QUEUING THEORY AND ITS APPLICATION: ANALYSIS OF THE SALES CHECKOUT OPERATION IN ICA SUPERMARKET

by

Azmat Nafees

A 'D level' essay in Statistics submitted in partial fulfillment of the requirements for the degree of

M. Sc.

Department of Economics and Society

June 2007

Presented to supervisor Martin Sköld

University of Dalarna

ABSTRACT

ANALYSIS OF THE SALES CHECKOUT OPERATION
IN ICA SUPERMARKET USING QUEUING
SIMULATION

by Azmat Nafees

Members of the Supervisory Committee: Martin Sköld & Richard Stridbek
Department of Economics & Society

This paper contains the analysis of Queuing systems for the empirical data of supermarket checkout service unit as an example. One of the expected gains from studying queuing systems is to review the efficiency of the models in terms of utilization and waiting length, hence increasing the number of queues so customers will not have to wait longer when servers are too busy. In other words, trying to estimate the waiting time and length of queue(s), is the aim of this research paper. We may use queuing simulation to obtain a sample performance result and we are more interested in obtaining estimated solutions for multiple queuing models.

This paper describes a queuing simulation for a multiple server process as well as for single queue models. This study requires an empirical data which may include the variables like, arrival time in the queue of checkout operating unit (server), departure time, service time, etc. A questionnaire is developed to collect the data for such variables and the reaction of the ICA Supermarket from the customers separately. This model is developed for a sales checkout operation in ICA supermarket, Borlänge. The model designed for this example is multiple queues multiple-server model. The model contains five servers which are checkout sales counters; attached to each server is a queue. In any service system, a queue forms whenever current demand exceeds the existing capacity to serve. This occurs when the checkout operation unit is too busy to serve the arriving costumers, immediately.

Keywords: multiple-server model, queuing simulation, steady-state condition, confidence intervals for arrival rate and service rate, estimated queue length, interarrival time.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank the staff of ICA. I am extremely grateful to my project advisors, Mr. Richard Stridbeck and specially Mr. Martin Sköld who assisted me from the beginning to the end of the project report and gave guidance and a sense of direction to me. I am very grateful to my classmate Liwen Liang who helped me collecting the data, and grateful to a choco-bar staff in ICA who let us sit in their bar to work for the data collection for my project.

Special thanks to my coordinator Ms. Catia Cialani who helped me in compilation of my survey report, and all those people (customers) who took time out of their busy schedules and gave me the information, which was essential for the completion of this project. These people have been instrumental in my research and project work

OBJECTIVES OF STUDY

The purpose of this study is to review Queuing Theory and its empirical analysis based on the observed data of checking out sales service unit of ICA Supermarket. The main idea in the application of a mathematical model is to measure the expected queue length in each checkout sales service unit (server) and the service rate provided to the customers while checking out. Another idea is to give insight view of the steady-state behavior of queuing processes and running the simulation experiments to obtain the required statistical results.

Descriptions of events are given i.e. the arrivals and service rate in each checkout unit and how they can be generated for any amount of working hour. The other important factor analyzed is about the comparison of two different queuing models: single-queue multiple-server and multiple-queue multiple-server model.

Chapter 1

INTRODUCTION

This paper is the review of queuing theory and for empirical study the sales checkout service unit of ICA supermarket is chosen as an example.

ICA AB is a Swedish corporate group in the retail business, which started in 1938 and operated 1,668 stores as of 2003. The stores have different profiles, depending on location, range of products and size:

a) ICA Nara ("ICA Near-by", convenience-type stores for daily retail needs);
b) ICA Supermarket (Mid-size supermarkets, located near where customers dwell or work carrying a wide range of products);
c) ICA Kvantum (Superstores for large, planned, purchases. Large spaces allocated for traffic and parking. Typically located outside of the cities);
d) MAXI ICA Stormarknad (these are Hypermarkets with a full range of groceries as well as fashions, home wares, entertainment and electrical. Smaller stores do not offer the fashion and electrical ranges while the largest stores also have a DIY and Garden department).

Each store is owned and operated separately, but operations are coordinated within the group. All feature ICA brand products.

There are two ICA Supermarkets in Borlänge; the bigger one was chosen to be the research object and to collect data from.

The main purpose of this paper is to review the application of queuing theory and to evaluate the parameters involved in the service unit for the sales checkout operation in ICA Supermarket. Therefore, a mathematical model is developed to analyze the performance of the checking out service unit. Two important results need to be known from the data collected in the supermarket by the mathematical model: one is the 'service rate' provided to the customers during the checking out process, and the other is the gaps between the arrival times (interarrival time) of each customer per hour. In order to get an

overall perspective of the customer's quality of service, the questionnaires which indicate the result in percentages, are also used to get the evaluation from the customers directly.

There are five counters in ICA Borlänge at one place, which means consisting of five servers with five queues in terms of Queuing Theory. A queue forms whenever current demand exceeds the existing capacity to serve when each counter is so busy that arriving customers cannot receive immediate service facility. So each server process is done as a queuing model in this situation.

The data used in the Queuing model is collected for an arrival time of each customer in two days by the questionnaire form. The observations for number of customers in a queue, their arrival-time and departure-time were taken without distracting the employees. The whole procedure of the service unit each day was observed and recorded using a time-watch during the same time period for each day. In addition, the questionnaires are conducted at the same timings for each day.

The aim of studying queuing system simulation is trying to detect the variability in a quality of service due to queues in sales checkout operating units, find the average queue length before getting served in order to improve the quality of the services where required, and obtain a sample performance result to obtain time-dependent solutions for complex queuing models. The defined model for this kind of situation where a network of queues is formed is time-dependent and needs to run simulation. The results obtained from ICA Supermarket using queuing model suggest that sales checkout operating unit is rather busy each day of a week but the service is satisfactory.

Chapter 2

BACKGROUND

Queuing Theory

Delays and queuing problems are most common features not only in our daily-life situations such as at a bank or postal office, at a ticketing office, in public transportation or in a traffic jam but also in more technical environments, such as in manufacturing, computer networking and telecommunications. They play an essential role for business process re-engineering purposes in administrative tasks. "Queuing models provide the analyst with a powerful tool for designing and evaluating the performance of queuing systems." (Bank, Carson, Nelson & Nicol, 2001)

Whenever customers arrive at a service facility, some of them have to wait before they receive the desired service. It means that the customer has to wait for his/her turn, may be in a line. Customers arrive at a service facility (sales checkout zone in ICA) with several queues, each with one server (sales checkout counter). The customers choose a queue of a server according to some mechanism (e.g., shortest queue or shortest workload). (Adan, 2000)

Sometimes, insufficiencies in services also occur due to an undue wait in service may be because of new employee. Delays in service jobs beyond their due time may result in losing future business opportunities. Queuing theory is the study of waiting in all these various situations. It uses queuing models to represent the various types of queuing systems that arise in practice. The models enable finding an appropriate balance between the cost of service and the amount of waiting.

Chapter 3

METHODOLOGY

Queuing Models with Single Stage (facility)

The term queuing system is used to indicate a collection of one or more waiting lines along with a server or collection of servers that provide service to these waiting lines. The example of ICA supermarket is taken for queuing system discussed in this chapter include: 1) a single waiting line and multiple servers (fig.1), 2) multiple waiting lines (arranged by priority) and multiple servers (fig.2) , and 3) a single waiting line and a single server (fig.3). All results are presented in next chapter assuming that FIFO is the queuing discipline in all waiting lines and the behavior of queues is jockey[1].

The supermarkets may consist of multiple units to perform same checkout operation of sales, which are usually set all together besides the entrance of the supermarket. Each unit contains one employee. This kind of a system is called a multiple-server system with single service facility, in other words multiple checkouts counters (service units) with sales checkout as a service available in a system. There are two possible models for multiple-server system: Single-Queue Multiple-Server model, and Multiple-Queue Multiple-Server model.

Using the same concept of model, the sales checkout operating units are all together taken as a series of servers that forms either single queue or multiple queues for sales checkout (single service facility) where the arrival rate of customers in a queuing system and service rate per busy server are constants regardless of the state of the system (busy or idle). For such a model the following assumptions are made:

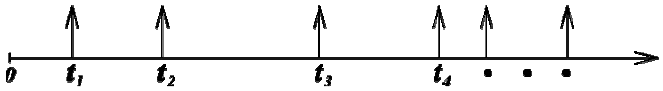
Assumptions

a) Arrivals of customers follow a Poisson process
   i. The number of the customers that come to the queue of sales checkout server during time period [t, t+s) only depends on the length of the time period 's' but no relationship with the start time 't'

[1] The customer enters one line and then switches to a shorter line to reduce the waiting time.

ii. If s is small enough, there will be at most one customer arrives in a queue of a server during time period [t, t+s)

Therefore, the number of customers that arrive in an interval [t, t+s) follows a Poisson distribution and the arrivals of them in a queue follows a Poisson process.



A Poisson process as a sequence of events 'randomly spaced in time'

b) Interarrival times of a Poisson process are exponentially distributed

Let τ_1 = the time until the next arrival from t_0 to t_1 i.e. $(t_1 - t_0)$

And $P(\tau_1 > t) = P_0(t) = e^{-\mu \cdot t}$

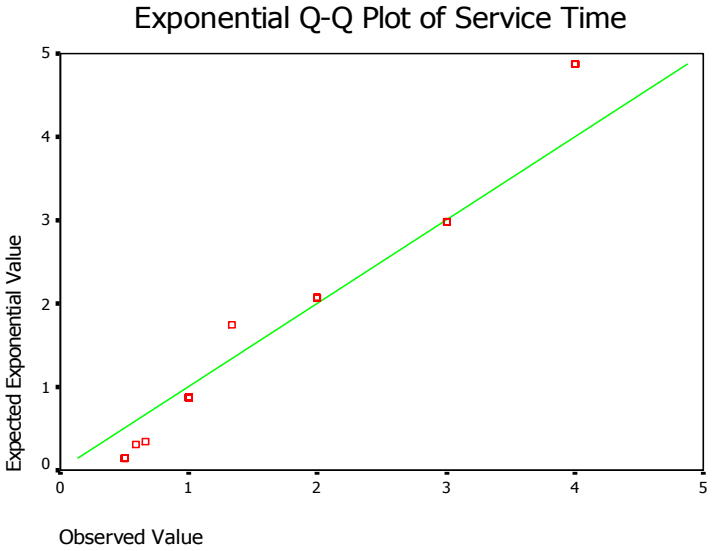
Then $P(\tau_1 \le t) = F_{\tau_1}(t) = 1 - e^{-\mu \cdot t}$ and $f_{\tau_1}(t) = \mu e^{-\mu \cdot t}$ $for\ t > 0$

Similarly, the random variables $\tau_1, \tau_2, \ldots \tau_n, \ldots$ of interarrival times are independent of each other and each has an exponential distribution with mean $1/\mu$

c) Service times are exponentially distributed

This has been examined by Q-Q plot of collected data given below. The length of the time between arrivals and departures contain the length of the queue and the service time. So the service times are exponentially distributed.

Q-Q plot shows the service time is exponentially distributed:



And there is one more thing to mention is that there are only a few points on the graph but the number of observations in the original data is nearly 100. The reason for this condition is that, the data was not observed per seconds, whereas service may vary per second. Therefore, some service time has identical value of time.

d) Identical service facilities (same sales checkout service on each server)

e) No customer leaves the queue without being served

f) Infinite number of customers in queuing system of ICA (i.e. no limit for queue capacity)

g) FIFO (First In First Out) or FCFS (First Come First Serve)

Customers arriving from different flows are treated equally by placing into the queues, respecting strictly, their arriving order. Already in the queue are served in the same order they entered, this means, first customer that comes in the queue is the first one that goes out.

All customers arriving in the queuing system will be served approximately equally distributed service time and being served in an order of first come first serve, whereas customer choose a queue randomly, or choose or switch to shortest length queue. There is no limit defined for number of customers in a queue or in a system.

Basic Queuing Process

Customers requiring service are generated over time by an input source. The required service is then performed for the customers by the service mechanism, after which the customer leaves the queuing system. We can have following two types of models: One model will be as **Single-queue Multiple-Servers model (fig.1)** and the second one is **Multiple-Queues, Multiple-Servers model (fig.2)** (Sheu, C., Babbar S. (Jun 1996)).

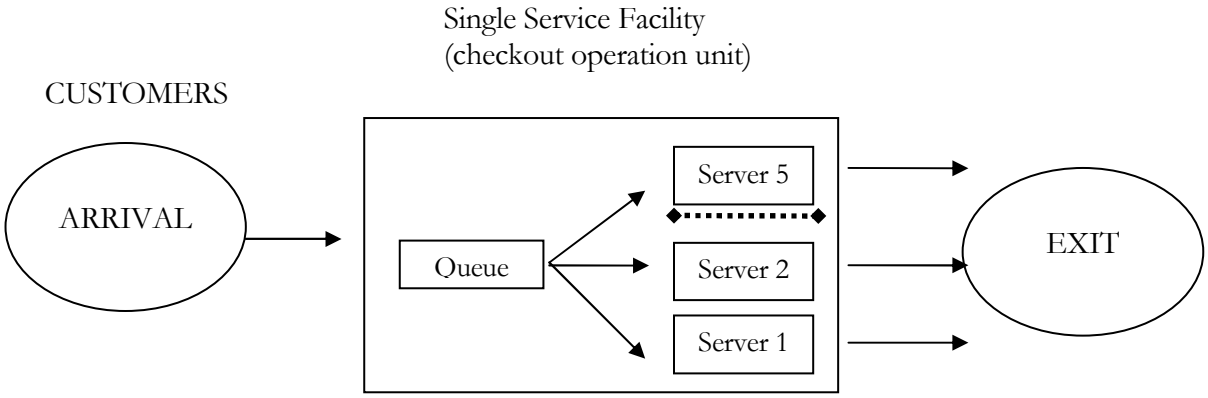


Fig. 1: Single Stage Queuing Model with Single-Queue and Multiple Parallel Servers

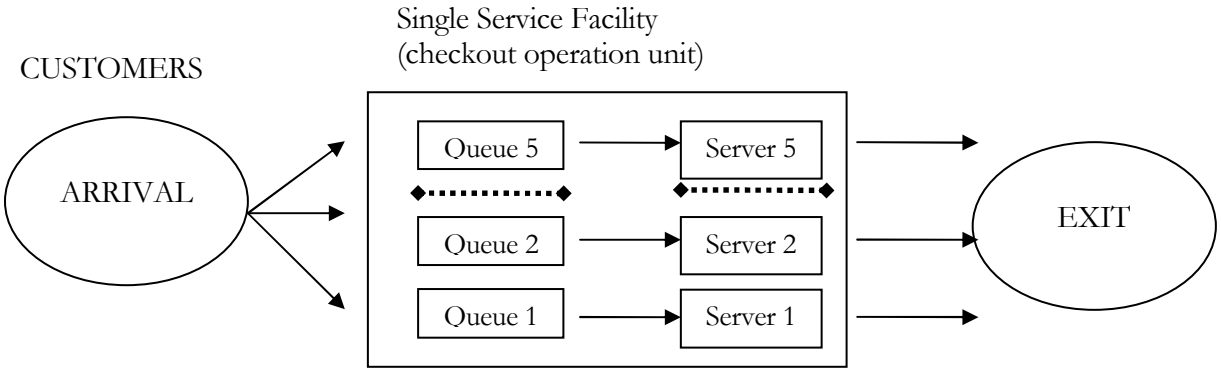


Fig. 2: Single Stage Queuing Model with Multiple Queues and Multiple Parallel Servers

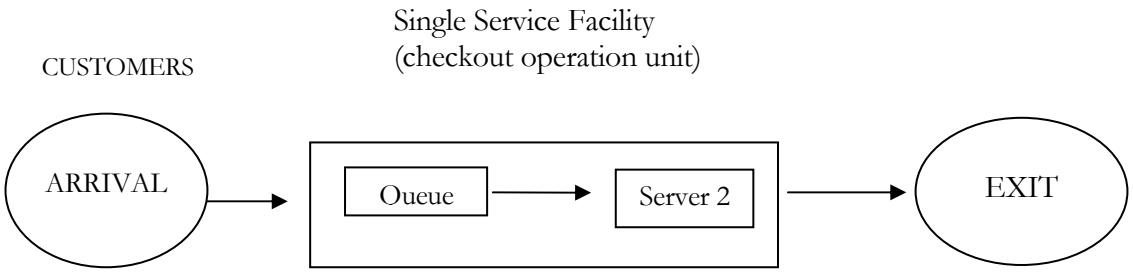


Fig. 3: Single Stage Queuing Model with Single-Queue and Single-Server

In these models, three various sub-processes may be distinguished:

- Arrival Process: includes number of customers arriving, several types of customers, and one type of customers' demand, deterministic or stochastic arrival distance, and arrival intensity. The process goes from event to event, i.e. the event "customer arrives" puts the customer in a queue, and at the same time schedules the event "next customer arrives" at some time in the future.
- Waiting Process: includes length of queues, servers' discipline (First In First Out). This includes the event "start serving next customer from queue" which takes this customer from the queue into the server, and at the same time schedules the event "customer served" at some time in the future.
- Server Process: includes a type of a server, serving rate and serving time. This includes the event "customer served" which prompts the next event "start serving next customer from queue". (Troitzsch, 2006)

The Queuing model is commonly labeled as M/M/c/K, where first **M** represents Markovian[2] exponential distribution of inter-arrival times, second **M** represents Markovian exponential distribution of service times, **c** (a positive integer) represents the number of servers, and **K** is the specified number of customers in a queuing system. This general model contains only limited number of K customers in the system. However, if there are unlimited number of customers exist, which means $K = \infty$, then our model will be labeled as M/M/c (Hillier & Lieberman, 2001.)

[2] A stochastic process is called Markovian (after the Russian mathematician Andrey Andreyevich Markov) if at any time t the conditional probability of an arbitrary future event given the entire past of the process—i.e., given $X(s)$ for all $s \leq t$—equals the conditional probability of that future event given only $X(t)$. Thus, in order to make a probabilistic statement about the future, Markovian processes are used.

Parameters in Queuing Models (Multiple Servers, Multiple Queues Model)

n Number of total customers in the system (in queue plus in service)

c Number of parallel servers (Checkout sales operation units in ICA)

λ Arrival rate (1 / (average number of customers arriving in each queue in a system in one hour))

μ Serving rate (1 / (average number of customers being served at a server per hour))

cμ Serving rate when c > 1 in a system

ρ System intensity or load, utilization factor (= λ/(cμ)) (the expected factor of time the server is busy that is, service capability being utilized on the average arriving customers)

Departure and arrival rate are state dependent and are in steady-state (equilibrium between events) condition.

Notations & their Description for single queue and parallel multiple servers model (fig.1) assuming the system is in steady-state condition

P_0 Steady-state Probability of all idle servers in the system, i.e. $P_0 = \left[\sum_{n=0}^{c-1} \frac{\gamma^n}{n!} + \frac{\gamma^c}{c!(1-\rho)}\right]^{-1}$

where $\gamma = \frac{\lambda}{\mu}$

P_n Steady-state Probability of exactly n customers in the system

$$P_n = \frac{\lambda^n}{c! c^{n-c} \mu^n} P_0 \quad n > c$$

Lq Average number of customers in the waiting line (queue) = $\frac{\gamma^c \rho}{(c)!(1-\rho)^2} \times P_0$

Wq Average waiting time a customer spends waiting in line excluding the service time = $\frac{L_q}{\lambda}$

There are no predefined formulas for networks of queues, i.e. multiple queues (fig.2). A complexity of the model is the main reason for that. Therefore, we use notations and formulas for single queue with parallel servers. In order to calculate estimates for multiple queues multiple servers' model, we may run simulation.

Expected length of each Queue

Besides service time, it is important to know the number of customers waiting in a queue to be served. It is possible that any customer would change his queue and choose another if find a shorter queue in another parallel server. In general, variability of interarrival and service time causes lines to fluctuate in length. Then question arises, what could be the estimated length of the queue in any server? Some papers describe the general criterion for counting the number of customers in a queue. These counts are a combination of input processes, that are: arrival point process, Poisson counting process (which counts only those units that arrive during the interarrival time and these units are conditionally independent on Poisson interval), and counting group of units being served within the Poisson interval. The above mentioned formula of Lq is defined for average queue length of the queuing system but does not evaluate a length of parallel queues.

We are next concerned about how to obtain solution for a queuing model with a network of queues? Such questions require running Queuing Simulation. Simulation can be used for more refined analysis to represent complex systems.

Queuing Simulation:

The queuing system is when classified as M/M/c with multiple queues where number of customers in the system and in a queue is infinite, the solution for such models are difficult to compute. When analytical computation of μ is very difficult or almost impossible, a Monte Carlo simulation is appealed in order to get estimations. A standard Monte Carlo simulation algorithm fix a regenerative state and generate a sample of regenerative cycles, and then use this sample to construct a likelihood estimator of state. (Nasroallah, 2004) Although supermarket sales do not have regenerative situation but simulation here is used to generate estimated solutions.

Simulation is the replication of a real world process or system over time. Simulation involves the generation of artificial events or processes for the system and collects the observations to draw any inference about the real system. A discrete-event simulation simulates only events that change the state of a system. Monte Carlo simulation uses the mathematical models to generate random variables for the artificial events and collect observations. (Banks, 2001)

Discrete models deal with system whose behavior changes only at given instants. A typical example occurs in waiting lines where we are interested in estimating such measures as the average waiting time or

the length of the waiting line. Such measures occur only when the customer enters or leaves the system. The instants at which changes in the system occurs identify the model's events, e.g. arrival and departure of the customers. The arrival events are separated by the 'interarrival time' (the interval between successive arrivals), and the departure events are specified by the service time in the facility. The fact that these events occur at discrete points is known as "Discrete-event Simulation." (Taha, 1997)

When the interval between successive arrivals is random then randomness arises in simulations. The time t between customers' arrivals at ICA is represented by an exponential distribution; to generate the arrival times of the next customers from this distribution, we have $t = -\left(\frac{1}{\mu}\right)\ln(1-R)$ where R = random number. (1 – R) is a compliment of R, so we can replace (1 – R) with R.

Chapter 4

ANALYSIS OF CHECKOUT SALES OPERATION SERVICE IN ICA

A sales checkout service has 5 waiting lines in a form of parallel cash counters (see fig.1 in the chapter 3). Customers are served on a first-come, first-served (FIFO) basis as a salesman of checkout operation unit becomes free. The data has been collected for only two out of five servers on Wednesday (weekday) by using questionnaires (Appendix A). It was assumed that the customers' crowd is more, on average, on weekday. Although the sales checkout unit has 5 parallel counters out of which 2 were observed (each of them has an individual salesman to deal with the customers in a queue), it is possible that some of the checkout units are idle. The data collected from questionnaires were tabulated in a spreadsheet in order to calculate the required parameters of queuing theory analysis (Appendix B). Firstly, the confidence intervals are computed to estimate service rate and arrival rate for the customers. Then the later first part of the analysis is done for the model involving one queue and 2 parallel servers (fig.1), whereas the second part is done by queuing simulation for second model involving 2 queues for each corresponding parallel server (fig.2).

We can estimate confidence intervals for average service rate and average arrival rate. Assuming service time and arrival time are iid with N(0,1), then the 95% confidence interval for arrival rate can be:

Confidence Intervals

$$\left[(mean\ arrival\ time + 1.96 \cdot SE(mean\ arrival\ time))^{-1}, (mean\ arrival\ time - 1.96 \cdot SE(mean\ arrival\ time))^{-1}\right]$$

where $SE(mean\ arrival\ time) = SD(mean\ arrival\ time)/\sqrt{n}$

Similarly, 95% confidence interval for service rate can be:

$$\left[(mean\ service\ time + 1.96 \cdot SE(mean\ service\ time))^{-1}, (mean\ service\ time - 1.96 \cdot SE(mean\ service\ time))^{-1}\right]$$

where $SE(mean\ servcie\ time) = SD(mean\ servcie\ time)/\sqrt{n}$

Confidence Intervals for weekday:

We have,

Mean (service time) = 01:06 minutes per customer (read clock as min:sec)

SD (service time) = 00:06 min

Mean (arrival time) = 00:37 min per customer

SD (arrival time) = 00:06 min

And n = 41 customers

95% Confidence Intervals for Service Time:

Mean(service time) - 1.96 (SE(service time)) =	54 sec/customer
Mean(service time) + 1.96 (SE(service time)) =	78 sec/customer

SE = SD/sqrt(n)

95% Confidence Intervals for Service Rate:

1/[Mean(service time) + 1.96 (SE(service time))] =	0.01282	= 46 customers/sec
1/[Mean(service time) - 1.96 (SE(service time))] =	0.01852	= 67 customers/sec**

** (0.01852 sec *60 *60)

95% Confidence Intervals for Arrival Time:

Mean(arrival time) – 1.96 (SE(arrival time)) =	24 sec /customer
Mean(arrival time) + 1.96 (SE(arrival time)) =	49 sec /customer

95% Confidence Intervals for Arrival Rate:

1/[Mean(arrival time) + 1.96 (SE(arrival time))] =	0.02041	= 73 customers/sec**
1/[Mean(arrival time) - 1.96 (SE(arrival time))] =	0.04167	= 150 customers/sec

** (0.02041 sec *60 *60)

Interpretation of confidence intervals

The confidence intervals show that 73 to 150 customers arrive in 2-server system within an hour whereas 46 to 67 customers are served. That means there are still some customers not being served and are waiting for their turn in a queue to be served. This is due to a service time provided by a server to the customers. The service time can vary between 54 sec to 78 sec per customer.

Expected Queue Length

We can find the expected length of queue by using empirical data. In survey, the number of customers waiting in a queue was observed (Appendix B). The average of that number in a system is (1+1+3+…+2+0)/41 = 2.07 customers per minute on average waiting in a queue in a system within 25 min of data collection time.

Queuing Analysis

On Wednesday (weekday), customers arrive at an average of 98 customers per hour, and an average of 55 customers can be served per hour by a salesperson.

Results for Weekday applying Queuing model 1 (fig.1)

The parameters and corresponding characteristics in Queuing Model M/M/2, assuming system is in steady-state condition, are:

c number of servers = 2

λ arrival rate = 98 customers per hour

μ serving rate = 55 customers per server per hour

$c\mu$ (2) (55) = 110 (service rate for 2 servers)

ρ $= \lambda/(c\mu) = 98 / 110 = 0.8909$

γ $= \lambda/\mu = 1.7818$

Overall system utilization = ρ = 89.09 %

The probability that all servers are idle (Po) = 0.5769

Average number of customers in the queue (Lq) = $\frac{\gamma^c \rho}{(c)!\,(1-\rho)^2} \times P_0 = 6.8560$

Average time customer spends in the queue (Wq) = Lq/λ = 0.0700 hours

Interpretation of results for queuing model 1

The performance of the sales checkout service on weekday is sufficiently good. We can see that the probability for servers to be busy is 0.8909, i.e. 89.09%. The average number of customers waiting in a queue is Lq = 6.8560 customers per 2-server. The waiting time in a queue per server is Wq = 4.2 min which is normal time in a busy server. This estimate is not realistic as the model shows that the customers make a single queue and choose an available server. Hence we can consider each server with a queuing

model as a single-server single-queue model to get the correct estimate of the length of queue. M/M/1 queue is a useful approximate model when service times have standard deviation approximately equal to their means.

Results for Weekday applying Queuing model 3 (fig.3)

The parameters and corresponding characteristics in Queuing Model M/M/1, assuming system is in steady-state condition, are:

c number of servers = 1

λ arrival rate = 98 customers per hour for 2 servers i.e. 49 customers

μ serving rate = 55 customers per server per hour

ρ $= \lambda/(c\mu) = (98 \div 2)/ 55 = 0.8909$

γ $= \lambda/\mu = 0.8909$ (= ρ in case of c = 1)

Overall system utilization = ρ = 89.09 %

The probability that all servers are idle (Po) = 0.1091

Average number of customers in the queue (Lq) = $\frac{\gamma^c \rho}{(c)!\,(1-\rho)^2} \times P_0 = 7.2758$

Average time customer spends in the queue (Wq) = Lq/λ = 0.1485 hours

Interpretation of results for queuing model 3

The performance of the sales checkout service remains same as for 2 servers on weekday. The number of customers in a queue is (7.2758) higher than a queue with two servers. Each customer in a queue has to wait for 8.9 minutes. This means, reducing the number of servers may lead a longer queue.

Queuing Simulation

It is not possible to obtain solutions for multi-queue models in closed form or by solving a set of equations, but they are readily obtained with simulation methods. The simulation has been run for the same empirical data as for model 1, using software WinQSB for Queuing System Simulation (Appendix C). The mean interarrival time and mean service time as taken same for both servers.

Results for Weekday applying Queuing model 2 (fig.2)

Server 1

Mean interarrival time = 0.6333 min

Mean Serving time = 1.1000 min

Server utilization = ρ = 99.00 %

Number customers served = 93 customers

Average number of customers in the queue (Lq) = 28.1820 customers

Average time customer spends in the queue (Wq) = 21.3131 min

Server 2

Mean interarrival time = 0.6333 min

Mean Serving time = 1.1000 min

Server utilization = ρ = 99.00 %

Number customers served = 77 customers

Average number of customers in the queue (Lq) = 39.3991 customers

Average time customer spends in the queue (Wq) = 28.8511 min

Overall for two servers

Mean Serving time = 1.1000 min

Server utilization = ρ = 99.00 %

Average number of customers in the queue (Lq) = 67.5812 customers

Average time customer spends in the queue (Wq) = 25.0821 min

Interpretation of Queuing Simulation results for model 2

A simulation process has clearly shown the performance of the sales checkout service of two servers including their corresponding queues. The simulation was run for 100 hours. The servers are found to be very busy (99%). The average number of customers waiting in a queue in overall two servers on weekday is Lq = 67.5812 whereas the waiting time in a queue in overall two servers is approximately Wq = 25.0821 min which is normal time in a very busy server. Such a longer queue can be reduced in size by a decrease in service time or server utilization. Although interarrival time and mean service time is same for both servers but there is a small difference in the value of Lq and Wq. This is possible when system has multiple queues and queues have jockey behavior. In other words, customers tend to switch to a shorter queue to reduce the waiting time.

Comparison of the results for Queuing model 1 and model 2

The actual structure of our survey example ICA has queuing model 2 (fig.2). A queuing model with single queue and multiple parallel servers (fig.1) does not clearly evaluate performance for each server. For instance, the utilization factor for both servers varies in each analysis, i.e. for model 1 its 89% whereas for model 2 its 99%. A simulation process shows the performance of each server with their corresponding queues (fig.2). For instance, in server 2 each customer has to wait for 15.67 minutes in case of 40 customers in a queue and in server 1 each customer has to wait for 21.87 minutes in case of 31 customers waiting in a queue for being served.

DISCUSSION

This paper reviews a queuing model for multiple servers. The average queue length can be estimated simply from raw data from questionnaires by using the collected number of customers waiting in a queue each minute. We can compare this average with that of queuing model. Three different models are used to estimate a queue length: a single-queue multi-server model, single-queue single-server and multiple-queue multi-server model. In case of more than one queue (multiple queue), customers in any queue switch to shorter queue (jockey behavior of queue). Therefore, there are no analytical solutions available for multiple queues and hence queuing simulation is run to find the estimates for queue length and waiting time.

The empirical analysis of queuing system of ICA supermarket is that they may not be very efficient in terms of resources utilization. Queues form and customers wait even though servers may be idle much of the time. The fault is not in the model or underlying assumptions. It is a direct consequence of the variability of the arrival and service processes. If variability could be eliminated, system could be designed economically so that there would be little or no waiting, and hence no need for queuing models.

With the increasing number of customers coming to ICA for shopping, either for usual grocery or for some house wares, there is a trained employee serving at each service unit. Sales checkout service has sufficient number of employees (servers) which is helpful during the peak hours of weekdays. Other than these hours, there is a possibility of short Queues in a model and hence no need to open all checkouts counters for each hour. Increasing more than sufficient number of servers may not be the solution to increase the efficiency of the service by each service unit.

When servers are analyzed with one queue for two parallel servers, the results are estimated as per server whereas when each server is analyzed with its individual queue, the results computed from simulation are for each server individually.

REFERENCES

Abolnikov, L., Dshalalow, J. E., Dukhovny, A. M. (1990), "On Some Queue Length Controlled Stochastic Processes," Journal of Applied Mathematics and Stochastic Analysis, Vol. 3, No. 4

Adan, I.J.B.F., Boxma1, O.J., Resing, J.A.C. (2000), "Queuing models with multiple waiting lines," Department of *Mathematics* and Computer Science, Eindhoven University of Technology,

Banks, J., Carson, J. S., Nelson, B. L., Nicol, D. M. (2001), Discrete-Event System Simulation, Prentice Hall international series, 3rd edition, p24–37

Bhatti, S. A., Bhatti, N. A. (1998), *Operations Research – an Introduction*, Department of Computer Science, Quad-e-Azam University, p.315–356

Hillier, F. S., Lieberman, G. J. (2001), *Introduction to Operations Research*, McGraw-Hill higher education, 7th edition, p834–8

Jensen, Paul A. (2004), "Queuing models," Operations Research Models and Methods, www.me.utexas.edu/~jensen/ORMM/models/unit/queue/index.html

Keith G. Calkins (May 2005), "Queuing theory and Poisson distribution," Statistical Probabilities and Distributions, Ch. 10, www.Andrews.edu/~calkins/math/webtexts/prod10.html

Nasroallah, A. (2004), "Monte Carlo Simulation of Markov Chain Steady-state Distribution," Extracta Mathematicae, Vol. 19, No. 2, p279-288

Sheu, C., Babbar S. (Jun 1996), "A managerial assessment of the waiting-time performance for alternative service process designs," *Omega, Int. J. Mgmt Sci.* Vol. 24, No. 6, pp. 689-703

Taha, Hamdy A. (1997), *Operations Research an Introduction*, PHIPE Prentice, 6th edition, p607–643

Tsuei, Thin-Fong; Yamamoto, W., "A Processing queuing simulation model for multiprocessor system performance analysis," *Sun Microsystems, Inc*

Troitzsch, Klaus G., Gilbert, Nigel (Sep 2006), "Queuing Models and Discrete Event Simulation," ZUMA Simulation Workshop 2006

_____, *Operations Management – Focusing on Quality & competitiveness*, EW University, Ch. 16 *education, 6th edition*, p834–8

APPENDICES

Appendix A: Questionnaire

This questionnaire is only for Employees serving at Sales checkout unit of ICA supermarket. Therefore, all information is intended specifically to the Service Unit.

Service provided to the customer at the spot (Quality of Service)

Please mark √ into ☐ and fill into the following questions:

1. At what time the customer arrived at the sales checkout unit?

   _____

2. Is there any other customer waiting for his turn, while the service unit is already serving another customer?

   ☐ No
   ☐ Yes

3. If yes, then how many?

   ☐ 1
   ☐ 2
   ☐ 3
   ☐ 4
   ☐ 5
   ☐ 6 or more

4. Mode of payment for purchase of items at checkout unit:

   ☐ Cash
   ☐ Credit card

5. The checking out operation service given to the customer for sales item was _____ for the customer.

   ☐ Sufficient
   ☐ Moderately Sufficient
   ☐ Insufficient / Incomplete

6. At what time the customer left the sales checkout unit after a successful purchase of items?

_____

Appendix B: Spreadsheets

Friday **Raw Data**

	Arrival Time	Interarrival Time	No. of People in a Queue	Departure Time	Service Time
1	16:50:00	0:00:00	2	16:51:00	0:01:00
2	16:51:00	0:01:00	3	16:53:30	0:02:30
3	16:53:00	0:02:00	3	16:55:00	0:02:00
4	16:54:00	0:01:00	4	16:55:00	0:01:00
5	16:55:00	0:01:00	3	16:56:00	0:01:00
6	16:55:00	0:00:00	2	16:56:00	0:01:00
7	16:56:00	0:01:00	1	16:56:45	0:00:45
8	16:56:00	0:00:00	1	16:57:00	0:01:00
9	16:57:00	0:01:00	0	16:59:10	0:02:10
10	16:58:00	0:01:00	2	16:59:00	0:01:00
11	16:59:10	0:01:10	1	16:59:40	0:00:30
12	17:00:00	0:00:50	0	17:01:00	0:01:00
13	17:01:00	0:01:00	0	17:03:00	0:02:00
14	17:03:00	0:02:00	1	17:03:30	0:00:30
15	17:03:00	0:00:00	3	17:03:40	0:00:40
16	17:03:30	0:00:30	2	17:04:00	0:00:30
17	17:04:00	0:00:30	1	17:05:00	0:01:00
18	17:04:00	0:00:00	0	17:05:00	0:01:00
19	17:05:00	0:01:00	0	17:06:00	0:01:00
20	17:07:00	0:02:00	2	17:17:00	
21	17:07:00	0:00:00	0	17:09:00	0:02:00
22	17:09:00	0:02:00	1	17:10:00	0:01:00
23	17:10:00	0:01:00	2	17:11:00	0:01:00
24	17:10:00	0:00:00	Enquiry not transaction		
25	17:11:00	0:01:00	1	17:12:00	0:01:00
26	17:12:00	0:01:00	2	17:12:50	0:00:50
27	17:12:50	0:00:50	2	17:14:00	0:01:10
28	17:14:00	0:01:10	3	17:14:30	0:00:30
29	17:14:30	0:00:30	3	17:15:00	0:00:30
30	17:15:00	0:00:30	4	17:18:00	0:03:00
31	17:15:00	0:00:00	5	17:16:00	0:01:00
32	17:15:00	0:00:00	2	17:16:00	0:01:00

33	17:16:00	0:01:00	0	17:17:00	0:01:00	
34	17:17:00	0:01:00	2	17:17:30	0:00:30	
35	17:18:00	0:01:00	0	17:18:30	0:00:30	
36	17:18:00	0:00:00	1	17:19:00	0:01:00	
37	17:19:00	0:01:00	2	17:20:00	0:01:00	
38	17:19:00	0:00:00	0	17:19:30	0:00:30	
39	17:20:00	0:01:00	2	17:21:00	0:01:00	
40	17:20:00	0:00:00	2	17:21:00	0:01:00	
41	17:21:00	0:01:00	2	17:21:40	0:00:40	
42	17:21:00	0:00:00	0	17:21:30	0:00:30	
43	17:21:00	0:00:00	0	17:23:00	0:02:00	
44	17:21:40	0:00:40	1	17:23:00	0:01:20	
45	17:23:00	0:01:20	1	17:23:30	0:00:30	
	17:08	**0:01:04**		**17:09**	**0:01:04**	Mean
	0:00:35	**0:00:35**			**0:00:35**	SD

mean arrival time =	0:00:44	seconds or	0.73333	minutes per customer
Arrival Rate =	81.818	customers per hour		
mean serving time =	1.0667	minutes per customer		
Service Rate =	56.250	customers per hour		
Duration of data collection =	0:33:00	min		

mean arrival time = 1/m

arrival rate = m or 1/(mean arrival time)

Wednesday **Raw Data**

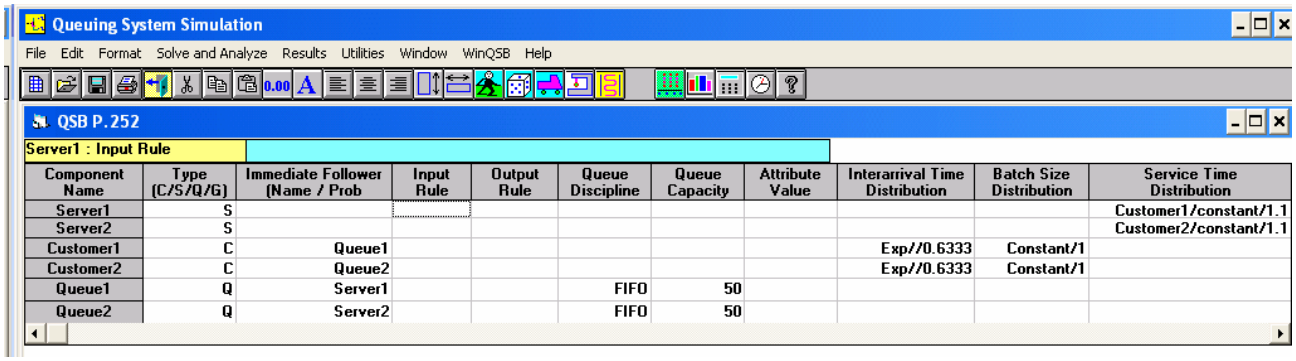
	Arrival Time	Interarrival Time	No. of People in a Queue	Departure Time	Service Time
1	16:25:00	0:00:00	1	16:26:00	0:01:00
2	16:26:00	0:01:00	1	16:27:00	0:01:00
3	16:27:00	0:01:00	3	16:27:30	0:00:30
4	16:27:00	0:00:00	2	16:28:00	0:01:00
5	16:28:00	0:01:00	2	16:28:30	0:00:30
6	16:28:30	0:00:30	3	16:29:00	0:00:30
7	16:29:00	0:00:30	4	16:31:00	0:02:00
8	16:29:00	0:00:00	3	16:30:00	0:01:00
9	16:30:00	0:01:00	4	16:31:00	0:01:00
10	16:30:00	0:00:00	1	16:31:00	0:01:00
11	16:31:00	0:01:00	2	16:32:00	0:01:00

12	16:31:00	0:00:00	3	16:32:00	0:01:00	
13	16:32:00	0:01:00	3	16:32:30	0:00:30	
14	16:32:00	0:00:00	5	16:33:00	0:01:00	
15	16:33:00	0:01:00	5	16:33:30	0:00:30	
16	16:33:00	0:00:00	5	16:36:00	0:03:00	
17	16:34:00	0:01:00	4	16:34:30	0:00:30	
18	16:34:00	0:00:00	3	16:35:00	0:01:00	
19	16:35:00	0:01:00	2	16:35:30	0:00:30	
20	16:36:00	0:01:00	1	16:37:00	0:01:00	
21	16:36:00	0:00:00	6	16:37:00	0:01:00	
22	16:37:00	0:01:00	0	16:37:30	0:00:30	
23	16:37:00	0:00:00	0	16:40:00	0:03:00	
24	16:37:00	0:00:00	4	16:38:00	0:01:00	
25	16:38:00	0:01:00	1	16:38:30	0:00:30	
26	16:38:30	0:00:30	1	16:40:00	0:01:30	
27	16:40:00	0:01:30	0	16:40:30	0:00:30	
28	16:40:00	0:00:00	0	16:42:00	0:02:00	
29	16:40:00	0:00:00	0	16:43:00	0:03:00	
30	16:41:00	0:01:00	1	16:43:00	0:02:00	
31	16:43:00	0:02:00	2	16:44:00	0:01:00	
32	16:43:00	0:00:00	1	16:44:00	0:01:00	
33	16:44:00	0:01:00	1	16:45:00	0:01:00	
34	16:45:00	0:01:00	2	16:46:00	0:01:00	
35	16:46:00	0:01:00	1	16:47:00	0:01:00	
36	16:47:00	0:01:00	2	16:48:00	0:01:00	
37	16:48:00	0:01:00	2	16:48:40	0:00:40	
38	16:48:00	0:00:00	1	16:49:00	0:01:00	
39	16:49:00	0:01:00	1	16:49:35	0:00:35	
40	16:49:00	0:00:00	2	16:50:00	0:01:00	
41	16:50:00	0:01:00	0	16:51:00	0:01:00	
	16:36	**0:00:38**		**16:37**	**0:01:06**	Mean
	0:00:40	**0:00:40**			**0:00:40**	SD

mean arrival time =	0:00:37	seconds or	0.61667	minutes per customer
Arrival Rate =	97.297	customers per hour		
mean serving time =	1.1000	minutes per customer		
Service Rate =	54.545	customers per hour		
Duration of data collection =	0:25 min			

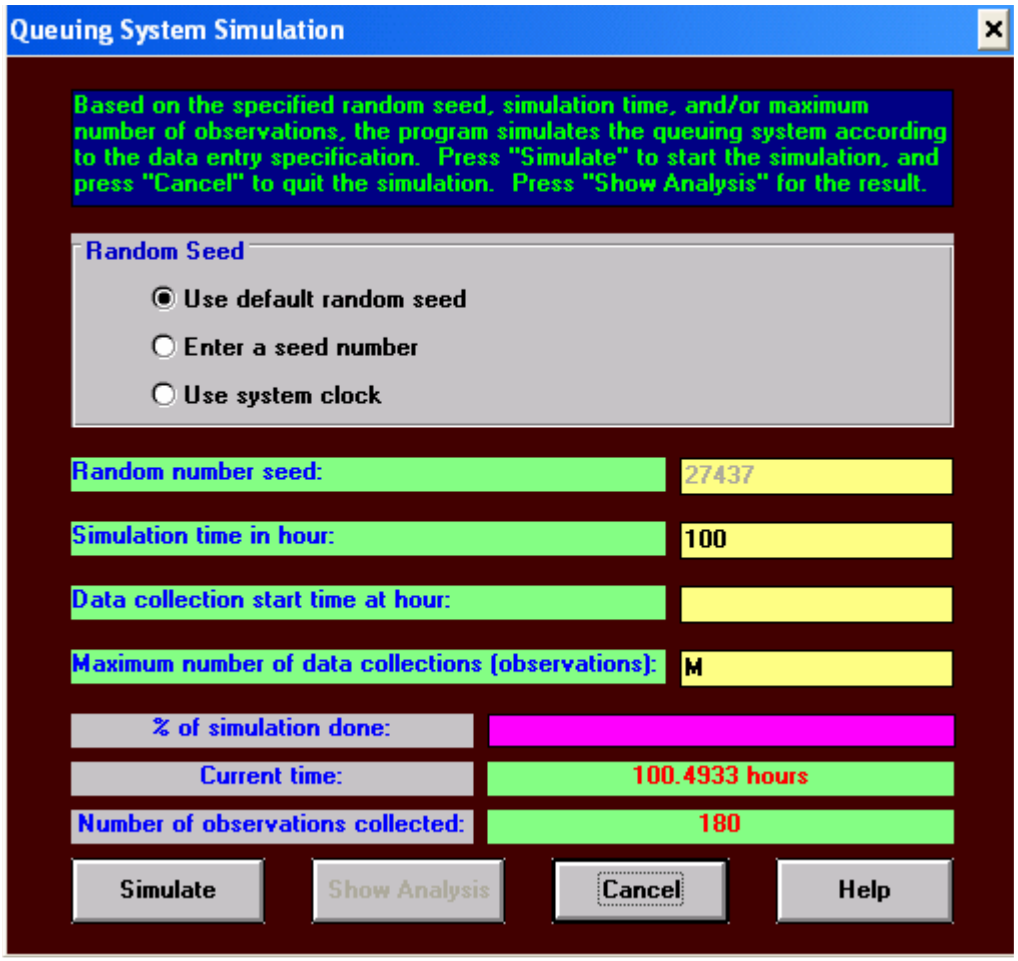
Appendix C: Queuing Software Input and Outputs

Data Input using mean service time and mean interarrival time:



Component Name	Type (C/S/Q/G)	Immediate Follower (Name / Prob	Input Rule	Output Rule	Queue Discipline	Queue Capacity	Attribute Value	Interarrival Time Distribution	Batch Size Distribution	Service Time Distribution
Server1	S									Customer1/constant/1.1
Server2	S									Customer2/constant/1.1
Customer1	C	Queue1						Exp//0.6333	Constant/1	
Customer2	C	Queue2						Exp//0.6333	Constant/1	
Queue1	Q	Server1			FIFO	50				
Queue2	Q	Server2			FIFO	50				

Simulation of the Data Input:



Overall result from simulation for 0 to 100 hours:

07-05-2007	Result	Customer1	Customer2	Overall
1	Total Number of Arrival	139	176	315
2	Total Number of Balking	1	35	36
3	Average Number in the System (L)	29.1828	40.3914	69.5742
4	Maximum Number in the System	51	51	102
5	Current Number in the System	48	51	99
6	Number Finished	90	90	180
7	Average Process Time	1.1000	1.1000	1.1000
8	Std. Dev. of Process Time	0.0012	0.0012	0.0012
9	Average Waiting Time (Wq)	21.1321	28.5634	24.8478
10	Std. Dev. of Waiting Time	11.6340	15.1996	14.0355
11	Average Transfer Time	0	0	0
12	Std. Dev. of Transfer Time	0	0	0
13	Average Flow Time (W)	22.2321	29.6634	25.9478
14	Std. Dev. of Flow Time	11.6340	15.1996	14.0355
15	Maximum Flow Time	39.5823	54.9761	54.9761
	Data Collection: 0 to	100 hours		
	CPU Seconds =	5.4490		

Servers result from simulation for 0 to 100 hours:

07-05-2007	Server Name	Server Utilization	Average Process Time	Std. Dev. Process Time	Maximum Process Time	Blocked Percentage	# Customers Processed
1	Server1	99.00%	1.1000	0.0007	1.1000	0.00%	90
2	Server2	99.00%	1.1000	0.0007	1.1000	0.00%	90
	Overall	99.00%	1.1000	0.0007	1.1000	0.00%	180
Data	Collection:	0 to	100	hours	CPU	Seconds =	5.4490

Queues result from simulation for 0 to 100 hours:

07-05-2007	Queue Name	Average Q. Length (Lq)	Current Q. Length	Maximum Q. Length	Average Waiting (Wq)	Std. Dev. of Wq	Maximum of Wq
1	Queue1	28.1820	47	50	21.3131	11.6965	38.4823
2	Queue2	39.3991	50	50	28.8511	15.3603	54.7423
	Overall	67.5812	97	50	25.0821	14.1626	54.7423
Data	Collection:	0 to	100	hours	CPU	Seconds =	5.4490