CUSTOMER-TELLER SCHEDULING SYSTEM FOR OPTIMIZING BANKS SERVICE

Osmond E. Agbo and **Thomas A. Nwodoh**

Department of Electrical Engineering, Faculty of Engineering, University of Nigeria, Nsukka, Nigeria

Abstract

Customer satisfaction is a concern to service industries as customers expect to be served promptly when they arrive. Demand for service is highly variable: depends on hour of the day, or day of the week, or even dates of the month. For a service industry like a Bank, there is a need for efficient Bank Teller scheduling system which takes into account varying customer service demand levels. This paper models Bank Teller scheduling system for optimizing a Banks customer service. The model takes into account real time system behaviour including changing customer arrival rates throughout the day and customer balking. It provides scheduling rules and the corresponding service levels when demand varies with cost minimisation as goals.

Keywords: scheduler, service optimisation, queuing theory, simulation, customer satisfaction, integer programming, probability theory

1. Introduction

In service industries, demand for service is variable and often depends on the day of the month, day of the week, or even on the time of the day. However, service need to be delivered promptly when it is demanded [1]. Daily work scheduling is required in many service companies such as Hospitals and Banks. Bad staff-customer scheduling results in long customer waiting times, long queues, and consequently, waiting cost. Bad scheduling can also result in loss of productively of Tellers due to idle times. On the other hand, good scheduling results in low waiting cost, good Teller utilization, customer satisfaction, and more profit. Operations managers are faced with the problem of recognizing the trade-off that must be taken between providing good service and the cost of waiting for such service.

As service speeds up, time spent waiting on queue decreases. Service cost however increases as the level of service increases. The goal of managers is to schedule as few employees as possible while maintaining a minimum customer service level. Managers want queues short enough so that customers do not become dissatisfied and either leave without transacting their business or transact and never return in the future. However, some waiting can be allowed if the waiting cost is balanced with significant saving in service cost.

A typical Teller-Customer scheduling system has two goals:

(i) to determine the minimum number of personnel to satisfy a set of service level requirements, and

(ii) to build a schedule that specifies when a staff should start shift so that, the staffing level requirements of each period of the day are met [2].

It is therefore important to recognize peak and non-peak periods to decide the staffing needs. However, very often, identifying the changes in the demand level from past or real-time data is not a straightforward matter because of the high variability. Using queuing theory and simulation with a control system that monitors the state of the system at all times, one can model and analyze a real time queuing situation in a Bank, compare scheduling alternatives and the corresponding service levels, and provide the best scheduling rule based on the desired service level.

This paper provides techniques for scheduling Tellers to service arriving customers at the open Teller windows. These techniques include:

1. Monitoring the system to obtain: i.) Customer information such as: the customers arrival times, inter-arrival times, start of service time, wait time, the departure time, service time duration, and total time each customer spends in the system. ii.) Teller service information (example, Teller idle times).

2. Use of the information generated to compute system parameters which are used to: i.) schedule Tellers, ii.) determine when the waiting time duration or time in the system exceeds certain predetermined threshold level so that service time is adjusted to meet this threshold.

3. Balancing the waiting cost with the service cost in order to optimise services.

4. An algorithm (a computer programme) for obtaining those information using Discrete Event Modelling and Simulation in conjunction with Queuing theory.

2. Literature Review

Several approaches have been used by researchers for Bank Teller scheduling. These approaches include the use of Operations Research based scheduling tools such as Statistics, Work measurement, Queuing theory, and Integer programming [2]. The use of these operations research based scheduling tools for the staff scheduling problem do not solve the problem adequately. Mathematical models are very cumbersome and seldom provide complete answers to real time problems rather; they provide partial solutions [3]. For instance, real time situations violate the assumptions of classical analytical waiting line models. Queuing theory approach may help Banks schedule services using a FIFO scheduling policy. However, this approach does not capture and take customer information or service information generated continuously with time into account. Queuing theory rely mainly on probability predictions which may be unreliable. The existing approaches forecast arrival pattern of customers based on historical data and use that to schedule Tellers [3]. Such approach does not perform real time scheduling using actual arrival of customers. It does not handle dynamic resource availability. In such a system, a sudden explosion in the arrival rate may cause a serious deviation in the forecasted schedule based on historical data. This model presents a scheme for monitoring the system at all times to identify immediately when changes have occurred in the

system and then generate corrective action to take (see section 4.2).

It is the practice of Banks to often staff their Tellers as if all Teller staff had the same service capabilities. Most applications of queuing models for recommending staffing levels assume that service times are exponentially distributed and that each staff member has the same service time distribution [4]. Previously, there has not been strong reason to abandon these assumptions given the relative scarcity of Teller performance data. This model in this paper shows how to tract Teller transaction activities to obtain the Teller performance data thus, allowing service time data to be routinely generated and integrated with staffing schedules.

In most Banks, customer and service information are identified generally based on manual observation and personal judgment (this is obtained through information from visits to some local Banks). This gives inaccurate results and wastes time. It also requires continuous observation by management personnel and thus results in additional cost. There is no correct procedure for tracking customers sojourn in the system therefore shunting and jockeying are typical occurrences. These lead to some customers being dissatisfied as a customers who came first may be served last. The present work presents a scheme for continuously monitoring the system changes, generating the information as the changes occur.

Existing approaches do not provide a differentiated service to customers based upon the customers volume of transaction (this is based on observations made to some local Banks in Nigeria). Most Banks keep all types of customers together in the same queue. Equal service time is assumed for all customers irrespective of the type and volume of transactions. This may lead to wrong computations of the number of Tellers required. The present model employs segmentation of customers according to volume of transaction and Tellers allocated to each volume depending on the number of customers in each group.

The USA Patent No. 4,700,295 issued on October 13, 1987 to Katsof et al described a system and method for forecasting Bank traffic and scheduling work assignments for Bank personnel [5]. The system uses data gathering as a means for sensing the arrival and departure of customers as well as when a Teller is at a station. A data processing means counts the arrivals and departures of customers and measures the amount of time that the Teller is active at each station. The system uses two detectors. One detector is placed at the entrance and the other at the exit of the queue, respectively. The forecasting method uses a queuing model to obtain forecast of waiting time per customer and Teller utilization. A record is kept of the number of arrivals observed during each interval and of the average service time per each day of the week. The problem noticed with this system is that one cannot keep accurately the number of the customers arriving or departing. This is because anyone, whether a customer or not who passes through the detector is counted as customer. This may cause a serious deviation in the real number of customers counted. Also, as mentioned earlier, forecasting without using real time data generated continuously with time is unreliable because of the variability in the arrival and service rates.

Another invention, the USA Patent 5,541,835 issued to Dextraze relates to a method and a system for monitoring and forecasting customer traffic and customer servicing at a location, where each customer may be served at any one of a plurality of available service stations [6]. The object of the invention was to provide a system and a method for monitoring and forecasting

customer traffic and customer service in a manner which greatly reduces mathematical computations, increases the accuracy in real time, and uses less components. This invention did not address the problems enumerated in the case of the USA patent No.4, 700,295 mentioned above. One of the objectives of this present work is to minimize the use of hardware components and increase the accuracy of data gathering.

3. System Modeling

The problem is *how can managers schedule the available tellers so that customers do not wait too long before receiving the desired service irrespective of the time of their arrival, and that tellers are not kept idle typically because there are no customers to be served irrespective of the periods ?* How to get these optimum service conditions is the objective of this work. The optimum point is reached when a balance is made between the minimum service cost and the minimum waiting cost.

During busy hours, such as lunch hours, and during busy days, such as pay days, customers must wait in a queue before being served by a Teller. Increasing the number of Tellers will actually reduce the waiting time for the customers, but it does so at the expense of the cost of provision of extra Teller stations and Tellers. Thus this solution has only a limited applicability. Even when there are a large number of Tellers, they must be utilized efficiently. Waiting for customers who are not there at Teller stations does not constitute such an efficient utilisation. Maximum efficiency together with maximum customer service is best accomplished by scheduling Teller staff based on accurate prediction of levels of traffic at the Bank during different hours of the day. This can only be achieved accurately if the system is monitored continuously.

Goals of this research:

1. To find a trade-off between the cost of waiting and the cost of providing service in Bank. This helps managers to eliminate long queues and reduce labour cost by providing efficient Teller scheduling.

2. To maintain a minimum wait time and service cost by knowing when to increase the number of Tellers or withdraw Tellers. This maximizes the level of customer satisfaction with the service provided thereby optimizing service.

3.1. Approach to solving the problem

An algorithm for a superimposed control system which can continuously sample the throughput system waiting line is developed. This algorithm is used to monitor the system at all times and to communicate changes in the system input and output to other distant systems which performs some action (example, a system that raises alarm when the number in the queue exceeds some threshold). This complementary system is incorporated to cross-check the queuing system continuously to actually see what is happening. This is necessary because of inaccuracy of queuing model which is based on probability theory.

Collection of data samples: The number of customers arriving in a given interval of time from Monday to Friday is collected continuously, starting from first working day of the month. Service time duration per-customer and Wait time duration per-customer are collected. The average time a customer spends while waiting for service from the time the customer enters the system is collected. Customers monthly income and transaction volume which would be used for cost analysis are also collected.

The problem is formulated as a series of relationships in a mathematical model. A mathematical queuing model is used to estimate the probability of waiting time exceed-

ing a given threshold when the arrival rate and service rates are known. A model of the Banks customer-Teller scheduling system is produced such that the customers waiting time does not exceed the threshold which is determined by operations managers taking into account changing arrival rates and service time durations while balancing the service cost with waiting cost.

3.2. Mathematical model of the problem

For a single goal where some variables are subject to control, the general form of operations research model [7] is:

$$E = f(X_i, Y_j) \tag{1}$$

Where E is a measure of effectiveness of the system, f is a function, X_i represents controllable variables, and Y_j represents uncontrollable variables. This model can be classified either as an optimizing model or a simulation model. For instance waiting time = f (service time, inter arrival time). If this model is used as optimizing model, values inserted as X_i and for Y_j are manipulated to optimize the measure of effectiveness E. To optimize service in a Bank, the uncontrollable variables include: competitors, arrival rates, etc. The controllable variables include: number of Tellers required, commissions allowed, Teller service times etc. A set of values may be assumed for the uncontrollable variables because often, they cannot be known. Users can compute various Es until they find one they believe to be satisfactory.

A complete modeling of the system requires the following inputs:

(i) Customer arrival times in the queue (that is, detecting when a customer arrives, giving identity to the arriving customer and registering the time of the day this event occurs).

(ii) Tellers request for next customer and identifying the particular Teller that has made the request. Identify the next customer to be served. Note the time this service started and register the time of the day this event occurs.

(iii) Identifying the end of service event for each particular customer, and registering the time of the day this event occurs.

(iv) Detecting when the arrival rate, queue length, or time in the system exceeds or falls below some thresholds and registering the time of the day this event occurs.

Events are used to model the start and end of activities.

A. When an event occurs, the state of the system can change in the following ways: (i) The values of one or more variables can alter. (ii) The relationships among values assigned to one or more variables through file manipulation can alter.

B. The states of the system are measured by the following events: (i) A customer arrival event at the waiting line. (ii) Addition of a Teller if queue or arrival rate exceeds a threshold. (iii) Removal of a Teller if queue or arrival rate falls below a threshold. (iv) Start of service event at a station. (v) End of service event at a station.

C. This model as presented uses: (i) A switch at the entrance of the queue which customers on arrival use to indicate arrival. This action initiates the issuing of identity to the customer. (ii) A means to detect when a Teller calls customers for service and identify which customer from the FIFO queue whose turn it is to be served. Another means of detecting when service actually starts and ends for each customer. (iii) A means to detect when arrival rate, queue length, waiting time or time in the system exceeds or falls below the threshold. (iv) A clock to generate in real time

the time of the day when each event occurs. (v) A processing means which includes:

- Means for registering the time when a customer arrival is detected.
- Means for registering the time each particular customer is called for service by a Teller and registering the identity of the customer.
- Means for registering when a customer departs from the system.
- Means to provide logical comparison between the types of events and the related times registered by the registering means.
- Means to display the information.

3.3. Mathematical queuing models

Customer arrivals are at random and described by Poison distribution [8]. There exists a constant time λ (which is independent of queue length, time or any other property of the queue) such that the probability of an arrival during an interval Δt (where Δt is very small) is given by $\lambda\Delta t$. That is, the probability [of arrival between time t and $t + \Delta t$]

$$Pn = \lambda\Delta t \tag{2}$$

By using probability theory, one can obtain an expression for the probability (Pn) of having n number of customers in the system at any time t given the arrival rate λ and service rate μ. The service rate that will keep the waiting time below a threshold can also be obtained together with the cost of this service rate. The probability of waiting times exceeding a given threshold when λ and μ are known can be computed.

3.4. Determination of queuing parameters

Given the arrival rate λ of customers and the service rate μ of the Tellers in a Bank, the following parameters are determined: (i) Average number in queuing system (L). (ii) Average number in the queue (L_q). (iii) Average waiting time in whole system (W). (iv) Average waiting time in the queue (W_q).

These parameters are determined as follows:

3.4.1. Use of poison process

For a Poisson process with average arrival rate λ the probability of seeing n arrivals in time interval Δt is calculated mathematically by the equation:

$$P_r(n) = e^{-\lambda\Delta t}(\lambda\Delta t)^n/n! \tag{3}$$

Expected value or mean of n is:

$$E(n) = \lambda\Delta t. \tag{4}$$

The probability of no customer in the system $P_r(0)$ is:

$$\begin{aligned} P_r(0) &= (e^{-\lambda\Delta t}(\lambda\Delta t)^0)/0! \\ &= e^{-\lambda\Delta t} = 1 - \lambda\Delta t \end{aligned} \tag{5}$$

$$P_r(1) = \lambda\Delta t e^{-\lambda\Delta t} = \lambda\Delta t \tag{6}$$

The inter arrival time t (that is the time between arrivals) follows exponential distribution [9] with rate parameter λ. i.e. $P_r(t) = \lambda e^{-\lambda t}$. $E(t) = \lambda^{-1}$, Where $E(t)$ is the mean (expected) inter arrival time.

3.4.2. Queuing parameters for a system consisting of one queue and one Teller (m/m/1 model)

Given the mean inter arrival rate $\lambda - 1$ and mean service rate $\mu - 1$, where customers are served in first come first served order, the occupational rate ρ is mathematically:

$$\rho = \text{mean service rate} \div \text{mean inter arrival rate} = \lambda/\mu \tag{7}$$

(This is the fraction of time the server is busy). It is required that $\rho = \lambda/\mu < 1$. Otherwise the queue length explodes. Consider the time dependent behavior and the limiting behavior of the system:

Time dependent behavior of the System: Exponential distribution can be used to describe the state of the system at time t as follows: Let $Pn(t)$ be the probability that at time t, there are n customers in the system, where $n = 0,\ 1,\ 2\ \ldots$

$$\begin{aligned} P(x < t + \Delta t | x > t) &= 1 - e^{-\mu \Delta t} \\ &= \mu \Delta t + 0(\Delta t). \\ (\Delta t \to 0) & \end{aligned} \tag{8}$$

Therefore as $\Delta t \to 0$,

$$\begin{aligned} P_0(t + \Delta t) = &\ (1 - \lambda \Delta t) P_0(t) \\ &+ \mu \Delta t P_1(t) + 0(\Delta t) \end{aligned}$$

$$P_n(t + \Delta t) = \lambda \Delta t P_{n-1}(t) + (1 - (\lambda + \mu)\Delta t) P_n(t) + \mu \Delta t P_{n+1}(t) + 0(\Delta t)$$

By letting $\Delta t \to 0$, the following infinite set of differential equations are obtained for the probabilities $P_n(t)$:

$$P_0^1(t) = -\lambda P_0(t) + \mu P_1(t) \tag{9}$$

$$\begin{aligned} P_n^1(t) &= \lambda P_{n-1}(t) - (\lambda + \mu) P_n(t) \\ &\quad + \mu P_{n+1}(t).\ \ n = 1, 2, \ldots \end{aligned} \tag{10}$$

Limiting behavior of the System: If $P_n(t)$ is the probability of having n tasks in the system at time t, then, $P_0(t + \Delta t)$, is the probability of having no customer in the system at time interval $t + \Delta t$

$$\begin{aligned} &= P_0(t)[(1 - \mu \Delta t)(1 - \lambda \Delta t) + \mu \Delta t \lambda \Delta t] \\ &+ P_1(t)[(\mu \Delta t)(1 - \lambda \Delta t)] \end{aligned}$$

$$\begin{aligned} P_n(t + \Delta t) = &\ P_n(t)[(1 - \mu \Delta t)(1 - \lambda \Delta t) \\ &+ \mu t + \lambda t] + P_{n+1}(t)[(\mu \Delta t)(1 - \lambda \Delta t)] \\ &+ P_{n-1}(t)[(\lambda \Delta t)(1 - \mu \Delta t)] \end{aligned}$$

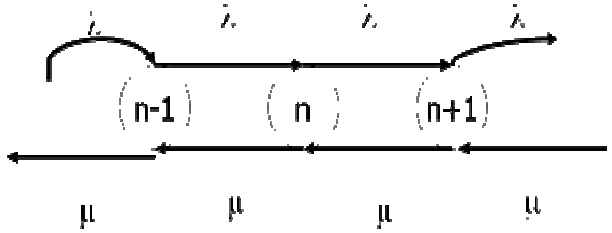

Figure 1: State transition diagram for the equilibrium conditions of M/M/1 queue.

$$\begin{aligned} &[P_0(t + \Delta t) - P_0(t)] \div \Delta t = \\ &\qquad -\lambda P_0(t) + \mu P_1(t) \end{aligned}$$

$$\begin{aligned} &[P_n(t + \Delta t) - P_n(t)] \div \Delta t = \\ &\lambda P_{n-1}(t)(\lambda + \mu) P_n(t) + \mu P_{n+1}(t) \end{aligned}$$

For the system to be stabilized the arrival rate must be less than or equal to the service rate [9] ($\lambda \leq \mu$). As $t \to \infty$, $P_n(t) \to 0$ and $P_n(t) \to Pn$. That is, $\lim_{t \to \infty} P_n(t) = P_n$ and $\lim_{t \to \infty} [P_n(t + \Delta t) - P_n(t)] \div \Delta t = 0$.

From (9), the limiting or equilibrium probabilities P_n satisfy the equations:

$$0 = -\lambda P_0 + \mu P_1 \tag{11}$$

$$\begin{aligned} 0 &= \lambda P_{n-1} - (\lambda + \mu) P_n + \mu P_{n+1}, \\ &\qquad n = 1, 2.3 \end{aligned} \tag{12}$$

Also the probabilities P_n also satisfy the normalization equation

$$\sum_{n}^{\infty} P_n = 1 \tag{13}$$

The arrows indicate possible transitions. The rate at which transition occurs is λ for a transition from n to $n+1$ (an arrival) and μ for a transition from $n+1$ to n (a departure). The number of transitions per unit time from n to $n+1$ (the flow from n to $n+1$) is equal to P_n, (the fraction of time the system is in state n) multiplied by λ (the rate at which arrivals occur while the system is in state n). One can derive the equilibrium equations (11) and (12)

by equating the flow out of state n to the flow into state n.

Solution to the equilibrium equations of the M/M/1 queue: The solutions to equations (11) and (12) can be obtained by any of the following approach: direct approach, recursion, global balance principle and the use of littles law. For the purpose of this model, the direct approach and Littles law are used.

Direct Approach

Equations (11) and (12) are a second order recurrence relation with constant coefficients. There general solution is of the form:

$$P_n = C_1 X_1^n + C_2 X_2^n \tag{14}$$

Where $n = 0,\ 1,\ 2,\ \ldots,\ X_1$ and X_2 are roots of the quadratic equation $\lambda - (\lambda + \mu)x + \mu x^2 = 0$. (See equation (11)). Consider now equation (12),

$$0 = \lambda P_{n-1} - (\lambda + \lambda)P_n + \mu P_{n+1} \quad \text{where } n = 1,\ 2,\ \ldots$$

Thus for $n = 1$, $0 = \lambda P_0 - (\lambda + \mu)P_1 + \mu P_2)$.

The equation has two zeros (at $x = 1$, and $x = \lambda/\mu = \rho$), so that all solutions to (12) are of the form $P_n = C_1 + C_2\rho^n$. Where $n = 0,\ 1,\ 2,\ \ldots$

Applying equation (13), means that C_1 must be equal to 0. That $C_1 = 0$ also follows from (11) by substituting the solution (13) into (11).The coefficient C_2 finally follows from the normalization equation (11), yielding that $C_2 = 1 - \rho$. Thus

$$P_n = (1 - \rho)\rho^n (\text{where } n = 0, 1, 2\ \ldots) \tag{15}$$

Therefore the equilibrium distribution depends upon λ and μ only through their ratio ρ. From (10), P_1 can be expressed in terms of P_0 as follow $0 = \lambda P_0 + \mu P_1$. Therefore,

$$P_1 = \lambda/\mu P_0 = \rho P_0 \tag{16}$$

Substituting P_1 in equation (12) gives: $0 = \lambda P_{n-1} - (\lambda + \mu)P_n + \mu P_{n+1} \quad n = 1,\ 2,\ 3,\ \ldots$. For $n = 1$, $0 = \lambda P_0 - (\lambda + \mu)P_1 + \mu P_2$. But $P_1 = \rho P_0$, and $\rho = \lambda/\mu$. Thus $0 = \lambda P_0 - (\lambda\rho P_0 + \mu\rho P_0) + \mu P_2$, and so, $P_2 = (\lambda/\mu)P_0\rho = \rho^2 P_0$.

One can also find $P_3,\ P_4,\ \ldots$ for $n = 3,\ 4\ \ldots$. Thus all probabilities can be expressed in terms of P_0. That is

$$P_n = \rho^n P_0, \qquad n = 0,\ 1,\ 2,\ \ldots \tag{17}$$

P_0 can be found from the normalization equation (13).

Littles Law

Little's law gives a relation between $E(L)$ - the mean number of customers in the system, $E(S)$ - the mean sojourn time and λ - the average number of customers entering the system per unit time. The law states that

$$E(L) = \lambda E(S) \tag{18}$$

This law applies to any system at equilibrium as long as nothing is creating or destroying tasks in the system [9]. It is assumed that the number of customers in the system does not grow to infinity. Application of Little's law to the system consisting of queue plus server yields relation (18). Application of Littles law to the queue (excluding the server) yields a relation between the queue length L_q and the waiting time W. That is:

$$E(L_q) = \lambda E(W) \tag{19}$$

Let the average number of customers in the system be L, the average time customer is in the system be W, the total time be T, the number of customers be N, then:

$$\begin{aligned} \lambda &= N/T \\ TL &= NW \end{aligned} \tag{20}$$

$$\begin{aligned} &L = \tfrac{N}{T}W \qquad \text{or } L = (\lambda)W \\ &\text{Also } L_q = \lambda W q \quad \text{(Steady state condition)} \end{aligned} \tag{21}$$

$$W = Wq + 1/\mu$$

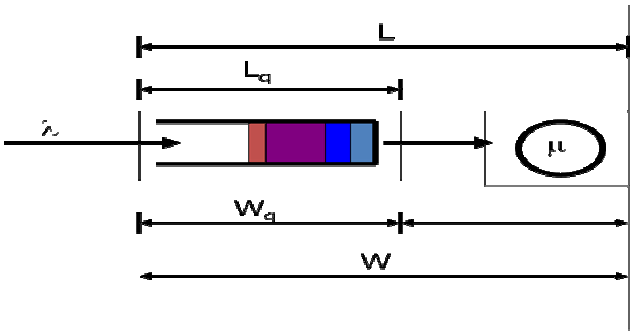

Figure 2: A queuing system.

The mean number of customers in the system is

$$L = E(L_q) = \sum_{n=0}^{\infty} nPn \tag{22}$$

But from (15), $P_n = (1-\rho)\rho^n$, where $n = 0, 1, 2, \ldots$

Thus

$$\begin{aligned} \sum_{n=0}^{\infty} nPn &= \sum_{n=0}^{\infty} n\rho n\,(1-\rho) \\ &= (1-\rho)\rho \sum_{n=1}^{\infty} n\rho^{n-1} \\ &= \rho/(1-\rho) = \lambda/(\mu-\lambda) \end{aligned}$$

Solving for W yields that

$$W = \frac{L}{\lambda} = \frac{\lambda}{\mu-\lambda}\frac{1}{\lambda} = \frac{1}{\mu-\lambda}$$

Solving for $E(s)$ by applying littles law [9] gives: $E(s) = \dfrac{1/\mu}{1-\mu}$

$$W_q = W - \frac{1}{\mu} = \frac{\lambda}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$$

$$L_q = \lambda W_q = \lambda\left[\frac{\lambda}{\mu(\mu-\lambda)}\right] = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

3.4.3. A system of one Queue and more than one teller (m/m/c)

For a system with parallel identical Tellers, exponential inter-arrival time and service time with mean $\lambda-1$ and $\mu-1$ respectively, where c is the number of tellers, the occupational rate per Teller [9] is:

$$\rho = \lambda/c\mu \tag{23}$$

To avoid the queue length explosion, it is required that $\rho < 1$. The expressions for the expected waiting time and queue lengths are fairly complicated and depend on the probability of there being no jobs in the system upon arrival, P_0, and on the probability of a job having to wait upon arrival, P_q.

$$P_o = \left[\sum_{k=0}^{c-1}\left(\frac{(c\rho)^k}{k!}\right) + \frac{(c\rho)^c}{c!}(1-\rho)\right]^{-1} \tag{24}$$

$$P_q = P_o(c\rho)^c/c!(1-\rho) \tag{25}$$

The mean waiting time:

$$EWq = \rho * P_q/(\lambda(1-\rho)) \tag{26}$$

The mean queue length:

$$ELq = \rho * P_q/(1-\rho) \tag{27}$$

The mean system time:

$$EW = (1/\mu) + \rho * P_q/(\lambda(1-\rho)) \tag{28}$$

The mean number in system:

$$EL = (c*\rho) + (\rho * P_q)/(1-\rho) \tag{29}$$

The waiting time distribution for the M/M/c system is also complicated and given by:

$$W_q(t) = \begin{cases} 1 - \frac{c(\lambda/\mu)^c}{c!(c-\lambda/\mu)}P_o & t = 0 \\ (\frac{\lambda}{\mu})^c \frac{1-e^{-(\mu c-\lambda)t}}{(c-1)!}(c-\frac{\lambda}{\mu})p_0 & \\ +W_q(0) & t > 0 \end{cases} \tag{30}$$

3.5. Cost structure and Types

There are two types of costs:

1. Cost of providing service: Service cost is directly proportional to service rate - as service throughput increases, cost of providing additional Tellers is added per throughput increase. Waiting time is inversely proportional to service rate-as service throughput increases, waiting time decreases.

2. Waiting cost This could be balking cost, cost of Teller idle times, or cost for the times customers wait or spend in the system. Waiting cost is directly proportional to wait times.

These two types of cost work in opposite direction. It is to be noted that as service and service cost increases, waiting cost typically decline. The Bank has to balance these two costs. Generally, managers will want to minimize cost. The two objectives: service cost (operational cost) and waiting cost (customer satisfaction) are therefore to be optimized.

3.5.1. Balancing cost of service with cost of waiting

The total cost can be computed in two ways:

$$\text{Cost} = \text{service cost} + (\text{number waiting x wait cost}) \quad (31)$$

$$\text{Cost} = \text{service cost} + (\text{number in system x wait cost}) \quad (32)$$

These costs are calculated using various numbers of Tellers. The total cost based on the number of customers waiting is compared with the total cost based on the number of customers in the system. The number of Tellers that give the least cost in both cases is the optimum.

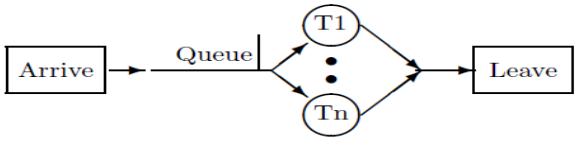

Figure 3: Customer-Teller Model (T represents Teller).

3.6. Model Algorithm

(i) Customers arrival is indicated when the customer depresses an arrival switch. The depressed switch corresponds with the volume of transaction the customer intends to transact. Customer transactions are segmented into volumes. For instance, customers for transactions of less than a certain amount are grouped into one queue with a corresponding arrival switch, and Tellers provided for them, while customers for transactions above the amount are placed on a different queue. This is because time to serve any customer greatly depends on the amount of money being deposited or withdrawn. A high volume transaction takes longer time than lower volume transactions. For instance assuming there are 5 customers in group A, with service rates of 5 minutes and 10 customers in group B with service rates of 20 minutes. The total time for group A customers is 50 minutes, while it is 200 minutes for group B. If both the group A and B customers are kept together on one queue, and served on first come first served order, their combined mean service rate will be 16.7. This means that every customer in the system spends 16.7 minutes i.e. 6.7 minutes extra for group A customers. Thus the 5 customers in group A will jointly spend 33.5 minutes extra. This theoretically looks as if there are about 3 more customers in the queue. This may lead to wrong computations of the number of Tellers required.

(ii) On depression of the switch, a processor generates an identity in the form of se-

rial (queue) number for the arriving customer. This triggers a card dispenser (printer) to produce the serial number for the customer. A real time clock is used by the processor to keep record of the time and date the event occurred. This time is recoded against the serial number of the customer. This customer is then placed in a FIFO Entrance. To avoid confusion, different numbering pattern is used for different volumes of transactions. For instance the first arriving customer in transaction volume A is given the Serial number A001, the second, A002, and so on. That of transaction volume B is given B001, B002.... The inter arrival time is calculated by subtracting the arrival time of the immediately preceding customers arrival time from the arrival time of a new arrival.

(iii) At the service stations, a Teller calls customers for service by the use of a service switch. When a Teller depresses this switch, the processor recognizes this time as the end of service for the customer in service and a call for service for the next customer at the head of the queue. The serial number of the customer at the head of the queue is sent to a voice prompt which announces this to the waiting customers in the waiting room. When a customer gets to the Teller, a card reader reads the serial number on the card. The processor by using the real time clock records the time and date this card was read and records this in memory against the customers serial number. This time indicates the start of service for that particular customer. The customers waiting time is calculated by subtracting the arrival time of each customer from the start of service time of each customer. The result is kept in memory against the corresponding serial number of the customers.

(iv) Depressing the service switch another time in successions indicates an end of service for the customer in service and a request for another customer to be served. The processor registers this time against the customer as the customers end of service time. A signal is made to the voice prompt; this contains the serial number of the next customer at the FIFO queue which the voice prompt announces. The service duration of this departing customer is calculated by subtracting the time the serial number was read from the time the service switch was depressed a second time. The result is registered against the departing customer as his service duration. The total time a customer spent in the system is calculated by subtracting the customers arrival time from the departure time.

(v) At a certain time interval, check is made of any customer that did not come up when called for service (such a customer is assumed to have balked or reneged).The total number of balked customers is registered. At the given interval, the times and types of events registered (that is, the mean arrival rate, mean time in the queue, mean service duration, and average time spent in the system) are computed. The results are displayed or printed out. The results are used to analyze the behavior of the system so as to adjust it to meet optimum service goals. The processor uses the results obtained to compare the arrival rate, queue length or waiting time, with the organizations goals (that is, the threshold levels), so that alarm is raised when these goals are not met so that service levels are adjusted. The number of Tellers to use is computed by balancing the service cost with the waiting cost as indicated in section 3.5.1.

3.7. Answers to research questions

The following are the reasons why queues form:

(i) Queues result when customers wait for service or servers wait for customers. Customers are forced to wait when the number

of service facilities could not take care of the increasing demand for service. The arriving customers do not receive service immediately on arrival and upon request but must have to wait. Servers are forced to wait when the number of servers exceeds the number of customers some servers must remain idle or unutilized.

(ii) Queues form when the arrival rate is faster than the service rate. If the arrival rate is less than the minimum service rate, the queue length varies in size but reaches steady state. If the arrival rate is greater than the maximum service rate, the system never reaches steady state and the queue length is continually increasing.

Factors that Determine Queue Length are: (i) Probability distribution of arrivals. (ii) Probability distribution of service times. (iii) Number of service channels. (iv) Line discipline.

The following factors make queue length to increase: (i) Competitive price may attract more customers. (ii) Better (improved or faster) service attracts customers. (iii) Rush periods. (iv) Sudden demand for special service (example students payment for school fees).

Factors that may increase waiting time are: (i) Shunting of queue i.e. truncating the FIFO discipline. (ii) Distraction of Tellers, example un-necessary interactions with a customer etc.

4. Model simulation and results

The operational procedure and results are as follows:

(A) The system keeps track of every customer that arrives in a Bank with a view to obtaining the following information; (i) Arrival time, (ii) Inter-arrival time. It performs the following functions: Gives identity to the arriving customers, this is given sequentially as the customers arrive based on the time of each arrival. Example, if the first customer arrives at 9.00am, this customer is given serial number 1.This means that the customers identity is 1, and the position on the FIFO queue is also 1. Place the customers in a First-In First-Out Queue as they arrive. (iii) The system also obtains: the time each customer starts to receive service, the time each customer waits in the queue before receiving service, the time service ends for each customer, the service time duration for each customer and the total time each customer spends in the system.

(B) The system uses the information obtained to compute at every given time interval the following parameters: The total number of customers that have arrived at the given time period for service, the total number of customers that have been served, the average inter-arrival time, the total number of customers that arrived at that given time period but are yet to receive service, the total number of customers that have balked (that is, the number of customers that arrived but did not wait to receive service), the mean wait time, the mean service time and the mean time in the system.

With the above information, a threshold is set so that if the mean wait time, mean arrival time, or the mean time in the system exceeds or falls below the threshold, the system raises alarm so that management can take action. A software package is developed to capture the above requirement specifications as a model using software development tools.

Customer action

An arriving customers first contact with the system is the customer user interface. The customer is asked on the screen to click on the type of transaction required. (Switches are used in place of the screen, each switch corresponding to the type or volume of trans-

action). The customer clicks on the type of transaction required. The system recognizes this click as an arrival of a new customer. It then generates the arrival time and serial number for this new customer.

The following are the actions of the customers and the response action of the processing system:

- Stage1; Initial screen display for the arriving customers is *Please click on the required volume of transaction to join the queue.*
- Stage2; On clicking on the required volume of transaction, the display changes to *Welcome, your queue number is (*queue number*) arrival time is (*arrival time).
- Stage3; The queue number and arrival time is then printed out by the printer.
- Stage4; The display initializes to stage 1.

Internal system operations

Customer action: Clicks on arrival switch. **Processor action:** The processor performs the following actions: Record the time of clicking the arrival switch event. Set this time as arrival time of new customer. Give a serial number to the click that is, this arrival time (this number represents the identity of the customer and the customers position on the queue). Place the customer on a FIFO queue. Calculate the inter-arrival time. Register the customer in memory. Send information to the printer or display.

Example:

Time of first click = 9.05am. Arrival time of first customer = 9.05am. Serial number/position of Customer in the queue = 1. Customer identity = 1. Inter-arrival time = 0. Information displayed or printed out to the Customer: *Welcome, your arrival Time is 9.05am, your queue number is 1.*

Time of second click = 9.08am. Arrival time of second Customer = 9.08am. Serial number and position of Customer in the queue = 2. Customer identity = 2. Inter-arrival time = time of n^{th} click minus time of $(n-1)^{th}$ click. (For n=2, n-1 = 1). Therefore inter-arrival time is 9.08 - 9.05 = 3minutes. Information displayed or printed out to the customer: *Welcome, your arrival Time is 9.08am, your queue number is 2.*

Teller Actions and Processor responses:

Teller: A Teller clicks on the service switch for the first time. **Processor:** Records the time this event occurred. Call the customer at the head of the queue for service. Send the queue number of this customer at the head of the queue for announcement by the voice prompt. **Customer action:** Customer number 1 approaches Teller. **Teller action:** The Teller enters the serial number of the customer in the space provided and clicks enter. **Processor action:** Record the time the serial number was entered as the start of service time for the customer whose serial number was entered. Compute wait time for the customer in service. Wait time = time of start service minus time of arrival. **Teller action:** The Teller clicks the service switch a second time. Processor action: Record the time this event occurred. Assign this time as the departure time (that is, end of service for customer in service). Calculate the service time duration of the departing customer (service duration = departure time minus start service time for the particular customer). Call the next customer at the head of the queue for service. Send the queue number of the customer at the head of the queue for announcement by the voice prompt.

Example: Time to enter serial number 1 for service = 9.06am. Time for Teller to click ser-

vice switch a second time = 9.10am. Departure time for last customer = 9.10am. Service time duration for customer number 1 = departure time minus start of service time (that is 9.10 - 9.06 = 4 minutes). Wait time duration for customer number 1= 9.06 9.05 = 1 minute. Time in the system for customer number 1 = Departure time minus Arrival time (that is, 9.10-9.05 = 5 minutes).

Print out or Displayed information: Customer identity = 1. Arrival time = 9.05am. Start service time = 9.06am. Wait time duration = 1 minute. Departure time = 9.10am. Service time duration = 4minutes. Time in the system = 5minutes. Next customer for service = *Customers number.*

4.1. Results

System model condition at the end of a given period

Table 1 is a simulation result for 30 minutes (based on manual timing of Customer-Teller transactions in a local Bank). The table shows that at 30 minutes interval, the following queuing parameters are obtained; the arrival rate λ is 15 customers every 30minutes(or $15/30 = 0.5$), while the service rate μ is 6 customers every 30 minutes (or $6/30 = 0.2$).This data is collected daily for one month so that average data can be computed for each period. To compute the system parameters, one uses the formulae derived in sections 3.3-3.4.3, in particular, the mean queue length, mean number in the system, and the occupational rate. $\rho = \lambda/c\mu$ and $W_q = L_q/\lambda$. Where: L_q = Average queue length, W_q = Average waiting time, ρ, λ, and μ maintain their usual meanings. Weekly data can be used for future prediction of the behavior of the system. Using analytical methods, the waiting time for various queuing systems are computed and used to relate a given waiting time to an occupational rate ρ (see Table 2). By keeping a record of the number of arrivals observed during each given time interval, and by keeping the average service time for each day of the week, one computes the number of Tellers needed.

Table 2: Number of Tellers with corresponding occupational rates, mean waiting time of customers, and mean time in the system.

Number of Tellers	**Occupational Rate**	**Mean waiting time**	**Mean time in system**
1	2.5	-1.67	-4.17
2	1.25	0.68	9.8
3	0.83	3.71	6.2
4	0.63	0.55	3.08
5	0.5	0.07	2.57

Table 3: Number of Tellers with corresponding occupational rates, mean waiting time of customers, and mean time in the system.

Number of Tellers	**Total cost based on mean waiting time**	**Total cost based on mean time in system**
1	-4.99	-2.9
2	22.04	49.40
3	41.13	48.60
4	41.65	49.54
5	50.21	57.71

Assuming that it costs 10 units to keep a Teller for the time period and wait cost is 3 units then Table 3 gives the costs for various numbers of Tellers using equations (31) and (32). It can be seen from Table 3 that three Tellers would be the number of Tellers that would give the lowest cost based on waiting and service. Three Tellers also gave the highest Teller utilization (83%) (See table 2).This is the optimum service condition. Two Tellers or lower cannot be used because the occupational rates for two Tellers and below are

Table 1: Result obtained from manually timing Customer-Teller transactions in a local Bank.

Customer Identity	Arrival Time, A	Inter-Arrival time, B	Start Service Time, C	Wait Time Duration, D $(C-A)$	Departure time, E	Service Time duration, F $(E-C)$	Time in the system, G $(E-A)$
1	9.05		9.06	1	9.10	4	5
2	9.06	1	9.10	4	9.15	5	9
3	9.08	2	9.15	7	9.19	4	11
4	9.09	1	9.19	10	9.22	3	13
5	9.10	1	9.22	12	9.26	4	16
6	9.12	2	9.26	14	9.30	4	18
7	9.13	1	9.31	18	9.36	4	23
8	9.15	2	9.36	21	9.40	4	25
9	9.18	3	9.40	22	9.43	3	25
10	9.20	2	9.43	23	9.48	5	28
11	9.21	1	9.48				
12	9.24	3					
13	9.26	2					
14	9.28	2					
15	9.30	2					

greater than 1, see equation (23). Using one or two Tellers at the same arrival rate will make the queue length to explode. Four Tellers and above would give higher costs based on both waiting and service. This Bank can start with three Tellers. Mean waiting Time and Mean Time in the system thresholds can be set at 3.71 and 6.2 respectively. If at the given interval say 30 minutes these thresholds are exceeded or not reached, the system raises alarm. By so doing control is placed on the system so that Tellers are added or removed as the case maybe. This guarantees that queue length will not explode nor any Teller kept unnecessarily idle. To determine the balked customers, every arriving customer is required to pick a queue number before knowing the queue length or entering the Bank. Consider another system as shown in table 4 which is the result obtained using a random time generator.

For a period of 37 minutes the arrival rate is 0.2703 or 10 customers every 37 minutes, while the service rate is 0.2432 or 9 customers every 37 minutes. In this case, Table 5 shows the occupational rate per Teller with the corresponding wait time and time in the system while Table 6 provides the cost analysis.

It can be seen that two Tellers is the ideal for the system in table 4 over the period considered. This corresponds to lower cost and

Table 5: Number of Tellers with corresponding occupational rates, mean waiting time of customers, and mean time in the system for the system in Table 4

Number of Tellers	Occupational Rate	Mean waiting time	Mean time in system
1	1.1110	-9.97	-36.9
2	0.5555	1.3135	5.425
3	0.3703	2.1750	6.287
4	0.2778	2.3340	6.446
5	0.2222	0.0062	4.118

Table 6: Total costs with corresponding number of Tellers for the system in table 4.

Number of Tellers	Cost of service	Total cost based on waiting	Total cost based on system
1	370	340	262
2	740	744	756
3	1110	1117	1129
4	1480	1487	1499
5	1850	1850	1862

Table 4: Results obtained by using a random time generator.

Customer ID	Arrival Time	Inter-Arrival time	Start of Service Time	Wait Time	Departure time	Service Time	Time in the system
1	3.2		3.2	0.0	7.0	3.8	3.8
2	10.9	7.7	10.9	0.0	14.4	3.5	3.5
3	13.2	2.3	14.4	1.2	18.6	4.2	5.4
4	14.8	1.6	18.6	3.8	21.7	3.1	6.9
5	17.7	2.9	21.7	4.0	24.1	2.4	6.4
6	19.8	2.1	24.1	4.3	28.4	4.3	8.6
7	21.5	1.7	28.4	6.9	31.1	2.7	9.6
8	26.3	4.8	31.8	4.8	33.2	2.1	6.9
9	32.1	5.8	33.2	1.1	35.7	2.5	3.6
10	36.6	4.5	36.6	0.0	40.0	3.4	3.4

high Teller Utilization.

4.2. Supervisory control system - Program Algorithm

5. Conclusion

A scheme for scheduling Tellers in Banks has been developed as discussed. This scheme optimizes Customer-Teller service. Due to high variable arrival rates, it is not sufficient to forecast the system conditions using queuing theory and simulation results only. In addition, the system should also be monitored to see that deviations in the forecasted values are compensated immediately. This avoids unnecessary costs. These deficiencies in hitherto presented models are what this model has solved. It seeks to reduce the waiting Time and hence the queue length by optimizing the number of Tellers at a reduced cost. This will build up the customer loyalty, who will return for similar business in the future. The analysis of the result can be used to improve the flow of customers through the queuing system (Bank).This would speed up services and provide competitive edge.

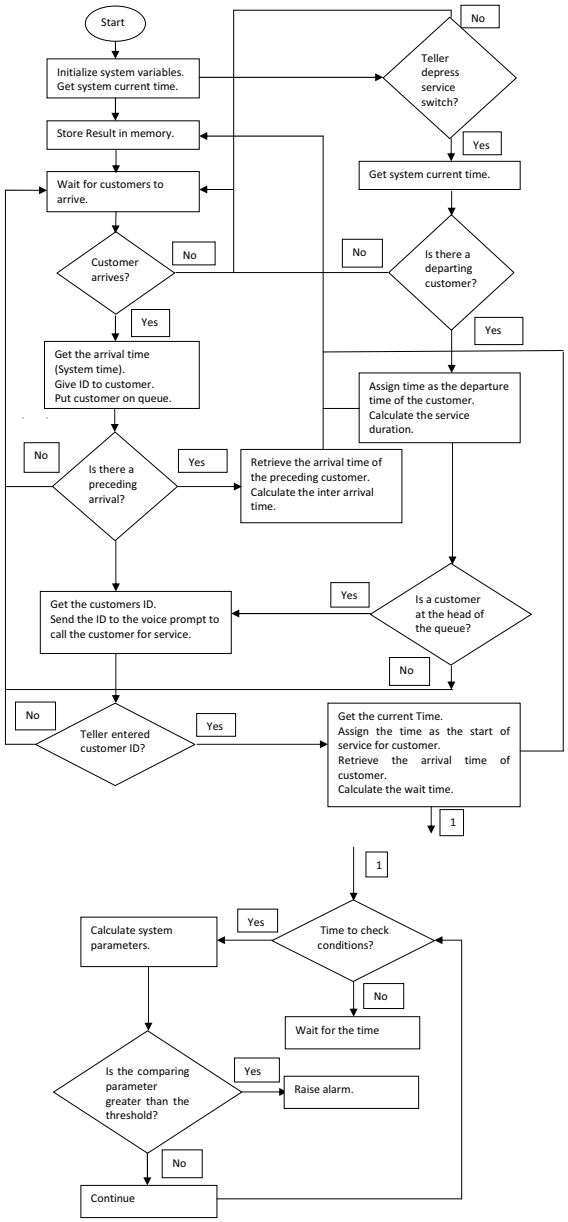

Figure 4: Supervisory control system - Program Algorithm

References

1. Thompson G. Accounting for the multi period impact of science when determining employee requirement for labour scheduling. *Journal of Operations Management*, vol. 11, 1993, pp. 269 287.

2. Martha A.C., Ronald G., Richard L., and Abdulla M.I. A Simulation ILP Based Tool for Scheduling ER Staff. In *Proceeding of the winter Simulation conference*, 2003, p. 1930.

3. Hammond D. A Simulation and Analysis of Bank Teller Manning. In *Proceeding of Winter Simulation Conference*, 1995, pp. 1077 1080.

4. Travis C., and Micheal M. Analysis of Teller service Times in Retail Banks. *CS BIGS* 1(1), 2007, pp. 15 25.

5. Katsof et al., *System and Method for forecasting Bank Traffic and Scheduling Work Assignment for Bank Personnel.* United State Patent 4,700,295., www.patentstorm.us/patents/., 1987.

6. Marcel D., Miguel A.M. *Monitoring and Forecasting Customer Traffic.* United State Patent 5,541,835., www.patentstorm.us/patents/ ., 1992.

7. Harold K., and Cyril O. *A System and Contingency Analysis of Managerial Function.* McGraw - Hill Kogakusha Ltd, Sixth Edition, 1992, chapter 5.

8. Latouche G., and Ramaswami V. *Introduction to Matrix Analytic Methods in Stochastic Modeling, Quassi - Birth - and - Death process.* ASA SIAM, First Edition, 1999.

9. Ivo A., and Jacques R., *Queuing Theory.* Department of Computing Science, Eindhoven University of Technology, Eindhoven, The Netherland, 2001, pp.11 48.

10. Balakrshnan N., and Basu A.P. *The Exponential distribution, Theory, Methods and Application.* Gordon and Breach publication, 1996.

11. Ronald A.N., and James P.W. Queuing Theory and Customer Satisfaction, A review of Terminology, Trends and Application to Pharmacy Practice. *Hospital Pharmacy Journal*, vol. 36, 2001, pp. 275 279.