Schwartz BL. 1974. Queuing models with lane selection: A new class of problems. Operations Research. 22:331-339. https://doi.org/10.1287/opre.22.2.331

[17] Talbi E. 2009. Metaheuristics: From design to implementation. 1st ed. Wiley. https://doi.org/10.1002/9780470496916

[18] Toth P. 2000. Optimization engineering techniques for the exact solution of NP-hard combinatorial optimization problems. European Journal of Operational Research 125:222-238. https://doi.org/10.1016/S0377-2217(99)00453-1

[19] Wallace RB, Whitt W. 2005. A staffing algorithm for call centers with skill-based routing. Manufacturing and Service Operations Management. 7: 276-294. https://doi.org/10.1287/msom.1050.0086

[20] Whitt W. 1999. Partitioning customers into service groups. Management Science. 45:1579-1592. https://doi.org/10.1287/mnsc.45.11.1579