mathematics

MDPI

Article

The Stability Analysis of a Double-X Queuing Network Occurring in the Banking Sector

Hong Zhang [1], Saviour Worlanyo Akuamoah [2], Wilson Osafo Apeanti [3], Prince Harvim [4], David Yaro [5] and Paul Georgescu [6,*]

[1] School of Economics and Management, Changzhou Institute of Technology, Changzhou 213032, China; zhanghong2018@czu.cn
[2] Department of Mathematics and Statistics, Ho Technical University, Ho Volta Region P.O. Box HP 217, Ghana; wakuamoah@htu.edu.gh
[3] Faculty of Science, University of Education, Winneba P.O. Box 25, Ghana; woapeanti@uew.edu.gh
[4] Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON K1N 6N5, Canada; princeharvim@yahoo.com
[5] Department of Mathematics and Statistics, School of Applied Sciences and Technology, Cape Coast Technical University, Cape Coast P.O. Box DL 50, Ghana; ortaega36@yahoo.com
[6] Department of Mathematics, Technical University of Iaşi, Bd. Copou 11A, 700506 Iaşi, Romania
* Correspondence: vpgeo@tuiasi.ro

Abstract: We model a common teller–customer interaction occurring in the Ghanaian banking sector via a Double-X queuing network consisting of three single servers with infinite-capacity buffers. The servers are assumed to face independent general renewal of customers and independent identically distributed general service times, the inter-arrival and service time distributions being different for each server. Servers, when free, help serve customers waiting in the queues of other servers. By using the fluid limit approach, we find a sufficient stability condition for the system, which involves the arrival and service rates in the form of a set of inequalities. Finally, the model is validated using an illustrative example from a Ghanaian bank.

Keywords: Double-X cascade network; Lyapunov function; fluid limit approach; interacting server

Citation: Zhang, H.; Akuamoah, S.W.; Apeanti, W.O.; Harvim, P.; Yaro, D.; Georgescu, P. The Stability Analysis of a Double-X Queuing Network Occurring in the Banking Sector. *Mathematics* **2021**, *9*, 1957. https://doi.org/10.3390/math9161957

Academic Editor: Mark Kelbert

Received: 7 July 2021
Accepted: 11 August 2021
Published: 16 August 2021

1. Introduction

The competition resulting from a decade of deregulation in the Ghanaian banking industry is now becoming more intense, due to the regulatory imperatives of worldwide banking and to the customers becoming increasingly aware of their rights [1,2]. On the one hand, bank customers have become increasingly demanding, requiring high quality, low priced and immediate service delivery [3], a key development being the entry of private banks into the market and the expansion of branches of existing banks. On the other hand, banks have found new methods to differentiate their products and services to attract new customers [4].

Queuing theory aims at developing mathematical and numerical models to investigate the formation and congestion of waiting lines when a service is requested, its basic ideas being proposed by A. K. Erlang. In this regard, there is a high probability for one to queue at least once during everyday activities before a service is rendered [5,6]. Common experience suggests that the act of queuing is associated with waiting, which is an inevitable part of modern life. This leads to the idea that queuing theory has very wide applications [7]. For example, customers waiting to be served at grocery stores, banks and post offices, people having to wait for an operator during a telephone call, people waiting for a taxi or a bus on the way to the workplace can be thought as subjects of queuing theory [8,9].

Even queues that are generally non-problematic may become a nuisance if not managed well. For the banking sector, this requires a swift management of the server–customer relationship since queuing cannot be excluded from the daily operation of a bank [10,11].

Service delivery in banks is personal, customers being either served immediately or joining a queue (waiting line) if servers are busy. A queue occurs if facilities are limited and cannot satisfy the demand during a particular period. However, most customers are not comfortable with waiting or queuing, the time wasted in the queue being perceived as an opportunity cost [12,13]. Performing financial activities can, however, be time consuming and tedious within the limitations of the traditional methods of banking. A recent contribution by [14] found that the banks in Ghana have grown through different eras and now comprise several branches across the breadth of the country, operating in almost every district. The majority of banks in Ghana have been listed on the Ghana stock exchange market since 1996 [15]. Their growth has been synonymous with the growth of their customer base, performance, innovative product and services profitability.

As far as the modeling of the banking sector is concerned, a queuing system consisting of two basic customer classes; two servers were discussed from a mathematical viewpoint in [16,17]. Class-j customers were primarily assigned to station-j, $j = 1, 2$. However, customers could move from station-1 to station-2 but not from station-2 to station-1. Additionally, an arriving class-2 customer would interrupt the current service of a customer who moved from station-1 to station-2, the former being served immediately, while the latter could not come back to station-1, even if Server-1 was free. The service of the moving customers would later be resumed from the point at which it was interrupted. Such a two-station system is called a cascade network [17].

In [18], a bi-objective optimization problem was considered, the first objective being to maximize the total profit and the second one being to minimize the average length of queuing at workstations, and optimal values for the decision variables were determined. According to [19], competitive queuing (CQ) models of working memory (WM) rely upon two underlying mechanisms, namely excitatory parallel planning and inhibitory competitive choice/response suppression. While the former mechanism is responsible for item activation, the latter is responsible for item recall followed by item suppression. Other models in [16,20,21] revealed closely related systems with flexible servers, for which a server might transfer the service capacity to accommodate the workload accumulated in another server. In addition, systems with cross-trained servers in which certain servers can handle a reduced set of customer types, whereas others accept all types, were also discussed in [22,23]. A variety of real-life systems conforming to the above paradigm, including service centers, production systems, computer networks with rescheduling of jobs, and parallel computing systems for which processors have overlapping capabilities, are described in [17]. The same considerations are applicable to manufacturing machines, which have certain primary functions and a number of secondary ones, as described in [24] via the so-called N-models with static priority rules. The particular case of a N-station cascade network in which each station is only able to support the previous one was analyzed in [16] from a stability viewpoint by using a fluid limit approach rather than a more standard Foster type argument. Further details regarding this class of models are given in [16].

However, most of these studies have limited practical success and little subsequent impact, the vast majority of the existing models remaining unvalidated against real waiting time and service time data. This happened since collecting these data were time-consuming, expensive and record error in manual operation (see [6,25,26]).

A bare-bones model of a queuing network consists of two single-server stations with independent general renewal inputs, independently identically distributed (i.i.d.) general service times and infinite capacity buffers. However, most banks in Ghana operate with three servers and with three buffers for customers [27], in which the first server only assists the second server, the second server assists the first and third server if available, and the third server in turn assists the second server only (see Figure 1). In view of these characteristics of Ghanaian banking, so far, there is no existing model to formulate a queuing network of immediate use in the Ghanaian banking sector.

Consequently, the main objective of this paper is to formulate and analyze a so-called Double-X queuing network model, and in this regard we have a prime opportunity to take advantage of the detailed customer arrival time and service time data, automatically collected by the Ghanaian Agricultural Development Bank to monitor compliance with targets. The model is then validated using an illustrative example from the real data, for which we find the arrival and service rate of customers over a given time slot. We then ascertain the appropriateness of bank queue utilization at different levels and also give a fair idea to customers on the amount of time they would be expected to spend in the banking hall.

In this model, each server has an unlimited-capacity buffer for awaiting customers. Customers arrive following an independent general renewal input and have i.i.d. general service times when served at the same station. Inter-arrival times and service times are possibly different for different customer classes. The service discipline is work-conserving and obeys a FIFO (first in first out) scheduling algorithm within each class. Whenever a buffer becomes empty while customers are awaiting service in other buffers, a customer moves to be served there, provided that this is allowed in the given model.

In order to establish the stability of the fluid limit model associated to the queuing network, we start by employing functional laws of large numbers for the renewal processes, as motivated by [28]. In this regard, the choice of an appropriate Lyapunov function is of paramount importance. Although the approaches are related, they do differ in certain technical details. Particularly, we are not prompted to consider a Skorokhod problem for our model [29], which means that we do not address the boundary problems of the state spaces of the job count and workload processes, provided that a service station in a queuing network is empty.

The remaining part of this paper is organized as follows. In Section 2, we introduce and describe our Double-X queuing model, the research assumptions and the equations that govern the processes associated with the network structure. Section 3 looks into the stability analysis of this model. In particular, Section 3.1 introduces the associated fluid limit model, and Section 3.2 is devoted to proving the stability result. Section 4 is concerned with an analysis of real data. Finally, we present the summary of our findings and conclusions in Section 5.

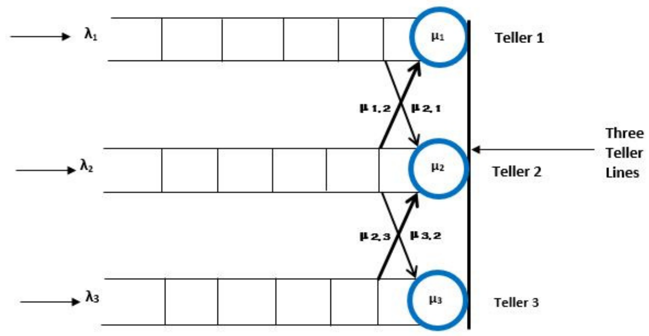

Figure 1. The Double-X cascade network.

2. The Model

2.1. Model Description

Servers

The teller position is a key one in banking, requiring qualified and experienced personnel since tellers are generally the employees that customers first see and interact with. We consider a queuing system consisting of three servers (tellers), which is a quite common occurrence in a Ghanaian bank. The main job tasks and responsibilities of a server towards customers are as follows:

1. Identify customers and process validation and cashing of checks.

2. Process cash and checks depositing and check the accuracy of subsequent deposit slips.
3. Process cash withdrawals.
4. Prepare and issue cashier's checks, traveler's checks and personal money orders, and also exchange foreign currency.
5. Receive and verify payments of utility bills, loans and mortgages.
6. Answer inquiries regarding checking and savings accounts and other banking products.
7. Attempt to resolve issues raised by the customers.
8. Open new accounts.
9. Explain, advise on and promote bank products and services to customers.

Servers typically work from a station, usually located on a server line as shown in Figure 1. In our model, the responsibilities of Server-1 are to serve customers who fall within categories 1–4, and as soon as Server-2 becomes free, customers waiting in Buffer-1 (if any) switch to Buffer-2 to be served there immediately. The responsibilities of Server-2 are to serve customers who fall within category 1–9 and, if Server-1 is free, customers who require services 1–4 switch to Buffer-1 for service, while those requiring services 5–9 move to Buffer-3, if Server-3 is free for service. The responsibilities of Server-3 are to serve customers who fall within categories 5–9; if Server-2 is free, customers waiting in Buffer-3 (if any) switch to start service there immediately. Servers are cross-trained so that they can serve customers from other buffers when they become free.

Customer classes

Class-j, $j = 1,2,3$ are the exogenous customers who arrive at Server-j for service, while Class-(i,j) customers are those moving from Server-i to Server-j for service, $(i,j) = (1,2)$, $(2,1)$, $(2,3)$ and $(3,2)$. In what follows, we use the index j (the double index (i,j)) to denote Class-j (respectively Class-(i,j)) customers.

Service and arrival times

Following the standard definitions on interarrival times and service times given in [28], we let $\xi_j(i), i \geqslant 2$ be the i.i.d. inter-arrival times of i-th Class-j customer entering the network after instant 0, where $j = 1,2,3$ and let $\eta_k(i), i \geqslant 2$ be the i.i.d. service times of the i-th Class-k customer being served after instant 0, $k = 1,2,3,(1,2),(2,1),(2,3)$ and $(3,2)$. All sequences (service and arrival time) are mutually independent. The residual arrival time $\xi_j(1)$ of the first Class-j customer entering the system after instant 0 is independent of $\{\xi_j(i), i \geqslant 2\}, j = 1,2,3$. Additionally, the residual service time $\eta_k(1)$, of a Class-k customer initially being served, if any, is independent of $\{\eta_k(i), i \geqslant 2\}$.

Following the idea from [16], we introduce the arrival rate $\lambda_j = \dfrac{1}{\mathsf{E}\xi_j}$ of Class-j customers, $j = 1,2,3$ and the service rate $\mu_k = \dfrac{1}{\mathsf{E}\eta_k}$ of Class-k customers, $k = 1,2,3,(1,2),(2,1),(2,3)$ and $(3,2)$ (see Figure 1). Additionally, the following standard conditions are imposed:

$$\begin{cases} \mathsf{E}\xi_j < \infty, & j = 1,2,3. \\ \mathsf{E}\eta_k < \infty, & k = 1,2,3,(1,2),(2,1),(2,3),(3,2). \\ \mathsf{P}(\xi_j \geqslant x) > 0, & j = 1,2,3, \text{ for any } x \in [0,\infty). \end{cases} \tag{1}$$

We also assume that inter-arrival times are spread out, that is, for some integers $s_j > 1$ and functions $f_j \geqslant 0$ with $\int_0^\infty f_j(y)dy > 0$ as follows:

$$\mathsf{P}\left(a \le \sum_{i=2}^{s_j} \xi_j(i) \le b\right) \geqslant \int_a^b f_j(y)dy, \quad 0 \le a \le b, j = 1,2,3. \tag{2}$$

2.2. Model Assumptions

The Double-X cascade model is an appropriate one to use for this study under the following assumptions.

(i) Service discipline is assumed to be non-idling, that is, servers cannot be idle if there are waiting customers in their allocated buffers.
(ii) Service discipline is also non-preemptive, which means that a running task cannot be disturbed, that is, customers who are in the process of being served cannot be interrupted.
(iii) Priority is given to customers waiting in their approved buffers over those who move from their rightful buffers to a different buffer.
(iv) Customers who enter the system do not leave unless they are served.

We begin by introducing several notations and primitive processes of the network. The exogenous arrival process is denoted by the following:

$$E(t) \doteq \big(E_1(t), E_2(t), E_3(t)\big),\ t \geq 0,$$

for which

$$E_j(t) \doteq \max\left\{ n \geq 1 : \sum_{i=1}^{n} \xi_j(i) \leq t \right\}$$

is the total number of Class-j arrivals in the interval $[0, t]$, $j = 1, 2, 3$. The served customers process is given by the following:

$$S \doteq \big\{ (S_1(t), S_2(t), S_3(t), S_{1,2}(t), S_{2,1}(t), S_{2,3}(t), S_{3,2}(t)), t \geq 0 \big\},$$

and the renewal process

$$S_k(t) \doteq \max\left\{ n \geq 1 : \sum_{i=1}^{n} \eta_k(i) \leq t \right\}$$

representing the total number of Class-k customers served in $[0, t]$, if the server devotes all time to Class-k customers, $k = 1, 2, 3, (1, 2), (2, 1), (2, 3)$ and $(3, 2)$. By definition, $E_j(0) = S_k(0) = 0$ for all k, j.

We now introduce descriptive processes which measure the performance of the network. Let us denote the following:

$$\begin{cases} D = \big\{ D(t) \doteq (D_1(t), D_2(t), D_3(t), D_{1,2}(t), D_{2,1}(t), D_{2,3}(t), D_{3,2}(t)), t \geqslant 0 \big\}, \\ Z = \big\{ Z(t) \doteq (Z_1(t), Z_2(t), Z_3(t), Z_{1,2}(t), Z_{2,1}(t), Z_{2,3}(t), Z_{3,2}(t)), t \geqslant 0 \big\}, \\ T = \big\{ T(t) \doteq (T_1(t), T_2(t), T_3(t), T_{1,2}(t), T_{2,1}(t), T_{2,3}(t), T_{3,2}(t)), t \geqslant 0 \big\}, \\ Q = \{ Q(t) \doteq (Q_1(t), Q_2(t), Q_3(t)), t \geqslant 0 \}, \\ Y = \{ Y(t) \doteq (Y_1(t), Y_2(t), Y_3(t)), t \geqslant 0 \}, \\ U = \{ U(t) \doteq (U_1(t), U_2(t), U_3(t)), t \geqslant 0 \}, \\ V = \big\{ V(t) \doteq (V_1(t), V_2(t), V_3(t), V_{1,2}(t), V_{2,1}(t), V_{2,3}(t), V_{3,2}(t)), t \geqslant 0 \big\}, \end{cases}$$

their definitions being summarized in Table 1 below.

The processes D, T and Y are nondecreasing and satisfy the initial conditions $D(0) = T(0) = Y(0) = 0$, being assumed that $Z(0)$ and $Q(0)$ are independent. Assume the processes U and V to be right continuous and define $V_k(t) = 0$ if $Z_k(t) = 0$. Additionally, note that $U_j(0) = \xi_j(1)$, while $V_k(0) = \eta_k(1)$ if $Z_k(0) = 1$. We then define the following process:

$$X(t) = \big\{ \big(Q(t), Z(t), U(t), V(t)\big), t \geqslant 0 \big\},$$

describing the dynamics of the network within the state space $\mathbb{Z}_+^3 \times \{0, 1\}^7 \times \mathbb{R}_+^3 \times \mathbb{R}_+^7$. Then the process X is a piecewise-deterministic Markov process.

Table 1. Variables used in the model and their definitions, where $j = 1,2,3$ and $k = 1, 2, 3, (1, 2), (2, 1), (2, 3), (3, 2)$.

Variable	Definition
$D_k(t)$	The number of Class-k customers who depart from the bank in the interval $[0, t]$.
$Z_k(t)$	The number of Class-k customers that are being served at time t ($Z_k(t) \in \{0, 1\}$).
$T_k(t)$	The total service time devoted to Class-k customers in the interval $[0, t]$.
$Y_j(t)$	The idle time of Server-j in the interval $[0, t]$.
$Q_j(t)$	The number of customers in Buffer$-j$ at time t.
$U_j(t)$	The remaining time t before the next exogenous customer arrival to Class-j.
$V_k(t)$	The remaining service time t before the next Class-k customer being served at time t.

2.3. The Queuing Network Equations

The following queuing network equations hold for $t \geqslant 0$:

$$D_k(t) = S_k(T_k(t)), \quad k = 1, 2, 3, (1, 2), (2, 1), (2, 3), (3, 2). \tag{3}$$

Equation (3) counts the total number of service completions for Class-k customers, equal to the total number of Class-k customers departing from the bank at time t.

$$Q_1(t) = Q_1(0) + E_1(t) - (D_1(t) + Z_1(t) + D_{1,2}(t) + Z_{1,2}(t)). \tag{4}$$

Equation (4) indicates how customers are queuing in Buffer-1, waiting for service from Server-1.

$$Q_2(t) = Q_2(0) + E_2(t) - (D_2(t) + Z_2(t) + D_{2,1}(t) + Z_{2,1}(t) + D_{2,3}(t) + Z_{2,3}(t)). \tag{5}$$

Equation (5) shows how customers are queuing in Buffer-2, waiting for service from Server-2.

$$Q_3(t) = Q_3(0) + E_3(t) - (D_3(t) + Z_3(t) + D_{3,2}(t) + Z_{3,2}(t)). \tag{6}$$

Equation (6) also captures how customers are queuing in Buffer-3, waiting for service from Server-3.

$$\begin{aligned}(T_1(t) + T_{2,1}(t)) + Y_1(t) &= (T_2(t) + T_{1,2}(t) + T_{3,2}(t)) + Y_2(t) \\ &= (T_3(t) + T_{2,3}(t)) + Y_3(t) = t.\end{aligned} \tag{7}$$

Equation (7) counts the total service time of Server-j, $j = 1, 2, 3$, including the idle time.

$$\int_0^\infty Q_j(t)dY_j(t) = 0, \quad j = 1, 2, 3. \tag{8}$$

Equation (8) corresponds to a work-conserving scheduler, which always tries to keep the scheduled resources busy.

$$\int_0^\infty Q_2(t)dY_1(t) = 0. \tag{9}$$

Equation (9) shows that Server-1 supports Server-2 and cannot be idle if there are customers who require services 1–4 waiting in Buffer-2.

$$\int_0^\infty (Q_1(t) + Q_3(t))dY_2(t) = 0. \tag{10}$$

Equation (10) means that Server-2 supports Server-1 and Server-3 and cannot be idle if there are customers waiting in Buffer-1 and Buffer-3 for service.

$$\int_0^\infty Q_2(t)dY_3(t) = 0. \tag{11}$$

Equation (11) shows that Server-3 cannot be idle if there are waiting customers in Buffer-2 demanding services 5–9.

$$\int_0^\infty (Q_1(t) + Z_1(t))dT_{2,1}(t) = 0. \tag{12}$$

Equation (12) means that Server-1 gives priority to Class-1 customers over Class-2 customers who move from Buffer-2 to Buffer-1.

$$\int_0^\infty (Q_3(t) + Z_3(t))dT_{2,3}(t) = 0. \tag{13}$$

Equation (13) means that Server-3 gives priority to Class-3 customers over customers who move from Buffer-2 to Buffer-3.

$$\int_0^\infty (Q_2(t) + Z_2(t) + Z_{3,2}(t))dT_{1,2}(t) = 0 \text{ and } \int_0^\infty (Q_2(t) + Z_2(t) + Z_{1,2}(t))dT_{3,2}(t) = 0. \tag{14}$$

Equation (14) means that Server-2 gives priority to Class-2 customers.

We still consider two priority phases for Server-2.

Phase 1 (priority $2 > 1 > 3$): from Equation (15), Server-2 serves customers in their rightful buffer, that is, Buffer-2, then proceeds to serve customers from Buffer-1 before serving customers from Buffer-3:

$$\int_0^\infty (Q_2(t) + Z_2(t) + Q_1(t) + Z_{1,2}(t))dT_{3,2}(t) = 0. \tag{15}$$

Phase 2 (priority $2 > 3 > 1$): also from Equation (16), Server-2 gives priority to Class-2 customers and customers who move from Buffer-3 to Buffer-2 over customers who move from Buffer-1 to Buffer-2 for service:

$$\int_0^\infty (Q_2(t) + Z_2(t) + Q_3(t) + Z_{3,2}(t))dT_{1,2}(t) = 0. \tag{16}$$

3. Results

Let us recall that a queuing network is said to be stable if its corresponding Markov process X is positive Harris recurrent, i.e. it has a unique invariant probability measure [16,30]. For instance, a Markov chain where the chain returns to a particular part of the state space an unbounded number of times is positive Harris recurrent. Harris chains are regenerative processes [16,30].

Lemma 1 ([16]). *For almost all sample paths, any sequence of initial states* $\{x_n\}_{n\geqslant 1} \subset X$ *with* $\lim_{n\to\infty} |x_n| = \infty$, *there exists a sub-sequence* $\{x_{n_r}\}_{r\geqslant 1} \subseteq \{x_n\}_{n\geqslant 1}$ *with* $\lim_{r\to\infty} |x_{n_r}| = \infty$, *such that*

$$\lim_{r\to\infty} \frac{1}{|x_{n_r}|} X^{x_{n_r}}(0) \doteq \overline{X}(0) \tag{17}$$

and the following limit exists uniformly on compact sets:

$$\begin{aligned}&\lim_{r\to\infty} \frac{1}{|x_{n_r}|} (X^{x_{n_r}}(|x_{n_r}|t), D^{x_{n_r}}(|x_{n_r}|t), T^{x_{n_r}}(|x_{n_r}|t), Y^{x_{n_r}}(|x_{n_r}|t)) \\ &\doteq (\overline{X}(t), \overline{D}(t), \overline{T}(t), \overline{Y}(t)),\end{aligned} \tag{18}$$

where

$$\overline{X}(t) = \big(\overline{Q}(t), \overline{Z}(t), \overline{U}(t), \overline{V}(t)\big). \tag{19}$$

Definition 1 ([28,31,32]). *The fluid limit model $(\overline{X}, \overline{D}, \overline{T}, \overline{Y})$ associated to the Double-X model, with $\overline{X} = (\overline{Q}, \overline{Z}, \overline{U}, \overline{V})$, is stable if there exists $t_1 \geqslant 0$ (depending on the arrival and service rates only) such that if $|\overline{Q}(0)| + |\overline{U}(0)| + |\overline{V}(0)| = 1$. Then, the following holds:*

$$\overline{Q}(t) = 0 \quad \forall \quad t \geqslant t_1. \tag{20}$$

3.1. The Fluid Limit Model

The following equations are satisfied for any $t \geq 0$, $\quad k = 1,2,3,(1,2),(2,1),(2,3),(3,2)$ and $j = 1,2,3,(1,2),(2,1),(2,3),(3,2)$. By using the (functional) Strong Law of Large Numbers for the renewal processes, we deduce the following Lemma 4.2 in [28] that the following (u.o.c) limits hold with probability 1 (w.p.1) as $r \to \infty$:

$$\begin{aligned}
\frac{E_j^{x_{n_r}}(|x_{n_r}|t)}{|x_{n_r}|} &\to \lambda_j(t - \overline{U}_j(0))^+, \\
\frac{S_k^{x_{n_r}}(|x_{n_r}|t)}{|x_{n_r}|} &\to \mu_k(t - \overline{V}_k(0))^+, \\
\frac{D_k^{x_{n_r}}(|x_{n_r}|t)}{|x_{n_r}|} &\to \overline{D}_k(t).
\end{aligned}$$

From Equation (3), we obtain the following:

$$\overline{D}_k(t) = \mu_k(\overline{T}_k(t) - \overline{V}_k(0)). \tag{21}$$

Because $Z_k^x(t) \leq 1$ for all x, t and k, where $Z_k(t) \in \{0,1\}$, then as $r \to \infty$, the number of Class-k customers that are being served at time t goes to zero.

$$\overline{Z}_k(t) = \lim_{r\to\infty} \frac{\left(Z_k^{x_{n_r}}(|x_{n_r}|t)\right)}{|x_{n_r}|} = 0. \tag{22}$$

We obtain equalities (21) and (22), which are used to deduce the fluid model Equations (23)–(25) and (31)–(36) from Equations (4)–(6) and (12)–(16) above:

$$\begin{aligned}
\overline{Q}_1(t) &= \overline{Q}_1(0) + \lambda_1(t - \overline{U}_1(0)) - (\overline{D}_1(t) + \overline{D}_{1,2}(t)) \\
&= \overline{Q}_1(0) + \lambda_1(t - \overline{U}_1(0)) - (\mu_1(\overline{T}_1(t) - \overline{V}_1(0)) + \mu_{1,2}(\overline{T}_{1,2}(t) - \overline{V}_{1,2}(0))).
\end{aligned} \tag{23}$$

$$\begin{aligned}
\overline{Q}_2(t) &= \overline{Q}_2(0) + \lambda_2(t - \overline{U}_2(0)) - (\overline{D}_2(t) + \overline{D}_{2,1}(t) + \overline{D}_{2,3}(t)), \\
&= \overline{Q}_2(0) + \lambda_2(t - \overline{U}_2(0)) - \big(\mu_2(\overline{T}_2(t) - \overline{V}_2(0)) + \mu_{2,1}(\overline{T}_{2,1}(t) - \overline{V}_{2,1}(0)) \\
&\quad + \mu_{2,3}(\overline{T}_{2,3}(t) - \overline{V}_{2,3}(0))\big).
\end{aligned} \tag{24}$$

$$\begin{aligned}
\overline{Q}_3(t) &= \overline{Q}_3(0) + \lambda_3(t - \overline{U}_3(0)) - (\overline{D}_3(t) + \overline{D}_{3,2}(t)), \\
&= \overline{Q}_3(0) + \lambda_3(t - \overline{U}_3(0)) - (\mu_3(\overline{T}_3(t) - \overline{V}_3(0)) + \mu_{3,2}(\overline{T}_{3,2}(t) - \overline{V}_{3,2}(0))).
\end{aligned} \tag{25}$$

It follows that Equations (26)–(30) can be obtained from Equations (7)–(11) above, respectively.

$$\begin{aligned}
(\overline{T}_1(t) + \overline{T}_{2,1}(t)) + \overline{Y}_1(t) &= (\overline{T}_2(t) + \overline{T}_{1,2}(t) + \overline{T}_{3,2}(t)) + \overline{Y}_2(t) \\
&= (\overline{T}_3(t) + \overline{T}_{2,3}(t)) + \overline{Y}_3(t) = t.
\end{aligned} \tag{26}$$

$$\int_0^\infty \overline{Q}_j(t)d\overline{Y}_j(t) = 0, \quad j = 1,2,3. \tag{27}$$

$$\int_0^\infty \overline{Q}_2(t)d\overline{Y}_1(t) = 0. \tag{28}$$

$$\int_0^\infty (\overline{Q}_1(t) + \overline{Q}_3(t))d\overline{Y}_2(t) = 0. \tag{29}$$

$$\int_0^\infty \overline{Q}_2(t)d\overline{Y}_3(t) = 0. \tag{30}$$

$$\int_0^\infty \overline{Q}_1(t)d\overline{T}_{2,1}(t) = 0. \tag{31}$$

$$\int_0^\infty \overline{Q}_3(t)d\overline{T}_{2,3}(t) = 0. \tag{32}$$

$$\int_0^\infty \overline{Q}_2(t)d\overline{T}_{1,2}(t) = 0. \tag{33}$$

$$\int_0^\infty \overline{Q}_2(t)d\overline{T}_{3,2}(t) = 0. \tag{34}$$

Phase 1 $\quad (2 > 1 > 3)$:

$$\int_0^\infty (\overline{Q}_2(t) + \overline{Q}_1(t))d\overline{T}_{3,2}(t) = 0. \tag{35}$$

Phase 2 $\quad (2 > 3 > 1)$:

$$\int_0^\infty (\overline{Q}_2(t) + \overline{Q}_3(t))d\overline{T}_{1,2}(t) = 0. \tag{36}$$

The residual life process associated with the delayed renewal process is given by the following:

$$\overline{U}(t) = (\overline{U}(0) - t)^+, \quad \overline{V}(t) = (\overline{V}(0) - t)^+. \tag{37}$$

For any $0 \le s \le t$, the following holds:

$$0 \le \overline{T}_k(t+s) - \overline{T}_k(s) \le t. \tag{38}$$

Without any loss of generality, one may assume that $\overline{U}(0) = \overline{V}(0) = 0$. By (37), this leads to $\overline{U}(t) = \overline{V}(t) = 0$ for all $t > 0$, which can be shortened as $\overline{U} = \overline{V} = 0$.

3.2. The Stability Analysis

In this section, we present a stability result for the Double-X model. We denote $r_{1,2} = \frac{\mu_2}{\mu_{(1,2)}}, \quad r_{2,1} = \frac{\mu_1}{\mu_{(2,1)}}, \quad r_{2,3} = \frac{\mu_3}{\mu_{(2,3)}}$ and $r_{3,2} = \frac{\mu_2}{\mu_{(3,2)}}$, all in the interval $(0, \infty)$.

Let conditions (1) and (2) hold and assume that either one of the assumptions $\mathcal{H}_1$, $\mathcal{H}_2$ takes place (Table 2).

Table 2. Stability conditions.

$\mathcal{H}_1$	$\mathcal{H}_2$
$r_{1,2}, r_{2,1}, r_{2,3}, r_{3,2} \geqslant 1$	$r_{1,2}, r_{2,1}, r_{2,3}, r_{3,2} < 1$
$r_{2,1}r_{1,2},\ r_{1,2}r_{3,2},\ r_{2,3}r_{3,2} \geqslant 1$	$r_{2,1}r_{1,2},\ r_{1,2}r_{3,2},\ r_{2,3}r_{3,2} < 1$
$r_{1,2}r_{3,2}r_{2,3},\ r_{2,1}r_{2,3}r_{3,2},\ r_{1,2}r_{2,1}r_{2,3} \geqslant 1$	$r_{1,2}r_{3,2}r_{2,3},\ r_{1,2}r_{2,1}r_{2,3},\ r_{2,1}r_{2,3}r_{3,2} < 1$

Then, the Double-X model is stable, provided that the following six conditions are satisfied:

$$\begin{cases} (A_1) & \lambda_1 - \mu_1 + \dfrac{\lambda_2 - \mu_2}{r_{1,2}} < 0, \\ (A_2) & \lambda_2 - \mu_2 + \dfrac{\lambda_1 - \mu_1}{r_{2,1}} + \dfrac{\lambda_3 - \mu_3}{r_{2,3}} < 0, \\ (A_3) & \lambda_3 - \mu_3 + \dfrac{\lambda_2 - \mu_2}{r_{3,2}} < 0, \\ (A_4) & \dfrac{\lambda_1 - \mu_1}{r_{2,1}} + \dfrac{\lambda_2 - \mu_2}{r_{1,2} r_{3,2}} + \dfrac{\lambda_3 - \mu_3}{r_{1,2} r_{3,2} r_{2,3}} < 0, \\ (A_5) & \dfrac{\lambda_2 - \mu_2}{r_{1,2} r_{3,2}} + \dfrac{\lambda_3 - \mu_3}{r_{2,3}} + \dfrac{\lambda_1 - \mu_1}{r_{2,3} r_{3,2} r_{2,1}} < 0, \\ (A_6) & \dfrac{\lambda_1 - \mu_1}{r_{2,1}} + \dfrac{\lambda_2 - \mu_2}{r_{2,1} r_{3,2}} + \dfrac{\lambda_3 - \mu_3}{r_{2,3}} < 0. \end{cases} \tag{39}$$

Remark 1. *Assumption $\mathcal{H}_1$ corresponds to the situation in which customers are served slower at a changed buffer, while assumption $\mathcal{H}_2$ reflects just the opposite situation.*

The detailed proof of the above-mentioned result is placed in the Appendix A.

4. Numerical Simulations

The data for the illustrative example were collected on 27 June 2017 (9:00 a.m.–12:00 p.m. (noon)), from the Agricultural Development Bank, Juapong branch, Ghana. The mean arrival rates and the mean service rates are calculated from the collected data, and the subsequent results are used to measure the performance of the system.

Table 3 shows the number of customers served within a given time interval. As shown in Table 3, we recorded 47 exogenous customers, with 6 customers moving from Buffer-1 to Buffer-2 for service, indicating that 41 customers stayed in Buffer-1 for service. The total time for service for customers in Buffer-1 is 177 min.

Table 3. Queue-1 (time in minutes).

$E_1(t)$	$D_1(t)$	$Z_1(t)$	$D_{1,2}(t)$	$Z_{1,2}(t)$	$T_1(t)$	$T_{2,1}(t)$
47	41	41	6	6	152	25

Table 4 shows the mean arrival rate for Buffer-1. According to Table 4, the highest number of customers (16) arrived within the time interval 9:30–10:00 a.m. and the lowest number of customers (4) arrived within the time interval 11:30 a.m.–12:00 p.m. The mean arrival rate for Buffer-1 is $\lambda_1 = \frac{1.57}{6} \approx 0.26$. The service rate of customers served in Buffer-1 is given by $\mu_1 = \frac{1}{E\eta_1}$, where η_1 is the expectation of the service times. $E\eta_1 = \frac{151.71}{41} \approx 3.70$. The service rate for Buffer-1 is $\mu_1 = \frac{1}{3.70} \approx 0.27$. The service rate for Buffer-2 customers served in Buffer-1 is given by $E\eta_{2,1} = \frac{24.79}{7} \approx 3.54$ and $\mu_{2,1} = \frac{1}{3.54} \approx 0.28$.

Table 5 shows that 37 out of 49 exogenous customers who arrived within the given time slot were served in their rightful buffer, whereas 7 customers moved from Buffer-2 to Buffer-1 for service and 5 customers moved from Buffer-2 to Buffer-3 for service. The total service time spent by customers in Buffer-2 is 174 min.

The mean arrival rate for Buffer-2 is shown in Table 6. According to Table 6, the highest number of customers (17) arrived within the time interval 9:00–9:30 a.m. and the lowest number of customers (2) arrived within the time interval 11:30 a.m.–12:00 p.m. The mean arrival rate for Buffer-2 is $\lambda_2 = \frac{1.63}{6} \approx 0.27$. The service rate of customers served in Buffer-2 is given by $\mu_2 = \frac{1}{3.76} \approx 0.28$. The service rate for Buffer-1 customers served in Buffer-2 is given by $E\eta_{1,2} = \frac{18.64}{6} \approx 3.11$ and $\mu_{1,2} = \frac{1}{3.11} \approx 0.32$. The service rate for Buffer-3 customers served in Buffer-2 is given by $E\eta_{3,2} = \frac{16.35}{6} \approx 2.73$ and $\mu_{3,2} = \frac{1}{2.73} \approx 0.37$.

Table 4. The mean arrival rate of Buffer-1.

Time Interval (a.m.)	Customers' Number	Inter-Arrival Rate
9:00–9:30	7	0.23
9:30–10:00	16	0.53
10:00–10:30	5	0.17
10:30–11:00	8	0.27
11:00–11:30	7	0.23
11:30–12:00	4	0.13
Total	47	1.57

Table 5. Queue-2 (Time in minutes).

$E_2(t)$	$D_2(t)$	$Z_2(t)$	$D_{2,1}(t)$	$Z_{2,1}(t)$	$D_{2,3}(t)$	$Z_{2,3}(t)$	$T_2(t)$	$T_{1,2}(t)$	$T_{3,2}(t)$
49	37	37	7	7	5	5	139	19	16

Table 6. The mean arrival rate of Buffer-2.

Time Interval (a.m.)	Customers' Number	Inter-Arrival Rate
9:00–9:30	17	0.57
9:30–10:00	11	0.37
10:00–10:30	6	0.20
10:30–11:00	7	0.23
11:00–11:30	6	0.20
11:30–12:00	2	0.07
Total	49	1.63

As shown in Table 7, we recorded 35 exogenous customers within the given time slot, with 6 customers joining Buffer-3 for service. The total service time recorded in Buffer-3 is 175 min.

Table 8 shows the mean arrival rate for Buffer-3. According to Table 8, the highest number of customers (12) arrived within the time interval 9:30–10:00 a.m. The mean arrival rate for Buffer-3 is $\lambda_3 = \frac{1.17}{6} \approx 0.19$. The service rate of customers served in Buffer-3 is $\mu_3 = \frac{1}{4.52} = 0.22 \approx 2$. The service rate for Buffer-2 customers served in Buffer-3 is given by $E\eta_{2,3} = \frac{17.43}{5} \approx 3.54$ and $\mu_{2,3} = \frac{1}{3.54} \approx 0.28$.

Table 7. Queue-3 (time in minutes).

$E_3(t)$	$D_3(t)$	$Z_3(t)$	$D_{3,2}(t)$	$Z_{3,2}(t)$	$T_3(t)$	$T_{2,3}(t)$
35	29	29	6	6	158	17

Table 8. The mean arrival rate of Buffer-3.

Time Interval (a.m.)	Customers' Number	Inter-Arrival Rate
9:00–9:30	7	0.23
9:30–10:00	12	0.53
10:00–10:30	3	0.17
10:30–11:00	7	0.27
11:00–11:30	3	0.23
11:30–12:00	3	0.13
Total	35	1.17

Figure 2 shows the movements of customers within the three buffers at a specific time intervals. It can be deduced that more customers are served in Buffer-1 from 9:30–10:00 a.m. and from 11:30 a.m.–12:00 p.m. In addition, more customers are also served in Buffer-2 from 9:00–9:30 a.m. followed by 10:00–10:30 a.m. Generally, Figure 2 depicts that Buffer-3 recorded the smallest number of customers served in the various specific time interval as compared to Buffer-1 and Buffer-2.

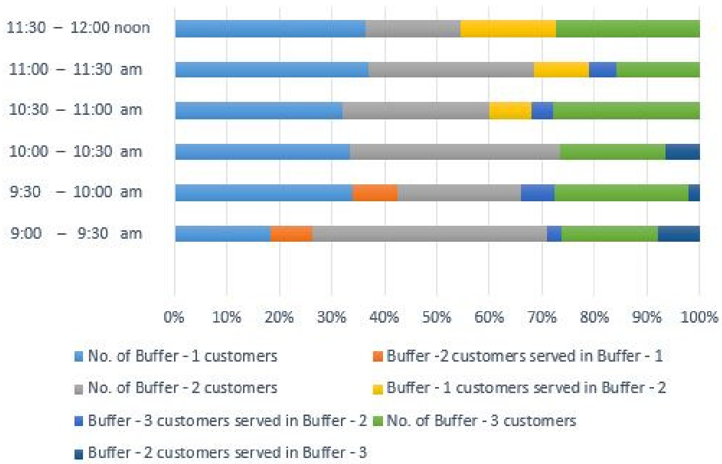

Figure 2. Comparison between the numbers of customers served in each buffer.

Table 9 shows the comparison between the arrival and service rates among customers. According to the results shown in Table 9, the mean arrival and service rates for Buffer-1 and Buffer-2 are approximately 3 minutes per customer, whereas the arrival and service rate is about 3 min for Buffer-3. Customers spent 3 min with a server while being served at a changed buffer; also, the average service time a customer spent in their rightful buffer is less than the average service time that customers spent in a changed buffer. This, therefore, supports Remark 1.

Table 9. Comparison of the arrival and service rates.

Queue	Arrival Rates	Service Rates		
Buffer-1	$\lambda_1 = 0.26$	$\mu_1 = 0.27$	$\mu_{2,1} = 0.28$	
Buffer-2	$\lambda_2 = 0.27$	$\mu_2 = 0.28$	$\mu_{1,2} = 0.32$	$\mu_{3,2} = 0.37$
Buffer-3	$\lambda_3 = 0.19$	$\mu_3 = 0.22$	$\mu_{2,3} = 0.28$	

Figures 3–5 show the relationship among customers arrival, service and departure times for the three buffers.

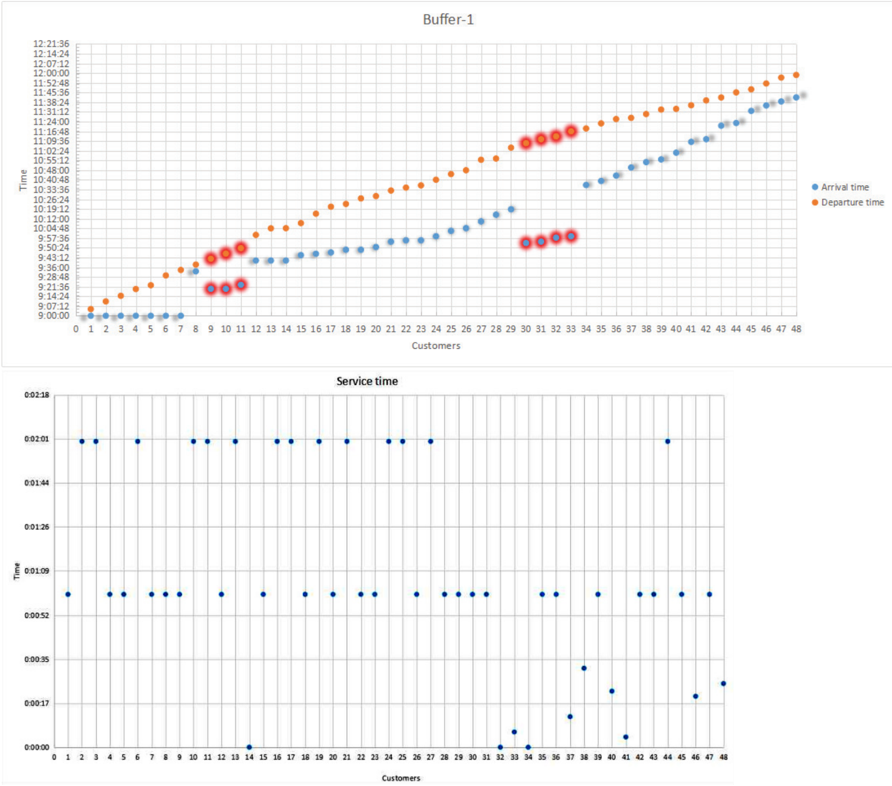

Figure 3. Arrival and departure times together with service times in Server-1 for Buffer-1. A point with red shadow represents a customer moving from Buffer-2 to Buffer-1 for service.

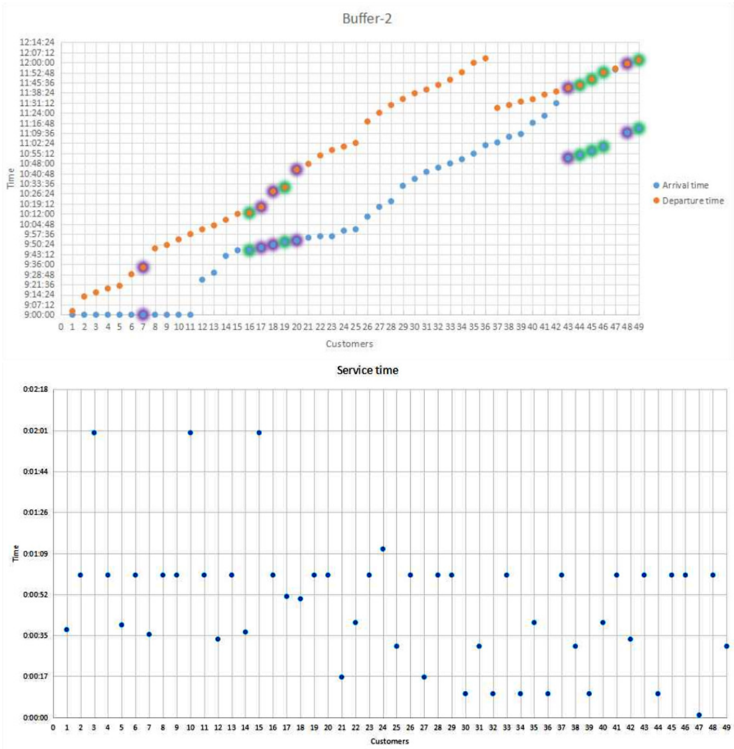

Figure 4. Arrival and departure times together with service times in Server-2 for Buffer-2. A point with green shadow represents a customer moving from Buffer-1 to Buffer-2 and a point with purple shadow represents a customer moving from Buffer-3 to Buffer-2 for service.

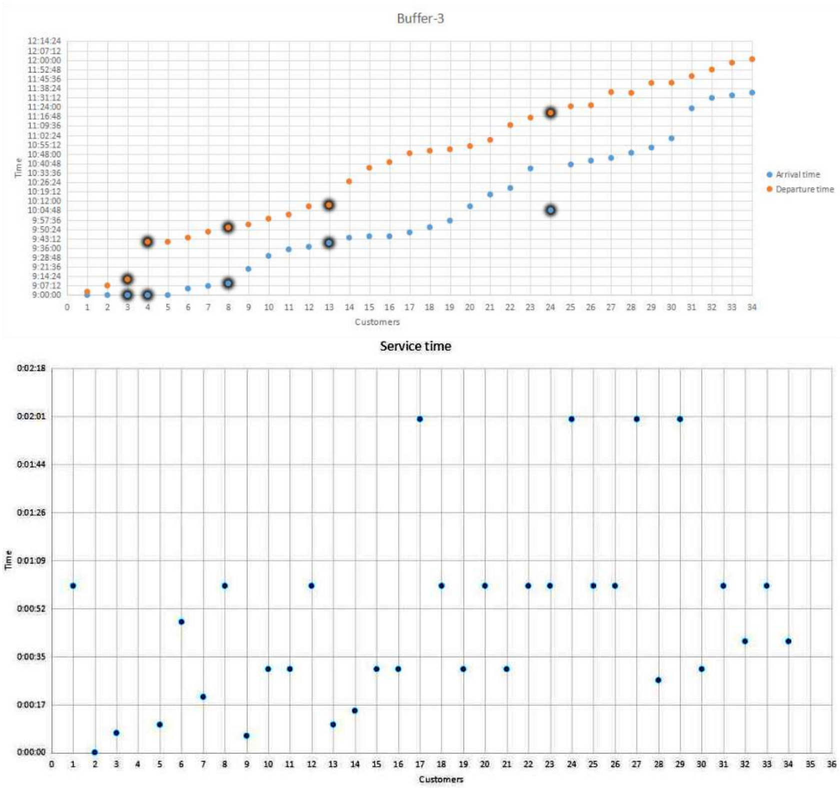

Figure 5. Arrival and departure times together with service times in Server-3 for Buffer-3. A point with black shadow represents a customer moving from Buffer-2 to Buffer-3 for service.

In comparing the arrival and service rates, it is to be noted that the performance measures from the study in Table 10 satisfy all the six stability conditions for the Double-X model. This, therefore, indicates that the model is fit to make predictions of the customers' arrival and service rate at the bank.

Table 10. Stability analysis on the data collected.

Sufficient Condition	**Stability**
$(A_1): \lambda_1 - \mu_1 + \frac{\lambda_2 - \mu_2}{r_{1,2}} = -0.01 < 0$	Stable
$(A_2): \lambda_2 - \mu_2 + \frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} = -0.02 < 0$	Stable
$(A_3): \lambda_3 - \mu_3 + \frac{\lambda_2 - \mu_2}{r_{3,2}} = -0.03 < 0$	Stable
$(A_4): \frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_2 - \mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{1,2}r_{3,2}r_{2,3}} = -0.001 < 0$	Stable
$(A_5): \frac{\lambda_2 - \mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} + \frac{\lambda_1 - \mu_1}{r_{2,3}r_{3,2}r_{2,1}} = -0.001 < 0$	Stable
$(A_6): \frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_2 - \mu_2}{r_{2,1}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} = -0.002 < 0$	Stable

5. Conclusions

In this paper, we model a server–customer interaction occurring in the banking sector, using a Double-X cascade network. In the model, each server has a renewal customer with general i.i.d. inter-arrival times and general i.i.d. service times. The Double-X cascade model contains three servers with three classes of customers awaiting in three buffers for service. Server-1, when free, serves customers who desire services 1–4 waiting in the queue of Buffer-2 to be served; Server-2, when free, serves customers in the queue of Buffer-1 and Buffer-3, while Server-3, when free, serves customers who require services 5–9 in queue of Buffer-2. The main component of the analysis is the finding of the Lyapunov function, which is used to characterize the positive recurrence of the basic Markov process describing stochastic dynamics of the network [16].

We find a sufficient stability condition for the Double-X cascade network system, using the fluid limit approach. Furthermore, we present an illustrative example describing the real-world functioning of the banking system. The average time a customer spends in the system with Server-j, $j = 1, 2, 3$ is determined to be 0.27, 0.28 and 0.22 h, respectively. Based on our mathematical findings and observations made for the case study, we find adequate arrival and service rates to meet customer demands, even in situations in which periods of high demand arise during business hours.

Author Contributions: Conceptualization, H.Z. and S.W.A.; formal analysis, H.Z., S.W.A., W.O.A., P.H., D.Y. and P.G.; investigation, H.Z., S.W.A., W.O.A., P.H., D.Y. and P.G.; writing—original draft preparation, H.Z., S.W.A., W.O.A., P.H. and P.G.; writing—review and editing, H.Z., S.W.A. and P.G.; supervision, H.Z.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: H.Z. acknowledges the financial support of the Chinese Ministry of Education of Humanities and Social Science Project 20YJA630088 and the Qinlan Project of Jiangsu Province.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this paper are available in the article.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding the publication of the research article.

Appendix A

This Appendix contains proofs of the stability conditions shown in Equation (39). Let us consider a fluid limit model $(\overline{X}, \overline{D}, \overline{T}, \overline{Y})$, where $\overline{X} = (\overline{Q}, \overline{Z}, \overline{U}, \overline{V})$ satisfies (23)–(37), and assume that $|Q(0)| = 1$. We introduce the following Lyapunov function:

$$f(t) = \frac{\overline{Q}_1(t)}{r_{2,1}} + \frac{\overline{Q}_2(t)}{r_{1,2}r_{3,2}} + \frac{\overline{Q}_3(t)}{r_{2,3}}, \quad t \geqslant 0.$$

Then, $\overline{Q}$ and f have the same differentiability points. It follows from (23)–(25) that for any regular point t of $\overline{Q}$, the following holds:

$$\dot{\overline{Q}}_1(t) = \lambda_1 - (\mu_1 \dot{\overline{T}}_1(t) + \mu_{1,2}\dot{\overline{T}}_{1,2}(t)), \tag{A1}$$

$$\dot{\overline{Q}}_2(t) = \lambda_2 - (\mu_2 \dot{\overline{T}}_2(t) + \mu_{2,1}\dot{\overline{T}}_{2,1}(t) + \mu_{2,3}\dot{\overline{T}}_{2,3}(t)), \tag{A2}$$

$$\dot{\overline{Q}}_3(t) = \lambda_3 - (\mu_3 \dot{\overline{T}}_3(t) + \mu_{3,2}\dot{\overline{T}}_{3,2}(t)), \tag{A3}$$

which altogether imply the following:

$$\begin{aligned} \dot{f}(t) = \frac{\lambda_1}{r_{2,1}} - \left(\frac{\mu_1}{r_{2,1}}\dot{\overline{T}}_1(t) + \frac{\mu_{1,2}}{r_{2,1}}\dot{\overline{T}}_{1,2}(t)\right) + \frac{\lambda_2}{r_{1,2}r_{3,2}} - \Big(\frac{\mu_2}{r_{1,2}r_{3,2}}\dot{\overline{T}}_2(t) \\ + \frac{\mu_{2,1}}{r_{1,2}r_{3,2}}\dot{\overline{T}}_{2,1}(t) + \frac{\mu_{2,3}}{r_{1,2}r_{3,2}}\dot{\overline{T}}_{2,3}(t)\Big) + \frac{\lambda_3}{r_{2,3}} - \left(\frac{\mu_3}{r_{2,3}}\dot{\overline{T}}_3(t) + \frac{\mu_{3,2}}{r_{2,3}}\dot{\overline{T}}_{3,2}(t)\right). \end{aligned} \tag{A4}$$

Let $t \geqslant 0$ be a differentiability point of f such that $f(t) > 0$. Then, $\overline{Q}_1(t) > 0$, $\overline{Q}_2(t) > 0$ or $\overline{Q}_3(t) > 0$. We show that in either case, $\dot{f}(t) \leqslant -C$, for a constant $C > 0$. Then, f is non-increasing and $f(t) = 0$ for all $t \geqslant \frac{f(0)}{C}$.

Because $|\overline{Q}(0)| = \overline{Q}_1(0) + \overline{Q}_2(0) + \overline{Q}_3(0) = 1$ from Definition (1), we have $\overline{Q}(t) = 0$ for all $t \geqslant t_1$. This happens since the following is true:

$$f(0) = \frac{\overline{Q}_1(0)}{r_{2,1}} + \frac{\overline{Q}_2(0)}{r_{1,2}r_{3,2}} + \frac{\overline{Q}_3(0)}{r_{2,3}} \leqslant \frac{1}{r_{2,1}} + \frac{1}{r_{1,2}r_{3,2}} + \frac{1}{r_{2,3}}.$$

Then (20) will follow:

$$t_1 = \frac{1}{C}\left(\frac{1}{r_{2,1}} + \frac{1}{r_{1,2}r_{3,2}} + \frac{1}{r_{2,3}}\right) > 0.$$

Case 1:Assume that $\overline{Q}_1(t) > 0$ and $\overline{Q}_2(t) = \overline{Q}_3(t) = 0$. Then, by (26) and (27), $\dot{\overline{Y}}_1(t) = \dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_3(t) = 0$ and by (31) and (34), $\dot{\overline{T}}_{2,1}(t) = \dot{\overline{T}}_{3,2}(t) = 0$. We obtain the following:

$$\dot{\overline{T}}_1(t) + \dot{\overline{T}}_{2,1}(t) + \dot{\overline{Y}}_1(t) = 1 \implies \dot{\overline{T}}_1(t) = 1.$$

From Equations (A2) and (A3), $\overline{Q}_2(t) = \overline{Q}_3(t) = 0$, so t is a local minimum. As t is a regular point, then by Fermat's theorem on stationary points, $\dot{\overline{Q}}_2(t) = \dot{\overline{Q}}_3(t) = 0$. Therefore, the following holds:

$$\begin{aligned}\lambda_2 &= (\mu_2\dot{\overline{T}}_2(t) + \mu_{2,1}\dot{\overline{T}}_{2,1}(t) + \mu_{2,3}\dot{\overline{T}}_{2,3}(t)) \\ &= \mu_2\dot{\overline{T}}_2(t) \implies \dot{\overline{T}}_2(t) = \frac{\lambda_2}{\mu_2},\end{aligned}$$

also from Equation (29). Further, we have from (26) the following:

$$\begin{aligned}\dot{\overline{T}}_2(t) + \dot{\overline{T}}_{1,2}(t) + \dot{\overline{T}}_{3,2}(t) =& 1 \\ \implies \dot{\overline{T}}_{1,2}(t) =& 1 - \dot{\overline{T}}_2(t) = 1 - \frac{\lambda_2}{\mu_2}.\end{aligned}$$

Moreover, from (A3), $\dot{\overline{Q}}_3(t) = 0$, therefore we have the following:

$$\lambda_3 = \mu_3\dot{\overline{T}}_3(t) + \mu_{3,2}\dot{\overline{T}}_{3,2}(t).$$

Finally, we obtain the following:

$$\begin{aligned}\dot{f}(t) &= \frac{\lambda_1}{r_{2,1}} - \left(\frac{\mu_1}{r_{2,1}} + \frac{\mu_{1,2}}{r_{2,1}}(1 - \frac{\lambda_2}{\mu_2})\right) + \frac{\lambda_2}{r_{1,2}r_{3,2}} - \left(\frac{\mu_2}{r_{1,2}r_{3,2}} \times \frac{\lambda_2}{\mu_2}\right), \\ &= \frac{\lambda_1 - \mu_1}{r_{2,1}} - \frac{\mu_{1,2}}{r_{2,1}}\left(\frac{\mu_2 - \lambda_2}{\mu_2}\right).\end{aligned}$$

Since $r_{1,2} = \frac{\mu_2}{\mu_{1,2}}$, this implies the following:

$$\dot{f}(t) = \frac{1}{r_{2,1}}\left((\lambda_1 - \mu_1) + \frac{\lambda_2 - \mu_2}{r_{1,2}}\right) = -C_1,$$

the constant

$$C_1 = -\frac{1}{r_{2,1}}\left((\lambda_1 - \mu_1) + \frac{\lambda_2 - \mu_2}{r_{1,2}}\right)$$

being positive by condition (A_1).

Case 2: Assume that $\overline{Q}_2(t) > 0$ and $\overline{Q}_1(t) = \overline{Q}_3(t) = 0$. Then, by (26) and (27), $\dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_1(t) = \dot{\overline{Y}}_3(t) = 0$. Equations (33)–(36) give $\dot{\overline{T}}_{1,2}(t) = \dot{\overline{T}}_{3,2}(t) = 0$. As a result, we obtain from (26) the following:

$$\begin{aligned}\dot{\overline{T}}_2(t) + \dot{\overline{T}}_{1,2}(t) + \dot{\overline{T}}_{3,2}(t) + \dot{\overline{Y}}(t) = 1\\ \implies \dot{\overline{T}}_2(t) =1,\end{aligned}$$

$$\dot{\overline{T}}_1(t) + \dot{\overline{T}}_{2,1}(t) = 1, \tag{A5}$$

$$\dot{\overline{T}}_3(t) + \dot{\overline{T}}_{2,3}(t) = 1. \tag{A6}$$

By Fermat's theorem on stationary points, one sees from Equation (A1) and (A3) that $\dot{\overline{Q}}_1(t) = \dot{\overline{Q}}_3(t) = 0$. Therefore, we have the following:

$$\lambda_1 = \mu_1 \dot{\overline{T}}_1(t) \implies \dot{\overline{T}}_1(t) = \frac{\lambda_1}{\mu_1},$$
$$\lambda_3 = \mu_3 \dot{\overline{T}}_3(t) \implies \dot{\overline{T}}_3(t) = \frac{\lambda_3}{\mu_3}.$$

From (A5), we have the following:

$$\dot{\overline{T}}_{2,1}(t) = 1 - \dot{\overline{T}}_1(t) = \left(1 - \frac{\lambda_1}{\mu_1}\right),$$

while from (A6), we have the following:

$$\dot{\overline{T}}_{2,3}(t) = 1 - \dot{\overline{T}}_3(t) = \left(1 - \frac{\lambda_3}{\mu_3}\right),$$

which implies the following:

$$\dot{f}(t) = \frac{1}{r_{1,2}r_{3,2}}\big((\lambda_2 - \mu_2) + (\lambda_1 - \mu_1)\frac{\mu_{2,1}}{\mu_1} + (\lambda_3 - \mu_3)\frac{\mu_{2,3}}{\mu_3}\big).$$

Since $r_{2,1} = \frac{\mu_1}{\mu_{2,1}}$ and $r_{2,3} = \frac{\mu_3}{\mu_{2,3}}$, this yields the following:

$$\dot{f}(t) = \frac{1}{r_{1,2}r_{3,2}}\big((\lambda_2 - \mu_2) + \frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_3 - \mu_3}{r_{2,3}}\big) = -C_2,$$

the constant

$$C_2 = -\frac{1}{r_{1,2}r_{3,2}}\left((\lambda_2 - \mu_2) + \frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_3 - \mu_3}{r_{2,3}}\right)$$

being positive by condition (A_2).

Case 3: Assume that $\overline{Q}_3(t) > 0$ and $\overline{Q}_1(t) = \overline{Q}_2(t) = 0$, then by (26), (27), (31) and (32), $\dot{\overline{Y}}_3(t) = \dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_1(t) = \dot{\overline{T}}_{2,1}(t) = \dot{\overline{T}}_{2,3}(t) = 0$. From Equation (26), we obtain the following:

$$\begin{aligned}\dot{\overline{T}}_3(t) + \dot{\overline{T}}_{2,3}(t) + \dot{\overline{Y}}_3(t) = 1,\\ \dot{\overline{T}}_3(t) = 1.\end{aligned}$$

From Equations (A1) and (A2), $\overline{Q}_1(t) = \overline{Q}_2(t) = 0$, so t is a local minimum. As t is a regular point, then by Fermat's theorem on stationary points, $\dot{\overline{Q}}_2(t) = \dot{\overline{Q}}_1(t) = 0$. Therefore, we have the following:

$$\begin{aligned}
\lambda_1 &= (\mu_1 \dot{\overline{T}}_1(t) + \mu_{1,2}\dot{\overline{T}}_{1,2}(t)). \\
\lambda_2 &= (\mu_2 \dot{\overline{T}}_2(t) + \mu_{2,1}\dot{\overline{T}}_{2,1}(t) + \mu_{2,3}\dot{\overline{T}}_{2,3}(t)) \\
&= \mu_2 \dot{\overline{T}}_2(t), \\
\dot{\overline{T}}_2(t) &= \frac{\lambda_2}{\mu_2}.
\end{aligned}$$

From (26), $\dot{\overline{T}}_2(t) + \dot{\overline{T}}_{1,2}(t) + \dot{\overline{T}}_{3,2}(t) = 1$, which implies the following:

$$\dot{\overline{T}}_{3,2}(t) = \left(1 - \frac{\lambda_2}{\mu_2}\right),$$

and consequently the following:

$$\begin{aligned}
\dot{f}(t) &= \frac{\lambda_3 - \mu_3}{r_{2,3}} - \frac{1}{r_{2,3}}\left(\mu_{2,3}(1 - \frac{\lambda_2}{\mu_2})\right), \\
\dot{f}(t) &= \frac{1}{r_{2,3}}\left((\lambda_3 - \mu_3) + \frac{\lambda_2 - \mu_2}{r_{3,2}}\right) = -C_3,
\end{aligned}$$

the constant

$$C_3 = -\frac{1}{r_{2,3}}\left((\lambda_3 - \mu_3) + \frac{\lambda_2 - \mu_2}{r_{3,2}}\right)$$

being positive by condition (A_3).

Case 4: Assume that $\overline{Q}_1(t) > 0, \quad \overline{Q}_2(t) = 0$ and $\overline{Q}_3(t) > 0$. Then by (26), (27), (31) and (32), $\dot{\overline{Y}}_1(t) = \dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_3(t) = \dot{\overline{T}}_{2,1}(t) = \dot{\overline{T}}_{2,3}(t) = 0$, which gives the following:

$$\begin{aligned}
\dot{\overline{T}}_1(t) + \dot{\overline{T}}_{2,1}(t) + \dot{\overline{Y}}_1(t) &= 1, \\
\dot{\overline{T}}_3(t) + \dot{\overline{T}}_{2,3}(t) + \dot{\overline{Y}}_3(t) &= 1,
\end{aligned}$$

$$\dot{\overline{T}}_1(t) = \dot{\overline{T}}_3(t) = 1.$$

Then, from Equation (A2), by Fermat's theorem on stationary points, $\dot{\overline{Q}}_2(t) = 0$. This implies the following:

$$\begin{aligned}
\lambda_2 &= \mu_2 \dot{\overline{T}}_2(t) + \mu_{2,1}\dot{\overline{T}}_{2,1}(t) + \mu_{2,3}\dot{\overline{T}}_{2,3}(t), \\
\dot{\overline{T}}_2(t) &= \frac{\lambda_2}{\mu_2}.
\end{aligned}$$

Again from (26), we have the following:

$$\dot{\overline{T}}_2(t) + \dot{\overline{T}}_{1,2}(t) + \dot{\overline{T}}_{3,2}(t) + \dot{\overline{Y}}_2(t) = 1.$$

From Phase 1:

$$\dot{\overline{T}}_{3,2}(t) = 0 \implies \dot{\overline{T}}_{1,2}(t) = 1 - \frac{\lambda_2}{\mu_2}.$$

Additionally, from Phase 2 we have the following:

$$\dot{\overline{T}}_{1,2}(t) = 0 \implies \dot{\overline{T}}_{3,2}(t) = 1 - \frac{\lambda_2}{\mu_2},$$

resulting in the following:

$$\begin{aligned}\dot{f}(t) &= \frac{\lambda_1}{r_{2,1}} - \left(\frac{\mu_1}{r_{2,1}} + \frac{\mu_{1,2}}{r_{2,1}}(1-\frac{\lambda_2}{\mu_2})\right) + \frac{\lambda_2}{r_{1,2}r_{3,2}} - \left(\frac{\mu_2}{r_{1,2}r_{3,2}} \times \frac{\lambda_2}{\mu_2}\right),\\ &= \frac{\lambda_1-\mu_1}{r_{2,1}} + \frac{1}{r_{2,1}}\left(\frac{\lambda_2-\mu_2}{\mu_2} \times \frac{\mu_{1,2}}{\mu_2}\right),\\ &= \frac{\lambda_1-\mu_1}{r_{2,1}} + \frac{\lambda_2-\mu_2}{r_{2,1}r_{1,2}} = -C_4.\end{aligned}$$

where the constant

$$C_4 = -\frac{1}{r_{2,1}}\left(\lambda_1-\mu_1+\frac{\lambda_2-\mu_2}{r_{1,2}}\right),$$

is positive by condition (A_1).

Case 5: Assume that $\overline{Q}_1(t) > 0,\quad \overline{Q}_2(t) > 0$ and $\overline{Q}_3(t) = 0$. Then, by (26), (27) and (31), (33), (34), $\dot{\overline{Y}}_1(t) = \dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_3(t) = \dot{\overline{T}}_{2,1}(t) = \dot{\overline{T}}_{1,2}(t) = \dot{\overline{T}}_{3,2}(t) = 0$, resulting in

$$\dot{\overline{T}}_1(t) = \dot{\overline{T}}_2(t) = 1.$$

From (A3),

$$\lambda_3 = \mu_3\dot{\overline{T}}_3(t) + \mu_{3,2}\dot{\overline{T}}_{3,2} \implies \dot{\overline{T}}_3(t) = \frac{\lambda_3}{\mu_3}.$$

Moreover, from (26),

$$\begin{aligned}\dot{\overline{T}}_3(t) + \dot{\overline{T}}_{2,3}(t) + \dot{\overline{Y}}_3(t) &= 1\\ \dot{\overline{T}}_{2,3}(t) &= 1-\frac{\lambda_3}{\mu_3}.\end{aligned}$$

Using (A4), this implies the following:

$$\begin{aligned}\dot{f}(t) &= \frac{\lambda_1}{r_{2,1}} - \left(\frac{\mu_1}{r_{2,1}}\right) + \frac{\lambda_2}{r_{1,2}r_{3,2}} - \left(\frac{\mu_2}{r_{1,2}r_{3,2}} + \frac{\mu_{2,3}}{r_{1,2}r_{3,2}}(1-\frac{\lambda_3}{\mu_3})\right),\\ &= \frac{\lambda_1-\mu_1}{r_{2,1}} + \frac{\lambda_2-\mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3-\mu_3}{r_{1,2}r_{3,2}r_{2,3}},\\ &= \frac{\lambda_1-\mu_1}{r_{2,1}} + \frac{\lambda_2-\mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3-\mu_3}{r_{1,2}r_{3,2}r_{2,3}} = -C_5,\end{aligned}$$

the constant

$$C_5 = -\left(\frac{\lambda_1-\mu_1}{r_{2,1}} + \frac{\lambda_2-\mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3-\mu_3}{r_{1,2}r_{3,2}r_{2,3}}\right),$$

being positive by condition (A_4).

Case 6: For $\overline{Q}_1(t) = 0,\quad \overline{Q}_2(t) > 0$ and $\overline{Q}_3(t) > 0$ Then, by (26), (27) and (32)–(34), $\dot{\overline{Y}}_1(t) = \dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_3(t) = \dot{\overline{T}}_{2,3}(t) = \dot{\overline{T}}_{1,2}(t) = \dot{\overline{T}}_{3,2}(t) = 0$, resulting in

$$\dot{\overline{T}}_2(t) = \dot{\overline{T}}_3(t) = 1.$$

From (A1),

$$\lambda_1 = \mu_1\dot{\overline{T}}_1(t) + \mu_{1,2}\dot{\overline{T}}_{1,2} \implies \dot{\overline{T}}_1(t) = \frac{\lambda_1}{\mu_1}.$$

Then Equation (26) gives us the following:

$$\dot{\overline{T}}_1(t) + \dot{\overline{T}}_{2,1}(t) + \dot{\overline{Y}}_1(t) = 1$$

$$\dot{\overline{T}}_{2,1}(t) = 1 - \frac{\lambda_1}{\mu_1},$$

$$\begin{aligned}\dot{f}(t) &= \frac{\lambda_2 - \mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} - \left(\frac{\mu_{2,1}}{r_{1,2}r_{3,2}}(1 - \frac{\lambda_1}{\mu_1})\right),\\ &= \frac{\lambda_2 - \mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} + \frac{\lambda_1 - \mu_1}{r_{2,3}r_{3,2}r_{2,1}},\\ &= \frac{\lambda_2 - \mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} + \frac{\lambda_1 - \mu_1}{r_{2,3}r_{3,2}r_{2,1}} = -C_6,\end{aligned}$$

the constant

$$C_6 = -\left(\frac{\lambda_2 - \mu_2}{r_{1,2}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} + \frac{\lambda_1 - \mu_1}{r_{2,3}r_{3,2}r_{2,1}}\right)$$

being positive by condition (A_5).

Case 7: Assume that $\overline{Q}_1(t) > 0$, $\overline{Q}_2(t) > 0$ and $\overline{Q}_3(t) > 0$. Then, by (26), (27) and (31)–(36), we have $\dot{\overline{Y}}_1(t) = \dot{\overline{Y}}_2(t) = \dot{\overline{Y}}_3(t) = \dot{\overline{T}}_{1,2}(t) = \dot{\overline{T}}_{2,1}(t) = \dot{\overline{T}}_{2,3}(t) = \dot{\overline{T}}_{3,2}(t) = 0$, implying the following:

$$\dot{\overline{T}}_1(t) = \dot{\overline{T}}_2(t) = \dot{\overline{T}}_3(t) = 1.$$

From (A1)–(A3), we see the following:

$$\dot{f}(t) = \frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_2 - \mu_2}{r_{2,1}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}} = -C_7,$$

the constant

$$C_7 = -\left(\frac{\lambda_1 - \mu_1}{r_{2,1}} + \frac{\lambda_2 - \mu_2}{r_{2,1}r_{3,2}} + \frac{\lambda_3 - \mu_3}{r_{2,3}}\right)$$

being positive by condition (A_6). Consequently, in either case, $\dot{f}(t) \leqslant -C < 0$, where $C := \min\{C_1, C_2, C_3, C_4, C_5, C_6, C_7\}$.

References

1. Afanaseva, L.G.; Grishunina, S.A. Stability conditions for a multiserver queuing system with a regenerative input flow and simultaneous service of a customer by a random number of servers. *Queueing Syst.* **2020**, *94*, 213–241. [CrossRef]
2. Marfo-Yiadom, E.; Ansong, A. Customers' perception of innovative banking products in Cape Coast Metropolis Ghana. *Int. J. Bus. Adm. Manag. Res.* **2012**, *7*, 162–172. [CrossRef]
3. Dungey, M.; Gajurel, D. Contagion and banking crisis-international evidence for 2007–2009. *J. Bank. Financ.* **2015**, *60*, 271–283. [CrossRef]
4. Owusu-Frimpong, N. Patronage behavoiur of Ghanaian bank customers. *Int. J. Bank Mark.* **1999**, *17*, 335–342. [CrossRef]
5. Bain, C.A.; Taylor, P.G.; McDonnell, G.; Georgiou, A. Myths of ideal hospital occupancy. *Med. J. Aust.* **2010**, *192*, 42–43. [CrossRef]
6. Barak, S.; Fallahnezhad, M.S. Cost analysis of fuzzy queuing systems. *Int. J. Appl. Oper. Res.* **2014**, *2*, 25–36.
7. Cascone, A.; Rarità, L.; Trapel, E. Simulation and analysis of a bank's multi-server queuing system. *J. Math. Sci.* **2014**, *196*, 23–29. [CrossRef]
8. Akwasi, A.G. Application of Queueing Model at the Driver and Vehicle Licensing Authority-Kumasi (Eye-Test Section). Master's Thesis, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, 2015. Available online: http://ir.knust.edu.gh/bitstream/123456789/8088/1/AKWASI%20ADU%20GYAMFI.pdf (accessed on 13 March 2017).
9. Yanting, C.; Richard, J.B.; Jasper, G. Necessary conditions for the compensation approach for a random walk in the quarter-plane. *Queueing Syst.* **2019**, *94*, 257–277.
10. Sanda, M.A.; Arhin, E. Using ATMs as workload relievers for Ghanaian bank servers: the customer behavioral challenge. *J. Econ. Behav. Organ.* **2011**, *3*, 13–21.
11. Nosek, R.A.; Wilson, J.P. Queuing theory and customer satisfaction: a review of terminology, trends and applications to pharmacy practice. *Hosp. Pharm.* **2001**, *36*, 275–279. [CrossRef]

12. Kristen, G.; Rhonda R. Product forms for FCFS queuing models with arbitrary server-job compatibilities: An overview. *Queueing Syst.* **2020**, *96*, 3–51.
13. Vasumathi, A.; Dhanavanthan, P. Application of simulation technique in queuing model for ATM facility. *Int. J. Appl. Eng. Res.* **2010**, *1*, 103–110.
14. Ibrahim, R.; Whitt, W. Real-time delay estimation in overloaded multiserver queues with abandonments. *Manag. Sci.* **2009**, *55*, 1729–1742. [CrossRef]
15. Boako, G.; Alagidede, P. Examining evidence of 'shift-contagion' in African stock markets: A CoVaR-copula approach. *Rev. Dev. Financ.* **2017**, *7*, 142–156. [CrossRef]
16. Delgado, R.; Morozov, E. Stability analysis of cascade networks via fluid models. *Perform. Eval.* **2014**, *82*, 39–54. [CrossRef]
17. Morozov, E.; Steyaert, B. Stability analysis of a two-station cascade queuing network. *Ann. Oper. Res.* **2013**, *202*, 135–160. [CrossRef]
18. Maleki, L.; Pasandideh, S.H.R.; Niaki, S.T.A.; Cárdenas-Barrón, L.E. Determining the prices of re-manufactured products, capacity of internal workstations and the contracting strategy within queuing framework. *Appl. Soft Comput.* **2017**, *54*, 313–321. [CrossRef]
19. Nafees, A. Queuing Theory and Its Application: Analysis of the Sales Checkout Operation in ICA Supermarket. Master's Thesis, University of Dalarna, Falun, Sweden, 2007. Available online: http://www.statistics.du.se/essays/D07E.Nafees.pdf (accessed on 13 March 2017).
20. Ahn, H.S.; Duenyas, I.; Zhang, Q.R. Optimal control of a flexible server. *Adv. Appl. Probab.* **2004**, *36*, 139–170. [CrossRef]
21. Kirkizlar, E.; Andradottir, S.; Ayhan, H. Robustness of efficient server assignment policies to service time distributions in finite-buffered lines. *Nav. Res. Logist.* **2010**, *57*, 563–582. [CrossRef]
22. Agnihothri, S.R.; Mishra, A.K.; Simmons, D.E. Workforce cross-training decisions in field service systems with two job types. *J. Oper. Res. Soc.* **2003**, *54*, 410–418. [CrossRef]
23. Terekhov, D.; Beck, J.C. An extended queuing control model for facilities with front room and back room operations and mixed-skilled workers. *Eur. J. Oper. Res.* **2009**, *198*, 223–231. [CrossRef]
24. Tezcan, T. Stability analysis of N-model systems under a static priority rule. *Queueing Syst.* **2013**, *73*, 235–259. [CrossRef]
25. Wang, K.; Li, N.; Jiang, Z. Queueing system with impatient customers: A review. In Proceedings of the 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2010, Qingdao, China, 15–17 July 2010; IEEE: Piscataway, NJ, USA, 2015; pp. 82–87.
26. Perry, O.; Whitt, H. Achieving rapid recovery in an overload control for large-scale service systems. *INFORMS J. Comput.* **2016**, *27*, 491–506. [CrossRef]
27. Udayabhanu, V.; Sunder, K.; Kannan, S. Optimizing capacity utilization in queuing systems: parallels with the EOQ model. *Calif. J. Oper. Manag.* **2010**, *8*, 54–59.
28. Dai, J.G. On positive Harris recurrence of multiclass queuing networks: A unified approach via fluid limit models. *Ann. Appl. Probab.* **1995**, *5*, 49–77. [CrossRef]
29. Burdzy, K.; Kang, W.; Ramanan, K. The Skorokhod problem in a time-dependent interval. *Stoch. Proc. Appl.* **2009**, *119*, 428–452. [CrossRef]
30. Dai, J.G.; Meyn, S.P. Stability and convergence of moments for multiclass queuing networks via fluid limit models. *IEEE Trans. Automat. Cont.* **1995**, *40*, 1889–1904. [CrossRef]
31. Bramson, M. Stability of queuing networks. *Probab. Surv.* **2008**, *5*, 169–345. [CrossRef]
32. Chen, H. Fluid approximations and stability of multiclass queuing networks: work-conserving disciplines. *Ann. Appl. Probab.* **1995**, *5*, 637–665. [CrossRef]