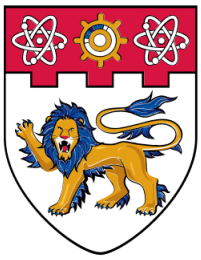


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

MH4702 Simulation Project

Decision-dependent non-stationary Queuing system

November 6, 2021

Author	*Student ID*
SHENG SHUNAN	U1840636E
TAN LIN MIN	U1940871K
WILLSON LIM WEI SEN	N2101078E

Contents

1 Introduction

Aviation is a vital component of Singapore's business. We hope to improve passengers' travel experiences by reducing their waiting time at the immigration counters. We approach this by modeling the airport's immigration counters as a form of the queuing system and therefore optimizing it. In the airport, arrival rates vary with time. As traditional queuing models assume stationary arrival, using them will affect the accuracy of our analysis [Green et al., 1991]. To tackle this, we consider different arrival rates independently and use numerical methods such as Pointwise Stationary Approximation (PSA) to analyze the performance of the queuing system [Green and Kolesar, 1991]. Furthermore, we also consider the effects of waiting time when opening up more counters (1 server per counter) at different times as a response to different arrival rates [Wang and Tai, 2000]. Hence, in this paper, we seek to optimize the queuing system of the airport's immigration by finding the optimal time to introduce new counters while subjected to cost constraints and non-stationary arrival rates.

2 Model Setup

In this report, we will first explore the $M/M/s$ system, then generalize to a $M(t)/M/s(t)$ system where λ_t is the arrival rate at time t, μ is the service rate, s is the number of servers, and queue discipline will follow the first-come-first-served policy.

To approach this, we first consider a simpler queuing system with only two different arrival rates then we generalize to a more complex queuing system with K different arrival rates and K different costs (for opening up new counters). Consider a time frame from $[0, T]$ and we let t_0 be an arbitrary time in $[0, T]$. From $[0, t_0]$, the arrival rate is λ_1 and from $[t_0, T]$ the arrival rate is λ_2, and we let $\lambda_2 > \lambda_1$. At the time $s \geq t_0$, a new counter (server) will be introduced to accommodate the influx of crowds. Here, we assume the service rate of any counter to be constant μ. Now, we seek to minimize average waiting queuing time in the queuing system W_q subjected to cost constraint, the optimization problem (P), as follows:

$$\arg\min_{s}(W_q) \tag{1}$$

such that

$$\begin{aligned} \frac{s}{T}c_1 + \frac{(T-s)}{T}c_2 &< C \\ s &\geq t_0 \end{aligned} \tag{2}$$

where c_1, c_2 and C are the costs of opening counters 1, 2 and the upper bound of cost respectively. We fixed $c_1 < C < c_2$ to make this problem well-defined.

Next, we generalise to include K different arrival rates and costs i.e., let $\lambda_1, \lambda_2, \cdots, \lambda_K$ be arrival rates during $[t_0, t_1], [t_1, t_2], ..., [t_{K-1}, T]$ respectively and $c_1, c_2, \cdots, c_K$ be costs of opening up counters $1, 2, \cdots, K$. Now, our objective is to find optimal $(s_1, .., s_{K-1})$ by enumerating them from 0 to T for different K. We define the problem as follow:

$$\arg\min_{S}(W_q) \tag{3}$$

such that

$$\frac{s_1}{T}c_1 + \frac{s_2 - s_1}{T}c_2 + \cdots + \frac{s_{K-1} - s_{K-2}}{T}c_{K-1} + \frac{T - s_{K-1}}{T}c_K < C \tag{4}$$

$$s_{i+1} \geq s_i \text{ and } s_i \geq t_{i-1} \text{ and } s_{K-1} \geq t_{K-2} \text{ for } 1 \leq i \leq K-2 \tag{5}$$

where $S = (s_1, s_2, ..., s_{K-1})$ and $c_1 < c_2 < \cdots < c_K$ and $C < c_K$.

3 Simulation

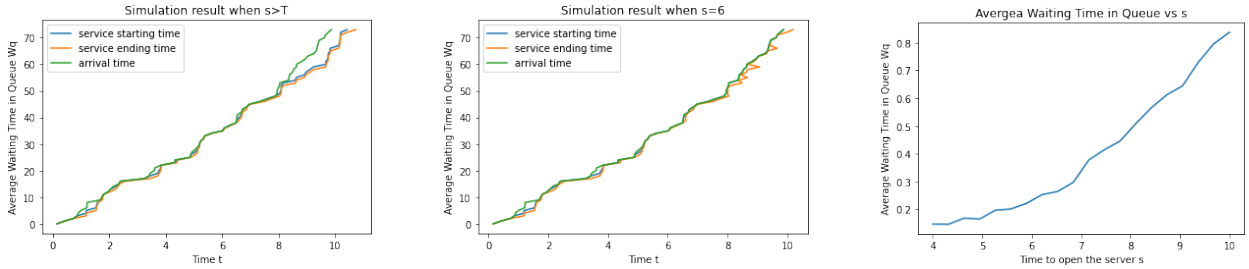
In this section, we used Monte-Carlo Simulation to study the behaviors of the decision-dependent of a non-stationary queuing systems when $K = 2, 3$. The simulation may also be applied to general K; however, the extension will be beyond the scope of this paper.

3.1 Simulation when $K = 2$

We start with the case when $K = 2$. We first study the behaviour of the model when $T = 10$, $t_0 = 4$, and we let the arrival rates $\lambda_1 = 6$ during $[0, t_0]$, $\lambda_2 = 9.9$ during $[t_0, T]$ and we fix constant service rate $\mu = 10$, the cost to maintain i servers is $c_i = 5i$ per hour for $i = 1, 2$, and the upper bound for the cost $C = 8$ per hour. In a real world scenario, this setup can be presented as follows: assume we consider the customer with one/two counters opening from 8:00 a.m. to 6:00 p. and at 12.00 p.m., there is an influx of crowd and in response the manager needs to decide the optimal time to open a new counter. Of course, in real life, the exact time of the influx change may only be observable with a certain delay, say half an hour, but it will not change the essence of our model.

In our simulation, the inter-arrival time of customers is exponentially distributed at varying arrival rates. The servers are dealing with customers with service time following exponential distribution at the given service rate. After a new server is open, the customer starts his/her service once one of the two predecessors who currently being served finished their service. The waiting time in the queue of each customer is computed by subtracting arrival time from the service starting time. The average waiting time in queue W_q for each iteration is calculated by taking the average among all customers served. We run our model for $10,000$ iterations and calculate the average of all W_qs to obtain a better estimate.

Next, we vary the time slot to open a new server. When $s > T$, that is, there is no additional server introduced during $[0, T]$, we found that $W_q = 0.411$; while when $s = 6$, $W_q = 0.108$, a significant decline compared to the previous case. For better illustration, we plot the service starting time, service ending time, and arrival time over $[0, T]$. Each increment indicates a new customer starts or ends his/her service or arrives at the office. We see that, in Figure 1a, when $s > T$, the final service ending time exceeds $T = 10$, that is the service is not finished even after the planned closing time. With $s = 6$, we see that, in Figure 1b, the service can be finished on time. Moreover, as s decreases, the average waiting time in queue W_q decreases, see Figure 1c.



(a) Single server (b) Varying server with $s = 6$ (c) W_q vs s

Figure 1: The Simulation Result when $K = 2$

3.2 Simulation when $K = 3$

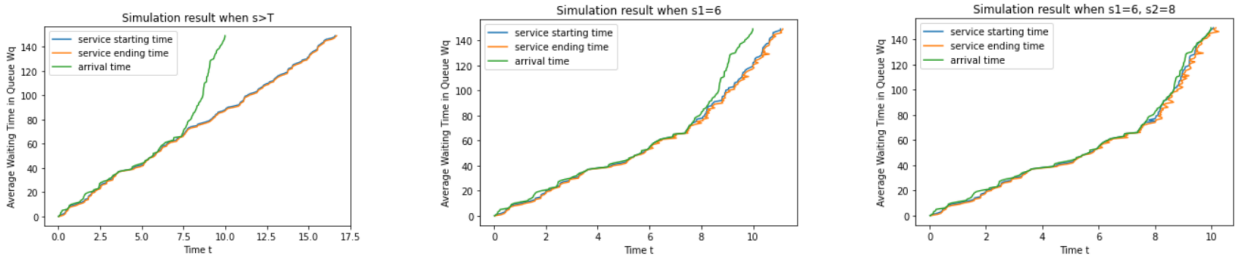
Next, we extend the simulation to consider the case when $K = 3$. We proceed to include a third arrival rate to obtain arrival rates $\lambda_1 = 6$ between $[0, t_0]$, $\lambda_2 = 9.9$ between $[t_0, t_1]$ and $\lambda_3 = 25$ over $[t_0, T]$. The model will be studied when $T = 10$, $t_0 = 4$ and $t_0 = 7$, with similar conditions as in $K = 2$, a constant service rate $\mu = 10$ and the cost to maintain i servers is $c_i = 5i$ per hour for $i = 1, 2, 3$, and the upper bound for the cost increased to $C = 13$ per hour to account for the third counter in the system. With these conditions, the manager now has to decide the optimal time to open two new counters, to give rise to a total of three

operating counters, with changes in customer influx rate at 12:00 p.m. and 3:00 p.m. We first simulate the system when the first additional counter will be opened at 2:00 p.m. and the second additional counter will be opened at 4:00 p.m., that is, $s_1 = 6$ and $s_2 = 8$ respectively.

The simulation carried out reflects the process that of when $K = 2$. We continue to run our model for 10,000 iterations before computing the average of all W_qs to obtain the estimates.

With up to three counters operating in this system, we first consider the case where there is no additional counter opened during $[0, T]$, that is, $s_1 > T$. Following this, we consider the case where the first additional counter is opened at $s_1 = 6$ and $s_2 > T$, like that of in $K = 2$, before finally opening the second additional counter at $s_2 = 8$. We observed that W_q consistently declines with every additional counter opened, from $W_q = 1.679$ in a system with only one counter, to $W_q = 0.144$ for $K = 2$, and to $W_q = 0.087$ for a system with three counters in $K = 3$. Similarly, we illustrate the service starting time, service ending time, and arrival time over $[0, T]$. From Figure 2a, the final service ending time far exceeds the planned closing time of $T = 10$, implying that a single server is under-equipped to accommodate the additional influx in arrivals. With only one new counter introduced at $s_1 = 6$, it can be seen from Figure 2b that although there is an improvement, the final service ending time still exceeds $T = 10$. Lastly, the introduction of a second server at $s_2 = 8$ as seen in Figure 2c shows that service can be completed on time as service starting time is brought earlier after the introduction of the third counter in the system at $s_2 = 8$.

By varying the time at which the two new servers are introduced in addition to the initial counter opened, we plot s_1 and s_2 against the resulting W_q to obtain Figure 3. As the results show, the earlier both counters are opened, the lower the average waiting time in the queue, W_q.



(a) Single server (b) New server at $s_1 = 6$, $s_2 > T$ (c) New server at $s_1 = 6$ and $s_2 = 8$

Figure 2: The Simulation Result when $K = 3$

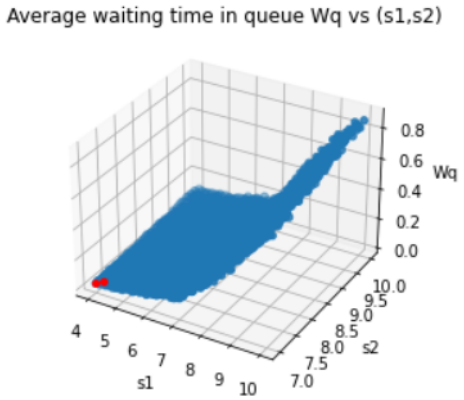


Figure 3: W_q against varying s_1 and s_2

4 Analysis

4.1 Sensitivity Analysis

In order to study how the model behaves under different conditions, we proceed to conduct sensitivity analysis based on the $K = 2$ model against a single server model. We first investigate a change in service rate, for $1 \leq \mu < 15$, and its effects on the average waiting time in queue, W_q, holding all other conditions constant. Figure 4 reflects diminishing marginal returns as service rate increases, with W_q tapering as service rate increases beyond $\mu = 4$ for the $K = 2$ system. On the other hand, for a system with only one counter throughout $[0, T]$, W_q continues to decrease until tapering when the service rate increases beyond $\mu = 11$. These results allow the manager to evaluate whether if it is more worthwhile to increase the service rates of the servers with respect to the number of servers in the system.

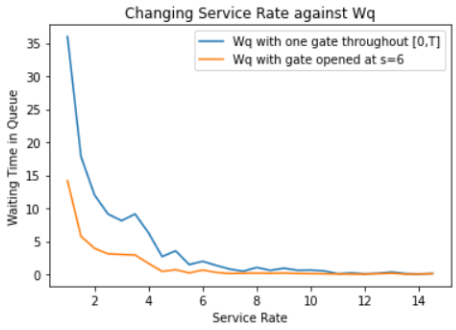


Figure 4: Varying μ against W_q

Next, we proceed to investigate a change in both arrival rates and its effects on W_q in the $K = 2$ system, for $1 \leq \lambda_1 < 6$ and for $1 \leq \lambda_2 < 15$, ceteris paribus. We vary both arrival rates in order to better reflect the behavior of such a queuing system in the real world where arrival rates may fluctuate. As can be seen in Figure 5, the result of varying λ_1 and λ_2 is intuitive as W_q increases as both λ_1 and λ_2 increase, with W_q increasing significantly for larger rates of λ_2. A possible application of this analysis is to hence allow the manager to estimate the potential waiting time in the queue to determine whether or not to introduce more servers into the system, given the range of possible arrival rates.

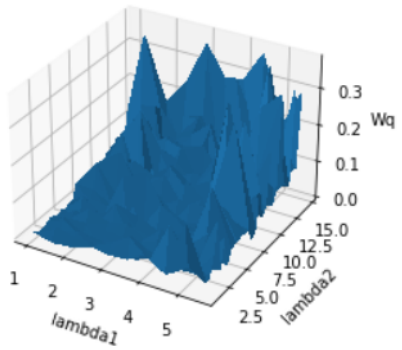


Figure 5: Varying λ_1 and λ_2 against W_q

4.2 Optimization

Though many papers are investigating the behavior of the proposed system, the average waiting time in queue W_q in a general Decision-dependent Non-stationary Queuing system remains analytically intractable. However, intuitively, the answer to the proposed optimization problem (P) is clear: as the time to open a

new server s approaches the boundary of feasibility region, i.e, t_0 or $T(c_2 - C)/(c_2 - c_1)$, the average waiting time in queue W_q will decrease, which is also illustrated in the simulation, see Figure 1c. Therefore, $s_{\text{opt}} = t_0$ or $T(c_2 - C)/(c_2 - c_1)$.

In the following parts, we would like to derive the approximated optimizer for the **point-wise stationary approximation (PSAs)** of the queuing system when $K = 2$. Point-wise stationary approximation, see details in [Green and Kolesar, 1991], is a strategy widely used to approximate the non-stationary queuing system by analyzing the system's performance conditioning on the timepoint t, then integrating the result over the period $[0, T]$. That being said, for a general non-stationary queuing system $M(t)/M/s(t)$ with arrival rates $\lambda(t)$ and the number of servers $s(t)$, let $W_q(t)$ denote the expected waiting time for a stationary $M/M/s$ with an arrival rate $\lambda(t)$, number of servers $s(t)$, and service rate μ as given, assume

$$\sup_t \frac{\lambda(t)}{s(t)\mu} < 1, \tag{6}$$

then the approximated waiting time in queue is given by

$$W_q^{\infty} = \frac{1}{\bar{\lambda} T} \int_0^T \lambda(t) W_q(\lambda(t)) dt, \tag{7}$$

where

$$\bar{\lambda} = \frac{1}{T} \int_0^T \lambda(t) dt. \tag{8}$$

In our model when $K = 2$, we have

$$\lambda(t) = \begin{cases} \lambda_1, & 0 \le t \le t_0 \\ \lambda_2, & t_0 < t \le T, \end{cases} \tag{9}$$

and

$$s(t) = \begin{cases} 1, & 0 \le t \le s \\ 2, & s < t \le T. \end{cases} \tag{10}$$

Therefore,

$$W_q^{\infty} = \frac{1}{\bar{\lambda} T} \Big(\lambda_1 W_q(\lambda_1, \mu, 1) t_0 + \lambda_2 W_q(\lambda_2, \mu, 1)(s - t_0) + \lambda_2 W_q(\lambda_2, \mu, 2)(T - s) \Big), \tag{11}$$

where $W_q(\lambda, \mu, s)$ denote the corresponding expected waiting time in the queue in the stationary queuing system with arrival rate λ, service rate μ, and s servers. Then the optimization problem (P^{∞}) with the approximated objective W_q^{∞} has the optimizer uniquely defined by $s_{\text{opt}}^{\infty} = \max\{t_0, T(c_2 - C)/(c_2 - c_1)\}$, matching our intuition for the original problem (P).

However, note that the PSA is only an upper bound for the original queuing system in the sense that $W_q < W_q^{\infty}$. The approximation is "accurate" if $\lambda(t)$ and μ go to infinity, as suggested in [Green et al., 1991].

5 Conclusion and Future Work

In summary, we empirically evaluated a non-stationary arrival rate $M(t)/M/s$ queuing system through simulation for $K = 2, 3$ by varying different λ_i and s. We also used PSAs to obtained an estimate for s_{opt}^{∞} which verified our intuition. However, our models have some limitations that we may seek for improvements in the future: firstly, the simulation is only carried out for the case when $K = 2, 3$, a more general simulation can be studied to accommodate the case for arbitrary K; moreover, PSAs only investigates approximated optimization problem, more research can be done to derive an analytical expression for W_q in the model. You may access our code for the simulation results in GitHub (code).

References

[Green and Kolesar, 1991] Green, L. and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Manage. Sci.*, 37(1):84–97.

[Green et al., 1991] Green, L., Kolesar, P., and Svoronos, A. (1991). Some effects of nonstationarity on multiserver markovian queueing systems. *Oper. Res.*, 39(3):502–511.

[Wang and Tai, 2000] Wang, K.-H. and Tai, K.-Y. (2000). A queueing system with queue-dependent servers and finite capacity. *Applied Mathematical Modelling*, 24(11):807–814.