FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 6, June, 2025, pp 222 – 227
DOI: https://doi.org/10.33003/fjs-2025-0906-3622

HEAVY TRAFFIC APPROXIMATION METHODS FOR OPTIMIZING TANDEM MULTISERVER QUEUING SYSTEMS IN ANTENATAL CLINICS

***[1]Okechukwu Ifeoma Chizoba, [2]Oruh Ben Ifeanyichukwu, [3]Omekara Chukwuemeka O. and [3]Enogwe Samuel Ugochukwu**

[1]Department of Mathematics/Statistics, Caritas University Amorji-Nike, Enugu, Enugu State.
[2]Department of Mathematics, Michael Okpara University of Agriculture, Umudike, Abia State.
[3]Department of Statistics, Michael Okpara University of Agriculture, Umudike, Abia State.

*Corresponding authors' email: ifystat4real@yahoo.com

ABSTRACT

There have been issues of long waiting hours just to get attended to in real-world systems, which is unacceptable and inappropriate. This prolonged waiting time is a challenge in the health sector, in particular the antenatal clinic. However, since the prolonged waiting time causes the system to approach full utilisation (traffic intensity close to one, $\rho \rightarrow 1$), heavy traffic approximation methods were employed. This study extended and integrated heavy traffic approximation to the healthcare sector to optimise the performance and behaviour of the tandem queueing system of the antenatal clinic for scheduled customers (without the idea of reneging or balking). The heavy traffic approximation methods were applied to the tandem queueing system with multiserver stations in which queue discipline is first come, first served, infinite queue capacity and customers' interarrival time and service time followed general distributions rather than exponential distributions. Real-world data that comprised arrival, service and waiting times were collected and analysed, and the results showed that the variabilities in the arrival and service processes significantly impacted the system performance as an increase (or decrease) in these processes brought about an increase (or decrease) in the expected waiting time. Also, high traffic intensity increased the expected waiting time and vice versa; and there is a significant decrease in the expected waiting time when the number of servers is increased by one. On average, the Whitt approximation method outperformed the Kingman approximation method in estimating the expected waiting time.

Keywords: Heavy traffic approximation methods, Tandem queueing system, Multiserver station, Expected waiting time, Traffic intensity

INTRODUCTION

A tandem queuing system is a type of queuing model in which customers or entities sequentially pass through multiple service stages, or stations, arranged in series. In this system, a customer must complete service at one stage before proceeding to the next. Each stage may have one or more servers operating simultaneously to accommodate incoming entities (Chen and Yao, 2001; Whitt, 2002). An extension of the basic tandem queuing model is the tandem multi-server queuing system, where multiple servers are deployed at each service stage. In these systems, each stage has several servers that operate in parallel to process arriving customers or entities, facilitating the simultaneous handling of multiple individuals. This model is particularly relevant in scenarios that require managing a high volume of customers while reducing waiting times through shared resources. In a typical tandem multi-server system, customers enter sequentially, passing through various stages where multiple servers are available (e.g., receptionists, medical staff, or diagnostic equipment). Customers queue at each stage and are served by the first available server, after which they advance to the next stage in the sequence. This structure introduces a level of parallelism at each service stage, which can significantly enhance overall throughput and decrease individual waiting times, assuming the system is well-balanced and effectively managed. In healthcare settings, especially in antenatal clinics, tandem queuing systems naturally arise due to the sequential nature of care delivery. However, any delays or inefficiencies at one stage can propagate downstream, leading to longer overall waiting times and diminished system efficiency. Effective management of such systems under heavy traffic conditions—marked by high arrival rates near the system's capacity—is crucial for optimizing patient flow and resource utilization (Green *et.al*, 2006). Heavy traffic approximation serves as an essential analytical tool for studying queuing systems, particularly when operations occur under high traffic conditions. In tandem multi-server queuing systems, this situation arises when the arrival rate of customers (e.g., patients in an antenatal clinic) approaches the system's capacity. Under these circumstances, the system's performance tends to deteriorate, as demand for services exceeds the servers' ability to provide them, resulting in bottlenecks at various service stages. Consequently, patients experience longer waits times, servers may become overburdened, and overall patient satisfaction declines. When patient arrival rates are at or exceed the system's service capacity, these bottlenecks become increasingly pronounced. Poorly managed queues in antenatal clinics can lead to prolonged waiting times, which may discourage patients from accessing essential services and increase maternal and neonatal risks (WHO, 2016). Additionally, such inefficiencies may lead to resource mismanagement, where healthcare workers are either overburdened during peak periods or underutilized during off-peak times (Azraii *et al*., 2017), and cause patient dissatisfaction, as extended waits and congested systems diminish trust in healthcare services and hinder compliance with antenatal care recommendations. One potential strategy for enhancing system performance under these conditions is implementing heavy traffic approximation models. These models enable researchers and practitioners to evaluate and optimize tandem multi-server systems by predicting and elucidating system behavior as traffic levels rise. Utilizing simulation, approximation methods, or analytical models allows for the assessment of system

performance, identification of inefficiencies, and the development of strategies to improve throughput while minimizing delays and costs. The literature has extensively examined these models. For instance, Whitt (1983; 2002) and Kingman (1964) provide a foundational framework for understanding queueing systems under heavy traffic conditions, while Green *et.al* (2006) discuss applications of these models to optimize service systems within healthcare contexts. By applying these principles to tandem multi-server queueing systems in antenatal clinics, it becomes feasible to develop solutions that enhance overall service delivery and patient experiences, even amid high patient loads. This work is aimed at identifying the bottlenecks, optimizing the performance and behaviour of a tandem queueing system in the antenatal clinic for scheduled customers (without the idea of reneging or balking); It holds significant potential for improving the efficiency of antenatal clinics by addressing the queueing challenges in tandem multiserver systems.

Literature Review

Whitt (1983) introduced the Queueing Network Analyzer (QNA), a software tool designed for evaluating congestion in networks of queues under flexible conditions. The QNA uses moment-based approximations to analyze individual queue nodes as standard GI/G/m models, providing essential insights into system performance under heavy traffic.

Medhi (2003) highlighted the challenges of deriving mean waiting times in more complex queueing systems, noting the availability of exact solutions for simpler models like M/M/c, but a lack of closed-form expressions for G/G/c systems. This has led to the development of various approximation techniques.

Whitt (1993) further advanced the analysis of GI/G/m queues with his approximations that rely on the first two statistical moments of service and interarrival times. His work emphasized the importance of understanding waiting times in steady state, which helps in integrating GI/G/m models into larger queueing networks.

Green (2006) noted that although exact formulas for non-Markovian multi-server queues do not exist, several effective and straightforward approximations that can be utilized. These approximations illustrate that as the coefficient of variation of service time increases, the average delay also increases; and to compute the coefficient of variation, it is necessary to obtain both the mean and the standard deviation of the interarrival and service times.

In their article, Wu and McGinnis (2013) introduced an approximation method grounded in the observed characteristics of tandem queue behaviour, specifically focusing on the concepts of intrinsic gap and intrinsic ratio. They capitalized on the nearly linear and heavy-traffic properties of the intrinsic ratio, which tend to be observed in practical production scenarios. The proposed approach outperformed existing approximation methods over a wide variety of cases and demonstrated significant potential for producing accurate estimates of mean queue times in real-world production environments when applied to historical data.

Moon and Shin (2019) focused on tandem queues, proposing an approximate analysis method that accounts for service times modelled with phase-type distributions and utilizing a decomposition approach. Their results suggest this method is effective for real-world applications.

Mala and Varma (2016) applied basic queueing theory to a single-server local healthcare clinic, assessing various performance metrics with the goal of minimizing patient wait times and optimizing clinic efficiency.

Elsewhere, Emenonye *et al*. (2022) examined queue models in relation to teletraffic

System in which there was a formulation of a tele-traffic problem and an established solution through the model. They thus, concluded that the model showed improved services leading to increased customers' satisfaction and also minimized cost

Moreso, Ailobhio *et al*. (2020) utilized queueing models to improve service rates for expectant mothers at a hospital, successfully reducing their waiting times and analyzing performance measures associated with the queueing system. This approach illustrates the practical implications of queueing theory in enhancing healthcare service delivery.

MATERIALS AND METHODS

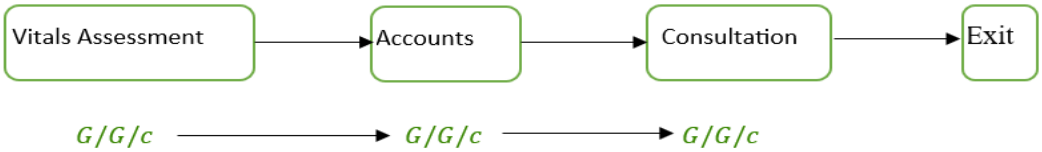
System Description

The antenatal clinic operates as a tandem multiserver system**.** We therefore consider a tandem queue system of n arrivals with k service stations. The customers from outside (at the arrival rate λ) arrive at the first station (i), after receiving service at the station $i, i < k$, moves to station $i + 1$ and after receiving service at the station $i + 1$ continues until they reach the station k and exit the system after receiving service. Thus, station k is the final stage in the processing of the customer.

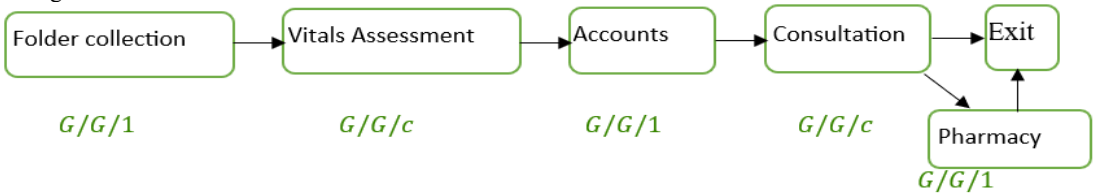
The buffer capacity of the stations is assumed to be infinite and the service time follows general distributions with parameter $\mu_i; i = 1, i + 1, \ldots k$. The service discipline at all stations is First Come First Served (FCFS); a job waiting or being processed at the station i incurs costs with the rate h_i, and we assume that h_i is non-decreasing in i.

The tandem queueing system of the antenatal clinic is modelled as a series of three sequential stations (stages) comprising: Vitals assessment, Accounts and Consultation.

Thus, the modified antenatal clinic's tandem queueing system is represented below:



The existing model in use for the antenatal clinic is



The $G/G/c$ Tandem Queueing model

Data for this study were collected from the ante-natal clinic of a state teaching hospital for four weeks time period; the collected data were on arrival time, service time, waiting time and number of servers at each station.

The assumptions for the $G/G/c$ tandem queueing system include:

i. General interarrival and service times
ii. Multiple service stations in series with Single or multiple servers at each station
iii. Queue discipline is FCFS
iv. No feedback, no balking, no reneging and no jockeying
v. Independent interarrival and service time
vi. The calling population is infinite
vii. Infinite queue capacity
viii. Stability condition (steady state)
ix. System operates under heavy traffic (utilization close to 1).
x. Servers are identical at each stage.

The basic notations include:

λ – the arrival rate
μ – the service rate/server
c – number of servers
$\rho = \frac{\lambda}{c\mu}$ – traffic intensity (system utilization)
$\frac{1}{\mu}$ – mean service time
$\frac{1}{\lambda}$ – interarrival time
c_a^2- squared coefficient of variation of interarrival time
c_s^2- squared coefficient of variation of service time

In situations where queueing systems are complex, like $G/G/c$ queueing systems, exact analytical solutions are generally (typically) intractable (difficult to derive). In such situations, performance measures are determined by the use of bounds and approximations methods. Approximations are methods used to estimate performance measures in many complex queueing systems where exact analytical solutions are intractable and the arrival process is assumed to be a renewal process. These methods like heavy traffic or diffusion approximations are important as they provide reasonable predictions for system behaviour. The approximations are either derived by interpolation or asymptotic analysis.

Heavy-traffic Approximation

A queueing system with traffic intensity ρ close to unity (that is, system approaches full utilization) is called a heavy-traffic queueing system. Kingman (1961) was the first to investigate the behaviour of a queueing system $G/G/1$ in the heavy traffic case, which was sort of central limit theorem for heavy traffic and the theorem stated that under heavy traffic, the steady-state waiting time distribution in a queue can be approximated by an exponential distribution.

According to Medhi (2003), the Kingman's approximation for waiting time distribution after inverting the Laplace transform (LST) is

$$W(t) \approx 1 - \exp\left\{-\frac{2(1-\rho)}{\lambda(\sigma_a^2+\sigma_s^2)}t\right\} \quad (1)$$

which gives the distribution function of the waiting time to an approximation; and the distribution is exponential with the mean given as

$$E(W) \approx \frac{\lambda(\sigma_a^2+\sigma_s^2)}{2(1-\rho)} \quad (2)$$

The result for E(W) for large $\rho(<1)$ according to Kingman (1964) can also be expressed as

$$EW(G/G/1) \approx \frac{(c_a^2+c_s^2)}{2} EW(M/M/1) \quad (3)$$

where

$$EW(M/M/1) \approx \frac{\rho}{\mu(1-\rho)} \quad (4)$$

and c_a^2, c_s^2 is the squared of the coefficient of variation of interarrival, service time.

Whitt (1983), provided an approximation for the mean waiting time in a $G/G/1$ queueing model given as:

$$EW(G/G/1) \approx \frac{\rho(c_a^2+c_s^2)g}{2\mu(1-\rho)} \quad (5)$$

where $g \equiv g(\rho, c_a^2, c_s^2)$ is defined as

$$g \equiv g(\rho, c_a^2, c_s^2) = \begin{cases} exp\left\{-\frac{2(1-\rho)}{3\rho}\frac{(1-c_a^2)}{c_a^2+c_s^2}\right\}; & c_a^2 \le 1 \\ 1 \quad ; & c_a^2 > 1 \end{cases} \quad (6)$$

Kingman (1964) made a conjecture from the result of $G/M/c$ system that, for heavy traffic, the waiting time for a $G/G/c$ queueing model should be exponentially distributed with mean;

$$EW(G/G/c) \approx \frac{\sigma_a^2+\frac{\sigma_s^2}{c^2}}{2(\frac{1-\rho}{\lambda})} \quad (7)$$

Köllerström (1974) proved Kingman's conjecture and the approximate distribution of waiting time in $G/G/c$ queueing model in heavy traffic is exponential and is given by $W(t) \approx$

$$1 - \exp\left\{-\frac{2\frac{(1-\rho)}{\lambda}}{\sigma_a^2+\frac{\sigma_s^2}{c^2}}t\right\} \quad (8)$$

According to Whitt (1983), in heavy–traffic as $\rho \to 1$, the approximation for the mean waiting time is

$$EW(\rho, c_a^2, c_s^2, c) \approx \frac{1}{c\mu(1-\rho)} \frac{(c_a^2+c_s^2)}{2} \quad (9)$$

and by virtue of heavy-traffic limit theorems, we know that (9) is also asymptotically correct for $G/G/c$ systems as $\rho \to 1$ for fixed c

RESULTS AND DISCUSSION

We considered a three nodes (stations) tandem queueing system, an antenatal clinic, with a general arrival and service time distribution. The arrival and the service processes of the $G/G/c$ model for the three stations in the antenatal clinic were computed and used to evaluate the Kingman and Whitt's approximation methods. The minimum number of servers at the three stations (nodes) are: three, two and five servers.

Approximations of Expected Waiting Time for a G/G/c Queueing Model in Tandem Queueing System

Table 1: A Comparison of Approximations of Expected Waiting Time for a G/G/c Queueing Model in Tandem Queueing System of Data Set 1

Nodes	Arrival / Service Process		Traffic Intensity, ρ	Methods	
				Kgm	Whitt
1	c_a^2	2.043	0.85	237.37	71.41
	σ_a^2	23.785			
	c_s^2	22.380	0.425	37.75	27.94
	σ_s^2	1770.113			
2	c_a^2	0.290	0.96	2616.68	3.53
	σ_a^2	12.902			
	c_s^2	0.018	0.479	92.86	0.27
	σ_s^2	1552.490			

3	c_a^2	4.076	0.97	253.52	49.36
	σ_a^2	51.422			
	c_s^2	0.248	0.483	14.50	7.16
	σ_s^2	62.967			

From the Table 1 above; it is observed that:

i. The traffic intensities are high in the three stations but very high in the second and the third stations, the system very close to overload (unstable). This may be due to high variability of the service processes in the stations.
ii. High traffic intensity and high variability in the service process has a negative impact on the expected waiting time (very high waiting time) especially with Kingman's approximation method which causes bottlenecks at the stations.
iii. The arrival and the service processes having different general distributions affects the expected waiting time – the two approximation methods, both have high waiting time.
iv. Whitt approximation method significantly outperformed the Kingman's approximation method in the approximation of the expected waiting time.
v. There is a significant decrease in the expected waiting time when there is an increase in number of servers by a unit.

Table 2: A Comparison of Approximations of Expected Waiting Time for a G/G/c Queueing Model in Tandem Queueing System of Data Set 2

Nodes	**Arrival / Service Process**		**Traffic Intensity, ρ**	**Methods**	
				Kgm	**Whitt**
1	c_a^2	1.782	0.92	99.38	481.06
	σ_a^2	49.693			
	c_s^2	76.496	0.46	14.57	106.94
	σ_s^2	11.433			
2	c_a^2	0.096	0.72	0.45	0.07
	σ_a^2	0.395			
	c_s^2	0.002	0.361	0.20	0.04
	σ_s^2	0.024			
3	c_a^2	3.485	0.70	51.37	5.56
	σ_a^2	152.347			
	c_s^2	1.203	0.352	23.63	6.44
	σ_s^2	83.074			

From Table 2, it is observed that:

i. There are high variabilities in the arrival and the service processes at stations one and three which explained the reason for high expected waiting time though the effect is more evident with Whitt approximation method.
ii. The traffic intensity is higher in station one ($\rho = 0.92$) as compared to station two and three. Thus, high traffic intensity brings about high expected waiting time
iii. In station two, the traffic intensity ($\rho = 0.72$) together with the arrival and service processes (all less than one) brought about minimized expected waiting time less than one for both approximation methods.
iv. A unit increase in the number of servers reduced the expected waiting time drastically except in station three where there is rather an increase in waiting time for Whitt approximation method.

Table 3: A Comparison of Approximations of Expected Waiting Time for a G/G/c Queueing Model in Tandem Queueing System of Data Set 3.

Nodes	**Arrival / Service Process**		**Traffic Intensity, ρ**	**Methods**	
				Kgm	**Whitt**
1	c_a^2	2.103	0.90	30.58	17.77
	σ_a^2	21.659			
	c_s^2	1.107	0.447	5.41	4.82
	σ_s^2	9.671			
2	c_a^2	0.443	0.72	0.18	0.70
	σ_a^2	0.053			
	c_s^2	0.059	0.358	0.05	0.31
	σ_s^2	0.682			
3	c_a^2	4.928	0.59	23.18	5.60
	σ_a^2	125.573			
	c_s^2	1.000	0.293	13.42	8.12
	σ_s^2	7.000			

It is observed from Table 3, that,

i. In station two, the arrival and service processes are less than one which actually brought about minimized expected waiting time that are very small (< 1) with the two approximation methods.
ii. The traffic intensity is highest in station one and together with high variability in the arrival process

caused very high expected waiting time amongst the three stations (higher with Kingman approximation method).

iii. An increase in the number of servers by a unit reduced the expected waiting time drastically except in station three where there is rather an increase in the expected waiting time for Whitt approximation method.

iv. On the average, Whitt approximation method outperformed Kingman approximation method.

In antenatal clinics, there is high traffic intensity ($\rho \to 1$) in each which possibly resulted from the arrival and service time distribution (general distribution). This high traffic intensity and arrival and service processes (their high variability) significantly affect the expected waiting time (high expected waiting time) with the approximation methods, though the effect is higher with the Kingman approximation method. Now, increasing the number of servers by a unit across the three stations significantly reduced not just the traffic intensity but also the expected waiting time; thus, the customers' (expectant mothers') are satisfied for the service rendered to them. Also, on the average, Whitt's approximation method is preferred over the Kingman method.

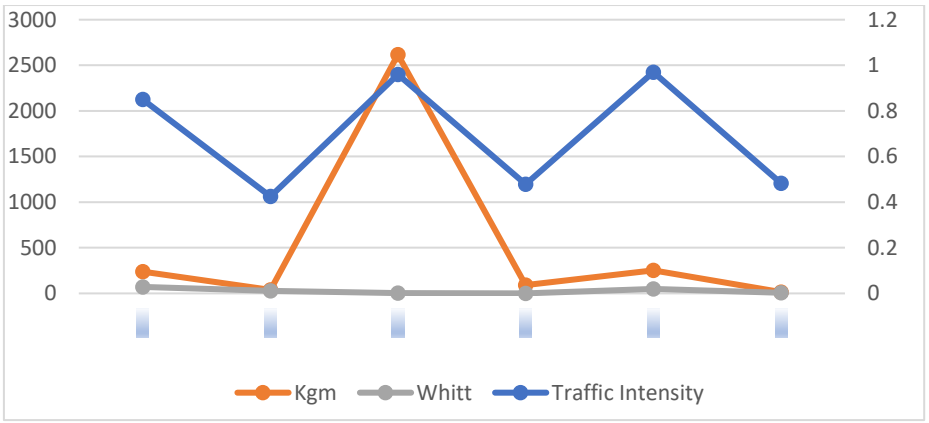


Figure 1: Approximation Methods and Traffic Intensity of a unit Increase of Servers Across the Three Stations for Data Set 1

Figure 2: Approximation Methods and Traffic Intensity of a unit Increase of Servers Across the Three Stations for Data Set 2

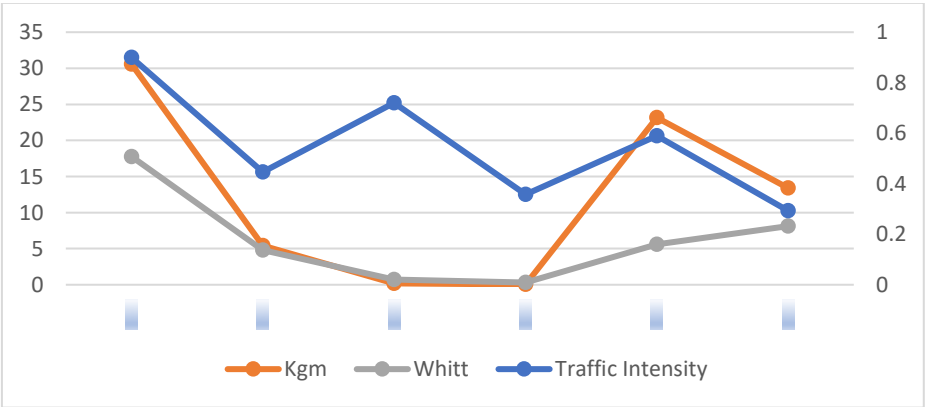


Figure 3: Approximation Methods and Traffic Intensity of a unit Increase of Servers Across the Three Stations for Data Set 3

The plotted graphs for the data sets clearly show that a unit increase in the number of servers across the three stations significantly minimized the expected waiting time and the traffic intensity, with the Whitt approximation method outperforming Kingman approximation method on the average.

CONCLUSION

This work dealt with the extension and integration of heavy traffic approximation methods to the health care sector, antenatal care in particular. It provides precise performance metrics predictions with varying arrival and service processes. The traffic intensity (ρ)and the average waiting time were the performance metrics analysed under steady-state.

There are bottlenecks at some stations in the tandem system where the traffic intensities are very close to unity. The results revealed the impact of variability in the interarrival and service times on system performance and high traffic intensity (ρ) increases the mean waiting time in the system. But a unit increase in the number of servers caused a significant decrease in the expected waiting time.

Furthermore, there is under-utilisation of available servers which resulted in inefficiency of the system, causing bottlenecks at the stations and consequently customers' dis-satisfaction.

Therefore, with this study, the results obtained could be of help to the hospital management and policy makers to estimate delays, minimize patient waiting time without compromise of care quality, optimise staffing (using faster servers where necessary) and allocate resources more effectively.

Altogether, this work provides empirical evidence on the efficacy of heavy-traffic approximation in complex multi-server systems.

Finally, the system should always have enough servers that are capable in terms of capacity and processing speed to ensure the stability of the system ($\rho < 1$) and drastically reduce the mean waiting time, improving the overall system efficiency and effectiveness. Also, the effectiveness of the heavy traffic approximation method in healthcare (tandem in nature) will encourage its adoption in other similar queueing systems.

REFERENCES

Ailobhio, D.T., Owolabi, T. I. and Ayoo, P.V. (2020). Application of Queuing Theory in Antenatal Clinics. *IOSR Journal of Mathematics (IOSR-JM),* 12(6), 42-47.

Azraii, A.B., Kamaruddin, K.N. and Ariffin, F. (2017). An assessment of patient waiting and consultation time in a primary healthcare clinic. Malaysian Family Physician, 12(1), 14–21.

Chen, H. and Yao, D. (2001), *Fundamentals of Queueing Networks.* Springer Series in Operations Research. New York. https://doi.org/10.1007/978-1-4757-5301-1

Emenonye, E. C., Nwakego, S. O. and Ehiwario, J. C. (2022). Queueing Theory in Solving Tele-Traffic Problem. FUDMA Journal of Sciences (FJS), 6(4), 191-194. https://doi.org/10.33003/f js-2022-0604-1063

Green, L. and Savin S. (2006). Providing Timely Access to Medical Care: A Queueing Model. Graduate School of Business, Columbia University. *www.researchgate.net*

Green, L., Savin, S. and Wang, B. (2006). Managing Patients Service in a Diagnostic Medical Facility. INFORMS, 54(1), 11-25. https://doi.org/10.1287/opre.1060.0242

Kingman, J.F.C. (1961) The Single Server Queue in Heavy Traffic. *Mathematical Proceedings of the Cambridge Philosophical Society*, 57, 902-904. http://dx.doi.org/10.1017/S0305004100036094

Kingman, J. F.C. 1964. The Heavy Traffic Approximation in the Theory of Queues. *Proceedings of the Symposium on Congestion Theory*.

Kollerström, J. (1974). Heavy Traffic Theory for Queues with Several Servers 1. J.Appl. Prob. 11, 544-552.

Mala and Varma, S.P.(2016). Waiting Time Reduction in a Local Health Care Centre Using Queueing Theory. *IOSR Journal of Mathematics (IOSR-JM), 12(1), 95-100.* https://doi.org/10.9790/5728-121495100

Medhi, J. (2003). *Stochastic Models in Queueing theory*. 2nd Edition, Academic Press, California, USA.

Moon, D. and Shin, Y. (2019). Approximation of Tandem Queues with Blocking. Proceedings of the 8th International Conference on Operations Research and Enterprise Systems (ICORES), 422-428. https://doi.org/10.5220/0007469504220428.

Whitt, W. (1983). The Queueing Network Analyzer. The Bell System Technical Journal. 62(9), 2779-2815.

Whitt, W. (1993). Approximations for the GI/G/m Queue. Production and Operations Management. 2(2), 114-161

Whitt, W. (2002). *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer Series in Operations Research. New York.

World Health Organization (WHO). (2016), 2016 WHO Annual Care Guidelines. www.mcsprogram.org

Wu, K. and McGinnis, L. (2013), Interpolation Approximations for Queues in Series. *IIE Transactions,* 45, 273–290.