Simulation approach to model queuing Problems

Conference Paper · December 2011

1 author:

Saad Talib Hasson Aljebori
University of Babylon

Simulation approach to model queuing Problems

Prof. Dr. Fadhil M. Al- Rubaee Assist. Prof. Dr. Saad Talib Hasson
University of Babylon

Abstract

Queuing theory utilizes mathematical analysis to determine the systems measures of effectiveness. Such Important and effective measures are then used as data in building an optimization model for determining the system efficiency and its requirements. In this paper we will discuss and provide the steps of building the computer simulation model and the necessary mathematical algorithms to analyze the complex service network as queuing systems.

Each queuing system is a typical problem of discrete event system, and the computer simulation becomes a quite effective way for solving the queuing problems and analyzing its performances in a feasible manner. Queuing techniques are differs from other mathematical techniques of Operations Research (OR) in that it does not deal with optimization.

Computer simulation is the feasible imitation of the operation of a real-world process or system over time. It represents a very helpful and valuable tool to analyze and evaluate each available or developed process. It can be used to learn, test and evaluate the behavior of the many field's problems. Simulation provides low cost, secure and fast analysis tool with required possible sensitivity analysis.

In this paper Markov chains were used to model the inputs and processing parameters required to perform a valid queuing network simulation. A method for modeling the dependencies and balances the required network parameters will be proposed.

Many configurations of production lines using Queuing approaches were suggested in this paper. Required Computer program was developed to evaluate each model (of the suggested systems) using stochastic random variables generated through scientific sampling process to be used as data to operate the models.

I. INTRODUCTION

One of our indications that the scientific world researches are directed towards services rather than designing and production industries. Queuing models are a versatile tool for investigating various characteristics either: in their project planning stage or; during their exploitation and extension of the service networks [Cooper, R. P.1981]. It was used to represent the properties and the behavior of each network with simulation tools. Simulation can offer a feasible mission to better understand the expected performance of the real system and to test any change with its effectiveness on the system design. Markov chains are discrete state space stochastic processes with an interesting feature and, the probabilities of the future states of a Markov chain depend only on their current probabilities [Gunter Bolch et al 2006].

Simulation has become an important and useful tool with the growing need better understand the behavior of complex systems. It was defined as an"*experimental and applied* *methodology"*. While the computer simulation represents a method that demonstrates dynamically the structure and the behaviors of the system with computer in order to evaluate and predict the effect of the behaviors of some system and provide information for decision. Simulation was used in many areas including, industries, medicine, mechanical and electrical design, engineering and service organizations, transportation systems, global systems, social and behavioral sciences [A. Horv´ath, 2000].

Simulation is always used to help decision makers and designers in their better understanding to the expected performance of the real system to test the effectiveness of the system design. Performance evaluation process represents the essential activity in design of a new service system as well as in the case of tuning an operating one. Performance evaluation of a service system can be implemented at a discrete event level. It can be employed and imitated through building or developing a suitable valid discrete event simulation model. This research will focuses on queuing theory and simulation as an effective tools to model the behavior of the service networks in order to optimize its performance without exceeding the given constraints. Its significance is to develop a simulation model to measure the performance of the suggested network utilization.

A Model is an abstraction of the system obtained by making a set of assumptions about how the system works. It must capture the essential characteristics of the system. When the system is being in the planning stage the model is important to test its performance and to test the effects of each variable on the system.

II. PERFORMANCE EVALUATION

The performance evaluation process is used to measure or estimate the system behavior through the implementation phase of its model. Each built or developed valid model can effectively be used for system performance prediction through planning, designing or operating stages. One Performance measure approach is to obtain the data by observing the events and activities on an existing system. Performance modeling means represent the system by a model and manipulate the model to obtain information about its behavior and performance.

Queuing theory offers reliable and acceptable estimation as an average value to each of the following important performance measure parameters which helps in decision making about each process [Lammer S, and Helbing D, 2008]; .

- The Server utilization (ρ); describes the fraction of time that the server is busy, or the mean fraction of active servers, in the case of multiple servers.
- Throughput (λ) describes the number of jobs, whose processing is completed in a single unit of time.
- Queue length is the number of jobs waiting in the queue at a given time.
- Waiting time (W) is the time that the jobs spend in the queue waiting to be served.
- The number of (k) jobs in the system at a given time.
- The probability of a given number of jobs i in the system (p_i) .

The most and important fundamental technique for the performance evaluation is the correct and reliable measurements. It commonly used to verify that a system meets its designed specifications and not exceed the performance tolerances. Measurements are used as a tool for the analytical and simulation modeling to obtain parametric values for the models to implement and validate the results.

III. QUEUING PROBLEMS

Queuing Theory is a scientific approach to analyze waiting lines. It was applied to many situations in which customers arrive at a system, wait, and receive service. The main objectives of queuing theory are to improve customer service and reduce the system operating costs. Queuing problem is a problem that deals with a balance between average waiting time, and idle time of the server. It helps to improve queuing facility for each entity and server [Cooper, R. P.,1981, Kishor S. Trivedi, 2001].

Due to the fast technology development the use of the simple Queues was chained to form queuing networks in which the departures from one queue enter the next queue. Queuing networks can be either open or closed. It has a wide application in real life such as planning process, manufacturing systems, industries, project management, computer networks, telecommunications, transport, and logistics. The probability and statistical methods are the most frequently tools that used in these systems performance determination. In queuing process the customers (not necessarily human) are arriving for service, waiting for service if it is not immediate, and leaving the system as soon as they are served [Cooper, R. P., 1981, Kishor S. Trivedi, 2001, Sheldon M. Ross, 2001].

There are six basic characteristics of queuing processes which provide an adequate description of a queuing system: (1) arrival pattern, (2) service pattern, (3) number of service channels, (4) system capacity, (5) the population size and, (6) queue discipline. Queuing theory used to measure the following important quantitative decision indications about the quality of service provided in each queuing system: (1) Waiting time in the queue, (2) Total time in the system required to complete the wanted service, (3) Completion by a deadline, (5) Average queue length, (6) Average number of customers in the system (customers in queue plus customers in service), (7) The rate at which customers are served, (8) Server utilization, (9) Percentage of lost customers [Curry, G., et al., 2003 ,Sheldon M. Ross, 2001].

IV. MARKOV CHAINS

Markov chains are becomes the most powerful analytic techniques for evaluating complex system performance. Most of the real-world systems contain uncertainty and evolve over time. Stochastic processes and Markov chains are the most suitable probability models for such systems. As it was known that the random process is called a Markov Process if, conditional on the current state of the process, its future is independent of its past. A Markov chain is a mathematical model for stochastic systems whose states, discrete or continuous, are governed by transition probability. Since all the states of the real systems changes randomly, it is impossible to predict the exact state of the system in the future. The future statistical properties of the system's can be predicted which are important in many applications. Transitions means the change of the state of the system, and the probabilities associated with various state-changes are called transition probabilities. The collection of the states and transition probabilities forms a Markov chain [Gunter Bolch et al , 2006, Nelson R., 2000].

Most of the queuing models are in fact represents Markov processes. There are a large number of real engineering, industrial, physical, biological, economics, and social phenomena that can be described and analyzed as Markov chains. To represent the behavior of a queuing network as a Markov chain first one have to choose a state space representation.

A state can be represented as a vector whose components described completely the state of each of the elements of the queuing network [Gunter Bolch et al , 2006]

V. PROBLEM FORMULATION

Several services models were developed in order to analyze different queuing dispatching and processing conditions. The main objectives here are: (1) Evaluate several layout production models, (2) Show the difference between these models in queuing, waiting (delay) time and idle time, and (3) Show which of this models is best than the other [A. Horv´ath, 2000].

Our suggested design depends on the fact that the status of the system changes each time an event occurs. During the time that elapses between two successive events, the system's status remains unchanged. In view of this, it suffices to monitor the changes in the system's status. In order to implement this idea, each event is associated with a clock. The value of this clock gives the time instance in the future that this event will occur. The simulation model, upon completion of processing an event, say at time t1, regroups all the possible events that will occur in the future and finds the one with the smallest clock value. It then advances the time, i.e., the master clock, to this particular time when the next event will occur, say at time t2. It takes appropriate action as dictated by the occurrence of this event, and then repeats the process of finding the next event (say at time t3). The simulation model, therefore, moves through time by simply visiting the time instances at which events occur. In view of this it is known as *event-advance design* [A. T. Andersen, 1998].

VI. MODELING PROCEDURE

The general possible procedure to analyse and constract any queuing model can be performed by following these listed stages [Karen C. Jones, 1992];

1. One must clearly Identify the parameters of the system, which are; the arrival rate, service time, Number of servers, queue capacity, and Queue discipline, in addition to draw the logic flow diagram of the system.

2. The possible system states must be indicate. (A state may generally represent the number of customers, people, jobs, calls, messages, etc. in the system).

3. a detailed state transition diagram may be sketched to represent the possible system states and to identify the rates to enter and leave each state. This diagram called Markov chain.

4. Because the state transition diagram represents the steady state situation between state there is a balanced flow between states so the probabilities of being in adjacent states can be related mathematically in terms of the arrival and service rates and state probabilities.

5. Using the inter-state transition relationships, one can express all the state probabilities in terms of the empty state probability.

6. The empty state probability can be determined by using the fact that all state probabilities always sum to 1.

The following notations are used to be, [Wenjing Xu, 2008].

λ = Expected capacity demand per unit time

μ = Expected capacity per time unit

N(t) = number of customers in the system at time t,

E[N(t)] = represents the expected number of customers in the system.

- A Condition for existence of a steady state solution is that $\rho = \lambda/(c\mu) < 1$

$$\rho = \frac{Capacity\ Demand}{Available\ Capacity} = \frac{\lambda}{\mu} \qquad \rho = \frac{Capacity\ Demand}{Available\ Capacity} = \frac{\lambda}{c * \mu}$$

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & for \quad n = 1, 2, \ldots, c \\ \dfrac{(\lambda/\mu)^n}{c! c^{n-c}} P_0 & for \quad n = c+1, c+2, \ldots \end{cases}$$

$$L_q = \sum_{n=c}^{\infty} (n-c) P_n = \ldots = \frac{(\lambda/\mu)^c \rho}{c!(1-\rho)^2} P_0$$

Little's Formula $\Rightarrow$ $W_q = L_q/\lambda$

Little's Formula $\Rightarrow$ $L = \lambda W = \lambda(W_q + 1/\mu) = L_q + \lambda/\mu$

VII. THE INVERSE TRANSFORMATION METHOD

This method is applicable only to cases where the cumulative density function can be inversed analytically. Assume that we wish to generate stochastic variates from a probability density function (pdf). Let F(x) be its cumulative density function. We note that F(x) is defined in the region [0,1]. We explore this property of the cumulative density function to obtain the following simple stochastic variates generator [Hong Lian, 2007] .

We first generate a random number r which we set equal to F(x). That is, F(x) = r. The quantity x is then obtained by inverting F. That is, $x = F^{-1}(r)$, where $F^{-1}(r)$ indicates the inverse transformation of F [Averill M.La, 1997].

VIII. SAMPLING FROM AN EXPONENTIAL DISTRIBUTION

we use the inverse transformation method to generate variates from a uniform distribution. The probability density function of the exponential distribution is defined as follows:

$f(x) = ae^{-ax}$, $a > 0$, $x > 0$.

The cumulative density function is [Hong Lian, 2007]:

$$F(x) = \int_0^x f(t)dt = \int_0^x ae^{-at}tdt = 1 - e^{-ax}.$$

The expectation and variance are given by the following expressions

$$E(X) = \int_0^\infty aet^{-at}dt\, dt = \frac{1}{a}$$

The inverse transformation method for generating random variable is as follows [Averill M.La, 1997]

$$r = F(x) = 1 - e^{-ax}$$

$$1 - r = e^{-ax}$$

$$x = -\frac{1}{a}\log(1\text{-}r) = -E(x)\log(1\text{-}r)$$

$$x = -\frac{1}{a}\log r.$$

IX. SINGLE-SERVER QUEUING MODEL

Single-server queues are, perhaps, the most commonly encountered queuing situation in real life. One encounters a queue with a single server in many situations, including business (e.g. sales clerk), industry (e.g. a production line), and transport (e.g. queues consequently, being able to model and analyze a single server queue's behavior is a particularly useful thing to do. M/M/1 represents a single server that has unlimited queue capacity and infinite calling population, both arrivals and service are Poisson (or random) processes, meaning the statistical distribution of both the inter-arrival times and the service times follow the exponential distribution. Because of the mathematical nature of the exponential distribution, a number of quite simple relationships are able to be derived for several performance measures based on knowing the arrival rate and service rate. This is fortunate because an M/M/1 queuing model can be used to approximate many queuing situations.

A simulation model was built to imitate the behavior of a single server Queuing system with unlimited number of customers and FCFSqueuing discplin, called (M/M/1) model . All the required performance measurements wwill be estimated and the inter arrival and service times are changed to observe the effect of each on the other parameters**.** The model layout was indicated in figure (1) , and its transition diagram (by Markov representation) is shown in fig(2). The main results was shown in the following cases [A. T. Andersen, 1998] :

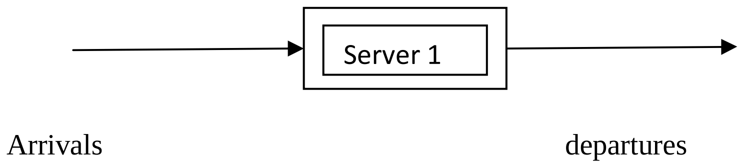


Fig (1): single server queuing model

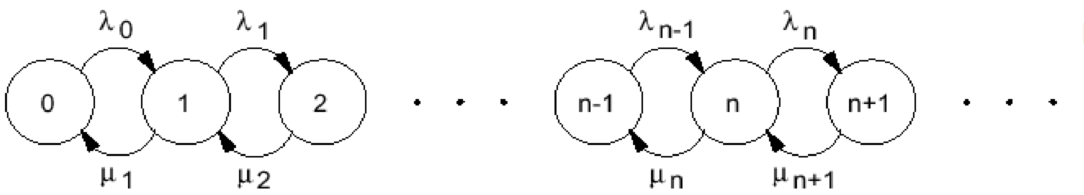


Fig (2): Transition Diagram for single server queuing model

$$\rho = \frac{\lambda}{\mu}$$

$$L_q = \frac{\rho^2}{1-\rho}, \quad L = \frac{\rho}{1-\rho}$$

$$W_q = \frac{\rho}{\mu(1-\rho)}, \quad W = \frac{1}{\mu(1-\rho)}$$

$$P_n = (1-\rho)\rho^n$$

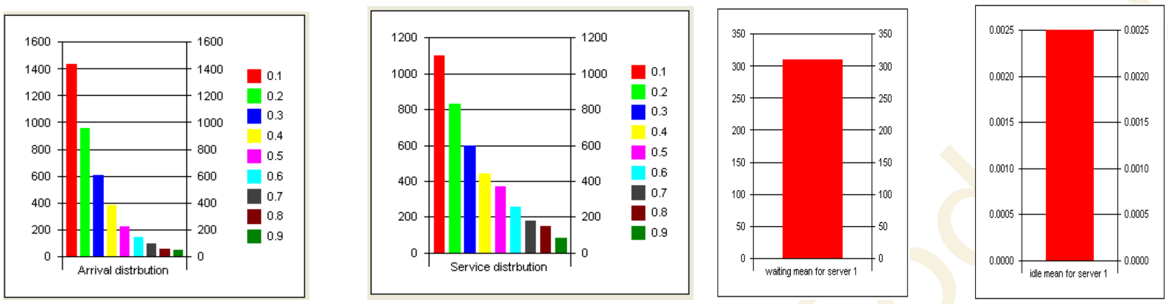
Such a system can be modeled by a (birth-death process), where each state represents the number of users in the system. As the system has an infinite queue and the population is unlimited, the number of states the system can occupy is infinite: state 0 (no users in the system), state 1 (1 user), state 2 (two users), etc. As the queue will never be full and the population size being infinite, the birth rate (arrival rate), λ, is constant for every state. The death rate (service rate), μ, is also constant for all states (apart from in state 0). The model can reveal interesting performance measures of the system being modeled, for example [Karen C. Jones, 1992];

- The mean time a user spends in the system
- The mean time a user spends waiting in the queue
- The expected number of users in the system
- The expected number of users in the queue
- The [[throughput]] (Number of users served per unit time)

After the simulation model was built, programmed and implemented, the following cases were evaluated to indicate the wanted parameters [F. Hosseinpour, 2007]:

In fig (3), fig (4) fig (5) fig (3)we collects the resulted arrival rate in (a), the resulted service or departure rate in (b), the average waiting time in (c) and the average idle time in (d).

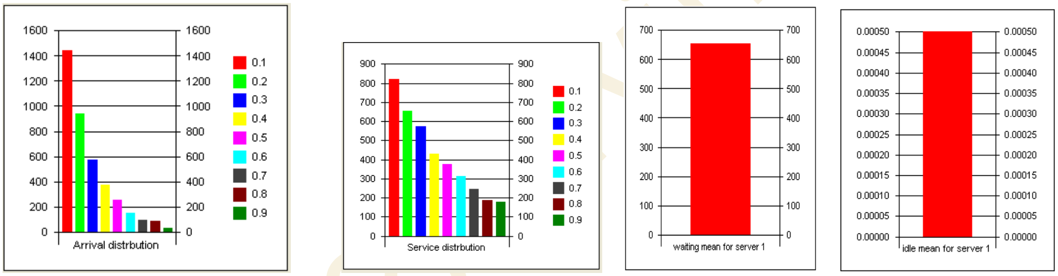
Case 1: Arrival rate **λ = 0.9 ,** service rate μ= 0.6



a: arrival rate distribution b: service rate distribution c: mean waiting time d: mean idle time

Fig(3) : case 1 results for single server model .

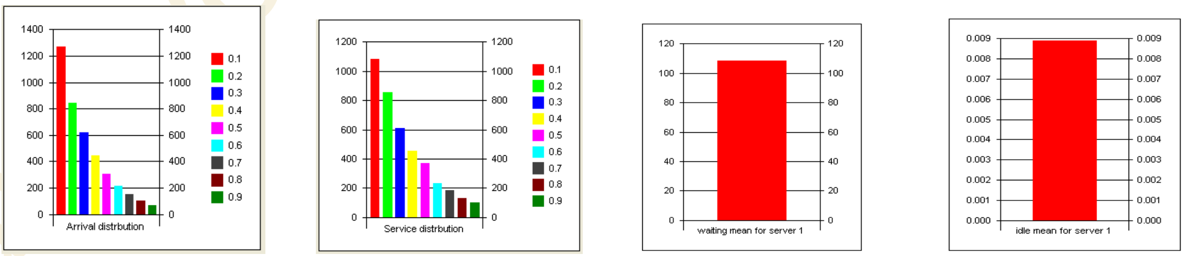
Case 2: λ= 0.9 , μ= 0.4



a: arrival rate distribution b: service rate distribution c: mean waiting time d: mean idle time

Fig(4) : case 2 results for single server model.

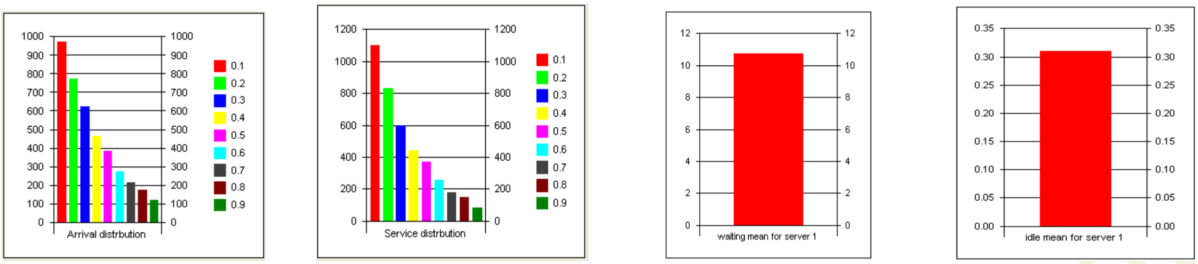
Case 3: λ= 0.7 μ= 0.6



a: arrival rate distribution b: service rate distribution c: mean waiting time d: mean idle time

Fig(5) : case 3 results for single server model.

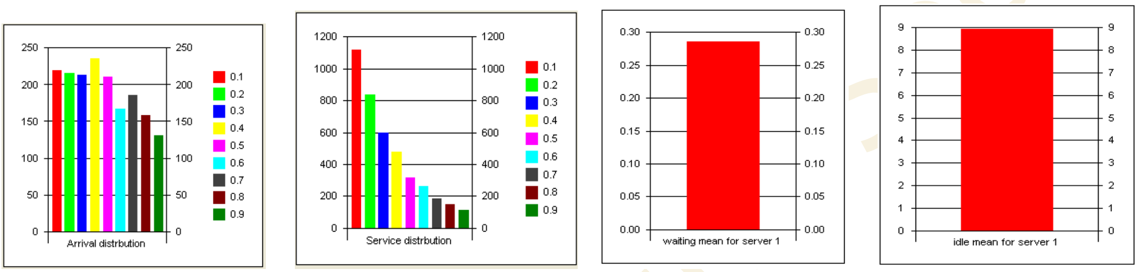
Case 4: $\lambda = 0.5$, $\mu = 0.6$



a: arrival rate distribution b: service rate distribution c: mean waiting time d: mean idle time

Fig(6) : case 4 results for single server model.

Case 5: $\lambda = 0.1$, $\mu = 0.6$



a: arrival rate distribution b: service rate distribution c: mean waiting time d: mean idle time

Fig(7) : case 5 results for single server model.

X. TANDEM QUEUE MODEL

The actual service networks configuration are different and its analysis mathematically seems difficult. "Fig. 8" presents the layout of an open network with four stages of single server connected in series. "Four stages tandem queuing model". The fundamentals of Markov chains were used to develop the transition rate diagram and all possible states for the four stages queuing model as shown in "Fig. 9". Part a, of "Fig. 9" shows the possible steady state transitions from a general state to all adjacent states, while part b, shows the possible transitions from all adjacent states to the suggested general state. The arrival rate is λ and the service rates are μi.

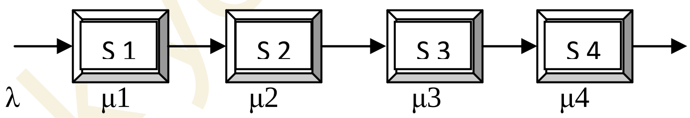
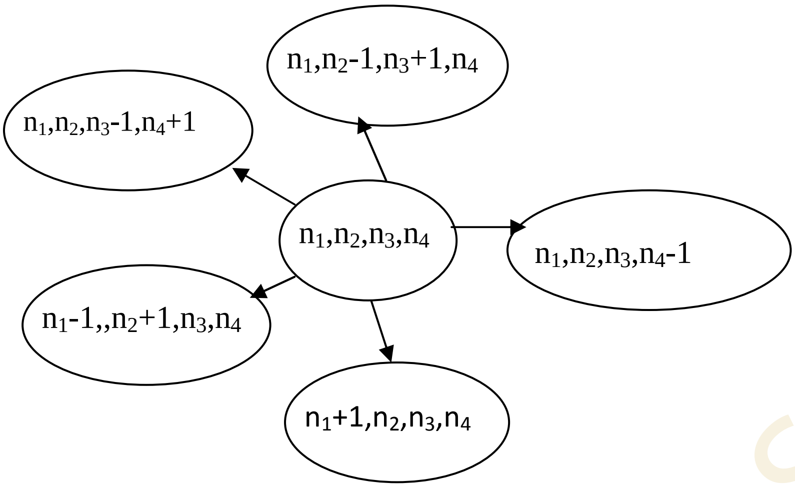


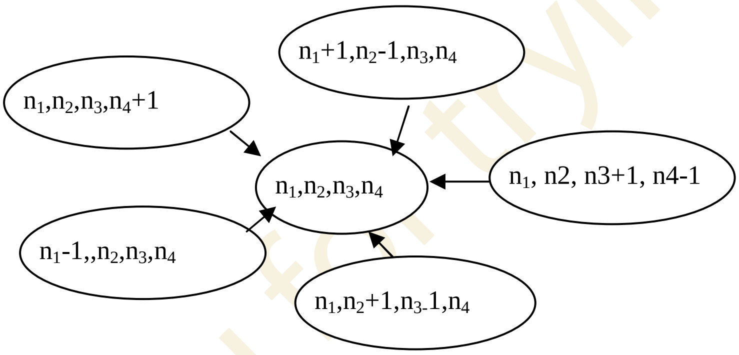
Fig (8) : Four stages open network tandem queuing model.

Markov chain is one of the possible techniques to analyze this model. The required differential equations can be developed from the transition rate diagram. The developed difference differential equations to represent this system steady state are not easy to deal with and the situation will be complicated if we have more complex service network. The process of developing such system of equations and to find their solution requires a very high skill and consumes time in build and implement each model. We suggested the following simulation approach to estimate the required parameters to assist the decision maker about any required information.

Our simulation approach suggests the exponential inter-arrival time and exponential service time with different parameters. These parameters can be changed and observe the effects on the final results. Any other distribution can be used after the process validation through the computer model which requires change the form of the inverse transform which used to generate the required random variable as a sampling process from the wanted distribution. The computer simulation model was built and implemented with long runs to imitate this open network model and we got the following results which summarized in “Fig. 10”. It shows the resulted arrival rate distribution in (a), the service rate distribution for each server in (b, c, d, and e), the waiting time distributions in (f), the mean idle time for each server in (g), and the mean average waiting time for the system in (h).



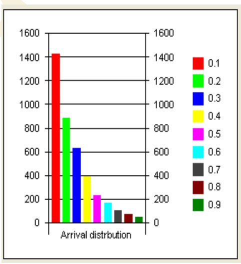
(a) : transition rate from genral state to each possible state.



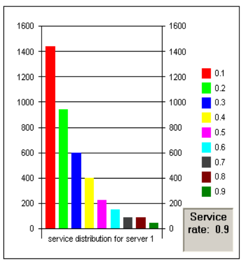
(b) : transition rate to genral state from possible states.

Fig (9): transition rate diagram to the four stages tandem queuing model.

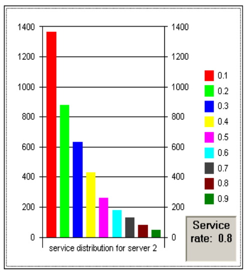
The computer simulation model was fed with the following suggested parameters value:
Inter- arrival time: λ= 0.9 , Server 1 service time is μ1= 0.9 , Server 2 service time is μ2= 0.8 , Server 3 service time is μ3= 0.6 , Server 4 service time is μ4= 0.4.



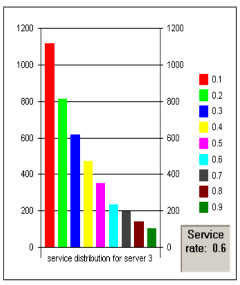
(a):arrival rate



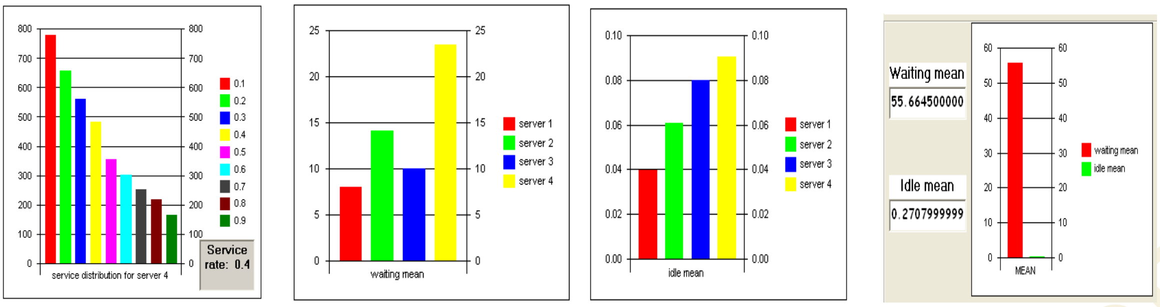
(b): service rate for server 1



(c):service rate for server 2



(d): service rate for server 3



(e): service rate for server 4 (f) waiting times (g): idle times (h): mean waiting time

Fig(10): Simulation results of the four stages tandem queuing model.

XI. CONCLUSIONS

The modeling and simulation Technique offers a room for questions and arguments about the validity of the performance numbers. How realistic are they? What are their levels of confidence? What interpretation and extrapolation is consistent with the assumed model? is another model should have been used? What is the optimum model for the network under given conditions?

Simulation as a powerful scientific technique aims in service process to evalute the suitable model, or testes the effects of changes in existing operated system.The outcome from this paper showed simulation models to real queuing problems developed using computer. This technique have benifets to both; the dicision maker, the planner and the manegers , due to the possible helps to take the right decision about the system and to evaluate the performance of each configration depending on possibility to minimazing cost by minimizing waiting time or idle time or both. Computer simulation allow us to test these models in virtual and change each factor each time and show all these effects on the system , before trying it in the real life to find out which case give us the optimal solution.

The main recommendation we have reached is to advice all the researchers to make use of the modeling process in describing the behavior of the system and in evaluating its performance. When using correct parameter values the results will be very close to the measured values. In the worst case when it does not provide accurate results, it helps in developing a creative understanding on the behavior and the performance of the system.

The basic performance modeling technique utilized in the paper can be applied for a wide variety of new service networks design alternatives, hardware capacity studies, the process of evaluating the system performance, and the requirements for data collection.

Reference

[A. Horv´ath, 2000]; A. Horv´ath, G.I.R´ozsa, M. Telek, “A map fitting method to approximate real traffic behaviour,” 8-th IFIP Workshop on Performance Modelling and Evaluation of ATM and IP Networks, Ilkley, U.K., July 2000.

[A. T. Andersen, 1998]; A. T. Andersen, B. F. Nielsen, “A markovian approach for modeling packet traffic with long-range dependence,” IEEE journal on selected areas in communications, Vol.16, No.5, pp. 719-732, 1998.

[Averill M.La, 1997] ;Averill M.Law , Michael G.McComas, '*Simulation of manufacturing systems*' Proceedings of the 1997 Winter Simulation Conference , ed. S.Andradottir , K. J. Healy , D. H Withers , and B. L. Nelson, Averill M. Law & Associates, Inc.P.O. Box 40996,Tucson, Arizona 85717, U.S.A.

[Cooper, R.. P, 1981]; Cooper, R. P., "Introduction to Queuing Theory", 1981, Elsevier, New York.

[Curry, G., B., et al, 2003]; Curry, G., B. A. Peters, and M. Lee. 2003. Queueing network model for a class of material-handling systems. *International Journal of Production Research* 41 (16): 3901–20.

[F. Hosseinpour, 2007]; F. Hosseinpour, and H. Hajihosseini, 'Importance *of Simulation in Manufacturing' World* Academy of Science, Engineering and Technology 51 2009.

[Gunter Bolch et al, 2006]; Gunter Bolch et al, Queuing Networks and Morkov Chains: Modeling and Performance Evaluation with Computer Science Applications, John Wiley & Sons, 2006.

[Hong Lian, 2007]; Hong Lian, and Zhenkai Wan , '*The Computer Simulation for Queuing System*', World Academy of Science, Engineering and Technology 34 2007.

[Karen C. Jones, 1992]; Karen C. Jones,' *Manufacturing Simulation Using Virtual Reality'*, MSc thesis in Industrial Engineering , University of Washington ,1992.

[Kishor S. Trivedi, 2001]; Kishor S. Trivedi : Probability and Statistics with Reliability, Queuing, and Computer Science Applications, John Wiley and Sons, 2001.

[Lammer S, and Helbing D, 2008]; Lammer S. and Helbing D., Self-control of traffic lights and vehicle flows in urban road networks, J. Stat. Mech, 1742-5468, 2008

[Nelson R, 2000]; Nelson R., "Probability, stochastic processes, and queuing theory : the mathematics of computer performance modeling", 2000, Springer-Verlag, New York

[Sheldon M. Ross, 2001]; Sheldon M. Ross : *Introduction to Probability Models*, Academic Press, 2001.

[Wenjing Xu, 2008]; Wenjing Xu, '*Application of Proxels to Queuing Simulation with Attributed Jobs*, MSc thesis in Computational Visualistics , Magdeburg, 10, February 2008.