QUEUES IN SERIES WITH BLOCKING

Thesis submitted to

Cardiff University

For the degree of

Doctor of Philosophy

by

Jack Martyn Alec Baber

April 2008,

School of Mathematics,

Cardiff University

ACKNOWLEDGMENTS

My thanks go out to Professor Jeff Griffiths, whose outstanding guidance as my supervisor could not have been any better, Dr. Janet Williams, without whom I would never have embarked on this academic route. I have much to thank these two people for, their support and encouragement, has been constant. Thank you.

I would also like to thank Danielle and my Mother for their encouragement and always being available to talk. My friends, I must thank for their support at work and more importantly away from work.

My thanks are also extended to the staff at the University Hospital of Wales, all of whom have been friendly and approachable throughout. I would finally like to thank the EPSRC, who provided the financial support for this thesis.

Jack Baber
April 2008.

SUMMARY

This piece of work describes a hospital's Critical Care Unit and uses different mathematical techniques to model the behaviour seen there. The main factor that is included in these models is the problem of bed blocking in the Unit. Blocking is defined as patients who are well enough not to be in the Critical Care Unit, but remain there, for any number of reasons. These patients are using up an expensive and limited resource.

The mathematical techniques that the models are built on are extensively reviewed and analysed. These are the Coxian Phase Type Distribution and Networks of Queues with Blocking Equations. Both techniques are described in detail and their distributions analysed under different circumstances.

The final chapter shows how the two distributions can be used to model a complex situation such as the one found in the Critical Care Unit. The models are tested and compared. Finally, the models are tested under a number of 'what if' scenarios to predict the effect of changing certain factors on the actual Unit.

CONTENTS

Chapter 1

INTRODUCTION

1.1 About This Thesis

Critical Care Units are important parts of modern hospitals. These departments cater for the seriously ill patients that need special treatments or levels of care that cannot be provided on general wards. Given the wide range of health care available in a Critical Care Unit, and the high level of qualified nurses required to look after ill patients, the Critical Care Unit is very expensive to run. These departments care for many patients that are admitted to hospitals for planned operations immediately after their surgery, but also have to care for emergency patients. Balancing the cost of running the department and keeping beds available for emergency patients, whilst also trying to provide a through-put of planned patients to reduce waiting lists, is a complex task. Within this thesis we shall look at a large teaching hospital's Critical Care Unit and attempt to use mathematical techniques to aid the managers of the department with this complex balancing act.

This chapter will provide an introduction to the modelling of Critical Care Units, and outline the theoretical methods we shall use to model the department, namely Blocking equations and the Phase Type distribution. We shall also provide a review of the existing published work in these fields. Further chapters will be devoted to each of these topics individually, and then finally we shall bring together the data and the methods established in this work to create realistic models of the CCU. These will enable the unit's management to have a greater insight when process planning and making decisions.

1.2 Queueing Theory In Health Care

Operational Research has been used in the field of health care for many years. Europe has it own research group in the field, Operational Research Applied to Health Services (ORAHS) which was established in 1975. Its many members along with other Operational Research practitioners are all dedicated to using Operational Research in the field of health care from problems such as staffing Intensive Care Units (Price-Lloyd 2003), to managing the flow of the treatment of renal complaints in Europe (Davies and Davies 1987), to optimal spread of ambulance provision within a county (Taylor 1989). Many different techniques are used within Operational Research in the field of healthcare; Discrete Event Simulation (Jun, Jacobson et al. 1999), Systems Dynamics (Worthington 1991), Bayesian Belief Networks (Marshall, McClean et al. 2001), amongst many others. Healthcare professionals are also looking towards Operational Research to help them solve problems within their area; (Buhaug 2002) is a short article that cites examples of how Operational Research can be used in the field of healthcare and medicine. Its publication in the British Medical Journal is sure to be seen by many people who can be instrumental in bringing Operational Research into the field. (Preater 2002) has created a bibliography of many health related Operational Research publications. The bibliography has been split into five categories;

Appointments, Outpatients and Waiting Lists,
Departments,
Ambulances,
Compartmental Modelling,
Miscellaneous.

Though this is a substantial Bibliography there does not appear to be many papers on the subject of modelling a Critical Care Unit.

In (Harper 2002), the author creates a high level simulation of an entire hospital. It is used to demonstrate to health professionals that widely used existing methods of bed management are insufficient and unrealistic. It is commonplace to use simple deterministic spreadsheet approximations based on average occupancy. These calculations typically under-estimate the requirements of a hospital due to the length of stay in a bed being highly variable. The model then creates a relationship between bed occupancy rates and the referral rate; the referral rate in this case means patients that are not accepted in to the hospital. This gives the managing staff at the hospital a useful tool to help with their decision making. (Vasilakis and Marshall 2005) also considers the inaccuracies of taking simple averages to analyse the length of stay in a hospital. Many statistical tools were used within this paper, Survival Analysis, Compartmental Modelling and a Discrete Event Simulation program. The paper shows how these techniques compare on a large data set of over 10,000 stroke patients. Conclusions are reached about all of the methods, giving the pros and cons. The paper summaries by saying that all of the tools have the possibility to be powerful if used in the appropriate conditions. Even though these papers do not specifically look at the problems within Critical Care Units, the same observations can be made about simple deterministic approaches to bed management.

A review of papers that specifically look at the modelling of a Critical Care Unit shall be undertaken and is now presented. (McManus, Long et al. 2003) states that "variability in the demand for any service is a significant barrier to efficient distribution of limited resources". In this paper the limited resources are the critical care beds. Analysis was undertaken between different groups of patients and the differences between them in the model explained. The authors split the admissions into unscheduled and scheduled arrivals and analyse the refusals in each group. A regression model was created, and the correlation coefficient between the groups of patients and the chance of refusal was calculated. As we might intuitively suppose, the correlation between scheduled patients and refused was high. The paper also suggested that peaks in the refusal rates due to the unit becoming full are linked closely with the peaks in admission rate from scheduled arrivals, and by making the scheduled arrives less erratic, the refusal rate would also become more stable.

(Costa, Ridley et al. 2003) creates a CCU model that can be used in any hospital to aid decision making concerning the number of beds that a CCU should have. The authors break down creation of the model into four parts. Methods, this lays down the rules of how patients can enter the department. Whether they can be accepted, or if there is no room then the policy on delaying patients or even transferring them to a different facility. Data required, this is any data which could affect the occupancy rate of the department. Third is data analysis which involves a process called

Classification and Regression Tree analysis (CART). CART analysis breaks down large data set into smaller sets of similarly behaving groups; length of stay is then found for each of these groups. Finally, the model structure is created. A queueing process was created in this work but a computer program was used as the theoretical equations were very complex due to the nature of the units being analysed. Once the model was created, three hospitals were analysed to show the effect which changes in bed numbers has on the occupancy, referral and deferral rates. In one of the hospitals, the model showed that by adding four extra beds the deferral rate halved and the transfer of patients to different hospitals rather than being deferred quartered.

(Ridge, Jones et al. 1998) also considers the capacity planning of an Intensive Care Unit. The method used in this paper was to set up a simulation using the programming language PASCAL to create a "virtual ICU" with "virtual beds" and "virtual patients". Patients were put into different groups, defined by the hospital, and the length of stay distribution for each group was calculated by a statistical package. A priority system of arrival was set up so that emergency patients had the highest level of importance followed by patients that had been referred multiple times before. This simulation was compared with a theoretical queue with priority rules. In this system the customer who has higher priority jumps to the front of the queue, but does not interfere with the existing service. This is known as a non pre-emptive queueing system. It was decided that the theoretical distribution was too inflexible to model the real life situation, but it was used to decide warm up times for the simulation. The model was then created and used to predict what would happen in a number of different scenarios such as; varying the number of beds, if an admission was referred what effect would varying how many weeks a patients would wait before the operation was rescheduled, the effect of including a number of ring fenced beds for emergency patients. By ring fencing beds the emergency transfers went down a little, but the planned transfers more than doubled. This may seem to be a negative outcome, but the planned patients that were transferred made up a small proportion of the total patients transferred, as the total amount of transfers went up by only a couple of percent.

(Price-Lloyd 2003) undertakes a very thorough investigation of an Intensive Care Unit and a Discrete Event Simulation model was created. The simulation was able to find optimal numbers of beds under current demand and predict what it would be given future predictions of demand. The simulation also had the ability to schedule nurses in the ICU. If there are not enough nurses for the number of patients, the hospital is forced to employ agency nurses, which are much more expensive than the nurses employed directly by the hospital. Cost analysis was used on the results of the simulation and the optimum number of nurses to employ on the unit at any time was found. These results, as well as being useful as a mathematical exercise, were implemented by the hospital in both the planning of resources and of beds.

1.3 Queues in series

The first mathematical method that we shall use to aid our understanding of the Critical Care Unit is Blocking equations. This is a specific application of Queueing Theory that can deal with customers that are blocking a service point. This is something that is very interesting within the context of a hospital. Patients that are blocking beds, especially in Critical Care Unit, are wasting capacity in a highly utilised department. We shall attempt to use these equations to see the effect of blocking on the department. We shall also be able to analyse the results of 'what If' questions. These provide a way of investigating what the unit's performance if different scenarios are adopted, without wasting time and money on trialling initiatives in a live environment.

Queueing theory is a well recognised approach to solving problems within Operational Research. It has applications theoretically and in real life situations. Queues in series are an application of queueing theory where regular queueing systems are put together so that the output of one system is the input, in part or in whole, for another queueing system. Jackson first brought the idea of queues in series to attention in a paper from 1957 (Jackson 1957).

Kendall created a method of expressing queueing systems using a simple notation. Consider a basic queueing system; 1 server who serves in a Poisson fashion, with an infinite queueing space. This space is used to house customers that arrive whilst the server is occupied. The customers arrive at a random and are seen in a first come first served manner. This system can be described in Kendal's notation as M|M|1|∞|FIFO queueing system and has been widely studied. A useful way to understand a queueing system is to find out what state the system is under specific conditions and with specific parameters. A system's state can be described as whether there is a customer being served or not and whether there is a queue and if there is what size it is. These states can be given representative mathematical symbols and the computed under certain conditions. They can be considered in, among other ways, a steady state. A steady state is when the system behaves in a stable manner as time passes. The queue size may alter in this stable manner but it will not grow beyond control and the most probable queue size will be steady. If the system is not in a steady state, the queue size will expand and as time increases the queue size will continue to increase, creating an unusable system. The M|M|1|∞|FIFO queue has been found to have a steady state solution with the condition that the arrival rate is less than the service rate. This makes sense intuitively; if customers arrive more frequently than they are served then we would expect the size of the queue to build up as time passed.

If there exists 2 or more M|M|1|∞|FIFO systems such that the output of the first system is the input for the next system, they are referred to as queues in series, or sometimes as tandem queues. Similar restrictions on the arrival rates are required in this tandem queue for this system to be in a steady state. It was thought that the additional systems would make computing this arrival rate more complicated however (Burke 1956) showed that the input for the second service point of 2 queues in series is the same as the input into the first server, implying that the output of the first queue is the same input as for our first queue. (Reich 1957) goes on to show a similar and more general result which allows for multiple servers in a single service station, and still shows that the output rate of the first server is the same rate as the input of that server. These results can of course be extended to the case when there are more than 2 queues in series. This makes finding the requirement for a steady state solution much simpler. These results lead us to a result of queues in series called product form.

For queues in series to be of product form the size of the queues in front of each server must be infinite. In this case, as customers pass from one phase to the next, there is always a space in the next queue for the arriving customer. This product form property is very useful but is limited. (Reich 1957) shows that when the service distributions are Exponential the system is of product form. By using the reversibility of a Markov Chain, in the case where service patterns are Erlang, the resulting system can not be shown to be of product form. (Jackson 1957) considers the system with k service points each with an infinite queueing space, and when a customer leaves one service point they go to any other with a fixed probability. The author goes on to show that this system has product form if each individual service point is in a steady state. For that to happen the service rate at each service point must be greater than the total arrival rate, including external sources and from all other service points. Systems that do not have the product form property maybe approximated by it or some adaptation of it to suit the differences in the system. Many papers show approximation of different scenarios as adaptations of networked queues so that they have the product form property.

To calculate the probability of what state a tandem queueing system is in, if it has the product form property, simply multiply the probability of the first queueing system being in the correct state by the probability of the second queueing system being in the correct state. We can do this because the product form states that these two queueing systems are independent. If we let $P_i(n)$ be the probability of there being n customers in the i-th service point in tandem queue and $P(n_1,n_2...n_i...)$ be the probability that the tandem queue has n_1 is the first system, n_2 in the second and n_i in the i-th point, then $P(n_1,n_2,...,n_i,...)=P(n_1).P(n_2).....P(n_i)....$. This result can be found in (Cox 1954).

Now suppose a limit is imposed on the size of the queue for a service station on any service station other than the initial one. When the capacity of that service station is reached, no more customers can enter it. In this system when a customer completes a service at the previous phase there is nowhere for that customer to go and the customer must remain in their current service facility until space is made available in the next service facility; this is called blocking. Blocking can occur between any stages in a tandem queue and can be dealt with in many ways.

(Akyildiz 1989) defines three different types of blocking; Transfer (or Manufacturing) blocking, Service (or Communication 1) blocking and Rejection (or Communication 2) blocking. In the transfer blocking situation, when a customer becomes blocked they remain in the service channel and no more customers can be served until space is available for the blocking customer to leave. Service blocking says a service can not commence if there is no space for the customer in the next facility once the service has finished. So in this case the customer waits for an available space before starting service and the blocking occurs before the service. This method has obvious advantages on transfer blocking but is less efficient. Finally with rejection blocking customers start their service as soon as possible but if there is no space on completion of service the customer restarts their service. All of these are available with multiple phases and with multiple routes in and out of phases and have been widely researched. Each of these types of blocking has many applications in telecommunication networks, production lines and computer processors

Queues in series with blocking, have been studied since around 1950. One of the first papers to consider them is (Hunt 1956). The blocking element of queues in series affects the utilisation of the system, as the system is made less efficient. In tandem queues, which allow an infinite queue in front of every service facility, each can be treated independently. Each service facility, which will be referred to as a node, is of the form M|M|1. For an M|M|1 system to remain in a steady state, the arrival rate must be less than the service rate. It is usual to say that customers arrive in a random fashion with a mean rate of λ and are served with a Poisson distribution with parameter μ. So for the system to remain in a steady state we have that $\frac{\lambda}{\mu} < 1$. The ratio of the arrival rate to the service rate is referred to as the traffic intensity or ρ. (Cox and Miller 1965). The traffic intensity provides the proportion of arrivals relative to the customers served. For a traffic intensity value less than one, the system is in a steady state as the rate of arrivals is less than the rate which the customers are served at. When the traffic intensity is greater than one, the system is no longer in a steady state. The queue to this service facility will keep growing. With networks of queues the utilisation is reduced because of blocking. The time which the initial node is occupied now consists of the service time and the time that node is blocked for. As the server is unable to accept customers during either of these states, the utilisation decreases. This means that the average inter-arrival time will have to be less than the sum of the average service time and the average blocked time. (Hillier and Boling 1967) extends Hunts results by considering systems with any number of service points. Hiller finds the maximum utilisation and mean number of customers in the system when all service points serve with Exponential or Erlang (excluding first point) service times. The author uses a matrix method accompanied with a 'special algorithm' as the transition matrix is so sparse. The paper gives extensive numerical results for the maximum utilisation and mean number in the system for a wide range of service points with different maximum queueing space and service rates.

(Hunt 1956) sets up time dependent equations of a 2 node system and steady state solutions are found. These equations will be set up in Chapter 3 but solved using a different method. Hunt uses a method of raising operators, a matrix approach to find ρ which he defines as the maximum utilisation for the system, "In the general N-state problem, blocking occurs more frequently in the first stage than any succeeding stage and the maximum possible utilisation for the first stage is the maximum possible utilisation for the entire system" , (Hunt 1956). To find the maximum arrival rate from the maximum utilisation it must be multiplied by the service rate in the first phase. This accords with intuition as the utilisation of the first node multiplied by the average rate at which customers, when not blocked, are served at will give the average rate that Phase 1 processes customers including their blocked time. Hunt then goes on to apply a similar theory to a three stage model with no intermediate queues. This thesis will also show results for the three stage case. Hunt goes on to look at the case of a fixed queue size in between phases but in a limited manner.

(Avi-Itzhak and Yadin 1965) looks at the case of a zero size queue in between 2 phases and also the case where there is a fixed size queue between the phases. The author starts by looking at Poisson arrivals and an arbitrary service distribution and gives examples of Exponential and regular service rates in detail. The author uses moment generating function of customer's time in system and shows that the result is symmetrical with respect to the services rates at each phase and is therefore indifferent to the order of the stations. Also found is the number of customers in the system by using Little's formula.

The concept of reversibility of phase is not something considered here though it is an interesting property that may be useful in applying queues in series to other applications. (Yamazaki, Kawashima et al. 1985) looked further into C-reversible queues, so called as the capacity remains invariant under reversal of the system. It had already been shown that C- Reversibility holds for single server blocking queues with non deterministic service times and for multi-server of deterministic service times. Yamazaki showed that this property can also be applied to multi-server nondeterministic service time but only in a 2 stage case and that it is not true for more than 2 phases. So not all systems can be rearranged or reversed to give the same capacity. One could look at this in production lines of any sort to try and create a maximum flow as an interesting problem.

(Hildebrand 1968) sets up the equations of queues in series in a different way from previous papers. He uses waiting, blocking and vacancy times to set up equations to find the maximum throughput. This gives a general method for queues in series with Exponential service and *m* phases which is more usable than previous papers. A useful bibliography has been prepared on the subject of queues with blocking by (Perros 1984), citing both theoretical papers and applications of tandem queues. In total 75 papers are cited with those mentioned so far amongst the earliest of these. Since this paper, another bibliographic paper has been produced, (Papadopoulus and Heavey 1996). This paper cites 257 papers on open queueing networks. Papers that are cited are generally in the area of manufacturing industry as the title of the paper suggests and are mostly attempting to solve manufacturing problems with theoretical reasoning. The authors create their own classification system to create the bibliography and suggest that it should be adopted it the wider field, though there is little evidence that it has been widely accepted.

As mentioned earlier some queues in series with blocking can be approximated using product form. (Akyildiz 1989) goes further and says that all queues where blocking occurs can be approximated by a product form solution. This is not an easy task and is done by normalising infeasible states that violate station capacities to produce mean number of jobs processed. For the throughput, a queueing network that does not allow any blocking is used as an approximation to the blocking system. The paper looks only at closed queueing networks. Closed networks do not allow customers to leave or join the system, the only customers in the system are those present at the creation of the network, these queues are often called cyclic as the same customers cycle around the system. These closed networks have restrictions on them to avoid a situation called deadlock. Deadlock as expected, is when the network cannot process customers as all available space is taken up, to avoid this situation there must be strictly more spaces in the system than there are customers, so that there is always a space to move into somewhere in the system. In (Akyildiz 1988), the author also looks at closed queueing networks but this time uses Mean Value Analysis as an approximation for a blocking network. Mean value analysis is a method of finding the total time spent in a phase of a non blocking network including queueing and service time. To adapt this for a blocking case the author includes an algorithm to include the blocked time. This method is preferred by the author as it has a very quick computation time and also does not require a lot of memory. Another issue with cyclic queues is finding the maximum utilisation. This is difficult as there are a fixed number of customers in the system, which can often be a set amount to avoid the deadlock situation. (Onvural and Perros 1989) cites other examples of cyclic systems that have product form solutions. The paper then goes on to find a function for the maximum utilisation in terms of the number of customers in the system. This is found by computing the throughput of the system and specific service points for an increasing number of customers in the system. This produces a curve which can be approximated to find the maximum utilisation for cyclic systems.

Some papers have used queues in series in conjunction with other theoretical queueing techniques so that more specific problems can be solved. In (Browning 1998), the author looks at dependent and independent service times in an *m* phase system. Dependent service times refer to customers that will have the same time of service at both service points. The paper goes on to show that in this specific case if there is a large enough queueing space before each phase then the throughput of the dependent system is greater than the throughput of the independent system, provided service time is not deterministic and has finite variance. (Pinedo and Wolff 1982) also looked at this specific case of queues in series and found that under similar conditions, but without a queue between the phases, the converse was true; that the maximum throughput for the dependent case is less than or equal to that of the independent case. (Moutzoukis and Langaris 2001) looked at another specific use of queues in series, those that have retrial customers. Initially considered is a 2 phase model with no queueing spaces at all. If a customer arrives to find the first phase busy, either occupied or blocked, then that arriving customer moves to a retrial area. Customers in this retrial area attempt to rejoin the system with a rate specified in the area; this could be, for example, μ or, as in this case, $\frac{\mu}{n}$. This allows for a fixed arrival rate no matter how many customers are in the retrial area. The author evaluates this system with general service rates, and solves it for Exponential service rates to find maximum process rates. (Rhee and Perros 1996) looks at a problem in a network queueing system with a semaphore queue. This is an open queueing system but limits the number of customers present at any time. A token procedure is defined to cope with this. As a customer arrives they take a token, and once all the tokens have been taken customers are forced to wait in a queue outside the system until a token becomes available from a customer leaving the system. The author find bounds on the mean waiting time for this system and shows that reversibility holds when service times are Poisson and inter-arrival times are random. (Ahn, Duenyas et al. 1999) looks at multiple servers in a 2 stage tandem queue with no blocking allowed and where servers are free to move between the 2 service points. The aim of this paper is to minimise the holding cost for the whole system by utilising the servers at the appropriate service points. (Hillier and So 1995) considers how to optimise a tandem queue by the altering of service rates, queue capacities and number of servers to change throughput. The self stated main aim of the paper is to "to help open up this research" as there is little published work in the area. The paper looks at altering each of the variables individually and in pairs. The authors cite other papers that have looked at these fields but are unable to find one that attempts to optimise a network of queues altering all three. The authors do not cite work that involves multiple servers other than their own work, (Hillier and So 1989), which allows a pool of servers to be assigned to different service points.

Whilst queues in series seem to be a very appropriate way of modelling health systems, there is very little literature on the subject. (Koizumi, Keno et al. 2005) considers the case of a mental health institution that has patients spending unnecessary days in intensive facilities when they are well enough to leave and move to a less intensive level of care. The model describes four levels of care; Acute hospitals, Extended Acute hospitals, Residential facilities and supported housing. Patients can leave the system and join the community at any stage; in fact the model allows for patients to move from almost any service point to every other service point. However, blocking is only allowed to occur between the extended acute hospital and the residential facilities, and between the residential facilities and supported housing. Equations were set up for this system and it was found that the system was not in a steady state. To overcome this problem, extra capacity was added to the over utilised areas and the results were then found. The results of a simulation model were also included in the paper, as validation for the mathematical approach.

(El-Darzi, Vasilakis et al. 1998) looks at a geriatric hospital and notes that the average length of stay in the acute department is artificially high, if long term patients are kept there until a bed becomes available in the less intensive facility, a residential or nursing home. The paper gives an overview of how geriatric bed planning is run in the UK and states that average lengths of stay, bed occupancy and emptiness are used, which as we have already discussed are unrealistic. Two, 3 stage simulations were set up and compared. The first was a Phase-Type model which did not allow blocking, but blocking was then included on the second simulation. Analysis was then undertaken on the sensitivity on the models. It showed that the model that did not allow blocking was much more sensitive to change in the parameters than the one with blocking. (Weiss and McClain 1987) also considers a case of geriatric health flow. Theoretical equations were set up and solved. The system was then tested on data from seven different hospitals. The simulation results were validated by a goodness of fit test. The authors then went on to show, if specific groups of patients spent less time in administrate days, what the effect would be on the hospital.

1.4 Phase Type Distributions

The second mathematical technique we shall use in the analysis of the department is the Phase Type Distribution. This distribution is very versatile and it has a proven record in health care modelling.

The Phase Type distribution consists of a number of Exponential service points which customers pass between. To exit the service there exists an absorbing phase which the customer can go to anytime throughout their service. We shall be looking at a specific subset of these Phase Type distributions, called Coxian Phase Type distributions, named after D.R.Cox and his paper (Cox 1954). In this class of distribution the service points must be in sequence and once completed the service point is not revisited. The absorbing phase still exists, which customers can go to at any point. This class of distribution is used because the estimation of parameters in the standard Phase Type distribution can be very complex, (Faddy and McClean 1999). In the standard case there is no maximum or fixed number of parameters to be estimated, whereas with the Coxian version there are $2n$-1 parameters to be estimated, where n is the maximum number of Exponential service points through which a customer can pass, excluding the absorbing phase.

In (Cox 1954) the author does not use the words Phase Type distribution to describe his model, but he does clearly describe the case. In this paper, the author show two different ways of establishing a Phase Type distribution and shows that they are equal. In the first case before the customer starts their service they have a probability of moving straight to the absorbing phase, or commencing their service. If they start their service the first point has an Exponential service time and then another chance of moving to the absorbing state or carrying on with their service. This happens at the end of every service point. This system is compared to the one where the customer is able to move to the absorbing phase at any point in time and does not have to wait until the service at any particular point comes to an end. These two systems are shown to be the same in this paper.

Phase Type distributions are used in (Faddy and McClean 1999) to model patients' length of stay times in a geriatric ward. In this paper the authors describes the length of stay of patient as short, medium or long stay. The parameters for each of the phases in the distribution are then calculated. The author makes an important point in the study of Phase Type distributions when noting that each phase could respond to an increase or decrease in the severity of the patient's condition but this is not essential, but it is useful especially in describing to the process to non mathematicians. This means that patients do not need to appear in different compartments or wards to be able to use the Phase Type distribution; even different levels of illness do not need to have official compartments. The Phase Type distribution's flexibility can account for the differences in the times. The paper goes on to find a length of stay distribution which is based on a sample of 2090 patients. It was found that a Phase Type distribution with four phases fitted the data. Five phases were also tried but the improvement did not justify the added complexity. The lengths of stay was also shown for short stay, medium and long stay patients, defined by the phase from which they joined the absorbing phase.

(Gorunescu, McClean et al. 2002) set up a bed management process that uses the Phase Type distribution to model patient length of stay. This process was created to allocate patients to beds without wasting hospital resources. The theoretical equations of a Phase Type distribution are analysed alongside a stock control model. Costs are given to holding an empty bed and of turning patients away, as well as the cost of care for a patient. These costs can then be optimised and the number of beds for a given reject rate can be found. Examples were given in the paper of how this generic model can be used. The analysis shows that for the hospital reviewed to get zero patient lost the unit has to run with average bed occupancy of 84%

(Harrison 2001) uses compartmental modelling to model length of stay in a hospital. The author compares two theoretical compartmental models, where a type of Hyper Exponential service point is compared to a Phase Type model. The Phase Type model was found to be more practical for altering patient flows as it is a more flexible distribution. It is natural ease to interpret, was a further an advantage of the distribution, though it was found to over complicate the fitting of data in some examples. The author gives two examples of American hospitals. A private hospital in which a Hyper-Exponential Distribution fitted the data as well as a Phase Type distribution, and a veteran's hospital which the Phase Type distribution did a much better job of fitting the data. The author comments that a similar effect can be seen in British hospitals, and when there are a small number of patients that stay for a long period, the Phase Type seems more appropriate.

(Marshall, McClean et al. 2002) uses Phase Type distributions coupled with Bayesian belief networks. These Bayesian belief networks are then used to classify patients so that a more accurate length of stay distribution can be used to model their length of stay. Factors considered are gender, admission source and measure of the patient's dependency on entry to the hospital. The distribution in the subsequent Phase Type model is then dependent on these factors when being calculated. In a paper by (Marshall and McClean 2004) looks at Phase Type distributions in a different manner. A Phase Type distribution was set up on data from a London hospital. The patients were then grouped depending from which phases they joined the absorbing phase. Patients' details were then compared to see if similar characteristics were seen in each group. This was done so that on a patients' arrival they could be assessed and an approximation for their length of stay could be found. This is to aid bed management schemes. The technique proved useful especially in the second and third phases. The patients that joined the absorbing phase from these phases had significantly different characteristics from those in the first phase. These patients that spend a long time in the unit consume many resources and this tool may help is predicting which patient maybe in the long stay group, which should aid with planning.

1.5 Summary

The aims of this thesis have been described in this chapter. We shall model the behaviour of a Critical Care Unit with mathematical techniques. The methodology uses Phase Type Distributions and Blocking equations. Once a model of the Critical Care Unit has been established it will be possible to alter variables in the model to simulate what might happen to the unit if a change was able to occur. An introduction to these two tools has been given and a review of the published literature in each field provided. Phase Type Distributions have been widely used in the study of healthcare; however, their penetration into the modelling of Critical Care Units seems limited. Blocking equations have not been implemented in the field of health care to any significant level. We shall go on to show how each of these techniques can be used to model Critical Care Units and how they can help the management of these departments.

Chapter 2

DATA REVIEW

2.1 Introduction

This chapter will provide an overview of the data of the system that shall be modelled by some of the methods that will be developed in this thesis. It will show trends in the data and an explanation of how we shall attempt to model various scenarios using mathematical techniques.

All the data for this piece of work has been generously supplied by the University of Wales Hospital (UHW), the largest hospital in Wales. Over many years Cardiff Mathematics department has built a thorough working relationship with UHW. The most recent work has been undertaken in conjunction with the Critical Care Unit's director and the unit's data analyst Mr Martyn Read. The director is an experienced doctor, working in both medical and managerial roles at UHW. The data analyst, who is also a medical consultant, has worked at UHW for many years and has a great understanding of all the processes in place at the hospital and of the immense amount of data that the hospital produces. Working with staff at the unit has proved to be invaluable; being able to visit the department to understand where the data is sourced, understand data inconsistencies, both medical and managerial reasons for decisions being undertaken, and countless other insights.

The Critical Care Unit (CCU) is a facility that hospitals use to care for the critically ill. Some hospitals use two separate units, an Intensive Care Unit (ICU) along with a High Dependency Unit (HDU). HDU patients are usually less severely ill than those in the ICU. Patients in the ICU must have a nurse to patient ratio of one to one, whereas the HDU operate at a ratio of one to two.

The UHW used to operate under the two unit system until they combined on the 1st April 2004. Patients are now situated in one ward, but are classified as either level 2 or 3. Level 2 patients require a nurse patient ratio of one to two, whereas those with a level 3 condition receive one to one care.

Patients can go to the CCU for many different reasons; frequently elective surgery patients spend some time in the CCU once their surgery is over to be closely monitored. Emergency operations could also result in the patient having to spend some time in the critical unit; patients can arrive to the CCU from other hospitals that do not have the facilities to look after them. As such the CCU is a much sought after resource. It is also very expensive to run due to the high staffing levels and the hi-tech equipment required around each bed to constantly monitor the patients. As mentioned above, many patients that under-go elective surgery require a bed in the CCU once their operation is completed. This bed is always confirmed before the patient is prepared for surgery, if the CCU is highly utilised then the elective patient's surgery may be cancelled so that the unit does not become unable to deal with any emergencies that may occur. This will obviously impact on hospital waiting lists and on customer service as operations can be cancelled at very short notice.

2.2 The Data Source

In the CCU, some live data is collected hourly for each patient, such as medical records and any interventions that occurred; however a more thorough data snapshot is recorded using the Riyadh Intensive Care Program Unit (RIP) only once a day at midday. The RIP is a large database program which is mounted on a networked computer at all the beds in the CCU.

Systems records data ranging from simple fields, such as temperature and blood pressure, to more complex fields such as the Glasgow Comma Score, a physiological score that is based on eye, motor and verbal response of the patient, and the

Therapeutic Intervention Score System, which is a method indicating how ill a patient is by recording any interventions the patient has, which are then weighted according to their severity. The RIP also records patients' personal and demographic information, which is collected only once either on the patients arrival or on their departure. These include; name, address, age, gender. The time of arrival and time of departure is also recorded in the RIP, enabling us to compute the length of stay of all the patients on the unit. This can be separated by arrival source or surgery type, providing a great deal of granularity. One thing that is not recorded by the RIP is what we shall refer to as the 'ill length of stay'.

Before a patient is able to leave the CCU they must be referred to a regular ward by their doctor. The doctor usually does this during their morning rounds at 08:00. Any patients that do not require the level of treatment provided on the CCU are referred to other hospital wards keeping the valuable CCU beds available for the seriously ill patients. The transfer can not happen instantaneously, as the ward bed may need to be prepared, the ward pharmacy needs to be able to serve the patient and porters need to be available to move the patients. After discussing this situation with the Director and after some data analysis, it was recommended that 7 hours should be enough time to complete these formalities. From the time a patient enters the CCU to the time that they are referred to leave the CCU, plus 7 hours, will be defined to be the 'ill length of stay'.

The time of referral out of the department is not recorded in the RIP so the ill length of stay can not be calculated directly from that source. However, as a patient processing initiative, undertaken by the CCU, the referral time has been recorded on a separate database. This data was recorded from 2nd February 2004 to 31st December 2005 and the database contained 3600 unique entries from the UHW and Llandough Hospital. Llandough hospital is a much smaller hospital within the same Trust, which shares its data with the UHW. After another discussion with The Director, it was recommended that this database should be used only for the time since the ICU and HDU amalgamated. Once the data in the time period prior to the amalgamation and the Llandough data was removed, and then some data cleaning was done, 2520 entries were remaining. For the same time period the RIP recorded 2876 different departures from the CCU. At first glance this seems very positive, with only 356 patient data not recorded. However, in the smaller database many patients have multiple entries. Every time a patient is referred, an entry is placed into this database, and as we will see, many patients do not leave on the day of the original referral resulting in many multiple entries. Removing these duplicates leaves 1474 entries. This means that around 50% of the data has the available fields to calculate the ill length of stay. The hospital's data analyst explained this observation by stating the data were only recorded in a sample of the beds on the CCU. In the next section the RIP data of the whole CCU will be analysed and compared to the smaller database. During this analysis the attributes of the whole population will be compared to the sub-population which have the time of referral recorded.

2.3 The Critical Care Unit Data

Figure 2.1 shows that over the 2 years analysed, 55% of the patients seen in the CCU are male. The majority of patients fall in the age bracket 50 to 80. By excluding the final 2 columns (age 90+), the chance of a patient surviving can be seen to decrease the older a patient is. The final 2 columns buck this trend, but as shown by the yellow line, the volume of patients over 90 is very small, resulting in unreliable statistics.

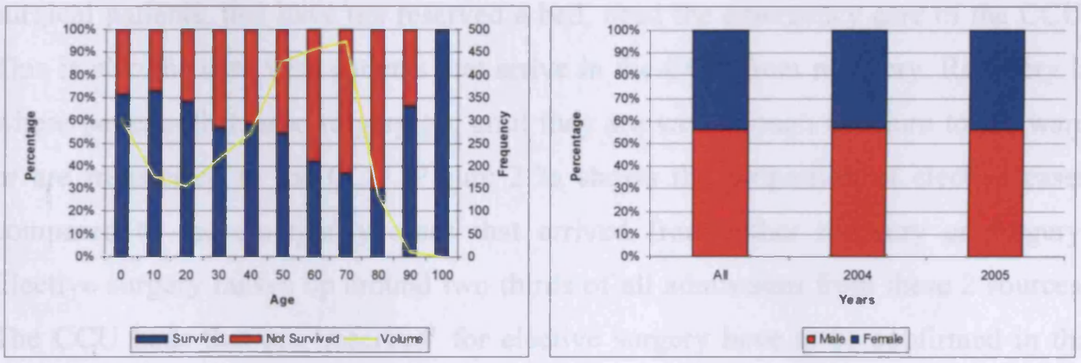


Figure 2.1 - Gender percentage by year (Right) & Percentage of patients that survived by age.

Due to the low number of patients aged 80+ that pass through the CCU, Figure 2.2, which shows the age distribution by year, has grouped patients above 80 years old into one class. The distribution of the ages is consistent across the 2 years of data that are being analysed.

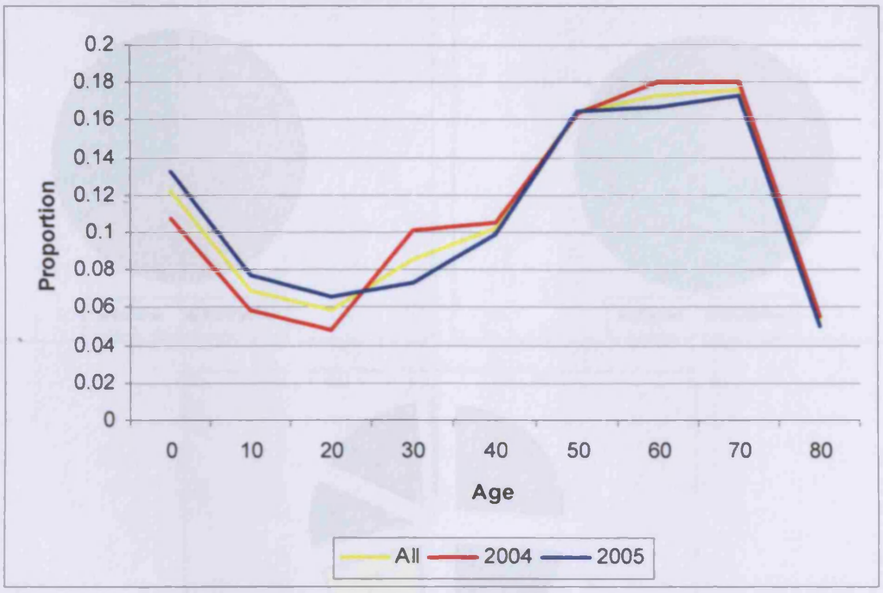


Figure 2.2 - Distribution of age by year.

Figure 2.3c shows where patients have come from when they arrive at the CCU. The top left chart shows the physical location from where they came. The largest proportion comes from the theatres, following a planned surgical procedure. Once the surgery has been completed, the patient needs to spend time with the intensive nursing staff and systems available in the CCU. This visit to the CCU can be planned in advance; in which case a bed is 'reserved'. Not all patients that have elective surgery require the use of the CCU, so a bed is not always reserved. Sometimes planned surgical patients, that have not reserved a bed, need the emergency care of the CCU. This is also the case with patients that arrive in the CCU from recovery. Recovery is where patients that have surgery go, until they are well enough to return to the ward or are transferred to the CCU. Figure 2.3a shows the proportion of elective cases compared to the emergency cases that arrived from either recovery or surgery. Elective surgery makes up around two thirds of all admissions from these 2 sources. The CCU beds that are 'reserved' for elective surgery have to be confirmed in the morning of the surgery, so that if the CCU is approaching capacity they have the right to stop the surgery so that the CCU does not become over-utilised and hence is not prepared for any emergencies that may occur.

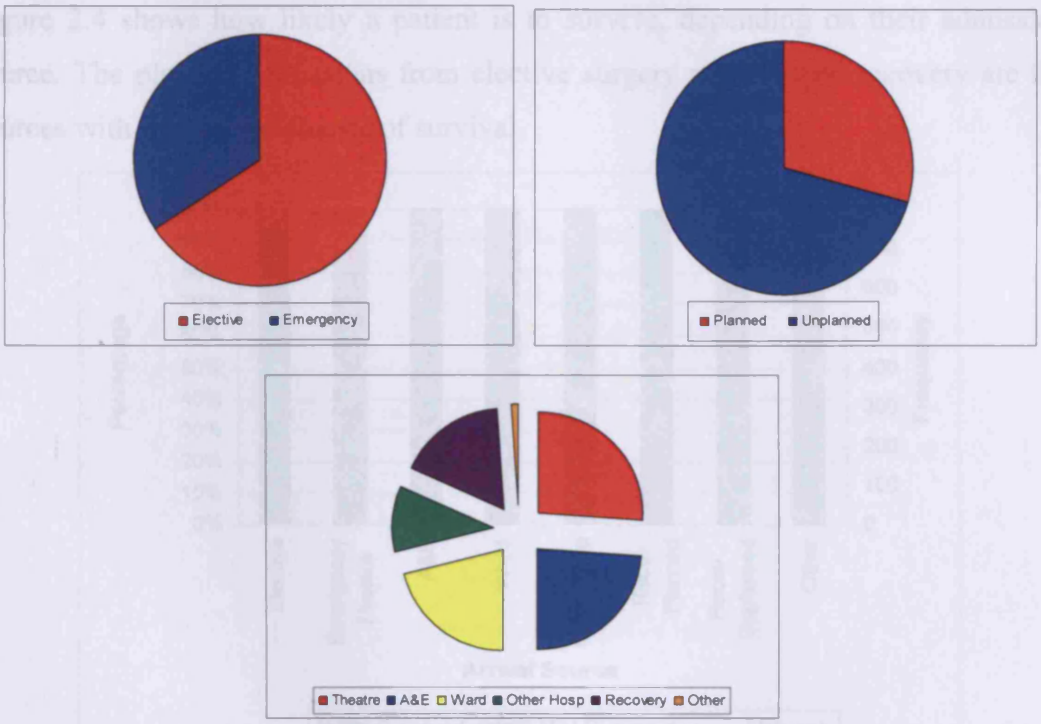


Figure 2.3a- Proportion of surgery arrivals that are elective or emergency (left), 2.3b - Proportion of planned and unplanned arrivals (right), 2.3c - Patient arrival source (below).

Figure 2.3b shows the proportion of planned and unplanned admissions to the CCU. The unplanned cases include those from the recovery or the theatre that had not reserved beds, plus the other sources of arrival seen in the chart on the left; A & E, Ward, other hospitals and the 'other' section, which includes X-ray and other sources of Radiography. In unplanned cases the CCU is not given much prior notice of the patient's arrival. The CCU is informed about patients that arrive from other hospitals, but without very much warning and are not given much opportunity to refuse them. This is because these patients are usually severally ill, and the smaller hospitals that these patients come from do not have the facilities to look after them. So in this definition, the planned cases are those that the CCU has the ability to cancel.

Figure 2.4 shows how likely a patient is to survive, depending on their admission source. The planned admissions from elective surgery and planned recovery are the sources with the highest chance of survival.

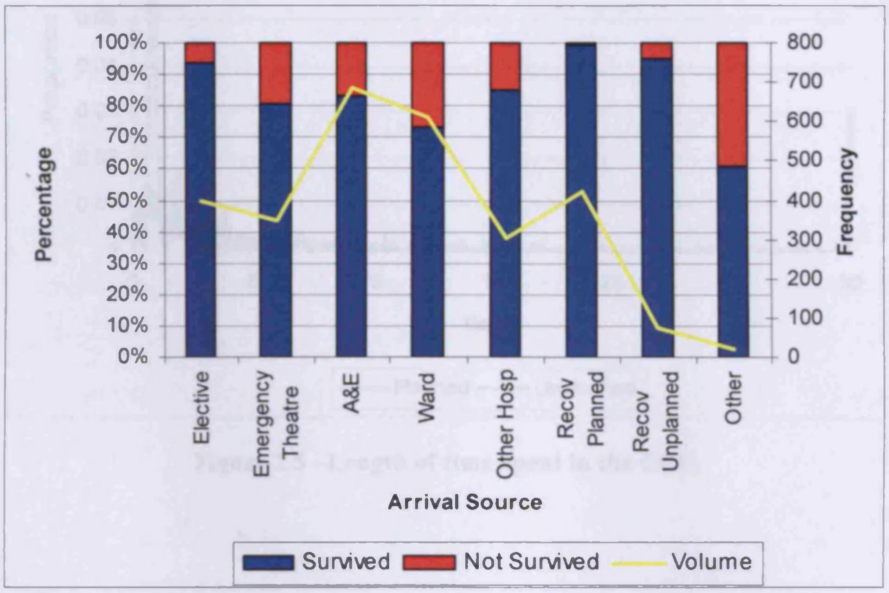


Figure 2.4 - Percentage of patients that survived by arrival source.

The length of stay for patients can be seen in Figure 2.5. This shows a clear spike in the planned cases at around one day. The modal value of the planned distribution is 21 hours and the mean being 40.3 hours. The high spike quickly drops to leave the distribution highly skewed. A long tail occurs in this data, with the longest recorded time a planned patient stayed in the CCU being 54 days. The unplanned cases have a much less peaked distribution. When the large peak at the end of this truncated distribution is ignored the mode is 19 hours, and the mean is 13.8 days. The large spike seen at the end is due to many small observations being grouped. The maximum length of stay recorded over the data period is 373 days!

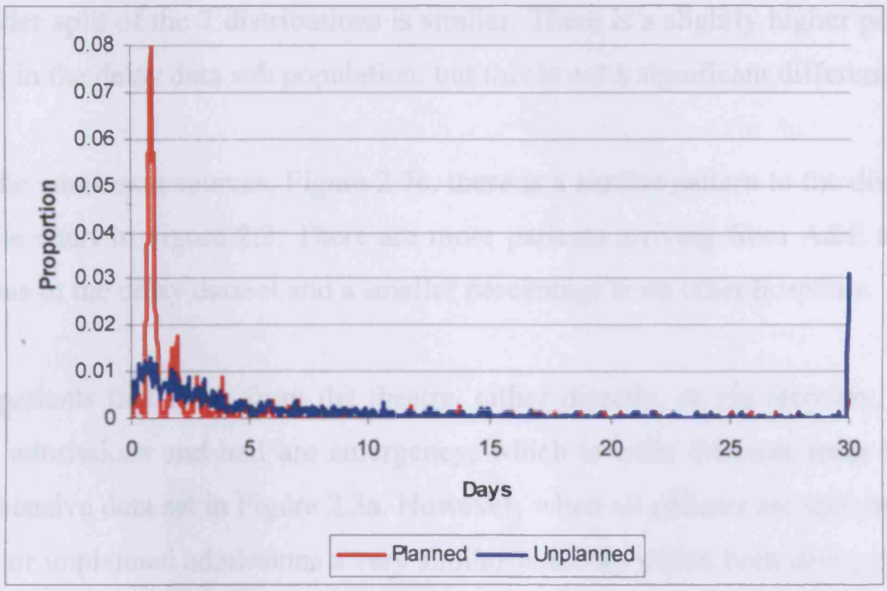


Figure 2.5 - Length of time spent in the CCU.

2.4 Delay Data

As previously discussed, there is only a limited amount of data for which the delay in leaving the department was recorded. So to start with, some of the same metrics will be computed to compare the 2 populations.

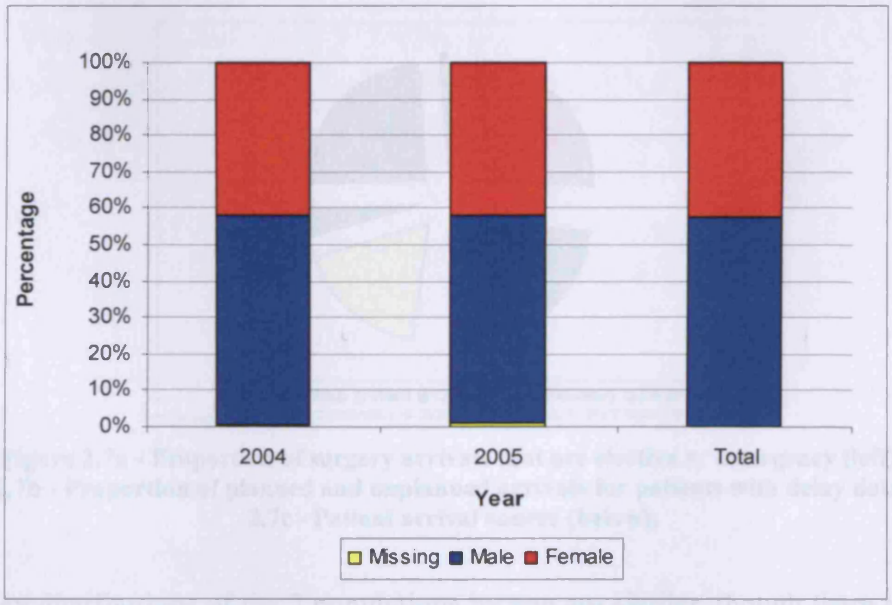


Figure 2.6 - Gender percentage for patient with delay data.

The gender split of the 2 distributions is similar. There is a slightly higher percentage of males in the delay data sub population, but this is not a significant difference.

As for the admission sources, Figure 2.7c, there is a similar pattern to the distribution in the pie chart in Figure 2.3. There are more patients arriving from A&E and from recoveries in the delay dataset and a smaller percentage from other hospitals.

Of the patients that come from the theatre, either directly, or via recovery, half are elective admissions and half are emergency, which is quite different from the more comprehensive data set in Figure 2.3a. However, when all patients are split into either planned or unplanned admissions a very similar make up within both data populations is seen.

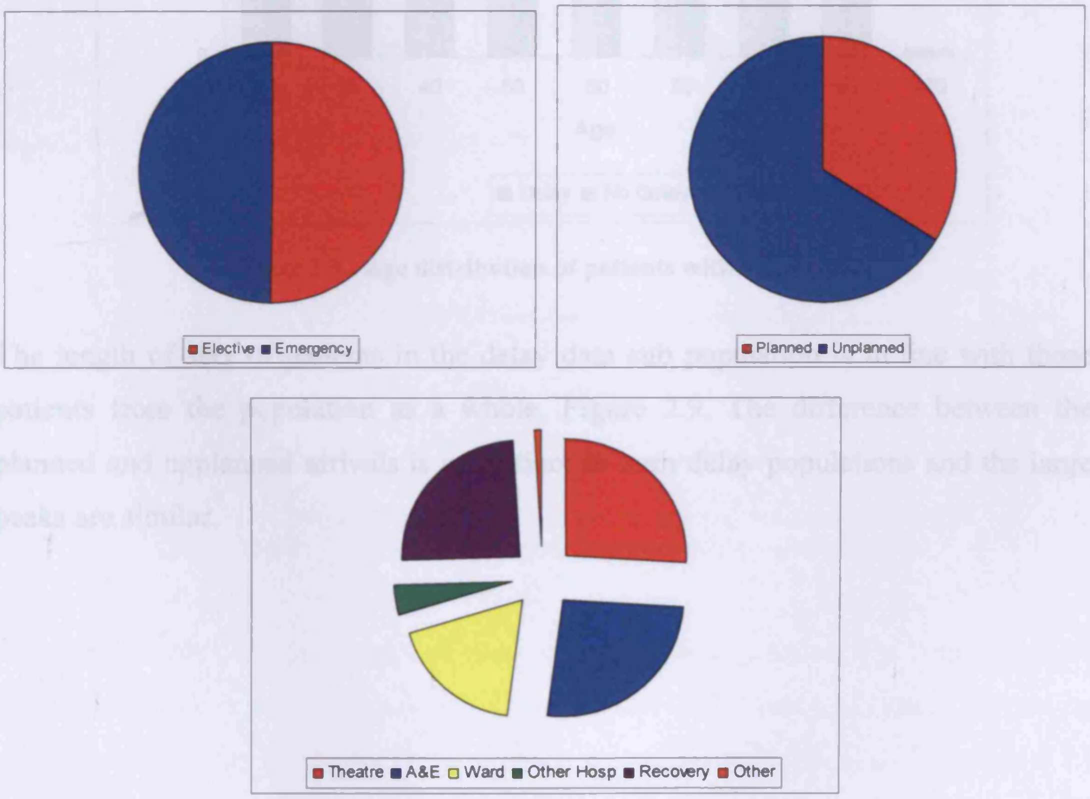


Figure 2.7a - Proportion of surgery arrivals that are elective or emergency (left), 2.7b - Proportion of planned and unplanned arrivals for patients with delay data, 2.7c - Patient arrival source (below),

The delay distributions of the 2 populations by age are similar, though there is a large spike in the whole population in the under 20 age group, Figure 2.8. This would suggest that the beds that delay data is taken from do not contain as many patients under the age of 20 as the rest of the CCU on average. It is also worth noting that none of the patients that were referred to leave the CCU, in this population, died whilst in the CCU. This would suggest that a large group of patient that use the CCU are not captured within this sub population, 20% of patients in the whole population did not survive their stay in the CCU.

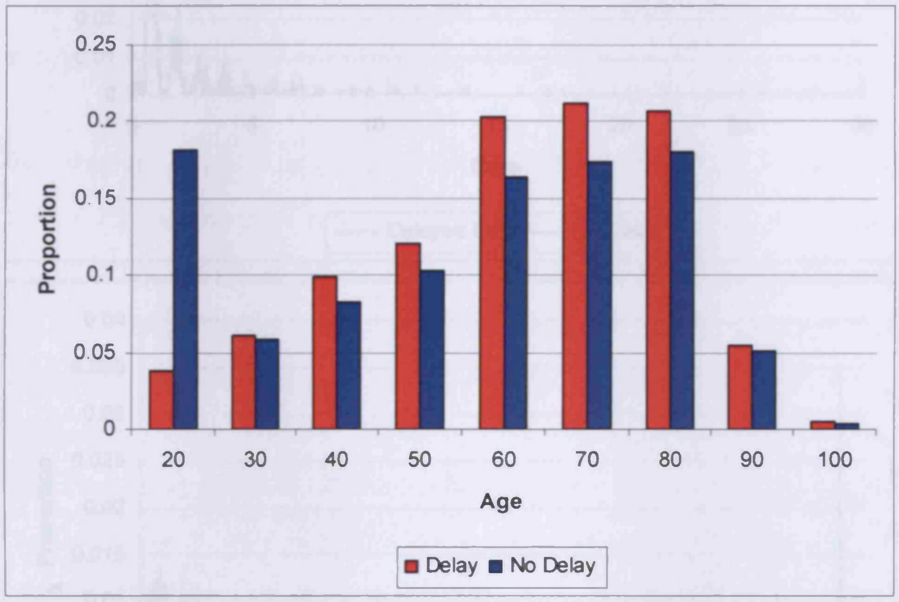


Figure 2.8 - Age distribution of patients with delay data.

The length of stay of patients in the delay data sub population is in line with those patients from the population as a whole, Figure 2.9. The difference between the planned and unplanned arrivals is as distinct in both delay populations and the large peaks are similar.

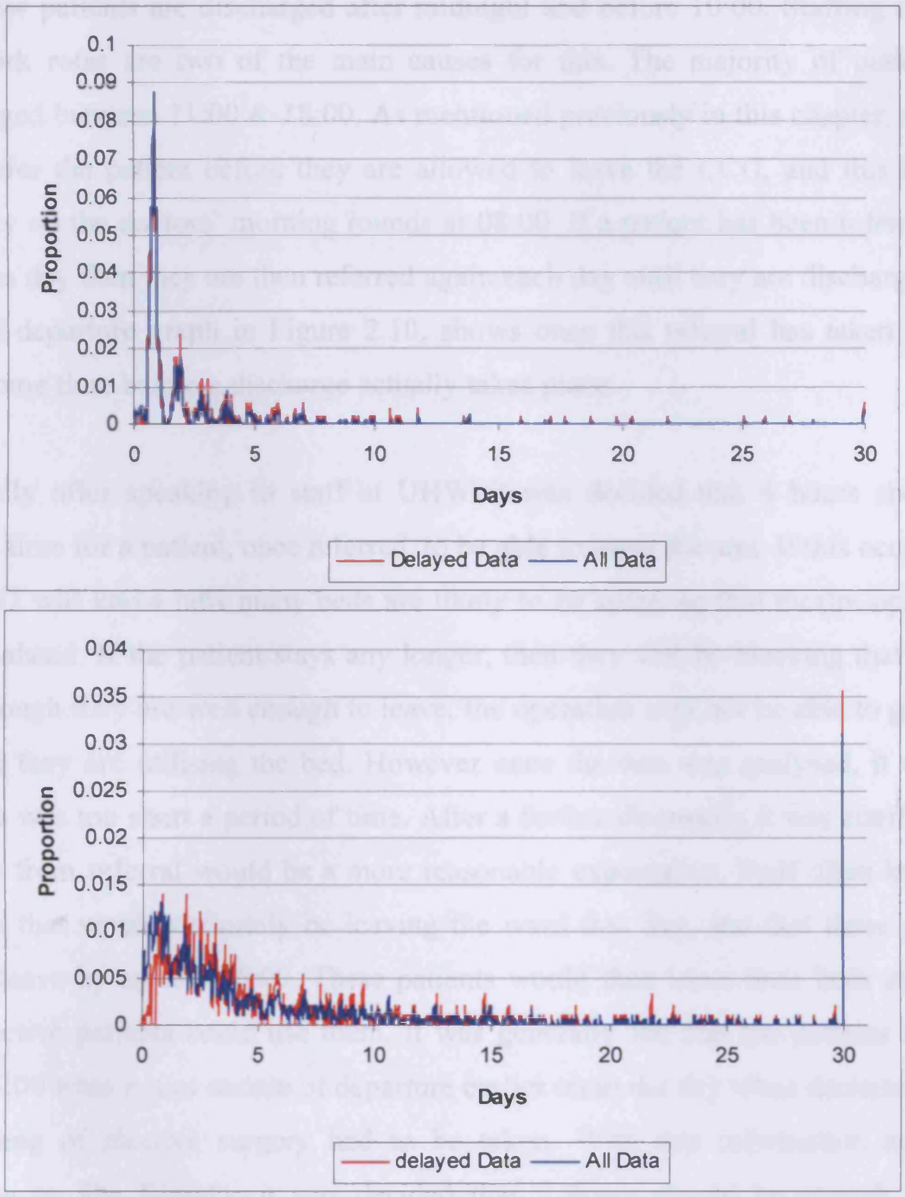


Figure 2.9a - Length of stay for planned patients, recorded each hour (top), 2.9b - Length of stay for unplanned patients, recorded hourly (below).

It is quite clear that the length of stay is cyclic in appearance for both types of patients. A peak is seen in every day then the probability that a customer leaves the unit decreases. This is due to the time dependent nature of arrivals and especially of departures, Figure 2.10. The arrival times have a more spread out distribution than that of departure times due to the fact that the hospital has no control over the arrivals of its unplanned patients. There is still a time dependent shape; more patients arrive during the early afternoon through to late evening as this the time when patients will be leaving theatre or recovery. There is also a smaller peak between 00:00 and 01:00.

Very few patients are discharged after midnight and before 10:00. Staffing numbers and work rotas are two of the main causes for this. The majority of patients are discharged between 11:00 & 18:00. As mentioned previously in this chapter, a doctor must refer the patient before they are allowed to leave the CCU, and this happens typically on the doctors' morning rounds at 08:00. If a patient has been referred on a previous day then they are then referred again each day until they are discharged. The time of departure graph in Figure 2.10, shows once this referral has taken place it takes some time before a discharge actually takes place.

Originally after speaking to staff at UHW it was decided that 4 hours should be enough time for a patient, once referred, to be able to leave the unit. If this occurs then the CCU will know how many beds are likely to be spare, so that theatre operations can go ahead. If the patient stays any longer, then they will be blocking that bed as, even though they are well enough to leave, the operation may not be able to go ahead because they are utilising the bed. However once the data was analysed, it was felt that this was too short a period of time. After a further discussion it was clarified that 7 hours from referral would be a more reasonable expectation. Staff often knew the patients that would definitely be leaving the ward that day, and that those patients would leave by around 15:00. These patients would then leave their beds empty so that elective patients could use them. It was generally felt that the patients that left after 15:00 were not as certain of departure earlier on in the day when decisions about scheduling of elective surgery had to be taken. With this information and after speaking to The Director it was decided that 7 hours should be enough time to discharge a patient. If the patient is occupying a bed in the CCU after this period of time then they are then identified as blocking that bed. The staff still believed that 4 hours should be long enough to discharge a referred patient, but understood that this rarely does happen.

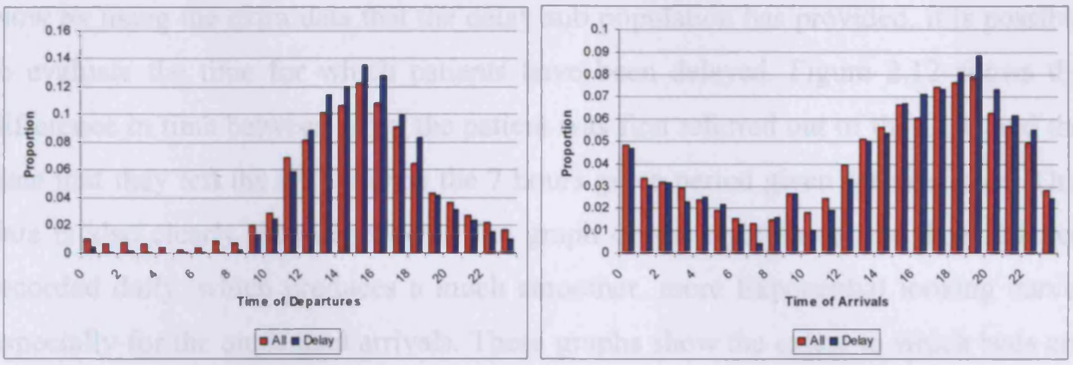


Figure 2.10 - Departure & Arrival times for the original and delay data sets.

As previously mentioned, the length of stay graphs, Figure 2.9 and Figure 2.5 are clearly time dependent. One of the methods we will use to model the CCU is steady state equations. To be able to do this the system that is being analysed clearly can not be time dependent. If a longer term approach is taken, the hour by hour state of the system is not as important as the day by day state. This will give a higher level, long term view of the CCU.

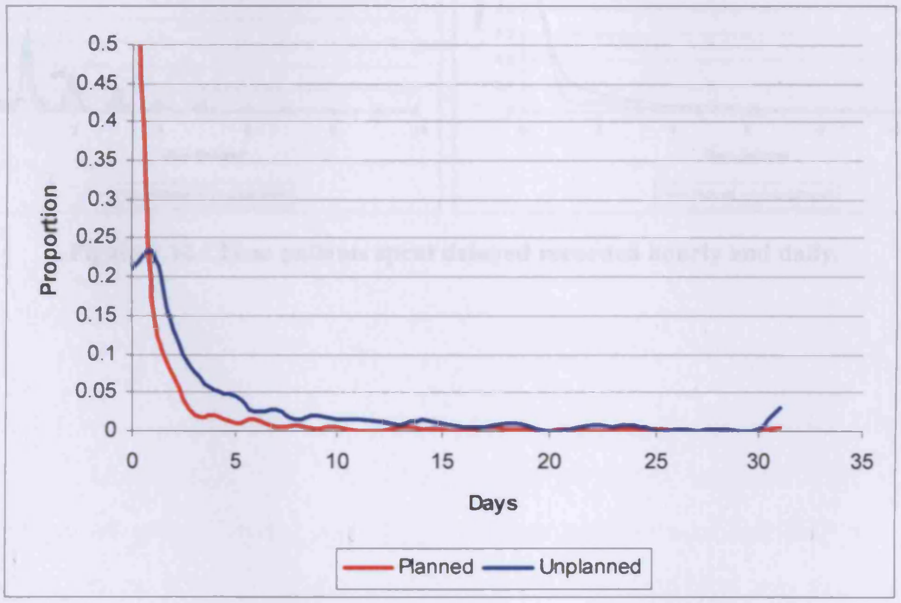


Figure 2.11 - Length of stay for patients with delay data.

Figure 2.11 shows the length of stay for patients in the delay data sub population, split by whether they were planned or not. This graph contains the same information as that in Figure 2.9, except that it is now summarised daily rather than hourly. It can be seen that this curve is much smoother and more appropriate to being modelled by a steady state equation.

Now by using the extra data that the delay sub population has provided, it is possible to evaluate the time for which patients have been delayed. Figure 2.12 shows the difference in time between when the patient was first referred out of the CCU and the time that they left the CCU minus the 7 hours grace period given for discharge. This data is also clearly time dependent. The graph on the right shows the days delayed recorded daily, which produces a much smoother, more Exponential looking curve, especially for the unplanned arrivals. These graphs show the extent to which beds are being blocked in the CCU. Some patients have spent 10 extra days in the CCU after their original date of referral. As well as this being an unpleasant experience for the patient and their family, the patient is using an expensive limited resource which could be better utilised. Within this thesis we shall attempt to model this blocking behaviour and show how changing it could alter the capacity of the CCU.

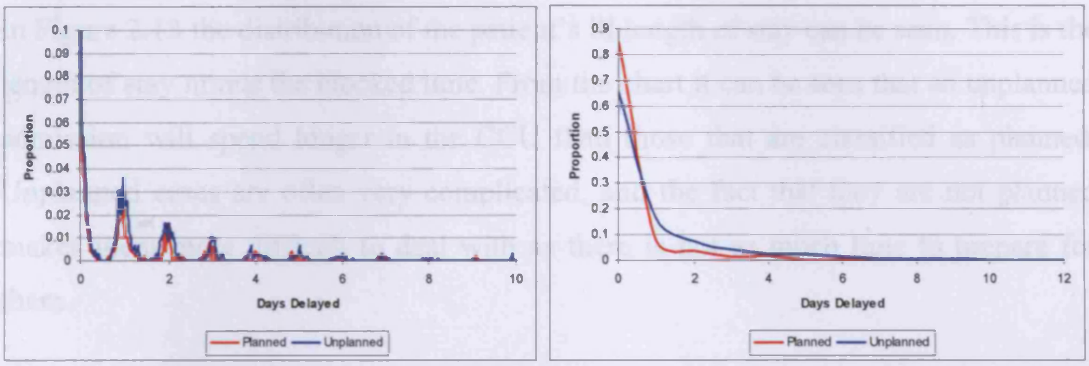


Figure 2.12 - Time patients spent delayed recorded hourly and daily.

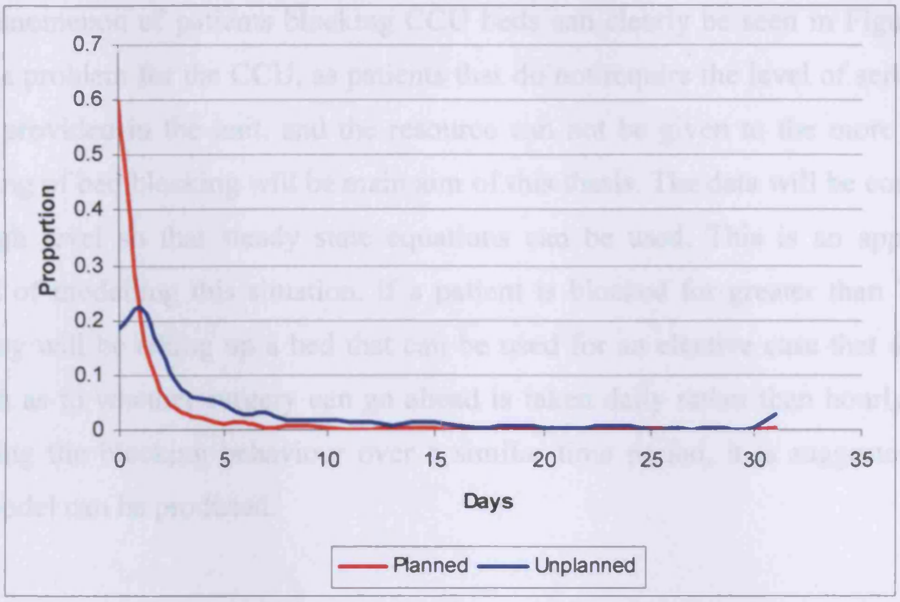


Figure 2.13 - Ill Length of stay

In Figure 2.13 the distribution of the patient's ill length of stay can be seen. This is the length of stay minus the blocked time. From the chart it can be seen that an unplanned admission will spend longer in the CCU than those that are classified as planned. Unplanned cases are often very complicated, and the fact that they are not planned makes them more difficult to deal with as there is not as much time to prepare for them.

2.5 Summary

This chapter has provided an overview of the data made available from the Critical Care Unit of UHW. The fact that not all of the data has been recorded for every patient provided some issues. Although the demographics of the data are not too different within the sub population for which the time of referral was recorded, there are some notable exceptions. The number of patients under the age of 20 is much higher compared to the whole population; this should be considered when modelling this department. Another factor that should be taken into account is that all of the patients in the sub population survived their time in the CCU. Even though the length of stay graphs look comparable in Figure 2.9, this is still a significant difference between the 2 populations.

The phenomenon of patients blocking CCU beds can clearly be seen in Figure 2.12. This is a problem for the CCU, as patients that do not require the level of service that can be provided in the unit, and the resource can not be given to the more ill. The modelling of bed blocking will be main aim of this thesis. The data will be considered at a high level so that steady state equations can be used. This is an appropriate method of modelling this situation. If a patient is blocked for greater than 7 hours, then they will be taking up a bed that can be used for an elective case that day. The decision as to whether surgery can go ahead is taken daily rather than hourly. So by modelling the blocking behaviour over a similar time period, it is suggested that a valid model can be produced.

The next three chapters will look at mathematical techniques that can be used to model a similar situation. Once the techniques have been established, we shall use this chapter's data to form models of the bed blocking situation to see what effect changing some variables may have on the unit.

Chapter 3

BLOCKING

3.1. Two Node Systems

3.1.1 Introduction

To attempt to model the bed blocking situation in the Critical Care Unit of the University Hospital of Wales, a number of theoretical and simulation techniques will be used. In this chapter an introduction to the theory of queues in series will be given and a detailed study of how they behave will be shown.

Queues are said to be in series when a service point's output accounts for, in all or in part, the input to another service point. We may visualise a system of k service points, with a possible queueing space in front of each service point. When two or more service points are placed in series they are often referred to as being in tandem, or as tandem, queues as well as being in series. Figure 3.1 illustrates a tandem queue with random arrivals at the first service point, at mean rate λ. The output from the first service point is the only input into the second service point. In this case both service points have a random or Poisson service distribution with parameter μ. For notational purposes a service point will from here on be referred to as a node.

In a series of queues it is possible for a node to become blocked. The term blocked in this sense means a node is unable to carry on processing customers because of an issue at a subsequent node. This could happen, for example, when Node x completes a service but the queue for Node $x+1$ has reached capacity and there is nowhere to place the processed customer from x. Another possibility might be that there is no queue at Node $x+1$ and Node $x+1$ is still serving and unable to take the processed customer from Node x. In either case the current node is unable to receive a new customer even though it has completed its work and is therefore blocked.

As mentioned in the introduction chapter there are three well recognised methods to deal with blocking within networks of queues; Transfer, Communication and Rejection Blocking techniques. For the remainder of the work in this thesis only Transfer Blocking will be considered. In this case, when a customer at Node x has completed service but Node $x+1$, its destination, is unable to take the customer, Node x becomes blocked and is unable to process any more customers until Node $x+1$ has completed a service and made way for Node $x's$ served customer.

3.1.2 Zero Queue Size, Equal Service Rates

The simplest case will be considered first. The initial system will consist of two nodes with no queue allowed between the nodes, and no queue allowed in front of the first node. Customers who arrive whilst the first node is servicing a customer or whilst the first node is blocked will leave the system instantly. Customers arrive with a Negative Exponential inter-arrival distribution with parameter, λ and are served at both nodes by a Negative Exponential distribution with mean rate μ. Figure 3.1 is a pictorial representation of this system.

μ
μ
λ
Node 1
Node 2

Figure 3.1 - A tandem queue with the same service rates

To be able to describe this model some mathematical notation is required;

$P_{0,0}(t)$ Probability of system being empty at time t,

$P_{1,0}(t)$ Probability a customer is in service at Node 1 and that Node 2 is empty at time t,

$P_{0,1}(t)$ Probability a customer is in service at Node 2 and that Node 1 is empty at time t,

$P_{1,1}(t)$ Probability Nodes 1 and 2 are both processing customers at time t,

$P_{1b,1}(t)$ Probability Node 1 is blocked and Node 2 is serving a customer at time t.

Initial time dependent equations are setup. These equations relate what could happen to the system at time $t+\delta t$ considering the state the system is in at time t;

$$\begin{aligned}
P_{0,0}(t+\delta t) &= (1-\lambda\delta t)P_{0,0}(t)+\mu\delta t(1-\lambda\delta t)P_{0,1}(t)\\
P_{1,0}(t+\delta t) &= \lambda\delta t P_{0,0}(t)+(1-\mu\delta t)P_{1,0}(t)+\mu\delta t(1-\mu\delta t)P_{1,1}(t)\\
P_{0,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu\delta t)P_{0,1}(t)+\mu\delta t P_{1,0}(t)+\mu\delta t P_{1b,1}(t)\\
P_{1,1}(t+\delta t) &= (1-\mu\delta t)(1-\mu\delta t)P_{1,1}(t)+\lambda\delta t(1-\mu\delta t)P_{0,1}(t)\\
P_{1b,1}(t+\delta t) &= (1-\mu\delta t)P_{1b,1}(t)+(1-\mu\delta t)\mu\delta t P_{1,1}(t)
\end{aligned} \tag{3.1}$$

These can be rearranged to give;

$$\begin{aligned}
\frac{dP_{0,0}(t)}{dt} &= -\lambda P_{0,0}(t) + \mu P_{0,1}(t) \\
\frac{dP_{1,0}(t)}{dt} &= \lambda P_{0,0}(t) - \mu P_{1,0}(t) + \mu P_{1,1}(t) \\
\frac{dP_{0,1}(t)}{dt} &= -(\lambda + \mu) P_{0,1}(t) + \mu P_{1,0}(t) + \mu P_{1b,1}(t) \\
\frac{dP_{1,1}(t)}{dt} &= -2\mu P_{1,1}(t) + \lambda P_{0,1}(t) \\
\frac{dP_{1b,1}(t)}{dt} &= -\mu P_{1b,1}(t) + \mu P_{1,1}(t)
\end{aligned} \tag{3.2}$$

Next, by setting the differential equal to zero the steady state equations can be found. This steady state system is so called because the rate at which each probability is changing over time (the differential) is equal to zero and hence the system is not changing and steady. Equations (3.3) are the steady state equation for this system;

$$\begin{aligned}
\lambda P_{0,0} &= \mu P_{0,1} \\
\mu P_{1,0} &= \lambda P_{0,0} + \mu P_{1,1} \\
(\lambda + \mu) P_{0,1} &= \mu P_{1,0} + \mu P_{1b,1} \\
2\mu P_{1,1} &= \lambda P_{0,1} \\
P_{1b,1} &= P_{1,1}
\end{aligned} \tag{3.3}$$

There is also the additional fact that all of these probabilities must sum to 1, providing another equation;

$$\sum_i \sum_j P_{i,j} = 1 \tag{3.4}$$

For this system there are five unknowns and six equations, but any one equation is redundant as it is a linear combination of the other five. The equation involving $P_{0,1}$ on the left hand side is rejected;

$$\begin{aligned}
P_{0,1} &= \frac{\lambda}{\mu} P_{0,0} \\
P_{1,1} &= \frac{\lambda}{2\mu} P_{0,1} \\
&= \frac{\lambda^2}{2\mu^2} P_{0,0} \\
P_{1b,1} &= P_{0,1} \\
&= \frac{\lambda^2}{2\mu^2} P_{0,0} \\
P_{1,0} &= \frac{\lambda}{\mu} P_{0,0} + P_{1,1} \\
&= \frac{\lambda}{\mu}\left(1 + \frac{\lambda}{2\mu}\right) P_{0,0}
\end{aligned} \tag{3.5}$$

$$\sum_i \sum_j P_{i,j} = 1$$

Substituting the four probabilities in (3.5) into the final equation;

$$\begin{aligned}
P_{0,0} + P_{1,0} + P_{0,1} + P_{1,1} + P_{1b,1} &= 1 \\
P_{0,0}\left(1 + 2\frac{\lambda}{\mu} + 3\frac{\lambda^2}{2\mu^2}\right) &= 1 \\
P_{0,0} &= \frac{2\mu^2}{3\lambda^2 + 4\lambda\mu + 2\mu^2}
\end{aligned} \tag{3.6}$$

Hence;

$$
\begin{aligned}
P_{0,1} &= \frac{2\lambda\mu}{3\lambda^2+4\lambda\mu+2\mu^2} \\
P_{1,1} &= \frac{\lambda^2}{3\lambda^2+4\lambda\mu+2\mu^2} \\
P_{1b,1} &= \frac{\lambda^2}{3\lambda^2+4\lambda\mu+2\mu^2} \\
P_{1,0} &= \frac{\lambda(\lambda+2\mu)}{3\lambda^2+4\lambda\mu+2\mu^2}
\end{aligned}
\qquad (3.7)
$$

3.1.3 Zero Queue Size, Different Service Rates

The next system to be set up is similar to the previous, but this time the service distribution parameters are different, denoted by μ_1 and μ_2 for Nodes 1 and 2 respectively. This system can be seen in Figure 3.2.

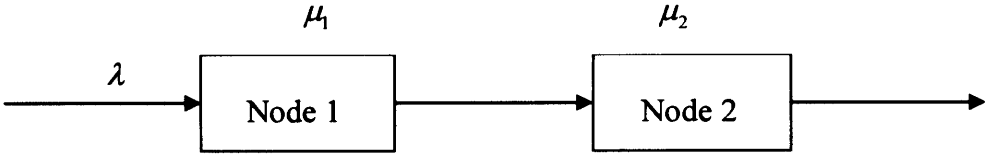


Figure 3.2 - A two node tandem queue with different service rates

The time dependent equations for this system are set up and solved in a similar way to before;

$$
\begin{aligned}
&P_{0,0}(t+\delta t)=(1-\lambda\delta t)P_{0,0}(t)+(1-\lambda\delta t)\mu_2\delta tP_{0,1}(t) \\
&P_{1,0}(t+\delta t)=\lambda\delta tP_{0,0}(t)+(1-\mu_1\delta t)P_{1,0}(t)+(1-\mu_1\delta t)\mu_2\delta tP_{1,1}(t) \\
&P_{0,1}(t+\delta t)=(1-\lambda\delta t)(1-\mu_2\delta t)P_{0,1}(t)+\mu_1\delta tP_{1,0}(t)+\mu_2\delta tP_{1b,1}(t) \\
&P_{1,1}(t+\delta t)=(1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1}(t)+\lambda\delta t(1-\mu_2\delta t)P_{0,1}(t) \\
&P_{1b,1}(t+\delta t)=(1-\mu_2\delta t)P_{1b,1}(t)+\mu_1\delta t(1-\mu_2\delta t)P_{1,1}(t)
\end{aligned}
\qquad (3.8)
$$

The probabilities are found in terms of $P_{0,0}$ in the same way;

$$
\begin{aligned}
P_{0,1} &= \frac{\lambda}{\mu_2} P_{0,0} \\
P_{1,1} &= \frac{\lambda^2}{\mu_2\left(\mu_1+\mu_2\right)} P_{0,0} \\
P_{1b,1} &= \frac{\lambda^2\mu_1}{\mu_2\left(\mu_1+\mu_2\right)} P_{0,0} \\
P_{1,0} &= \frac{\lambda}{\mu_1}\left(1+\frac{\lambda}{\left(\mu_1+\mu_2\right)}\right) P_{0,0}
\end{aligned}
\tag{3.9}
$$

There is also the additional equation;

$$\sum_i \sum_j P_{i,j} = 1 \tag{3.10}$$

These can be solved to give;

$$
\begin{aligned}
P_{0,0} &= \frac{\mu_1\mu_2^2\left(\mu_1+\mu_2\right)}{\mu_2^3\left(\lambda+\mu_1\right)+\mu_2^2\left(\lambda+\mu_1\right)^2+\mu_2\left(\lambda+\mu_1\right)\left(\lambda\mu_1\right)+\lambda^2\mu_1^2} \\
P_{1,1} &= \frac{\lambda\mu_1\mu_2\left(\mu_1+\mu_2\right)}{\mu_2^3\left(\lambda+\mu_1\right)+\mu_2^2\left(\lambda+\mu_1\right)^2+\mu_2\left(\lambda+\mu_1\right)\left(\lambda\mu_1\right)+\lambda^2\mu_1^2} \\
P_{1,1} &= \frac{\lambda^2\mu_1\mu_2}{\mu_2^3\left(\lambda+\mu_1\right)+\mu_2^2\left(\lambda+\mu_1\right)^2+\mu_2\left(\lambda+\mu_1\right)\left(\lambda\mu_1\right)+\lambda^2\mu_1^2} \\
P_{1b,1} &= \frac{\lambda^2\mu_1^2}{\mu_2^3\left(\lambda+\mu_1\right)+\mu_2^2\left(\lambda+\mu_1\right)^2+\mu_2\left(\lambda+\mu_1\right)\left(\lambda\mu_1\right)+\lambda^2\mu_1^2} \\
P_{1,0} &= \frac{\lambda\left(\lambda+\mu_1+\mu_2\right)\mu_2^2}{\mu_2^3\left(\lambda+\mu_1\right)+\mu_2^2\left(\lambda+\mu_1\right)^2+\mu_2\left(\lambda+\mu_1\right)\left(\lambda\mu_1\right)+\lambda^2\mu_1^2}
\end{aligned}
\tag{3.11}
$$

If $\mu_1 = \mu_2$, then the result can be seen to be the same as the previous case.

The transient solution of this system has been coded in a Visual Basic program as a validation technique. The transient results when μ_1 and μ_2 are set equal to 1, λ is equal to 0.5 and δt equals 0.01, has been plotted and can be seen in Figure 3.3.

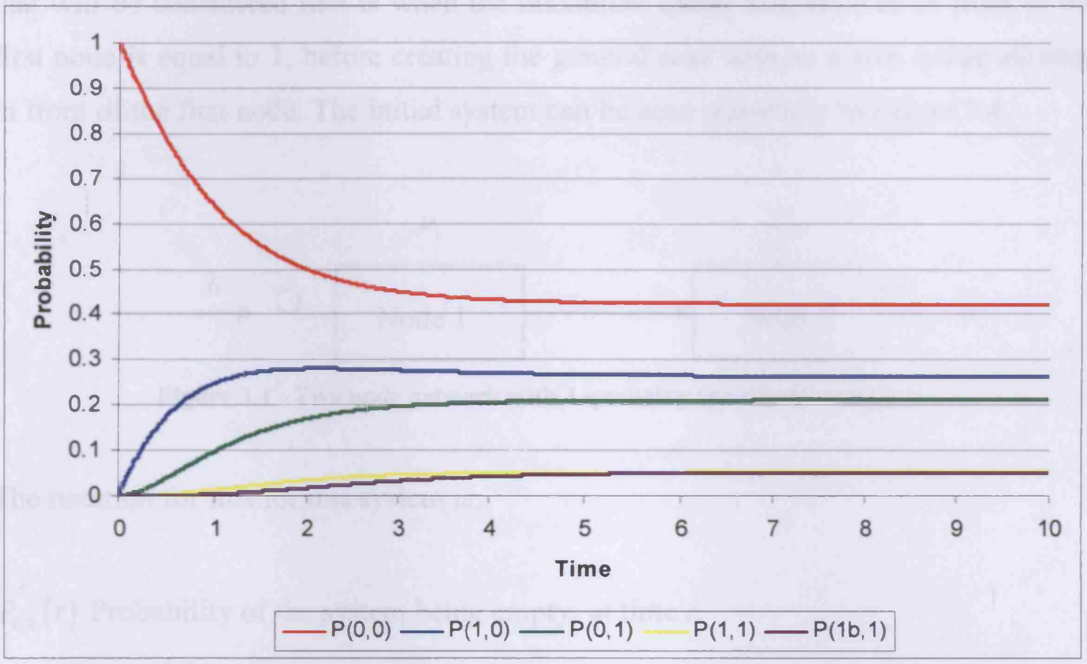


Figure 3.3 - Graph showing transient solutions for a two node system with λ= 0.5 and $\mu_1 = \mu_2 = 1$.

Figure 3.3 shows that the transient solution tends to fixed value for these service and arrival rates, and by substituting these values in to the steady state equation for this system they can be seen to be the same.

3.1.4 Queue Size One

The next step is to find the equations and solutions for the case when a queue is allowed in front of the first node, but no queue allowed between nodes. The system that will be considered first is when the maximum queue size allowed in front of the first node is equal to 1, before creating the general case with an n size queue allowed in front of the first node. The initial system can be seen pictorially in Figure 3.4.

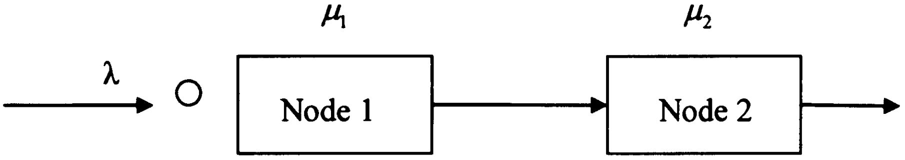


Figure 3.4 - Two node network with 1 queueing space before Node 1.

The notation for this for this system is;

$P_{0,0}(t)$ Probability of the system being empty, at time t.

$P_{1,0}(t)$ Probability that a customer is being served in Node 1 and Node 2 is empty and there is no queue in front of Node 1, at time t.

$P_{2,0}(t)$ Probability that a customer is being served in Node 1, Node 2 is empty, and there is a (full) queue of one in front of Node 1, at time t.

$P_{0,1}(t)$ Probability that Node 1 is empty, Node 2 is currently serving a customer and there is no queue in front of Node 1, at time t.

$P_{1,1}(t)$ Probability that there is a customer at both nodes and no queue, at time t.

$P_{2,1}(t)$ Probability that there is a customer at both nodes and a (full) queue of one customer in front of Node 1, at time t.

$P_{1b,1}(t)$ Probability Node 2 is currently serving a customer, Node 1 has completed its service and is blocked and there is no queue in front of Node 1, at time t.

$P_{2b,1}(t)$ Probability Node 2 is currently serving a customer, Node 1 has completed its service and is blocked and there is a (full) queue of one customer in front of Node 1, at time t.

The time dependent equations of the system with space for one to queue are;

$$\begin{aligned}
P_{0,0}(t+\delta t) &= (1-\lambda\delta t)P_{0,0}(t)+(1-\lambda\delta t)\mu_2\delta tP_{0,1}(t)\\
P_{1,0}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)P_{1,0}(t)+\lambda\delta tP_{0,0}(t)\\
&\quad+(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta tP_{1,1}(t)\\
P_{2,0}(t+\delta t) &= (1-\mu_1\delta t)P_{2,0}(t)+\lambda\delta t(1-\mu_1\delta t)P_{1,0}(t)\\
&\quad+(1-\mu_1\delta t)\mu_2\delta tP_{2,1}(t)\\
P_{0,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{0,1}(t)+(1-\lambda\delta t)\mu_1\delta tP_{1,0}(t)\\
&\quad+(1-\lambda\delta t)\mu_2\delta tP_{1b,1}(t)\\
P_{1,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1}(t)\\
&\quad+\lambda\delta t(1-\mu_2\delta t)P_{0,1}(t)+\mu_1\delta tP_{2,0}(t)+\mu_2\delta tP_{2b,1}(t)\\
P_{2,1}(t+\delta t) &= (1-\mu_1\delta t)(1-\mu_2\delta t)P_{2,1}(t)\\
&\quad+\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1}(t)\\
P_{1b,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{1b,1}(t)\\
&\quad+(1-\lambda\delta t)\mu_1\delta t(1-\mu_2\delta t)P_{1,1}(t)\\
P_{2b,1}(t+\delta t) &= (1-\mu_2\delta t)P_{2b,1}(t)+\mu_1\delta t(1-\mu_2\delta t)P_{2,1}(t)\\
&\quad+\lambda\delta t(1-\mu_2\delta t)P_{1b,1}(t)
\end{aligned}\tag{3.12}$$

These equations can be manipulated to give the following steady state equations;

$$P_{0,1} = \frac{\lambda}{\mu_2} P_{0,0}$$

$$P_{1,0} = \frac{\lambda\left(\left(\lambda+\mu_2\right)^2 + \mu_1\mu_2\right)}{\mu_1\mu_2\left(2\lambda+\mu_1+\mu_2\right)} P_{0,0}$$

$$P_{1b,1} = \frac{\lambda^2\left(\lambda+\mu_1+\mu_2\right)}{\mu_2^2\left(2\lambda+\mu_1+\mu_2\right)} P_{0,0}$$

$$P_{1,1} = \frac{\left(\lambda+\mu_2\right)\lambda^2\left(\lambda+\mu_1+\mu_2\right)}{\mu_1\mu_2^2\left(2\lambda+\mu_1+\mu_2\right)} P_{0,0}$$

$$P_{2,1} = \frac{\lambda^3\left(\lambda+\mu_2\right)\left(\lambda+\mu_1+\mu_2\right)}{\mu_1\mu_2^2\left(\mu_1+\mu_2\right)\left(2\lambda+\mu_1+\mu_2\right)} P_{0,0}$$

$$P_{2b,1} = \frac{\lambda^3\left(\lambda+\mu_1+\mu_2\right)\left(\lambda+\mu_1+2\mu_2\right)}{\mu_2^3\left(2\lambda+\mu_1+\mu_2\right)\left(\left(\mu_1+\mu_2\right)\right)} P_{0,0}$$

$$P_{2,0} = \frac{\lambda^2\left(\lambda^3+\lambda^2\left(2\mu_1+3\mu_2\right)+\lambda\left(3\mu_1\mu_2+3\mu_2^2\right)+\mu_1^2\mu_2+2\mu_1\mu_2^2+\mu_2^3\right)}{\mu_1^2\mu_2\left(\mu_1+\mu_2\right)\left(2\lambda+\mu_1+\mu_2\right)} P_{0,0} \tag{3.13}$$

From the fact that all the probabilities sum to 1 $P_{0,0}$ can be found;

$$\begin{aligned}
P_{0,0} &= \frac{N}{D} \\
N &= \mu_1^2\mu_2^3\left(\mu_1+\mu_2\right)\left(2\lambda+\mu_1+\mu_2\right) \\
D &= \lambda^5\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right) \\
&\quad +\lambda^4\left(2\mu_1+3\mu_2\right)\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right) \\
&\quad +\lambda^3\left(\mu_1+3\mu_2\right)\left(\mu_1+\mu_2\right)\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right) \\
&\quad +\lambda^2\left(\mu_1^2+3\mu_1\mu_2+\mu_2^2\right)\left(\mu_1+\mu_2\right)^2 \\
&\quad +\lambda\mu_1\mu_2^2\left(\mu_1+\mu_2\right)\left(\mu_1^2+4\mu_1\mu_2+\mu_2^2\right) \\
&\quad +\mu_1^2\mu_2^3\left(\mu_1+\mu_2\right)^2
\end{aligned} \tag{3.14}$$

All of the probabilities can now be found in terms of λ, μ_1 and μ_2. The transient solutions of this system, when both μ's are equal to 1, and λ equal to 0.5 have also been found, the graph of which can be seen in Figure 3.5. As before, and as expected, the values that the transient probabilities tend to are the same as those derived from the equations with the appropriate values substituted in.

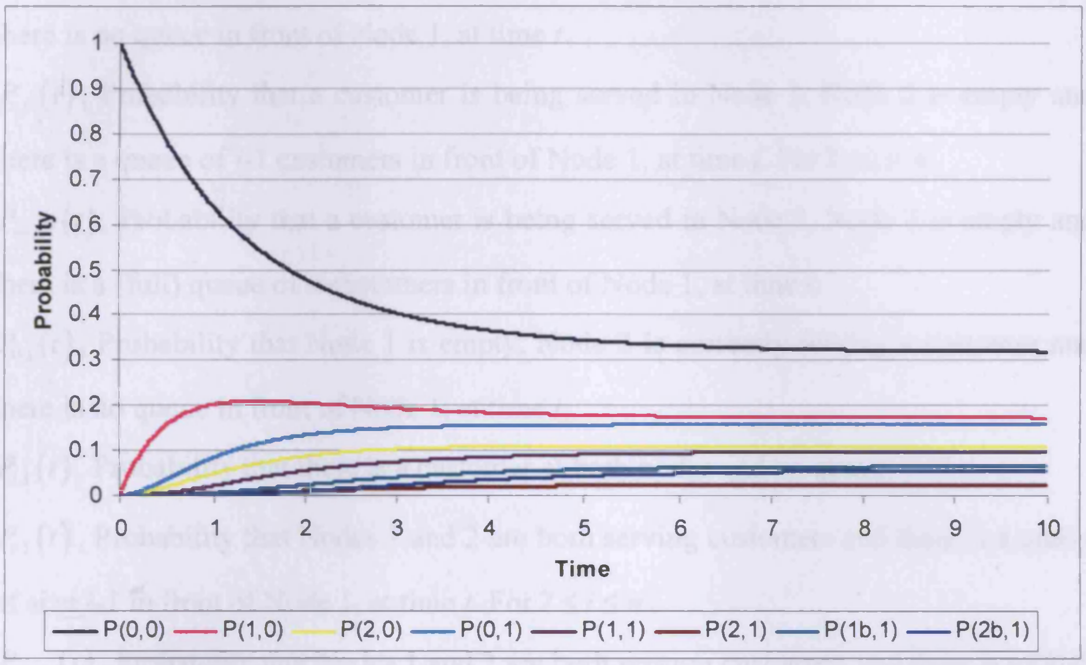


Figure 3.5 - Graph showing transient solutions for a two node system with λ= 0.5 and $\mu_1 = \mu_2 = 1$.

Again we note that the transient solution tends to the steady state values, provided by equations (3.13), as time increases.

3.1.5 'n' Size Queue

The next case to be studied is the two node system with both nodes having different service rates, as above. However, in this case a maximum queue length of size n is imposed in front of the initial node. This can be seen in Figure 3.6.

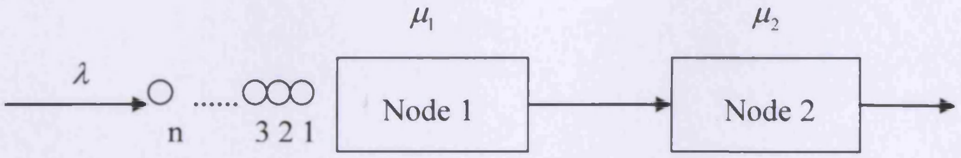


Figure 3.6 - Two node network with n spaces to queue in front of Node 1.

The notation for this system is similar to that of the previous section, with extra probabilities to allow for the increase in queue size;

$P_{0,0}(t)$, Probability system is empty, at time *t*.

$P_{1,0}(t)$, Probability that a customer is being served in Node 1, Node 2 is empty and there is no queue in front of Node 1, at time *t*.

$P_{i,0}(t)$, Probability that a customer is being served in Node 1, Node 2 is empty and there is a queue of *i*-1 customers in front of Node 1, at time *t*. For $2 \le i \le n$.

$P_{n+1,0}(t)$, Probability that a customer is being served in Node 1, Node 2 is empty and there is a (full) queue of *n* customers in front of Node 1, at time *t*.

$P_{0,1}(t)$, Probability that Node 1 is empty, Node 2 is currently serving a customer and there is no queue in front of Node 1, at time *t*.

$P_{1,1}(t)$, Probability that there is a customer at both nodes and no queue, at time *t*.

$P_{i,1}(t)$, Probability that Nodes 1 and 2 are both serving customers and there is a queue of size *i-1* in front of Node 1, at time *t*. For $2 \le i \le n$.

$P_{n+1,1}(t)$, Probability that Nodes 1 and 2 are both serving customers and there is a (full) queue of size *n* in front of Node 1, at time *t*.

$P_{1b,1}(t)$, Probability Node 2 is currently serving a customer; Node 1 has completed its service and is blocked, with no queue in front of Node 1, at time *t*.

$P_{ib,1}(t)$, Probability Node 2 is currently serving a customer, Node 1 has completed its service and is blocked, with a queue size *i*-1 in front of the initial node, at time *t*. $2 \le i \le n$.

$P_{(n+1)b,1}(t)$, Probability Node 2 is currently serving a customer, Node 1 has completed its service and is blocked, with a (full) queue size *n* in front of the initial node, at time *t*.

Using the same method as with the previous systems the following time dependent equations are created;

$$
\begin{aligned}
P_{0,0}(t+\delta t) &= (1-\lambda\delta t)P_{0,0}(t)+(1-\lambda\delta t)\mu_2\delta t P_{0,1}(t) \\
P_{1,0}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)P_{1,0}(t)+\lambda\delta t P_{0,0}(t) \\
&\quad +(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta t P_{1,1}(t) \\
P_{i,0}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)P_{i,0}(t)+\lambda\delta t(1-\mu_1\delta t)P_{i-1,0}(t) \\
&\quad +(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta t P_{i,1}(t) \\
P_{n+1,0}(t+\delta t) &= (1-\mu_1\delta t)P_{n+1,0}(t)+\lambda\delta t(1-\mu_1\delta t)P_{n,0}(t) \\
&\quad +(1-\mu_1\delta t)\mu_2\delta t P_{n+1,1}(t) \\
P_{0,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{0,1}(t)+(1-\lambda\delta t)\mu_1\delta t P_{1,0}(t) \\
&\quad +(1-\lambda\delta t)\mu_2\delta t P_{1b,1}(t) \\
P_{1,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1}(t) \\
&\quad +\lambda\delta t(1-\mu_2\delta t)P_{0,1}(t) \\
&\quad +(1-\lambda\delta t)\mu_1\delta t P_{2,0}(t)+(1-\lambda\delta t)\mu_2\delta t P_{2b,1}(t) \\
P_{i,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{i,1}(t) \qquad\qquad i=2,3,\ldots,n-1 \\
&\quad +\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{i-1,1}(t) \\
&\quad +(1-\lambda\delta t)\mu_1\delta t P_{i+1,0}(t)+(1-\lambda\delta t)\mu_2\delta t P_{(i+1)b,1}(t) \\
P_{n,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{n,1}(t) \\
&\quad +\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{n-1,1}(t) \\
&\quad +\mu_1\delta t P_{n+1,0}(t)+\mu_2\delta t P_{(n+1)b,1}(t)
\end{aligned}
\tag{3.15}
$$

$$
\begin{aligned}
P_{n+1,1}(t+\delta t) &= (1-\mu_1\delta t)(1-\mu_2\delta t)P_{n+1,1}(t) \\
&\quad +\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{n,1}(t) \\
P_{1b,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{1b,1}(t) \\
&\quad +(1-\lambda\delta t)\mu_1\delta t(1-\mu_2\delta t)P_{1,1}(t) \\
P_{ib,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{ib,1}(t) \qquad\qquad i=2,3,\ldots,n \\
&\quad +\lambda\delta t(1-\mu_2\delta t)P_{(i-1)b,1}(t) \\
&\quad +(1-\lambda\delta t)\mu_1\delta t(1-\mu_2\delta t)P_{i,1}(t) \\
P_{(n+1)b,1}(t+\delta t) &= (1-\mu_2\delta t)P_{(n+1)b,1}(t)+\mu_1\delta t P_{n+1,1}(t) \\
&\quad +\lambda\delta t(1-\mu_2\delta t)P_{nb,1}(t)
\end{aligned}
$$

In this case it is difficult to find all the probabilities in terms of $P_{0,0}(t)$ without first knowing the size of *n*, the capacity of the queue, as there are an unknown number of equations to solve. Instead the probabilities have been found in terms of previously calculated probabilities or 'lower order` probabilities. This makes computing the probabilities easier when *n* is defined. Probabilities can be found in order with the higher queue probabilities being computed from lesser ones until the maximum capacity has been reached.

$$
\begin{aligned}
P_{1,0} &= \frac{\lambda\left((\lambda+\mu_2)^2+\mu_1\mu_2\right)}{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)}P_{0,0} \\
P_{i,0} &= \frac{\left((\lambda+\mu_2)^2+\mu_1(\lambda+\mu_2)\right)P_{i-1,1}-\lambda\mu_2P_{(i-1)b,1}+\lambda\mu_1P_{(i-1),0}-\lambda(\lambda+\mu_2)P_{i-2,1}}{\mu_1(2\lambda+\mu_1+\mu_2)} \qquad i=2,3,\ldots n \\
P_{n+1,0} &= \frac{\lambda\left(\mu_2P_{n,1}+(\mu_1+\mu_2)P_{n,0}\right)}{\mu_1(\mu_1+\mu_2)} \\
P_{0,1} &= \frac{\lambda}{\mu_2}P_{0,0}
\end{aligned}
\tag{3.16}
$$

$$P_{1,1}=\frac{\lambda^2(\lambda+\mu_2)(\lambda+\mu_1+\mu_2)}{\mu_1\mu_2^2(2\lambda+\mu_1+\mu_2)}P_{0,0}$$

$$P_{i,1}=\frac{\begin{array}{r}(\lambda+\mu_1)(\lambda+\mu_2)\left((\lambda+\mu_1+\mu_2)P_{i-1,1}-\lambda(\lambda+\mu_1)(\lambda+\mu_2)P_{i-2,1}\right)\\ -\lambda\left(\mu_2(\lambda+\mu_1)P_{(i-1)b,1}+\mu_1(\lambda+\mu_2)P_{i-1,0}\right)\end{array}}{\mu_1\mu_2(2\lambda+\mu_1+\mu_2)}\qquad i=2,3,\ldots,n$$

$$P_{n+1,1}=\frac{\lambda}{\mu_1+\mu_2}P_{n,1}$$

$$P_{1b,1}=\frac{\lambda^2(\lambda+\mu_1+\mu_2)}{\mu_2^2(2\lambda+\mu_1+\mu_2)}P_{0,0}$$

$$P_{ib,1}=\frac{(\lambda+\mu_1)(\lambda+\mu_1+\mu_2)P_{i-1,1}-\lambda(\lambda+\mu_1)P_{i-2,1}+\lambda\mu_2P_{(i-1)b,1}-\lambda\mu_1P_{i-1,0}}{\mu_2(2\lambda+\mu_1+\mu_2)}\qquad i=2,3,\ldots,n$$

$$P_{(n+1)b,1}=\frac{\lambda\left(\mu_1P_{n,1}+(\mu_1+\mu_2)P_{nb,1}\right)}{\mu_2(\mu_1+\mu_2)}$$

It is not possible to find $P_{0,0}$ in terms of λ, μ_1 and μ_2 without first knowing the value of n. As this is a general result the solution will be left in this form.

3.1.6 Infinite Size Queue

Now that equations have been created for an '*n*' size queue all that is left to further this type of queueing system with no queueing space between the nodes is to add an infinite queueing space in front of the first node. This will then provide a comprehensive overview of the different possibilities in this two node system. This can be seen in Figure 3.7.

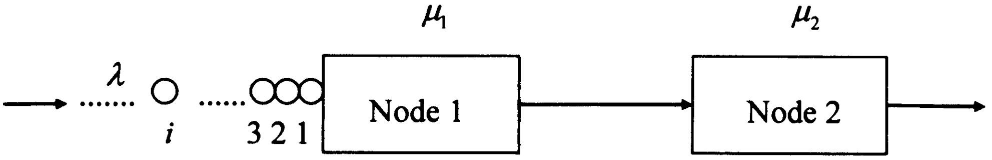


Figure 3.7 - Two node network with infinite space to queue in front of first node.

No new notation needs to be defined as there are no extra probabilities that can not be described by using the notation from the previous system, however the probabilities involving n will not be required as there is no limit on the possible queue size. The time dependent solutions for this system are;

$$
\begin{aligned}
P_{0,0}(t+\delta t) &= (1-\lambda\delta t)P_{0,0}(t)+\mu_2\delta t(1-\lambda\delta t)P_{0,1}(t) \\
P_{1,0}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)P_{1,0}(t)+\lambda\delta t P_{0,0}(t) \\
&\quad +(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta t P_{1,1}(t) \\
P_{i,0}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)P_{i,0}(t)+\lambda\delta t(1-\mu_1\delta t)P_{i-1,0}(t) && i=2,3,\ldots \\
&\quad +(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta t P_{i,1}(t) \\
P_{0,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{0,1}(t)+(1-\lambda\delta t)\mu_1\delta t P_{1,0}(t) \\
&\quad +(1-\lambda\delta t)\mu_2\delta t P_{1b,1}(t) \\
P_{1,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1}(t) \\
&\quad +\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{0,1}(t)+(1-\lambda\delta t)\mu_1\delta t P_{2,0}(t) \\
&\quad +(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta t P_{2b,1}(t) \\
P_{i,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{i,1}(t) && i=2,3,\ldots \\
&\quad +\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{i-1,1}(t)+(1-\lambda\delta t)\mu_1\delta t P_{i+1,0}(t) \\
&\quad +(1-\lambda\delta t)(1-\mu_1\delta t)\mu_2\delta t P_{(i+1)b,1}(t) \\
P_{1b,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{1b,1}(t)+(1-\lambda\delta t)\mu_1\delta t(1-\mu_2\delta t)P_{1,1}(t) \\
P_{ib,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{ib,1}(t)+\lambda\delta t(1-\mu_2\delta t)P_{(i-1)b,1}(t) && i=2,3,\ldots \\
&\quad +(1-\lambda\delta t)\mu_1\delta t(1-\mu_2\delta t)P_{n,1}(t)
\end{aligned}
\tag{3.17}
$$

The steady state solutions of these equations have been found in terms of lower order probabilities, as with the previous system.

$$
\begin{aligned}
P_{0,1} &= \frac{\lambda}{\mu_2} P_{0,0} \\
P_{1,0} &= \frac{\lambda\left(\lambda^2 + 2\lambda\mu_2 + \mu_2^2 + \mu_1\mu_2\right)}{\mu_1\mu_2\left(2\lambda + \mu_1 + \mu_2\right)} P_{0,0} \\
P_{1,1} &= \frac{\lambda^2\left(\lambda + \mu_2\right)\left(\lambda + \mu_1 + \mu_2\right)}{\mu_1\mu_2^2\left(2\lambda + \mu_1 + \mu_2\right)} P_{0,0} \\
P_{1b,1} &= \frac{\lambda^2\left(\lambda + \mu_1 + \mu_2\right)}{\mu_2^2\left(2\lambda + \mu_1 + \mu_2\right)} P_{0,0} \\
P_{i,0} &= \frac{\left(\lambda + \mu_2\right)\left(\lambda + \mu_1 + \mu_2\right)P_{i-1,1} - \lambda\left(\lambda + \mu_2\right)P_{i-2,1} - \lambda\mu_2 P_{(n-1)b,1} + \mu_1\lambda P_{n-1,0}}{\mu_1\left(2\lambda + \mu_1 + \mu_2\right)} \qquad i = 2,3,\ldots \\
P_{ib,1} &= \frac{\left(\lambda + \mu_1\right)\left(\lambda + \mu_1 + \mu_2\right)P_{i-1,1} - \lambda\left(\lambda + \mu_1\right)P_{i-2,1} - \lambda\mu_1 P_{i-1,0} + \lambda\mu_2 P_{(i-1)b,1}}{\mu_2\left(2\lambda + \mu_1 + \mu_2\right)} \qquad i = 2,3,\ldots \\
P_{i,1} &= \frac{\begin{array}{c}\left(\lambda + \mu_1\right)\left(\lambda + \mu_2\right)\left(\lambda + \mu_1 + \mu_2\right)P_{i-1,1} - \lambda\left(\lambda + \mu_1\right)\left(\lambda + \mu_2\right)P_{i-2,1} \\ - \lambda\mu_1\left(\lambda + \mu_2\right)P_{i-1,0} - \lambda\mu_2\left(\lambda + \mu_1\right)P_{(i-1)b,1}\end{array}}{\mu_1\mu_2\left(2\lambda + \mu_1 + \mu_2\right)} \qquad i = 2,3,\ldots
\end{aligned}
\tag{3.18}
$$

3.1.7 'Drip Feed' Queue

For all these equations to be in a steady state there has to be a constraint on the size of λ. The constraint on λ prevents a large queue forming and thus making the system erratic and not in a steady state. It does this by making sure that the mean arrival rate does not exceed a value that the service nodes can cope with. This constraint is found by finding the maximum throughput that the system can handle. This is the same as finding the maximum rate at which Node 1 can process customers. In a single node system, steady state equations hold when $\lambda \leq \mu$. This means that the arrival rate λ has to be less than μ, the serving rate, because μ is the fastest average rate that the system can process customers. In the two node case, Node 1 can become blocked which can affect the rate at which it can process customers.

Initially, to find the maximum rate at which customers could arrive into this two node case, a 'drip feed' queueing system was considered. This is when the first node is never allowed to be empty; it is either working or blocked. As soon as a customer leaves the first node a new customer is fed in to the system. This can also be thought of as an infinite queue of customers in front of the first node waiting to enter the system. This means that the probabilities required will be the same as when modelling the situation in Figure 3.2 but those probabilities which allow an empty first node can be disregarded and new steady state equations can be derived. These are the initial time dependent equations;

$$
\begin{aligned}
&P_{1,0}(t+\delta t)=(1-\mu_1\delta t)P_{1,0}(t)+(1-\mu_1\delta t)\mu_2\delta tP_{1,1}(t)\\
&P_{1,1}(t+\delta t)=(1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1}(t)+\mu_1\delta tP_{1,0}(t)+\mu_2\delta tP_{1b,1}(t)\\
&P_{1b,1}(t+\delta t)=(1-\mu_2\delta t)P_{1b,1}(t)+\mu_1\delta t(1-\mu_2\delta t)P_{1,1}(t)
\end{aligned}
\tag{3.19}
$$

Solving these in a similar way to previously, the following steady state equations are found;

$$
\begin{aligned}
P_{1,0}&=\frac{\mu_2^2}{\mu_1^2+\mu_1\mu_2+\mu_2^2}\\
P_{1,1}&=\frac{\mu_1\mu_2}{\mu_1^2+\mu_1\mu_2+\mu_2^2}\\
P_{1b,1}&=\frac{\mu_1^2}{\mu_1^2+\mu_1\mu_2+\mu_2^2}
\end{aligned}
\tag{3.20}
$$

It is worth noting that when the service rates are equal, the probability of being in any of the three states is equal, and therefore equal to a third.

3.1.8 Maximum Utilisation

Customers in Node 1 who pass through into Node 2 without first becoming blocked will spend an average of $\frac{1}{\mu_1}$ in the first node. Customers who become blocked in the first node after completing their service will spend an average of $\frac{1}{\mu_2}$ in the first node. This will be the case because in this two node system both services commence simultaneously, as Figure 3.8 represents. If a customer does not become blocked they will take Route 1, as soon as their service ends in Node 1 they move to the second node and commence service, at the same time a new customer moves into the first node and starts a service. If blocking has taken place, Route 2 is used and the customer waits in Node 1 until the customer in Node 2 has completed their service, then as in the previous case they move into the second node to commence service and simultaneously a new customer enters Node 1 to start their service. This ensures that all services start simultaneously.

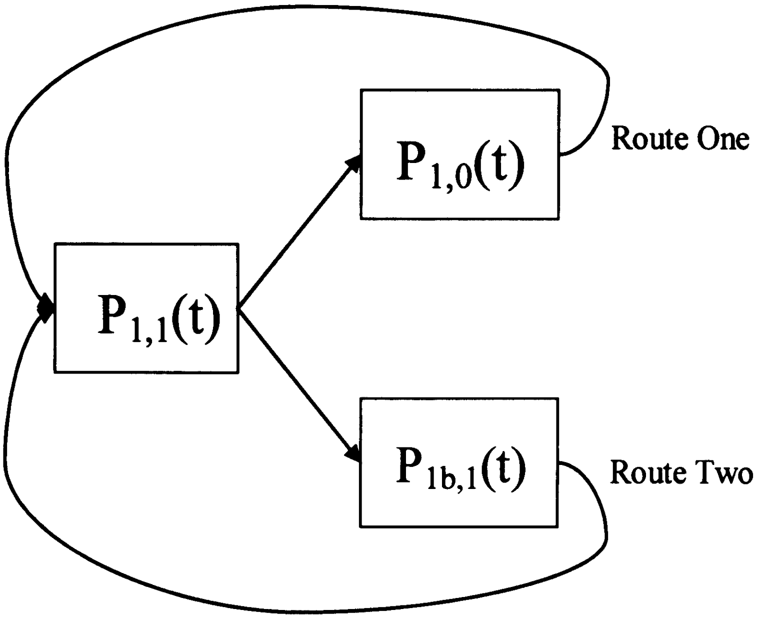


Figure 3.8 - Route map of two node `drip feed' system.

To compute the average time spent in the first node, the average time it takes to serve a customer in the first node (this is the part that a single node system takes) is taken, then the average time spent in Node 2 multiplied by the proportion of blocked customers is added to it;

Average time spent in Node 1+ $P_{1b,1}(t)$ *average time spent in Node 2

$$\frac{1}{\mu_1}+\frac{1}{\mu_2}\frac{\mu_1^2}{\mu_1^2+\mu_1\mu_2+\mu_2^2}=\frac{\mu_2\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right)+\mu_1^2}{\mu_1\mu_2\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right)} \tag{3.21}$$

From this the reciprocal is taken to give the average rate for which the first node processes customers and hence the maximum arrival rate.

$$\frac{\mu_1\mu_2\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right)}{\mu_2\left(\mu_1^2+\mu_1\mu_2+\mu_2^2\right)+\mu_1^2} \tag{3.22}$$

This does not agree with literature (Hunt 1956) or with simulated values. To find out where this formula is incorrect it is important to understand what these steady state probabilities actually represent. If this system was running and it was paused at an arbitrary moment in time, the probability of it being in a specific state is given by its relevant steady state probability. This is because the steady state probability is a proportion of how much time the system spends in each state. This however is not the same as the proportion of customers who become blocked. A simple example of this is a 'drip feed' type queue with two nodes with the same Negative Exponential service times, both having an average service time of one unit. In this example it would be expected that half of the customers would become blocked because half of the time Node 1 would finish before Node 2, and the other half of the time Node 2 would finish before Node 1. The steady state probabilities for this system are $P_{1,1}(t)=\frac{1}{3}$, $P_{1,0}(t)=\frac{1}{3}$ and $P_{1b,1}(t)=\frac{1}{3}$. This shows that an even amount of time was spent in each of the different states and that $P_{1b,1}(t)$ is not equal to the probability of a customer becoming blocked.

So to calculate the actual maximum process rate the following expression must be used;

$$\frac{1}{\mu_1}+P(Blocked)\frac{1}{\mu_2} \tag{3.23}$$

This is similar to the previous formula but with $P_{1b,1}(t)$ replaced by $P(Blocked)$, the probability of a customer being blocked. To find the probability of a customer becoming blocked, the probability that $T_1 < T_2$ has to be found, where T_1 and T_2 are the times customers spend in Node 1 and 2 respectively.

$$\begin{aligned}
&P(T_1<T_2)=\int_0^\infty P(T_1<t_2 \mid T_2=t_2)f(t_2)\delta t_2\\
&P(T_1<t_2 \mid T_2=t_2)=\int_0^{t_2}\mu_1 e^{-\mu_1 t_1}dt_1\\
&P(T_1<t_2 \mid T_2=t_2)=1-e^{-\mu_1 t_2}\\
&P(T_1<T_2)=\int_0^\infty\left(1-e^{-\mu_1 t_2}\right)\mu_2 e^{-\mu_2 t_2}dt_2\\
&P(T_1<T_2)=\int_0^\infty \mu_2 e^{-\mu_2 t_2}dt_2-\int_0^\infty \mu_2 e^{-t_2(\mu_1+\mu_2)}dt_2\\
&P(T_1<T_2)=1-\frac{\mu_2}{\mu_1+\mu_2}=\frac{\mu_1}{\mu_1+\mu_2}
\end{aligned} \tag{3.24}$$

This is the probability that, when there are services taking place in Node 1 and Node 2, Node 1's service will be completed first, thus leading to a blocked situation. This result is expected from a common sense point of view. From this the average time spent in Node 1 can be found as follows;

$$\frac{1}{\mu_1}+\frac{\mu_1}{\mu_1+\mu_2}\frac{1}{\mu_2}=\frac{\mu_2(\mu_1+\mu_2)+\mu_1^2}{\mu_1\mu_2(\mu_1+\mu_2)}=\frac{\mu_1^2+\mu_1\mu_2+\mu_2^2}{\mu_1\mu_2(\mu_1+\mu_2)} \tag{3.25}$$

The average rate at which Node 1 can process customers is the reciprocal of the average time;

$$\frac{\mu_1\mu_2\left(\mu_1+\mu_2\right)}{\mu_1^2+\mu_1\mu_2+\mu_2^2} \tag{3.26}$$

This is the maximum rate at which customers can arrive so that the system is in a steady state; this value is called $\lambda_{\max}$. The maximum utilisation $\rho_{\max}$ is defined as being;

$$\begin{aligned}\rho_{\max} &= \frac{\lambda_{\max}}{\mu_1} \\ &= \frac{\mu_2\left(\mu_1+\mu_2\right)}{\mu_1^2+\mu_1\mu_2+\mu_2^2}\end{aligned} \tag{3.27}$$

This should not be interpreted as the rate at which the system outputs its customers. This is because this only accounts for the first node. To find the rate of output for the system as a whole, first sum the average time spent in each node, then take the reciprocal of that sum. The average time spent in the system is;

$$\frac{\mu_1^2+\mu_1\mu_2+\mu_2^2}{\mu_1\mu_2\left(\mu_1+\mu_2\right)}+\frac{1}{\mu_2}=\frac{\mu_1^2+\mu_1\mu_2+\mu_2^2+\mu_1\left(\mu_1+\mu_2\right)}{\mu_1\mu_2\left(\mu_1+\mu_2\right)} \tag{3.28}$$

So the average output rate of the entire system is the reciprocal of this, i.e.

$$\frac{\mu_1\mu_2\left(\mu_1+\mu_2\right)}{\mu_1^2+\mu_1\left(\mu_1+2\mu_2\right)+\mu_2^2} \tag{3.29}$$

These values agree with the existing literature, but were derived using a different method. Hunt uses a raising operator on the system equations, and then creates a matrix of these equations from which these results are then derived. The method shown in this paper appears to be more intuitive method of producing these results.

3.1.9 Simulation

A Visual Basic program has been created to verify these results. The pseudo code for this program is as below. The state of the system is defined to be whether the nodes have customers present and their current status i.e. whether they are blocked or undergoing a service.

- Only data for four customers at a time is stored. This reduces computation time by having smaller arrays. The customers recorded are the two being served at the nodes, the newly arriving customer and the next customer
- The recorded data is arrival time, start time of each service at each node, end time of each service at each node and the time that the customer leaves Node 1.
- Customer arrival time is sampled from a negative Exponential distribution such that the value is greater than the previous customer's first node finish time ensuring that there is no queue.
- Customer service times for both nodes are calculated.
- End of service times for consecutive customers at consecutive nodes are compared. If Node 2 finishes first, the time that the customer leaves Node 1 is set to the end of Node 1's service. If the first node finishes its service first, then the time the customer leaves Node 1 is set to the end of Node 2's service.
- The time a customer leaves Node 2 is set as the time Node 2's service completes, as there is no blocking possible there.
- A loop is set up such that each time a new customer arrives into the system the time spent in each state is summed, until another customer arrives. This is done by a series of 'if' statements to see whether a node is empty, working or blocked.
- When a new customer arrives into the system, the times are all shifted on in the array making space for the new customer's data, and the loop repeats until the required number of customers have passed through the system.

For an example of the maximum utilisation of the system, both μ 's will be set equal to 1. In this case the $\rho_{\max} = \frac{2}{3}$, so this system would remain in a steady state for all values of $\rho_{\max} < \frac{2}{3}$. This is in comparison to an M|M|1 system where $\rho_{\max} < 1$ for the system to be in a steady state. Practically this means that, for the queue size to remain stable, $\rho_{\max} < 1$ and when the maximum utilisation is breached the queue size will become unstable and start to grow.

A two node blocking situation simulation has been created in Visual Basic, which allows an infinite queue to build up in front of the first node and no space for a queue in between the two nodes. A run size of 1 million customers was set, the queue size was recorded every time; a new service started, a current service ended and with a new arrival into the system. Each of these is defined as an event. Figure 3.9 is a graph of how the queue size changes for different values of λ around $\lambda_{\max}$ as more events occur. The queue size is recorded for every thousandth event.

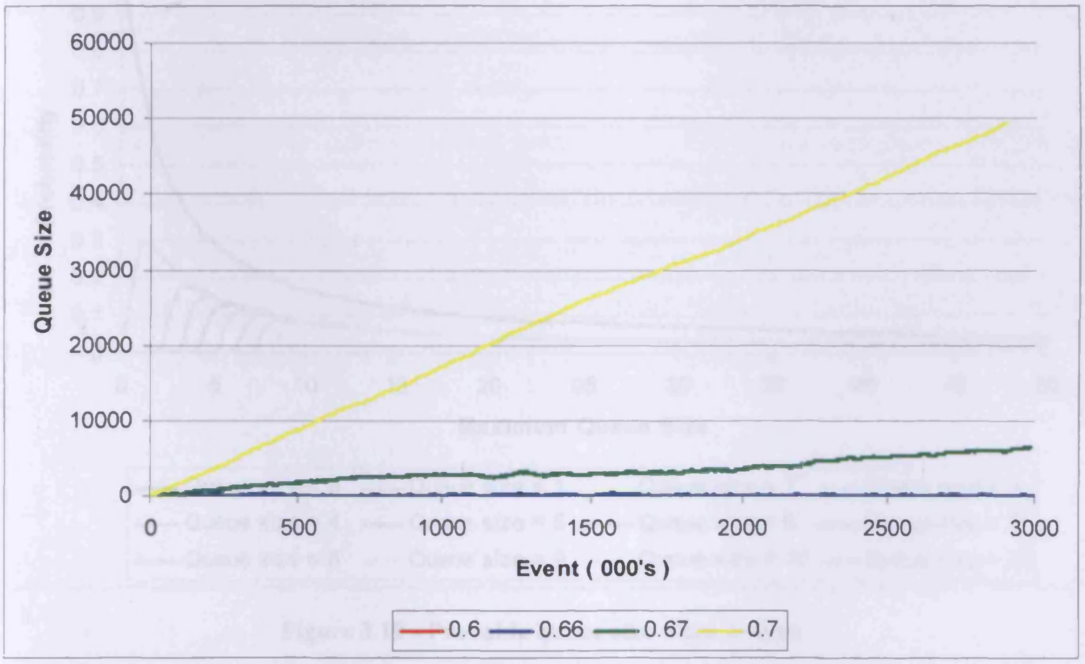


Figure 3.9 - Queue size for differing values of λ

As shown with the theoretical results it can be seen that the queue size remains steady when $\lambda < \lambda_{\max}$ and the queue size starts to increase as soon as λ is greater than the maximum utilisation of $\frac{2}{3}$.

Another way of looking at the maximum utilisation is with using the steady state. In this method there is no physical queue to count but it is possible to sum all the states that have an *n* size queue and say that the summation is equal to the probable queue size. The limitation with this technique is that to be able to solve these steady state equations there must be a fixed value for the queueing capacity as seen in equation (3.29). When the arrival rate is less than the maximum utilisation the most probable queue size should be constant, so that the queue size would always have the same most likely size value, creating a steady state. As the arrival rate increases above the maximum utilisation, the queue size would increase and correspondingly the most probable queue size would change as the conditions of steady state are breached.

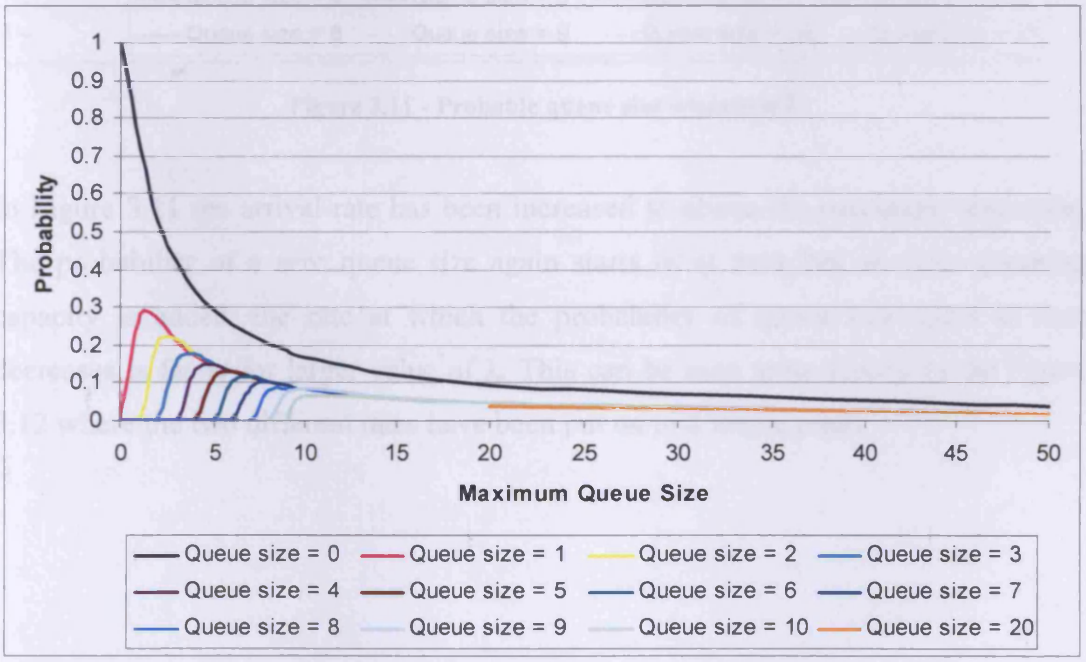


Figure 3.10 - Probable queue size when λ=0.66

Figure 3.10 shows that the probability of a particular queue size for a specific maximum queueing capacity. When there is no queue allowed before the first node, the probability of a queue size of zero is unsurprisingly 1. A queue capacity of 1 is introduced and the probability of a queue size zero is reduced but is still higher than the probability the queue has size 1. If this pattern carries on as more queueing capacity is added, a zero size queue is always the most probable. This has not been tested any further than a maximum queue size of 50 because of computational time.

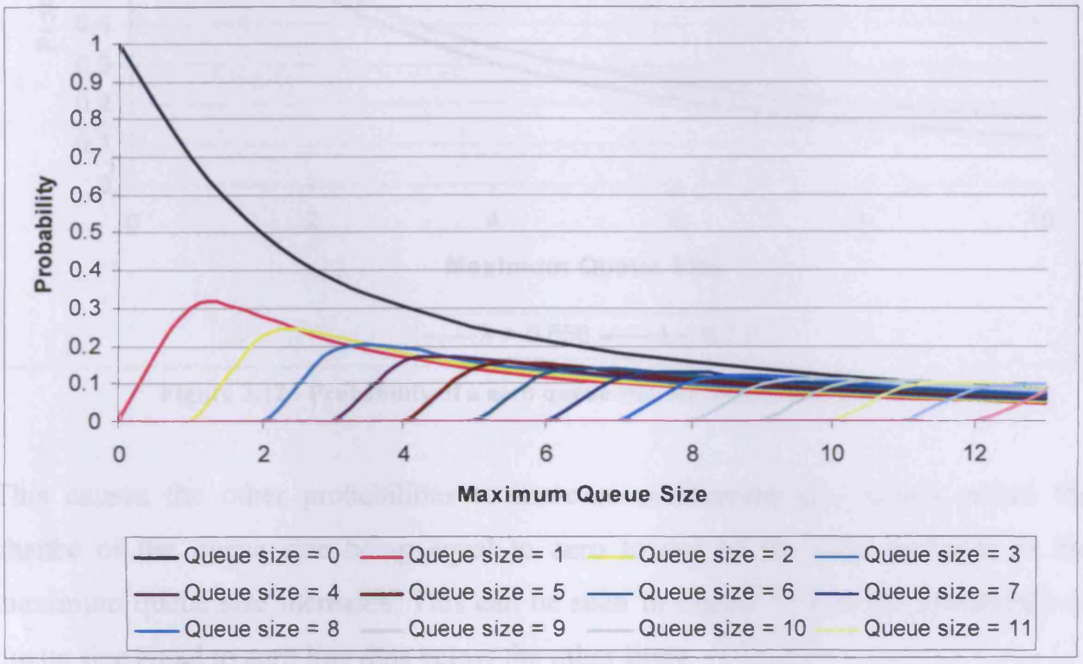


Figure 3.11 - Probable queue size when λ=0.7

In Figure 3.11 the arrival rate has been increased to above the maximum utilisation. The probability of a zero queue size again starts of at zero, but as more queueing capacity is added, the rate at which the probability of queue size equal to zero decreases is faster for larger value of λ. This can be seen more clearly in the Figure 3.12 where the two different rates have been put on to a single graph.

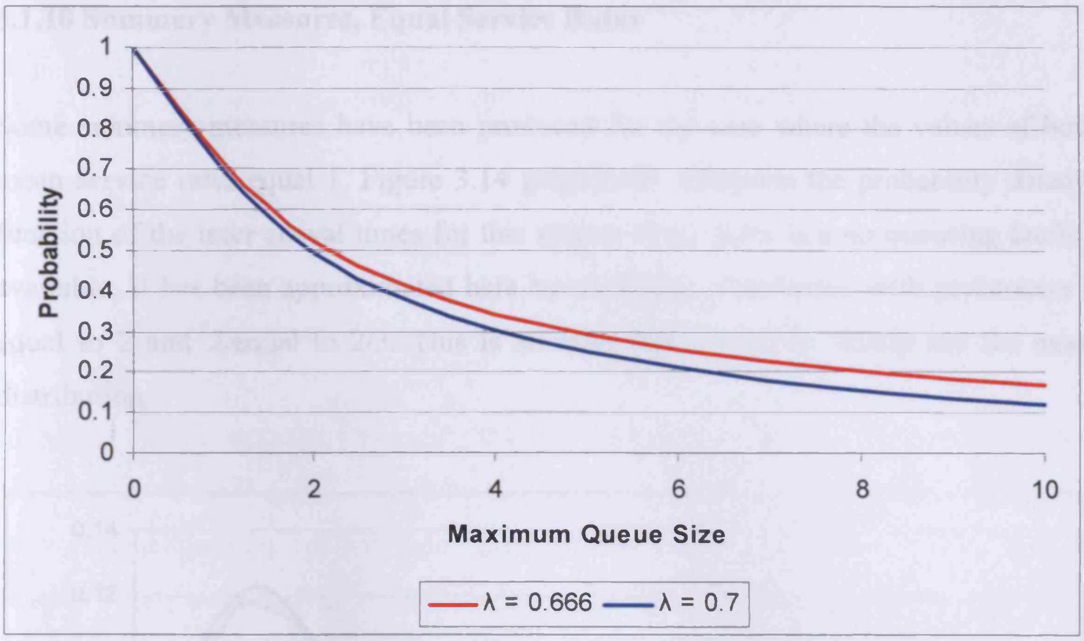


Figure 3.12 - Probability of a zero queue size for different values of λ

This causes the other probabilities to increase correspondingly, which makes the chance of the queue size being equal to zero to not be the most probable as the maximum queue size increases. This can be seen in Figure 3.13 as the probability of queue size equal to zero line dips below the other lines.

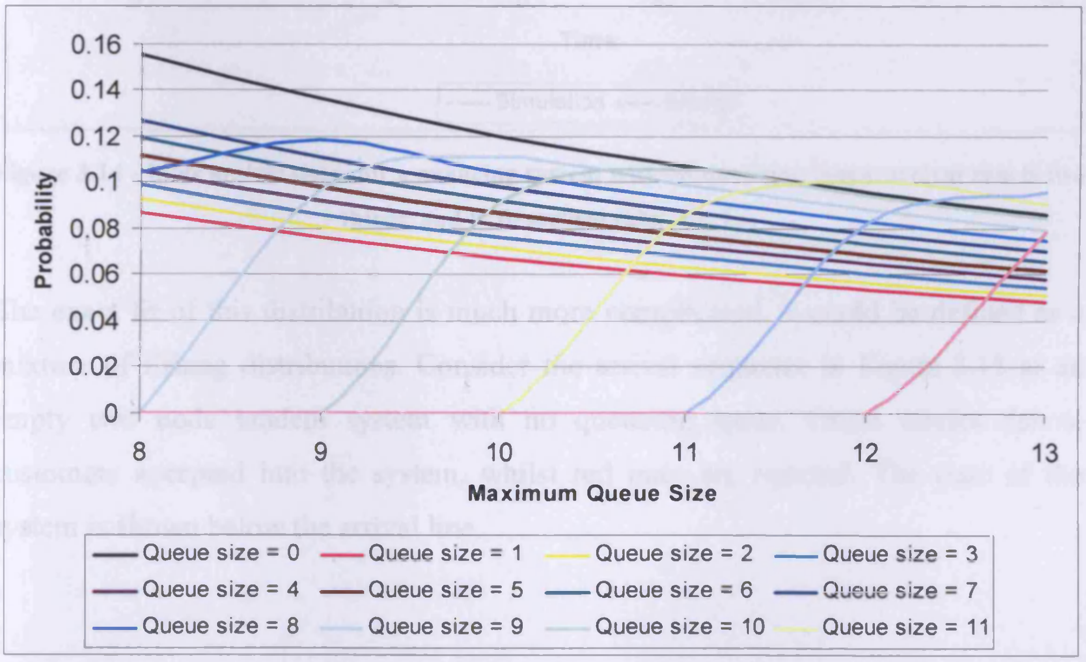


Figure 3.13 - Probable queue size when λ=0.7

3.1.10 Summary Measures, Equal Service Rates

Some summary measures have been produced for the case where the values of both mean service rates equal 1. Figure 3.14 graphically interprets the probability density function of the inter arrival times for this system when there is a no queueing facility available. It has been approximated here by an Erlang distribution with parameters k equal to 2 and λ equal to 2/3. This is an estimated fit and is clearly not the exact distribution.

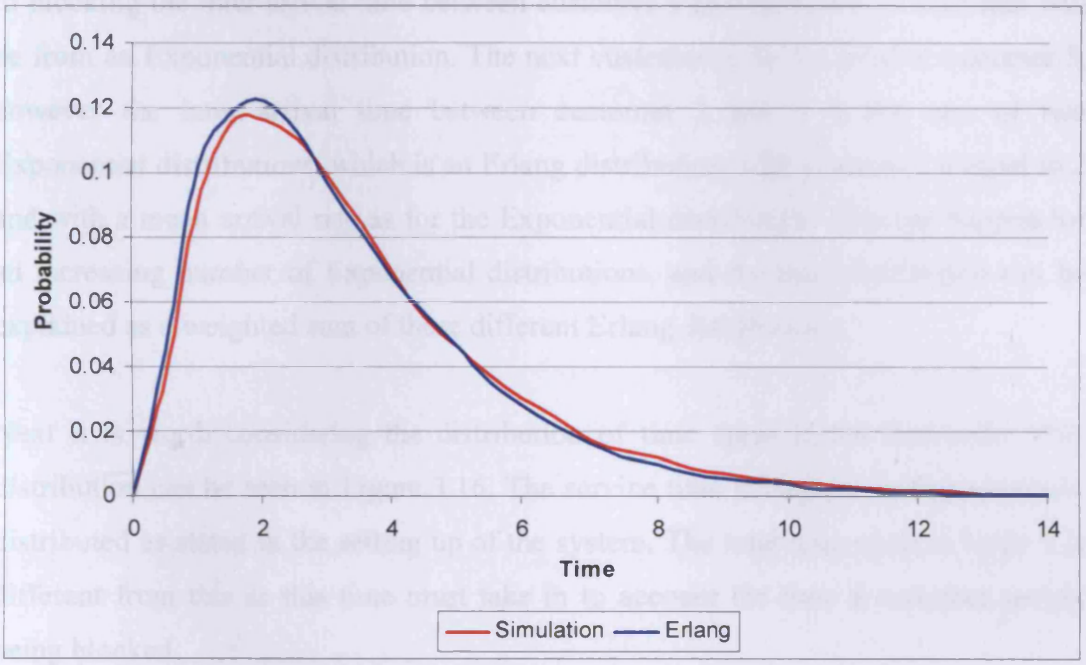


Figure 3.14 - Inter arrival times for a queueing system with no queueing space, arrival rate is two thirds, and both service rates are 1.

The exact fit of this distribution is much more complicated. It could be defined as a mixture of Erlang distributions. Consider the arrival sequence in Figure 3.15 at an empty two node tandem system with no queueing space. Green circles denote customers accepted into the system, whilst red ones are rejected. The state of the system is shown below the arrival line.

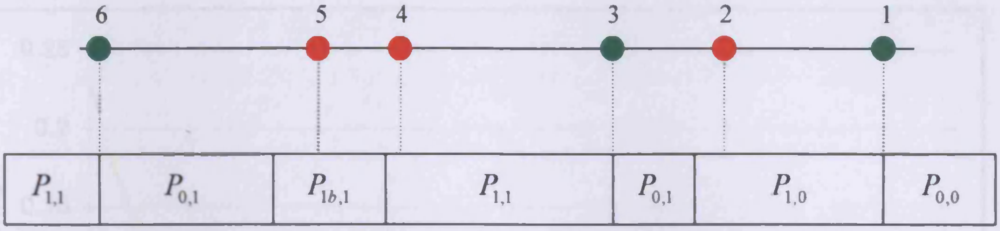


Figure 3.15 - Arriving sequence at a two node system with no queueing space

Customer 1 is accepted and assuming that the previous customer was not rejected due to blocking the inter arrival time between customer 1 and the previous customer will be from an Exponential distribution. The next customer to be accepted is customer 3, however the inter arrival time between customer 1 and 3 is the sum of two Exponential distributions, which is an Erlang distribution, with parameter k equal to 2 and with a mean arrival rate as for the Exponential distribution. This can happen for an increasing number of Exponential distributions, and the final distribution can be explained as a weighted sum of these different Erlang distributions.

Next it is worth considering the distribution of time spent in the first node. This distribution can be seen in Figure 3.16. The service time is Negatively Exponentially distributed as stated in the setting up of the system. The total time spent in Node 1 is different from this as this time must take in to account the time a customer spends being blocked.

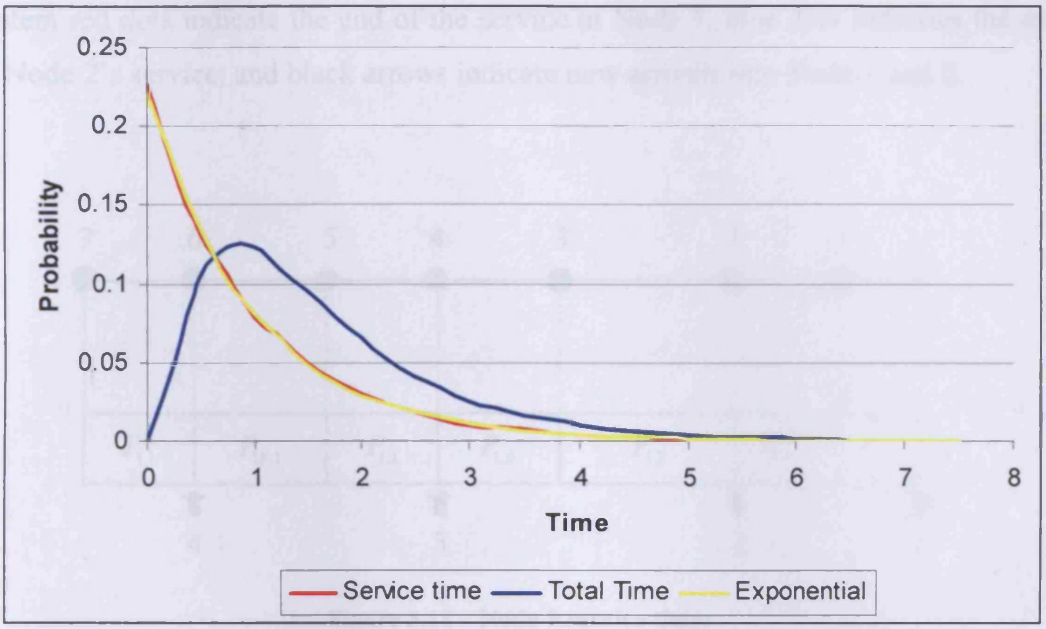


Figure 3.16 - Time spent in the first node of a drip feed system with both service rates equal to 1, alongside an Exponential distribution with parameter equal to 1.

Figure 3.17 breaks the total time spent in Node 1 into customers that were blocked and those that were not. It shows that these customers have the same distribution of time spent in this node.

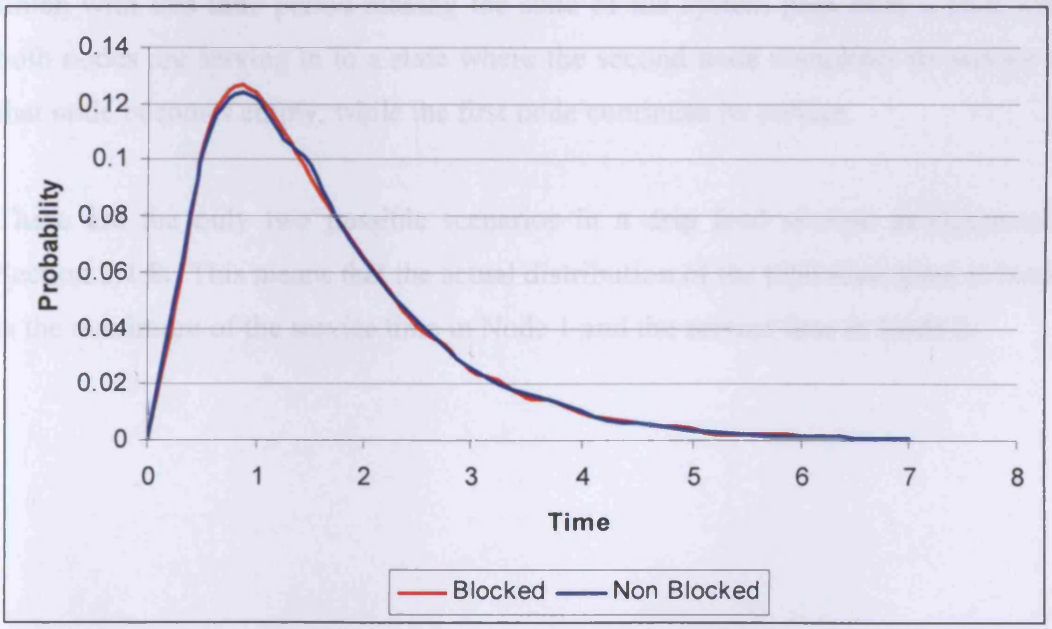


Figure 3.17 - Time spent in Node 1 of a drip feed system for blocked and non blocked customers, with service rates equal to 1.

This may not be intuitive; Figure 3.18 helps to explain this situation. In this drip feed system red dots indicate the end of the service at Node 1, blue dots indicates the end of Node 2's service, and black arrows indicate new arrivals into Node 1 and 2.

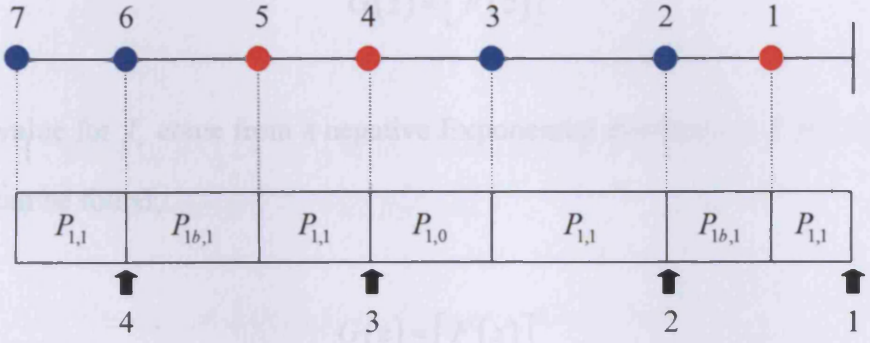


Figure 3.18 – Node 1 service times

In Figure 3.18 between arrival point 1 (indicated by the black arrow 1) and arrival point 2, both service 1 and 2 must end. Within this time the state of the system changes from both nodes serving to the first node becoming blocked whilst the second node continues its service. As soon as the second service ends new customers enter into each node. Between arrivals 2 and 3, 2 services come to an end. Service 3 and 4 finish with this time period making the state of the system pass from a state where both nodes are serving in to a state where the second node completes its service and that node becomes empty, while the first node continues its service.

These are the only two possible scenarios in a drip feed system, as discussed in Section 3.1.8. This means that the actual distribution of the total time spent in Node 1 is the maximum of the service time in Node 1 and the service time in Node 2.

If, as in this example, the service rates are equal, then let $Z = \max(T_1, T_2)$ where T_i is the service time at service point i. It is a well known result that if the cumulative density function (CDF) of T is $F(t)$ then the CDF of $Z, G(z)$ is;

$$G(z) = [F(z)]^2 \tag{3.30}$$

As the value for T_i come from a negative Exponential distribution, $F(t)$ is known so $G(z)$ can be found;

$$\begin{aligned} G(z) &= [F(z)]^2 \\ &= \left[1 - e^{-\mu z}\right]^2 \end{aligned} \tag{3.31}$$

Figure 3.19 shows the distribution of the maximum of two Exponential distributions along with the data of time spent in the first node from the simulation. It can be seen to be a realistic description of the time spent in Node 1.

The probability density function (PDF) of $Z, g(z)$ can also be derived from well known results.

$$\begin{aligned} g(z) &= 2f(z)F(z) \\ &= 2\mu e^{-\mu z}\left(1 - e^{-\mu z}\right) \end{aligned} \tag{3.32}$$

Using (3.32) the expected value of Z can be found.

$$\begin{aligned} E(z) &= \int_0^\infty z g(z)\,dz \\ &= 2\int_0^\infty z\mu e^{-\mu z}\left(1-e^{-\mu z}\right)dz \qquad (3.33) \\ &= \frac{3}{2\mu} \end{aligned}$$

When the value of μ is equal to 1, as in this example the, the mean time spent in Node 1 is 1.5. This value is consistent with the data.

Figure 3.19 also has an Erlang PDF displayed. The Erlang PDF is shown as it is a close approximation to the maximum distribution. The Erlang distribution it is often easier to deal with and derive. It can be seen to be a very reasonable approximation. The value for the shape parameter, k, was set equal to 2 and λ was then found by equating the means of the Erlang distribution with that of the maximum distribution.

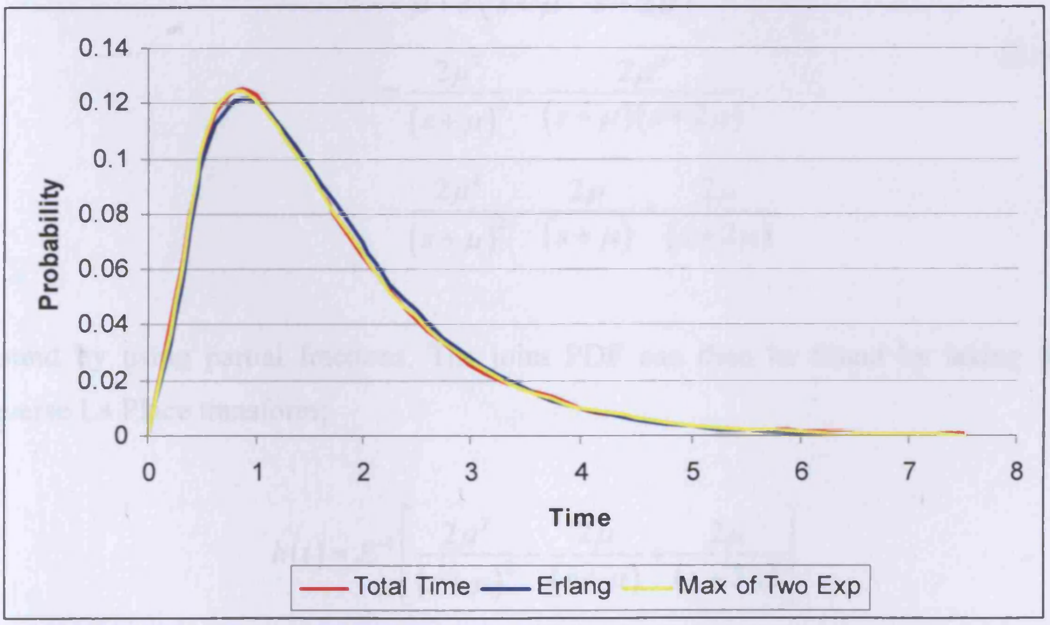


Figure 3.19 - Time spent in Node 1 of a drip feed tandem queue, with service rates equal to 1, compared with; an Erlang distribution with λ = 4/3 and k = 2 and the distribution of the maximum of two Exponential distributions with λ =1.

The mean total time in the system can easily been seen to be the sum of the total time spent in Node 1 and Node 2. The time spent in Node 2 is a standard negative Exponential distribution as no blocking occurs at the second service point. So the mean total time spent in the system (excluding any queueing) in this example is 2.5. To find out the distribution for the total time spent in system (excluding any queueing) the Convolution Theorem is required;

$$\mathcal{L}\left[h(t)\right]=\mathcal{L}\left[f(t)\right]\mathcal{L}\left[g(t)\right] \qquad (3.34)$$

Using this it is possible to find the joint PDF, $h(t)$, for the time spent in the service system. Letting $f(t)$ be the PDF of total time in Node 1 and $g(t)$ be the PDF of the time spent in Node 2, the joint distribution can be found;

$$\begin{aligned}\mathcal{L}\left[h(t)\right]&=\mathcal{L}\left[2\mu e^{-\mu t}\left(1-e^{-\mu t}\right)\right]\mathcal{L}\left[\mu e^{-\mu t}\right]\\&=\frac{\mu}{\mu+s}\left(\frac{2\mu}{s+\mu}-\frac{2\mu}{s+2\mu}\right)\\&=\frac{2\mu^2}{(s+\mu)^2}-\frac{2\mu^2}{(s+\mu)(s+2\mu)}\\&=\frac{2\mu^2}{(s+\mu)^2}-\frac{2\mu}{(s+\mu)}+\frac{2\mu}{(s+2\mu)}\end{aligned} \qquad (3.35)$$

Found by using partial fractions. The joint PDF can then be found by taking the inverse La Place transform;

$$\begin{aligned}h(t)&=\mathcal{L}^{-1}\left[\frac{2\mu^2}{(s+\mu)^2}-\frac{2\mu}{(s+\mu)}+\frac{2\mu}{(s+2\mu)}\right]\\&=2\mu^2te^{-\mu t}-2\mu e^{-\mu t}+2\mu e^{-2\mu t}\\&=2\mu e^{-\mu t}\left(\mu t-1+e^{-\mu t}\right)\end{aligned} \qquad (3.36)$$

The expectation of this distribution can be found

$$
\begin{aligned}
E(t) &= \int_0^\infty 2t\mu e^{-\mu t}\left(\mu t - 1 + e^{-\mu t}\right)dt \\
&= \int_0^\infty 2t^2\mu^2 e^{-\mu t}dt - \int_0^\infty 2t\mu e^{-\mu t}dt + \int_0^\infty 2t\mu e^{-2\mu t}dt \\
&= 2\mu\left[2\int_0^\infty t e^{-\mu t}dt\right] - \int_0^\infty 2t\mu e^{-\mu t}dt + \int_0^\infty 2t\mu e^{-2\mu t}dt \\
&= 2\int_0^\infty t\mu e^{-\mu t}dt + \int_0^\infty 2t\mu e^{-2\mu t}dt \qquad (3.37) \\
&= 2\int_0^\infty e^{-\mu t}dt + \int_0^\infty e^{-2\mu t}dt \\
&= \frac{2}{\mu} + \frac{1}{2\mu} \\
&= \frac{5}{2\mu}
\end{aligned}
$$

Figure 3.20 shows the PDF of the total time in the system with the data and again with an Erlang approximation with parameters gathered in a similar way to previously.

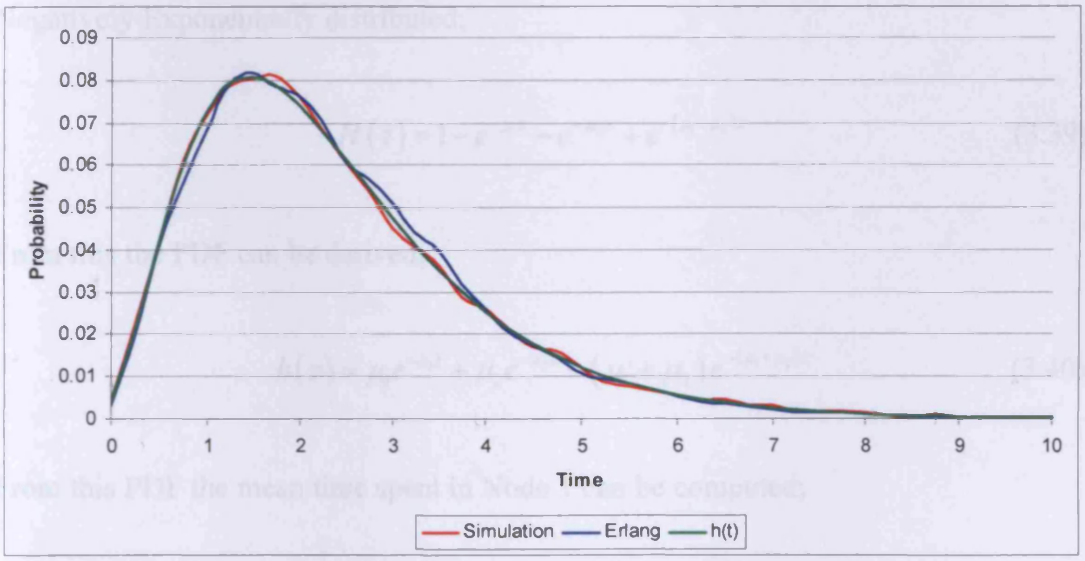


Figure 3.20 - Total time spent in a two node drip feed tandem system with service rate equal to 1, compared with an Erlang distribution with $\lambda = 6/5$ and $k = 3$, and the PDF of the Total time as given by (3.35).

3.1.11 Summary Measures, Unequal Service Rates

Expanding the example above, the PDF of the total time spent in the system and in Node 1, excluding queueing, can be calculated. The PDF of the total time spent in Node 1 is the maximum, *Z*, of two Exponential distributions. The solution has already been shown for the case when the service rates are equal, equations (3.35), so consider the case when the service rates are different but still independent. Letting *X* be the service time in Node 1 and *Y* being the service time in Node 2 (the change in notation is to make clear that these two values may come from negative Exponential distributions with different parameters);

$$\begin{aligned} P(Z<z) &= P(\max(X,Y)<z) \\ &= P(X<z, Y<z) \\ &= P(X<z)P(Y<z) \\ H(z) &= F(z)G(z) \end{aligned} \tag{3.38}$$

Where *F*, *G* and *H* are the CDF of *X*, *Y* and *Z* respectively. As these are both from a Negatively Exponentially distributed;

$$H(z) = 1 - e^{-\mu_1 z} - e^{-\mu_2 z} + e^{-(\mu_1+\mu_2)z} \tag{3.39}$$

From this the PDF can be derived;

$$h(z) = \mu_1 e^{-\mu_1 z} + \mu_2 e^{-\mu_2 z} - (\mu_1 + \mu_2) e^{-(\mu_1+\mu_2)z} \tag{3.40}$$

From this PDF the mean time spent in Node 1 can be computed;

$$
\begin{aligned}
E(z) &= \int_0^\infty z h(z)\,dz \\
&= \int_0^\infty z\mu_1 e^{-\mu_1 z}\,dz + \int_0^\infty z\mu_2 e^{-\mu_2 z}\,dz - \int_0^\infty z(\mu_1+\mu_2)e^{-(\mu_1+\mu_2)z}\,dz \\
&= \frac{1}{\mu_1} + \frac{1}{\mu_2} + \frac{1}{\mu_1+\mu_2} \\
&= \frac{\mu_1^2+\mu_1\mu_2+\mu_2^2}{\mu_1\mu_2(\mu_1+\mu_2)}
\end{aligned}
\tag{3.41}
$$

This result confirms the earlier result as seen in equation (3.27). Now that not only the mean time in Node 1 has been found but also the distribution of the time in Node 1 and the distribution of the total time can be found;

$$\mathcal{L}\left[h(t)\right]=\mathcal{L}\left[\mu_1 e^{-\mu_1 t}+\mu_2 e^{-\mu_2 t}-(\mu_1+\mu_2)e^{-(\mu_1+\mu_2)t}\right]\mathcal{L}\left[\mu_2 e^{-\mu_2 t}\right]$$

$$=\frac{\mu_2}{s+\mu_2}\left(\frac{\mu_1}{s+\mu_1}+\frac{\mu_2}{s+\mu_2}-\frac{\mu_1+\mu_2}{s+(\mu_1+\mu_2)}\right)$$

$$=\frac{\mu_1\mu_2}{(s+\mu_1)(s+\mu_2)}+\frac{\mu_2^2}{(s+\mu_2)^2}-\frac{\mu_2(\mu_1+\mu_2)}{(s+(\mu_1+\mu_2))(s+\mu_2)}$$

$$h(t)=\mathcal{L}^{-1}\left[\begin{array}{l}\dfrac{\mu_1\mu_2}{(\mu_2-\mu_1)}\left(\dfrac{1}{s+\mu_1}-\dfrac{1}{s+\mu_2}\right)\\ +\dfrac{\mu_2^2}{(\mu_1+\mu_2)^2}\\ -\dfrac{\mu_2(\mu_1+\mu_2)}{\mu_1}\left(\dfrac{1}{s+\mu_2}-\dfrac{1}{s+(\mu_1+\mu_2)}\right)\end{array}\right]$$

$$=\frac{\mu_1\mu_2}{\mu_2-\mu_1}\left(e^{-\mu_1 t}-e^{-\mu_2 t}\right)+\mu_2^2 t e^{-\mu_2 t}-\frac{\mu_2(\mu_1+\mu_2)}{\mu_1}\left(e^{-\mu_2 t}-e^{-(\mu_1+\mu_2)t}\right)$$

$$=\mu_2 e^{-\mu_2 t}\left(\mu_2 t-\frac{(\mu_1+\mu_2)}{\mu_1}-\frac{\mu_1}{\mu_2-\mu_1}\right)+\frac{\mu_1\mu_2}{\mu_2-\mu_1}e^{-\mu_1 t}+\frac{\mu_2(\mu_1+\mu_2)}{\mu_1}e^{-(\mu_1+\mu_2)t}$$

$$=\mu_2^2 e^{-\mu_2 t}\left(t-\frac{\mu_2}{\mu_1(\mu_2-\mu_1)}\right)+\frac{\mu_1\mu_2}{\mu_2-\mu_1}e^{-\mu_1 t}+\frac{\mu_2(\mu_1+\mu_2)}{\mu_1}e^{-(\mu_1+\mu_2)t}$$

(3.42)

The mean time spent in the total system can be found by finding the expected value of this function.

$$
\begin{aligned}
E(t) &= \int_0^\infty t h(t)\,dt \\
&= \mu_2 \int_0^\infty \mu_2 t^2 e^{-\mu_2 t}\,dt - \frac{\mu_2^2}{\mu_1(\mu_2-\mu_1)} \int_0^\infty \mu_2 t e^{-\mu_2 t}\,dt \\
&\quad + \frac{\mu_2}{\mu_2-\mu_1} \int_0^\infty \mu_1 t e^{-\mu_1 t}\,dt + \frac{\mu_2}{\mu_1} \int_0^\infty (\mu_1+\mu_2) t e^{-(\mu_1+\mu_2)t}\,dt
\end{aligned}
\qquad (3.43)
$$

Using integration by parts the expected value becomes;

$$
\begin{aligned}
E(t) &= \frac{2}{\mu_2} - \frac{\mu_2}{\mu_1(\mu_2-\mu_1)} + \frac{\mu_2}{\mu_1(\mu_2-\mu_1)} + \frac{\mu_2}{\mu_1\mu_2(\mu_1+\mu_2)} \\
&= \frac{2\mu_1^2 + 2\mu_1\mu_2 + \mu_2}{\mu_1\mu_2(\mu_1+\mu_2)}
\end{aligned}
\qquad (3.44)
$$

In the example when both mean service rates are the same and equal to 1, the average time taken to pass through the system was 2.5. Substituting these values into the expected value for the time spent in the whole system the same value is obtained.

3.2 Three Node Systems

3.2.1 Introduction

The three node tandem queue will be briefly studied here as it provides a useful insight into how queues in series behave though they will not be looked at in much depth due to the unnecessary complications that it would provide for the application in this piece of work, namely applying this model to a Critical Care Unit.

The three node case has the same structure as the two node case. There is of course an extra service node which customers enter once they have completed their service at the second node and if there is sufficient space for a new service to start in the third node. However only the limiting case when there is no queueing facility allowed in between any service nodes will be considered.

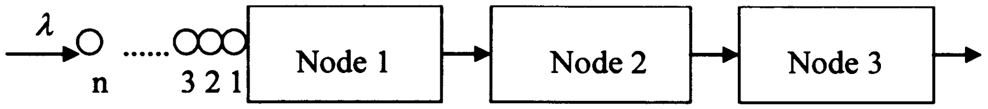


Figure 3.21 – A three node tandem queue

3.2.2 'Drip Feed' Case

The first case to be considered is the system with a 'drip feed` queue into the system described above. As there is now an extra node, an adaptation of the current notation is required for this type of system.

$P_{1,0,0}(t)$ Probability of Node 1 serving and all other nodes are empty.

$P_{1,1,0}(t)$ Probability of nodes 1 and 2 serving customers and Node 3 is empty.

$P_{1,1,1}(t)$ Probability all nodes are partaking in a service.

$P_{1,0,1}(t)$ Probability that nodes 1 and 3 are serving customers and Node 2 is empty.

$P_{1b,1,0}(t)$ Probability that Node 2 is serving a customer, Node 1 has completed its service but the customer is blocked and Node 3 is empty.

$P_{1b,1,1}(t)$ Probability that nodes 2 and 3 are serving customers and Node 1 has completed its service but the customer is blocked.

$P_{1b,1b,1}(t)$ Probability Node 3 is engaged in a service, with nodes 1 and 2 having completed their service but both are blocked.

$P_{1,1b,1}(t)$ Probability nodes 1 and 3 are serving customers and Node 2 has completed its service but the customer remains blocked.

The resulting initial time dependent equations of this three node system are;

$$\begin{aligned}
P_{1,0,0}(t+\delta t) &= (1-\mu_1\delta t)P_{1,0,0}(t)+(1-\mu_1\delta t)\mu_3\delta t P_{1,0,1}(t)\\
P_{1,1,0}(t+\delta t) &= (1-\mu_1\delta t)(1-\mu_2\delta t)P_{1,1,0}(t)+(1-\mu_1\delta t)(1-\mu_2\delta t)\mu_3\delta t P_{1,1,1}(t)\\
&\quad +\mu_1\delta t P_{1,0,0}(t)\\
P_{1,1,1}(t+\delta t) &= (1-\mu_1\delta t)(1-\mu_2\delta t)(1-\mu_3\delta t)P_{1,1,1}(t)+\mu_1\delta t(1-\mu_3\delta t)P_{1,0,1}(t)\\
&\quad +\mu_3\delta t P_{1b.1b.1}(t)+\mu_2\delta t P_{1b,1,0}(t)\\
P_{1,0,1}(t+\delta t) &= (1-\mu_1\delta t)(1-\mu_3\delta t)P_{1,0,1}(t)+(1-\mu_1\delta t)\mu_2\delta t P_{1,1,0}(t)\\
&\quad +(1-\mu_1\delta t)\mu_3\delta t P_{1,1b,1}(t)\\
P_{1b,1,0}(t+\delta t) &= (1-\mu_2\delta t)P_{1b,1,0}(t)+\mu_1\delta t(1-\mu_2\delta t)P_{1,1,0}(t)\\
&\quad +(1-\mu_2\delta t)\mu_3\delta t P_{1b,1,1}(t)\\
P_{1b,1,1}(t+\delta t) &= (1-\mu_2\delta t)(1-\mu_3\delta t)P_{1b,1,1}(t)+\mu_1(1-\mu_2\delta t)(1-\mu_3\delta t)P_{1,1,1}(t)\\
P_{1b,1b,1}(t+\delta t) &= (1-\mu_3\delta t)P_{1b,1b,1}(t)+\mu_1\delta t(1-\mu_3\delta t)P_{1,1b,1}(t)\\
&\quad +\mu_2\delta t(1-\mu_3\delta t)P_{1b,1,1}(t) \qquad (3.45)\\
P_{1,1b,1}(t+\delta t) &= (1-\mu_1\delta t)(1-\mu_3\delta t)P_{1,1b,1}(t)+(1-\mu_1\delta t)\mu_2\delta t(1-\mu_3\delta t)P_{1,1,1}(t)
\end{aligned}$$

When finding the steady state solutions, as well as using the equations above, we use the fact that all probabilities sum to 1;

$$\sum_i\sum_j\sum_k P_{i,j,k}=1 \qquad (3.46)$$

These can be found in terms of the 'lowest order` probability $P_{1,0,0}(t)$ and then using the final equation, the probabilities can all be found in terms of λ, μ_1, μ_2 and μ_3.

$$P_{1,0,1} = \frac{\mu_1}{\mu_3} P_{1,0,0}$$

$$P_{1,1,1} = \frac{\mu_1^2 (\mu_1 + \mu_3)(\mu_1 + \mu_2 + \mu_3)}{\mu_2 \mu_3^2 (2\mu_1 + \mu_2 + \mu_3)} P_{1,0,0}$$

$$P_{1,1,0} = \frac{\mu_1 \left((\mu_1 + \mu_2)^2 + \mu_2 \mu_3 \right)}{\mu_2 \mu_3 (2\mu_1 + \mu_2 + \mu_3)} P_{1,0,0}$$

$$P_{1b,1,0} = \frac{\mu_1^2 \left((\mu_1 + \mu_3)^3 + \mu_2 \left(2\mu_1^2 + 3\mu_1 \mu_3 + 2\mu_3^2 + \mu_1 \mu_3 \right) \right)}{\mu_2^2 \mu_3 (2\mu_1 + \mu_2 + \mu_3)(\mu_2 + \mu_3)} P_{1,0,0}$$

$$P_{1b,1,1} = \frac{\mu_1^3 (\mu_1 + \mu_3)(\mu_1 + \mu_2 + \mu_3)}{\mu_2 \mu_3^2 (2\mu_1 + \mu_2 + \mu_3)(\mu_2 + \mu_3)} P_{1,0,0}$$

$$P_{1b,1b,1} = \frac{\mu_1^3 (\mu_1 + \mu_2 + \mu_3)(\mu_1 + \mu_2 + 2\mu_3)}{\mu_3^3 (2\mu_1 + \mu_2 + \mu_3)(\mu_2 + \mu_3)} P_{1,0,0}$$

$$P_{1,1b,1} = \frac{\mu_1^2 (\mu_1 + \mu_2 + \mu_3)}{\mu_3^2 (2\mu_1 + \mu_2 + \mu_3)} P_{1,0,0} \tag{3.47}$$

The following value for $P_{1,0,0}$ can be found by summing the probabilities;

$$
\begin{aligned}
P_{1,0,0} &= \frac{N}{D} \\
N &= \mu_2^2\mu_3^3\left(\mu_2+\mu_3\right)\left(2\mu_1+\mu_2+\mu_3\right) \\
D &= \mu_1^5\left(\mu_2^2+\mu_2\mu_3+\mu_3^2\right) \\
&\quad +\mu_1^4\left(2\mu_2^3+5\mu_2^2\mu_3+5\mu_2\mu_3^2+3\mu_3^3\right) \\
&\quad +\mu_1^3\left(\mu_2^4+5\mu_2^3\mu_3+8\mu_2^2\mu_3^2+7\mu_2\mu_3^3+3\mu_3^4\right) \\
&\quad +\mu_1^2\left(\mu_2^4\mu_3+5\mu_2^3\mu_3^2+8\mu_2^2\mu_3^3+5\mu_2\mu_3^4+\mu_3^5\right) \\
&\quad +\mu_1\left(\mu_2^4\mu_3^2+5\mu_2^3\mu_3^3+5\mu_2^2\mu_3^4+\mu_2\mu_3^5\right) \\
&\quad +\mu_2^4\mu_3^3+2\mu_2^3\mu_3^4+\mu_2^2\mu_3^5
\end{aligned}
\tag{3.48}
$$

This value of $P_{1,0,0}$ can be substituted into (3.48) to find all the probabilities for this system in terms of λ, μ_1 and μ_2.

3.2.3 Maximum Utilisation

To find the maximum arrival rate, $\lambda_{\max}$, as with the previous system, it is not possible to use the probabilities in (3.48) as they do not give the proportion of customers that are blocked but the amount of time that the system spends in a blocked state. To work out the maximum arrival rate, the probability that a customer can become blocked in Node 1 is required. Two probabilities are required for this. The probability that Node 1 completes its service before Node 2 completes its service, is as in the previous system with two nodes. The second probability required is when the customer at Node 1 is blocked and the customer in Node 2 is also blocked because both their services have ended before the service in Node 3 is completed. So the required formula for the time spent in Node 1 will be the average time spent for a service in Node 1 added to the probability that Node 1 completes service before Node 2 multiplied by the average service time in Node 2 then this is added to the probability that Nodes 1 and 2 complete their service before Node 3 completes its service. T_1 is defined to be the time Node 1 takes to complete a service, T_2 to be the time Node 2 takes to complete a service and T_3 the time Node 3 takes to complete a service. Then the formula becomes;

$$\frac{1}{\mu_1}+P(T_1<T_2)\frac{1}{\mu_2}+P(T_1<T_3)P(T_2<T_3)\frac{1}{\mu_3} \qquad (3.49)$$

This is somewhat more complicated than for the two node system as not all nodes start their services simultaneously. Below is map of possible routes this system could take.

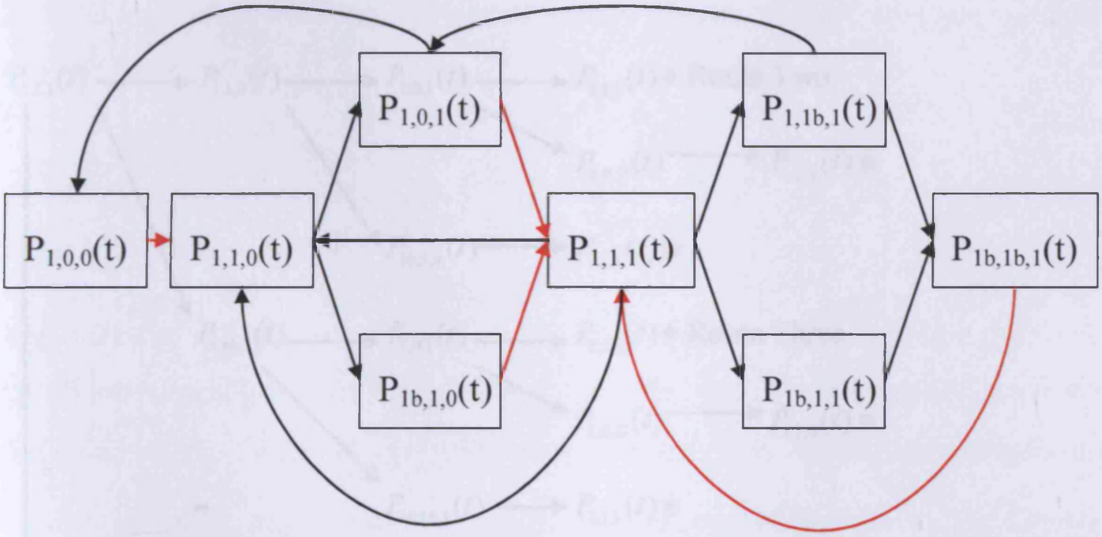


Figure 3.22 - Route map for three node `drip feed' network.

The red arrows on the diagram indicate when new customers arrive into the system. Outlined below are which routes have customers who start simultaneously at each node.

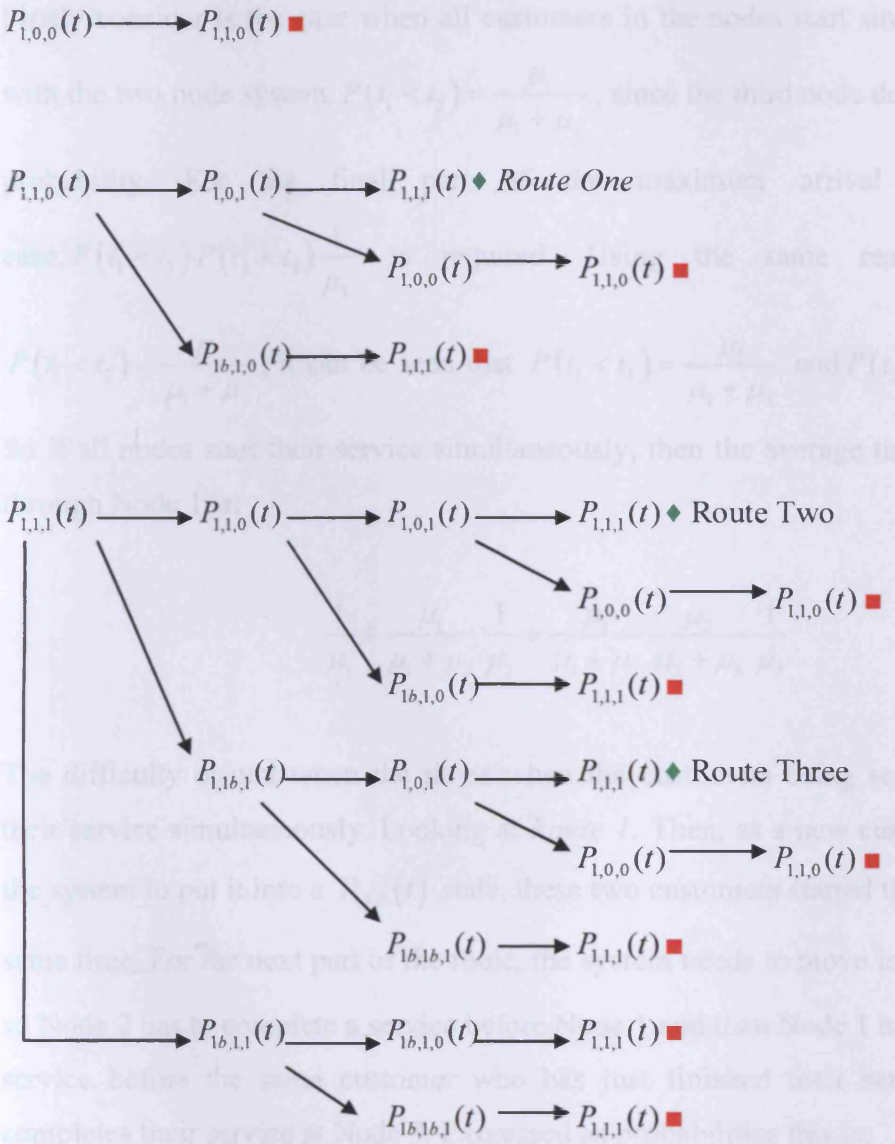

$P_{1,0,0}(t)$ → $P_{1,1,0}(t)$ ■
$P_{1,1,0}(t)$ → $P_{1,0,1}(t)$ → $P_{1,1,1}(t)$ ♦ Route One
$P_{1,0,0}(t)$ → $P_{1,1,0}(t)$ ■
$P_{1b,1,0}(t)$ → $P_{1,1,1}(t)$ ■
$P_{1,1,1}(t)$ → $P_{1,1,0}(t)$ → $P_{1,0,1}(t)$ → $P_{1,1,1}(t)$ ♦ Route Two
$P_{1,0,0}(t)$ → $P_{1,1,0}(t)$ ■
$P_{1b,1,0}(t)$ → $P_{1,1,1}(t)$ ■
$P_{1,1b,1}(t)$ → $P_{1,0,1}(t)$ → $P_{1,1,1}(t)$ ♦ Route Three
$P_{1,0,0}(t)$ → $P_{1,1,0}(t)$ ■
$P_{1b,1b,1}(t)$ → $P_{1,1,1}(t)$ ■
$P_{1b,1,1}(t)$ → $P_{1b,1,0}(t)$ → $P_{1,1,1}(t)$ ■
$P_{1b,1b,1}(t)$ → $P_{1,1,1}(t)$ ■

Figure 3.23 - Route maps showing possible route until next arrival.

These are the complete set of routes that end with a new arrival. All the routes that end with a red square are ones when all the nodes start simultaneously when a customer has just joined the system. All the routes that end in a green diamond are routes where a new customer has arrived in the system but these nodes do not start their service simultaneously. These are called routes 1, 2 and 3. As can be seen the only state in which the system, that has just had an arrival, has all the customers not starting simultaneously is when all the nodes are serving a customer.

Firstly consider is the case when all customers in the nodes start simultaneously. As with the two node system, $P(t_1 < t_2) = \frac{\mu_1}{\mu_1 + \mu_2}$, since the third node does not affect the probability. For the final part of the maximum arrival rate, in this case, $P(t_1 < t_3)P(t_2 < t_3)\frac{1}{\mu_3}$ is required. Using the same reasoning as for $P(t_1 < t_2) = \frac{\mu_1}{\mu_1 + \mu_2}$, it can be seen that $P(t_1 < t_3) = \frac{\mu_1}{\mu_1 + \mu_3}$ and $P(t_2 < t_3) = \frac{\mu_2}{\mu_2 + \mu_3}$. So if all nodes start their service simultaneously, then the average time taken to pass through Node 1 is;

$$\frac{1}{\mu_1} + \frac{\mu_1}{\mu_1 + \mu_2}\frac{1}{\mu_2} + \frac{\mu_1}{\mu_1 + \mu_3}\frac{\mu_2}{\mu_2 + \mu_3}\frac{1}{\mu_3} \tag{3.50}$$

The difficulty occurs when the times when the customers being served do not start their service simultaneously. Looking at *Route 1*. Then, as a new customer has joined the system to put it into a $P_{1,1,0}(t)$ state, these two customers started their service at the same time. For the next part of the route, the system needs to move into a $P_{1,0,1}(t)$ state so Node 2 has to complete a service before Node 1 and then Node 1 has to complete its service before the same customer who has just finished their service in Node 2, completes their service at Node 3. Expressed as probabilities this is;

$$P(t_2 < t_1)P(t_1 < t_2 + t_3 \mid t_1 > t_2) \tag{3.51}$$

The first part of (3.51) is similar to the probabilities that have already been calculated, as $P(t_2 < t_1) = \frac{\mu_2}{\mu_1 + \mu_2}$. For the second part, more probability theory is required;

$$P(t_1 < t_2 + t_3 \mid t_1 > t_2) = \frac{P(t_2 < t_1 < t_2 + t_3)}{P(t_1 > t_2)} \tag{3.52}$$

If the two probabilities are multiplied together;

$$P(t_1 > t_2)\frac{P(t_2 < t_1 < t_2 + t_3)}{P(t_1 > t_2)} = P(t_2 < t_1 < t_2 + t_3) \qquad (3.53)$$

From Bayesian theory.

$$P(T_2 < T_1 < T_2 + T_3) = P(0 < T_1 - T_2 < T_3) = P(0 < T_1 - T_2 < t_3 \mid T_3 = t_3)P(T_3 = t_3) \qquad (3.54)$$

Considering $P(0 < T_1 - T_2 < t_3)$ the area to be integrated over is;

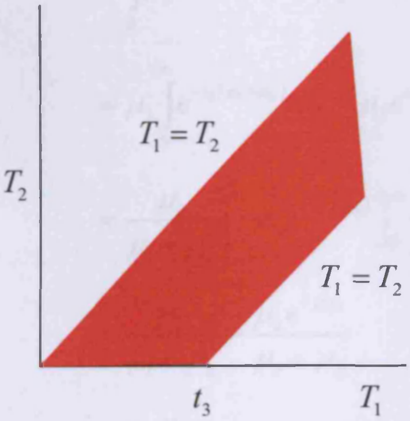


Figure 3.24 – Area to integrate over

$$P\left(0<T_1-T_2<t_3\,|\,T_3=t_3\right)=\int_{T_2=0}^{\infty}\int_{T_1=t_2}^{t_2+t_3}\mu_1 e^{-\mu_1 t_1}\mu_2 e^{-\mu_2 t_2}\,dt_1 dt_2$$

$$=\int_0^{\infty}\mu_2 e^{-\mu_2 t_2}\int_{t_2}^{t_2+t_3}\mu_1 e^{-\mu_1 t_1}\,dt_1 dt_2$$

$$=\int_0^{\infty}\mu_2 e^{-\mu_2 t_2}\left[-e^{-\mu_1 t_1}\right]_{t_2}^{t_2+t_3}dt_2$$

$$=\int_0^{\infty}\mu_2 e^{-\mu_2 t_2}\left[e^{-\mu_1 t_2}-e^{-\mu_1(t_2+t_3)}\right]dt_2$$

$$=\int_0^{\infty}\mu_2 e^{-t_2(\mu_1+\mu_2)}-\mu_2 e^{-t_2(\mu_1+\mu_2)}e^{-\mu_1 t_3}dt_2$$

$$=\mu_2\int_0^{\infty}e^{-t_2(\mu_1+\mu_2)}dt_2-\mu_2 e^{-\mu_1 t_3}\int_0^{\infty}e^{-t_3(\mu_1+\mu_2)}dt_2$$

$$=\frac{\mu_2}{\mu_1+\mu_2}\left[-e^{-t_2(\mu_1+\mu_2)}\right]_0^{\infty}-\frac{\mu_2 e^{-\mu_1 t_3}}{\mu_1+\mu_2}\left[-e^{-t_2(\mu_1+\mu_2)}\right]_0^{\infty}$$

$$=\frac{\mu_2}{\mu_1+\mu_2}-\frac{\mu_2 e^{-\mu_1 t_3}}{\mu_1+\mu_2} \tag{3.55}$$

$$=\frac{\mu_2}{\mu_1+\mu_2}\left(1-e^{-\mu_1 t_3}\right)$$

From this it can be seen.

$$\begin{aligned}
&P\left(0<T_1-T_2<T_3\right)=P\left(0<T_1-T_2<t_3 \mid T_3=t_3\right)P\left(T_3=t_3\right)\\
&=\int_0^{\infty}\frac{\mu_2}{\mu_1+\mu_2}\left(1-e^{-\mu_1 t_3}\right)\mu_3 e^{-\mu_3 t_3}dt_3\\
&=\frac{\mu_1}{\mu_1+\mu_2}\left[\left[-e^{-\mu_3 t_3}\right]_0^{\infty}-\mu_3\int_0^{\infty}e^{-t_3(\mu_1+\mu_3)}dt_3\right]\\
&=\frac{\mu_2}{\mu_1+\mu_2}-\frac{\mu_3\mu_2}{\mu_1+\mu_2}\left[\frac{-e^{-t_3}}{\mu_1+\mu_3}\right]_0^{\infty}\\
&=\frac{\mu_2}{\mu_1+\mu_2}-\frac{\mu_2\mu_3}{\left(\mu_1+\mu_2\right)\left(\mu_1+\mu_3\right)}\\
&=\frac{\mu_2}{\mu_1+\mu_2}\left(1-\frac{\mu_3}{\mu_1+\mu_3}\right)\\
&=\frac{\mu_2}{\mu_1+\mu_2}\frac{\mu_1}{\mu_1+\mu_3}\\
&=\frac{\mu_1\mu_2}{\left(\mu_1+\mu_2\right)\left(\mu_1+\mu_3\right)}
\end{aligned}\qquad(3.56)$$

Once this calculation was completed it was noted that is was not necessary to compute as the probability is equal to;

$$\frac{\mu_2}{\mu_1+\mu_2}\frac{\mu_1}{\mu_1+\mu_3}=P\left(t_2<t_1\right)P\left(t_1<t_3\right)\qquad(3.57)$$

where t_i is the service time at node i. So the probability of Route $1, P_{1,1,0}(t)\rightarrow P_{1,0,1}(t)\rightarrow P_{1,1,1}(t)$, is just the probability that Node 2's service ends before Node 1's service is completed, multiplied by the probability that Node 1 's service time is less than Node 3's service time. Even though the customer in that starts this route in Node 2 goes through two services. This is due to the memoryless property of Exponential distribution.

This probability then needs to be multiplied by the probability that States 1 and 2are both occupied and serving and the final node is unoccupied. This is different to $P_{1,1,0}$, which is the probability that the system is in that state, i.e. the proportion of time that the system spends in that state. This is the same distinction as with the two node case and can be difficult to compute with many different combinations for each probability. A new method would be preferable.

If instead of attempting to work out these probabilities, we take the time that the initial node is working, not just occupied as it is occupied all the time in a drip feed model, and divide this by the total time Node 1 is occupied in this drip feed system. This will be equal to the rate which the first node can process customers in a non drip feed system. This is because in the drip system the initial node can only be working or blocked. In a situation with an infinite initial queue, if customers are allowed to arrive at a faster average rate than the initial node can serve then a queue will build up and the system would not be in a steady state. So the maximum utilisation rate of the first node is;

$$\rho_{\max} = \frac{\sum \text{Time First Node Working}}{\sum \text{Total Time}} \tag{3.58}$$

Using the notation for the proportion of time spent in each node $\rho_{\max}$ becomes;

$$\rho_{\max} = \frac{P_{1,0,0}(t)+P_{1,1,0}(t)+P_{1,0,1}(t)+P_{1,1,1}(t)+P_{1,1b,1}(t)}{P_{1,0,0}(t)+P_{1,1,0}(t)+P_{1,0,1}(t)+P_{1,1,1}(t)+P_{1,1b,1}(t)+P_{1b,1,1}(t)+P_{1b,1b,1}(t)+P_{1b,1,0}(t)} \tag{3.59}$$

Since all probabilities sum to 1 the denominator of the above sums to 1.

$$\rho_{\max} = P_{1,0,0}(t)+P_{1,1,0}(t)+P_{1,0,1}(t)+P_{1,1,1}(t)+P_{1,1b,1}(t) \tag{3.60}$$

These can be found in terms of $P_{1,0,0}(t)$

$$
\begin{aligned}
\rho_{\max} &= P_{1,0,0}(t)\left[\begin{array}{l} 1+\dfrac{\left((\mu_1+\mu_3)^2+\mu_2\mu_3\right)\mu_1}{\mu_2\mu_3(2\mu_1+\mu_2+\mu_3)} \\ +\dfrac{\mu_1}{\mu_3}+\dfrac{\mu_1^2(\mu_1+\mu_3)(\mu_1+\mu_2+\mu_3)}{\mu_2\mu_3^2(2\mu_1+\mu_2+\mu_3)} \\ +\dfrac{(\mu_1+\mu_2+\mu_3)\mu_1^2}{(2\mu_1+\mu_2+\mu_3)\mu_3^2} \end{array}\right] \\
&= P_{1,0,0}(t)\left[\dfrac{\begin{array}{l}\mu_1^4+\mu_1^3(2\mu_2+3\mu_3)+\mu_1^2\left(\mu_2^2+4\mu_2\mu_3+3\mu_3^2\right) \\ \mu_1\left(\mu_2^2\mu_3+4\mu_2\mu_3^2+\mu_3^3\right)+\mu_2\mu_3^2(\mu_2+\mu_3)\end{array}}{\mu_2\mu_3^2(2\mu_1+\mu_2+\mu_3)}\right] \\
&= P_{1,0,0}(t)\frac{N_1}{D_1} \qquad\qquad (3.61) \\
&= \frac{N}{D}\frac{N_1}{D_1} \\
&= \frac{\mu_2\mu_3(\mu_2+\mu_3)N_1}{D}
\end{aligned}
$$

where the values for N and D are the same as in Equation (3.48) and $N_1=\mu_1^4+\mu_1^3(2\mu_2+3\mu_3)+\mu_1^2\left(\mu_2^2+4\mu_2\mu_3+3\mu_3^2\right)+\mu_1\left(\mu_2^2\mu_3+4\mu_2\mu_3^2+\mu_3^3\right)+\mu_2\mu_3^2(\mu_2+\mu_3)$ and $D_1=\mu_2\mu_3^2(2\mu_1+\mu_2+\mu_3)$. This method gives the same result as (Hunt 1956) using a different method

This method can also be used for finding the value of $\rho_{\max}$ for the two node case;

$$
\begin{aligned}
\rho_{\max} &= \frac{P_{1,0}+P_{1,1}}{P_{1,0}+P_{1,1}+P_{1b,1}} \\
&= \frac{\mu_2^2}{\mu_1^2+\mu_1\mu_2+\mu_2^2}+\frac{\mu_1\mu_2}{\mu_1^2+\mu_1\mu_2+\mu_2^2} \qquad\qquad (3.62) \\
&= \frac{\mu_2(\mu_2+\mu_1)}{\mu_1^2+\mu_1\mu_2+\mu_2^2}
\end{aligned}
$$

This value for $\rho_{\max}$ is the same as previously derived, equation (3.27).

This simple method can now be used to compute the maximum utilisation rate for a tandem queueing systèm. This method will be used in future chapters when extra routes are added to the system.

Chapter 4

PHASE TYPE DISTRIBUTIONS

4.1 Introduction

Phase Type probability distributions are formed by a number of Poisson processes in sequence ending in an absorbing phase. This type of probability distribution is very flexible, as the number of different phases and the opportunities to join these phases by any number of pathways are available. This flexibility is very useful but leads to difficulties in creating a theoretical system. A more practical distribution that is used is the Coxian Phase Type, which is also very flexible, but it is simpler to create a theoretical distribution. The Coxian distribution has a fixed number of Poisson processes in a sequential order. When one process comes to an end either the next phase in the sequence starts or the customer being served goes into the absorbing phase and the service ends. This distribution is quite similar to the Hyper-Exponential distribution, as it also has a fixed number of Poisson processes in a sequence, but in this case the absorbing phase can only be accessed once the customer has passed through all of the processes. In the Coxian Phase Type distribution the absorbing phase can be accessed from all phases.

The theoretical Coxian Phase Type distribution can be set up in several ways, as seen in (Cox 1954). Two of these will now be highlighted. First of all, the time spent in any phase is set to come from a single Poisson process. Once the service in Phase i is over, the customer moves to Phase j with probability α_i, or to the absorbing phase with probability $1-\alpha_i$. Another option is to have two Poisson processes happening simultaneously in each phase. One of these processes is attached to the absorbing phase whilst the other is used for the next phase in the sequence. Figure 4.1 shows a Coxian Phase Type distribution with two phases with the second of these two methods. This piece of work will use the second method; the majority of the papers that use Coxian Phase Type distributions in the medical field also use this system.

Phase Type distributions have been used widely for modelling hospital environments, their flexible nature make them very suited to the task. The phases can be linked to actual areas or more commonly phases that correspond to the health or status of the patient rather than the physical phase that they are in. In this chapter Phase Type equations are set up and solved, using two different methods. As this type of distribution has a history of successfully modelling medical situations, though not specifically in bed blocking, once the equations have been set up they will be used to model the same bed blocking situation that the blocking equations will be used for. The two methods can then be compared.

4.2 Two Phases

A two phase Coxian Phase Type system is illustrated in Figure 4.1 the two phases. Both phases have a Negative Exponential service time distributions with parameters as indicated.

$$f_1(t) = (\lambda + \mu_1)e^{-(\lambda+\mu_1)t} \qquad \text{Phase 1}$$

$$f_2(t) = \mu_2 e^{-\mu_2 t} \qquad \text{Phase 2}$$

where λ is the average rate at which customers move from Phase 1 to Phase 2, μ_i average rate which customers leave from Phase i to enter the Exit Phase where i=1,2.

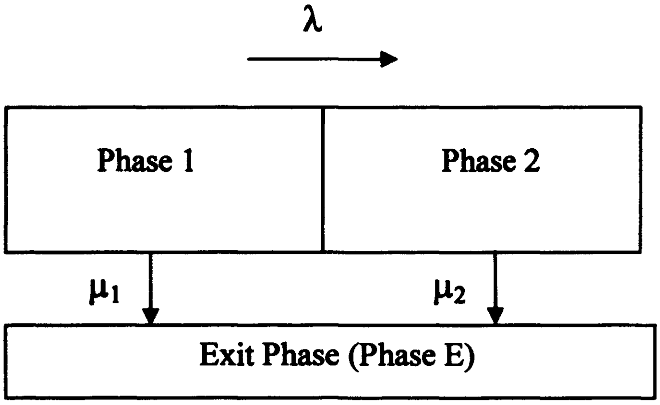


Figure 4.1 – Two phase Coxian Phase Type system

At time $t=0$, a single customer is introduced into Phase 1; at this time both the second phase and the Exit Phase are empty. The probabilities of a customer moving from Phase *i* to *j* in the time interval $(t,t+\delta t)$, where δt is a small interval of time is;

Phase i	Phase j	Probability
1	2	$\lambda\delta t+o(\delta t)^2$
1	E	$\mu_1\delta t+o(\delta t)^2$
2	E	$\mu_2\delta t+o(\delta t)^2$

The probability a customer will go to the Exit Phase from Phase 1 is;

$$\frac{\mu_1}{\lambda+\mu_1}$$

The probability a customer will go to Phase 2 from Phase 1 is;

$$\frac{\lambda}{\lambda+\mu_1}$$

From these results the average length of stay before joining the Exit Phase can be calculated;

$$\begin{aligned}&\frac{1}{\lambda+\mu_1}+\frac{1}{\mu_2}\frac{\lambda}{\lambda+\mu_1}\\&=\frac{\lambda+\mu_2}{\mu_2(\lambda+\mu_1)}\end{aligned} \tag{4.1}$$

Using the notation,

$P_n(t)$ Probability that a customer is in Phase n at time t. $n = 1,2$.

$P_E(t)$ Probability that a customer is in the Exit Phase at time t.

Using these, the following equations can be set up.

$$\begin{aligned}&P_1(t+\delta t)=P_1(t)(1-\lambda\delta t)(1-\mu_1\delta t)\\&\frac{dP_1(t)}{dt}=-(\lambda+\mu_1)P_1(t)\\&\int\frac{dP_1(t)}{dt}\frac{1}{P_1(t)}dt=-\int(\lambda+\mu_1)dt\\&\ln P_1(t)=-(\lambda+\mu_1)t+C,\quad \text{where } C \text{ is an arbitary constant.}\\&P_1(t)=Ce^{-(\lambda+\mu_1)t}\end{aligned} \tag{4.2}$$

To find C, the fact that customers always start in Phase 1 is used. So at time zero the probability of the customer being in Phase 1 is 1.

$$\begin{aligned}&P_1(0)=1\\&1=C\\&\text{So,}\quad P_1(t)=e^{-(\lambda+\mu_1)t}\end{aligned} \tag{4.3}$$

Now the probability that Phase 2 is occupied at time t can be found;

$$P_2(t+\delta t)=P_2(t)(1-\mu_2\delta t)+P_1(t)\lambda\delta t(1-\mu_1\delta t)$$
$$\frac{dP_2(t)}{dt}+\mu_2P_2(t)=\lambda P_1(t) \tag{4.4}$$

This differential equation can be solved using the integrating factor method; in this case the integrating factor. is $e^{\mu_2 t}$.

$$e^{\mu_2 t}P_2(t)=\int\lambda e^{(\mu_2-\lambda-\mu_1)t}dt$$
$$e^{\mu_2 t}P_2(t)=\frac{\lambda}{\mu_2-\lambda-\mu_1}e^{(\mu_2-\lambda-\mu_1)t}+C,\quad \text{where C is an arbitary constant.} \tag{4.5}$$

In this case we use the fact that in the initial state the probability of being in Phase 2 is zero.

$$P_2(0)=0$$
$$0=\frac{\lambda}{\mu_2-\lambda-\mu_1}+C$$
$$C=-\frac{\lambda}{\mu_2-\lambda-\mu_1} \tag{4.6}$$
$$P_2(t)=\frac{\lambda}{\mu_2-\lambda-\mu_1}\left(e^{-(\lambda+\mu_1)t}-e^{-\mu_2 t}\right)$$

Next the probability of being in the Exit Phase at time t is found;

$$P_E(t+\delta t)=P_2(t)\mu_2\delta t+(1-\lambda\delta t)\mu_1\delta tP_1(t)+P_E(t)$$

$$\frac{dP_E(t)}{dt}=\mu_2P_2(t)+\mu_1P_1(t)$$

$$\frac{dP_E(t)}{dt}=\frac{\mu_2\lambda}{\mu_2-\lambda-\mu_1}\left(e^{-(\lambda+\mu_1)t}-e^{-\mu_2t}\right)+\mu_1e^{-(\lambda+\mu_1)t}$$

$$P_E(t)=\int\left\{\frac{\lambda\mu_2+\mu_1\mu_2-\mu_1^2-\lambda\mu_1}{(\mu_2-\lambda-\mu_1)}e^{-(\lambda+\mu_1)t}-\frac{\mu_2\lambda}{\mu_2-\lambda-\mu_1}e^{-\mu_2t}\right\}dt$$

$$P_E(t)=\frac{\mu_1^2+\lambda\mu_1-\lambda\mu_2-\mu_1\mu_2}{(\lambda+\mu_1)(\mu_2-\lambda-\mu_1)}e^{-(\lambda+\mu_1)t}+\frac{\lambda\mu_2}{(\mu_2-\lambda-\mu_1)\mu_2}e^{-\mu_2t}+C,\ \text{where C is an arbitary cc}$$

$$P_E(t)=\frac{(\lambda+\mu_1)(\mu_1-\mu_2)}{(\lambda+\mu_1)(\mu_2-\lambda-\mu_1)}e^{-(\lambda+\mu_1)t}+\frac{\lambda}{\mu_2-\lambda-\mu_1}e^{-\mu_2t}+C$$

$$P_E(t)=\frac{\mu_1-\mu_2}{\mu_2-\lambda-\mu_1}e^{-(\lambda+\mu_1)t}+\frac{\lambda}{\mu_2-\lambda-\mu_1}e^{-\mu_2t}+C \tag{4.7}$$

Using the initial condition that the Exit Phase was empty C can be calculated;

$$P_E(0)=0,\ \text{so}$$

$$C=\frac{\mu_2-\mu_1-\lambda}{\mu_2-\lambda-\mu_1}$$

$$C=1 \tag{4.8}$$

$$P_E(t)=\frac{\mu_1-\mu_2}{\mu_2-\lambda-\mu_1}e^{-(\lambda+\mu_1)t}+\frac{\lambda}{\mu_2-\lambda-\mu_1}e^{-\mu_2t}+1$$

It may be checked that the probabilities sum to 1.

$$\begin{aligned}\sum_{i=1}^{3} P_i(t) &= e^{-(\lambda+\mu_1)t}\left(\frac{\mu_1-\mu_2}{\mu_2-\lambda-\mu_1}+\frac{\lambda}{\mu_2-\lambda-\mu_1}+1\right)\\ &+e^{-\mu_2 t}\left(-\frac{\lambda}{\mu_2-\lambda-\mu_1}+\frac{\lambda}{\mu_2-\lambda-\mu_1}\right)+1 \qquad (4.9)\\ &=1\end{aligned}$$

4.3 Three Phases

Now an extra phase will be added to the distribution. As before phases have Negative Exponential service times with parameters as shown below;

$$\begin{aligned}f_1(t) &= (\lambda_1+\mu_1)e^{-(\lambda_1+\mu_1)t}\\ f_2(t) &= (\lambda_2+\mu_2)e^{-(\lambda_2+\mu_2)t} \qquad (4.10)\\ f_3(t) &= \mu_3 e^{-\mu_3 t}\end{aligned}$$

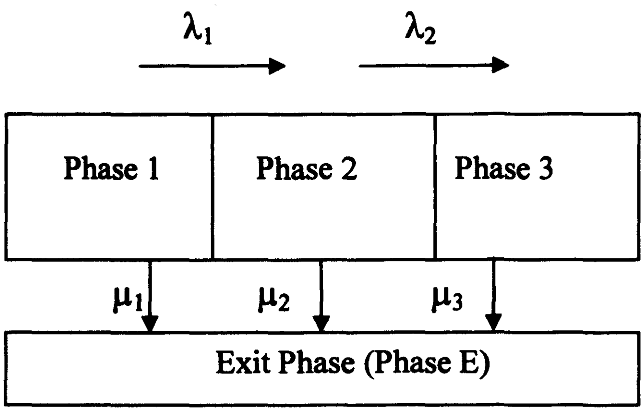


Figure 4.2 – A three phase Coxian Phase Type system

The probability that a customer will go to the Exit Phase from Phase 1 is;

$$\frac{\mu_1}{\lambda + \mu_1} \tag{4.11}$$

The probability that a customer will join the Exit Phase from Phase 2 is;

$$\frac{\lambda_1}{\lambda_1 + \mu_1} \frac{\mu_2}{\lambda_2 + \mu_2} \tag{4.12}$$

The probability that a customer will go to the Exit Phase from Phase 3 is;

$$\frac{\lambda_1}{\lambda_1 + \mu_1} \frac{\lambda_2}{\lambda_2 + \mu_2} \tag{4.13}$$

Using these probabilities an expression for the average length of stay before the customer joins the Exit Phase can be found;

$$\frac{1}{\lambda_1 + \mu_1} + \left(\frac{\lambda_1}{\lambda_1 + \mu_1}\right)\frac{1}{\lambda_2 + \mu_2} + \left(\frac{\lambda_1 \lambda_2}{(\lambda_1 + \mu_1)(\lambda_2 + \mu_2)}\right)\frac{1}{\mu_3} \tag{4.14}$$

Next, as with the two phase situation, equations relating the probabilities will be set up. Phase 1 is independent of the other phases so:

$$\begin{gathered} P_1(t + \delta t) = P_1(t)(1 - \mu_1 \delta t)(1 - \lambda_1 \delta t) \\ \frac{dP_1(t)}{dt} = -(\lambda_1 + \mu_1)P_1(t) \end{gathered} \tag{4.15}$$

This is the same equation as for the two phase case, so we have:

$$P_1(t) = e^{-(\lambda_1+\mu_1)t} \tag{4.16}$$

The subsequent phases are dependent on previous phases.

$$\begin{aligned}
P_2(t+\delta t) &= P_2(t)(1-\mu_2\delta t)(1-\lambda_2\delta t) \\
&\quad +P_1(t)\lambda_1\delta t \\
P_3(t+\delta t) &= P_3(t)(1-\mu_3\delta t) \\
&\quad +P_2(t)\lambda_2\delta t \\
P_E(t+\delta t) &= P_E(t) \\
&\quad +P_3(t)\mu_3\delta t \\
&\quad +P_2(t)\mu_2\delta t \\
&\quad +P_1(t)\mu_1\delta t
\end{aligned} \tag{4.17}$$

Taking Phase 2;

$$\frac{dP_2(t)}{dt} + (\lambda_2+\mu_2)P_2(t) = \lambda_1 P_1(t) \tag{4.18}$$

By using the integrating factor method on equation (4.18) ;

$$\begin{aligned}
e^{(\lambda_2+\mu_2)t}P_2(t) &= \lambda_1 \int e^{(\lambda_2+\mu_2)t} P_1(t)\,dt \\
&= \lambda_1 \int e^{(\lambda_2+\mu_2-\lambda_1-\mu_1)t}\,dt \\
&= \frac{\lambda_1}{\lambda_2+\mu_2-\lambda_1-\mu_1} e^{(\lambda_2+\mu_2-\lambda_1-\mu_1)t} + C, \quad \text{where C is an arbitary constant}
\end{aligned} \tag{4.19}$$

Using the fact that, in the initial system, Phase 2 is empty an explicit expression for $P_2(t)$ can be found;

$$\begin{aligned}
P_2(0) &= 0 \\
C &= -\frac{\lambda_1}{\lambda_2+\mu_2-\lambda_1-\mu_1} \\
e^{(\lambda_2+\mu_2)t}P_2(t) &= \frac{\lambda_1}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(e^{(\lambda_2+\mu_2-\lambda_1-\mu_1)t}-1\right) \\
P_2(t) &= \frac{\lambda_1}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(e^{-(\lambda_1+\mu_1)t}-e^{-(\lambda_2+\mu_2)t}\right)
\end{aligned} \tag{4.20}$$

This may be expressed in the following form

$$P_2(t)=\lambda_1\left(\frac{e^{-(\lambda_1+\mu_1)t}}{\lambda_2+\mu_2-\lambda_1-\mu_1}+\frac{e^{-(\lambda_2+\mu_2)t}}{\lambda_1+\mu_1-\lambda_2-\mu_2}\right) \tag{4.21}$$

Next looking at Phase 3;

$$\begin{aligned}
\frac{dP_3(t)}{dt} &= -\mu_3P_3(t)+\lambda_2P_2(t) \\
\frac{dP_3(t)}{dt}+\mu_3P_3(t) &= \lambda_2P_2(t)
\end{aligned} \tag{4.22}$$

Once again, by using the integrating factor method this equation can be solved;

$$\begin{aligned}
e^{\mu_3t}P_3(t) &= \frac{\lambda_1\lambda_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\int\left\{e^{(\mu_3-\lambda_1-\mu_1)t}-e^{(\mu_3-\lambda_2-\mu_2)t}\right\}dt \\
&= \frac{\lambda_1\lambda_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\frac{e^{(\mu_3-\lambda_1-\mu_1)t}}{\mu_3-\lambda_1-\mu_1}-\frac{e^{(\mu_3-\lambda_2-\mu_2)t}}{\mu_3-\lambda_2-\mu_2}\right)+C
\end{aligned} \tag{4.23}$$

Using the fact that the third phase is empty initially;

$$P_3(0)=0$$

$$C=-\frac{\lambda_1\lambda_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\frac{1}{\mu_3-\lambda_1-\mu_1}-\frac{1}{\mu_3-\lambda_2-\mu_2}\right)$$

$$e^{\mu_3 t}P_3(t)=\frac{\lambda_1\lambda_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\frac{e^{(\mu_3-\lambda_1-\mu_1)t}-1}{\mu_3-\lambda_1-\mu_1}-\frac{e^{(\mu_3-\lambda_2-\mu_2)t}-1}{\mu_3-\lambda_2-\mu_2}\right)$$

$$P_3(t)=\frac{\lambda_1\lambda_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\frac{e^{-(\lambda_1+\mu_1)t}-e^{-\mu_3 t}}{\mu_3-\lambda_1-\mu_1}-\frac{e^{-(\lambda_2+\mu_2)t}-e^{-\mu_3 t}}{\mu_3-\lambda_2-\mu_2}\right)$$

$$P_3(t)=\frac{\lambda_1\lambda_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\frac{e^{-(\lambda_1+\mu_1)t}}{\mu_3-\lambda_1-\mu_1}-\frac{e^{-\lambda_2+\mu_2 t}}{\mu_3-\lambda_2-\mu_2}+\frac{(\lambda_2+\mu_2-\lambda_1-\mu_1)e^{-\mu_3 t}}{(\mu_3-\lambda_1-\mu_1)(\mu_3-\lambda_2-\mu_2)}\right)$$

(4.24)

Note that this maybe expressed as;

$$\lambda_1\lambda_2\left(\begin{array}{r}\dfrac{e^{-(\lambda_1+\mu_1)t}}{(\mu_3-\lambda_1-\mu_1)(\lambda_2+\mu_2-\lambda_1-\mu_1)}+\dfrac{e^{-(\lambda_2+\mu_2)t}}{(\mu_3-\lambda_2-\mu_2)(\lambda_1+\mu_1-\lambda_2-\mu_2)}\\ +\dfrac{e^{-\mu_3 t}}{(\lambda_1+\mu_1-\mu_3)(\lambda_2+\mu_2-\mu_3)}\end{array}\right)$$

(4.25)

The Exit Phase probability may be expressed as follows;

$$\frac{dP_E(t)}{dt}=\mu_3P_3(t)+\mu_2P_2(t)+\mu_1P_1(t)$$

$$P_E(t)=\frac{\lambda_1\lambda_2\mu_3}{\lambda_2+\mu_2-\lambda_1-\mu_1}\int\left\{\frac{e^{-(\lambda_1+\mu_1)t}-e^{-\mu_3 t}}{\mu_3-\lambda_1-\mu_1}-\frac{e^{-(\lambda_2+\mu_2)t}-e^{-\mu_3 t}}{\mu_3-\lambda_2-\mu_2}\right\}dt$$

$$+\frac{\lambda_1\mu_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\int\left\{e^{-(\lambda_1+\mu_1)t}-e^{-(\lambda_2-\mu_2)t}\right\}dt$$

$$+\mu_1\int e^{-(\lambda_1+\mu_1)t}dt$$

$$=\frac{\lambda_1\lambda_2\mu_3}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\begin{array}{r}-\frac{e^{-(\lambda_1+\mu_1)t}}{(\lambda_1+\mu_1)(\mu_3-\lambda_1-\mu_1)}+\frac{e^{-\mu_e t}}{\mu_3(\mu_3-\lambda_1-\mu_1)}+\frac{e^{-(\lambda_2+\mu_2)t}}{(\lambda_2+\mu_2)(\mu_3-\lambda_2-\mu_2)}\\ -\frac{e^{-\mu_3 t}}{\mu_3(\mu_3-\lambda_2-\mu_2)}\end{array}\right)$$

$$+\frac{\lambda_1\mu_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(-\frac{e^{-(\lambda_1+\mu_1)t}}{(\lambda_1+\mu_1)}+\frac{e^{-(\lambda_2+\mu_2)t}}{(\lambda_2+\mu_2)}\right)$$

$$-\frac{\mu_1}{\lambda_1+\mu_1}e^{-(\lambda_1+\mu_1)t}+C,$$ where C is an arbitary constant.

(4.26)

The value of the constant is found using the initial conditions. In this case the probability of the Exit Phase being occupied at time zero is zero.

$$P_E(0)=0$$

$$
\begin{aligned}
C = & -\frac{\lambda_1\lambda_2\mu_3}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\begin{array}{l} -\dfrac{1}{(\lambda_1+\mu_1)(\mu_3-\lambda_1-\mu_1)}+\dfrac{1}{\mu_3(\mu_3-\lambda_1-\mu_1)} \\ +\dfrac{1}{(\lambda_2+\mu_2)(\mu_3-\lambda_2-\mu_2)}-\dfrac{1}{\mu_3(\mu_3-\lambda_2-\mu_2)} \end{array}\right) \\
& -\frac{\lambda_1\mu_2}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(-\frac{1}{(\lambda_1+\mu_1)}+\frac{1}{(\lambda_2+\mu_2)}\right) \\
& +\frac{\mu_1}{\lambda_1+\mu_1} \\
& =\frac{\mu_1}{\lambda_1+\mu_1}+\frac{\lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)}-\frac{\lambda_1\lambda_2\mu_3}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\begin{array}{l} \dfrac{1}{(\mu_3-\lambda_1-\mu_1)}\left(-\dfrac{1}{(\lambda_1+\mu_1)}+\dfrac{1}{\mu_3}\right) \\ +\dfrac{1}{(\mu_3-\lambda_2-\mu_2)}\left(\dfrac{1}{(\lambda_2+\mu_2)}-\dfrac{1}{\mu_3}\right) \end{array}\right) \\
& =\frac{\mu_1}{\lambda_1+\mu_1}+\frac{\lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)}-\frac{\lambda_1\lambda_2\mu_3}{\lambda_2+\mu_2-\lambda_1-\mu_1}\left(\frac{1}{\mu_3(\lambda_2+\mu_2)}-\frac{1}{\mu_3(\lambda_1+\mu_1)}\right) \\
& =\frac{\mu_1}{\lambda_1+\mu_1}+\frac{\lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)}+\frac{\lambda_1\lambda_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} \\
& =\frac{\lambda_2\mu_1+\mu_1\mu_2+\lambda_1\mu_2+\lambda_1\lambda_2}{\lambda_1\lambda_2+\lambda_1\mu_2+\lambda_2\mu_1+\mu_1\mu_2} \\
& =1
\end{aligned}
\tag{4.27}
$$

The final expression for $P_E(t)$ is thus:

$$
\begin{aligned}
P_E(t) = &-\mu_1\left(\frac{e^{-(\lambda_1+\mu_1)t}}{\lambda_1+\mu_1}\right) \\
&-\lambda_1\mu_2\left(\frac{e^{-(\lambda_1+\mu_1)t}}{(\lambda_1+\mu_1)(\lambda_2+\mu_1-\lambda_1-\mu_1)}+\frac{e^{-(\lambda_2+\mu_2)t}}{(\lambda_2+\mu_2)(\lambda_1+\mu_1-\lambda_2-\mu_2)}\right) \\
&-\lambda_1\lambda_2\mu_3\left(\begin{array}{l}\dfrac{e^{-(\lambda_1+\mu_1)t}}{(\lambda_1+\mu_1)(\lambda_2+\mu_2-\lambda_1-\mu_1)(\mu_3-\lambda_1-\mu_1)} \\ +\dfrac{e^{-(\lambda_2+\mu_2)t}}{(\lambda_2+\mu_2)(\lambda_1+\mu_1-\lambda_2-\mu_2)(\mu_3-\lambda_2-\mu_2)} \\ +\dfrac{e^{-\mu_3 t}}{\mu_3(\lambda_1+\mu_1-\mu_3)(\lambda_2+\mu_2-\mu_3)}\end{array}\right) \\
&+1
\end{aligned}
\tag{4.28}
$$

4.4 M Phases

Figure 4.3 illustrates an '*m*' phase Coxian Phase Type system, where λ_i represents the rate at which a customer moves from Phase *i* to Phase *i*+1, and μ_i is the rate at which customer a passes from Phase *i* into the absorbing Exit Phase.

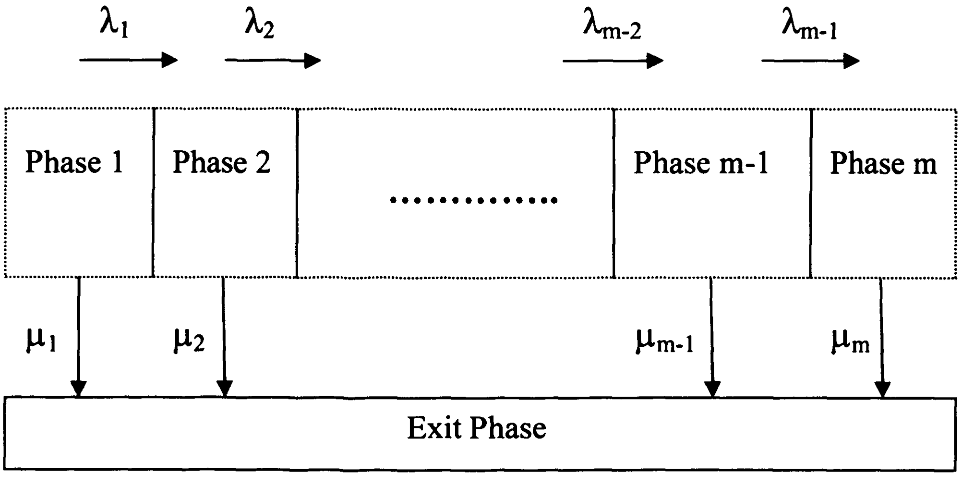


Figure 4.3 – An *m* phase Coxian Phase Type system

All phases have a Negative Exponential service distribution.

$$f_i(t) = (\lambda_i + \mu_i)e^{-(\lambda_i + \mu_i)t} \qquad \textit{for } 1 \le i \le m-1$$

$$f_m(t) = \mu_m e^{-\mu_m t}$$

The probability of a customer joining the Exit Phase from Phase 1 is the same as before;

$$\frac{\mu_1}{\lambda_1 + \mu_1} \tag{4.29}$$

The probability of a customer going to the Exit Phase from Phase i, where $2 \le i \le m-1$, is:

$$\prod_{j=1}^{i-1}\left(\frac{\lambda_j}{\lambda_j + \mu_j}\right)\frac{\mu_i}{\lambda_i + \mu_i} \tag{4.30}$$

The probability of a customer joining the Exit Phase from the m-th Phase is;

$$\prod_{j=1}^{m-1}\frac{\lambda_j}{\lambda_j+\mu_j} \tag{4.31}$$

The average length of stay can then be calculated to be;

$$\begin{aligned}&\frac{1}{\lambda_1+\mu_1}+\frac{\lambda_1}{(\lambda_1+\mu_1)}\frac{1}{(\lambda_2+\mu_2)}+\frac{\lambda_1}{(\lambda_1+\mu_1)}\frac{\lambda_2}{(\lambda_2+\mu_2)}\frac{1}{(\lambda_3+\mu_3)}+\ldots+\frac{\lambda_1\ldots\lambda_{m-1}}{(\lambda_1+\mu_1)\ldots(\lambda_{m-1}+\mu_{m-1})}\frac{1}{\mu_m}\\&=\frac{1}{\lambda_1+\mu_1}+\sum_{j=1}^{m-2}\left(\prod_{i=1}^{j}\frac{\lambda_i}{(\lambda_i+\mu_i)}\frac{1}{(\lambda_{j+1}+\mu_{j+1})}\right)+\prod_{i=1}^{m-1}\frac{\lambda_i}{(\lambda_i+\mu_i)}\frac{1}{\mu_m}\end{aligned} \tag{4.32}$$

Next, equations relating the state probabilities will be derived.

$$\begin{aligned}&P_1(t+\delta t)=P_1(t)(1-\lambda_1\delta t)(1-\mu_1\delta t)\\&P_i(t+\delta t)=P_i(t)(1-\lambda_i\delta t)(1-\mu_i\delta t) \qquad For\ 2\le i\le m-1\\&\qquad +P_{i-1}(t)\lambda_{i-1}\delta t\\&P_m(t+\delta t)=P_m(t)(1-\mu_m\delta t)\\&\qquad +P_{m-1}(t)\lambda_{m-1}\delta t\\&P_E(t+\delta t)=P_E(t)+P_m(t)\mu_m\delta t+P_{m-1}(t)\mu_{m-1}\delta t+\ldots+P_1(t)\mu_1\delta t\end{aligned} \tag{4.33}$$

To make the notation a little simpler here, the following substitutions will be used.

$$\begin{aligned}&\alpha_i=\lambda_i+\mu_i, \quad \text{for i=1, 2, 3, ..., m-1}\\&\alpha_m=\mu_m\end{aligned} \tag{4.34}$$

Using equations (4.33) the following differential equations can be formed;

$$\begin{aligned}
&\frac{dP_1(t)}{dt}=-\alpha_1 P_1(t)\\
&\frac{dP_i(t)}{dt}=-\alpha_i P_i(t)+\lambda_{i-1}P_{i-1}(t), \qquad 2\le i\le m-1\\
&\frac{dP_m(t)}{dt}=-\alpha_m P_m(t)+\lambda_{m-1}P_{m-1}(t)\\
&\frac{dP_E(t)}{dt}=\mu_m P_m(t)+\mu_{m-1}P_{m-1}(t)+\ldots+\mu_1 P_1(t)
\end{aligned} \qquad (4.35)$$

The first equation may be solved as previously

$$\begin{aligned}
&\int\frac{dP_1(t)}{dt}\frac{1}{P_1(t)}dt=-\int(\alpha_1)dt\\
&\ln P_1(t)=-(\alpha_1)t+C\\
&P_1(t)=e^{-(\alpha_1)t}
\end{aligned} \qquad (4.36)$$

In the set of equations in (4.35) the equation involving $P_i(t)$ is:

$$\frac{dP_i(t)}{dt}+(\alpha_i)P_i(t)=\lambda_{i-1}P_{i-1}(t) \qquad (4.37)$$

This can be solved using the integrating factor method. The integrating factor for this equation will be $e^{\alpha_i t}$.

$$\begin{aligned}
&e^{\alpha_i t}\frac{dP_i(t)}{dt}+\alpha_i e^{\alpha_i t}P_i(t)=\lambda_{i-1}P_{i-1}(t)e^{\alpha_i t}\\
&e^{\alpha_i t}P_i(t)=\lambda_{i-1}\int P_{i-1}(t)e^{\alpha_i t}dt
\end{aligned} \qquad (4.38)$$

To find probability $P_2(t)$ the expression for $P_1(t)$ is substituted in.

$$\begin{aligned} e^{(\alpha_2)t}P_2(t) &= \lambda_1 \int e^{(\alpha_2-\alpha_1)t}dt \\ &= \frac{\lambda_1}{\alpha_2-\alpha_1}e^{(\alpha_2-\alpha_1)t}+C \end{aligned} \tag{4.39}$$

Using the initial conditions of this system it can be seen that;

$$P_2(0)=0$$

$$C=-\frac{\lambda_1}{\alpha_2-\alpha_1}$$

$$e^{(\alpha_2)t}P_2(t)=\frac{\lambda_1}{\alpha_2-\alpha_1}\left(e^{(\alpha_2-\alpha_1)t}-1\right) \tag{4.40}$$

$$\begin{aligned} P_2(t) &= \frac{\lambda_1}{\alpha_2-\alpha_1}\left(e^{-\alpha_1 t}-e^{-\alpha_2 t}\right) \\ &= \frac{\lambda_1 e^{-\alpha_1 t}}{\alpha_2-\alpha_1}+\frac{\lambda_1 e^{-\alpha_2 t}}{\alpha_1-\alpha_2} \end{aligned}$$

An expression for $P_3(t)$ can now be found similarly;

$$
\begin{aligned}
e^{(\alpha_3)t}P_3(t) &= \lambda_2 \int e^{(\alpha_3)t} P_2(t)\,dt \\
&= \frac{\lambda_1\lambda_2}{\alpha_2-\alpha_1}\int e^{(\alpha_3-\alpha_1)t}dt + \frac{\lambda_1\lambda_2}{\alpha_1-\alpha_2}\int e^{(\alpha_3-\alpha_2)t}dt \\
&= \frac{\lambda_1\lambda_2 e^{(\alpha_3-\alpha_1)t}}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{\lambda_1\lambda_2 e^{(\alpha_3-\alpha_2)t}}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} + C \\
P_3(0) &= 0 \\
C &= -\left(\frac{\lambda_1\lambda_2}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{\lambda_1\lambda_2}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)}\right) \\
e^{(\alpha_3)t}P_3(t) &= \frac{\lambda_1\lambda_2\left(e^{(\alpha_3-\alpha_1)t}-1\right)}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{\lambda_1\lambda_2\left(e^{(\alpha_3-\alpha_2)t}-1\right)}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} \\
P_3(t) &= \frac{\lambda_1\lambda_2\left(e^{-\alpha_1 t}-e^{-\alpha_3 t}\right)}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{\lambda_1\lambda_2\left(e^{-\alpha_2 t}-e^{-\alpha_3 t}\right)}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} \\
&= \lambda_1\lambda_2\left(\frac{e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{e^{-\alpha_2 t}}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} - \frac{e^{-\alpha_3 t}}{(\alpha_2-\alpha_1)}\left(\frac{1}{\alpha_3-\alpha_1}-\frac{1}{\alpha_3-\alpha_2}\right)\right) \\
&= \lambda_1\lambda_2\left(\frac{e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{e^{-\alpha_2 t}}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} + \frac{e^{-\alpha_3 t}}{(\alpha_3-\alpha_1)(\alpha_3-\alpha_2)}\right) \\
&= \lambda_1\lambda_2\left(\frac{e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)} + \frac{e^{-\alpha_2 t}}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} + \frac{e^{-\alpha_3 t}}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)}\right)
\end{aligned}
\tag{4.41}
$$

These results are valid when the values of $\alpha_i \neq \alpha_j$, that is when the two phases have the same average service rate. If this were to occur with the data to be modelled different equations would have to be used.

It can be seen that a general form for the probability of the first m-1 phases being occupied at time t is being constructed;

$$P_i(t)=\prod_{j=0}^{i-1}\lambda_j\sum_{k=1}^{i}\frac{e^{-\alpha_k t}}{A_k} \qquad 1\le \mathrm{i}\le \mathrm{m\text{-}1} \qquad (4.42)$$

Where $A_k=\left\{\begin{array}{ll}\prod\limits_{\substack{l=1\\ l\ne k}}^{i}(\alpha_l-\alpha_k) & \text{For } i\ge 2\\ 1 & \text{For } i=1\end{array}\right\}$ and $\lambda_0=1$.

This formula can be shown to be correct by computing the general case using induction;

If $P_i(t)=\prod\limits_{j=1}^{i-1}\lambda_j\sum\limits_{k=1}^{i}\dfrac{e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\ne k}}^{i}(\alpha_l-\alpha_k)}$ for $2\le \mathrm{i}\le \mathrm{m\text{-}1}$, the probability of being in Phase *i*+1 at time t is:

$$P_{i+1}(t) = P_{i+1}(t)(1-\lambda_{i+1}\delta t)(1-\mu_{i+1}\delta t)$$
$$+P_i(t)\lambda_i\delta t$$

$$\frac{dP_{i+1}(t)}{dt} + \alpha_{i+1}P_{i+1}(t) = \lambda_i P_i(t)$$

$$e^{\alpha_{i+1}t}P_{i+1}(t) = \lambda_i \int P_i(t)e^{\alpha_{i+1}t}dt$$

$$= \lambda_i \left[\begin{array}{l} \displaystyle\int \frac{\lambda_1 \ldots \lambda_{i-1} e^{(\alpha_{i+1}-\alpha_1)t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)}dt \\ +\ldots \\ \displaystyle+\int \frac{\lambda_1 \ldots \lambda_{i-1} e^{(\alpha_{i+1}-\alpha_i)t}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)} \end{array}\right]$$

$$= \lambda_1 \ldots \lambda_i \left[\begin{array}{l} \displaystyle\frac{e^{(\alpha_{i+1}-\alpha_1)t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)} \\ +\ldots \\ \displaystyle+\frac{e^{(\alpha_{i+1}-\alpha_i)t}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)(\alpha_{i+1}-\alpha_i)} \end{array}\right] + C,$$

where C is an arbitary constant (4.43)

Using the initial condition that $P_i(0) = 0$ for $i > 0$;

$$C=-\lambda_1\ldots\lambda_i\left[\begin{array}{l}\dfrac{e^{(\alpha_{i+1}-\alpha_1)t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{(\alpha_{i+1}-\alpha_i)t}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)(\alpha_{i+1}-\alpha_i)}\end{array}\right]$$

$$e^{\alpha_{i+1}t}P_{i+1}(t)=\lambda_1\ldots\lambda_i\left[\begin{array}{l}\dfrac{e^{(\alpha_{i+1}-\alpha_1)t}-1}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{(\alpha_{i+1}-\alpha_i)t}-1}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)(\alpha_{i+1}-\alpha_i)}\end{array}\right]$$

$$P_{i+1}(t)=\lambda_1\ldots\lambda_i\left[\begin{array}{l}\dfrac{e^{-\alpha_1t}-e^{-\alpha_{i+1}t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{-\alpha_it}-e^{-\alpha_{i+1}t}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)(\alpha_{i+1}-\alpha_i)}\end{array}\right]$$

$$=\lambda_1\ldots\lambda_i\left[\begin{array}{l}\dfrac{e^{-\alpha_1t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{-\alpha_it}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)(\alpha_{i+1}-\alpha_i)}\end{array}\right]$$

$$-\lambda_1\ldots\lambda_ie^{-\alpha_{i+1}t}\left[\begin{array}{l}\dfrac{1}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)}\\ +\ldots\\ +\dfrac{1}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)(\alpha_{i+1}-\alpha_i)}\end{array}\right] \qquad (4.44)$$

Now the $e^{-\alpha_{i+1}t}$ term will be simplified;

$$\frac{1}{(\alpha_2-\alpha_1)(\alpha_3-\alpha_1)(\alpha_4-\alpha_1)...(\alpha_{i+1}-\alpha_1)}+\frac{1}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)(\alpha_4-\alpha_2)...(\alpha_{i+1}-\alpha_2)}$$
$$+\frac{1}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)(\alpha_4-\alpha_3)...(\alpha_{i+1}-\alpha_3)}+...+\frac{1}{(\alpha_1-\alpha_i)(\alpha_2-\alpha_i)(\alpha_3-\alpha_i)...(\alpha_{i+1}-\alpha_i)} \tag{4.45}$$

To confirm that the equation for $P_{i+1}(t)$ is of the general form, the $e^{\alpha_{i+1}(t)}$ terms should be equal.

$$-\frac{1}{\prod\limits_{\substack{j=1\\J\neq i+1}}^{i+1}(\alpha_j-\alpha_{i+1})} \tag{4.46}$$

All that is left is to show that;

$$\sum_{k=1}^{i}\frac{1}{\prod\limits_{\substack{j=1\\j\neq k}}^{i+1}(\alpha_j-\alpha_k)}=-\frac{1}{\prod\limits_{\substack{j=1\\J\neq i+1}}^{i+1}(\alpha_j-\alpha_{i+1})} \tag{4.47}$$

This can also be written as;

$$\sum_{k=1}^{i+1}\frac{1}{\prod\limits_{\substack{j=1\\j\neq k}}^{i+1}(\alpha_j-\alpha_k)}=0 \tag{4.48}$$

Consider the polynomial;

$$L(\alpha)=\sum_{j=1}^{n}\frac{L_j(\alpha)}{L_j(\alpha_j)} \tag{4.49}$$

where,

$$L_j(\alpha)=\prod_{\substack{i=1\\ i\neq j}}^{n}(\alpha-\alpha_i) \tag{4.50}$$

It is required that,

$$\frac{1}{\sum_{j=1}^{n}L_j(\alpha_j)}=0 \tag{4.51}$$

Firstly noting that $L(\alpha)$ is identically equal to 1, because it is a polynomial of degree $(n-1)$ in α, but equals 1 when α takes the n values $\alpha_1,\alpha_2,\alpha_3,\dots,\alpha_n$. This can only be so if $L(\alpha)$ is equal to 1 everywhere. So the coefficient of α^{n-1} in $L(\alpha)$ must be 0 (there is no α^{n-1} term, since $L(\alpha)\equiv 1$). But the coefficient of α^{n-1} in the expansion of $L(\alpha)$ is;

$$\frac{1}{\sum_{j=1}^{n}L_j(\alpha_j)} \tag{4.52}$$

Actually $L(\alpha)$ is a special case of the Lagrange interpolation formula used in numerical analysis. The Lagrange Interpolation Formula is;

$$f(x)=\sum_{i=1}^{n}\frac{L_i(x)}{L_i(x_i)}f_i \tag{4.53}$$

In this case all the $f_i=1$, since $f(x)=1$.

As $P_{i+1}(t)$ is of the general form described we have by induction that:

$$P_i(t)=\prod_{j=1}^{i-1}\lambda_j\sum_{k=1}^{i}\frac{e^{-\alpha_k t}}{\prod_{\substack{l=1\\ l\neq k}}^{i}(\alpha_l-\alpha_k)} \qquad 2\leq \text{i} \leq \text{m-1} \tag{4.54}$$

Next the m-th phase is considered, The m-th phase is similar to all the previous phases except for the fact that it only has one output transition. Equation (4.55) is the differential equation for the *m*-th phase.

$$\frac{dP_m(t)}{dt}+\mu_m P_m(t)=\lambda_{m-1}P_{m-1}(t) \tag{4.55}$$

The integrating factor of this equation is $e^{\alpha_m t}$, remembering that as the m-th phase only has one output $\alpha_m = \mu_m$. Using this gives;

$$e^{\alpha_m t}P_m(t) = \lambda_{m-1}\int P_{m-1}(t)e^{\alpha_m t}dt$$

$$= \lambda_{m-1}\prod_{j=1}^{m-2}\lambda_j \int\sum_{k=1}^{i}\frac{e^{(\alpha_m-\alpha_k)t}}{\prod_{\substack{l=1\\ l\neq k}}^{i}(\alpha_l-\alpha_k)}dt$$

$$= \prod_{j=1}^{m-1}\lambda_j\begin{bmatrix}\int\frac{e^{(\alpha_m-\alpha_1)t}}{(\alpha_2-\alpha_1)\dots(\alpha_{m-1}-\alpha_1)}dt \\ +\dots. \\ +\int\frac{e^{(\alpha_m-\alpha_{m-1})t}}{(\alpha_1-\alpha_{m-1})\dots(\alpha_{m-2}-\alpha_{m-1})}\end{bmatrix}$$

$$= \prod_{j=1}^{m-1}\lambda_j\begin{pmatrix}\frac{e^{(\alpha_m-\alpha_1)t}}{(\alpha_2-\alpha_1)..(\alpha_{m-1}-\alpha_1)(\alpha_m-\alpha_1)} \\ +\dots \\ +\frac{e^{(\alpha_m-\alpha_{m-1})t}}{(\alpha_1-\alpha_{m-1})\dots(\alpha_{m-2}-\alpha_{m-1})(\alpha_m-\alpha_{m-1})}\end{pmatrix}+C$$

$$P_m(0)=0$$

$$C = -\prod_{j=1}^{m-1}\lambda_j\begin{pmatrix}\frac{1}{(\alpha_2-\alpha_1)\dots(\alpha_{m-1}-\alpha_1)(\alpha_m-\alpha_1)} \\ +\dots \\ +\frac{1}{(\alpha_1-\alpha_{m-1})\dots(\alpha_{m-2}-\alpha_{m-1})(\alpha_m-\alpha_{m-1})}\end{pmatrix}$$

$$e^{\alpha_m t}P_m(t)=\prod_{j=1}^{m-1}\lambda_j\left(\begin{array}{l}\dfrac{e^{(\alpha_m-\alpha_1)t}-1}{(\alpha_2-\alpha_1)\ldots(\alpha_{m-1}-\alpha_1)(\alpha_m-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{(\alpha_m-\alpha_{m-1})t}-1}{(\alpha_1-\alpha_{m-1})\ldots(\alpha_{m-2}-\alpha_{m-1})(\alpha_m-\alpha_{m-1})}\end{array}\right)$$

$$P_m(t)=\prod_{j=1}^{m-1}\lambda_j\left(\begin{array}{l}\dfrac{e^{-\alpha_1 t}-e^{-\alpha_m t}}{(\alpha_2-\alpha_1)\ldots(\alpha_{m-1}-\alpha_1)(\alpha_m-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{-\alpha_{m-1}t}-e^{-\alpha_m t}}{(\alpha_1-\alpha_{m-1})\ldots(\alpha_{m-2}-\alpha_{m-1})(\alpha_m-\alpha_{m-1})}\end{array}\right)$$

$$=\prod_{j=1}^{m-1}\lambda_j\left(\begin{array}{l}\dfrac{e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)\ldots(\alpha_{m-1}-\alpha_1)(\alpha_m-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{-\alpha_{m-1}t}}{(\alpha_1-\alpha_{m-1})\ldots(\alpha_{m-2}-\alpha_{m-1})(\alpha_m-\alpha_{m-1})}\\ -e^{-\alpha_m t}\left(\begin{array}{l}\dfrac{1}{(\alpha_2-\alpha_1)\ldots(\alpha_m-\alpha_1)}+\ldots\\ +\dfrac{1}{(\alpha_1-\alpha_{m-1})\ldots(\alpha_m-\alpha_{m-1})}\end{array}\right)\end{array}\right)$$

$$=\prod_{j=1}^{m-1}\lambda_j\left(\begin{array}{l}\dfrac{e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)\ldots(\alpha_{m-1}-\alpha_1)(\alpha_m-\alpha_1)}\\ +\ldots\\ +\dfrac{e^{-\alpha_{m-1}t}}{(\alpha_1-\alpha_{m-1})\ldots(\alpha_{m-2}-\alpha_{m-1})(\alpha_m-\alpha_{m-1})}\\ +\dfrac{e^{-\alpha_m t}}{(\alpha_1-\alpha_m)\ldots(\alpha_{m-2}-\alpha_m)(\alpha_{m-1}-\alpha_m)}\end{array}\right)$$

$$=\prod_{j=1}^{m-1}\lambda_j\sum_{k=1}^{m}\frac{e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{m}(\alpha_l-\alpha_k)} \tag{4.56}$$

The final stage of this can be proved in a similar way to $P_i(t)$ for $1<i<m$ above.

This is of the same form as all previous phases probabilities, which as stated earlier is not surprising as it is of a similar form to all the previous phases. So when the sum of the output rates of each phase is taken, called α, the general form is obtained;

$$P_i(t)=\prod_{j=1}^{i-1}\lambda_j\sum_{k=1}^{i}\frac{e^{-\alpha_k t}}{\prod_{\substack{l=1\\l\neq k}}^{i}(\alpha_l-\alpha_k)} \qquad 1<i\leq m \tag{4.57}$$

For the final stage of this m phase system, the probability that the Exit Phase is occupied at any time has to be computed. The following differential-difference equation has already been created in equations (4.35).

$$\frac{dP_E(t)}{dt}=\mu_m P_m(t)+\mu_{m-1}P_{m-1}(t)+\ldots+\mu_1 P_1(t) \tag{4.58}$$

Substituting in the previously calculated probabilities;

$$\begin{aligned}P_E(t)&=\mu_m\int P_m(t)\,dt+\mu_{m-1}\int P_{m-1}(t)\,dt+\ldots+\mu_1\int P_1(t)\,dt\\&=\sum_{n=1}^{m}\int\mu_n P_n(t)\,dt\\&=\sum_{n=1}^{m}\mu_n\prod_{j=1}^{n-1}\lambda_j\sum_{k=1}^{n}-\frac{1}{\alpha_k}\frac{e^{-\alpha_k t}}{\prod_{\substack{l=1\\l\neq k}}^{n}(\alpha_l-\alpha_k)}+C_m\end{aligned} \tag{4.59}$$

To find the constant C the usual initial condition that $P_E(0)=0$ is used;

$$C_m=-\sum_{n=1}^{m}\mu_n\prod_{j=1}^{n-1}\lambda_j\sum_{k=1}^{n}-\frac{1}{\alpha_k}\frac{1}{\prod_{\substack{l=1\\l\neq k}}^{n}(\alpha_l-\alpha_k)} \tag{4.60}$$

This can be evaluated for $M = 2$ and then for $M = 3$ to see if there is any pattern.

$M = 2$

$$\begin{aligned}
C_2 &= -\left[\mu_1\left(-\frac{1}{\alpha_1}\right) + \lambda_1\mu_2\left(-\frac{1}{\alpha_1}\left(\frac{1}{\alpha_2-\alpha_1}\right) - \frac{1}{\alpha_2}\left(\frac{1}{\alpha_1-\alpha_2}\right)\right)\right] \\
&= \left[\frac{\mu_1}{\alpha_1} + \frac{\lambda_1\mu_2}{\alpha_1(\alpha_2-\alpha_1)} + \frac{\lambda_1\mu_2}{\alpha_2(\alpha_1-\alpha_2)}\right] \\
&= \frac{\mu_1}{\alpha_1} + \frac{\lambda_1\mu_2}{\alpha_2-\alpha_1}\left(\frac{1}{\alpha_1} - \frac{1}{\alpha_2}\right) \\
&= \frac{\mu_1}{\alpha_1} + \frac{\lambda_1\mu_2}{\alpha_1\alpha_2} \\
&= \frac{\mu_1\mu_2 + \lambda_1\mu_2}{\lambda_1\mu_2 + \mu_1\mu_2} \\
&= 1
\end{aligned}$$

M=3

$$\begin{aligned}
C_3 &= \frac{\mu_1}{\alpha_1} + \lambda_1\mu_2\left(\frac{1}{\alpha_1}\frac{1}{(\alpha_2-\alpha_1)} + \frac{1}{\alpha_2}\frac{1}{(\alpha_1-\alpha_2)}\right) \\
&\quad + \lambda_1\lambda_2\mu_3\left(\begin{array}{r} \frac{1}{\alpha_1}\frac{1}{(\alpha_3-\alpha_1)(\alpha_2-\alpha_1)} + \frac{1}{\alpha_2}\frac{1}{(\alpha_1-\alpha_2)(\alpha_3-\alpha_2)} \\ + \frac{1}{\alpha_3}\frac{1}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)} \end{array}\right) \\
&= \frac{\mu_1}{\alpha_1} + \frac{\lambda_1\mu_2}{\alpha_1\alpha_2} + \lambda_1\lambda_2\mu_3\left(\begin{array}{r} \frac{1}{\alpha_2-\alpha_1}\left(\frac{1}{\alpha_1(\alpha_3-\alpha_1)} - \frac{1}{\alpha_2(\alpha_3-\alpha_2)}\right) \\ + \frac{1}{\alpha_3}\frac{1}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)} \end{array}\right) \\
&= \frac{\lambda_2\mu_1 + \mu_1\mu_2 + \lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} + \lambda_1\lambda_2\mu_3\left(\begin{array}{r} \frac{1}{\alpha_2-\alpha_1}\left(\frac{\alpha_2\alpha_3 - \alpha_2^2 - \alpha_1\alpha_3 + \alpha_1^2}{\alpha_1\alpha_2(\alpha_3-\alpha_1)(\alpha_3-\alpha_2)}\right) \\ + \frac{1}{\alpha_3}\frac{1}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)} \end{array}\right) \\
&= \frac{\lambda_2\mu_1 + \mu_1\mu_2 + \lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} + \lambda_1\lambda_2\mu_3\left(\begin{array}{r} \frac{1}{\alpha_2-\alpha_1}\left(\frac{(\alpha_1-\alpha_3)(\alpha_1+\alpha_2)}{\alpha_1\alpha_2(\alpha_3-\alpha_1)(\alpha_3-\alpha_2)}\right) \\ + \frac{1}{\alpha_3}\frac{1}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)} \end{array}\right) \\
&= \frac{\lambda_2\mu_1 + \mu_1\mu_2 + \lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} + \lambda_1\lambda_2\mu_3\left(\frac{1}{(\alpha_3-\alpha_2)}\left(\frac{\alpha_1+\alpha_2}{\alpha_1\alpha_2(\alpha_2-\alpha_1)} - \frac{1}{\alpha_3(\alpha_1-\alpha_3)}\right)\right) \\
&= \frac{\lambda_2\mu_1 + \mu_1\mu_2 + \lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} + \lambda_1\lambda_2\mu_3\left(\frac{1}{(\alpha_3-\alpha_2)}\left(\frac{\alpha_1^2\alpha_3 - \alpha_1\alpha_3^2 + \alpha_1\alpha_2\alpha_3 - \alpha_2\alpha_3^2 - \alpha_1\alpha_2^2 + \alpha_1^2\alpha_2}{\alpha_1\alpha_2\alpha_3(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)}\right)\right) \\
&= \frac{\lambda_2\mu_1 + \mu_1\mu_2 + \lambda_1\mu_2}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)} + \frac{\lambda_1\lambda_2\mu_3}{\alpha_1\alpha_2\alpha_3} \\
&= \frac{\lambda_2\mu_1\mu_3 + \mu_1\mu_2\mu_3 + \lambda_1\mu_2\mu_3 + \lambda_1\lambda_2\mu_3}{(\lambda_1+\mu_1)(\lambda_2+\mu_2)\mu_3} \\
&= 1
\end{aligned} \qquad (4.61)$$

C_3 Is equal to 1, and it could be shown by induction that $C_i = 1$ for all values of *i*.

The following results can now be constructed for an M phase system:

$$P_1(t) = e^{-(\lambda_1+\mu_1)t}$$

$$P_i(t) = \prod_{j=1}^{i-1}\lambda_j \sum_{k=1}^{i} \frac{e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{i}(\alpha_l-\alpha_k)} \qquad \text{For } 2 < \mathrm{i} \leq m$$

$$P_E(t) = \sum_{n=1}^{m}\mu_n \prod_{j=1}^{n-1}\lambda_j \sum_{k=1}^{n} -\frac{1}{\alpha_k}\frac{e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{n}(\alpha_l-\alpha_k)} + 1$$

$$P_{i+1}(t) = P_i(t) + \prod_{j=1}^{i-1}\lambda_j \sum_{k=1}^{i+1} \frac{(\lambda_i+\alpha_k-\alpha_{i+1})e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{i+1}(\alpha_l-\alpha_k)}$$

$$\begin{aligned}
P_{i+1}(t) &= \prod_{j=1}^{i-1}\lambda_j \sum_{k=1}^{i} \frac{e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{i}(\alpha_l-\alpha_k)} + \prod_{j=1}^{i-1}\lambda_j \sum_{k=1}^{i+1} \frac{(\lambda_i+\alpha_k-\alpha_{i+1})e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{i+1}(\alpha_l-\alpha_k)} \\
&= \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)} + \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_2 t}}{(\alpha_1-\alpha_2)\ldots(\alpha_i-\alpha_2)} + \ldots + \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_i t}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)} \\
&+ \frac{\lambda_1 \ldots \lambda_{i-1}(\lambda_i+\alpha_1-\alpha_{i+1})e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)(\alpha_{i+1}-\alpha_1)} + \frac{\lambda_1 \ldots \lambda_{i-1}(\lambda_i+\alpha_2-\alpha_{i+1})e^{-\alpha_2 t}}{(\alpha_1-\alpha_2)\ldots(\alpha_i-\alpha_2)(\alpha_{i+1}-\alpha_2)} + \ldots \\
&\ldots + \frac{\lambda_1 \ldots \lambda_{i-1}(\lambda_i+\alpha_i-\alpha_{i+1})e^{-\alpha_i t}}{(\alpha_1-\alpha_i)\ldots(\alpha_i-\alpha_i)(\alpha_{i+1}-\alpha_i)} + \frac{\lambda_1 \ldots \lambda_{i-1}(\lambda_i+\alpha_{i+1}-\alpha_{i+1})e^{-\alpha_{i+1} t}}{(\alpha_1-\alpha_{i+1})\ldots(\alpha_{i-1}-\alpha_{i+1})(\alpha_i-\alpha_{i+1})} \\
&= \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_1 t}}{(\alpha_2-\alpha_1)\ldots(\alpha_i-\alpha_1)}\left(1+\frac{\lambda_n+\alpha_1-\alpha_{i+1}}{\alpha_{i+1}-\alpha_1}\right) \\
&+ \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_1 t}}{(\alpha_1-\alpha_2)\ldots(\alpha_i-\alpha_2)}\left(1+\frac{\lambda_n+\alpha_2-\alpha_{i+1}}{\alpha_{i+1}-\alpha_2}\right) \\
&+ \ldots \\
&+ \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_i t}}{(\alpha_1-\alpha_i)\ldots(\alpha_{i-1}-\alpha_i)}\left(1+\frac{\lambda_n+\alpha_i-\alpha_{i+1}}{\alpha_{i+1}-\alpha_i}\right) \\
&+ \frac{\lambda_1 \ldots \lambda_{i-1} e^{-\alpha_{i+1} t}}{(\alpha_1-\alpha_{i+1})\ldots(\alpha_i-\alpha_{i+1})} \\
&= \prod_{j=1}^{i}\lambda_j \sum_{k=1}^{i+1} \frac{e^{-\alpha_k t}}{\prod\limits_{\substack{l=1\\ l\neq k}}^{i+1}(\alpha_l-\alpha_k)} = P_{i+1}(t)
\end{aligned}$$

(4.62)

which is the result required, i.e. (i+1) replaces i in the proposed general form for $P_i(t)$.

4.5 Transitional Matrix Approach

4.5.1 M|M|1

A more usual way of looking at the Phase Type equations is to use a transitional matrix; this will now be considered to validate the steady state method. The transitional matrix approach involves setting up the equations into a matrix and vector form to aid the solution method. To introduce the idea of transitional matrices, we firstly consider a traditional M|M|1|∞|FIFO system. The first three differential equations that can be created for this system are as follows;

$$\begin{aligned}
\frac{dP_0(t)}{dt} &= -\lambda P_0(t) + \mu P_1(t) \\
\frac{dP_1(t)}{dt} &= -(\lambda+\mu)P_1(t) + \lambda P_0(t) + \mu P_2(t) \\
\frac{dP_2(t)}{dt} &= -(\lambda+\mu)P_2(t) + \lambda P_1(t) + \mu P_3(t)
\end{aligned} \tag{4.63}$$

The rest of the equations follow on in the same form. Now by defining two vectors $\underline{P}$ and $\underline{P}'$ and a matrix Q these equations can be set up in transitional matrix form;

$$\underline{P}' = \underline{P}\underline{Q} \tag{4.64}$$

Where $\underline{P} = \left[P_0(t), P_1(t), P_2(t), \ldots\right]$ and $\underline{P}' = \left[P_0'(t), P_1'(t), P_2'(t), \ldots\right]$. Considering equations (4.63) as a Markovian process, then Q is the matrix of the rates at which customers go from state *i* to state *j*, these can be summarised below;

To State

$$\begin{array}{cc|cccc} & & 0 & 1 & 2 & 3 \\ \hline & 0 & -\lambda & \lambda & & \\ \text{From State} & 1 & \mu & -(\lambda+\mu) & \lambda & \\ & 2 & & \mu & -(\lambda+\mu) & \lambda \\ & 3 & & & \mu & -(\lambda+\mu) \end{array}$$

Next consider an alternate form of equation (4.64);

$$\frac{dy}{dx} = yf(x) \tag{4.65}$$

This equation can be solved in the following way;

$$\begin{gathered} \int \frac{dy}{y} = \int f(x)\,dx \\ \ln y = F(x) + C \qquad (4.66) \\ y = ke^{F(x)} \end{gathered}$$

Using the same logic, the solution to equation (4.64) is;

$$\underline{P} = \underline{P}_0 e^{Qt} \tag{4.67}$$

Where $\underline{P}_0 = [1,0,0,...]$. A feature of this method is that in the solution, there is an Exponential raised to the power of Q, which is a matrix. A way of interpreting this is by using the Taylor series for the expansion of an Exponential;

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + ... \tag{4.68}$$

In Matrix notation;

$$e^{Qt} = I + Qt + \frac{(Qt)^2}{2!} + ... \tag{4.69}$$

Now substituting this into equation(4.67), a solution for $\underline{P}$ and hence the system can be found;

$$\begin{aligned}\underline{P} &= [1,0,0,...]\left\{\begin{pmatrix} 1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{pmatrix} + \begin{pmatrix} -\lambda & \lambda & & \\ \mu & -(\lambda+\mu) & \lambda & \\ & \mu & -(\lambda+\mu) & \lambda \\ & & \mu & \ddots \end{pmatrix} t + \frac{Q^2t^2}{2!} + ...\right\} \\ &= [1,0,0,...,0] + [-\lambda t, \lambda t, 0,...,0] + \frac{t^2}{2}\left[\lambda(\lambda+\mu), -\lambda(2\lambda+\mu), \lambda^2, 0,...,0\right] + ...\end{aligned} \tag{4.70}$$

As $\underline{P} = \left[P_0(t), P_1(t), P_2(t),...\right]$ to obtain an expression for $P_0(t)$, the sum of the first element in each of the vectors on the right hand side is taken; similarly for all other states.

4.5.2 Two Phases

The method outlined above will now be used to analyse the Phase-Type distribution with two phases. The differential equations required have already been set up in equations (4.2), (4.4) and (4.7). The equation for the absorbing or Exit Phase will be left out of the transitional matrix and will be computed at the end, using the fact that all probabilities must sum to 1. We define the following vectors and matrices.

$$\begin{aligned} \underline{P} &= \left[P_0(t), P_1(t)\right] \\ \underline{P}' &= \left[P_0'(t), P_1'(t)\right] \\ Q &= \begin{pmatrix} -(\lambda+\mu_1) & \lambda \\ 0 & -\mu_2 \end{pmatrix} \end{aligned} \tag{4.71}$$

As before, by using equation (4.64) the solution to this system can be found and is of the form;

$$\underline{P} = \underline{P}_0 e^{Qt} \tag{4.72}$$

Where $e^{Qt} = I + Qt + \frac{(Qt)^2}{2!} + \frac{(Qt)^3}{3!} + \ldots$ To find Q to high powers can be time consuming and tedious, but can be much simplified by using the fact that $Q = RDR^{-1}$, where D is the diagonal matrix whose elements are of the Eigenvalues, and R is the matrix made up of the corresponding Eigenvectors of Q. The Eigenvalues, α_i, are worked out as follows;

$$\begin{aligned} |\alpha I - Q| &= \begin{vmatrix} \alpha+\lambda+\mu_1 & -\lambda \\ 0 & \alpha+\mu_2 \end{vmatrix} \\ &= \left(\alpha+(\lambda+\mu_1)\right)\left(\alpha+\mu_2\right) \end{aligned} \tag{4.73}$$

this gives the Eigenvalues:

$$\begin{aligned} \alpha_1 &= -(\lambda+\mu_1) \\ \alpha_2 &= -\mu_2 \end{aligned} \tag{4.74}$$

The Eigenvectors can now be calculated. The two Eigenvectors for this system are;

$$v_1 = \begin{pmatrix} x \\ \dfrac{\lambda+\mu_1-\mu_2}{\lambda} x \end{pmatrix}, \quad v_2 = \begin{pmatrix} x \\ 0 \end{pmatrix} \tag{4.75}$$

By letting *x*=1 resulting matrix made up of Eigenvectors is;

$$R = \begin{pmatrix} 1 & 1 \\ 0 & \dfrac{(\mu_1+\lambda-\mu_2)}{\lambda} \end{pmatrix} \tag{4.76}$$

The transitional matrix now becomes;

$$\begin{aligned} Q &= \begin{pmatrix} 1 & 1 \\ 0 & \dfrac{\lambda+\mu_1-\mu_2}{\lambda} \end{pmatrix} \begin{pmatrix} -(\lambda+\mu_1) & 0 \\ 0 & -\mu_2 \end{pmatrix} \begin{pmatrix} 1 & \dfrac{\lambda}{\mu_2-\lambda-\mu_1} \\ 0 & -\dfrac{\lambda}{\mu_2-\lambda-\mu_1} \end{pmatrix} \\ &= \begin{pmatrix} -(\lambda+\mu_1) & -\mu_2 \\ 0 & \dfrac{-\lambda\mu_2}{\lambda+\mu_1-\mu_2} \end{pmatrix} \begin{pmatrix} 1 & \dfrac{\lambda}{\mu_2-\lambda-\mu_1} \\ 0 & -\dfrac{\lambda}{\mu_2-\lambda-\mu_1} \end{pmatrix} \\ &= \begin{pmatrix} -(\lambda+\mu_1) & \lambda \\ 0 & -\mu_2 \end{pmatrix} \end{aligned} \tag{4.77}$$

The square of Q can now be easily found;

$$Q^2=\begin{pmatrix}1 & 1\\ 0 & \dfrac{\lambda+\mu_1-\mu_2}{\lambda}\end{pmatrix}\begin{pmatrix}(\lambda+\mu_1)^2 & 0\\ 0 & \mu_2^2\end{pmatrix}\begin{pmatrix}1 & \dfrac{\lambda}{\mu_2-\lambda-\mu_1}\\ 0 & -\dfrac{\lambda}{\mu_2-\lambda-\mu_1}\end{pmatrix}$$

$$=\begin{pmatrix}(\lambda+\mu_1)^2 & \dfrac{-\lambda(\lambda+\mu_1)^2+\lambda\mu_2^2}{\lambda+\mu_1-\mu_2}\\ 0 & \mu_2^2\end{pmatrix} \tag{4.78}$$

Using the same method the cube of Q can be found to be;

$$Q^3=\begin{pmatrix}-(\lambda+\mu_1)^3 & \dfrac{\lambda(\lambda+\mu_1)^3-\lambda\mu_2^3}{\lambda+\mu_1-\mu_2}\\ 0 & -\mu_2^3\end{pmatrix} \tag{4.79}$$

We may proceed in a similar fashion as the power of Q increases. It is worth noting that the numerator of the top right corner of these matrices can be expressed in a different form;

$$(-1)^{n-1}\lambda(x^n-y^n)=(-1)^{n-1}\lambda(x-y)(x^{n-1}+x^{n-2}y+x^{n-3}y^2+\ldots+xy^{n-2}+y^{n-1}) \tag{4.80}$$

where $x=\lambda+\mu_1$ and $y=\mu_2$. Now that a factor has been removed it can cancel with the denominator. The resulting form of Q to any power greater than or equal to 2 is;

$$Q^n=(-1)^n\begin{pmatrix}(\lambda+\mu_1)^n & -\lambda(x^{n-1}+x^{n-2}y+x^{n-3}y^2+\ldots+xy^{n-2}+y^{n-1})\\ 0 & \mu_2^n\end{pmatrix} \tag{4.81}$$

A value of $\underline{P}$ can now be found;

$$P=\begin{bmatrix}1 & 0\end{bmatrix}\left(\begin{array}{l}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}+\begin{pmatrix}-(\lambda+\mu_1) & \lambda\\ 0 & -\mu_2\end{pmatrix}t+\dfrac{1}{2}\begin{pmatrix}(\lambda+\mu_1) & -\lambda(\lambda+\mu_1+\mu_2)\\ 0 & \mu_2^2\end{pmatrix}t^2\\ +\dfrac{1}{6}\begin{pmatrix}-(\lambda+\mu_1)^3 & \lambda\left((\lambda+\mu_1)^2+(\lambda+\mu_1)\mu_2+\mu_2^2\right)\\ 0 & -\mu_2^3\end{pmatrix}t^3+\ldots\end{array}\right)$$

$$=(1\quad 0)+(-(\lambda+\mu_1)\quad \lambda)t+\left(\frac{(\lambda+\mu_1)^2}{2}\quad -\frac{\lambda(\lambda+\mu_1+\mu_2)}{2}\right)t^2$$

$$+\left(\frac{-(\lambda+\mu_1)^3}{6}\quad \frac{\lambda\left((\lambda+\mu_1)^2+(\lambda+\mu_1)\mu_2+\mu_2^2\right)}{6}\right)t^3+\ldots$$

(4.82)

From this the values for each element in $\underline{P}$ can be obtained;

$$\begin{aligned}P_1(t)&=1-(\lambda+\mu_1)t+\frac{(\lambda+\mu_1)^2}{2!}t^2-\frac{(\lambda+\mu_1)^3}{3!}t^3+\ldots\\ &=e^{-(\lambda+\mu_1)t}\end{aligned} \tag{4.83}$$

The value for $P_2(t)$ can also be found;

$$P_2(t)=\lambda t\left[1-\frac{\lambda+\mu_1+\mu_2}{2}t+\frac{(\lambda+\mu_1)^2+(\lambda+\mu_1)\mu_2+\mu_2^2}{6}t+\ldots\right] \tag{4.84}$$

Though initially it does not seem that this can be reduced further, the change that was introduced to simplify the matrices in equations (4.80) is undone $P_2(t)$ can be written more simply as;

$$
\begin{aligned}
P_2(t) &= 0 + \lambda t - \frac{\lambda\left((\lambda+\mu_1)^2 - \mu_2^2\right)}{2(\lambda+\mu_1-\mu_2)}t^2 + \frac{\lambda\left((\lambda+\mu_1)^3 - \mu_2^3\right)}{6(\lambda+\mu_1-\mu_2)}t^3 - \dots \\
&= \frac{\lambda}{\lambda+\mu_1-\mu_2}\left(t(\lambda+\mu_1-\mu_2) - \frac{(\lambda+\mu_1)^2 - \mu_2^2}{2}t^2 + \frac{(\lambda+\mu_1)^3 - \mu_2^3}{6}t^3 - \dots\right) \\
&= \frac{\lambda}{\lambda+\mu_1-\mu_2}\begin{pmatrix} t(\lambda+\mu_1) - \frac{(\lambda+\mu_1)^2}{2}t^2 + \frac{(\lambda+\mu_1)^3}{6}t^3 - \dots \\ -\mu_2 t + \frac{\mu_2^2}{2}t^2 - \frac{\mu_3^2}{6}t^3 + \dots \end{pmatrix} \\
&= \frac{\lambda}{\lambda+\mu_1-\mu_2}\left(-e^{-(\lambda+\mu_1)t} + 1 + e^{\mu_2 t} - 1\right) \\
&= \frac{\lambda}{\mu_2-\lambda-\mu_1}\left(e^{-(\lambda+\mu_1)t} - e^{\mu_2 t}\right)
\end{aligned}
\tag{4.85}
$$

Now that solution have been found for the first two states the probability for the absorbing/Exit Phase can be calculated as 1 minus the sum of the other two states. This gives the solution of;

$$
P_E(t) = 1 - e^{-(\lambda+\mu_1)t} - \frac{\lambda}{\mu_2-\lambda-\mu_1}\left(e^{-(\lambda+\mu_1)t} - e^{-\mu_2 t}\right) \tag{4.86}
$$

These results correspond to the solutions gained in section 4.2.

4.5.3 Three Phases.

The computation of the system with three phases is similar to that of the two phase case, expect the matrices and vectors are larger in dimension; in the three phase system the vectors and matrices are;

$$\begin{aligned}
&\underline{P}=\left(P_1(t),P_2(t),P_3(t)\right)\\
&\underline{P}'=\left(\underline{P}_1'(t),\underline{P}_2'(t),\underline{P}_3'(t)\right)\\
&Q=\begin{pmatrix}-(\lambda_1+\mu_1) & \lambda_1 & 0\\ 0 & -(\lambda_2+\mu_2) & \lambda_2\\ 0 & 0 & -\mu_3\end{pmatrix}
\end{aligned}\qquad(4.87)$$

Again the Exit Phase differential equation has been left out of the transitional matrix as it will be calculated at the end of the process. The matrices R and D which are used in calculation powers of matrix Q are;

$$\begin{aligned}
&R=\begin{pmatrix}x & x & x\\ 0 & \dfrac{\left((\lambda_1+\mu_1)-(\lambda_2+\mu_2)\right)x}{\lambda_1} & \dfrac{\left((\lambda_1+\mu_1)-\mu_3\right)x}{\lambda_1}\\ 0 & 0 & \dfrac{\left((\lambda_1+\mu_1)-\mu_3\right)\left((\lambda_2+\mu_2)-\mu_3\right)x}{\lambda_1\lambda_2}\end{pmatrix}\\
&D=\begin{pmatrix}-(\lambda_1+\mu_1) & 0 & 0\\ 0 & -(\lambda_2+\mu_2) & 0\\ 0 & 0 & -\mu_3\end{pmatrix}
\end{aligned}\qquad(4.88)$$

The matrices produced when Q is raised to different powers greater than or equal to three are;

$$\begin{pmatrix} (-\alpha_1)^n & (-1)^{n+1}\dfrac{\lambda_1}{\gamma_{12}}\left(\alpha_1^n-\alpha_2^n\right) & (-1)^n\lambda_1\lambda_2\left(\dfrac{\mu_1^n}{\gamma_{12}\gamma_{13}}-\dfrac{\mu_2^n}{\gamma_{12}\gamma_{23}}+\dfrac{\mu_3^n}{\gamma_{13}\gamma_{23}}\right) \\ 0 & (-\alpha_2)^n & (-1)^{n+1}\dfrac{\lambda_2}{\gamma_{23}}\left(\alpha_2^3-\alpha_3^3\right) \\ 0 & 0 & (-\alpha_3)^n \end{pmatrix} \tag{4.89}$$

Here $\gamma_{ij}=\alpha_i-\alpha_j$ for conciseness. When Q is squared the matrix produced is;

$$\begin{pmatrix} \alpha_1^2 & \dfrac{-\lambda_1}{\alpha_1-\alpha_2}\left(\alpha_1^2-\alpha_2^2\right) & \lambda_1\lambda_2 \\ 0 & \alpha_2^2 & \dfrac{-\lambda_2}{\alpha_2-\alpha_3}\left(\alpha_2^2-\alpha_3^2\right) \\ 0 & 0 & \alpha_3^2 \end{pmatrix} \tag{4.90}$$

Using these matrices to find the values for the specific states, the solution for $P_1(t)$ is;

$$\begin{aligned} P_1(t) &= 1-\alpha_1 t+\frac{\alpha_1^2}{2}t^2-\frac{\alpha_1^3}{6}t^3+\ldots \\ &= e^{-\alpha_1 t} \end{aligned} \tag{4.91}$$

Similarly for $P_2(t)$;

$$\begin{aligned}
P_2(t) &= \lambda_1 t - \frac{\lambda_1}{\alpha_1-\alpha_2}\frac{t^2}{2}\left(\alpha_1^2-\alpha_2^2\right) + \frac{\lambda_1}{\alpha_1-\alpha_2}\frac{t^3}{6}\left(\alpha_1^3-\alpha_2^3\right) - \ldots \\
&= \frac{\lambda_1}{\alpha_1-\alpha_2}\begin{pmatrix} t\alpha_1 - \alpha_1^2\dfrac{t^2}{2} + \alpha_1^3\dfrac{t^3}{6} - \ldots \\ \\ -t\alpha_2 + \alpha_2^2\dfrac{t^2}{2} - \alpha_2^3\dfrac{t^3}{6} - \ldots \end{pmatrix} \\
&= \frac{\lambda_1}{\alpha_1-\alpha_2}\left(-e^{-\alpha_1 t} + 1 + e^{-\alpha_2 t} - 1\right) \\
&= \frac{\lambda_1}{\alpha_1-\alpha_2}\left(e^{-\alpha_2 t} - e^{-\alpha_1 t}\right)
\end{aligned} \tag{4.92}$$

$P_3(t)$ can be worked out in a similar way;

$$P_3(t) = \lambda_1\lambda_2\left(\frac{t^2}{2} - \frac{t^3}{6}\left(\frac{\alpha_1^3}{\gamma_{13}\gamma_{12}} - \frac{\alpha_2^3}{\gamma_{12}\gamma_{23}} + \frac{\alpha_3^3}{\gamma_{13}\gamma_{23}}\right) + \ldots\right) \tag{4.93}$$

This can be expressed as;

$$P_3(t) = \lambda_1\lambda_2\begin{pmatrix} \dfrac{1}{\gamma_{12}\gamma_{13}}\left(e^{-\alpha_1 t} - \dfrac{\alpha_1^2}{2}t^2 + \alpha_1 t - 1\right) \\ -\dfrac{1}{\gamma_{12}\gamma_{23}}\left(e^{-\alpha_2 t} - \dfrac{\alpha_2^2}{2}t^2 + \alpha_2 t - 1\right) \\ +\dfrac{1}{\gamma_{13}\gamma_{23}}\left(e^{-\alpha_3 t} - \dfrac{\alpha_3^2}{2}t^2 + \alpha_3 t - 1\right) \\ +\dfrac{t^2}{2} \end{pmatrix} \tag{4.94}$$

Taking the summation of the *t* terms that are of the power 2;

$$\begin{aligned}
&-\frac{\alpha_1^2}{\gamma_{12}\gamma_{13}}+\frac{\alpha_2^2}{\gamma_{12}\gamma_{23}}-\frac{\alpha_3^2}{\gamma_{13}\gamma_{23}} \\
&=\frac{-\alpha_1^2\gamma_{23}+\alpha_2^2\gamma_{13}}{\gamma_{12}\gamma_{13}\gamma_{23}}-\frac{\alpha_3^2}{\gamma_{13}\gamma_{23}} \\
&=\frac{\alpha_3(\alpha_1-\alpha_2)(\alpha_1+\alpha_2)-\alpha_1\alpha_2(\alpha_1-\alpha_2)}{\gamma_{12}\gamma_{13}\gamma_{23}}-\frac{\alpha_3^2}{\gamma_{13}\gamma_{23}} \qquad (4.95)\\
&=\frac{\alpha_3(\alpha_1+\alpha_2)-\alpha_1\alpha_2}{\gamma_{13}\gamma_{23}}-\frac{\alpha_3^2}{\gamma_{13}\gamma_{23}} \\
&=\frac{(\alpha_1-\alpha_3)(\alpha_3-\alpha_2)}{\gamma_{12}\gamma_{23}} \\
&=-1
\end{aligned}$$

Now take the summation of the *t* terms;

$$\begin{aligned}
&\frac{\alpha_1}{\gamma_{12}\gamma_{13}}-\frac{\alpha_2}{\gamma_{12}\gamma_{23}}+\frac{\alpha_3}{\gamma_{13}\gamma_{23}} \\
&=\frac{-\alpha_1\alpha_3+\alpha_2\alpha_3}{\gamma_{12}\gamma_{13}\gamma_{23}}+\frac{\alpha_3}{\gamma_{13}\gamma_{23}} \qquad (4.96)\\
&=-\frac{\alpha_3}{\gamma_{23}\gamma_{13}}+\frac{\alpha_3}{\gamma_{23}\gamma_{13}} \\
&=0
\end{aligned}$$

Finally taking the sum of the constant terms;

$$\begin{aligned}&-\frac{1}{\gamma_{12}\gamma_{13}}+\frac{1}{\gamma_{12}\gamma_{23}}-\frac{1}{\gamma_{13}\gamma_{23}}\\&=\frac{\alpha_1-\alpha_2}{\gamma_{12}\gamma_{13}\gamma_{23}}-\frac{1}{\gamma_{13}\gamma_{23}}\\&=\frac{1}{\gamma_{13}\gamma_{23}}-\frac{1}{\gamma_{13}\gamma_{23}}\\&=0\end{aligned}\tag{4.97}$$

Now the value for $P_3(t)$ can be simplified;

$$\begin{aligned}P_3(t)&=\lambda_1\lambda_2\left(\frac{e^{-\alpha_1 t}}{\gamma_{12}\gamma_{13}}-\frac{e^{-\alpha_2 t}}{\gamma_{12}\gamma_{23}}+\frac{e^{-\alpha_3 t}}{\gamma_{13}\gamma_{23}}+\frac{t^2}{2}-\frac{t^2}{2}\right)\\&=\lambda_1\lambda_2\left(\frac{e^{-\alpha_1 t}}{\gamma_{12}\gamma_{13}}-\frac{e^{-\alpha_2 t}}{\gamma_{12}\gamma_{23}}+\frac{e^{-\alpha_3 t}}{\gamma_{13}\gamma_{23}}\right)\end{aligned}\tag{4.98}$$

The value for $P_E(t)$ can be found easily from these formulae as previously:

$$\begin{aligned}P_E(t)&=1-e^{-\alpha_1 t}-\frac{\lambda_1}{\alpha_1-\alpha_2}\left(e^{-\alpha_2 t}-e^{-\alpha_1 t}\right)\\&-\lambda_1\lambda_2\left(\frac{e^{-\alpha_1 t}}{(\alpha_1-\alpha_2)(\alpha_1-\alpha_3)}-\frac{e^{-\alpha_2 t}}{(\alpha_1-\alpha_2)(\alpha_2-\alpha_3)}+\frac{e^{-\alpha_3 t}}{(\alpha_1-\alpha_3)(\alpha_2-\alpha_3)}\right)\end{aligned}\tag{4.99}$$

With some algebraic manipulation this can be seen to be the same as the result in section 4.3. This second, original, method of solving Phase Type equations confirms the result gained in the first part of this chapter. In Chapter 6 Phase Type models will be used to model a bed blocking situation. The results of which can then be compared to bed blocking equations.

Chapter 5

BLOCKING WITH MULTIPLE ROUTES

5.1 One Phase

5.1.1 Introduction

The next scenario to be considered is an extension of the blocking scenario seen in Chapter 3. In this chapter, an extra route or service channel will be added and the effect of service rates and throughput will be examined.

A comparison will be drawn between multiple servers in parallel in which routes do not share a queue and the shared queue system, to highlight which of these systems is more productive for the customer.

The simulation used in Chapter 3 will also be extended to allow for multiple routes and the results that this simulation produces will be analysed in comparison with the theoretical results also covered within this chapter.

There will be a distinct change in the notation of this chapter in that the term 'node' will be used less, as it implies a single service facility, and will be replaced by 'phase'. A phase encompasses many nodes that are in the same section of the network.

5.1.2 Zero Queueing Space

Consider a single phase queueing system; single phase refers to each customer only having to pass through a single service facility, with two servers that serve at the same rate. The system has no queueing space and no blocking occurs as there is no second service point. This is a standard queueing model and using Kendal's notation it is described as M|M|2 (System)|2|FIFO and can be see in Figure 5.1.

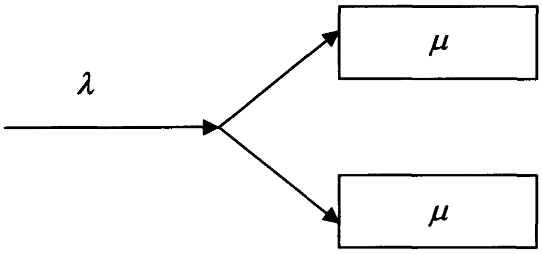


Figure 5.1 - An M|M|2 queueing system

The time dependent equations for this system are set up as follows;

$$
\begin{aligned}
&P_0(t+\delta t)=(1-\lambda\delta t)P_0(t)+\mu\delta t(1-\lambda\delta t)P_1(t)\\
&P_1(t+\delta t)=(1-\lambda\delta t)(1-\mu\delta t)P_1(t)+\lambda\delta t P_0(t)+2\mu\delta t P_2(t)\\
&P_2(t+\delta t)=(1-2\mu\delta t)P_2(t)+\lambda\delta t(1-\mu\delta t)P_1(t)
\end{aligned}
\tag{5.1}
$$

where $P_i(t)$ is the probability of there being *i* customers in the system at time *t*. These can then be rearranged to give;

$$
\begin{aligned}
&\frac{dP_0(t)}{dt}=-\lambda P_0(t)-\mu P_1(t)\\
&\frac{dP_1(t)}{dt}=-(\lambda+\mu)P_1(t)+\lambda P_0(t)+2\mu P_2(t)\\
&\frac{dP_2(t)}{dt}=-2\mu P_2(t)+\lambda P_1(t)
\end{aligned}
\tag{5.2}
$$

The left hand side is set to zero to obtain the steady state solution;

$$\begin{aligned} P_1 &= \frac{\lambda}{\mu} P_0 \\ P_2 &= \frac{\lambda}{2\mu} P_1 \\ &= \frac{\lambda^2}{2\mu^2} P_0 \end{aligned} \tag{5.3}$$

As the sum of the three probabilities is 1, the following solutions can easily be found;

$$\begin{aligned} P_0 &= \frac{2\mu^2}{(\lambda+\mu)^2+\mu^2} \\ P_1 &= \frac{2\lambda\mu}{(\lambda+\mu)^2+\mu^2} \\ P_2 &= \frac{\lambda^2}{(\lambda+\mu)^2+\mu^2} \end{aligned} \tag{5.4}$$

5.1.3 Distinguish Between Service Points

In this traditional way of setting up two server equations, it is not possible to distinguish which server is active. For example in equations (5.1) it is only possible to tell whether both servers are quiet, one is in service, or both are in service. Before it is possible to add another phase to this system, a method of distinguishing, if only one route is busy, which one it is. To do this, new notation must be introduced, $P_{n,a,b}(t)$. This is the probability that, at time *t*, there are *n* customers in the pooled queue, *a* customers with server number 1 and *b* customers with server number 2. Within the bound of this model *a* and *b* will be equal to either 1 or zero i.e. each server will only be able to serve one customer at a time. The following system is more general than that of the previous section. A single phase will still be considered, but an infinite queue will be allowed to form and the service rates of the servers will able to be different. This more general system can be seen in Figure 5.2.

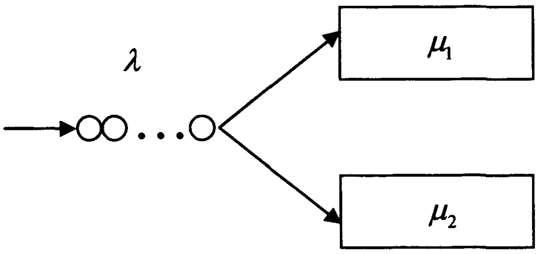


Figure 5.2 - An M|M|2 queueing system with a space to wait before start of service

For this system to distinguish between two service routes, when a customer starts their service they will need to be sent to a specific route. To achieve this, the chance of a customer going to a specific route is given a probability. If Route 1 is given the probability σ of being the selected service point then Route 2 will obviously have the probability $(1-\sigma)$ of being selected. This new system can be seen in Figure 5.3

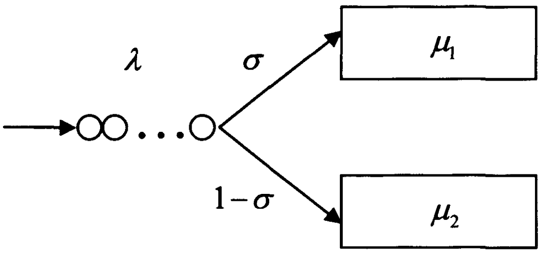


Figure 5.3 - An M|M|2 system with route probabilities

It is worth noting that this queueing system looks similar to the system with Hyper-Exponential service distribution. However in the system with a Hyper-Exponential service distribution, only 1 customer is able to be served at a time between the two routes. In this model both service points can be occupied at the same time.

The time dependent equations of this system are;

$$\begin{aligned}
P_{0,0,0}(t+\delta t) &= (1-\lambda\delta t)P_{0,0,0}(t)+\mu_1\delta t(1-\lambda\delta t)P_{0,1,0}+\mu_2\delta t(1-\lambda\delta t)P_{0,0,1}(t)\\
P_{0,1,0}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)P_{0,1,0}(t)+\sigma\lambda\delta t P_{0,0,0}(t)+\mu_2\delta t(1-\lambda\delta t)(1-\mu_1\delta t)P_{0,1,1}(t)\\
P_{0,0,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_2\delta t)P_{0,0,1}(t)+(1-\sigma)\lambda\delta t P_{0,0,0}(t)+\mu_1\delta t(1-\lambda\delta t)(1-\mu_2\delta t)P_{0,1,1}(t)\\
P_{0,1,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{0,1,1}(t)+\sigma\lambda\delta t(1-\mu_2\delta t)P_{0,0,1}(t)\\
&\quad +(1-\sigma)\lambda\delta t(1-\mu_1\delta t)P_{0,1,0}(t)+\mu_1\delta t(1-\lambda\delta t)(1-\mu_2\delta t)P_{1,1,1}(t)\\
&\quad +\mu_2\delta t(1-\lambda\delta t)(1-\mu_1\delta t)P_{1,1,1}(t)\\
P_{n,1,1}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_1\delta t)(1-\mu_2\delta t)P_{n,1,1}(t)+\lambda\delta t(1-\mu_1\delta t)(1-\mu_2\delta t)P_{n-1,1,1}(t) \qquad n\geq 1\\
&\quad +\mu_1\delta t(1-\lambda\delta t)(1-\mu_2\delta t)P_{n+1,1,1}(t)+\mu_2\delta t(1-\lambda\delta t)(1-\mu_1\delta t)P_{n+1}(t)
\end{aligned} \tag{5.5}$$

These can be rearranged to give;

$$\begin{aligned}
\frac{dP_{0,0,0}(t)}{dt} &= -\lambda P_{0,0,0}(t)+\mu_1 P_{0,1,0}(t)+\mu_2 P_{0,0,1}(t)\\
\frac{dP_{0,1,0}(t)}{dt} &= -(\lambda+\mu_1)P_{0,1,0}(t)+\sigma\lambda P_{0,0,0}(t)+\mu_2 P_{0,1,1}(t)\\
\frac{dP_{0,0,1}(t)}{dt} &= -(\lambda+\mu_2)P_{0,0,1}(t)+(1-\sigma)\lambda P_{0,0,0}(t)+\mu_1 P_{0,1,1}(t)\\
\frac{dP_{0,1,1}(t)}{dt} &= -(\lambda+\mu_1+\mu_2)P_{0,1,1}(t)+\lambda P_{0,0,1}(t)+\lambda P_{0,1,0}(t)+\mu_1 P_{1,1,1}(t)+\mu_2 P_{1,1,1}(t)\\
\frac{dP_{n,1,1}(t)}{dt} &= -(\lambda+\mu_1+\mu_2)P_{n,1,1}(t)+\lambda P_{n-1,1,1}(t)+(\mu_1+\mu_2)P_{n+1,1,1}(t) \qquad n\geq 1
\end{aligned} \tag{5.6}$$

By setting the differential equal to zero the steady state solutions can be found;

$$\begin{aligned}
\lambda P_{0,0,0} &= \mu_1 P_{0,1,0} + \mu_2 P_{0,0,1} \\
(\lambda+\mu_1) P_{0,1,0} &= \sigma\lambda P_{0,0,0} + \mu_2 P_{0,1,1} \\
(\lambda+\mu_2) P_{0,0,1} &= (1-\sigma)\lambda P_{0,0,0} + \mu_1 P_{0,1,1} \\
(\lambda+\mu_1+\mu_2) P_{0,1,1} &= \lambda P_{0,0,1} + \lambda P_{0,1,0} + (\mu_1+\mu_2) P_{1,1,1} \\
(\lambda+\mu_1+\mu_2) P_{n,1,1} &= \lambda P_{n-1,1,1} + (\mu_1+\mu_2) P_{n+1,1,1} \qquad n \geq 1
\end{aligned} \tag{5.7}$$

By setting $\mu_1 = \mu_2$ and $n = 0$;

$$\begin{aligned}
\lambda P_{0,0,0} &= \mu\left(P_{0,1,0} + P_{0,0,1}\right) \\
(\lambda+\mu) P_{0,1,0} &= \sigma\lambda P_{0,0,0} + \mu P_{0,1,1} \\
(\lambda+\mu) P_{0,0,1} &= (1-\sigma)\lambda P_{0,0,0} + \mu P_{0,1,1} \\
(\lambda+2\mu) P_{0,1,1} &= \lambda\left(P_{0,0,1} + P_{0,1,0}\right)
\end{aligned} \tag{5.8}$$

Comparing this case with the system described in section 5.1.2 we see that;

$$\begin{aligned}
P_0 &= P_{0,0,0} \\
P_1 &= P_{0,0,1} + P_{0,1,0} \\
P_2 &= P_{0,1,1}
\end{aligned} \tag{5.9}$$

Equations(5.9) relate the two systems just described. The probabilities on the left hand side are from the system that does not distinguish which routes are occupied. On the right hand side the occupied route is noted. By substituting the values from equations (5.9) into equations (5.8) the equations of the original two route system are obtained, (equations (5.4)).

5.2 Two Phases

5.2.1 Two Phase, Two Route System

To be able to look at blocking equations with more than 1 service route, a new phase needs to be added to the system describe above. Once this new phase has been added, the description of how customers can move between phases must be clarified. In most of the current research involving multiple routes and multiple phases, once customers have finished their service they move into a queueing facility for the second phase. If this queueing facility becomes full and a customer completes their service, they become blocked. In the case when there does not exist a queue between the two service phases, the customers behave as if there still is. This means that once a customer has completed their service they will try to go to any server in the next phase and will only become blocked if all servers in the next phase are busy. Many papers are produced on this pooled queue type of blocking; (Akyildiz 1989) is an example. Many of these papers attempt to either show that these systems are of product form or can be approximated by product form. This is the property that says that each individual set of servers is independent, and the probability of the system being in a specific state is the product of each set of servers being in the appropriate state, as discussed in Chapter 1.

Within this chapter, we shall consider a new method of dealing with the customer will be considered. When a customer leaves a phase they have a unique specific destination in the subsequent phase, if this destination is busy then the customer becomes blocked. This allows for the first phase to continue serving customers even if some routes in that phase are blocked. The fact that customers can still be processed and are able to leave upon completion of their service whilst other customers are blocked, leads to a different system which maybe more appropriate to a hospital environment, where one patient could remain blocking a bed even though their service was completed before another patient who has left the department before the. This system can be seen in Figure 5.4.

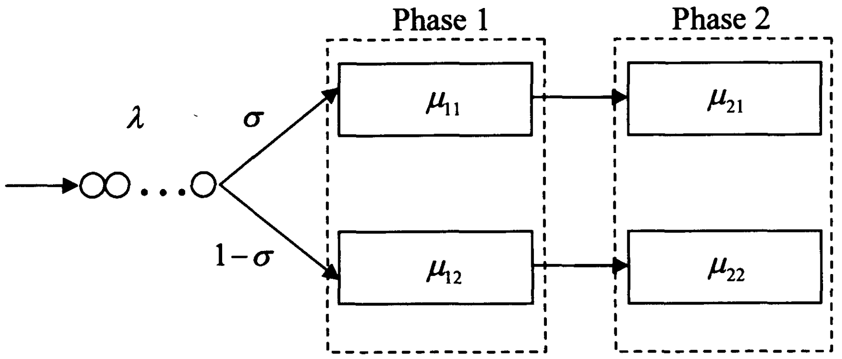


Figure 5.4 - A two phase, two route queueing system

In this system μ_{ij} is the service rate at *i*,j where *i* is the phase and *j* is the route. In this system, it is not possible for the customer to change routes once they have started and as with previous system, the probability of a customer selecting Route 1 when both are available is σ, however if there is already a customer being served in one of the routes in Phase 1 then the arriving customer will go to the empty server. If the arriving customer finds both service points occupied, then the customer will join the queue.

A flaw with using this model is that if a customer arrives to find both service points empty then they will go down Route 1 or 2 randomly (with probability σ or $1-\sigma$). However, if there is a customer in Phase 2, Route 1 it might be preferable for the arriving customer to join Route 2 to reduce the chance of them becoming blocked. It would be possible to build this into the model but it was felt that it would be over complicating the model and the added flexibility would not be very significant.

When trying to adapt the current notation for more phases it may get confusing which number is equivalent to which service point. So some further, simpler, notation will be introduced. In the previous blocking chapter it could be seen that there are 5 possible states that each route can be in, assuming a queue size of zero. These are listed in Figure 5.5. A dot indicates a customer and a "B" after a dot indicates that the customer is blocked.

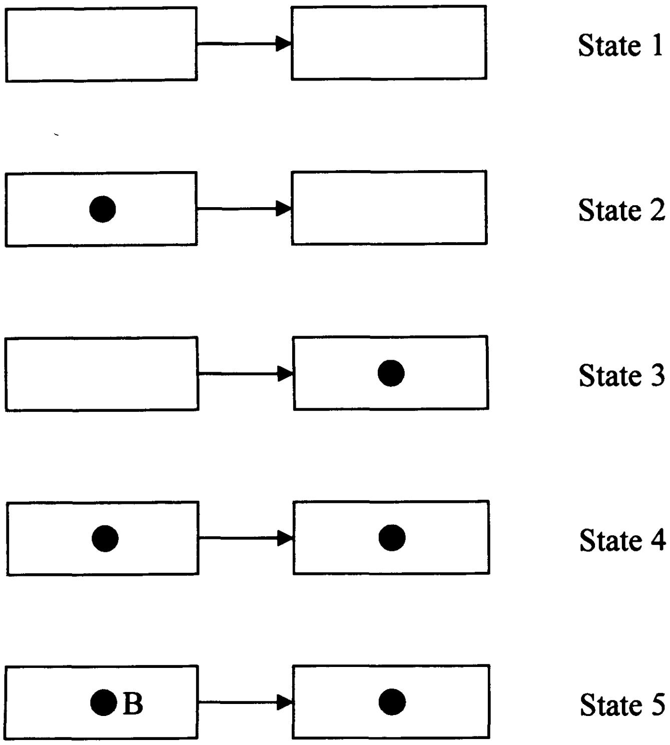


Figure 5.5 - Five possible route states

For the two route, two phase system that is now being considered, the notation will be used, $P_{n_1,n_2}(t)$. This is the probability of the first route being in State n_1 and Route 2 being in State n_2, at time t. As can be seen, the restriction on n_i is that it must be greater than zero and less than or equal to five. If more routes are to be added then a further n is added to the suffices.

It can be seen that for each extra route that is added to this system where there is no queueing space, the number of possible states increases by a factor of five. First consider the first phase, ignoring the blocking cases. With n routes there can be from 0 to n customers present in Phase 1. Counting the times that each number of customers can occur; $\begin{pmatrix} n \\ 0 \end{pmatrix}$ with no customers, $\begin{pmatrix} n \\ 1 \end{pmatrix}$ with one customer, $\begin{pmatrix} n \\ 2 \end{pmatrix}$ with two customers, ... and $\begin{pmatrix} n \\ n \end{pmatrix}$ with n customers. So in total there are $\sum_{k=0}^{n}\begin{pmatrix} n \\ k \end{pmatrix} = 2^n$ possible states for Phase 1, when excluding the chance of blocking. The same logic can be applied to the second phase which gives 2^n different arrangement of customers in the second phase. So if no blocking occurred, the total number of different states would be the product of the number of arrangements in each phase, namely $\left(2^n\right)^2$. Next take into account the blocked cases. If one route is blocked, then n-1 are unblocked, and these remaining n-1 routes would have $\left(2^{(n-1)}\right)^2$ arrangements available. If only one of the routes is blocked then it is of course possible that it could be any, single, one of them. So the there are $\binom{n}{1}$ different times this could happen. It can be worked out similarly when there are two blocked routes and so on, by summing these different arrangements, the total number of different states that n route system has:

$$\begin{aligned}\left(2^n\right)^2+\sum_{k=1}^{n}\binom{n}{k}\left(2^{2(n-k)}\right) &= \sum_{k=0}^{n}\binom{n}{k}\left(2^{2(n-k)}\right)\\ &= \sum_{k=0}^{n}\binom{n}{k}\left(2^{2(n-k)}\right)\left(1^k\right)\\ &= \left(2^2+1\right)^n\\ &= 5^n\end{aligned}$$

It can also be seen intuitively by the fact that there are five possible states for each route, and that for n routes there should be 5^n different states. For every route that is added to the amount of states is multiplied by five, there will be 25 equations when there are 2 routes and 125 for 3 routes. The two route case will be considered for the rest of this chapter, but the methods can be used on the n route system if the reader wishes.

5.2.2 Zero Queue Size

Time dependent equations for the two route, two phase system with a queue size of zero can now be set up;

$$P_{1,1}(t+\delta t)=(1-\lambda\delta t)P_{1,1}(t)+\mu_{21}\delta t(1-\lambda\delta t)P_{3,1}(t)+\mu_{22}\delta t(1-\lambda\delta t)P_{1,3}(t)$$

$$\begin{aligned}P_{2,1}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{11}\delta t)P_{2,1}(t)+\sigma\lambda\delta tP_{1,1}(t)+\mu_{21}\delta t(1-\lambda\delta t)(1-\mu_{11}\delta t)P_{4,1}(t)\\&+\mu_{22}\delta t(1-\lambda\delta t)(1-\mu_{11}\delta t)P_{3,2}(t)\end{aligned}$$

$$\begin{aligned}P_{1,2}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{12}\delta t)P_{2,2}(t)+(1-\sigma)\lambda\delta tP_{1,1}(t)+\mu_{21}\delta t(1-\lambda\delta t)(1-\mu_{12}\delta t)P_{3,2}(t)\\&+\mu_{22}\delta t(1-\lambda\delta t)(1-\mu_{12}\delta t)P_{1,4}(t)\end{aligned}$$

$$\begin{aligned}P_{2,2}(t+\delta t)=&(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)P_{2,2}(t)+\lambda\delta t(1-\mu_{11}\delta t)P_{2,1}(t)+\lambda\delta t(1-\mu_{11}\delta t)P_{2,1}(t)\\&+\lambda\delta t(1-\mu_{12}\delta t)P_{1,2}(t)+\mu_{21}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)P_{4,2}(t)\\&+\mu_{22}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)P_{2,4}(t)\end{aligned}$$

$$\begin{aligned}P_{3,1}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{3,1}(t)+\mu_{11}\delta t(1-\lambda\delta t)P_{2,1}(t)+\mu_{21}\delta t(1-\lambda\delta t)P_{5,1}(t)\\&+\mu_{22}\delta t(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{3,3}(t)\end{aligned}$$

$$\begin{aligned}P_{1,3}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{22}\delta t)P_{1,3}(t)+\mu_{12}\delta t(1-\lambda\delta t)P_{1,2}(t)+\mu_{22}\delta t(1-\lambda\delta t)P_{1,5}(t)\\&+\mu_{21}\delta t(1-\lambda dt)(1-\mu_{22}\delta t)P_{3,3}(t)\end{aligned}$$

$$\begin{aligned}P_{4,1}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{4,1}(t)+\mu_{22}\delta t(1-\lambda\delta t)(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{4,3}(t)\\&+\sigma\lambda\delta t(1-\mu_{21}\delta t)P_{3,1}(t)\end{aligned}$$

$$\begin{aligned}P_{3,2}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{3,2}(t)+(1-\sigma)\lambda\delta t(1-\mu_{21}\delta t)P_{3,1}(t)\\&+\mu_{2,1}\delta t(1-\mu_{12}\delta t)P_{5,2}(t)+\mu_{11}\delta t(1-\mu_{12}\delta t)P_{2,2}(t)\\&+\mu_{22}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{3,4}(t)\end{aligned}$$

$$\begin{aligned}P_{2,3}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{2,3}(t)+\sigma\lambda\delta t(1-\mu_{22}\delta t)P_{1,3}(t)\\&+\mu_{22}\delta t(1-\mu_{11}\delta t)P_{2,5}(t)+\mu_{12}\delta t(1-\mu_{11}\delta t)P_{2,2}(t)\\&+\mu_{21}\delta t(1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{4,3}(t)\end{aligned}$$

$$\begin{aligned}P_{1,4}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{1,4}(t)+(1-\sigma)\lambda\delta t(1-\mu_{21}\delta t)P_{1,3}(t)\\&+\mu_{21}\delta t(1-\lambda dt)(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{3,4}(t)\end{aligned}$$

$$\begin{aligned}P_{5,1}(t+\delta t)=&(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{5,1}(t)+\mu_{11}\delta t(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{4,1}(t)\\&+\mu_{22}\delta t(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{5,3}(t)\end{aligned}$$

$$\begin{aligned}
P_{1,5}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_{22}\delta t)P_{1,5}(t)+\mu_{12}\delta t(1-\lambda\delta t)(1-\mu_{22}\delta t)P_{1,4}(t)\\
&\quad +\mu_{21}\delta t(1-\lambda\delta t)(1-\mu_{22})\delta t P_{3,5}(t)\\
P_{4,2}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{4,2}(t)+\lambda\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{4,1}(t)\\
&\quad +\lambda\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{3,2}(t)+\mu_{22}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{4,4}(t)\\
P_{2,4}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{2,4}(t)+\lambda\delta t(1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{2,3}(t)\\
&\quad +\lambda\delta t(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{1,4}(t)+\mu_{21}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{4,4}(t)\\
P_{5,2}(t+\delta t) &= (1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{5,2}(t)+\mu_{11}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{4,2}(t)\\
&\quad +\lambda\delta t(1-\mu_{21}\delta t)P_{5,1}(t)+\mu_{22}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{5,4}(t)\\
P_{2,5}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{2,5}(t)+\mu_{12}\delta t(1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{2,4}(t)\\
&\quad +\lambda\delta t(1-\mu_{22}\delta t)P_{1,5}(t)+\mu_{21}\delta t(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{4,5}(t)\\
P_{3,3}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,3}(t)+\mu_{11}\delta t(1-\lambda\delta t)(1-\mu_{22}\delta t)P_{2,3}(t)\\
&\quad +\mu_{12}\delta t(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{3,2}(t)+\mu_{21}\delta t(1-\lambda\delta t)(1-\mu_{22}\delta t)P_{5,3}(t)\\
&\quad +\mu_{22}\delta t(1-\lambda\delta t)(1-\mu_{21}\delta t)P_{3,5}(t)\\
P_{4,3}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,3}(t)\\
&\quad +\sigma\lambda\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,3}(t)+\mu_{12}\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{4,2}(t)\\
&\quad +\mu_{22}\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{4,5}(t)\\
P_{3,4}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_{1,2}\delta t)(1-\mu_{2,1}\delta t)(1-\mu_{2,2}\delta t)P_{3,4}(t)\\
&\quad +\mu_{21}\delta t(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{5,4}(t)+\mu_{11}\delta t(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{2,4}(t)\\
&\quad +(1-\sigma)\lambda\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{33}(t)\\
P_{5,3}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{5,3}(t)+\mu_{12}\delta t(1-\mu_{21}\delta t)P_{5,2}(t)\\
&\quad +\mu_{22}\delta t(1-\mu_{21}\delta t)P_{5,5}(t)+\mu_{11}\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,3}(t)\\
P_{3,5}(t+\delta t) &= (1-\lambda\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,5}(t)+\mu_{11}\delta t(1-\mu_{21}\delta t)P_{2,5}(t)\\
&\quad +\mu_{21}\delta t(1-\mu_{22}\delta t)P_{5,5}(t)+\mu_{12}\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,4}(t)
\end{aligned}$$

$$
\begin{aligned}
P_{4,4}(t+\delta t) &= (1-\mu_{12}\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,4}(t) \\
&\quad +\lambda\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,4}(t) \\
&\quad +\lambda\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,3}(t) \\
P_{5,4}(t+\delta t) &= (1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{5,4}(t)+\lambda\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{5,3}(t) \\
&\quad +\mu_{11}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,4}(t) \\
P_{4,5}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,5}(t)+\lambda\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,5}(t) \\
&\quad +\mu_{12}\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,4}(t) \\
P_{5,5}(t+\delta t) &= (1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{5,5}(t)+\mu_{11}\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{4,5}(t) \\
&\quad +\mu_{12}\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{5,4}(t)
\end{aligned}
\tag{5.10}
$$

There are a total of 25 equations for the 25 different states that it is possible for this system to be in. The computer program Maple was set up to solve them, the values could be solved in terms of the arrival and service rates for each individual phase of each route as the computing power required when solving these equations is large. However when all of the service rates are equal the result is more simply achieved. In this case the steady state solution of $P_{1,1}$ is;

$$
P_{1,1}=\frac{8\mu^4\left(36\mu^3+56\lambda\mu^2+23\lambda^2\mu+3\lambda^3\right)}{288\mu^7+1024\lambda\mu^6+1656\lambda^2\mu^5+1688\lambda^3\mu^4+1156\lambda^4\mu^3+480\lambda^5\mu^2+105\lambda^6\mu+9\lambda^7}
\tag{5.11}
$$

All of the other steady state probabilities are multiples of $P_{1,1}$. A solution of transient behaviour has been created, starting at time = 0 and increasing by 0.01 each step. This validates the solutions for specific arrival and service rates, the behaviour can be seen in Figure 5.6.

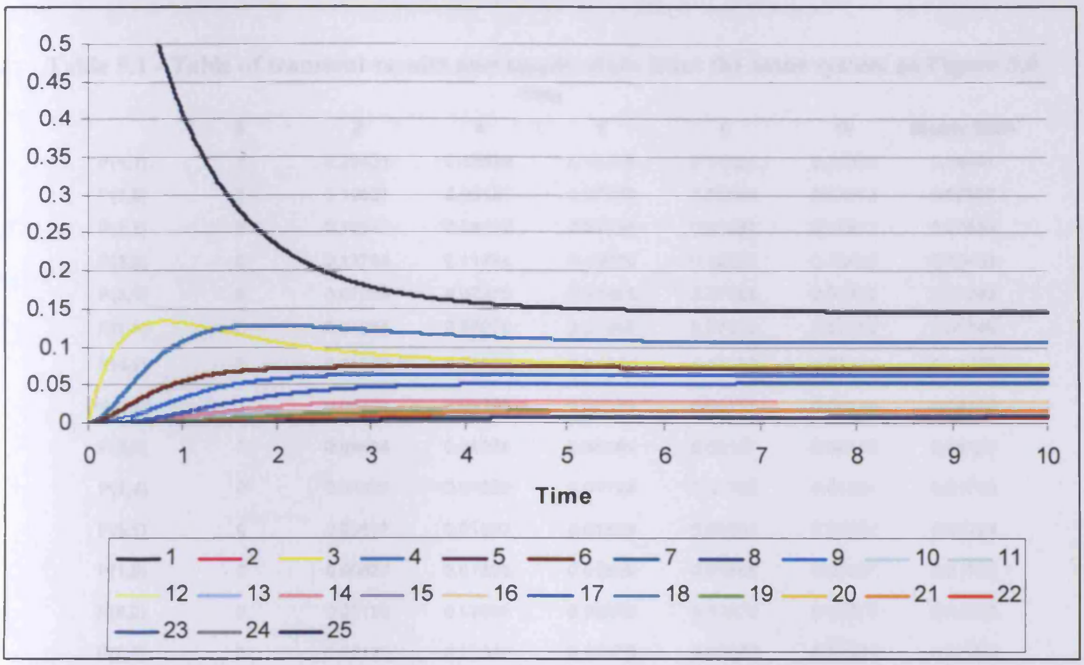


Figure 5.6 - The transient solutions of a two phase, two route system with service and arrival rates equal to one and delta t equal to 0.01

As it is difficult to read the final values of this graph they have been summarised in the Table 5.1 along with the theoretical steady state solutions produced by Maple.

Table 5.1 - Table of transient results and steady state from the same system as Figure 5.6

	Time						
	0	2	4	6	8	10	Steady State
P(1,1)	1	0.23421	0.15828	0.14778	0.14621	0.14598	0.14497
P(1,2)	0	0.10527	0.08197	0.07720	0.07629	0.07612	0.07552
P(2,1)	0	0.10527	0.08197	0.07720	0.07629	0.07612	0.07552
P(2,2)	0	0.12766	0.11334	0.10679	0.10521	0.10489	0.10434
P(3,1)	0	0.07085	0.07276	0.07256	0.07252	0.07252	0.07249
P(1,3)	0	0.07085	0.07276	0.07256	0.07252	0.07252	0.07249
P(4,1)	0	0.01329	0.01682	0.01724	0.01730	0.01731	0.01735
P(3,2)	0	0.05694	0.06266	0.06189	0.06157	0.06150	0.06120
P(2,3)	0	0.05694	0.06266	0.06189	0.06157	0.06150	0.06120
P(1,4)	0	0.01329	0.01682	0.01724	0.01730	0.01731	0.01735
P(5,1)	0	0.00637	0.01383	0.01629	0.01686	0.01697	0.01724
P(1,5)	0	0.00637	0.01383	0.01629	0.01686	0.01697	0.01724
P(4,2)	0	0.02123	0.02846	0.02878	0.02872	0.02870	0.02882
P(2,4)	0	0.02123	0.02846	0.02878	0.02872	0.02870	0.02882
P(5,2)	0	0.01000	0.02252	0.02595	0.02666	0.02679	0.02720
P(2,5)	0	0.01000	0.02252	0.02595	0.02666	0.02679	0.02720
P(3,3)	0	0.03541	0.04964	0.05166	0.05197	0.05201	0.05222
P(4,3)	0	0.00867	0.01460	0.01554	0.01568	0.01570	0.01582
P(3,4)	0	0.00867	0.01460	0.01554	0.01568	0.01570	0.01582
P(5,3)	0	0.00471	0.01330	0.01613	0.01677	0.01689	0.01712
P(3,5)	0	0.00471	0.01330	0.01613	0.01677	0.01689	0.01712
P(4,4)	0	0.00347	0.00705	0.00766	0.00775	0.00776	0.00791
P(5,4)	0	0.00183	0.00626	0.00776	0.00809	0.00815	0.00834
P(4,5)	0	0.00183	0.00626	0.00776	0.00809	0.00815	0.00834
P(5,5)	0	0.00093	0.00532	0.00744	0.00797	0.00808	0.00834

It can be seen that these transient solutions are tending towards the steady state solution for these values of $\lambda, \mu_{11}, \mu_{12}, \mu_{21}$ and μ_{22}. The restrictions on the values of the parameters for the steady state solutions to exist will be computed in section 5.2.4.

5.2.3 Drip Feed Model

To analyse the two route system further, the next system to be considered will be the 'drip feed' situation. In this system the first phase of neither server can be empty, as there is an imaginary infinite queue of customers in front of the first phase, and as soon as a server in Phase 1 becomes free a customer commences their service. The notation will remain the same as before, using a number between one and five to indicate what state each route is in at any particular time. In this 'drip feed' example, there are only three phases available; 2, 4 and 5. These numbers will be reallocated so that they are sequential. In this system State 1 will be equivalent to State 2 in the previous example, State 2 is equivalent to State 4 and State 3 is equivalent to State 5. These can be seen in Figure 5.7

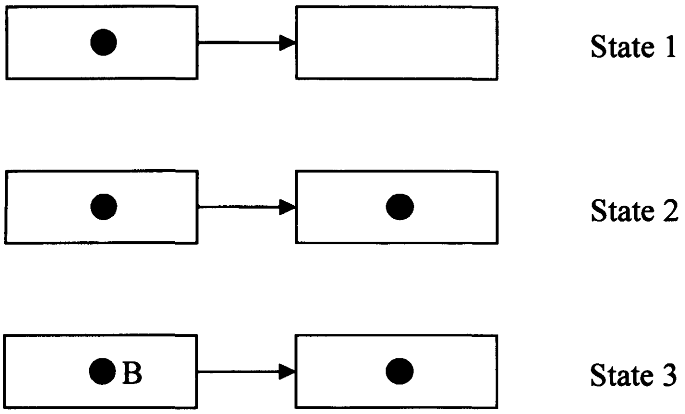


Figure 5.7 - Drip feed states

The time dependent equations follow;

$$
\begin{aligned}
P_{1,1}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{12})P_{1,1}(t)+\mu_{21}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)P_{2,1}(t) \\
&\quad +\mu_{22}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)P_{1,2}(t) \\
P_{2,1}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{2,1}(t)+\mu_{21}\delta t(1-\mu_{12}\delta t)P_{3,1}(t) \\
&\quad +\mu_{22}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{2,2}(t)+\mu_{11}\delta t(1-\mu_{12}\delta t)P_{1,1}(t) \\
P_{1,2}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{1,2}(t)+\mu_{22}\delta t(1-\mu_{11}\delta t)P_{1,3}(t) \\
&\quad +\mu_{21}\delta t(1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{2,2}(t)+\mu_{12}\delta t(1-\mu_{11}\delta t)P_{1,1}(t) \\
P_{2,2}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{2,2}(t) \\
&\quad +\mu_{21}\delta t(1-\mu_{12}\delta t)(1-\mu_{22}\delta t)P_{3,2}(t)+\mu_{22}\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{1,2}(t) \\
&\quad +\mu_{11}\delta t(1-\mu_{12}\delta t)(1-\mu_{22}\delta)P_{2,1}(t)+\mu_{12}\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)P_{2,1}(t) \\
P_{3,1}(t+\delta t) &= (1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{3,1}(t)+\mu_{11}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{2,1}(t) \\
&\quad +\mu_{22}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)P_{3,2}(t) \\
P_{1,3}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{1,3}(t)+\mu_{12}\delta t(1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{1,2}(t) \\
&\quad +\mu_{21}\delta t(1-\mu_{11}\delta t)(1-\mu_{22}\delta t)P_{2,3}(t) \\
P_{3,2}(t+\delta t) &= (1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,2}(t) \\
&\quad +\mu_{11}\delta t(1-\mu_{12}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{2,2}(t)+\mu_{22}\delta t(1-\mu_{21}\delta t)P_{3,3}(t) \\
&\quad +\mu_{12}\delta t(1-\mu_{21}\delta t)P_{3,1}(t) \\
P_{2,3}(t+\delta t) &= (1-\mu_{11}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{2,3}(t) \\
&\quad +\mu_{12}\delta t(1-\mu_{11}\delta t)(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{2,2}(t)+\mu_{21}\delta t(1-\mu_{22}\delta t)P_{3,3}(t) \\
&\quad +\mu_{11}\delta t(1-\mu_{22}\delta t)P_{1,3}(t) \\
P_{3,3}(t+\delta t) &= (1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,3}(t)+\mu_{12}\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{3,2}(t) \\
&\quad +\mu_{11}\delta t(1-\mu_{21}\delta t)(1-\mu_{22}\delta t)P_{2,3}
\end{aligned}
\tag{5.12}
$$

These can be rearranged to give the steady state equations of the form below;

$$
\begin{aligned}
(\mu_{11}+\mu_{12})P_{1,1} &= \mu_{21}P_{2,1}+\mu_{22}P_{1,2}\\
(\mu_{11}+\mu_{12}+\mu_{21})P_{2,1} &= \mu_{21}P_{3,1}+\mu_{22}P_{2,2}+\mu_{11}P_{1,1}\\
(\mu_{11}+\mu_{12}+\mu_{22})P_{1,2} &= \mu_{22}P_{1,3}+\mu_{21}P_{2,2}+\mu_{12}P_{1,1}\\
(\mu_{11}+\mu_{12}+\mu_{21}+\mu_{22})P_{2,2} &= \mu_{21}P_{3,2}+\mu_{22}P_{2,3}+\mu_{11}P_{1,2}+\mu_{12}P_{2,1}\\
(\mu_{12}+\mu_{21})P_{3,1} &= \mu_{11}P_{2,1}+\mu_{22}P_{3,2}\\
(\mu_{11}+\mu_{22})P_{1,3} &= \mu_{12}P_{1,2}+\mu_{21}P_{2,3}\\
(\mu_{12}+\mu_{21}+\mu_{22})P_{3,2} &= \mu_{11}P_{2,2}+\mu_{22}P_{3,3}+\mu_{12}P_{3,1}\\
(\mu_{11}+\mu_{21}+\mu_{22})P_{2,3} &= \mu_{12}P_{2,2}+\mu_{21}P_{3,3}+\mu_{11}P_{1,3}\\
(\mu_{21}+\mu_{22})P_{3,3} &= \mu_{12}P_{3,2}+\mu_{11}P_{2,3}
\end{aligned}
\tag{5.13}
$$

By using the fact that the sum of the probabilities has to equal 1, then these equations can be solved in terms of $\lambda, \mu_{11}, \mu_{12}, \mu_{21}$ and μ_{22}. These equations have been solved using Maple to give;

$$P_{1,1} = \frac{\mu_{21}^2 \mu_{22}^2}{D}$$

$$P_{2,1} = \frac{\mu_{11}\mu_{21}\mu_{22}^2}{D}$$

$$P_{1,2} = \frac{\mu_{12}\mu_{21}^2\mu_{22}}{D}$$

$$P_{2,2} = \frac{\mu_{11}\mu_{12}\mu_{21}\mu_{22}}{D}$$

$$P_{3,1} = \frac{\mu_{11}^2 \mu_{22}^2}{D}$$

$$P_{1,3} = \frac{\mu_{12}^2 \mu_{21}^2}{D}$$

$$P_{3,2} = \frac{\mu_{11}^2\mu_{12}\mu_{22}}{D}$$

$$P_{2,3} = \frac{\mu_{11}\mu_{12}^2\mu_{21}}{D} \qquad (5.14)$$

$$P_{3,3} = \frac{\mu_{11}^2 \mu_{12}^2}{D}$$

where D has the value $\left(\mu_{11}^2 + \mu_{11}\mu_{21} + \mu_{21}^2\right)\left(\mu_{12}^2 + \mu_{12}\mu_{22} + \mu_{22}^2\right)$. These are not entirely unexpected results. These solutions can be seen to be the product of appropriate equations as seen in Chapter 3 equation (3.20), the drip feed steady state solutions for a single route. As these routes are independent the probability of different routes being in a particular state at any specific time is just the probability of one route being in the correct state multiplied by the probability that the other route is in the correct state.

When all the values for the service rates are equal the result for the steady state probabilities are;

$$P_{1,1} = P_{1,2} = P_{1,3} = P_{2,1} = P_{2,2} = P_{2,3} = P_{3,1} = P_{3,2} = P_{3,3} = \frac{1}{9} \qquad (5.15)$$

For the single route case the state probabilities are equal to a third under the same circumstances.

5.2.4 Maximum Utilisation

It is possible to imagine both the case where the maximum arrival rate will increase in a linear fashion and where the rate of increase will alter as more servers are added. For example, if a single route case has a maximum utilisation of U, then it might be expected that if another server is added, the rate would increase to $2U$ and for this to continue to as more and more servers are added. It is also possible to suppose that, as another server is added, the chance that the system will become completely blocked and therefore unproductive, is reduced. This could result in the utilisation being greater than $2U$. The aim of this section is therefore to decide if the rate will increase in a linear fashion or if as more routes are added, the rate at which customers are dealt with increases above that linear fashion.

The maximum utilisation of this two route model will have similar properties to the maximum utilisation for the single route case. The maximum rate at which a service point in Phase 1 can process customers is less than the average service rate, as some customers will be blocked, making the first service point run more slowly. The arrival rate, λ, will have to be less than the maximum throughput of the first phase. In the single route blocking system this was first investigated by using a 'drip feed' type system. It, originally, proved unsuccessful in finding the capacity of the first phase, but when three nodes were considered a method of finding the maximum throughput of the first phase was found using the drip feed results. This method will be used to find the maximum utilisation and hence the maximum arrival that this system can deal with. The original method for calculating the maximum throughput will also be considered.

The drip feed method of finding the average time a customer spends in the first phase involves summing the proportion of times that the first phase is working, multiplied by the service rate. This gives the maximum rate at which Phase 1 can process customers, so in effect the maximum average rate which customers can arrive with out a queue growing uncontrollably. In this two route system the states that have Phase 1 working are; (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1) and (3,2). Where the first number indicates the state of the first route and second number is the state of the second route. All of the states other than (3,3) have at least one of the routes serving a customer rather than being blocked. Not all of these states are as 'productive' as others. For example state (1,1) customers in both routes are in a service whereas any of the states that contain a three have one route blocked. It is obviously important to take this into account when finding the maximum rate at which Phase 1 can serve.

$$\begin{aligned}
\lambda_{\max} &= P_{1,1}(\mu_{11}+\mu_{12}) + P_{2,1}(\mu_{11}+\mu_{12}) + P_{1,2}(\mu_{11}+\mu_{12}) + P_{2,2}(\mu_{11}+\mu_{12}) \\
&\quad + \mu_{12}P_{3,1} + \mu_{11}P_{1,3} + \mu_{12}P_{3,2} + \mu_{11}P_{2,3} \\
&= \frac{\begin{array}{c}\mu_{11}\mu_{21}\left(\mu_{21}\mu_{22}^2+\mu_{11}\mu_{22}^2+\mu_{12}\mu_{21}\mu_{22}+\mu_{11}\mu_{12}\mu_{22}+\mu_{12}^2\mu_{21}+\mu_{11}\mu_{12}^2\right)\\ +\mu_{12}\mu_{22}\left(\mu_{21}^2\mu_{22}+\mu_{11}\mu_{21}\mu_{22}+\mu_{12}\mu_{21}^2+\mu_{11}\mu_{12}\mu_{21}+\mu_{11}^2\mu_{22}+\mu_{11}^2\mu_{12}\right)\end{array}}{\left(\mu_{11}^2+\mu_{11}\mu_{21}+\mu_{21}^2\right)\left(\mu_{12}^2+\mu_{12}\mu_{22}+\mu_{22}^2\right)} \\
&= \frac{\begin{array}{c}\mu_{11}\mu_{21}\left(\mu_{22}^2(\mu_{21}+\mu_{11})+\mu_{12}\mu_{22}(\mu_{21}+\mu_{11})+\mu_{12}^2(\mu_{21}+\mu_{11})\right)\\ +\mu_{12}\mu_{22}\left(\mu_{21}^2(\mu_{22}+\mu_{12})+\mu_{11}\mu_{21}(\mu_{22}+\mu_{12})+\mu_{11}^2(\mu_{22}+\mu_{12})\right)\end{array}}{\left(\mu_{11}^2+\mu_{11}\mu_{21}+\mu_{21}^2\right)\left(\mu_{12}^2+\mu_{12}\mu_{22}+\mu_{22}^2\right)} \\
&= \frac{\mu_{11}\mu_{21}(\mu_{21}+\mu_{11})\left(\mu_{22}^2+\mu_{12}\mu_{22}+\mu_{12}^2\right)+\mu_{12}\mu_{22}(\mu_{22}+\mu_{12})\left(\mu_{21}^2+\mu_{11}\mu_{21}+\mu_{11}^2\right)}{\left(\mu_{11}^2+\mu_{11}\mu_{21}+\mu_{21}^2\right)\left(\mu_{12}^2+\mu_{12}\mu_{22}+\mu_{22}^2\right)} \\
&= \frac{\mu_{11}\mu_{21}(\mu_{21}+\mu_{11})}{\left(\mu_{11}^2+\mu_{11}\mu_{21}+\mu_{21}^2\right)} + \frac{\mu_{12}\mu_{22}(\mu_{22}+\mu_{12})}{\left(\mu_{12}^2+\mu_{12}\mu_{22}+\mu_{22}^2\right)}
\end{aligned} \tag{5.16}$$

When all the service rates in Phase i are set equal to μ_i;

$$\frac{\mu_1\mu_2(\mu_2+\mu_1)}{\mu_1^2+\mu_1\mu_2+\mu_2^2}+\frac{\mu_1\mu_2(\mu_2+\mu_1)}{\mu_1^2+\mu_1\mu_2+\mu_2^2} \tag{5.17}$$

Both elements in this expression are now clearly equal; they are also equal to the value of $\lambda_{\max}$ in Chapter 3, equation (3.26). This can also be seen using the other method of calculation. Consider the single Route 2 phase case considered in Chapter 3. The mean time spent in Phase 1 was calculated in the following way.

$$\frac{1}{\mu_1}+P(B)\frac{1}{\mu_2} \tag{5.18}$$

where μ_1 the rate of service is in the first phase, μ_2 is the rate of service in the second phase and $P(B)$ is the probability of a customer becoming blocked. To adapt this for a two server case, the same notation for the service rates will used as in the rest of this chapter and $P(B_i)$ will be used for the probability of a customer becoming blocked in route i. So using the equation from the single route system it is possible to say that Route 1 in the two route system has an average process time of;

$$\frac{1}{\mu_{11}}+P(B_1)\frac{1}{\mu_{21}} \tag{5.19}$$

Route 2;

$$\frac{1}{\mu_{12}}+P(B_2)\frac{1}{\mu_{22}} \tag{5.20}$$

As the customers arrive they have a probability, σ, of going to Route 1 but this is only when the system is empty so this is not taken into consideration. So the average time Phase 1 will be serving for is;

$$\frac{1}{\mu_{11}}+\frac{1}{\mu_{12}}+P(B_1)\frac{1}{\mu_{21}}+P(B_2)\frac{1}{\mu_{22}} \tag{5.21}$$

Using the single route example again, it has been shown in Chapter 3 that $P(B)=\frac{\mu_1}{\mu_1+\mu_2}$ so it can be seen that $P(B_1)=\frac{\mu_{11}}{\mu_{11}+\mu_{21}}$ and $P(B_2)=\frac{\mu_{12}}{\mu_{12}+\mu_{22}}$. So the average time spent in Phase 1 is;

$$\begin{aligned}&\frac{\mu_{21}(\mu_{11}+\mu_{21})+\mu_{11}^2}{\mu_{11}\mu_{21}(\mu_{11}+\mu_{21})}+\frac{\mu_{12}(\mu_{12}+\mu_{22})+\mu_{22}^2}{\mu_{12}\mu_{22}(\mu_{12}+\mu_{22})}\\&=\frac{\mu_{21}^2+\mu_{11}\mu_{21}+\mu_{11}^2}{\mu_{11}\mu_{21}(\mu_{11}+\mu_{21})}+\frac{\mu_{12}^2+\mu_{12}\mu_{22}+\mu_{22}^2}{\mu_{12}\mu_{22}(\mu_{12}+\mu_{22})}\end{aligned} \tag{5.22}$$

which can be seen to be the reciprocal of equation (5.16). When all the service rates are equal;

$$\begin{aligned}\rho_{\max}&=\frac{\lambda_{\max}}{\mu}\\&=2\left(\frac{2}{3}\right)\end{aligned} \tag{5.23}$$

This can be seen to be twice the maximum rate which customers can arrive at in the single route case, as expected from equations (5.22) and (5.17). Figure 5.8 displays what happens to the maximum process rate of customers in Phase 1 as the ratio of service rate in Phase 1 to the service rate in Phase 2 increases.

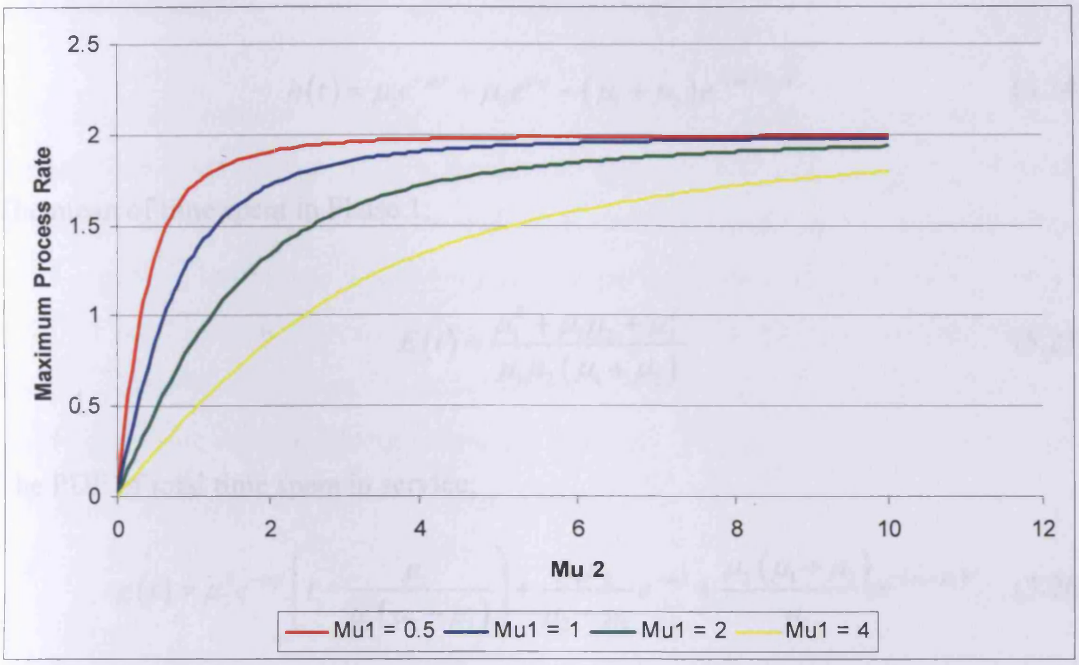


Figure 5.8 - The maximum rate customers are served at, as a proportion of the service rate in Phase 1, in a two Route 2 phase system

Figure 5.8 shows that as service rate 2 increases the maximum process rate increase to a value approaching 2. This is because as the service rate increases in the second phase the amount of blocking decreases and so the service in Phase 1 becomes less affected. As there are two servers in Phase 1, the maximum rate which customers can be processed when blocking decreases will be twice the average service rate in Phase 1.

5.2.5 Summary Measures

The average time that customers spend in this drip feed system is the same as with the single route system. The reason is that, at the time a customer starts their service in the first phase, they have entered a route which they do not leave. This means that the probability density function of the time spent in Phase 1, and the time spent in service in total is the same as the single route case. The equations below are taken straight from Chapter 3; equation (3.38), equation (3.39), equation (3.40) and equation (3.42). They are repeated here only for completeness.

The time spent in Phase 1 in a multi-route, 2 phase system has PDF;

$$h(t) = \mu_1 e^{-\mu_1 t} + \mu_2 e^{\mu_2 t} - (\mu_1 + \mu_2) e^{-(\mu_1+\mu_2)t} \qquad (5.24)$$

The mean of time spent in Phase 1;

$$E(t) = \frac{\mu_1^2 + \mu_1\mu_2 + \mu_2^2}{\mu_1\mu_2(\mu_1 + \mu_2)} \qquad (5.25)$$

The PDF of total time spent in service;

$$g(t) = \mu_2^2 e^{-\mu_2 t}\left(t - \frac{\mu_2}{\mu_1(\mu_2 - \mu_1)}\right) + \frac{\mu_1\mu_2}{\mu_2 - \mu_1} e^{-\mu_1 t} + \frac{\mu_2(\mu_1 + \mu_2)}{\mu_1} e^{-(\mu_1+\mu_2)t} \qquad (5.26)$$

The mean total time spent in service;

$$E(t) = \frac{2\mu_1^2 + 2\mu_1\mu_2 + \mu_2^2}{\mu_1\mu_2(\mu_1 + \mu_2)} \qquad (5.27)$$

5.2.6 Simulation

A Visual Basic simulation similar to the one created in Chapter 3 as been created for the multiple route case. Space for an infinite queue to build up in front of the first phase is allowed. A run size of one million customers was set, and as with the single route model, a record was kept every time a new service started, a current service ends, or when a new customer enters the system. Each time something happens, which is defined as an event, the queue size is noted. Figure 5.9 is a graph of how the queue size changes for different values of λ around $\lambda_{\max}$ as more events occur. The queue size is graphed for every thousandth event.

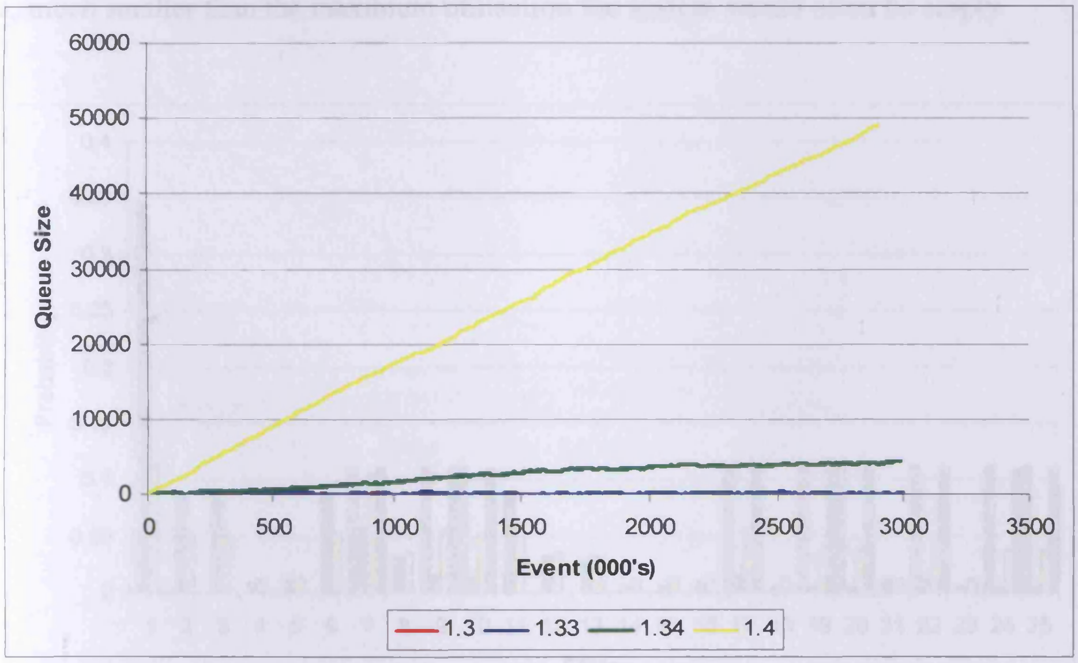


Figure 5.9 - Queue size for different values of λ

As expected in Figure 5.9 a queue starts to build up once the value of λ is greater than $\lambda_{\max}$.

To be able to give each state a numerical value so that they are easier to distinguish, the following calculation will be used to derive the state's number; if Route 1 is in state *i* and Route 2 is in state *j,* as defined in Figure 5.5, then the state of the two route system is $5(i-1)+j$. This gives all 25 possible routes. In the following simulation it was decided to use these 25 different states no matter what the queue size, so if a queue built up during the trial, only one of the 25 states was recorded. This is due to the large number of equations. If the queue size was recorded the number of different states would increase dramatically, making any graphical representation over complicated and confusing.

Figure 5.10 shows the simulation results of the probability of being in different states for different values of λ. The simulation is run for 100,000 customers and an average of 5 runs is taken for all stated values of λ, all service rates are set equal to 1. When λ is equal to 0.5 the probability of being in State 1 is almost 4 times more probable than any other state, State 1 is the case when both routes are empty. With a small value of λ, much smaller than the maximum utilisation the system would often be empty.

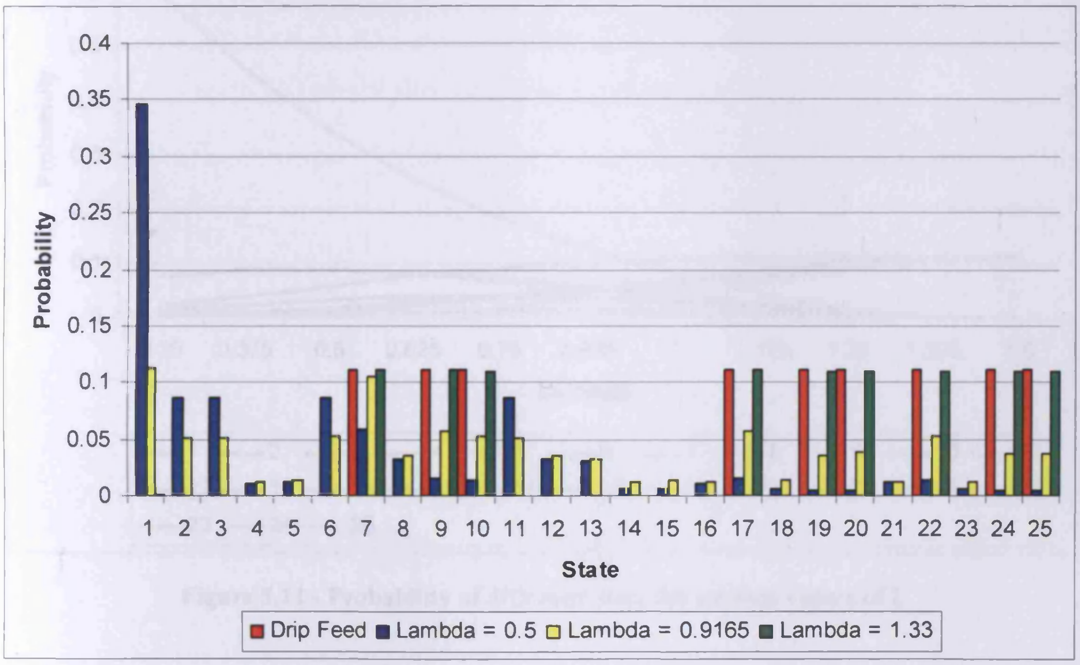


Figure 5.10 - Probability of different state for various values of λ including the drip feed case

It is also worth noting that most of the states are in pairs. In pairs meaning that spikes are of the same size occur in a multiples of 2. This is a useful verification for the simulation. For example a similar probability would be expected for the case when Route 1 is in State 3 and Route 2 in State 4, to the case when Route 1 in State 4 and Route 2 in State 3. The states which do not appear to be in pairs are 1, 7, 13, 19 and

25. In these states both routes are in the same state and therefore are not members of a pair.

As λ increases the probability of being in State 1 decreases dramatically and is more evenly spread amongst all the states. As λ approaches the maximum utilisation the probability of being in a particular state approaches that of the drip feed solution. As λ increases the system will have a queue more often, and as it approaches the maximum utilisation of the system a queue is likely to become more permanent, recreating the drip feed scenario. This can be seen more clearly in Figure 5.11.

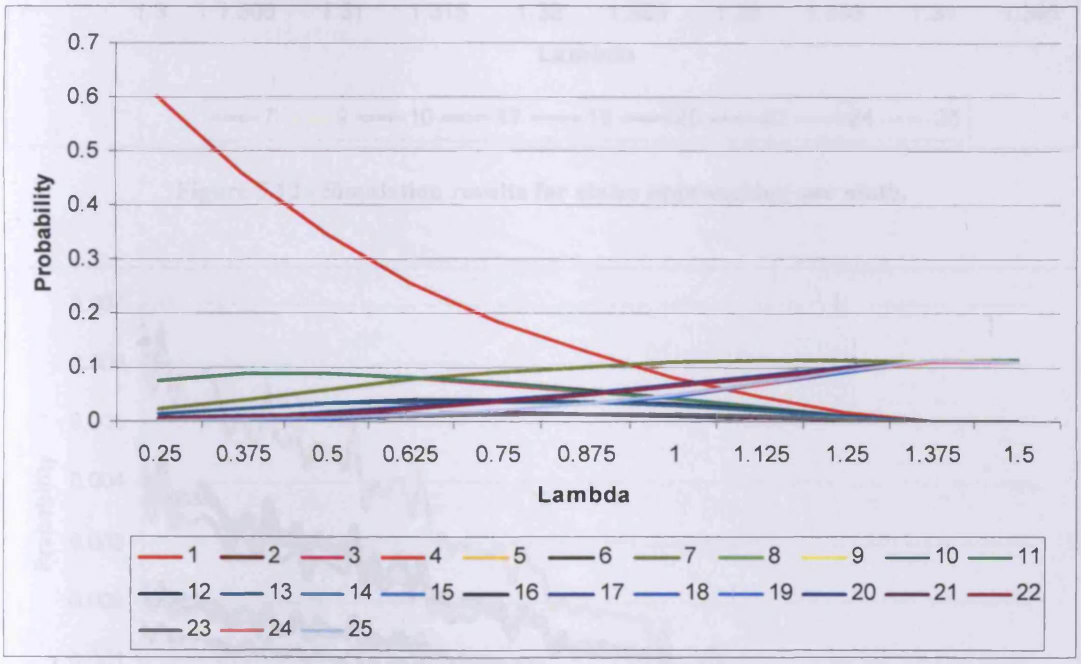


Figure 5.11 - Probability of different state for various values of λ

Figure 5.11 uses the same values for the simulation as in Figure 5.10 with more values of λ included. In this graph the change in state probability as λ increases can be seen more clearly. As λ increase the values of the probability tend to either 1/9 or zero, depending on whether both routes in Phase 1 are occupied or not.

The simulation was tested further when it was run for a million customers, and a sample of 50 runs was taken, the results of which can be seen in Figure 5.12 and Figure 5.13.

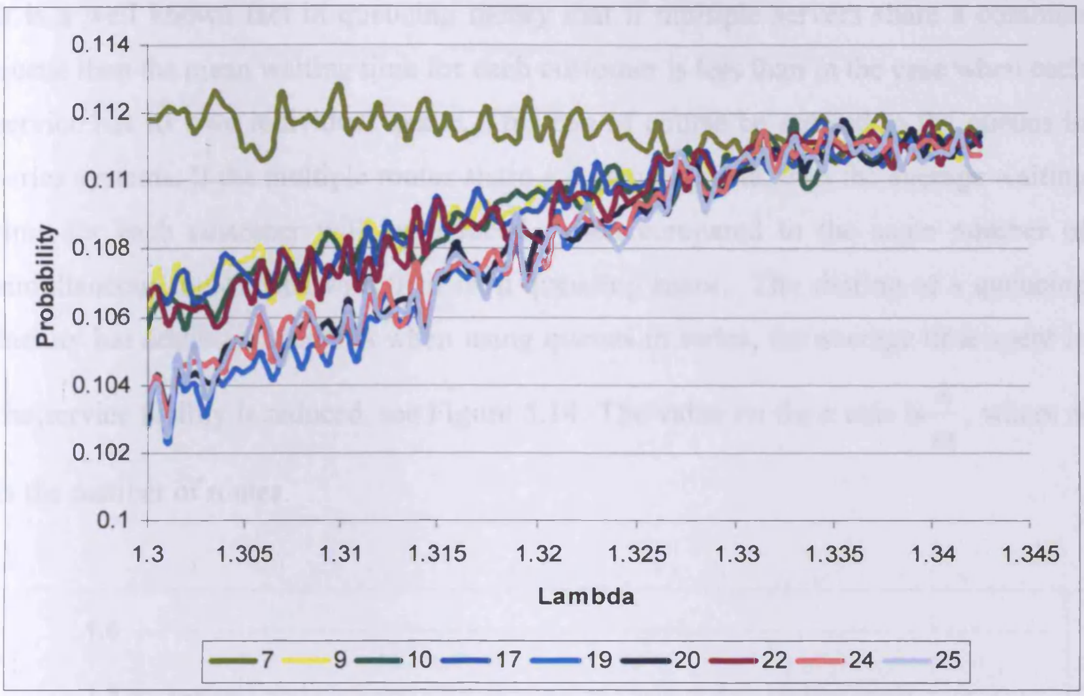


Figure 5.12 - Simulation results for states approaching one ninth.

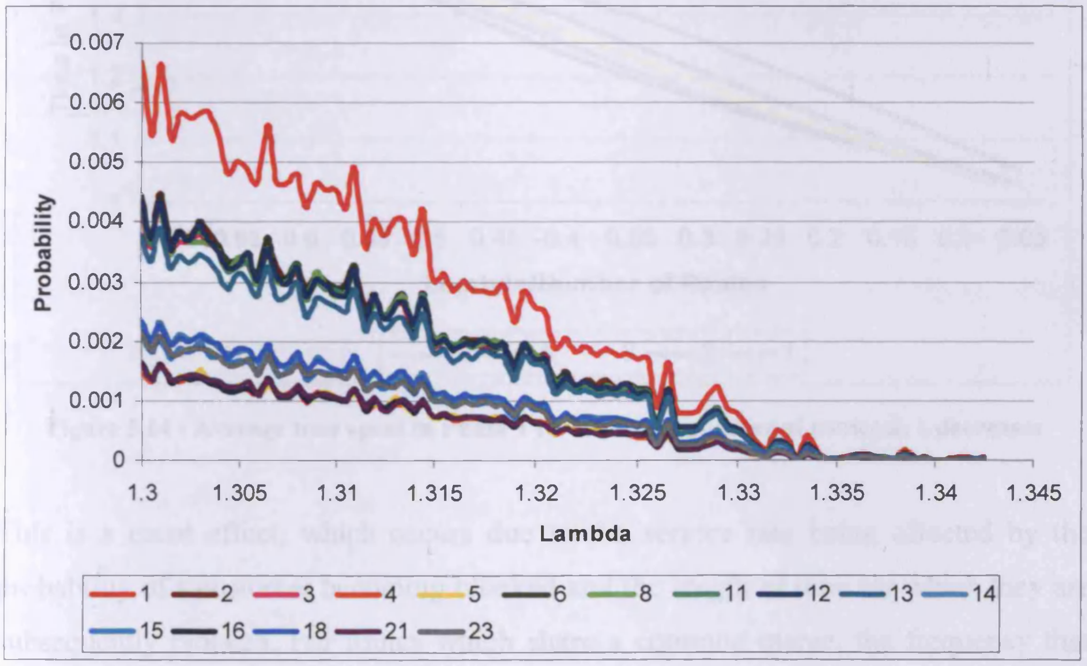


Figure 5.13 - Simulation results for states approaching zero.

These figures show the probabilities tending to the value of zero or 1/9 at values close to $\lambda_{\max}$. The scale on these axes is very small. It shows that the simulation behaves very similarly to the expected results even around the extreme points.

It is a well known fact in queueing theory that if multiple servers share a common queue then the mean waiting time for each customer is less than in the case when each service has its own individual queue. This can of course be applied to the queues in series systems. If the multiple routes share a common queue, then the average waiting time for each customer will be reduced, when compared to the same number of simultaneous routes each with their own queueing space. The sharing of a queueing facility has additional benefits when using queues in series, the average time spent in the service facility is reduced, see Figure 5.14. The value on the x axis is $\frac{\lambda}{m}$, where m is the number of routes.

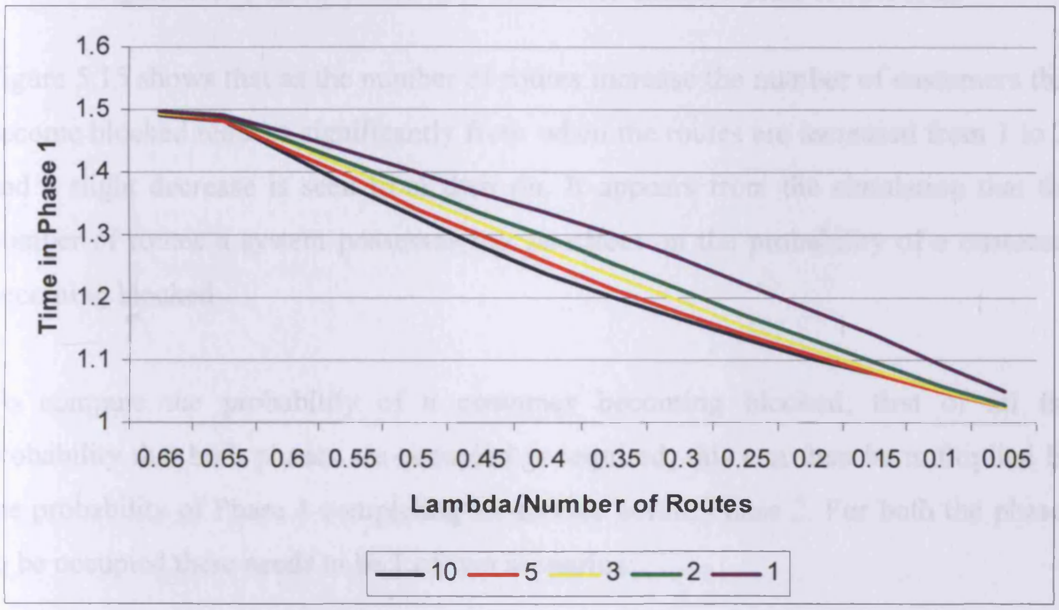


Figure 5.14 - Average time spent in Phase 1 for different number of routes as λ decreases

This is a usual effect, which occurs due to the service rate being affected by the probability of a customer becoming blocked and the length of time for which they are subsequently blocked. For routes which share a common queue, the frequency that customers are blocked is reduced, except for extreme values of λ. This can be seen in Figure 5.15.

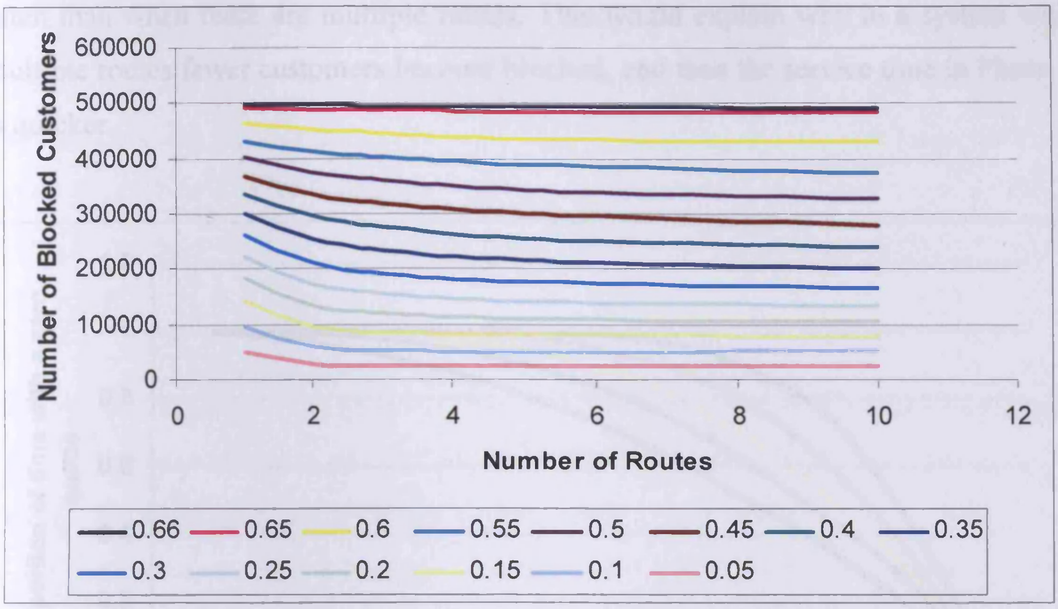


Figure 5.15 - Number of customers blocked for different values of λ per route

Figure 5.15 shows that as the number of routes increase the number of customers that become blocked reduces significantly from when the routes are increased from 1 to 2, and a slight decrease is seen from then on. It appears from the simulation that the number of routes a system possesses has an effect on the probability of a customer becoming blocked.

To compute the probability of a customer becoming blocked, first of all the probability that both phases are occupied is required, this can then be multiplied by the probability of Phase 1 completing its service before Phase 2. For both the phases to be occupied there needs to be 1 of two scenarios;

i. A queue is present when a customer leaves Phase 1. So when the customer leaves Phase 1 and starts their service at Phase 2, there is also a customer starting a service in Phase 1.

ii. When the queue size is equal to zero, an arrival needs to occur during a customer's service in Phase 2.

Figure 5.16 shows the proportion of time that each route spends with a queue size equal to zero. It clearly shows that the more routes there are, the more likely it is to have a zero size queue for comparable arrival rates. Bearing this in mind, the first scenario above, which is the main driver of blocked customer, is likely to happen less often than when there are multiple routes. This would explain why in a system with multiple routes fewer customers become blocked, and thus the service time in Phase 1 is quicker.

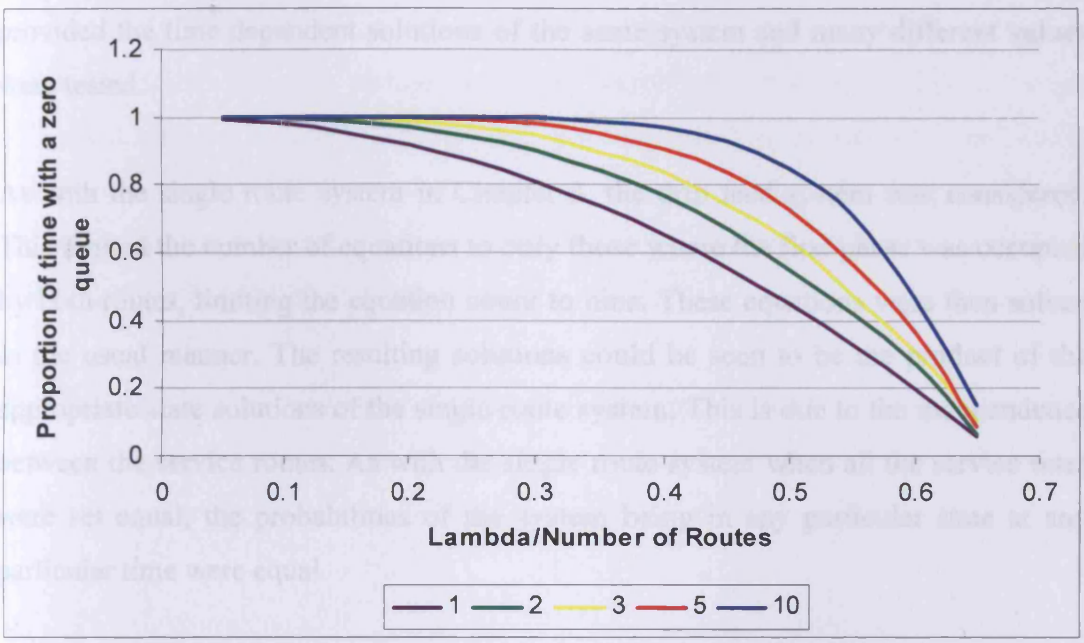


Figure 5.16 - A simulation of the proportion of time with a zero queue size for increasing values of λ and different number of routes, service rates all equal to one.

5.3 Summary

This chapter has focused on queueing systems with more than one service route. Initially a standard queue was analysed. Using Kendal's notation this was an M|M|2 queueing system. This queueing system was analysed by the usual method, by finding time dependent equations and solving them under steady state conditions. The detail of the model was then increased so that it was possible to distinguish in which service point the customers was situated. This was important so that when a second stage was to be added, it would be possible to tell which route would be blocked. The more detailed model was analysed in the same method as for the more traditional example and was shown to give the same results.

The second phase was then introduced; it was shown that for a queueing system with n routes that there would be 5^n possible states that the system could be in excluding any queue. This would also mean that there would be 5^n equations to solve. The time dependent equations of a two phase, two route system without a queue were set up, and solved using the computer program Maple when all the service rates were set equal. To validate these solutions a program in Visual Basic was created which provided the time dependent solutions of the same system and many different values were tested.

As with the single route system in Chapter 3, the drip feed system was considered. This limited the number of equations to only those where the first phase was occupied by both routes, limiting the equation count to nine. These equations were then solved in the usual manner. The resulting solutions could be seen to be the product of the appropriate state solutions of the single route system. This is due to the independence between the service routes. As with the single route system when all the service rates were set equal, the probabilities of the system being in any particular state at any particular time were equal.

The maximum utilisation of this system was then found, by using two different methods established in Chapter 3. The first method involved finding the proportion of the total time that customers spent in Phase 1. In practical terms this means summing the steady state probabilities that have the first phase in a non blocked state. This is because the steady state probability is the probability of finding the system in a particular state at a specified time i.e. the normalised time that the system would be in that state. The second method used was to calculate the probability of a customer becoming blocked then multiplied by time spent in the spent in Phase 1 if blocked, summed with the time spent in the route if not blocked.

The summary measures within this chapter were just those repeated from Chapter 3, as the time spent within the drip feed solution is the same for the multiple route system as for the single Route 1; they were provided for completeness only.

Finally a Visual Basic simulation was created using a similar method to the single route chapter. The various queue lengths were analysed using long runs for different values of λ. It could be seen that whilst λ remained under the value of maximum utilisation no significant queue formed, whereas when the arrival rate increased above the value of $\lambda_{\max}$ a queue appeared that never disappeared, showing that for these values the system was no longer in a steady state. The proportion of time that customers spent in each phase was recorded. It could be seen that as λ tended toward the maximum utilisation the results tended to a drip feed type system; the simulation demonstrated this for quite precise values, as seen on the zoomed in charts in Figure 5.12 and Figure 5.13. Finally the time that customer spent in service was found not to be dependent solely on the service rate but also on the number of routes. As the number of routes increased the proportion of customers that were blocked decreased, allowing for a faster average service rate.

This study of multiple route queueing networks that are capable of becoming blocked will now be used, along with the previous chapter's work, to see how applicable these types of distributions are for modelling a Hospital's Critical Care Unit.

Chapter 6

CRITICAL CARE MODEL

6.1. Introduction

This chapter will combine the techniques that we have used in previous chapters to model the Critical Care Unit of the University Hospital Wales. The CCU was discussed in Chapter 2. We model the flow of patients through the system using a number of different models and discuss the advantages and disadvantages of the various techniques.

6.2 The Blocking Model

When modelling real life situations, it is rare to be able to use theoretical results to exactly reflect the events seen. However, by using mathematical approximations, we can see what is likely to happen to a real life example under similar conditions. Here we model unplanned arrivals to the Critical Care Unit over the 2 year period from 1st April 2004 to the 31st March 2006. A number of different methods are applied using the techniques described in the previous chapters. To model the unit we take into account different aspects of the CCU. The aim of modelling a unit like this is to mimic the behaviour seen in the Unit as accurately as possible. This then enables us to alter the model to see what effect change on the Unit would have, without having the expense of a test and learn experiment.

The model provides an estimate for a patient's actual length of stay. Different models will do this in different ways, and we give the advantages and disadvantages of each method. As well as modelling patients' total length of stay, we also use information related to the blocking of patients in the CCU to see how changing this can effect the length of stay and hence the capacity of the unit.

To be able to use blocking equations to model the patient flow through the Critical Care Unit, we first need to understand the variables within the unit that would be included in the model. These include; numbers of service points (beds), inter arrival time, service time (length of stay) and probability of a patient becoming blocked.

In the blocking model we have 2 phases through which the patient must pass. If the customer in Phase 1 finishes their service before there is space in Phase 2 they become blocked. Once there is space available the patient in Phase 1 then moves into Phase 2 and makes space for a new customer in Phase 1. In reality the Critical Care Unit does not behave exactly like this. When a patient becomes blocked, it can be due to many different factors including; low staff numbers, no medicine available or limited ward bed space. Also, when a patient leaves the CCU they are not necessarily going to occupy the same bed that the previous incumbent of that bed in the CCU occupied, as the theoretical blocking equations describe.

To overcome this we build a model that only calculates the time spent in Phase 1, in this case Phase 1 is the CCU. Phase 2 shall be a dummy phase that patients will not actually enter. As a patient enters the system in Phase 1, a dummy patient enters Phase 2 simultaneously. Then, if the dummy patient's service exceeds the real patient, the patient will become blocked and a blocked time will be recorded for that patient. If the real patient completes their service after the dummy patient, then no blocking will occur, and once the service is complete, the real patient leaves the system. The flow path of this system can be seen in Figure 6.1. The left square in each of the rectangles refers to a bed in the CCU and the right hand side to the dummy station. The figure starts at 1, where both are empty. At point 2 a patient has arrived and a dummy patient has also started simultaneously. If the real patient finishes first, they then become blocked, indicted by the green dot changing to red in 3. They then stay in the bed until the dummy patient completes their service and then the bed becomes empty, and the patient immediately leaves the CCU. If the dummy patient completes their service first, then the real patient completes their service and then leaves the system as expected. The time that the customer spends in this phase will therefore come from the distribution of the maximum of two Exponential distributions as seen in Chapter 3 Equations (3.40). Once the customer has completed their service, the system becomes empty. As we do not allow a queue to build up, the bed always becomes empty before a new patient is admitted.

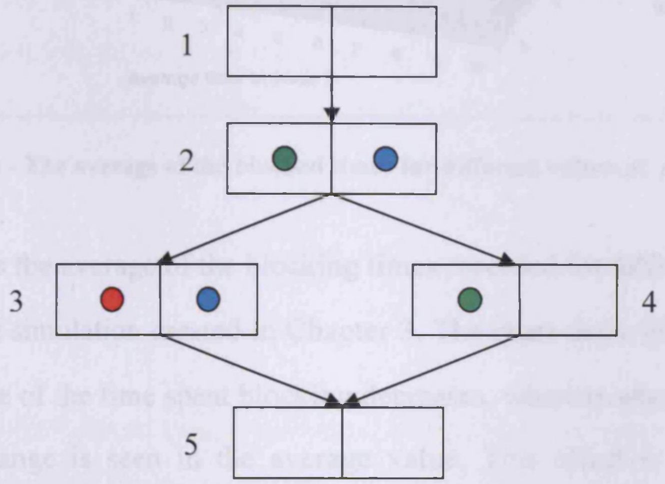


Figure 6.1 - Patient flow through CCU bed model.

The memory-less property of the Exponential distribution was discussed in Chapter 3. If a customer has spent time *t* in a phase so far, then the rest of the time spent in that phase has the same parameter as the original time, so the distribution does not remember what has happened previously. This helps us with the parameter value in the dummy phase. If a patient becomes blocked, the time that they will spend in that blocked state will come from the same distribution as that of Phase 2. No matter how much time has passed (the time that the patient has already spent in Node 1), the remainder of time in the second phase will have the same distribution. A result related to this can be seen in Figure 6.2. Therefore the distribution of the dummy patient is the same as that of the blocked time.

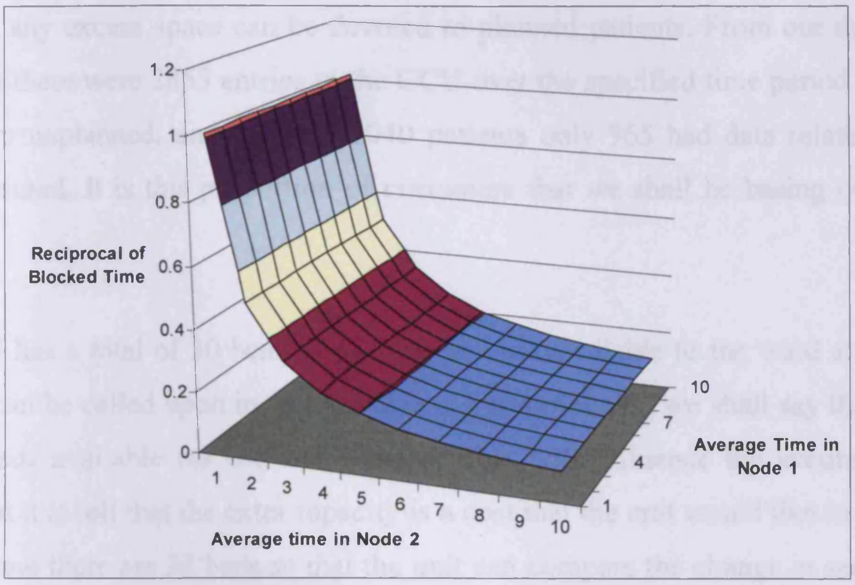


Figure 6.2 - The average of the blocked times for different values of μ_1 and μ_2.

Figure 6.2 shows the average of the blocking times recorded for different values of μ_1 and μ_2 from the simulation created in Chapter 3. The chart show that as μ_2 increases the average value of the time spent blocking decreases, whereas when the value of μ_1 increases no change is seen in the average value. This effect is also true of the standard deviation since the standard deviation of the Exponential distribution is the same as the mean. The distribution of the time spent blocking is dependent on the value of μ_2 due to the memoryless property of the Exponential distribution.

6.3.1 The Data

Within this analysis we model only the unplanned cases. These are the patients where the hospital has no control over their time of arrival, although this is not strictly true. It can be argued that there is some level of control; the hospital is able to divert some patients that are destined to the hospital, though this rarely happens. Likewise it can be argued that the hospital has only limited control over planned patient arrivals. The hospital has to serve patients that are on their waiting list, and the Unit could not turn these patients away for continued periods of time. However, in this case, we shall say that the patients that are not from expected elective cases cannot be planned for, and there needs to be adequate provision for them. Once the demand for these patients is met, then any excess space can be devoted to planned patients. From our data set we know that there were 2853 entries to the CCU over the specified time period. Of these, 2040 were unplanned, and of these 2040 patients only 965 had data related to their delay captured. It is this proportion of customers that we shall be basing our models on.

The CCU has a total of 30 beds. 6 of these are not available to the ward at all times, but they can be called upon in an emergency. For simplicity, we shall say that the unit has 24 beds available for use at all times. This will influence the accuracy of the model, but it is felt that the extra capacity is a cost that the unit would like to avoid; we shall assume there are 24 beds so that the unit can compare the change in service rates more accurately.

6.3.2 Inter Arrival Time

The inter arrival time distribution is estimated using the entire unplanned data set. The whole dataset is used as we are attempting to model the 24 standard beds in the Unit. The delay dataset comes from the same time period as the remainder of the data but accounts for less than half of the total arrivals. By using only that data, we are likely to under estimate the inter arrival time. The inter arrival time of all the unplanned patients has been calculated directly from the data and been summarised in Figure 6.3.

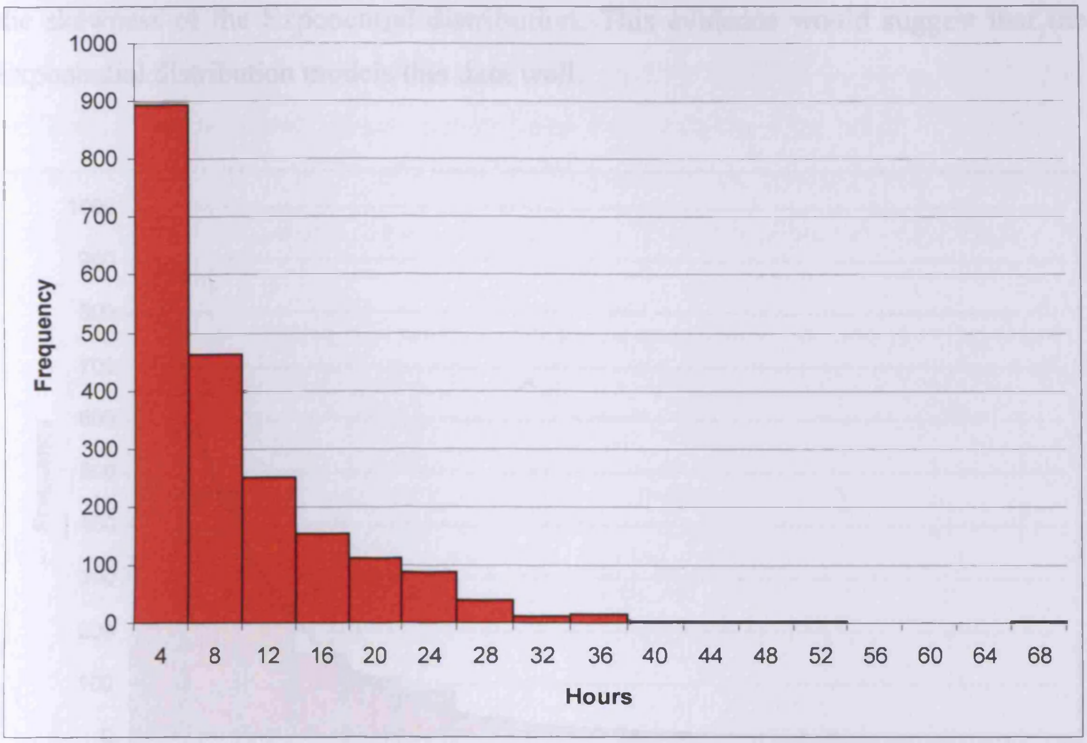


Figure 6.3 - Frequency of inter arrival time of all unplanned patients.

The data appears to have a negative Exponential distribution type shape. This is what we would expect intuitively. As these patients are unplanned, the time between arrivals would be random. The summary measures are all calculated from the data before they were grouped.

Table 6.1 - Inter arrival time summary measures (hours).

Statistic	Value
Average	7.50
Minimum	0.00
Maximum	67.00
Median	4.75
Standard Deviation	7.68
Co efficient of Variation	1.02
Skewness	2.02

The data in Table 6.1 gives us a good understanding of the inter arrival time distribution. It can be seen that the average time between arrivals of unplanned patients is 7.5 hours with a similar standard deviation of 7.68 hours. This gives a Coefficient of Variation of around 1. The Coefficient of Variation for the Exponential distribution is 1. The skewness of the data is approximately 2, which is the same as the skewness of the Exponential distribution. This evidence would suggest that the Exponential distribution models this data well.

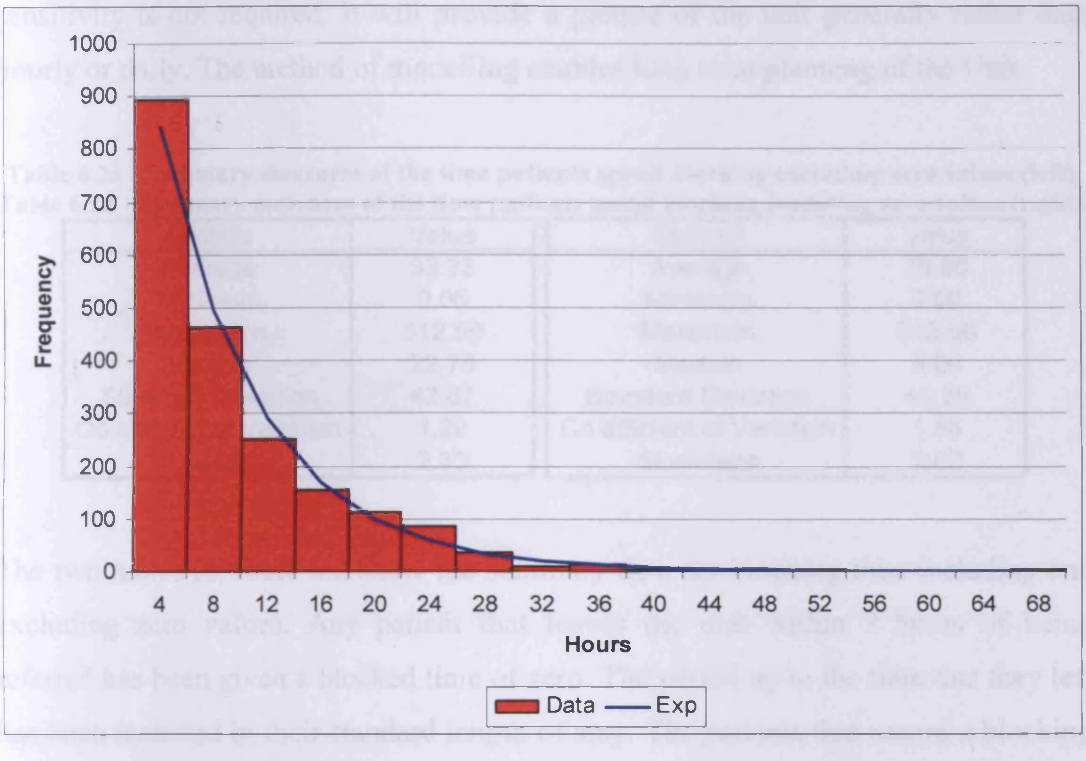


Figure 6.4 - Inter arrival distribution of unplanned patients with an Exponential distribution.

Figure 6.4 shows the inter arrival times overlaid with the fitted Exponential distribution. The distribution has the same mean as the data before it was grouped, i.e. the mean given in Table 6.1. The Exponential distribution captures the general shape of the distribution very well.

6.3.3 Blocking Time

As mentioned above, the time a patient spends blocked in this model comes from the same distribution as the second node distribution, assuming an Exponential distribution can be fitted. Table 6.2 is a summary of the time a patient spends blocked within the CCU. The definition of blocking is the same as given in Chapter 2. Once a patient has been referred from the CCU to a different ward by their doctor, they are then given 7 hours before they are classified as blocking a bed in the Critical Care Unit. This data is very time dependent as shown in Figure 2.12, and has been analysed in Figure 6.5 in a daily form, giving the data a non time dependent form. This may appear to lose some of the sensitivity and accuracy of the data, but we shall be looking at the results of the model over long run averages where this level of sensitivity is not required. It will provide a picture of the unit generally rather than hourly or daily. The method of modelling enables long term planning of the Unit.

Table 6.2a - Summary measures of the time patients spend blocking excluding zero values (left), Table 6.2b - Summary measures of the time patients spend blocking including zero values (right).

Statistic	Value
Average	33.33
Minimum	0.05
Maximum	312.50
Median	22.75
Standard Deviation	42.87
Co efficient of Variation	1.29
Skewness	2.33

Statistic	Value
Average	25.90
Minimum	0.00
Maximum	312.50
Median	5.00
Standard Deviation	40.26
Co efficient of Variation	1.55
Skewness	2.62

The two tables in Table 6.2 show the summary data for blocking time including and excluding zero values. Any patient that leaves the unit within 7 hours of being referred has been given a blocked time of zero. The period up to the time that they left has been included in their standard length of stay. The patients that record a blocking time have also been given the extra 7 hours on their standard length of stay, and then the blocked time has been recorded. As can be seen from the very different figures (especially the average and the median) the non blocked customers make up a large proportion of the patients. 215 out of the 965 patients that we have delay data records for did not block the CCU under the 7 hour rule. This means that 77.7% of patients become blocked. This may seem like a high proportion, but this highlights the magnitude of the issue facing the department. Figure 6.5 show the distribution of time spent blocking by patients (excluding zeros) with an Exponential distribution approximation.

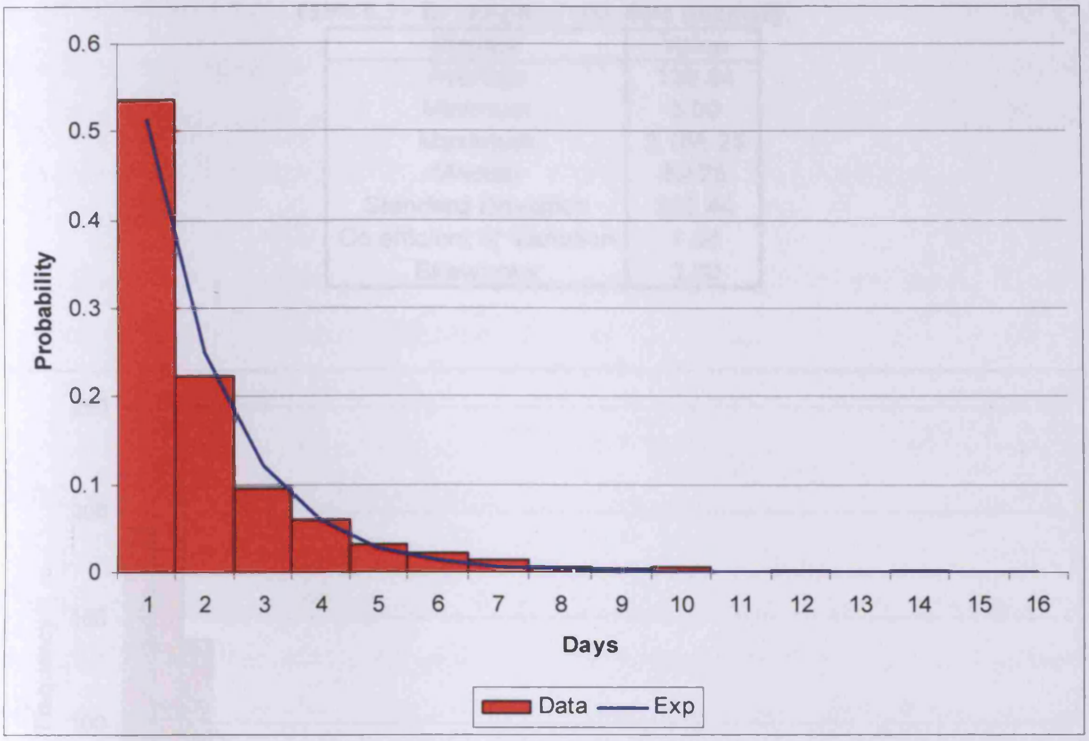


Figure 6.5 - Time patients spent blocking with an Exponential distribution overlaid.

Both the Coefficient of Variation and skewness are slightly higher than we would expect for an Exponential distribution. To be able to use the blocking equations that have been built, we require an Exponential distribution in both phases. Figure 6.5 shows that the distribution fits the data well.

6.3.4 Ill Length of Stay

Next we shall consider the ill length of stay. The ill length of stay is the time that a patient spends on the ward before they are referred out, plus the 7 hours grace period that is given to prepare the patient destination ward. If the patient leaves the CCU before the end of this 7 hour grace period, then only the time that they spend in the CCU is used. Table 6.3 show the summary data for the ill length of stay for the 965 patients in our data set, and Figure 6.6 show the frequency of these times with an overlaid Exponential distribution with the same mean as the data.

Table 6.3 - Ill Length of stay data summary.

Statistic	Value
Average	139.54
Minimum	3.00
Maximum	2,165.25
Median	59.25
Standard Deviation	220.44
Co efficient of Variation	1.58
Skewness	3.80

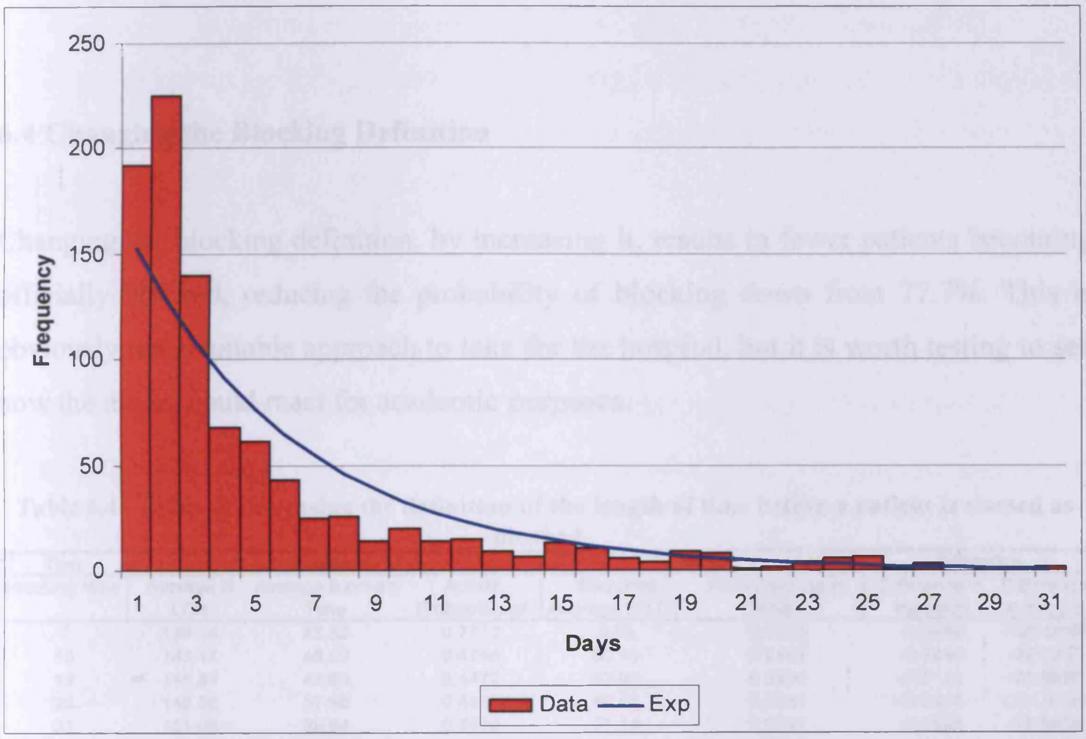


Figure 6.6 - Ill length of stay chart with Exponential distribution.

The pattern of this data is obviously not Exponentially distributed, though it does show a similar shape with the high peak, and a skew to the right. The Exponential distribution line in Figure 6.6 certainly does not fit the data very well. To be able to use the blocking equations, we require the distribution of the time spent in the first phase to be Exponential. However, a larger problem lies with the values provided by the data. The probability of blocking within this model is directly related to the parameters of the two Exponential distributions;

$$P(Blocking) = \frac{\mu_1}{\mu_1 + \mu_2} \tag{6.1}$$

If we accept the Exponential distribution parameters for the ill length of stay and the blocking time, we would expect a probability of blocking of 0.1928. We have already seen that the actual probability of blocking is 0.7772. This is clearly a limitation for this type of blocking model and may make it difficult to apply in this case.

There are a number of things that can be done to attempt to adapt this model to make it more applicable to real data. We shall consider 3 of these.

6.4 Changing the Blocking Definition

Changing the blocking definition, by increasing it, results in fewer patients becoming officially blocked, reducing the probability of blocking down from 77.7%. This is obviously not a suitable approach to take for the hospital, but it is worth testing to see how the model could react for academic purposes.

Table 6.4 - Effect of increasing the definition of the length of time before a patient is classed as blocked.

Time	Data			Model		Difference	
Blocking Hour	Average Ill LOS	Average Blocked Time	Actual Probability of	Required Average Ill LOS	Prob Blocking in Model	Difference in Blocking	Difference in Ill LOS
7	139.54	33.33	0.7772	9.55	0.1928	-0.5844	-129.9865
13	143.14	46.50	0.4798	50.41	0.2452	-0.2346	-92.7230
19	145.89	43.68	0.4477	53.90	0.2304	-0.2172	-91.9936
25	148.58	37.86	0.4456	47.11	0.2031	-0.2425	-101.4699
31	151.08	39.94	0.3596	71.14	0.2091	-0.1505	-79.9459
37	152.86	48.97	0.2570	141.58	0.2426	-0.0144	-11.2779
40	153.61	48.80	0.2425	152.45	0.2411	-0.0014	-1.1653
43	154.34	46.21	0.2404	146.01	0.2304	-0.0100	-8.3229
49	155.77	40.76	0.2373	131.00	0.2074	-0.0299	-24.7712
55	157.10	44.76	0.1865	195.21	0.2217	0.0352	38.1094
61	158.02	56.90	0.1306	378.87	0.2647	0.1342	220.8577
67	158.77	53.21	0.1254	371.17	0.2510	0.1256	212.3973
73	159.53	47.21	0.1254	329.32	0.2284	0.1030	169.7937

The effect of increasing the definition of the time before a patient is classed as blocking a bed is shown in Table 6.4. The first column indicates the number of hours after the original referral before a patient is classified as blocking a bed, currently 7 hours. The table is summarised in increments of 6 hours, though the data was originally analysed by increasing the time by an hour at a time. There is also an entry for 40 hours as this provided the optimal entry for this model.

The next 3 columns show the effect on the data of changing the blocking definition. The average ill length of stay can be seen to increase; this is an expected result, and a direct result of classifying more time as being included in the ill length of stay. The average blocked time is less consistent, though the trend is definitely in an upwards direction. This may seem counter intuitive, but can be explained by the long tail of the time spent blocking. Figure 6.6 shows that a large proportion of the blocking patients are blocking for a short period of time. As these patients are removed, the patients that are delayed for longer make up more of the population and the average time increases. The final effect on the data is that the probability of blocking can be seen to decrease, even though the amount of time customers are blocked for increases, there are fewer of them.

In the 'Model' section of the table, the required average ill length of stay and required probability of blocking are given. Both of these fields are calculated from equation (6.1). The required ill length of stay, μ_1, is calculated from the re-arrangement of equation (6.1) in terms of μ_2. The other terms are taken from the data; the probability of blocking, and the average amount of time spent blocking. The required blocking probability uses equation (6.1) in its current form and uses the average length of time spent blocking and the average ill length of stay provided by the data, to give the proportion of people who would be blocking if those parameters were used in the model.

The final two columns are the difference between the calculated field and the real data values. It is a crude approximation to see how many hours blocking are required to get the model to more accurately reflect the data. It turns out that the best fit for both parameters results in the same value of blocking time i.e. 40 hours.

This value is unfortunately not very applicable for the CCU, though 24% of patients are blocking for periods greater than this length of time. Also when this value is used for the definition of blocking, the resulting distribution of the ill length of stay of the patients and of their blocking times appear to be even less Exponential, Figure 6.7. This would make the total length of stay, found by running patients through the model, rather inaccurate. As a result, this method is not considered further.

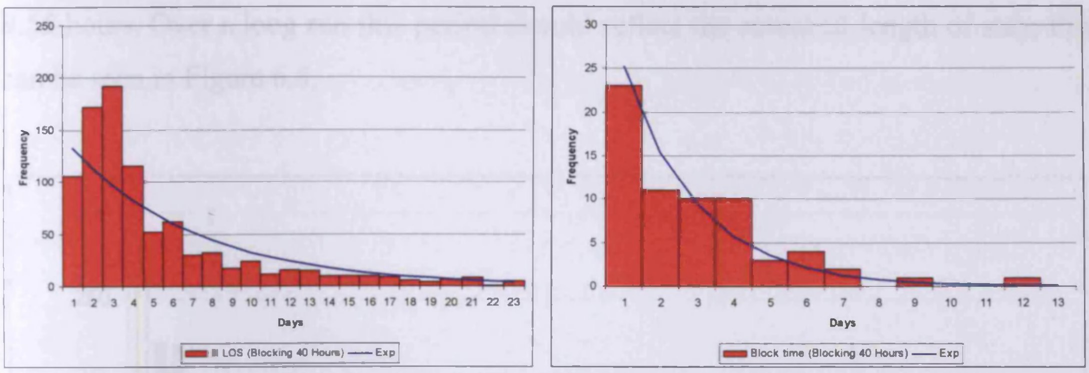


Figure 6.7 - Ill length of stay (left) and blocking time (right) frequencies.

6.5 Adding an Extra Node

Another possible solution to the problem maybe to add an extra service point before the blocking phases. All patients will be forced to go through this service and then move into the blocking service point, where the distributions can be fixed to give the correct probability of blocking, and then the model may fit the data more precisely.

The difficulty with this approach is finding the distribution of time spent in the extra node. The mean of this distribution would be the mean of the ill length of stay of the patients minus the mean time spent in the first node of the blocking system. It can be seen from the first row of data in Table 6.4, that the required value of the average time spent in the first node is 9.55 hours. The table also shows that the mean ill length of stay is 139.54 hours. The extra service time that patient would have to go through should have an average time of 130 hours.

To generate some data to model with, it was decided that the average time from the first node, 9.55 hours, would be taken off of the actual data and that would be the basis from which to work. If the actual data was less than 9.55 hours, then the patient's data was removed from the calculations. This removed the possibility of having a negative service time. This is not ideal as it will somewhat distort the shape of the distribution; but only 12 patients times were affected by this editing. Once the patients have gone through this service point, they will move into the blocking system, where they will have a random service time with an average service time of

9.55 hours. Over a long run this period should reflect the actual ill length of stay; this can be seen in Figure 6.8.

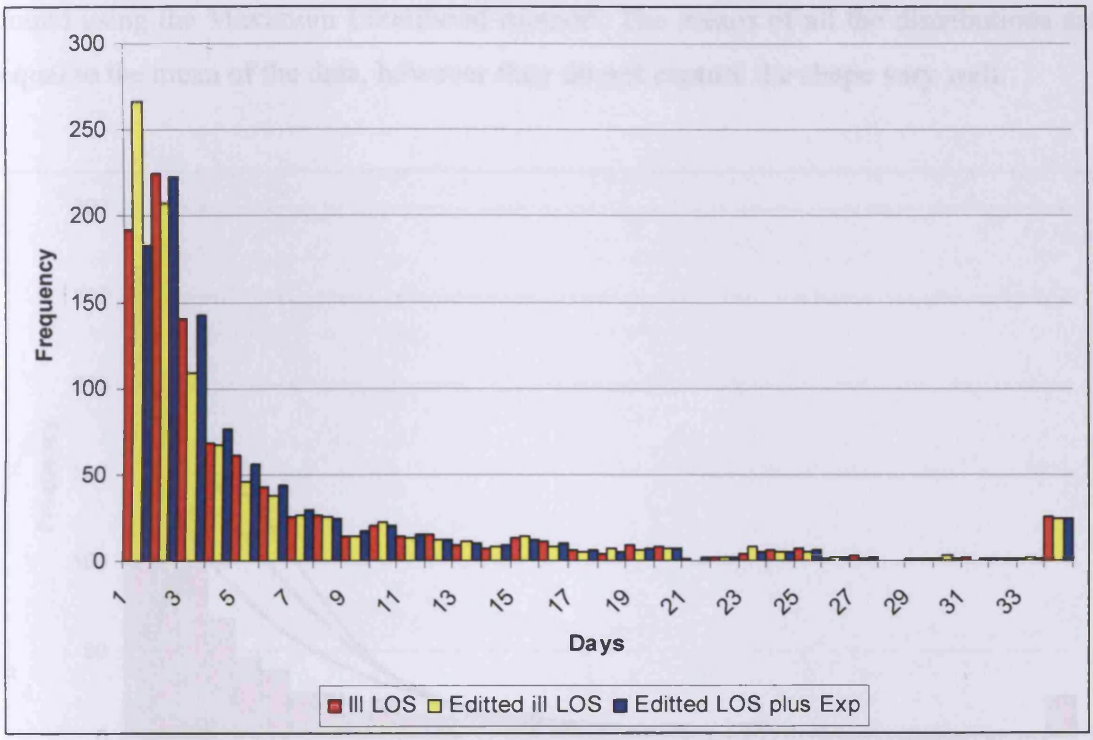


Figure 6.8 - Ill length of stay, edited by -9.55 hours, and the edited data with an added Exponential distribution.

Figure 6.8 shows the real ill length of stay data in red, and the ill length of stay data that has been edited by removing 9.55 hours in yellow, and finally the edited data plus the a time taken from an Exponential distribution with a mean time of 9.55 hours in blue. The edited ill length of stay plus an Exponential distribution is taken from an average 1000 runs. As expected when taking an average of these, the resulting chart looks similar to the original data. As we are looking at simulating this situation over a long period, the ill length of stay should be accurately modelled, assuming a suitable distribution can be found for the extra node.

Now the edited data needs to be approximated by a distribution to see how good a fit to the data the model is. Initially, the data was approximated by an Exponential distribution to be in keeping with the rest of the model. However, this was not adequate as it did not fit the shape of the data very well and did not take into account the long tail of the distribution. An extension of the Exponential distribution, the

Erlang distribution was tried. This distribution is the sum of a number of Exponential distributions with the same parameters. The results of these distributions being used to fit the data can be seen in Figure 6.9. The parameters for these distributions were found using the Maximum Likelihood method. The means of all the distributions are equal to the mean of the data, however they do not capture the shape very well.

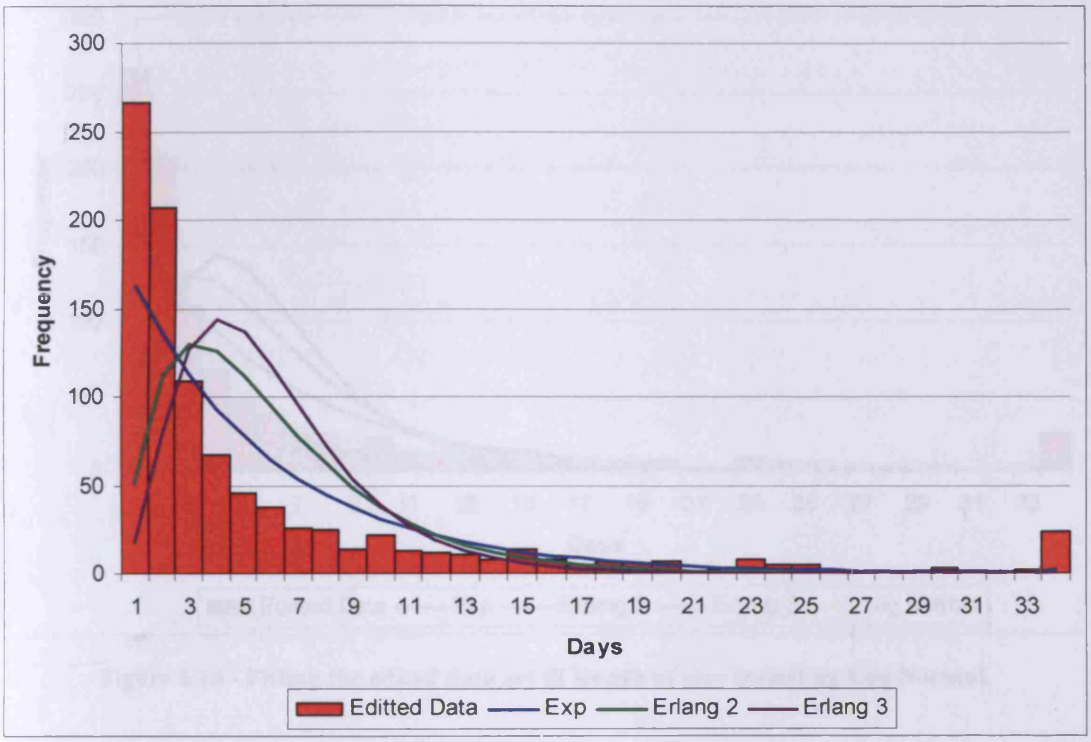


Figure 6.9 - Fitting the edited data set ill length of stay.

Another distribution that is often used to model length of stay is the Log Normal distribution; it is especially useful when the data has very long tails, (Strum, May et al. 2000) and (Spangler, Strum et al. 2004). It can be fitted to the data in the same way as the other distributions, using the Maximum Likelihood method. The Log Normal distribution has two parameters, μ and σ, which are the mean and standard deviation of the associated Normal distribution. The maximum likelihood estimators for this distribution are;

$$\hat{\mu} = \frac{\sum_{i=1}^{n} \ln(x_i)}{n} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \ln(x_i - \hat{\mu})^2}{n}$$

Using the edited data, these are found to give values of $\hat{\mu} = 3.95$ and $\hat{\sigma}^2 = 1.43$. The graph of this Log Normal distribution can be seen in Figure 6.10, alongside the Exponential and Erlang distributions.

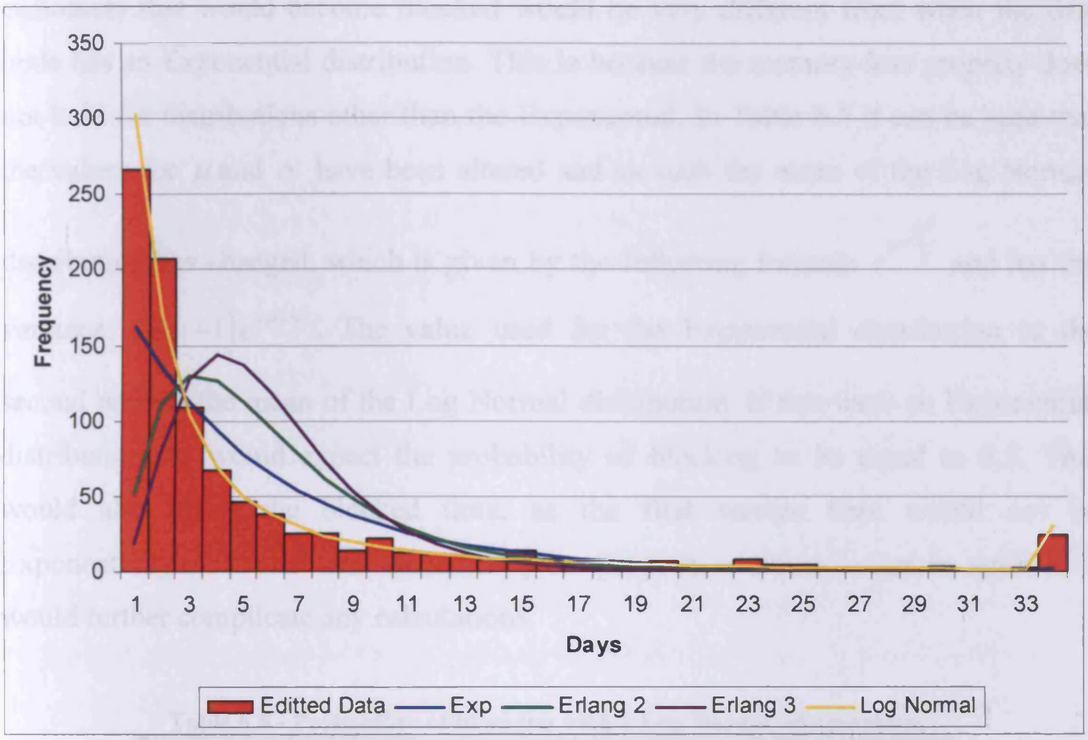


Figure 6.10 - Fitting the edited data set ill length of stay including Log Normal.

The Log Normal distribution is clearly the best fit, even the long tail is accounted for by the this distribution. However, the expected value of this Log Normal distribution is 144 hours; the real data has an average of 130 hours. This will mean that this model will overestimate the length of stay of patients in the Unit. We shall continue to use this model, as making all patients go through the extra node using the Log Normal distribution will enable us to use the blocking system to analyse the data, but this over estimation should be noted in all future calculations. This new method of adding an extra service point before the blocking service point adds flexibility to the otherwise quite restricted distribution.

6.6 Changing Distribution of the First Node

A further possibility is to use the two node blocking case, but to model the first node's length of stay with a distribution other than the Exponential. The proportion of customers that would become blocked would be very different from when the first node has an Exponential distribution. This is because the memory-less property does not hold for distributions other than the Exponential. In Table 6.5 it can be seen that the values for μ and σ have been altered and as such the mean of the Log Normal distribution has changed, which is given by the following formula $e^{\mu+\frac{\sigma^2}{2}}$ and has the variance, $\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$. The value used for the Exponential distribution in the second node is the mean of the Log Normal distribution. If this were an Exponential distribution we would expect the probability of blocking to be equal to 0.5. This would also affect the blocked time, as the first service time would not be Exponentially distributed and the memoryless property could no longer be applied. It would further complicate any calculations.

Table 6.5 - Probability of blocking with a Log Normal distribution.

Mu	Sigma	Mean	Variance	P(B)
0.1	0.1	1.11	0.01	0.3571
0.1	0.5	1.25	0.45	0.4049
0.1	1	1.82	5.70	0.5156
0.1	2	8.17	3,574.26	0.7317
0.5	0.1	1.66	0.03	0.3682
0.5	0.5	1.87	0.99	0.4135
0.5	1	2.72	12.70	0.5096
0.5	2	12.18	7,954.67	0.7337
1	0.1	2.73	0.08	0.3736
1	0.5	3.08	2.69	0.4105
1	1	4.48	34.51	0.5133
1	2	20.09	21,623.04	0.7359
2	0.1	7.43	0.55	0.3689
2	0.5	8.37	19.91	0.4128
2	1	12.18	255.02	0.5091
2	2	54.60	159,773.83	0.7347
5	0.1	149.16	223.59	0.3667
5	0.5	168.17	8,032.96	0.4180
5	1	244.69	102,880.65	0.5054
5	2	1096.63	64,457,364.85	0.7366

The variability in this method would make it very difficult to alter the parameters within the distribution and still keep a reliable estimate of the real data.

For the blocking model we shall use the second method discussed, that is, adding the extra service point in front of the blocking facility. This will allow us to alter the distribution fairly accurately; however, we must be careful when interpreting the results.

6.7 Phase Type Distribution Model

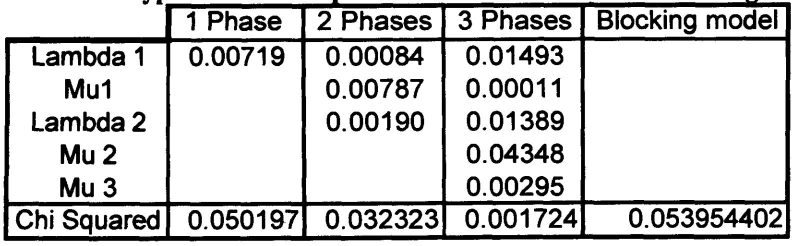
As well as analysing the CCU using a blocking model, we shall also consider a Coxian Phase Type in modelling to the Unit. This distribution is a very flexible distribution that can adapt to many different situations, due to the large number of parameters it has. The Coxian Phase Type distribution does have less parameters than the standard Phase Type distribution, but there are still many to estimate.

One statistic to see if the Phase Type distribution is an appropriate model for the data is if the square of the coefficient of variation of the data is greater than the reciprocal of the number of phases. In this data the coefficient of variation is greater than one so this will always be the case. A program has been built to estimate the parameters of the Coxian Phase Type distribution. If there are *n* phases in the distribution then there are $2(n\text{-}1) + 1$ parameters to estimate. The cumulative real data is first normalised and then the cumulative distribution of the Phase Type distribution is calculated. This is calculated from the equation for $P_E(t)$ in Chapter 4, equation (4.62). It is the probability that a customer will be in the Exit Phase at time *t*, and can be interpreted as the cumulative density function of the service time distribution. A Chi-squared goodness - of – fit is then calculated. The parameters with the smallest Chi-squared value are then accepted as the best fit to the data. The parameters have the constraints that the average of the distribution must be equal that of the data and also that the sum of each phase's 2 parameters, λ_i and μ_i, are not to equal the sum of the parameter in another phase. This is so that the denominator of $P_E(t)$is not equal to zero, equation (4.62). The final constraint was the maximum size of the average time of each Exponential distribution, which was limited to 10,000 hours. This was done so that the program had a limited period over which it could run.

It was decided to create the estimation system in Visual Basic so that the program was transferable, and could be controlled in a user friendly manner. It soon became apparent that this would not be the case. The program takes much too long to run to be able to be used as a regular tool. The three phase system can take over 10 hours. This is due to the large number of iterations over a number of different parameters. It was for this reason that the Phase Type model was not used in the blocking model. In the Blocking model, each time either of the first blocking phases parameters changed, new parameters were required for the extra service point. This would have been very time consuming and not practical for regular changes to the model.

The results of the program can be seen in Table 6.6 and they are illustrated graphically in Figure 6.11. The tables give a value for Chi-squared, which is not a true chi-squared value and could not be interpreted as one; as it is the value that has been minimised to find the optimal parameters using the Chi-squared formula. It does show how an increase to the fit is achieved with more phases, as well as the Blocking models. It looks to be a much worse fit to the data than the Phase Type distribution, but this distribution is much more flexible, and parameters are easy to estimate. The chart show the best estimates for the Phase Type for up to 3 phases. It was decided that as the fit of the 3 phase model seemed very good, it was not worth while extending to a 4 phase model, when the estimation of the parameters took so long to calculate.

Table 6.6 - Phase Type distribution parameter estimates of the total length of stay.



	1 Phase	2 Phases	3 Phases	Blocking model
Lambda 1	0.00719	0.00084	0.01493	
Mu1		0.00787	0.00011	
Lambda 2		0.00190	0.01389	
Mu 2			0.04348	
Mu 3			0.00295	
Chi Squared	0.050197	0.032323	0.001724	0.053954402

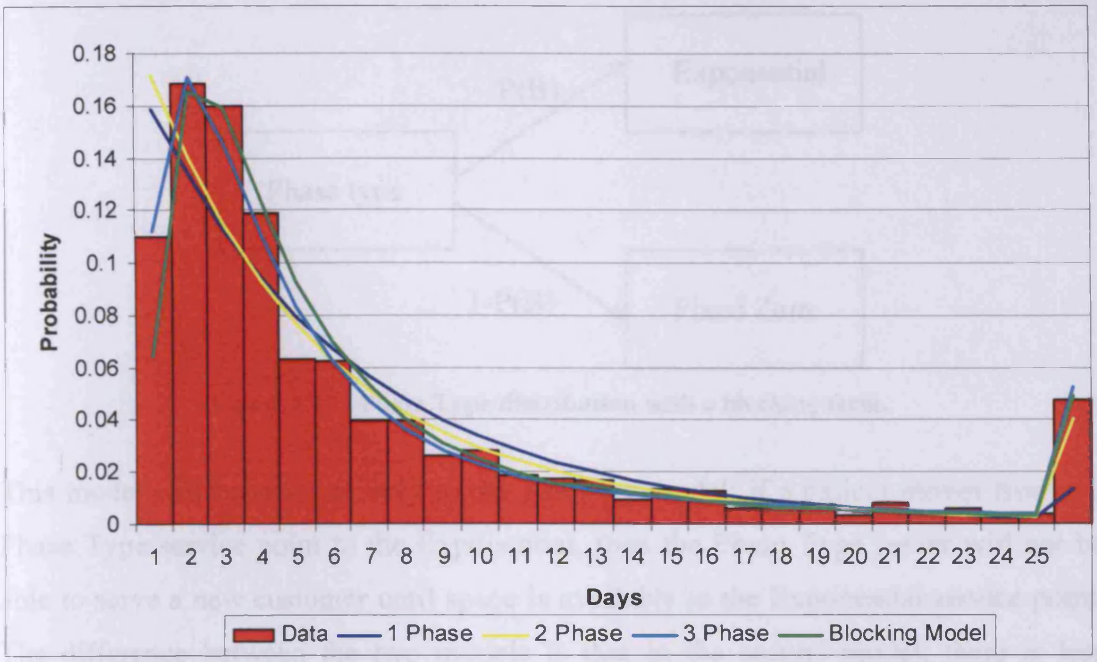


Figure 6.11 - Total length of stay with the Blocking model and the Phase Type distributions.

We would like to be able to use this distribution to build a separate model and compare the results to the Blocking model. However, using this method does not enable us to change the proportion of blocked customers or the average time spent blocking. We need to alter the Phase Type model to be able to take these effects into account.

If the Phase Type distribution is used to model a patient's ill length of stay rather than the total length of stay, it would then be possible to have the patient moving (with a probability equal to the probability of blocking) to a Exponential distribution with an average time of blocking equal to the mean of the data. This type of distribution can be seen in Figure 6.12. The distribution for the blocking time will be the same as that for the blocking model, an Exponential distribution with mean time 33.33 hours, and the probability of blocking will be set to 0.7772.

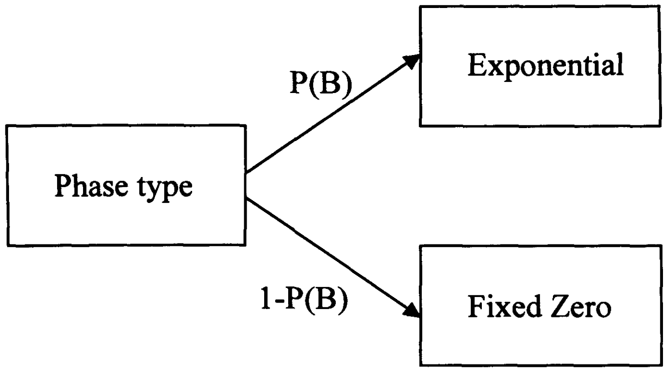


Figure 6.12 - Phase Type distribution with a blocking term.

This model will behave similarly to the blocking model; if a patient moves from the Phase Type service point to the Exponential, then the Phase Type server will not be able to serve a new customer until space is available in the Exponential service point. The difference between the two models is that in the second model, there is less dependence between the two phases. Adding an extra Poison process to the back of a phase type distribution creates a new adaption of the existing Phase Type work. This allows the total distribution of length of stay to be altered without affecting all of the parameters in the distribution by altering the blocking time. The effect of altering a parameter in the blocking model will automatically affect the other parameter, whereas in this case they can be altered with out affecting each other. The fit of the Phase Type distribution with 3 phases can be seen in Figure 6.13. The parameters for this distribution are; $\lambda_1 = 0.02128$, $\mu_1 = 0.00202$, $\lambda_2 = 0.02632$, $\mu_2 = 0.05882$ and $\mu_3 = 0.00329$.

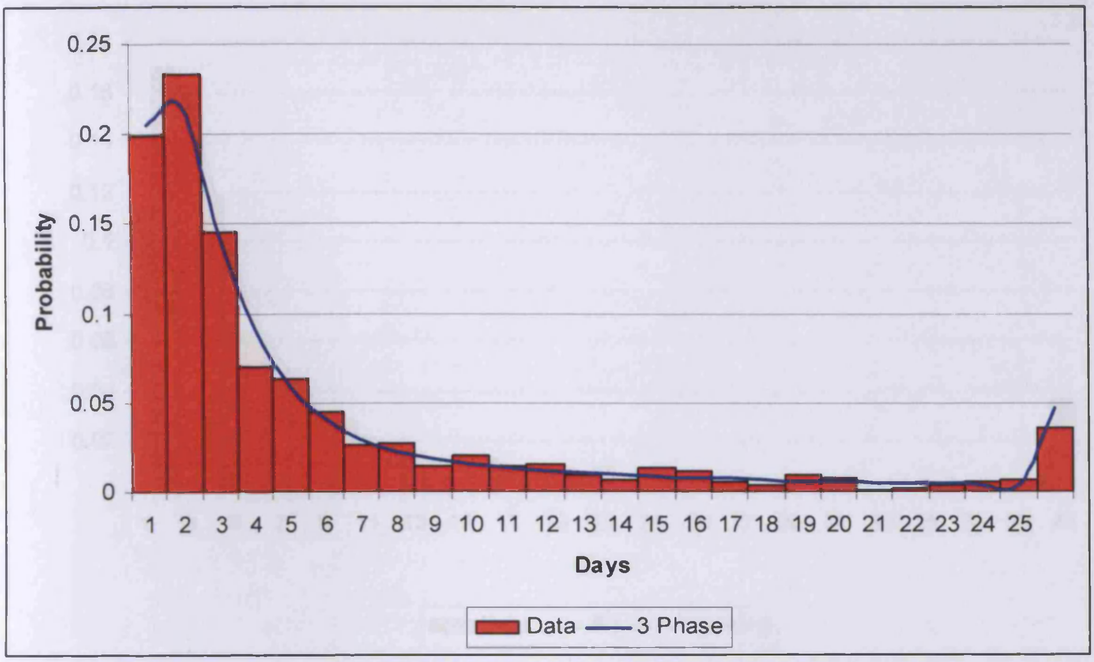


Figure 6.13 - Phase Type distribution estimation of the ill length of stay.

6.8 Model Analysis

6.8.1 Number of Beds

The models were all built using the Visual Basic program in Microsoft Excel. Patients go through different service points and their total length of stay is recorded. The results are shown in Figure 6.14. The results are based on averages of 50 trials of 100,000 patients. A warm up period of 1,000 patients was used.

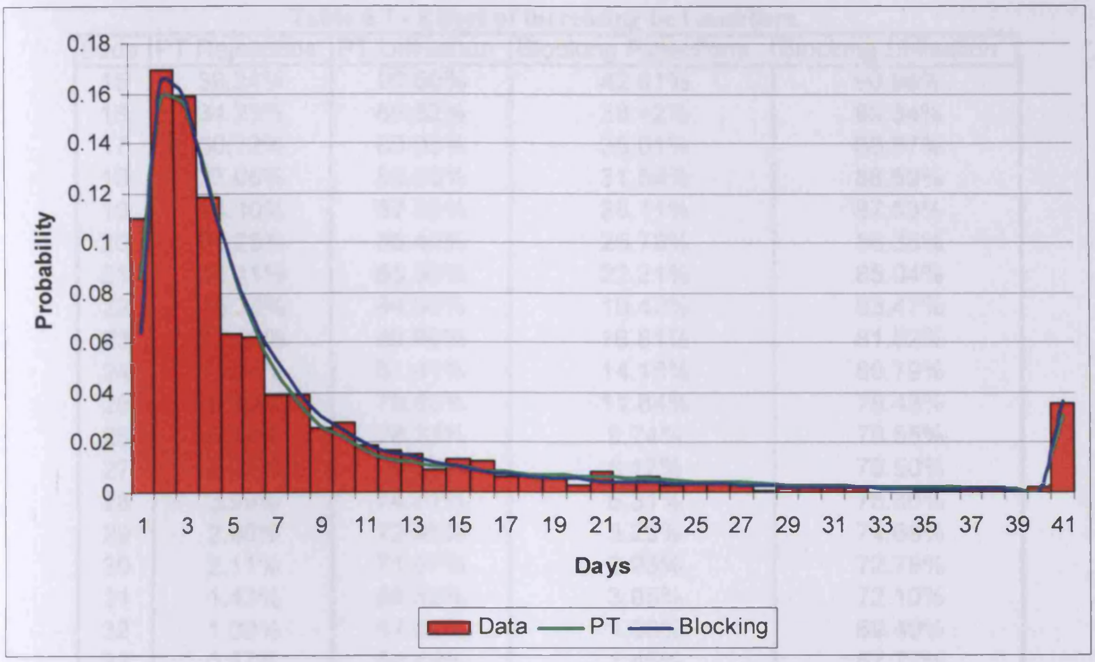


Figure 6.14 - Length of stay of different simulations over the total length of stay data.

We can see that both distributions capture the shape of the total length of stay very well. The chart has been compressed to be able to compare the results more easily. It can be clearly seen that both estimations of the data take into account the long tail of the data.

We will now alter some of the external factor variables in this situation to see how the two models predict what would happen to the Critical Care Unit. We shall start by altering the number of beds the Unit has to investigate the effect on the Utilisation of the Unit, and on the proportion of customers that would be rejected.

Table 6.7 - Effect of increasing bed numbers.

Beds	PT Rejections	PT Utilisation	Blocking Rejections	Blocking Utilisation
15	38.24%	90.60%	42.01%	90.96%
16	34.23%	89.82%	38.42%	89.34%
17	30.72%	89.05%	35.01%	88.87%
18	27.06%	88.05%	31.54%	88.69%
19	24.10%	87.32%	28.11%	87.53%
20	21.25%	86.46%	25.78%	86.35%
21	18.11%	85.30%	22.21%	85.04%
22	15.24%	84.05%	19.42%	83.47%
23	12.82%	82.98%	16.81%	81.52%
24	10.54%	81.41%	14.18%	80.79%
25	8.30%	79.62%	11.84%	79.43%
26	6.77%	78.33%	9.74%	78.55%
27	5.02%	76.20%	8.17%	76.50%
28	3.99%	74.71%	6.31%	75.80%
29	2.96%	72.95%	5.29%	74.68%
30	2.11%	71.07%	3.93%	72.79%
31	1.43%	69.32%	3.05%	72.10%
32	1.09%	67.62%	1.98%	69.40%
33	0.67%	65.48%	1.45%	67.77%
34	0.49%	63.91%	1.13%	66.27%
35	0.27%	62.13%	0.67%	64.69%
36	0.20%	60.77%	0.43%	63.41%
37	0.09%	58.78%	0.35%	61.57%
38	0.06%	57.49%	0.23%	59.42%
39	0.02%	55.81%	0.13%	57.60%
40	0.03%	54.42%	0.07%	56.94%

As expected, both the proportion of rejected customers and the utilisation of the ward decrease as the number of beds available increases. The rate that each of these changes across the two models as the number of beds increases can be seen in Figure 6.15. The effect of the Blocking model having a higher than average length of stay can be seen in this figure; the level of rejected customers is significantly higher in the simulation with lower numbers of beds.

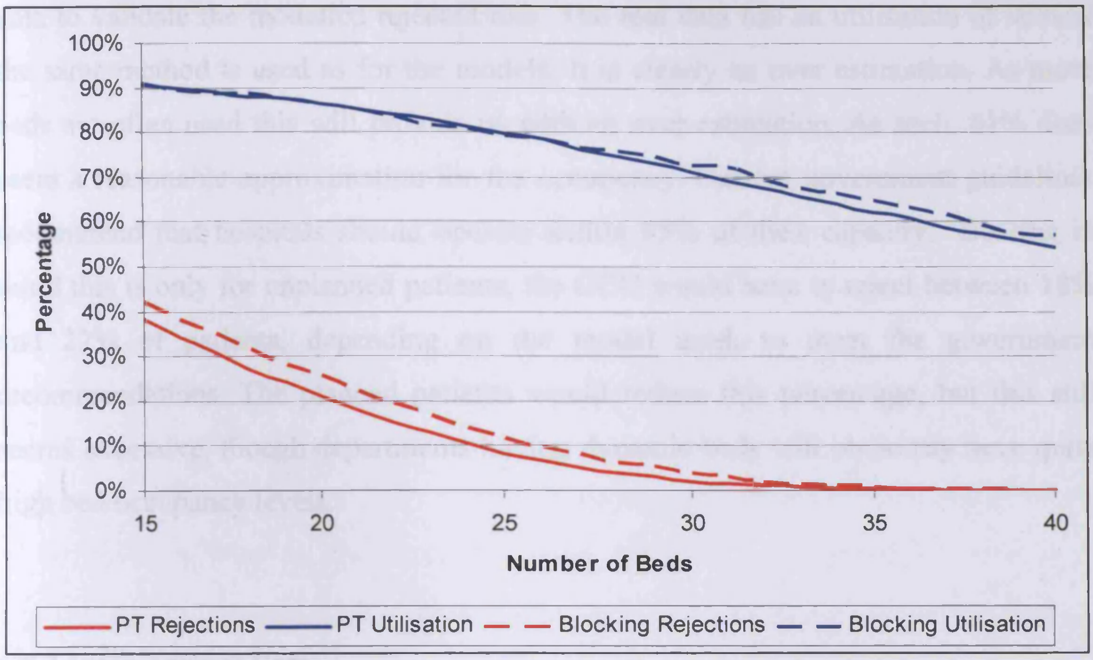


Figure 6.15 - Change in rejections and utilisation as the number of beds alters.

Currently there are 24 constant beds in the Critical Care Unit, with 6 extra that are available if required. These beds serve both planned and unplanned patients.

Using this blocking model it can be seen that with 24 beds, the utilisation of the department beds is around 81% with either model. This is only taking into account the unplanned patients. The utilisation has been defined as the sum of the length of stays, divided by the total length of the time the system runs, multiplied by the number of beds.

The percentage of patients that were rejected is slightly different for the 2 models. The Blocking model estimates it to be 14% whereas the Phase Type model says that about 11% will be rejected. As already discussed the Blocking model overestimates the average length of stay. This explains the difference between the two models in this variable.

Both projections will be an over estimate compared the real department. The CCU has a dynamic number of beds, and as such will have decreased the number of rejected patients by increasing the capacity if the demand rises. The data set is based on patients that where actually admitted to the CCU, and therefore we do not have any data to validate the modelled rejected rate. The real data has an utilisation of 95% if the same method is used as for the models. It is clearly an over estimation. As more beds are often used this will provide us with an over estimation. As such, 81% does seem a reasonable approximation for the occupancy. Current government guidelines recommend that hospitals should operate within 85% of their capacity. Bearing in mind this is only for unplanned patients, the CCU would have to reject between 18% and 22% of patients, depending on the model used, to meet the government recommendations. The planned patients would reduce this percentage, but this still seems excessive, though departments having dynamic beds will obviously have quite high bed occupancy levels.

6.8.2 Inter Arrival Rate

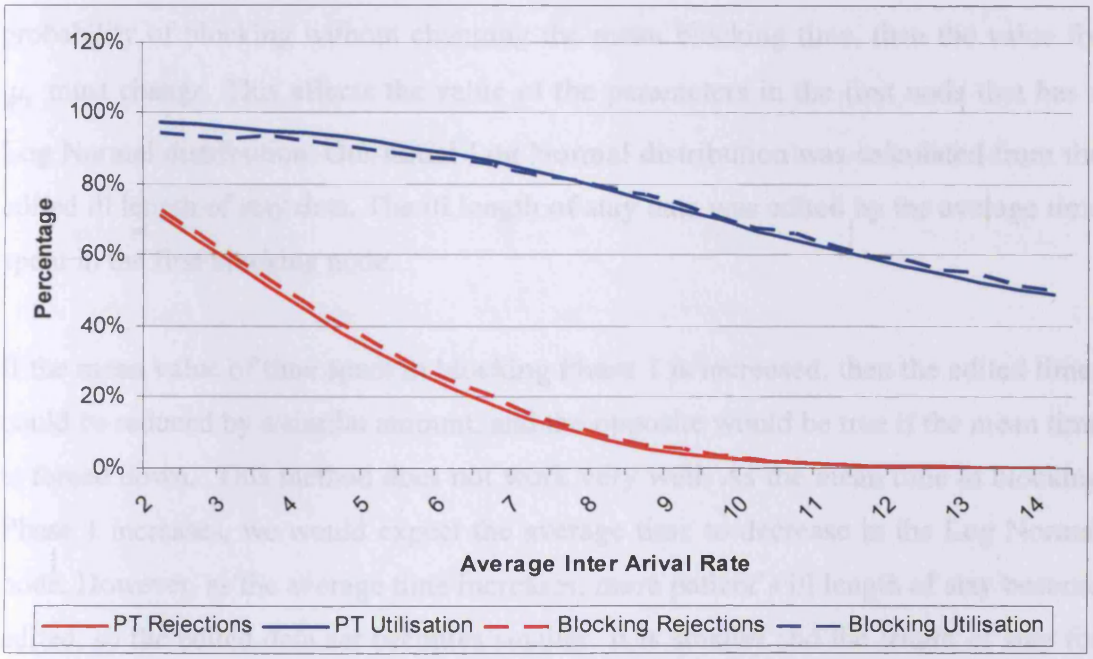


Figure 6.16 - Effect of changing Λ on a system with 24 beds.

We next address the question of ‘what if’ the arrival rate changes in the current system. Figure 6.16 shows the effect of changing the mean inter arrival time on the current system when that system is limited to 24 beds. Both models appear to behave similarly when changing the mean inter arrival time. The current mean inter arrival time of unplanned patients is 7.5 hours. We can see that if arrival rates increase even by a relatively small amount, the percentage of rejected patients increases dramatically. The utilisation is not affected as much. It would seem that the department could accept an increase in admissions so that the mean inter arrival time would be 6.5 hours. This would increase the number of rejected patients to 20% before the maximum utilisation recommended by the government is reached.

6.8.3 Probability of Blocking

It is more complex to test the effect of changing the proportion of patients that block, or the blocking time. As discussed, if the probability of blocking is fixed, then μ_1 is directly related to the value of μ_2. Likewise, if the value of either service rate is fixed, then there is a direct relationship between the two, and the probability of blocking. If we would like to see the effect of changing the effect of reducing or increasing the probability of blocking without changing the mean blocking time, then the value for μ_1 must change. This affects the value of the parameters in the first node that has a Log Normal distribution. Our initial Log Normal distribution was calculated from the edited ill length of stay data. The ill length of stay data was edited by the average time spent in the first blocking node.

If the mean value of time spent in blocking Phase 1 is increased, then the edited times could be reduced by a similar amount, and the opposite would be true if the mean time is forced down. This method does not work very well. As the mean time in blocking Phase 1 increases, we would expect the average time to decrease in the Log Normal node. However, as the average time increases, more patient's ill length of stay become edited, so the edited data set becomes smaller. It is smaller and the length of stay for these patients are significantly longer than for the main bulk of the patients, and so the average time spent in the log normal node does not change by much and not in the expected manner. A simple weighted mean approach to solve this problem was considered, taking the patients with a zero edited ill length of stay. They cannot be included in the calculation of the mean and standard deviation directly as the natural logarithm must be taken, which does not exist for zero values. The average time of the edited data was multiplied by the proportion of patients for whom a time is recorded.

All other patients return a time of zero for this data. The standard deviation was not altered due to the complex nature of mixing standard deviations of different distributions. This is not a very realistic assumption as when the number of patients with a zero blocking increases, the standard deviation should decrease. It would have affected the accuracy of the results. This method still proved to be unreliable. As less data points were used to calculate the values, they became more unstable. The results of this method were totally unrealistic so the method was abandoned.

A second method of altering the Log Normal distribution was required. To have a comparable result with previous simulations it would be useful to have a consistent average ill length of stay. As the probability of blocking changes, the value of the mean time spent in the first blocking nodes alters dramatically, the values can be seen in Table 6.8. To have the same average ill length of stay, the average time spent in the Log Normal service point would have to be changed by the same amount by which that the first blocking node changed. The average of the Log Normal distribution is given by $e^{\mu+\frac{\sigma^2}{2}}$. As we know what the average time should be, it is possible to rearrange this formula to give a value, in terms of the existing parameter σ and the required expected value; $\mu = \ln\left(E(X)\right)-\frac{1}{2}\sigma^2$. This approach requires keeping the same parameter for the standard deviation, as we did in the previous method, but this time a much smoother and more reasonable prediction is produced. The resulting graph for the Blocking model alongside the Phase Type model can be seen in Figure 6.17.

Table 6.8 - Change in average Length of stay in first blocking node as the probability of blocking changes.

Probability of Blocking	Average time
0.2	133.32
0.3	77.77
0.4	50.00
0.5	33.33
0.6	22.22
0.7	14.28
0.8	8.33
0.9	3.70

Figure 6.17 shows again that the two models perform quite similarly. Again the Blocking model has the percentages of both the utilisation and of the rejections slightly higher than the Phase Type distribution. This can be attributed to the higher average ill length of stay forecasted by the blocking method. The results show that by increasing the blocking rate to nearly 100%, a small increase in the utilisation of the unit of around 83% utilisation would occur. If the chance of blocking were reduced to 25%, then the utilisation would be reduced to 76% using the Phase Type model and down to 80% if the Blocking model is used.

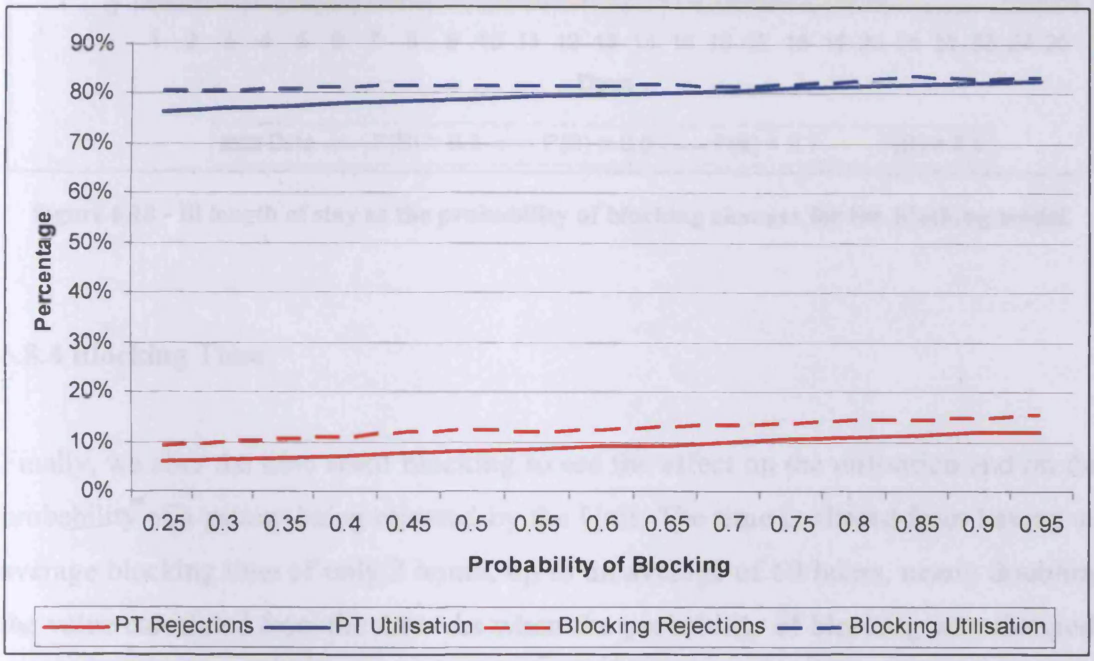


Figure 6.17 - The effect of changing the average blocked time.

The lowest probability of blocking that was considered was 25%; below this value the average time spent in the first blocking phase exceeded the patient's average ill length of stay. For this reason the values for the lower probabilities are less reliable for the blocking model. The first Blocking phase takes up a larger proportion of time, but as we have not altered the standard deviation in the Log Normal service point, there are some larger values than we may wish being produced. The change in the ill length of stay can be seen in Figure 6.18. The value for the probability of blocking gets further from the actual value, and the fit of the distribution becomes much worse. As discussed at lower values of the probability of blocking the variance shifts the shape of the distribution dramatically.

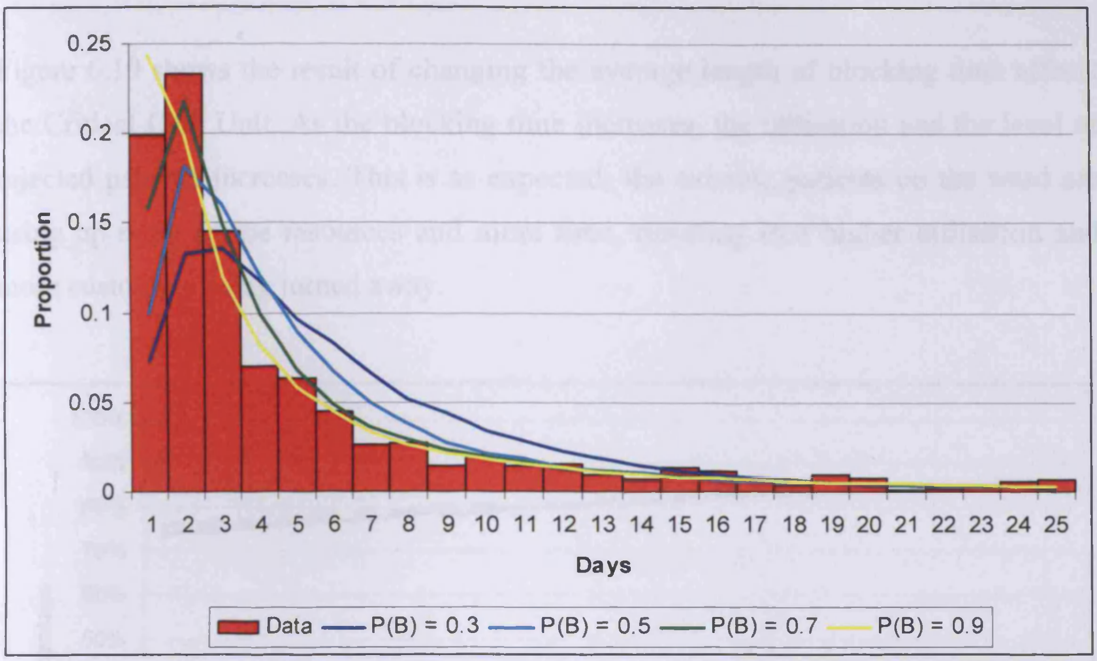


Figure 6.18 - Ill length of stay as the probability of blocking changes for the Blocking model.

6.8.4 Blocking Time

Finally, we alter the time spent blocking to see the effect on the utilisation and on the probability of a patient being rejected by the Unit. The time is altered from having an average blocking time of only 2 hours, up to an average of 60 hours, nearly doubling the value calculated from the data. As when the probability of blocking was changed, the Log Normal and the first blocking service point's parameter need to be recalculated for each change. The values of the average time required in the first blocking node to maintain a steady probability of blocking are shown in Table 6.9. The values can be seen to have a much smaller range than for the situation when the probability of blocking was altered.

Table 6.9 - How the first blocking phases average time changes as the average blocked time increases.

Average blocked time	Average node 1 time
2	0.57
10	2.87
20	5.73
30	8.60
40	11.47
50	14.33
60	17.20

Figure 6.19 shows the result of changing the average length of blocking time affects the Critical Care Unit. As the blocking time increases, the utilisation and the level of rejected patients increases. This is as expected; the existing patients on the ward are using up more of the resources and more time, resulting in a higher utilisation and more customers being turned away.

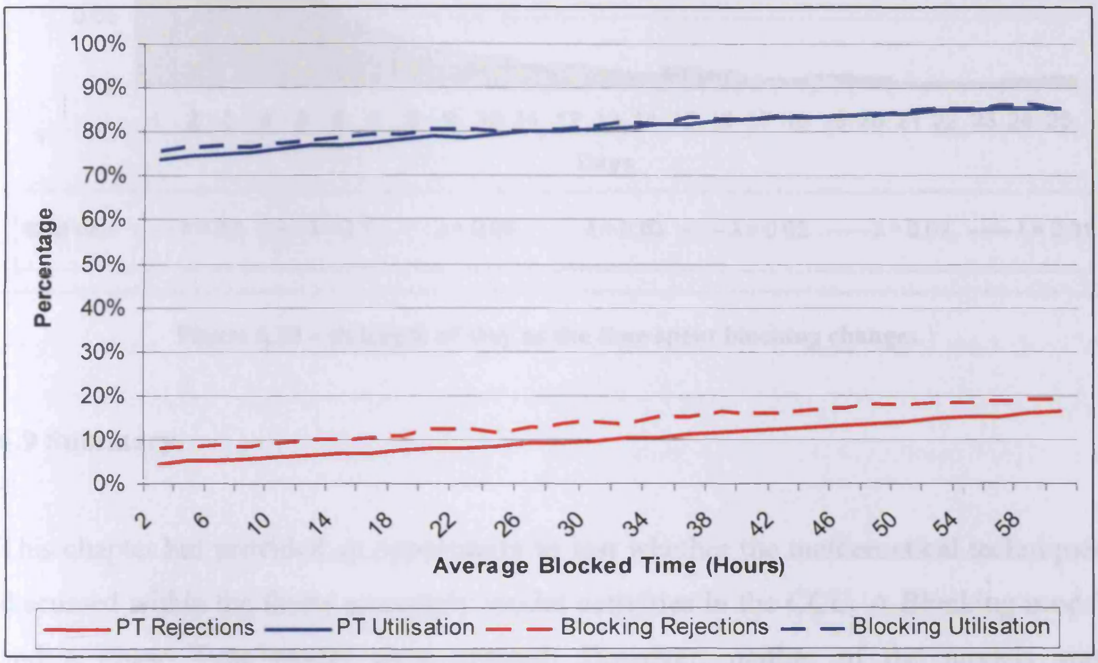


Figure 6.19 - The effect of change the time spent blocking.

As with the other analysis of these models, the Blocking model predicts a higher rejection value. The utilisation is closer to that of the Phase Type model. This is due to the parameter in the first blocking phase not having such extreme values to cope with the changes being demanded. The Ill length of stay distribution for the distribution in this situation is much more stable than those where the blocking probability changed, as can be seen in Figure 6.20.

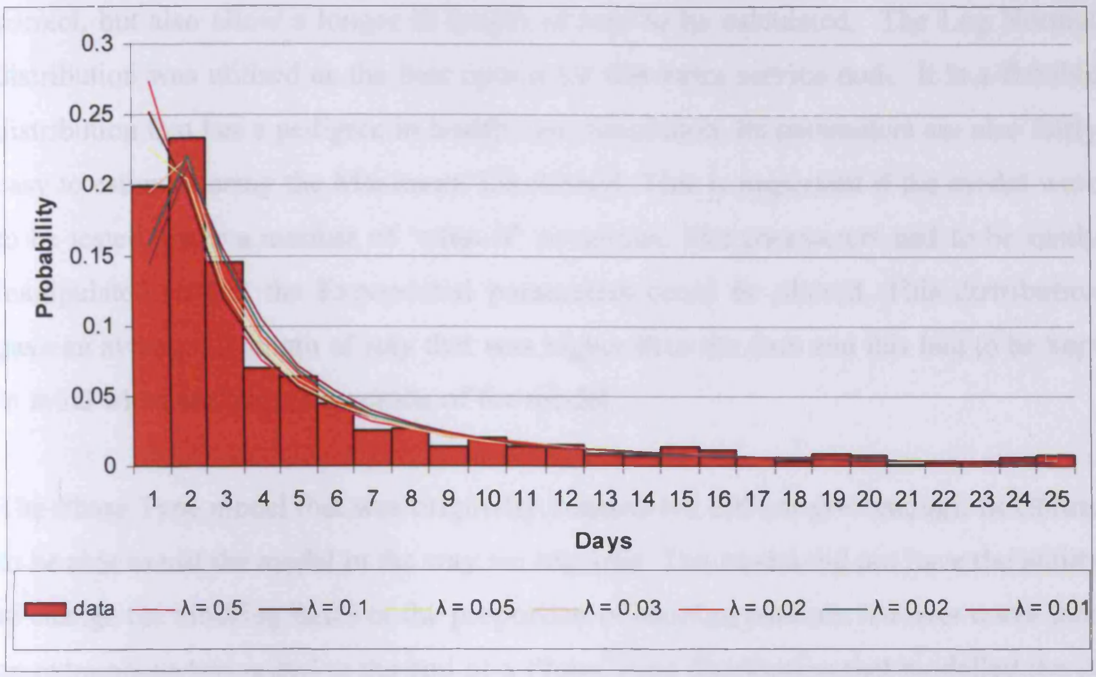


Figure 6.20 – Ill length of stay as the time spent blocking changes.

6.9 Summary

This chapter has provided an opportunity to test whether the mathematical techniques discussed within the thesis accurately model activities in the CCU. A Blocking model and a Phase Type model were created. However, neither of the models was satisfactory for modelling the real-life situation in their original form.

The blocking distribution was restricted by the fact that the probability of blocking is directly related to the two parameters that form the distribution. This was a problem as the data did not conform to these rules. A number of solutions were examined. It was possible to change the definition of when a patient was classified as blocking. This resulted in a very large value for the amount of time a patient was allowed to occupy the bed before they were officially classed as blocking, which would have been unacceptable to the hospital. The data also showed little sign of having an Exponential distribution, which is required for this blocking method. Changing the distribution in the first node was also considered. This would have made calculating the relationship between the parameters very complicated. The solution taken forward was to add an extra node in front of the blocking phases. This would enable us to fix the average length of stay in the first blocking node so that the probability of blocking was correct, but also allow a longer ill length of stay to be calculated. The Log Normal distribution was utilised as the best option for this extra service node. It is a flexible distribution that has a pedigree in health care simulation. Its parameters are also fairly easy to estimate using the Maximum Likelihood. This is important if the model were to be tested under a number of 'what if' situations. The parameters had to be easily manipulated so that the Exponential parameters could be altered. This distribution gave an average ill length of stay that was higher than the data and this had to be kept in mind when analysing the results of the model.

The Phase Type model that was originally constructed did not give enough flexibility to be able to test the model in the way we required. The model did not have the ability to change the blocking times or the proportion of blocked patients. To over come this, an extra phase was added to the end of a Phase Type distribution that modelled the ill length of stay. The extra phase was outside the Phase Type model and was used to model the blocked times. It was separate from the Phase Type model so that it could easily be manipulated without having to re-estimate all the other parameters in the distribution.

The two models were built in Visual Basic, so that they could be used by others. The models were both run for 100,000 patients, and averaged over 50 runs. The simulations were then analysed for a number of 'what if' questions. To start with, the numbers of beds available on the Unit were altered. The models results were shown by producing the Utilisation of the Unit and the proportion of patients that turned up to the department but did not get admitted. This is not how the real department behaves, as there are a flexible number of beds available. If the demand is increased, then more beds are made available so that patients are not rejected or forced to queue. The number of beds in the models created was not dynamic. This provided a more practical way to compare the results of changing different factors. As well as altering the number beds available to the department, the arrival rate was also altered.

We also wanted to change the probability of blocking and the average time spent blocking. To be able to do this, the parameters of the distributions in the blocking model had to be changed. This is due to the relationship between the parameters in the Blocking section of the model. A method was created which altered the average time spent in the Log Normal service point dependent on the average time spent in the first blocking phase. This was successful, in that the average time was now comparable for all probabilities of blocking changes in the average time spent blocking. However, the distribution of the ill length of stay time did not remain consistent, due to the large variance of the combined distributions. For this reason the blocking model was less reliable when the value of the average time spent in the first blocking phase gets too large.

Once this new method was established for the Blocking model, the probability of, and the average time spent blocking was altered and the results compared. Generally, the blocking model predicted a higher percentage utilisation and percentage of patients rejected across all the different 'what if' scenarios. This was expected due to the blocking method having a higher average ill length of stay than the Phase Type model, whose average ill length of stay was defined when the parameters were being calculated. The Blocking model could have been more comparable to the Phase Type results if the average time could have been fixed. As the method of Maximum Likelihood was used to estimate the parameters for this distribution, this was not possible. The Blocking model was also difficult to use due to the need for parameters for all distributions to be recalculated. The Blocking model does not appear to be the best technique to model this Critical Care Unit. It is more appropriate for situations where the interaction between the two phases is more directly related and changing the length of stay does directly affect the probability of blocking. The Blocking distribution is not totally inappropriate in this situation, as it did behave in a similar fashion to the Phase Type model. Though it did produce slightly higher estimates of occupancy and number of rejected patients due to the higher average ill length of stay. The idea of blocking was also incorporated into the Phase Type model, so that it was also altered to include a blocking phase on completion. This was a more appropriate technique, as the Phase Type distribution can accurately model the length of stay of patients and the extra phase can be altered more simply than for the standard Phase Type model, and without affecting the variance in the Blocking model.

The Phase Type model showed that if the maximum occupancy is set at 85% and if only unplanned patients are considered, then the unit could still keep within this limit.

If the average inter arrival time decreased, from the current 7.5 hours to 6.5 hours, it could also keep within this limit by lowering the current number of beds to 21. The results as stated are only for unplanned patients. In reality these alterations could not be made, as the Unit has a responsibility to serve patients that are critically ill after surgery. They do provide an insight into the demand of these patients on the current system, and how a change in behaviour may alter the effectiveness of the Unit. The effect of blocking can be seen in the possible utilisation of the Unit. If the problem of blocking were all but eliminated, and the average blocking time taken down so that the average time was only 2 hours over the official recommendation, then the systems occupancy dropped to 74%. This would free a lot of capacity for the processing of planned patients. A similar effect could be seen if only a quarter of patients, rather than the current 77.7%, become blocked.

These are results that the Unit could use to drive change. The models that have been created could be used to show that by reducing patients' blocking time, or the probability of them becoming blocked in the first place, then more planned patients could have access to the unit, and depending on restrictions to theatre times more operations could go ahead, thus leading to a reduction in the waiting list. If the theatres are all already running at capacity, then the cost of keeping patients in the one of the most expensive wards in the hospital could be reduced by taking these considerations into account.

Chapter 7

SUMMARY AND FURTHER WORK

7.1 Summary

This chapter provides a summary of the work presented in this thesis. Each of the chapters is reviewed, and finally the suitability of the models and results are discussed.

The data was provided by the University Hospital of Wales. The Critical Care Unit serves patients for many different reasons. These patients can be described as either 'planned' or 'unplanned' admissions. Planned patients are known about before their arrival at the Unit. They are elective surgical patients that require a stay within the Unit after their surgery, so that they can be closely monitored at a higher level, receiving the nursing attention that is not available on a general ward. Other patients that do not have planned admission can also come from the elective surgical cases. Not all Elective surgical cases require a stay in the Critical Care Unit after their surgery, but if a complication occurs then they may be admitted. These patients were not planned for by the Unit, but still require care. As well as these patients, unplanned cases can come from emergency surgery, A&E, other hospital, or from other departments inside the hospital.

The concept of blocking was introduced; patients are blocking if they are well enough to leave the Unit but for some reason were remaining there, using up a valuable, expensive and limited resource. This puts the Critical Care Unit under additional pressure. In the Critical Care Unit, a proportion of the patients had their blocked time recorded. This data was analysed and compared to the whole Unit's data. After a discussion with the Clinical Director of the Unit blocking was formally defined. A grace period of 7 hours was considered appropriate. Patients were unable to leave the Unit as soon as they had been referred, as provision had to be made for them and a bed found else where in the hospital. It was decided that 7 hours was adequate period of time for these arrangements to be made.

We then moved on to look at the different mathematical techniques that could be used, and applied to this situation to incorporate a blocking period. A methodology that considers blocking in the field of Operational Research maybe found in the Network of Queues system. This is when the output of a single queueing system leads into queueing system. When the queue size between these queueing systems is limited, service points can become blocked. If a service point completes its service, but there is no space for the customer to wait in the next queue, then they will be blocking their current service point.

The case when there is no queueing space in between the service points was investigated. This seemed applicable to the hospital environment, as when patients cannot queue when they are moved between departments, they must be moved from one bed to another. This situation was set up with time dependent queueing equations. These equations were then solved for the steady state situation. These results were then compared to a transient system that was built in Visual Basic.

These theoretical equations were set up for three cases: when there is a zero queue size in front of the first node, an '*n*' size queue, and infinite queueing space available. The through-put of these systems is of interest. The fact that blocking can occur would obviously decrease the productivity of the systems, as time that could be spent processing customers was being used up whilst customers were blocking service points. The 'drip feed' model was used to calculate the maximum throughput under the restricted conditions. The drip feed model never allows the first node in the Network of Queues to be empty. This ensures that the first service point is working at its maximum capacity and from this the maximum rate at which customers can arrive can be determined.

The probability of blocking in this system was used to calculate the average time spent in the first service point, and it was noted that this was different from the probability of a customer being blocked at any specific point in time. Using this value $\rho_{\max}$ was found. It was also noted that this is not the rate at which customer are processed through the entire system, but only the first service point. The capacity of the whole system was also calculated.

A simulation was set up within Visual Basic to examine what happens to the tandem queueing system with an infinite queueing space, around the value of $\rho_{\max}$ calculated from the 'drip feed'. It was seen that for values of average inter arrival time slightly under the value of $\rho_{\max}$, a queue does not appear to build up, whereas when the value increased over $\rho_{\max}$, a queue did build up. The probability of different queue sizes were also analysed and when the value of λ was less than $\rho_{\max}$, the most probable queue size was zero, whereas when the value increases above $\rho_{\max}$, the most probable queue size altered depending on the queueing space available and the time.

Theoretical summary measures were also calculated for these blocking queues, for both equal and unequal service times in each node. The average time spent in Node 1 as well as the whole system was calculated. The Probability Density Function was established. The PDF of time spent in the first node was shown to be the maximum of the two times calculated from the Exponential distributions in the two nodes.

The three node case was considered next in this theoretical analysis. The drip feed model was created with time dependent queueing equations. This system was solved for the steady state. When attempting to calculate the maximum utilisation by using the existing method, difficulties were found and a new methodology was established. This involved finding the proportion of time that the first node was working (not blocked) under the drip feed system. This is equal to the maximum rate at which customers could arrive at this service point and a value for $\rho_{\max}$ could be calculated. The same summary measures were calculated for this system as for the previous one.

This methodology was furthered by adding a second route that the customers could be served by. For theoretical reasons, it was important to be able to distinguish which route each customer was in. New notation was set up, and examples given of single queueing points with multiple servers. Then the time dependent equation of a two phase system, with two routes was set up. It was shown that there would be 5^n equations to solve when there are n routes. The computer program Maple was used to solve the 25 equations for two routes. A Visual Basic simulation was built and the results compared to the theoretical results from Maple. The transient results were also provided. The results found from both of these methods were compared well to the theoretical values calculated.

The drip feed equations were constructed and solved for the steady state. It was shown that for this system, the routes were independent. The maximum utilisation was found using the same method as for the single route system with three nodes. It was shown to be the same value as for the single route, multiplied by the number of routes available. It was also shown that for a non drip feed system, the routes are not independent. That is, the probability of the system being in State ij is not equal to the probability of Route 1 being in state i multiplied by the probability of Route 2 being in State j.

A method other than networks of queues was also considered, namely the Coxian Phase Type distribution. This distribution is a collection of Poisson processes, through which the customer must pass. The customer progresses from one phase to the next in order, and during the service at each phase, there is a probability of moving to an absorbing phase. When this occurs, the service is finished. If the customer moves through all the phases, then once the final service is over they will move to the absorbing phase.

The two phase system was initially considered. The probability of moving from Phase 1 to the absorbing phase was calculated, and an equation for the average time spent in the system computed. The time dependent equations for this system were solved so that the probability of a customer being in Phase i at time t could be calculated. This was then repeated for the 3 phase system, and then finally for the M phase system.

This gave generic equations, so that the probability of being in phase *i* at time *t* for a Phase Type system with any number of phases could be calculated. The probability of being in the absorbing phase at any particular time could be interpreted as the Cumulative Density Function. This states that as a customer enters the absorbing phase, their service comes to an end.

The Transitional Matrix approach of solving these equations was also discussed in this thesis. The approach was shown for the simple example of the M|M|1 queue and then used for the Phase Type distribution. The method involved using a Taylor Series expansion to solve the equations. The Transitional Matrix approach was applied to both the two and three phase systems and the result then compared to the time dependent equation approach.

Finally, parts of the mathematical techniques were used to build a simulation of the Critical Care Unit of the University Hospital Wales. The data for the models was collated. Only the unplanned arrivals were modelled; these are patients that the hospital has to take, and has little control over their arrival. The assumption behind modelling unplanned arrivals was that any extra capacity that could be found could be used to serve planned patients. The models provide an overview of results, as the models were run for long periods of time so that, general performance of the Unit could be analysed.

The average inter arrival time for unplanned admissions was calculated from the entire data set. This is due to the patients for whom delay data was recorded being a subset of the whole population. This provided a more realistic demand on the Unit. The distribution of the time spent blocking by these patients was then produced, along with the distribution for the patients' ill length of stay.

Appling the theoretical Blocking model discussed proved difficult. The inflexibility of the parameters meant that interpretation of the Critical Care Unit's data in a blocking model would have to be altered. The difficulties occurred because the probability of blocking is dependent on the average time spent in the two phases. A number of methods for altering the model were given, but it was decided that the best solution would be to add an extra service point through which the patient passed before they entered the blocking phase. The values for the probability of blocking and of the second parameter in the Blocking phase were calculated directly from the data. The probability of was blocking taken to be the proportion of patients that stayed on beyond the 7 hour grace period. The average time in the second phase, through which a dummy patient passed, had the same distribution as that of the blocked time of a patient. The first parameter was then calculated from these two known elements in the system. The service distribution for the extra node that was placed in front of this blocking system then had to be calculated. The Log Normal distribution was found to be the most suitable in this situation, and the parameters were estimated using the Maximum Likelihood method.

The Phase Type distribution was also altered slightly from the theoretical equations previously set up. The Phase Type distribution did model the patients' total length of stay very well, but it was inflexible, as the probability of blocking and the distribution of time spent blocking could not be altered. It was decided that the Phase Type distribution should be used to model the ill length of stay of the patients. Once the service was complete, patients would then have a probability of becoming blocked. Those that become blocked would then be served by an extra phase with the same parameter as the blocking model; those that did not exited the system straight away. The parameters for the Phase Type distribution were estimated by a program built in Visual Basic. The sum of the absolute values of the differences between the simulation and the data was calculated, and the smallest value was used as the best fit.

The two models were then put to the test under the current conditions of the Critical Care Unit. Both models predicted the distribution shape of the data very well, though the blocking model had a higher average length of stay than the Phase Type distribution, and the actual data. This is due to the fact that the Maximum Likelihood method was used to estimate the parameters for the Log Normal distribution, whereas the Phase Type distribution's average length of time was fixed to be the same as the data when the parameters for the distributions were being estimated. The models were then tested under a number of 'what if' scenarios. The metrics used to assess how the department might alter were: the utilisation of the Unit, and the proportion of customers that would be rejected. In reality the Unit does not reject patients. The flexible nature of the number of beds available to them in emergencies increased the number of patients allowed to enter the ward. It was used as a measure as it showed the demand that was put on the unit. The number of beds were altered, as well as changing the average inter arrival rate. The effects on the utilisation and the number of rejected patients were then compared between the two models. The Blocking model predicted higher results than the Phase Type model. This result was expected due to the higher average length of stay of this model. The fact that patients stayed in their beds for longer would stop other patients from being able to enter them during busy periods. Even though the two models had a discrepancy in the accuracy of their results, the general trend seen was very similar.

The models were then subject to change around the blocking time and the probability of blocking. To alter either of these in the Blocking model meant it was necessary to recalculate the parameters in the Blocking section and of the Log Normal distribution in the previous service point. By decreasing the probability of blocking, the average time spent in Node 1 had to increase, as the time spent blocking remained the same. Similarly, if the time spent blocking were to decrease, the amount of time spent in the first phase would have to decrease so that the probability of blocking remained constant. When the values for the average length of stay in the first blocking node changes, the average time in the Log Normal node also has to change. This caused more variable and less reliable results in the Blocking model than the Phase Type model. The Models were run for various values of blocking probabilities and times spent blocking and the results were compared. As expected the utilisation and the proportion of rejected patients decreased when the probability of blocking or the average blocking time decreased. These models enable to quantify the rate at which this could happen.

7.2 Further Work

The model of the Critical Care Unit could be extended by including planned admissions. This would provide a more precise model of how the Unit currently behaves. By including this aspect in the model, the mix of planned and unplanned admissions could be altered to see what would happen if the hospital attempted to clear a back log in the waiting list. Also, if an increase in the demand for the hospital's Critical Care Unit increased, possibly due to another local hospital reducing its capacity, the affect on the planned admissions would be seen more clearly if they were in incorporated in the model.

A further possibility would be to see what may happen if the hospital were to 'ring fence' some of their beds for planned cases. This would enable planned surgery to proceed more often. Currently, patients that require a stay in the Critical Care Unit after their surgery, must first be given the green light by the Critical Care Unit, before their surgery can take place, This ensures that there is adequate provision for them. The Unit is able to cancel planned operations that would result in shortages or over capacity. The possibility of 'ring fencing' beds could reduce the cancellations of operations, and help reduce the waiting list. 'Ring fencing' beds would of course produce a reduction in the capacity of the remainder of the Unit. However, with the inclusion of blocking in this model, it could be seen that this detrimental effect could be decreased by reducing the blocking which occurs, and thus increasing the capacity in the remaining beds.

The theoretical aspects of this thesis could also be developed further. The Blocking equations could be developed to include a queueing space between the service points, thereby increasing the capacity of the system. A further development that may help with the modelling of this Critical Care Unit would be to change the type of distributions used in the nodes. If the nodes could serve with an Erlang, Log Normal or any general distribution, then the model would become much more flexible, and better suited to a wider range of applications. This would complicate the relationship between the service points, but would definitely have benefits. The relationship between the two service rates and the probability of blocking have caused difficulties when applying this model to the Critical Care Unit, but this relationship could be used in other situations advantageously.

The Phase Type distribution already has a well established status in the field of healthcare modelling. Adding the extra node to the back of the system to account for the blocking is a useful adaption of the distribution. Including an extra phase, which can operate outside of the distribution, makes this very useful distribution even more flexible. This type of distribution successfully modelled the Critical Care Unit at the

University Hospital Wales. The current algorithm for estimation of the parameter of the Phase Type distribution uses an exhaustive method. This method could be adapted to create a smarter algorithm to cut down the calculation times. The distribution could be altered to take blocking problems of other hospital areas into account. A common problem in a hospital, is that elderly patients block ward beds (that they do not require for medical care), due to the limited availability of nursing home beds. The Phase Type model that was created could be adapted to this situation. A useful application would be to see the cost or capacity savings to the hospital if these patients could be processed with a smaller average blocking time, or a smaller chance of becoming blocked. This cost can then be compared to the cost of providing the extra space and of the extra cost that the hospital is currently experiencing by providing space for these patients. It could also be used to see the effect on waiting list times, as there would be more beds available for elective surgery.

REFERENCES

AHN, H., I. DUENYAS, ET AL. (1999). "OPTIMAL STOCHASTIC SCHEDULING OF A TWO STAGE TANDEM QUEUE WITH PARALLEL SERVERS." ADVANCES IN APPLIED PROBABILITY **31**(4): 1095-1117.

AKYILDIZ, I. F. (1988). "MEAN VALUE ANALYSIS FOR BLOCKING QUEUEING NETWORKS." IEEE TRANSACTIONS ON SOFTWARE ENGINEERING **14**(4): 418.

AKYILDIZ, I. F. (1989). "PRODUCT FORM APPROXIMATIONS FOR QUEUEING NETWORKS WITH MULTIPLE SERVERS AND BLOCKING." COMPUTERS, IEEE TRANSACTIONS ON **38**(1): 99-114.

AVI-ITZHAK, B. AND M. YADIN (1965). "A SEQUENCE OF TWO SERVERS WITH NO INTERMEDIATE QUEUE." MANAGEMENT SCIENCE **11**(5): 553.

BROWNING, S. G. (1998). "TANDEM QUEUES WITH BLOCKING: A COMPARISON BETWEEN DEPENDENT AND INDEPENDENT SERVICE." OPERATIONS RESEARCH **46**(3): 424.

BUHAUG, H. (2002). "LONG WAITING LISTS IN HOSPITALS." BMJ **324**: 252.

BURKE, P. (1956). "THE OUTPUT OF A QUEUING SYSTEM." OPERATIONS RESEARCH **4**(6): 699.

COSTA, A., S. RIDLEY, ET AL. (2003). "MATHEMATICAL MODELLING AND SIMULATION FOR PLANNING CRITICAL CARE CAPACITY." ANAESTHESIA **58**: 320.

COX, D. (1954). "A USE OF COMPLEX PROBABILITIES IN THE THEORY OF STOCHASTIC PROCESSES." PROCEEDINGS OF THE CAMBRIDGE PHILOSOPHICAL SOCIETY **51**: 313.

COX, D. AND H. MILLER (1965). THE THEORY OF STOCHASTIC PROCESSES. LONDON, METHUEN & CO LTD.

DAVIES, H. T. O. AND R. M. DAVIES (1987). "A SIMULATION MODEL FOR PLANNING SERVICES FOR RENAL PATIENTS IN EUROPE." JORS **38**(8): 693.

EL-DARZI, E., C. VASILAKIS, ET AL. (1998). "A SIMULATION MODELLING APPROACH TO EVALUATING LENGTH OF STAY, OCCUPANCY, EMPTINESS AND BED BLOCKING IN A HOSPITAL GERIATRIC DEPARTMENT." HEALTH CARE MANAGEMENT SCIENCE **1**(2): 143.

FADDY, M. J. AND S. I. MCCLEAN (1999). "ANALYSING DATA ON LENGTHS OF STAY OF HOSPITAL PATIENTS USING PHASE TYPE DISTRIBUTIONS." APPLIED STOCHASTIC MODELS IN BUSINESS AND INDUSTRY **15**(4): 311.

GORUNESCU, F., S. I. MCCLEAN, ET AL. (2002). "A QUEUEING MODEL FOR BED-OCCUPANCY MANAGEMENT AND PLANNING OF HOSPITALS." JORS **53**: 19.

Harper, P. (2002). "A Framework for Operational Modelling of Hospital Resources." Health Care Management Science **5**: 165.

Harrison, G. (2001). "Implications of Mixed Exponential Occupancy Distributions and Patient Flow for Health Care Planning." Health Care Management Science **4**: 37-45.

Hildebrand, D. (1968). "On the Capacity of Tandem Server, Finite Queue, Service Systems." Operations Research **16**(1): 72.

Hillier, F. and R. Boling (1967). "Finite Queues in Series with Eponential or Erlang Service Times - A Numerical Approach." Operations Research **15**(2): 286.

Hillier, F. and K. So (1989). "Thwe Assignment of Extra servers to stations in Tandem Queueing Systemswith Small or No Buffers." Perform. Eval **10**: 219-231.

Hillier, F. and K. So (1995). "On the Optimal Design of Tandem Queueing Systems with Finite Buffers." Queueing Systems **21**: 245-266.

Hunt, G. (1956). "Sequential Arrays of Waiting Lines." Operations Research **4**(6): 674.

Jackson, J. R. (1957). "Networks of Waiting Lines." Operations Research **5**(4): 518.

Jun, J., S. Jacobson, et al. (1999). "Application of discrete-event simulation in health care clinics: A survey." JORS **50**: 109.

Koizumi, N., E. Keno, et al. (2005). "Modelling Patient Flows Using a Queueing Network with Blocking." Health Care Management Science **8**(1): 49.

Marshall, A. and S. I. McClean (2004). "Using Coxian Phase Type Distribution to Identify Patient Characteristics for Duration of Stay in Hospital." Health Care Management Science **7**(4): 285.

Marshall, A., S. I. McClean, et al. (2001). "Developing a Bayesian blief network for the management of geriatric hospital care." Health Care Management Science **4**(1): 25.

Marshall, A., S. I. McClean, et al. (2002). "Modelling Patient Duration of Stay to Facilitate Resource Management of Geriatric Hospitals." Health Care Management Science **5**(4): 313.

McManus, M., M. Long, et al. (2003). "Variability in Surgical Caseload and Access to Intensive Care Services." Anesthesiology **98**: 1491.

MOUTZOUKIS, E. AND C. LANGARIS (2001). "TWO QUEUES IN TANDEM WITH RETAIL CUSTOMERS." PROBABILITY IN THE ENGINEERING AND INFORMATIONAL SCIENCES **15**(3): 311.

ONVURAL, R. AND H. PERROS (1989). "APPROXIMATE THROUGHPUT ANALYSIS OF CYCLIC QUEUEING NETWORKS WITH FINITE BUFFERS." IEEE TRANSACTIONS ON SOFTWARE ENGINEERING **15**(6): 800.

PAPADOPOULUS, H. AND C. HEAVEY (1996). "QUEUEING THEORY IN MANUFACTURING SYSTEMS ANALYSIS AND DESIGN: A CLASSIFICATION OF MODELS FOR PRODUCTION AND TRANSFER LINES." EJORS **92**(1): 1-27.

PERROS, H. (1984). "QUEUEING NETWORKS WITH BLOCKING: A BIBLIOGRAPHY." ACM SIGMETRICS PERFORMANCE EVALUATION REVIEW **12**(2): 8.

PINEDO, M. AND R. WOLFF (1982). "A COMPARISON BETWEEN TANDEM QUEUES WITH DEPENDENT AND INTERDEPENDENT TIMES." OPERATIONS RESEARCH **30**(3): 464-479.

PREATER, J. (2002). "A BIBLIOGRAPHY OF QUEUES IN HEALTH AND MEDICINE." HEALTH CARE MANAGEMENT SCIENCE **5**(4): 283.

PRICE-LLOYD, N. (2003). "STOCHASTIC MODELS FOR AN INTENSIVE CARE UNIT."

REICH, E. (1957). "WAITING TIMES WHEN QUEUES ARE IN TANDEM." THE ANNALS OF MATHEMATICAL STATISTICS **28**(3): 768-773.

RHEE, Y. AND H. PERROS (1996). "ANALYSIS OF AN OPEN TANDEM QUEUEING NETWORK WITH POPULATION CONSTRAINT AND CONSTANT SERVICE TIMES." EJORS **92**: 99.

RIDGE, J., S. JONES, ET AL. (1998). "CAPACITY PLANNING FOR INTENSIVE CARE UNITS." EJORS **105**: 346.

SPANGLER, W. E., D. P. STRUM, ET AL. (2004). "ESTIMATING PROCEDURE TIMES FOR SURGERIES BY DETERMINING LOCATION PARAMETERS FOR THE LOG NORMAL MODEL." HEALTH CARE MANAGEMENT SCIENCE **7**(2): 97-104.

STRUM, D. P., J. H. MAY, ET AL. (2000). "MODELLING THE UNCERTAINTY OF SURGICAL PROCEDURE TIMES: COMPARISION OF LOG NORMAL AND NORMAL MODELS." ANESTHESIOLOGY **92**: 1160-1167.

TAYLOR, B. (1989). "A NON-LINEAR MULTI_CRITERA PROGRAMMING APPROACH FOR DETERMINING COUNTY EMERGENCY MEDICAL SERVICE AMBULANCE ALLOCATIONS." JORS **40**(5): 423-432.

VASILAKIS, C. AND A. MARSHALL (2005). "MODELLING NATIONWIDE HOSPITAL LENGTH OF STAY: OPENING THE BLACKBOX." JORS **56**: 862-869.

WEISS, E. AND J. MCCLAIN (1987). "ADMINISTRATIVE DAYS IN ACUTE CARE FACILIITIES: A QUEUEING-ANALYTIC APPROACH." OPERATIONS RESEARCH **35**(1): 35-44.

WORTHINGTON, D. (1991). "HOSPITAL WAITING LIST MANAGEMENT MODELS." JORS **42**(10): 833.

YAMAZAKI, G., T. KAWASHIMA, ET AL. (1985). "REVERSIBILTY OF TANDEM BLOCKING QUEUING SYSTEMS." MANAGEMENT SCIENCE **31**(1): 78.