Estimation of Buffer Overflow Probabilities and Economies of Scale in ATM Multiplexers by Analysis of a Model of Packetized Voice Traffic

by

Paul J. Farrell

M Sc

A Thesis submitted to
Dublin City University
for the degree of

Doctor in Philosophy

School of Mathematical Sciences,
Dublin City University

January, 1999

Internal Supervisor Dr E Buffet

External Supervisor Dr N G Duffield

I declare that this thesis is based
on my own work

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor in Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work

Signed Paul Farrell

Paul J Farrell

Date 24/9/99.

Acknowledgements

I would like to thank my supervisor Dr Nick Duffield for all of his help and advice during the writing of this thesis and the preceding years

I would also like to thank Dr Emmanuel Buffet for his supervision in the early stages and for assistance at the end I would also like to thank Dr Jurgen Burzlaff for his assistance during the writing up

I would like to thank Kevin for all of his invaluable advice and help with the computer and for advice on the thesis, and Brendan and Grainne for their encouragement and questions about progress

Finally, I would like to thank my parents for all of their constant support, encouragement and advice throughout it all

Contents

List of Figures

Abstract

Estimation of Buffer Overflow Probabilities and Economies of Scale in ATM Multiplexers by Analysis of a Model of Packetized Voice Traffic

by Paul J. Farrell

We obtain upper bounds on the probability of buffer overflow for an ATM multiplexer of L identical packetized voice sources The multiplexer is modelled by a FCFS single server queue The arrivals at the multiplexer are a homogenous superposition of the arrivals from L independent identical sources, with each source modelled by a copy of a discrete time Markov Chain which we call the **Cell Level Model** Throughout, appropriate parameters are scaled with L, to maintain a constant load over all superposition sizes

The probability that, the queue-length (q_b^L) of the queue in a finite buffer exceeds the buffer size b, is bounded above by the probability that the queue-length (q^L) of the queue in an infinite buffer exceeds length b In order to bound the former above, we find upper bounds or approximations for the latter by using the theory of,

- Large Deviations, to determine its asymptotics for large b,
- Martingales, to obtain upper bounds, valid for all positive b,
- Large Deviations, to determine its asymptotics for large L for time rescaled (proportional to L) arrival processes

These demonstrate the *multiplexing gain* and *economies of scale* obtainable from **large and small buffers** and **large multiplexers**, respectively

Chapter 1

Introduction

1.1 Integrated Services Digital Network (ISDN)

The developement of ATM networks presents new problems in queueing theory, problems that cannot be solved by classical queueing theory ATM networks transmit packets of data or cells, as they have come to be called The asynchronous nature of the network means that in an ATM multiplexer cells from different sources can compete for available transmission bandwidth This leads inevitably to buffering of queueing cells awaiting transmission The nature of the arriving streams of cells at the buffer means that they cannot be modelled effectively by a Poisson process unlike the modelling of traffic at the call level arriving at an exchange in the classical theory as it is applied to teletraffic The arrivals (calls) in that case can be assumed to be independent and hence the arrival streams can be modelled by a Poisson process But the arrivals at an ATM multiplexer are highly correlated They are produced in bursts and are not well modelled by a Poisson process [1]

The asynchronicity and the need for queueing in an ATM multiplexer result in an inability to predetermine the delay that will be experienced by an arriving cell The delay will be random due to the randomness of the queue length A further difficulty arises, should the buffer be full when a cell arrives then that cell will be lost i e will not be transmitted The possibility of buffer overflow and cell loss has therefore to be addressed It is not possible to guarantee with certainty that the buffer will not overflow or that delays will not become excessively long. But it is possible to guarantee that neither will occur with a probability greater than some prescribed value. This value must be extremely small, of the order of 10^{-9}, because of the high transmission bandwidth. The problem arises of how to dimension the buffer in order to guarantee that buffer overflow will not occur with a probability greater than this. This problem is resistant to exact treatment because of the bursty nature of the arrival streams [1].

Because of the nature of the traffic as we have pointed out this problem cannot be solved by classical queueing theory; we need more recently developed techniques. The small probabilities mean we are dealing with rare events [1]. The theory of large deviations is a theory of rare events [2]. The correlations between arrivals mean we are dealing with non-independent arrivals. The theory of Martingales has been used to extend results applying to independent random variables to results for non-independent random variables [3].

We use the theory of large deviations and the theory of Martingales to obtain upper bounds or approximations for the tail of the queue length distribution in a model of a packetized voice ATM multiplexer that consists of a homogeneous superposition of independent sources feeding arrivals into an infinite buffer served on a first come first served basis. The traffic from each of the sources is modelled using a discrete time Markov Chain.

1.2 Asynchronous Transfer Mode (ATM)

Information and its transmission or communication is becoming of greater and greater significance in a shrinking and increasingly fast paced and changing world.

Information technology advances, such as the development of digital technology and the rapid development of optical fibre transmission technology, are leading to the development of high speed or broadband digital communications networks that are capable of providing new types of communication services [4].

These services would traditionally be carried by separate networks, specialised for the particular service that they carry. It is, however, more economical to integrate all of these services onto a single network, called an Integrated Services Digital Network (ISDN), thus avoiding the need for overlaying networks and allowing increased flexibility in the introduction and evolution of services This increased economy is motivating the development of the Integrated Services Digital Network (ISDN) [4]

The new services that will use such a network will involve the transmission of information in entirely digital form It is this fact, that all the different services are transmitted as digital streams, that makes integration possible The network need be able to handle nothing other than bit streams For example, in digital telephony, the initial analogue voice signal is sampled at discrete instants by digital technology which represents the sampled value in digital form The encoded signals are then divorced from the analogue waveforms of the source Thus, the digital transmission and switching equipment of a digital telephone or voice network is inherently capable of servicing any traffic of a digital nature The services, and the technologies used to provide them, are termed broadband, their speeds ranging from 1 Mb/s (10^6 bits per second) to 100 Mb/s and greater [4]

For transmission through the network, different bit streams are multiplexed onto a single Transmission medium to form a single bit stream The type of multiplexing used with such digital bit streams is termed Time Division Multiplexing (TDM) [4]

The large disparity between transmission bit rate, for example Gb/s (10^9 bits per second) where fibre optics and laser technologies are used in transmission, and the bit rate of, say, data terminals or telephones, which range from less than 1 kb/s to hundreds of kb/s depending on the encoding algorithms used, suggests that substantial economies should be achieved by using large multiplexers and also suggests the utilisation of new broadband services in order to capitalise on the increased bandwidth capability [4]

1.2.1 Multiplexing

The Transmission medium used for broadband services, and hence for the integrated services digital network, is optical fibre, because of its high bandwidth and transmission reliability Both constant bit rate and variable bit rate communications must be capable of sharing this transmission medium, and so the question of how this can be done efficiently and flexibly must be addressed.

The form of Time Division Multiplexing (TDM) that provides the greatest flexibility and the most efficient use of bandwidth, when different variable bit rate or bursty communications share the same transmission medium, is Asynchronous Time Division Multiplexing (ATDM) more commonly called Asynchronous Transfer Mode(ATM) [4].

Time Division Multiplexing (TDM) involves the multiplexing of different bit streams onto a common transmission medium in time slots of a predetermined length. There are two basic forms of TDM, namely, Synchronous Transfer Mode (STM) and Asynchronous Transfer Mode (ATM) [4, 5].

The STM approach involves the assumption of a common time frame of reference for all of the sources. Within this frame of reference each source is assigned its slot or slots. These slots, once assigned, are termed circuits and are said to be owned by the source, with each source having exclusive use of its assigned slot or slots in the reference frame [4].

In the case of single rate traffic, i.e. all the sources having the same bit rate, single slot TDM is used, with all the slots being of the same fixed length and, as the name suggests, exactly one slot being assigned to each source [4].

In the case of multi-rate traffic, slot assignment is more complicated, and multi-window TDM is used. For multi-window TDM the transmission channel capacity is divided into windows within the frame of reference, one window for each of the different bit rates, and each window is further divided into slots, each of the same fixed duration particular to that window and differing in duration from slots in other windows. One slot is assigned to each source in the window corresponding to its bit rate [4].

This multi-window TDM approach is inflexible for the following reasons. An initial standardized set of source bit rates must be chosen as must the number of sources of each bit rate type. But in a network with evolving services it may be difficult or impossible to predict what standard bit rates should be fixed and how many of them should be chosen The approach does not allow for the evolution of new services and is in this sense inflexible There are also difficulties in formatting and synchronising slot, window and frame lengths, and there can be as many windows as source types Sharing of capacity is inflexible as any change in the number of sources of more than one type can necessitate changing many window boundaries [4]

A more flexible form of TDM for dealing with multi-rate sources, is multi-slot TDM, where a source can be assigned, if needed, more than one slot in a reference frame A reference bit rate is chosen belonging to one of the source types We then proceed to derive the single-slot TDM format just as we would if all sources had this reference bit-rate Then any source with a bit-rate less than this reference rate is assigned one slot, and any source with a bit-rate greater than the reference rate is assigned the number of slots it requires, given its bit-rate Slots assigned then have a fixed owner within the frame of reference [4]

Bandwidth is wasted in this case by sources of bit-rate lower than the reference rate Choosing a small reference rate results in large numbers of slots, and high bit rate sources then require large numbers of slots, which all leads to greater complexity in tracking the assigned slots A balance must therefore be struck in this case between wasting bandwidth and increasing slot assignment complexity, and, in order to strike this balance, we again need to be able to predict the traffic mix, a task which may be difficult or impossible and which is in any case an imposition resulting from a lack of flexibility in the multi-slot TDM format [4]

The Synchronous Transfer Mode thus lacks flexibility when dealing with an evolving network This is due to having to choose a reference frame structure, both the choice and the structure itself are inherently incompatible with the flexible evolution of network services

The Asynchronous Transfer Mode abandons the idea of a common time reference frame for all of the sources Sources simply seize bandwidth when they have generated a sufficient number of bits The data to be transmitted is segmented into packets or cells of a fixed length Time is divided into slots of a fixed length, the length being the time taken to transmit a cell on the transmission line There is no reference frame, and hence each slot has no implicit owner, unlike STM where each slot is assigned an owner within the frame of reference. Thus in the ATM format each cell must be labelled. A cell then consists of data, plus a header, which contains the label that identifies the cells source and the time slot it was transmitted in. The header will contain other information, such as the type and priority of the data being transmitted, and possibly other routing information. The absence of a reference frame in an ATM network means new services can be introduced flexibly as they evolve. ATM also allows for more efficient use of bandwidth, particularly in the case of variable bit rate (VBR) sources, such as silence suppressed packetized voice sources, where speech activity detection is used. The use of ATM therefore provides for greater flexibility and more efficient use of bandwidth than the use of STM [4].

However, in the ATM format several sources may attempt to seize the same time slot, something which cannot happen with STM where slots are assigned to individual sources. If this occurs ,then, as only one cell can be transmitted per slot, one cell will be transmitted and the remaining cells will have to queue in a buffer until a slot becomes available for each. Thus, in the ATM format cells may suffer random delay depending on the length of the queue in the buffer, unlike the STM format where delay is fixed by the framing. Further, cells will be lost should the buffer capacity be exceeded by the length of the queue of waiting cells [4].

ATM thus allows for flexible sharing of a transmission medium, without the scheduling complexity of STM, but at the expense of this random delay, and the possibility of cell loss. Quality of service can only be guaranteed statistically in an ATM network [4, 5].

ATM is the multiplexing technique of choice for broadband ISDN primarily because of this flexibility, which is of importance due to the fact that it is not possible to predict what future traffic will be carried on the evolving ISDN. And due to the fact that different types of traffic, much of it from variable bit rate sources, will be carried on the ISDN [4, 5].

ATM multiplexing is used in the operation of ATM switches in the ISDN Network. The type of switching used is called packet switching, to distinguish it from circuit switching Circuit switching provides a dedicated path between two points carrying any information that fits into the available bandwidth Thus, for example, switching using STM is circuit switching, as slot ownership is assigned to sources within the reference frame, the slots being termed circuits [4]

1.2.2 Buffering and QoS

As the discussion of the previous subsection pointed out, competition for available transmission bandwidth is inevitable in an ATM multiplexer Buffering and buffer dimensioning are thus an essential part of the operation of an ATM multiplexer Should the queue become larger than the buffer can cope with, then cells will be lost, also excessively long queues can lead to unacceptable degradation of the service

The ATM forum proposed three parameters to measure the quality of service experienced by traffic as it passes through a queueing system [6] They are

- the cell loss ratio
- the mean cell delay
- the cell delay variance

These parameters depend on a fourth,

- buffer overflow frequency

What is required in order to guarantee a prescribed QoS is an upper bound on each of these parameters This requires us to have an upper bound on the tail of the queue length distribution for the queue in the buffer The queue length distribution for the queue in an infinite buffer will provide us with an upper bound on all of the parameters The reasons for this become clear when the meaning of the three parameters is explained The cell loss ratio is the ratio of the expected number of cells lost to the expected number of cells arriving at the queue The expected number of cells lost per tick is,

$$\mathbf{E}[\text{number of cells lost}] = \mathbf{E}[\text{number of cells arriving at a full buffer}]\mathbf{P}[\text{buffer overflow}]$$

The cell loss ratio is then,

$$\begin{aligned}\text{cell loss ratio} &= \frac{\mathbf{E}[\text{number of cells lost}]}{\mathbf{E}[\text{number of cells arriving at a full buffer}]} \\ &= \mathbf{P}[\text{buffer overflow}]\end{aligned}$$

The probability that the buffer overflows is bounded above by the probability that the queue length in an infinite buffer exceeds the buffer size. The mean cell delay can be bounded above in a similar manner, as the length of time that a cell will wait in the queue is approximately the length of the queue it finds on arrival at the buffer divided by the service rate. The cell delay variance can also be bounded once the tail of the queue length distribution is bounded. The typical values envisaged as upper limits on the cell loss ratio for an ATM multiplexer range between 10^{-8} to 10^{-11}. These are extremely small probabilities, so small that in many applications events with such probabilities of occurring would be regarded as never occurring [1]. As we stated earlier transmission bandwidth can be of the order of Gb/s (10^9 bits per second) if fibre optic technology is used [4]. But in a multiplexer with a transmission rate of one Gigabit per second and a cell loss ratio of 10^{-8} we would lose on average one cell per minute [1]. Thus we see the reason for requiring such low buffer overflow probabilities.

In order to be able to guarantee to the user a prescribed Quality of Service (QoS) for an ATM based ISDN it is thus necessary to approximate the tail of the queue length distribution of the queue at an ATM multiplexer or to be able to put an upper bound on the probability that such a queue will exceed any given length.

Chapter 2

Markov Chains

The models for an ATM multiplexer which we will be concerning ourselves with in this thesis involve modelling traffic from a single transmission source using Markov Chains. In this Chapter we outline the more important results and ideas relating to Markov Chains.

2.1 Stochastic Processes and Markov Chains

A stochastic process with state space E is a collection $\{X_t | t \in T\}$ of random variables X_t defined on the same probability space and taking values in E. T is referred to as the parameter set. If T is countable, for example, $T = N$, then the process is said to be a discrete parameter process. Otherwise it is referred to as a continuous parameter process. Usually t represents time, and X_t is thought of as the state of the process at time t [7].

Markov Processes are a particular class of stochastic processes, with the defining feature, that given the present state of the stochastic process the future evolution is independent of the past. A Markov process is called a Markov Chain if the parameter set is discrete.

Thus the stochastic process $\{X_n | n \in N\}$ on the state space E which we will take to be discrete is a Markov chain provided that,

$$\mathbf{P}[X_{n+1} = j | X_0, ...X_n] = \mathbf{P}[X_{n+1} = j | X_n] \quad \text{for all } j \in E$$

The probabilities

$$\mathbf{P}[X_{n+1} = j | X_n = i] = \mathbf{P}_n(i, j) \quad i, j \in E$$

are called the transition probabilities for the Markov chain X And the matrix of all such probabilities is called the transition matrix of the Markov chain

If,

$$P_n(i, j) = P(i, j) \quad \text{independent of n}$$

then the Markov chain is referred to as time-homogeneous or as having stationary transition probabilities [7]

2.1.1 Transition Matrix Properties

The transition matrix of a Markov chain X is a square matrix with the properties that all of its entries are positive, less than or equal to one, and that its rows sum to 1 i e ,

$$\sum_{j \in E} P(i, j) = 1 \quad \text{for each } i \in E$$

Given the initial distribution $P_0(i_0)$ the joint distribution of $X_0, \quad , X_n$ for any n is ,

$$\mathbf{P}[X_0 = i_0, \quad , X_n = i_n] = P_0(i_0)P(i_0, i_1)P(i_1, i_2) \quad P(i_{m-1}, i_m) \tag{2 1}$$

From this we can prove,

$$\mathbf{P}[X_{n+m} = j | X_n = i] = \mathbf{P}^m(i, j) \quad i, j \in E \text{ and } m \in N$$

and this in turn implies,

$$\mathbf{P}^{(m+n)}(i, j) = \sum_{k \in E} \mathbf{P}^{(m)}(i, k)\mathbf{P}^{(n)}(k, j) \quad i, j \in E$$

This is known as the Chapman-Kolmogorov equation, and it says that, starting in state i, in order to reach state j in exactly $m + n$ steps, X must enter some intermediate state k after m steps, and then reach state j from state k in n steps The right hand side of equation 2 6 above can also be written,

$$\mathbf{P}^m(i, j) = \mathbf{P}[X_m = j | X_0 = i] \tag{2 2}$$

and hence we can write, for a stationary Markov chain,

$$\mathbf{P}[X_{n+m} = j | X_n = i] = \mathbf{P}[X_m = j | X_0 = i] \tag{2.3}$$

This tells us that the evolution of the process X after time n, from fixed state i, is the same as the evolution of the process after time 0, from the same state i. In other words from all times of entry into state i, the process evolution will be the same, from that time, independent of that time, or how it reached this state i.

An important property of Markov chains is called the strong Markov property which holds for certain random present times T, instead of fixed present times, These random times have the property that for every time n the following holds,

$$I_{\{T \leq n\}} = I_{\{T \leq n\}}(X_0, \ldots.., X_n) \tag{2.4}$$

$$= \begin{cases} 1 & \text{if } T \leq n \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

Such a random time is called a stopping time or a Markov time. The occurrence or not of the event $\{T \leq n\}$ can be determined from the values of $X_0, ..., X_n$ alone. The occurrence of T does not anticipate the future evolution of the Markov Chain. The strong Markov Property then states, for any stopping time T,

$$\mathbf{P}[X_{T+m} | X_n; n \leq T] = P^m(X_T, j) \quad \text{for all } m \in N \text{ and } j \in E \tag{2.6}$$

$$\mathbf{P}[X_{T+m} = j | X_T = i] = P^m(i, j) \tag{2.7}$$

It tells us that the evolution of the Markov Chain starts afresh at time T if T is a stopping time [7].

Next we discuss the classification of the states of a Markov Chain.

2.1.2 Classification of States

In this section we describe how the states of a Markov chain are classified. This is covered in detail in [7]. The states of a Markov Chain are divided into classes according to properties of the time of first visit to the state, given that the Markov Chain is initially in the same state. These times are sometimes called the times of first return to the state or recurrence times for the state. The successive returns to a particular state constitute a recurrent event.

The question arises as to whether a return to a particular state is certain or not. And, further, if it is certain, the question of the finiteness of its mean recurrence time arises. States are classified according to the answer to these questions for each state.

Let T be the time of first visit to state j given that $X_0 = j$ then, state j is called recurrent if

$$\mathbf{P}[T < \infty | X_0 = j] = 1 \tag{2.8}$$

In words the state is recurrent if a return to the state is certain. Such a state will be visited infinitely often. State j is called transient if

$$\mathbf{P}[T = \infty | X_0 = j] > 0 \tag{2.9}$$

i.e. if a return to the state is not certain. Such a state will be visited only finitely many times; there will be a last visit to the state after which the state will not be entered again, hence the name transient. A recurrent state j is called null if

$$\mathbf{E}[T | X_0 = j] = \infty \tag{2.10}$$

Otherwise j is called non-null. Thus, recurrent states are called null or non-null, according as their mean recurrence times are infinite or finite respectively.

There is one further classification of the recurrent states of a Markov Chain, namely, whether or not a state is periodic or not i.e. whether $T \in \{\delta, 2\delta, 3\delta, ...\}$ with probability 1 for some integer $\delta > 1$, called the period of the state, or not. A state that is not periodic is called aperiodic.

All results pertaining to aperiodic states can be applied to periodic states j, with period δ, by considering the Markov Chain $\{Y_n | n \in N\}$, where $Y_n = X_{n\delta}$, in which j is aperiodic. The times of successive returns to a recurrent state j, given $X_0 = j$, form an increasing sequence of Stopping times. Thus, in particular, the Strong Markov Property holds at every element of the sequence.

We can classify sets of states according to whether states outside them can be reached by states inside the set or not. A state j can be reached from a state i if there exists an integer $n \geq 0$ such that $P^n(i,j) > 0$ If a state can be reached from another state, then, there is a sequence of intermediate states, each of which can be reached from the preceding state in one time step

A set of states is said to be closed if no state outside it can be reached from any state inside it Such a set containing only one state is called an absorbing state A closed set containing no closed proper subsets is called irreducible And a Markov Chain is called irreducible if the set of all states is irreducible A simple criterion for determining whether or not a Markov Chain is irreducible follows directly from these definitions

A Markov Chain is irreducible if and only if every state can be reached from every other state

Inspection of the transition matrix or the transition diagram for the Markov Chain will tell us if the chain is irreducible or not This follows from the fact that if j can be reached from i and k can be reached from j, then k can be reached from i Thus, by inspection of the Transition matrix, in an iterative fashion, using this fact we can find all the closed sets, and in particular, determine whether or not the chain is irreducible

In fact if we find a closed set C, and delete from the transition matrix all rows and columns corresponding to states not in C, then, the resulting matrix is again a Markov matrix, in fact it is the transition matrix for the Markov chain with state space C

From a recurrent state only recurrent states can be reached The reason for this is as follows Say, for example, that from a recurrent state j it is possible to reach a state i i e there is a positive probability of going from j to i in a finite number of steps Then, in order for j to be recurrent, the probability of going, eventually, from i to j, must be 1 After visiting state i, the chain will eventually visit state j, after which it will return to j infinitely often But after each visit, there is the positive probability that it will visit i, thus i will be visited infinitely often, i e i is also recurrent

The set of all recurrent states of a Markov chain is a closed set, and can be divided, in a unique manner, into irreducible closed sets This is because, from a recurrent state only recurrent states (and not transient states) can be reached, hence the set of all recurrent states must be closed The division into irreducible closed sets follows from the fact that the set of all states that can be reached from a fixed state j is by definition a closed set, and if that fixed state is a recurrent state, then it can be reached from any state in this set, and hence, every state in this set can be reached from every other state in this set via the intermediate state j, implying that the set is irreducible The set is unique for j, hence, the partition of the set of all recurrent states is unique

For an irreducible Markov Chain we can thus say that all states are transient, or all states are recurrent Since the set of all recurrent states is an irreducible set, it must be the set of all states or it must be the empty set, and in the latter case all states must be transient We can go further and say that if all states are recurrent then they must all be recurrent null, or all be recurrent non-null, and further, all must be aperiodic or all must be periodic with the same period

For an irreducible finite closed set C we can add that there are no recurrent null or transient states If there were one recurrent null state, then by the earlier statements all states would be recurrent null From any state in C, and for any given number of time steps, we can say that the probability of a transition to some state in C from that state, in that number of time steps, is 1 This is because C is irreducible But, If every state in C were recurrent null, then there would be some time step for which all of the transition probabilities from that state were $< 1/N$, where N is the size C, contradicting the last statement Put simply, in order for the chain to spend a fantastically long time between returns to every state in the set, it would have to leave the set, as the set is finite, but this is impossible, as the set is an irreducible closed set Similarly, if one state were transient, all states would be transient, due to irreducibility, and again, in order for the chain to leave all of these states, which are finite in number, never to return, the chain would have to leave the closed set

We can now say that in any Markov Chain with a finite state space, there are no recurrent null states and not all states are transient

In order to classify the states of a Markov Chain with finite state space it is necessary to first identify the irreducible closed sets, then all states belonging to these sets are recurrent non-null, and all other states are transient In order to determine if the recurrent state j is periodic or not, we simply find the greatest common divisor of the set of all $n \geq 1$, such that, $P^n(j,j) > 0$ If this is equal to 1, then the state is aperiodic, and hence, all recurrent states are aperiodic If it is greater than 1, then it is the period of the state, and of all the recurrent states In order to determine the g c d of the set of all the above n, it is necessary to look only at the sequences of states through which the chain can pass to return to j, and to count the number of states in enough of the sequences to be able to evaluate the g c d and hence the period For example if two such sequences differ in length by 1, then all of the states are aperiodic, as the g c d of the set of all $n \geq 1$, such that $P^n(j,j) > 0$, is 1

2.1.3 Limiting probabilities and the invariant measure

In 2 1 we described how the states of a Markov Chain may be classified The usefulness of this classification is due essentially to the fact that we can restrict our attention to states of one particular type This is always the case for irreducible Markov Chains, and we showed how the states of Markov Chain with finite state space can be easily classified The chains which we will be dealing with will have finite state spaces, but we will state the following theorem which can be used to classify the states of a Markov Chain with infinite state space

The theorem is of interest to us for another reason A corollary to this theorem, which applies to irreducible aperiodic Markov Chains (ergodic) with finite state spaces, guarantees the existence of a so-called invariant distribution for such Markov Chains [7] This invariant distribution gives us the probability, as n tends to infinity ,that starting in an initial state i we are in state j, n steps later This probability depends only on j, it is independent of the initial state i In the long run the chain forgets the initial distribution The independence from the initial state also means that the absolute probability of being in state j also tends to the invariant probability of being in this state The process settles into this invariant distribution The term invariant derives from the fact that if the the Markov Chain has this initial distribution, then it will have this distribution at all subsequent times The invariant distribution is also called the stationary distribution, and is sometimes referred to as the equilibrium distribution, this referring to the equilibrium reached by a large ensemble of such identical processes, where the number of processes in any state, at any time, tends to a constant, for large enough times, with the proportion of processes in that state being (approximately) the stationary probability of being in that state [8]

We now state the theorem and its corollary [7]

Theorem 1 *Let X be irreducible and aperiodic Then all states are recurrent non-null if and only if the system of linear equations,*

$$\pi(j) = \sum_{i \in E} \pi(i) P(i, j) \tag{2.11}$$

$$\sum_{j \in E} \pi(j) = 1 \tag{2.12}$$

where $j \in E$, has a solution π If the solution exists, then it is strictly positive and there are no other solutions, and further,

$$\pi(j) = \lim_{n \to \infty} P^n(i, j) \tag{2.13}$$

for all $i, j \in E$

Corollary 1 *If X is an irreducible aperiodic Markov Chain with finitely many states, then*

$$\pi(j) = \sum_{i \in E} \pi(i) P(i, j) \tag{2.14}$$

$$\sum_{j \in E} \pi(j) = 1 \tag{2.15}$$

has a unique solution The solution π is strictly positive, and

$$\pi(j) = \lim_{n \to \infty} P^n(i, j) \tag{2.16}$$

for all $i, j \in E$

Thus an irreducible aperiodic Markov Chain with finite state space has an invariant distribution. It is possible to interpret the invariant distribution probability for a state j as the rate at which state j is visited [7].

If we write $m(j)$ for the mean recurrence time for j then, we can say the following [7],

Proposition 1 *Let j be an aperiodic recurrent non-null state. Then,*

$$\pi(j) = \lim_{n\to\infty} P^n(j,j) = \frac{1}{m(j)} \tag{2.17}$$

We also have the following [7],

Proposition 2 *Let j be a recurrent non-null aperiodic state, and let $\pi(j)$ be as before. Then,*

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} I_j(X_m) = \pi(j) a.s. \tag{2.18}$$

This tells us that the fraction of time spent by the chain in state j is $\pi(j)$ and it has the following corollary [7],

Corollary 2 *Let X be an irreducible recurrent Markov Chain with stationary distribution π. Let f be a bounded function on E. Then,*

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} f(X_m) = \sum_{j\in E} \pi(j) f(j) \quad a.s. \tag{2.19}$$

And a corollary to this is [7],

Corollary 3 *Let X be an irreducible recurrent Markov Chain with stationary distribution π. Let f be a bounded function on E. Then,*

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} \mathbf{E}[f(X_m)] = \sum_{j\in E} \pi(j) f(j) \quad \textit{independent of } i \tag{2.20}$$

We could summarise some of the more important facts concerning Markov Chains that we have described in 2.1.2 and 2.1.3 as follows. An irreducible Markov Chain has at most one invariant (stationary) distribution. Its states are either all transient, all null recurrent or all non-null recurrent. Further, all states are periodic or all are aperiodic. They are all non-null recurrent if and only if the chain admits one invariant distribution, and this is certainly the case if the chain has finite state space.

2.1.4 Reversed Markov Chains and reversed time

Up to this point we have been considering Markov Chains where we know something about the "present" state of the chain, and we wish to know something about the "future" states But, in some situations, it is desirable to know something about the "past" developement of the Markov Chain given knowledge of the "present"

If a Markov Chain has a stationary distribution, then it behaves as a Markov Chain if its evolution is observed in reversed time If the stationary distribution of the Markov Chain is π, then the transition matrix for the reversed chain has transition probabilities related to the transition probabilities of the forward chain in the following manner [8],

$$Q(i,j) = \frac{\pi(j)P(j,i)}{\pi(i)} \tag{2 21}$$

The nth step transition probabilities for this chain can be calculated in exactly the same fashion as they are calculated for the forward chain They are related to the nth step transition probabilities of the forward chain as follows,

$$Q^n(i,j) = \frac{\pi(j)P^n(j,i)}{\pi(i)} \tag{2 22}$$

Thus, the study of the past developement of the original Markov Chain reduces to the study of the reversed chain In the special case when $Q(i,j) = P(i,j)$ the chain is said to be time-reversible, and all the probability relations for such a chain are symmetric in time It can be shown that if $P(i,j) > 0$ if and only if $P(j,i) > 0$ and if all sets of such pairs (i,j) are reachable from all others, then the chain is time reversible

Chapter 3

The Models

We described in chapter 1, broadly the operation of ATM. We will now describe in greater detail the situation we wish to model.

We begin by assuming that we have L independent sources or lines. Traffic from all of these sources is to be multiplexed onto a single transmission line using asynchronous time division multiplexing. Therefore we discretise time into fixed length slots. Each source produces digital information in the form of fixed length packets called cells. A single cell can be transmitted in each time slot. Thus the length of a time slot is equal to the transmission period of the multiplexer. Cells arrive at the multiplexer at the beginning of a time slot and are transmitted at the end of a time slot. In practice a cell is of size 48 bytes with an additional 5 bytes for the header giving a total size of 53 bytes (424 bits) [5].

3.1 Packetized Voice

With packetized voice a continuous time signal is generated by the source and is digitally sampled at discrete instants. The standard sampling rate used in digital telephony is 8 kHz and a sample is 8 bits thus giving us a constant bit rate of 64 kb/s. This digital information is then filled into the fixed length packets. Information is not sent by the source to the multiplexer until a packet has been filled. This process introduces its own fixed delay or packetization period [9]. It is measured in units of the multiplexer transmission period and we will label it s Thus s will be the ratio of the packetization period to the multiplexer transmission period

Speech contains, typically, 50 percent silence and hence digital telephony uses silence detection and suppression [5] This means that no cells are generated during periods of silence This produces bursty traffic from the source, what would otherwise be a constant bit rate output becomes a variable bit rate output, with periods of activity during talk spurts alternating with periods of inactivity during silences We will refer to active periods as burst periods and inactive periods as silences A burst period begins with the arrival of a cell at the multiplexer and it ends s ticks after the arrival at the multiplexer of the last cell in the burst This period of s ticks after the last cell of a burst is referred to as the overhang period One tick after this period has ended a silence is said to have begun, and it continues until the next burst begins with the arrival of another cell at the multiplexer Thus burst periods are multiples of s ticks in length A burst period then, consists a period of cells arriving periodically every s ticks, plus the overhang period of s ticks We illustrate this in figure 3-1

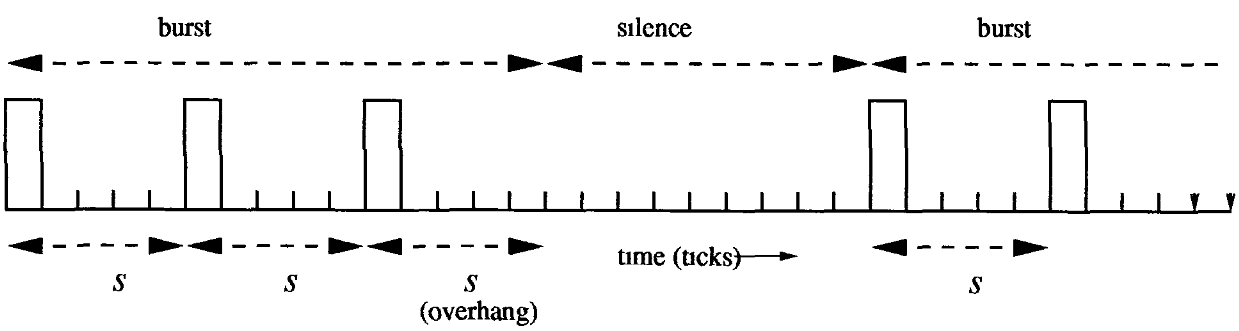

Figure 3-1 *Sample of traffic on a single line*

The group of cells arriving during a burst period we will refer to as a burst The length of a burst will then be measured as the number of cells in the burst Bursts and silences are of random length During normal conversation the duration of talk-spurts fits the exponential distribution reasonably well while the duration of inactive periods is approximated less well by the exponential distribution But we will assume, as others have, that both active and inactive (real-time) periods are exponentially distributed [10]. Thus bursts and silences are of random length and, since we are operating on a discrete time scale, we assume that burst and silence lengths are each geometrically distributed, and that burst lengths are i.i.d. and silence lengths are i.i.d., and further that silence and burst lengths are independent of each other. We will define the probability that a burst continues for another cell to be α i.e.

$$\alpha = \mathbf{P}[\text{burst continues}] \tag{3.1}$$

and we define the probability that a silence continues for another tick to be β i.e.

$$\beta = \mathbf{P}[\text{silence continues for another tick}] \tag{3.2}$$

Then,

$$\mathbf{P}[\text{burst} = n \text{ cells}] = (1-\alpha)\alpha^{n-1} \tag{3.3}$$

$$\mathbf{P}[\text{silence} = m \text{ ticks}] = (1-\beta)\beta^{m-1} \tag{3.4}$$

and hence the expected numbers of cells in a burst, and the expected number of ticks in a silence are,

$$\mathbf{E}[\text{burst length}] = \frac{1}{1-\alpha} \tag{3.5}$$

$$\mathbf{E}[\text{silence length}] = \frac{1}{1-\beta} \tag{3.6}$$

3.2 Superposition of Packetized voice sources

In the situation that we are modelling, there will be L independent identical packetized voice sources. The arrival process at the multiplexer will be the superposition of the L individual arrival processes.

The multiplexer serves one cell per tick, and the ratio of the packetization period to the multiplexer transmission period is s ticks. Thus s active voice sources will just saturate the multiplexer, or to put it another way, the multiplexer will be not be idle at any time during a period of length at least equal to the length of the shortest burst period among the s active sources. And, during the period when s sources are active, at the end of each interval of s ticks starting from the time of the first arrival, an arriving cell will find the buffer empty. In other words over this period when s sources are active, cells will not accumulate in the multiplexer buffer. Conversely, during periods where more than s sources are active, cells will accumulate in the multiplexer buffer; the length of the queue in the multiplexer buffer will grow over this period. After s ticks from the time of the first arrival, all arriving cells will find the buffer occupied. However, once the number of active sources falls below s the back-log of cells in the buffer can be cleared. And, during any such period when the number of active sources is less than s, the buffer queue length decreases to zero, except for periodic fluctuations, after which, arriving cells find the buffer to be empty. These high frequency periodic fluctuations are the only contribution to the queue length in the multiplexer buffer when the number of active sources is less than or equal to s, and are simply due to the simultaneous arrival of two or more cells at the multiplexer. In other words, these high frequency fluctuations in queue length are due to an instantaneous increase in the cell arrival rate at the multiplexer above the service rate, as opposed to a temporary increase in the average arrival rate, over a period of time greater than s, to above the service rate, as occurs in the case where the number of active sources exceeds s. The former leads to short length, short term queues, the latter can result in long queues lasting long periods of time. Large queues are also subject to these small high frequency changes. But these fluctuations are unimportant in the growth of large queues. They are however, an important consideration in buffer dimensioning. In order to accurately estimate the queue length distribution, the contribution to the queue-length from the cell-level component must be taken into account. The queue in the multiplexer buffer can be viewed as having two components: the burst-level component and, added to it, a cell-level component.

When the number of active lines over a period greater than s ticks exceeds s or equivalently when the average arrival rate over a period greater than s ticks exceeds the service rate, we are in what is referred to as a heavy traffic regime. A heavy traffic regime results in long term large queues. This is referred to as burst level congestion. When the number of active lines over a period greater than s is less than s then we are in a low traffic regime and we have short term, short length queues i e high frequency fluctuations of small amplitude in small queue-lengths This is referred to as cell-level congestion A defect m many of the models used to study ATM multiplexer congestion has been the failure of the models to properly take these high frequency, low amplitude queue-length fluctuations into account For example Daigle and Langford [10] describe three models A Semi Markov Process model, a Continuous time Markov Chain model, and a Uniform Arrival and Service model The first and last of their models ignore the high frequency fluctuations completely, and as a result, underestimate cell level congestion by assuming no queue length changes under a low traffic regime Their Continuous time Markov Chain model overestimates cell level congestion by implicitly assuming higher frequency fluctuations in queue length than can occur in the system being modelled These observations were confirmed by their simulations

Each individual source generates an arrival process that is in fact an alternating renewal process But because of the bursty and periodic nature of individual sources, the superposed arrival process is not a renewal process, the inter-arrival times of the superposed process are negatively correlated and the average arrival intensities over periods longer than s are positively correlated Only a superposition of Poisson processes is itself a renewal process, in fact again a Poisson process, and modelling the superposed process by a Poisson process has been shown to be unsatisfactory [11]

3.3 Modelling traffic from a single source

In this section we will describe how we model traffic from an individual source The model we introduce in 3 3 1, the **Cell Level Model**, is new and is the model which we will use throughout the thesis The model we describe in 3 3 2 the **Block Level Model** has been studied in detail by Buffet and Duffield [3] We will make a comparison between the two models in Chapter 4

3.3.1 The Cell Level Model

We will model traffic from a single source in the following manner Define the random variable $X(t)$ where $t \in N$ and $X(t)$ takes values in the state space $E = \{0, 1, 2, \quad , s\}$ by,

$$X(t) \quad = \quad \min\{s, \text{time since last cell arrival from the source at the multiplexer}\}$$

Then $X(t) = 0$ means an arrival from this source at time t And $X(t)$ is a Markov Chain Its transition diagram is shown in Figure 3-2

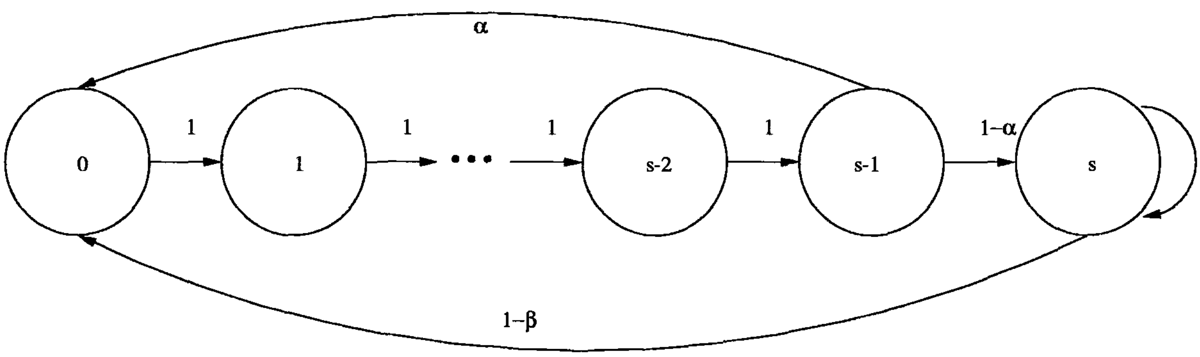

Figure 3-2 *Cell level transition diagram for single line*

The periodic nature of the arrivals from a single source means that transitions from state 0 through to state $s-1$ each occur with probability 1 At state $s-1$ either a cell arrives with probability α (i e the burst continues) and the state of the source makes the transition from state $s-1$ to state 0 or the transition from $s-1$ to s occurs i e the burst ends and a period of silence begins with probability $1-\alpha$ From state s transitions can occur to state s i e the silence continues with probability β, or to state 0 i e the silence ends with the arrival of another cell and the beginning of a new burst with probability $1-\beta$ The following is the $(s+1)$ by $(s+1)$ transition matrix for this Markov Chain model for a single line,

$$
P = \begin{pmatrix}
0 & 1 & \cdot & 0 & 0 & 0 \\
0 & 0 & & 0 & 0 & 0 \\
0 & 0 & \cdot & 0 & 0 & 0 \\
 & \cdot & \cdot \; \cdot \; \cdot & & & \\
 & & \cdot \; \cdot \; \cdot & & & \\
0 & 0 & & 0 & 1 & 0 \\
\alpha & 0 & \cdot & 0 & 0 & 1-\alpha \\
1-\beta & 0 & & 0 & 0 & \beta
\end{pmatrix}
$$

The states of this Markov Chain form an irreducible closed set and are recurrent non-null aperiodic (ergodic) Recall form Chapter 2 that they form an irreducible closed set because every state can be reached from every other state i e for each pair of states i and j there exists an integer n such that $P^n(i,j) > 0$, this can also be seen from the transition diagram Recall such a Markov Chain is termed irreducible Further, from 2 1 3, since the state space is finite all states are recurrent non-null (since none are transient and all states must be either transient or recurrent non-null by irreducibility) Finally from 2 1 3 all states are periodic or aperiodic and since a transition is possible from s to itself, i e $P(s,s) > 0$ the g c d of all $n \geq 1$ such that $P^n(s,s) > 0$ is 1 that is s cannot be periodic, hence all states are aperiodic All of this implies from 2 1 3 that the Markov Cham has a unique stationary distribution In fact it has stationary distribution,

$$
\pi = \frac{1}{s + \frac{1-\alpha}{1-\beta}} \begin{pmatrix} 1 & 1 & \cdot & 1 & 1 & \frac{1-\alpha}{1-\beta} \end{pmatrix} \tag{3 7}
$$

We will refer to this model from now on as the **cell-level model**

3.3.2 The Block Level Model

The cell level model described above is the model of greatest interest to us But a simplified model (studied in detail in [3]) which we will call the Block Level Model can be derived from it, and, in chapter 4 we compare the effective bandwidth decay rates for the two models The block level model is derived from the cell-level model by looking at the cell-level model on a different time scale, we look at the number of arrivals in each block of s ticks Thus we define a new random variable Y and define a block to be a unit of time of size s ticks during which a cell may or may not arrive We define Y as follows,

$$Y_j \in \{0,1\} \tag{3 8}$$

$$Y_j = 0 \quad \text{if no cell arrival in block number } j \tag{3 9}$$

$$Y_j = 1 \quad \text{if one cell arrival in block number } j \tag{3 10}$$

We could relate this process to the X process by writing X_{js} for the state of the source at the jth tick in the kth block Then,

$$Y_j = 0 \;\Rightarrow\; X_{js-1} = s \tag{3 11}$$

$$Y_j = 1 \;\Rightarrow\; X_{js-1} < s \tag{3 12}$$

But this process will not be a Markov Chain unless we make the incorrect assumption that the end of one burst and the beginning of the next cannot occupy the same block This is clearly possible as is illustrated in the figure 3-3

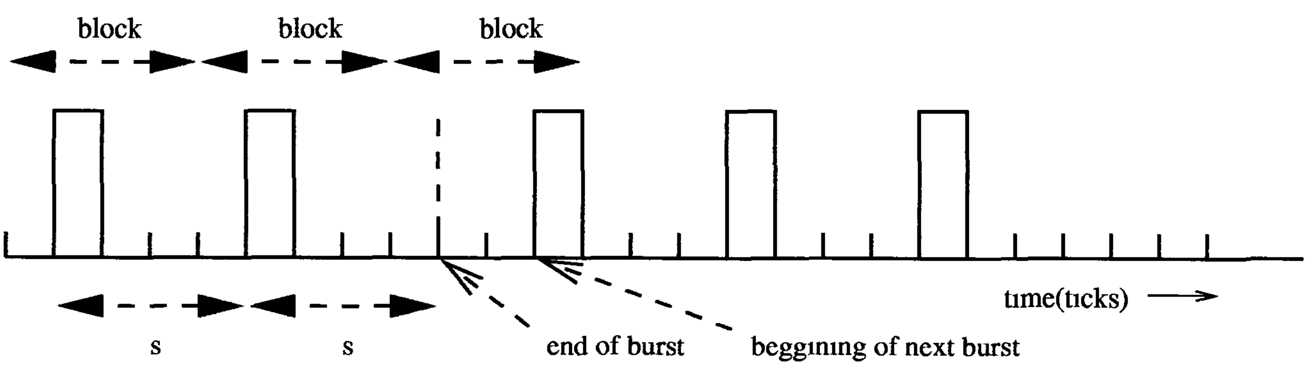

Figure 3-3 *Bursts overlapping in a block*

Instead we will define Y as in and assume that Y is a Markov Chain with the transition diagram shown in Figure 3-4 The transition matrix for this Markov Cham is,

$$P = \begin{pmatrix} 1-a & a \\ d & 1-d \end{pmatrix} \tag{3 13}$$

Where a and d are defined by,

$$a = \mathbf{P}[Y_{j+1} = 1 | Y_j = 0] \tag{3 14}$$

$$d = \mathbf{P}[Y_{j+1} = 0 | Y_j = 1] \tag{3 15}$$

We can see from these equations that a is the probability that a silent source

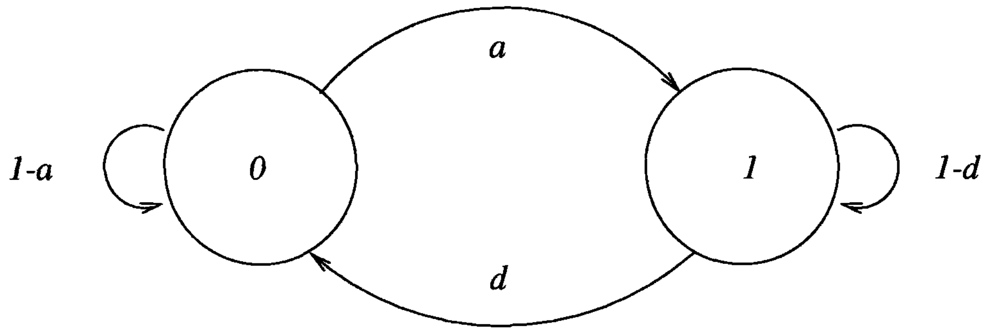

Figure 3-4 *Block Level Transition Diagram*

becomes active and d is the probability that an active source becomes silent Thus,

$$\begin{aligned} a &= \mathbf{P}[\text{silence} \leq s] & (3\ 16) \\ &= 1 - \beta^s & (3\ 17) \\ d &= \mathbf{P}[\text{burst ends}] & (3\ 18) \\ &= 1 - \alpha & (3\ 19) \end{aligned}$$

This Markov chain has unique stationary distribution,

$$\pi = \frac{1}{a+d} \begin{pmatrix} d & a \end{pmatrix} \tag{3 20}$$

As described earlier, the traffic presented to the multiplexer, will be the superposition of the traffic from all of the individual sources We will be interested in homogeneous superpositions, where all of the sources produce traffic which generates the same arrival process at the multiplexer for each individual source

Chapter 4

The effective bandwidth approximation

4.1 Effective bandwidth approximation

In this chapter we use the theory of large deviations to find an approximation for for the tail of the queue length distribution for the queue in an infinite buffer served at deterministic service rate where the arrivals are a homogeneous superposition of arrivals from sources modelled by the cell level model this can be used to approximate the the probability of buffer overflow from a finite buffer fed by the same arrivals process with the same service rate. The approximation is known as the effective bandwidth approximation [2] and is of the form,

$$\mathbf{P}[q \geq b] \approx e^{-\gamma b}$$

where γ is a constant. We also calculate the effective bandwidth approximation for the queue length distribution for the queue in an infinite buffer served at deterministic service rate where the arrivals are a homogeneous superposition of arrivals from sources modelled by the block level model [3]. We compare the decay rate constants for each model. In this we apply the work of Glynn and Whitt [12], Lewis and Russell [2] and Duffield et al [13] to our new cell level model and to the block level model already studied in detail in [3] to obtain the decay rate constants in each case. We then compare the two constants. In Section 4.1.2 we state the Large Deviations result of Glynn and Whitt [12] and explain their result using the work of Lewis and Russell [2]. We also state with proof an expression for the decay rate constant γ which is to be found in [2]. We give a proof from [2] of the Glynn and Whitt result [12]. In Section 4.1.3 we explain the term *effective bandwidth approximation* [2]. In Section 4.2 we give results from [13] on calculating the decay rate constant. In Section 4.3 we apply these results to calculate the decay rate constant for our new model the cell level model. In Section 4.4 we do the same for the block level model of [3]. In Section 4.5 we compare the two decay rate constants and prove that the decay rate constant for the cell level model is smaller than that for the block level model.

4.1.1 The equilibrium queue-length

Our problem consists of finding an upper bound or approximation for the tail of the queue-length distribution of a single server queue operating in discrete time with a FCFS service discipline and an infinite buffer, where there are non-independent arrivals to the queue. Let q_n be the queue-length at time $-n$. Then the queue-length at time 0 will be the sum of the queue-length at time -1 and the arrivals at time -1 minus the work done by the server at time -1. The queue is never negative, thus, if the server can do more work than the work presented to it at time -1, then the queue-length will be 0. Let U_n be the arrivals at time $-n$. Let Y_n be the work the server can do at time $-n$. Then we can write the following Lindley equation for the queue-length at time 0,

$$q_0 = \sup\{0, U_1 - Y_1 + q_1\}$$

We can iterate this equation as follows,

$$\begin{aligned} q_0 &= \sup\{0, U_1 - Y_1 + q_1\} \\ &= \sup\{0, U_1 - Y_1 + \sup\{U_2 - U_2 + q_2\}\} \\ &= \sup\{0, U_1 - Y_1 + U_2 - Y_2 + q_2\} \\ &= \sup\{0, U_1 + U_2 - (Y_1 + Y_2) + q_2\} \end{aligned}$$

Let $A_n = \sum_i^n U_i$ and $S_n = \sum_i^n Y_i$; i.e. A_n is the number of arrivals up to time $-n$ and S_n is the service that can be performed up to time $-n$. Let $W_n = A_n - S_n$, then W_n

is called the workload process We will define $W_0 = 0$ Then iterating equation 5 1 $t-1$ times we get,

$$q_0 = \sup\{W_0, W_1, \quad , W_t + q_t\}$$

We are interested in the equilibrium queue-length Any queue with a FCFS service discipline and stationary ergodic arrivals U_t and services Y_t has a unique stationary distribution if the load is less than 1 [14] The equilibrium queue-length is then given by,

$$q = \sup_{t \geq 0} W_t$$

4.1.2 Large deviations

We are interested in finding an approximation to, or upper bound on, the tail of the queue-length distribution of a queue with non-independent arrivals We can do this using the theory of Large Deviations Glynn and Whitt [12] showed that if $(\frac{W_t}{t}, t)$, where t is discrete, satisfies a Large Deviation principle with rate function I, i e ,

$$\mathbf{P}[\frac{W_t}{t} \geq w] \approx e^{-tI(w)}$$

then,

$$\mathbf{P}[q \geq b] \approx e^{-\gamma b}$$

and,

$$\gamma = \inf_w \frac{I(w)}{w}$$

The reason for this is indicated by the following [2],

$$\begin{aligned} \mathbf{P}[q > b] &= \mathbf{P}[\sup_{t \geq 0} W_t > b] \\ &= \mathbf{P}[\bigcup_{t \geq 0} \{W_t > b\}] \\ &\leq \sum_{t \geq 0} \mathbf{P}[W_t > b] \end{aligned}$$

but,

$$\begin{aligned}\mathbf{P}[W_t > b] &= \mathbf{P}[\frac{W_t}{t} > \frac{b}{t}] \\ &\approx e^{-tI(b/t)} \\ &= e^{-b\frac{I(b/t)}{b/t}}\end{aligned}$$

Thus,

$$\begin{aligned}\mathbf{P}[q > b] &\approx e^{-b\frac{I(b)}{b}} + e^{-b\frac{I(b/2)}{b/2}} + \qquad + e^{-b\frac{I(b/t)}{b/t}} + \\ &\approx e^{-b\inf_w \frac{I(w)}{w}} \\ &= e^{-\gamma b}\end{aligned}$$

with the second equality due to the fact that the term that dominates the right hand side of the first for large b, is the term for which the exponent is smallest We can derive a further expression for the decay constant γ [2],

$$\begin{aligned}\theta \leq \gamma &\Leftrightarrow \theta \leq \inf_w \frac{I(w)}{w} \\ &\Leftrightarrow \theta \leq \frac{I(w)}{w} \text{ for all } w \\ &\Leftrightarrow \theta w - I(w) \leq 0 \text{ for all } w \\ &\Leftrightarrow \sup\{w | \theta w - I(w) \leq 0\} \\ &\Leftrightarrow \lambda(\theta) \leq 0 \\ &\Leftrightarrow \gamma = \sup\{\theta | \lambda(\theta) \leq 0\}\end{aligned}$$

where $\lambda(\theta)$ is the scaled cumulant generating function (CGF) for the workload process

The following is a rigorous proof of the result for the asymptotic decay rate of the queue-length distribution tail [2] Recall,

$$q = \sup_{t \geq 0} W_t$$

Thus,

$$\{q \geq b\} = \bigcup_{t \geq 0} \{W_t \geq b\}$$

Thus for each $t \geq 0$

$$\{q \geq b\} \supset \{W_t \geq b\}$$

which implies, for all $t \geq 0$

$$\mathbf{P}[q \geq b] \geq \mathbf{P}[W_t \geq b]$$

Let $b = tw$ for fixed $w > 0$ Then,

$$\begin{aligned}\{W_t \geq b\} &= \{\frac{W_t}{t} \geq \frac{b}{t}\} \\ &= \{\frac{W_t}{t} \geq w\}\end{aligned}$$

hence,

$$\begin{aligned}\mathbf{P}[q \geq b] &\geq \mathbf{P}[W_t \geq b] \\ &\geq \mathbf{P}[\frac{W_t}{t} \geq w]\end{aligned}$$

which implies,

$$\frac{1}{b}\log \mathbf{P}[q \geq b] \geq \frac{1}{w}\frac{w}{b}\mathbf{P}[\frac{W_t}{t} \geq w]$$

and hence,

$$\begin{aligned}\liminf_{b\to\infty}\frac{1}{b}\log \mathbf{P}[q \geq b] &\geq \liminf_{b\to\infty}\frac{1}{w}\frac{w}{b}\log \mathbf{P}[\frac{W_t}{t} \geq w] \\ &= \frac{1}{w}\liminf_{t\to\infty}\frac{1}{t}\log \mathbf{P}[\frac{W_t}{t} \geq w] \\ &\geq -\frac{I(w)}{w}\end{aligned}$$

This is true for all $w > 0$ thus,

$$\begin{aligned}\liminf_{b\to\infty}\frac{1}{b}\log \mathbf{P}[q \geq b] &\geq \sup_{w>0}\{-\frac{I(w)}{w}\} \\ &= -\inf_{w>0}\frac{I(w)}{w}\end{aligned}$$

In order to complete the proof it is necessary to get an upper bound on the lim sup

Let the sequence of random variables $\{W_t\}_{t\geq 0}$ satisfy the conditions of the Gartner-Ellis theorem [15] That is, for all real θ let,

$$\begin{aligned}\lambda_t(\theta) &= \frac{1}{t}\log \mathbf{E}[e^{\theta W_t}] \\ \lambda(\theta) &= \lim_{t\to\infty}\lambda_t(\theta)\end{aligned}$$

and let the limit exist and be finite Further, let $\gamma = \sup\{\theta | \lambda(\theta) \leq 0\}$

Then consider the set $\{\theta | \lambda(\theta) \leq 0\}$ If this set is empty then $\gamma = -\infty$ and then,

$$\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[q \geq b] \;\leq\; -\gamma$$

If the set is not empty then let $\hat{\theta} < \gamma$ be an element of this set Then from Chernoff's bound we have,

$$\mathbf{P}[W_t \geq b] \;\leq\; e^{-\hat{\theta}b}\mathbf{E}[e^{\hat{\theta}W_t}]$$

But $\mathbf{E}[e^{\hat{\theta}W_t}]$ is finite, thus,

$$\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[W_t \geq b] \;\leq\; -\hat{\theta}$$

Now for each integer N,

$$\begin{aligned}
\mathbf{P}[\sup_{t\leq N} W_t \geq b] &\leq \sum_{t=1}^{t=N} \mathbf{P}[W_t \geq b] \\
&\leq N \sup_{t\leq N} \mathbf{P}[W_t \geq b]
\end{aligned}$$

But this implies,

$$\begin{aligned}
\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t\leq N} W_t \geq b] &\leq \sup_{t\leq N} \limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[W_t \geq b] \\
&\leq -\hat{\theta}
\end{aligned}$$

We also have,

$$\begin{aligned}
\mathbf{P}[\sup_{t\leq N} W_t \geq b] &\leq \sum_{t=1}^{t=N} \mathbf{P}[W_t \geq b] \\
&\leq e^{-\hat{\theta}b} \sum_{t>N} \mathbf{E}[e^{\hat{\theta}W_t}] \\
&= e^{-\hat{\theta}b} \sum_{t>N} e^{t\lambda_t(\hat{\theta})}
\end{aligned}$$

Now the $\lim_{t\to\infty} \lambda_t(\hat{\theta}) = \lambda(\hat{\theta}) < 0$, thus there exists $\epsilon > 0$ such that $\lambda < -\epsilon$, and there exists an integer $N(\hat{\theta})$ depending on $\hat{\theta}$ such that $\lambda_t(\hat{\theta}) < -\epsilon$ for all $t > N(\hat{\theta})$ This implies

$$\begin{aligned}
\mathbf{P}[\sup_{t\leq N} W_t \geq b] &\leq e^{-\hat{\theta}b} \sum_{t>N} e^{t\epsilon} \\
&< e^{-\hat{\theta}b} \frac{1}{1 - e^{-\epsilon}}
\end{aligned}$$

implying,

$$\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t\leq N} W_t \geq b] \quad \leq \quad -\hat{\theta}$$

Thus, as

$$\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t\geq 0} W_t \geq b]$$

is equal to,

$$\max\{\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t\leq N(\hat{\theta})} W_t \geq b], \limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t> N(\hat{\theta})} W_t \geq b]\}$$

we have,

$$\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t\geq 0} W_t \geq b] \quad \leq \quad -\hat{\theta}$$

and this is true for all $\hat{\theta} < \gamma$ and hence,

$$\limsup_{b\to\infty} \frac{1}{b} \log \mathbf{P}[\sup_{t\geq 0} W_t \geq b] \quad \leq \quad -\gamma$$

Thus together with the lower bound we have,

$$\lim_{b\to\infty} \frac{1}{b} \log \mathbf{P}[q \geq b] \quad = \quad -\gamma$$

as required

This tells us that the tail of the queue length distribution is asymptotically log-linear with slope $-\gamma$ [2]

4.1.3 Effective Bandwidths

The approximation, for large b,

$$\mathbf{P}[q \geq b] \approx e^{-\gamma b}$$

is called the effective bandwidth approximation [2], for the following reason Consider the situation which is of concern to us We have an arrival process served at deterministic service rate r, i e the workload process is,

$$W_t = A_t - rt$$

thus the scaled CGF is,

$$\begin{aligned}\lambda(\theta) &= \lim_{t\to\infty}\frac{1}{t}\log\mathbf{E}[e^{\theta W_t}]\\ &= \lim_{t\to\infty}\frac{1}{t}\log\mathbf{E}[e^{\theta A_t}] - r\theta\\ &= \hat{\lambda}(\theta) - r\theta\end{aligned}$$

where $\hat{\lambda}(\theta)$ is the scaled CGF of the arrivals process. Thus we can write the following,

$$\gamma(r) = \sup\{\theta|\hat{\lambda}(\theta) \leq r\theta\}$$

which gives us γ as a function of the service rate r. Now if our queue has finite waiting space, i.e. we have a finite buffer, then we can use the effective bandwidth approximation to give us an upper bound on the probability of the buffer overflowing. This is because the probability of finite buffer overflow is bounded above by the probability that the infinite buffer queue length exceeds the finite buffer size. This leads us to the reason for the term, effective bandwidth. If we have a prescribed probability of buffer overflow in an ATM network of, say y, and we have a buffer of fixed capacity b, then we would want to know what is the minimum service rate needed to guarantee that the probability of buffer overflow will not exceed our prescribed value y. Using our upper bound we get the following for this service rate $r(y)$

$$r(y) = \inf\{r|e^{-\gamma(r)b} \leq y\}$$

From which we get,

$$r(y) = \frac{\lambda(\theta_y)}{\theta_y}$$

with $\theta_y = \frac{\log(y)}{b}$. We call $r(y)$ the effective bandwidth of the arrivals. It is the minimum transmission bandwidth needed to guarantee that the probability of buffer overflow will not exceed the prescribed value y.

4.2 Calculating the decay rate constant

The decay rate constant γ can be calculated for our models of an ATM multiplexer by using the scaled CGF for the workload process [13]. Our model is an example of a finite state Markov Additive Process for which γ can be found using the following technique from [13]. The workload process for the homogeneous superposition of L independent sources served at rate r, is,

$$W_t^L = \sum_{l=1}^{L}(A_t^{(l)} - r/L)$$

Where $A_t^{(l)}$ is the number of arrivals from source l up to time $-t$. The first thing we note is that if we let $\mathbf{X}_t$ be the vector of states $(X^{(1)}, ..., X^{(L)})$ in the state space $\mathbf{E} = E^{\times L}$, then $\mathbf{X}_t$ is the state of the system of L sources or lines at time $-t$, and is a Markov Chain with transition matrix $\mathbf{P} = P^{\otimes L}$ where this means the outer product of the transitions matrices for the L lines. If we let the increment in the workload be $Z(\mathbf{x_t})$ when $\mathbf{X}_t = \mathbf{x}_t$ then,

$$\mathbf{E}[e^{\theta W_t^L}] = \sum_{\mathbf{x}_1 \in \mathbf{E}} \cdots \sum_{\mathbf{x}_t \in \mathbf{E}} e^{\theta \sum_{n=1}^{t} Z(\mathbf{x}_t)} \prod_{n=2}^{t} \mathbf{P}(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{n-1} = \mathbf{x}_{n-1}) \pi(\mathbf{x}_1)$$

where $\pi(\mathbf{x}_1)$ is the probability that $\mathbf{X}_1 = \mathbf{x}_1$. The product of the $t-2$ transition probabilities with $\pi(\mathbf{x}_1)$ is just the joint probability of the t state vectors. If we write,

$$\mathbf{P}(\theta)(\mathbf{x}_n, \mathbf{x}_{n-1}) = e^{\theta Z(\mathbf{x}_n)} \mathbf{P}(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{n-1} = \mathbf{x}_{n-1})$$

and write,

$$\pi(\theta)(\mathbf{x}_1) = e^{\theta Z(\mathbf{x}_n)} \pi(\mathbf{x}_1)$$

then we have,

$$\mathbf{E}[e^{\theta W_t^L}] = \pi(\theta)\mathbf{P}(\theta)^t 1^T$$

And hence,

$$\begin{aligned} \lambda(\theta) &= \lim_{t\to\infty} \frac{1}{t} \log[\pi(\theta)\mathbf{P}(\theta)^t 1^T] && (4.1) \\ &= \log[\mathrm{sp}(\mathbf{P}(\theta))] && (4.2) \end{aligned}$$

where $\mathrm{sp}(A)$ means spectral radius of the matrix A. Note we can write $\mathbf{P}(\theta)$ as $\mathbf{P}\mathbf{D}e^{-\theta}$ where $\mathbf{D} = D^{\otimes L}$ and D is the diagonal matrix with $e^{\theta n(i)}$ in the (i, i) position where $n(i)$ is the number of arrivals if a source is in state i. The above result follows from a result of Frobenius [7] which says, if a matrix A is positive; i.e. all its entries are non-negative and at least one entry is positive, and if the matrix raised to some power is such that all its entries are positive, then, $\lim_{n\to\infty}\frac{A^n}{(\mathrm{sp}(A))^n} = B$ where all entries of B are non-negative. Now if the Markov Chain $\mathbf{X}_t$ is irreducible, recurrent, non-null and aperiodic then $\mathbf{P}(\theta)$ satisfies the conditions of this theorem because $\mathbf{P}$ does and because $\mathbf{D}$ is a diagonal matrix with positive diagonal entries. Thus we can say,

$$\begin{aligned}
\lim_{t\to\infty}\frac{1}{t}\log[\pi(\theta)\mathbf{P}(\theta)^t 1^T] - \lim_{t\to\infty}\frac{1}{t}\log[(\mathrm{sp}(\mathbf{P}(\theta)))^t] &= \lim_{t\to\infty}\frac{1}{t}\log[\frac{\pi(\theta)\mathbf{P}(\theta)^t 1^T}{(\mathrm{sp}(\mathbf{P}(\theta)))^t}] \\
&= \lim_{t\to\infty}\frac{1}{t}\log[\pi(\theta)\mathbf{B}(\theta)1^T] \\
&= 0
\end{aligned}$$

Where $\mathbf{B}$ plays the same part here as B in the theorem of Frobenius. We have therefore,

$$\begin{aligned}
\lambda(\theta) &= \lim_{t\to\infty}\frac{1}{t}\log[\pi(\theta)\mathbf{P}(\theta)^t 1^T] \\
&= \lim_{t\to\infty}\frac{1}{t}\log[(\mathrm{sp}(\mathbf{P}(\theta)))^t] \\
&= \log[\mathrm{sp}(\mathbf{P}(\theta))]
\end{aligned}$$

as required.

Now if $\mathbf{X}_t$ is stationary and recurrent and irreducible, then $\mathrm{sp}(\mathbf{P}(\theta))$ is in fact the maximum of the moduli of the eigenvalues of $\mathbf{P}(\theta)$. For our models the individual sources are modelled by Markov Chains which satisfy this condition and, hence $\mathbf{X}_t$ satisfies this condition. The problem of finding γ reduces therefore to finding positive θ for which the log of the maximum of the moduli of the eigenvalues of $\mathbf{P}(\theta)$ is 0.

4.3 The cell level model

In order to find the decay rate constant γ for the queue q^L produced by the homogeneous superposition of L sources modelled by the cell level model, we need to find the maximum of the moduli of the eigenvalues of the transformed matrix, $\mathbf{P}(\theta)$ for the superposed process.

Define $\hat{v}$ as follows,

$$\begin{aligned}\hat{v}(\theta) &= \mathrm{sp}(\mathbf{P}(\theta)) \\ &= \mathrm{sp}(\mathbf{PD}e^{-\theta})\end{aligned}$$

where,

$$\begin{aligned}\mathbf{P} &= P^{\otimes L} \\ \mathbf{D} &= D^{\otimes L}\end{aligned}$$

Then define v as follows,

$$\begin{aligned}v(\theta) &= \mathrm{sp}(PDe^{-\theta/L}) \\ &= e^{-\theta/L}\mathrm{sp}(PD)\end{aligned}$$

Recall that γ is given by,

$$\gamma = \sup\{\theta|\lambda(\theta) \leq 0\}$$

But,

$$\begin{aligned}\lambda(\theta) &= \log \mathrm{sp}(\mathbf{P}(\theta)) \\ &= \log \hat{v}(\theta)\end{aligned}$$

Thus,

$$\gamma = \sup\{\theta|\hat{v}(\theta) \leq 1\}$$

But,

$$\hat{v}(\theta) = (v(\theta))^L$$

Thus,

$$\begin{aligned}\gamma &= \sup\{\theta|\hat{v}(\theta) \leq 1\} \\ &= \sup\{\theta|v(\theta) \leq 1\}\end{aligned}$$

Now,

$$v(\theta) = e^{-\theta/L}\mathrm{sp}(PD)$$

And sp(PD) is just the largest eigenvalue of PD In order to find $v(\theta)$ we need to find the largest eigenvalue of matrix PD Recall that the transition matrix A for the forward Markov Chain for a single line in the cell level model is,

$$A = \begin{pmatrix} 0 & 1 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ 0 & 0 & \cdot & 0 & 0 & 0 \\ & & \cdot & \cdot\ \cdot & & \\ \cdot & & \cdot & & & \\ 0 & 0 & \cdot & 0 & 1 & 0 \\ \alpha & 0 & & 0 & 0 & 1-\alpha \\ 1-\beta & 0 & \cdot & 0 & 0 & \beta \end{pmatrix}$$

We are working in reversed time, that is we are looking at the arrival process reversed in time Thus we want the reversed Markov chain transition matrix P with entries given by,

$$P_{ij} = \frac{\pi_j}{\pi_i} A_{ji}$$

Where π is the unique stationary distribution of the Markov Chain

$$\begin{aligned} \frac{\pi_j}{\pi_i} &= 1 \quad \text{for } i, j \in \{0, \ldots, s-1\} \text{ and for } i = j \\ \frac{\pi_j}{\pi_i} &= \frac{1-\alpha}{1-\beta} \quad \text{for } i = s \text{ and } j \in \{0, \ldots, s-1\} \\ \frac{\pi_j}{\pi_i} &= \frac{1-\beta}{1-\alpha} \quad \text{for } j = s \text{ and } i \in \{0, \ldots, s-1\} \end{aligned}$$

Thus we have,

$$P = \begin{pmatrix} 0 & 0 & & \cdot & 0 & \alpha & 1-\alpha \\ 1 & 0 & & & 0 & 0 & 0 \\ 0 & 1 & \cdot & \cdot & 0 & 0 & 0 \\ & \cdot & \cdot & & & \cdot & \\ \cdot & & & \cdot\ \cdot & & & \cdot \\ 0 & 0 & \cdot & & 0 & 0 & 0 \\ 0 & 0 & & & 1 & 0 & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 1-\beta & \beta \end{pmatrix}$$

and the $(s+1) \times (s+1)$ matrix D is,

$$D = \begin{pmatrix} e^{\theta} & 0 & & & 0 & 0 & 0 \\ 0 & 1 & 0 & & 0 & 0 & 0 \\ 0 & 0 & & \cdot\ \cdot & 0 & 0 & 0 \\ & \cdot & \cdot & & & & \\ \cdot & & & & \cdot & \cdot & \\ 0 & 0 & & & 1 & 0 & 0 \\ 0 & 0 & \cdot & \cdot\ \cdot & 0 & 1 & 0 \\ 0 & 0 & & & 0 & 0 & 1 \end{pmatrix}$$

Thus the transformed matrix is,

$$P(\theta) \ = \ PDe^{-\theta/L}$$

where,

$$PD = \begin{pmatrix} 0 & 0 & \cdot & 0 & \alpha & 1-\alpha \\ e^{\theta} & 0 & & 0 & 0 & 0 \\ 0 & 1 & & 0 & 0 & 0 \\ \cdot & & & & & \\ \cdot & \cdot & & & & \\ 0 & 0 & & 0 & 0 & 0 \\ 0 & 0 & & 1 & 0 & 0 \\ 0 & 0 & & 0 & 1-\beta & \beta \end{pmatrix}$$

The characteristic equation of matrix PD is,

$$\begin{aligned} \mathrm{Det}(x - PD) &= -x^{s+1} + \beta x^{s} + \alpha e^{\theta} x + e^{\theta}(1-\alpha-\beta) \\ &= g(x, \theta, s) \end{aligned}$$

Thus,

$$\begin{aligned} v(\theta) &= e^{-\theta/L} \sup\{x | g(x, \theta, s) = 0\} \\ &= \sup\{x^{\text{‘}} | g(x^{\text{‘}} e^{\theta/L}, \theta, s) = 0\} \end{aligned}$$

Where we have put $x^{\text{‘}} = xe^{-\theta/L}$ Thus we have the following for γ,

$$\gamma \ = \ \sup\{\theta | \sup\{x^{\text{‘}} | g(x^{\text{‘}} e^{\theta/L}, \theta, s) = 0\} \leq 1\}$$

Thus to find γ we should attempt to solve,

$$g(e^{\theta/L}, \theta, s) = 0$$

Then γ will be the supremum of the resulting solution set This gives us the following,

$$-e^{(s+1)\theta/L} + \beta e^{s\theta/L} + \alpha e^{\theta} e^{\theta/L} + e^{\theta}(1 - \alpha - \beta) = 0 \tag{4 3}$$

Note that the γ which we would obtain if we could solve this equation and find the supremum of the solution set would depend on L Now we are interested in finding the decay rate constant for the queue q^L produced by the superposition of L sources for large L i e the queue for a large multiplexer The load ρ of the multiplexed system is,

$$\rho = \frac{L}{s + (1-\alpha)/(1-\beta)}$$

This is dependent on L The requirement for stable queuing is that $\rho < 1$ Thus in order to study the queue for different values of L we will have to scale the parameters of the model to ensure that ρ remains constant To this end we define the constants (w r t L) σ and τ as follows,

$$\begin{aligned} \sigma &= \frac{s}{L} \\ \tau &= L(1-\beta) \end{aligned}$$

That is we scale s and β in a manner that makes ρ constant w r t L Putting the rescaled parameters into our equation for θ we get,

$$-e^{\sigma\theta} e^{\theta/L} + (1 - \frac{\tau}{L}) e^{\sigma\theta} + \alpha e^{\theta} e^{\theta/L} + -\alpha e^{\theta} + \frac{\tau}{L} = 0$$

Rearranging this so that all the L dependent terms appear on the same side we get,

$$e^{(\sigma-1)\theta} = \frac{\alpha(e^{\theta/L} - 1) + \frac{\tau}{L}}{(e^{\theta/L} - 1) + \frac{\tau}{L}} \tag{4 4}$$

Now consider the behaviour of $\gamma(L)$ w r t L We know $\gamma(L)$ solves the above equation Thus consider the right hand side of that equation,

$$\frac{\alpha(e^{\theta/L} - 1) + \frac{\tau}{L}}{(e^{\theta/L} - 1) + \frac{\tau}{L}} = \frac{\alpha(e^{\theta/L} - 1) + \alpha\frac{\tau}{L} + (1-\alpha)\frac{\tau}{L}}{(e^{\theta/L} - 1) + \frac{\tau}{L}}$$

$$
\begin{aligned}
&= \frac{\alpha((e^{\theta/L}-1)+\frac{\tau}{L})}{(e^{\theta/L}-1)+\frac{\tau}{L}} + \frac{(1-\alpha)\frac{\tau}{L}}{(e^{\theta/L}-1)+\frac{\tau}{L}} \\
&= \alpha + \frac{(1-\alpha)\frac{\tau}{L}}{(e^{\theta/L}-1)+\frac{\tau}{L}} \\
&> \alpha
\end{aligned}
$$

Thus for all $L > 0$ we have,

$$
e^{(1-\sigma)\gamma(L)} < \frac{1}{\alpha} \tag{4 5}
$$

Hence the sequence of gammas depending on L, $\gamma(L)$, is bounded above as follows,

$$
\gamma(L) < \frac{1}{(1-\sigma)} \log[1/\alpha]
$$

For typical values of α and σ the right hand side of this inequality will be much smaller than 1 Thus $\gamma(L)$ will be much smaller than 1 Now consider the following,

$$
\begin{aligned}
\alpha + \frac{(1-\alpha)\frac{\tau}{L}}{(e^{\gamma(L)/L}-1)+\frac{\tau}{L}} &= \alpha + \frac{(1-\alpha)\tau}{L(e^{\gamma(L)/L}-1)+\tau} \\
&< \alpha + \frac{(1-\alpha)\tau}{\gamma(L)+\tau} \qquad (4\ 6) \\
&= \frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau} \qquad (4\ 7)
\end{aligned}
$$

In fact if $\gamma(L)$ is small (and/or L is large), we have,

$$
\begin{aligned}
\alpha + \frac{(1-\alpha)\frac{\tau}{L}}{(e^{\gamma(L)/L}-1)+\frac{\tau}{L}} &= \alpha + \frac{(1-\alpha)\tau}{L(e^{\gamma(L)/L}-1)+\tau} \\
&\approx \alpha + \frac{(1-\alpha)\tau}{\gamma(L)+\tau} \\
&= \frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}
\end{aligned}
$$

Thus we obtain the following equation for an approximation for $\gamma(L)$

$$
e^{(\sigma-1)\gamma(L)} = \frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}
$$

But the solutions to this equation are independent of L and are much smaller than 1 for typical values of α and σ, hence we can write the following equation for γ an approximation for the decay rate constant for the queue-length q^L that is independent of L

$$
\frac{1}{1+(1-\sigma)\gamma} = \frac{\alpha\gamma+\tau}{\gamma+\tau}
$$

giving us the quadratic equation,

$$\alpha(1-\sigma)\gamma^2 + (\tau(1-\sigma) - (1-\alpha))\gamma = 0$$

Which gives us,

$$\gamma = \frac{(1-\alpha) - \tau(1-\sigma)}{\alpha(1-\sigma)} \tag{4.8}$$

This is positive as required because,

$$\begin{aligned} \rho &= \frac{1}{\sigma + \frac{1-\alpha}{\tau}} \\ &< 1 \end{aligned}$$

which implies,

$$\sigma + \frac{1-\alpha}{\tau} > 1$$

and this implies,

$$1 - \alpha - \tau(1-\sigma) > 0$$

As noted earlier we can use the effective bandwidth approximation to give us an upper bound on the probability of overflow of a finite buffer in an ATM multiplexer In the next section we find γ for the block level model

4.4 The block level model

In chapter 3 we described a 2 state Markov model Y from [3] as follows We let the state space $E = \{0, 1\}$ and define the transition matrix to be,

$$P = \begin{pmatrix} 1-a & a \\ d & 1-d \end{pmatrix}$$

Where,

$$\begin{aligned} a &= \mathbf{P}[Y_{t+1} = 1 | Y_t = 0] \\ d &= \mathbf{P}[Y_{t+1} = 0 | Y_t = 1] \end{aligned}$$

This has unique stationary distribution,

$$\pi = \frac{1}{a+d}\begin{pmatrix} d & a \end{pmatrix}$$

And

$$\pi_i P_{ij} = \pi_j P_{ji}$$

implying that the Markov Chain Y is reversible, i e the matrix for the reversed chain is also P As with the cell level model,

$$\mathbf{P} = P^{\otimes L}$$

and,

$$\mathbf{D} = D^{\otimes L}$$

and for the states $j \in E$ we have for the number of arrivals z_j when Y is in state j,

$$\begin{aligned} j &\in \{0,1\} \\ \mathbf{P}[z(Y_t) = z_j | Y_t = j] &= \begin{cases} 1 & \text{If } z_j = j \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Hence,

$$\begin{aligned} P(\theta) &= P_{ij}(\theta) \\ &= P_{ij} e^{\theta(z_j - s/L)} \end{aligned}$$

That is, the diagonal 2×2 matrix D is,

$$D = \begin{pmatrix} 1 & 0 \\ 0 & e^{\theta} \end{pmatrix}$$

and therefore the transformed matrix is,

$$P(\theta) = PDe^{-\theta s/L}$$

Where,

$$PD = \begin{pmatrix} 1-a & ae^{\theta} \\ d & (1-d)e^{\theta} \end{pmatrix}$$

The characteristic equation of PD is,

$$\begin{aligned}\text{Det}(xI - PD) &= x^2 - ((1-a) + (1-d)e^\theta)x + (1-a)(1-d)e^\theta - ade^\theta \\ &= x^2 - ((1-a) + (1-d)e^\theta)x + (1-a-d)e^\theta \\ &= h(x,\theta)\end{aligned}$$

As in the cell-level model in order to find γ we now solve, with $\sigma = \frac{s}{L}$

$$h(e^{\gamma\sigma}, \gamma) = 0$$

Then assuming that γ is small, we have the following,

$$(1+\gamma\sigma)^2 - ((1-a) + (1-d)(1+\gamma))(1+\gamma\sigma) + (1-a-d)(1+\gamma) = 0$$

Then,

$$\gamma^2(\sigma(\sigma - (1-d)) + \gamma((a+d)\sigma - a) = 0$$

hence,

$$\gamma = \frac{((a+d)\sigma - a)}{\sigma((1-d) - \sigma)} \tag{4 9}$$

4.5 Comparison of block and cell level models

The cell level model captures more of the features of the situation we wish to model than the block level model But we would like to know if the bound on the tail of the distribution of the block level model queue is more, or less conservative than the bound on the the tail of the distribution of the cell level model queue If it is more conservative then we could use it for dimensioning the buffer in an ATM multiplexer, in place of the cell level model upper bound If it is less conservative then we could not The answer to this question is the latter, as $\gamma_{\text{Block}} > \gamma_{\text{Cell}}$

$$\begin{aligned}\gamma_{\text{Cell}} &= \frac{((a+d)\sigma - a)}{\sigma((1-d) - \sigma)} \\ \gamma_{\text{Block}} &= \frac{(1-\alpha) - \tau(1-\sigma)}{\alpha(1-\sigma)}\end{aligned}$$

In order to compare the two models we must first relate the parameters from each. The defining equation for a means that a is the probability that a silent line becomes active, which for the cell level model means $\mathbf{P}[\text{silence} \leq s]$. But this is just $1 - \beta^s$, and β is close to 1 i.e. $1 - \beta$ is close to 0. Thus we have,

$$\begin{aligned} a &= 1 - \beta^s \\ &= 1 - (1 - (1 - \beta))^s \\ &\approx 1 - 1 + s(1 - \beta) \\ &= s(1 - \beta) \end{aligned}$$

Similarly d is the probability that an active line becomes inactive which for the cell level model is $1 - \alpha$. Thus,

$$d = 1 - \alpha$$

Hence,

$$\begin{aligned} \tau &= L(1 - \beta) \\ &= L\frac{a}{s} \\ &= \frac{a}{\sigma} \end{aligned}$$

Writing both γ_{Block} and γ_{Cell} in terms of a, d and σ we get,

$$\begin{aligned} \gamma_{\text{Block}} &= \frac{(a + d)\sigma - a}{\sigma(1 - \sigma) - d\sigma} \\ \gamma_{\text{Cell}} &= \frac{(a + d)\sigma - a}{\sigma(1 - \sigma) - d\sigma(1 - \sigma)} \end{aligned}$$

and $0 < \sigma < 1$. Thus, $d\sigma > d\sigma(1 - \sigma)$ and we have,

$$\gamma_{\text{Block}} > \gamma_{\text{Cell}}$$

Hence the block level upper bound on buffer overflow is less conservative than the cell level upper bound.

The effective bandwidth approximation deals with large b. It tells us nothing about small b. It also tells us nothing about the economies of scale that may be possible in large multiplexers And in fact the effective bandwidth approximation overestimates the probability of cell loss from a finite buffer with bursty arrivals On the other hand the effective bandwidth approximation shows the multiplexing gain in large buffers

Chapter 5

An Upper Bound Via Martingales

In this chapter we use the theory of Martingales to prove two new upper bounds of the form,

$$\mathbf{P}[q > b] \leq \phi e^{-\gamma b} \tag{5.1}$$

one for the full queue for the cell level model which includes the cell level queue the other for the burst level queue of the cell level model. The cell level queue is due to arrivals over a period shorter than the packetization periods in a multiplexer with a service rate of 1 cell per tick. This gives rise to short queues when the arrival rate temporarily exceeds the service rate. The burst level queue is due to arrivals over periods longer than s and is due to the average arrival rate over such a period exceeding the service rate. This gives rise to longer queues. The upper bound for the full queue does not exhibit the economies of scale seen in large multiplexers for any parameter values. The second upper bound does exhibit the economies of scale and is an improvement over the effective bandwidth approximation in terms of bounding the tail of the queue length distribution of the burst level queue. That is, $\phi < 1$ and $\phi < \Phi^L < 1$ for some parameter values and for Φ independent of L for this queue.

5.1 Martingales

Martingales were first studied by Levy but the development of the theory of Martingales is due to Doob [16]. The term first appeared in connection with gambling and the basic idea underlying the concept is that of a game being fair [17], in the sense that a players conditional expected future fortune is the players current fortune In this context the terms submartingales and supermartingales correspond to favourable and unfavourable games respectively [17] Martingale theory has developed a scope far beyond its gambling origins In the context of finding upper bounds queue length distributions Kingman [18] used the theory of Martingales to obtain exponential bounds for the queue length in the $GI|G|1$ queue Motivated by Kingman's result Buffet and Duffield [3] used Martingale methods to obtain an upper bound on the queue length distribution for the block level model of our Chapter 3, which can be viewed as an approximation for the burst level component of the queue in an ATM multiplexer We will use similar methods to obtain an upper bound on the tail of the queue length distribution for the cell level model We begin with the definition of a Martingale Then we define a Markov Time (stopping time) Then we state two theorems which are used or appear in this chapter, due to Doob, an Optional Stopping Theorem for non-negative martingales and the Maximal Inequality for Positive Submartingales [16]

Definition 1 *Let $\{M_n\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$ Let $\{\mathcal{F}_n\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$ with $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ for all n Then $\{M_n\}$ is called a submartingale with respect to $\{\mathcal{F}_n\}$ if,*

- *Each M_n is $\mathcal{F}_n$ -measurable*
- $\mathbf{E}[M_n^+] < \infty$ *for all n*
- $\mathbf{E}[M_{n+1}|\mathcal{F}_n] \geq M_n$ *for all n*

where $M_n^+ = \max\{M_n, 0\}$ If $\{-M_n\}$ is a submartingale, then $\{M_n\}$ is called a supermartingale If both $\{M_n\}$ and $\{-M_n\}$ are submartingales then $\{M_n\}$ is called a martingale with respect to $\{\mathcal{F}_n\}$

Definition 2 *Let $\{\mathcal{F}_n\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$ with $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ for all n A random variable T taking values in $\{0, 1, \quad, \infty\}$ is called a Markov time with respect to $\{\mathcal{F}_n\}$, if for every $n = 0, 1, 2, \quad$, the event $\{T = n\}$ is in $\mathcal{F}_n$ i e ,*

$$\{\omega \in \Omega | T(\omega) = n\} \quad \in \quad \mathcal{F}_n \quad \text{for all } n \tag{5 2}$$

This can be rewritten in the form of Definition 2 4,

Definition 3 *Let $\{\mathcal{F}_n\}$ be a sequence of sub-σ-algebras of $\mathcal{F}$ with $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ for all n A random variable T taking values in $\{0, 1, \quad, \infty\}$ is called a Markov time with respect to $\{\mathcal{F}_n\}$, if for every $n = 0, 1, 2, \quad,$*

$$I_{\{T \leq n\}} = \begin{cases} 1 & \text{if } T \leq n \\ 0 & \text{otherwise} \end{cases} \tag{5 3}$$

is $\mathcal{F}_n$ -measurable

Next we state an Optional Stopping Theorem [16] for positive martingales, (Theorem 4 2 page 267 in [16])

Theorem 1 *Let $\{M_n\}$ be a martingale with respect to $\{\mathcal{F}_n\}$ and let T be Markov time with respect to $\{\mathcal{F}_n\}$ If $M_n \geq 0$ for all n, then*

$$\mathbf{E}[M_T I_{\{T<\infty\}}] \leq \mathbf{E}[M_0] \tag{5 4}$$

The following is called Doob's Maximal Inequality for Positive Submartingales [16],

Theorem 2 *Let $\{M_n\}$ be a positive submartingale Then for any positive m,*

$$P[\sup\{M_0, \quad, M_n\} \geq m] \leq \frac{1}{m}\mathbf{E}[M_n] \tag{5 5}$$

5.2 Motivation

Kingman [18] used the theory of Martingales to obtain exponential bounds for the queue length in the $GI|G|1$ queue Basically this involved proving that if $\{Y_n\}_{n\geq 0}$ is a sequence of i i d random variables, then

$$\mathbf{P}[\sup_{n\geq 1}(Y_0 + \quad + Y_n) \geq x] \leq e^{-\theta x}$$

Where θ is a real number such that $\mathbf{E}[e^{\theta Y}] \leq 1$ The proof of this involves constructing the Martingale,

$$M_n = \frac{e^{\theta(Y_0 + \quad + Y_n)}}{(\mathbf{E}[e^{\theta Y}])^{n+1}}$$

This is a Martingale because, $\mathbf{E}[\frac{e^{\theta Y_n}}{\mathbf{E}[e^{\theta Y_n}]}] = 1$ for all n, and $(\frac{e^{\theta Y_n}}{\mathbf{E}[e^{\theta Y_n}]})_{n\geq 0}$ is a sequence of independent non-negative random variables, hence,

$$\mathbf{E}[M_{n+1}|Y_0, \quad , Y_n] = M_n$$

Then Doob's maximal inequality for positive submartingale tells us that,

$$\begin{aligned} \mathbf{P}[\sup\{M_0, \quad , M_n\} \geq m] &\leq \frac{1}{m}\mathbf{E}[M_n] \\ &= \frac{1}{m} \end{aligned}$$

This result is then used as follows to obtain the required exponential bound, for $\theta \geq 0$

$$\begin{aligned} \mathbf{P}[\sup_{n\geq 0} Y_0 + \quad + Y_n \geq x] &= \mathbf{P}[e^{\theta \sup_{n\geq 0}(Y_0 + \quad + Y_n)} \geq e^{\theta x}] \\ &= \mathbf{P}[\sup_{n\geq 0} e^{\theta(Y_0 + \quad + Y_n)} \geq e^{\theta x}] \\ &\leq \mathbf{P}[\sup_{n\geq 0} \frac{e^{\theta(Y_0 + \quad + Y_n)}}{(\mathbf{E}[e^{\theta Y}])^{n+1}} \geq e^{\theta x}] \end{aligned}$$

where $\mathbf{E}[e^{\theta Y}] \leq 1$,

Thus,

$$\begin{aligned} \mathbf{P}[\sup_{n\geq 0} Y_0 + \quad + Y_n \geq x] &\leq \mathbf{P}[\sup_{n\geq 0} M_n \geq e^{\theta x}] \\ &\leq e^{-\theta x} \end{aligned}$$

Now this result holds for i i d random variables But in the situation of interest to us we are dealing with dependent random variables Martingale methods can sometimes be used to extend results that hold for independent random variables to results for dependent random variables This is what Buffet and Duffield did in [3] motivated by Kingman We will use similar methods to obtain our upper bound

We will use a method for constructing Martingales for stationary Markov chains which says that if we have a stationary Markov chain Y_t with transition matrix P and we have an eigenfunction $f \quad E \to \mathbf{R}$, (where E is the state space for the Markov chain), with eigenvalue μ i e ,

$$\sum_{i\in E} f(i)P(i,j) = \mu f(j)$$

then,

$$M_t = f(Y_t)\mu^{-t}$$

is a Martingale w.r.t. the filtration $\mathcal{F}$ generated by $\mathcal{F}_t = \sigma(Y(0), \ldots\ldots\ldots\ldots, Y(t))$. Using the constructed Martingale and the Optional Sampling theorem (Theorem 1) we can obtain an upper bound on the tail of the queue length distribution of the form,

$$\mathbf{P}[q > b] \leq \phi e^{-b\gamma}$$

for the queue in an infinite buffer served at deterministic service rate on a FCFS basis, produced by the homogeneous superposition of L sources each modelled by an identical copy of the cell level model. Unlike the effective bandwidth approximation this upper bound holds for all values of $b \geq 0$ not just for large b.

5.3 The Martingale

We begin by briefly recalling the situation of interest to us. We have a queue with an infinite buffer with arrivals to the server processed on a first come first served basis. Let A_t be the time reversed arrival process at the queue for discrete time, t. We define $A_0 = 0$, and A_t to be the number of arrivals between time $-t$ and time 0. Arrivals are served at deterministic service rate. The workload process W_t is defined by $W_t = A_t - rt$. Then recall from Chapter 4 that under certain conditions the queue length has a unique stationary distribution [14]. The equilibrium queue length is given by,

$$q = \sup_{t \geq 0} W_t$$

As in Chapter 4 we define, for real θ and for $t > 0$,

$$\begin{aligned} \lambda_t(\theta) &= \frac{1}{t} \log \mathbf{E}[e^{\theta W_t}] \\ \lambda(\theta) &= \lim_{t \to \infty} \lambda_t(\theta) \end{aligned}$$

and assume the limit exists. We note that λ and λ_t are both strictly convex and essentially smooth. In the situation of concern to us the increments of the workload process are controlled by the states of an underlying Markov Process X The situation is an example of a Markov Additive Process (MAP) With such a process the workload is a function of the underlying Markov process in such a manner that the pair (X, W) is also a Markov Process More precisely on probability space $(\Omega, \mathcal{F}, \mathbf{P})$ let X_t be a stationary ergodic Markov Process on a state space E with σ-algebra $\mathcal{E}$ Let W_t be an additive component adjoined to it with $W_0 = 0$ and such that(X, W) is a Markov Process on the state space $E \times R^+$ (R^+ being the positive real numbers) Let Z be the increment process of W Then the joint distribution of Z_{t+1} and X_{t+1} conditioned on (X_u, W_u) for all $u \in [0, t]$ depends only on X_t, and this is expressed through the kernel,

$$P(x, G \times B) = \mathbf{P}[X_{t+1} \in G, Z_{t+1} \in B | X_t = x] \tag{5 6}$$

for $G \in \mathcal{E}$ and B a Borel set of R^+ [19]

Returning to the cell level model, recall, we have a homogeneous superposition of L independent sources or lines served at deterministic service rate r The workload process for the superposition is,

$$W_t^L = \sum_{l=1}^{L} (A_t^{(l)} - r/L)$$

where, as before, $A_t^{(l)}$ is the number of arrivals from source l up to time $-t$ Again we let $\mathbf{X}_t$ be the vector of states $(X^{(1)}, \quad , X^{(L)})$ in the state space $\mathbf{E} = E^{\times L}$, where state $X^{(l)}$ is a Markov chain for a single source Then $\mathbf{X}_t$ is the state of the system of L sources or lines at time $-t$, and is also a Markov Chain If the transition matrix for the individual source Markov Chain is P, then the transition matrix for $\mathbf{X}_t$ is the outer product of L copies of P which we denote as before by $\mathbf{P} = P^{\otimes L}$ Recall from Chapter 4 that if we define the transformed transition matrix, $\mathbf{P}(\theta)$ by,

$$\mathbf{P}(\theta)(\mathbf{x}_n, \mathbf{x}_{n-1}) = e^{\theta Z(\mathbf{x}_n)} \mathbf{P}(\mathbf{X}_n = \mathbf{x}_n | \mathbf{X}_{n-1} = \mathbf{x}_{n-1})$$

Then,

$$\begin{aligned} \lambda(\theta) &= \lim_{t \to \infty} \frac{1}{t} \log[\pi(\theta) \mathbf{P}(\theta)^t 1^T] \\ &= \log[\mathrm{sp}(\mathbf{P}(\theta))] \end{aligned}$$

where $\mathrm{sp}(A)$ means the spectral radius of the matrix A Then the Perron-Frobenius Theorem [7] tells us that for all θ in the effective domain of λ, $e^{\lambda(\theta)}$ is the unique, real, positive and simple maximal eigenvalue with corresponding strictly positive eigenvector, which we will denote by, $\hat{v}(\mathbf{x},\theta)$ Recall that γ is the unique positive solution of $\lambda(\theta)=0$ Then,

1 $\mathbf{P}(\gamma)$ has a maximal eigenvalue $e^{\lambda(\gamma)}=1$ with corresponding right eigenvector $\hat{v}(\ ,\gamma)$

2 Normalising $\hat{v}(\ ,\gamma)$ so that $\mathbf{E}[\hat{v}(\mathbf{X}_0,\gamma)]=1$ then $\mathbf{M}_t=e^{\gamma W_t^L}\hat{v}(\mathbf{X}_t,\gamma)$ is a positive martingale with respect to the canonical filtration $\mathcal{F}$ generated by $(\mathbf{X},W^L)$, and we also have $\mathbf{E}[\mathbf{M}_0]=1$

The proof of 2 is the following Firstly,

$$W_t^L = W_{t-1}^L + Z(\mathbf{X}_t)$$

Thus,

$$e^{\gamma W_t^L} = e^{\gamma W_{t-1}^L} e^{\gamma Z(\mathbf{X}_t)}$$

Thus we have,

$$\begin{aligned}
\mathbf{E}[\mathbf{M}_{t+1}(\gamma)|\mathcal{F}_t] &= \mathbf{E}[e^{\gamma W_{t+1}^L}\hat{v}(\mathbf{X}_{t+1},\gamma)|\mathcal{F}_t] \\
&= e^{\gamma W_t^L}\mathbf{E}[e^{\gamma Z(\mathbf{X}_{t+1})}\hat{v}(\mathbf{X}_{t+1},\gamma)|\mathcal{F}_t] \\
&= e^{\gamma W_t^L}\sum_{\mathbf{x}_{t+1}\in\mathbf{E}} e^{\gamma Z(\mathbf{x}_{t+1})}\hat{v}(\mathbf{X}_{t+1},\gamma)\mathbf{P}(\mathbf{X}_{t+1}=\mathbf{x}_{t+1}|\mathbf{X}_t=\mathbf{x}_t) \\
&= e^{\gamma W_t^L}\sum_{\mathbf{x}_{t+1}\in\mathbf{E}} \hat{v}(\mathbf{X}_{t+1},\gamma)\mathbf{P}(\gamma)(\mathbf{x}_{t+1},\mathbf{x}_t) \\
&= e^{\gamma W_t^L}\hat{v}(\mathbf{X}_t,\gamma) \\
&= \mathbf{M}_t
\end{aligned}$$

For the last part,

$$\begin{aligned}
\mathbf{E}[\mathbf{M}_0] &= \mathbf{E}[\hat{v}(\mathbf{X}_0)] \\
&= 1
\end{aligned}$$

concluding the proof This Martingale can be used to prove an upper bound for the queue length distribution of the form,

$$\mathbf{P}[q > b] \leq \phi e^{-\gamma b}$$

where, the prefactor ϕ is defined by the following equation,

$$\phi^{-1} = \inf_{\mathbf{y} \in \mathbf{E}, c > 0} \mathbf{E}[e^{\gamma(W_1^L - c)} v(\mathbf{X}_1) | W_1^L - c > 0 | \mathbf{X}_0 = \mathbf{y}]$$

The proof of this result is the following Define the stopping time τ by,

$$\tau = \inf\{t > 0 | \{W_t^L > b\}\}$$

Then,

$$\mathbf{P}[\sup_t W_t^L > b] = \mathbf{P}[\tau < \infty]$$

Then applying the Optional Stopping Theorem (Theorem 1) we get the following,

$$\begin{aligned} 1 &= \mathbf{E}[\mathbf{M}_0] \\ &\geq \mathbf{E}[\mathbf{M}_\tau, \tau < \infty] \\ &= \sum_{n \geq 0} \mathbf{E}[\mathbf{M}_n, \tau = n] \end{aligned} \tag{5 7}$$

But we can write the event $\{\tau = n\}$ as follows,

$$\{\tau = n\} = \bigcup_{c \geq 0, \mathbf{x} \in \mathbf{E}} \{G_n(c) \cap \{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}\} \tag{5 8}$$

Where $G_n(c) = \{\max_{1 \leq m \leq n-1} W_m^L \leq b, W_{n-1}^L = b - c\}$ Now this is a disjoint union for the following reasons, Firstly c is an integer and $\mathbf{E}$ is countable Let the integer $c_1, c_2 \geq 0$ and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{E}$ be such that $(c_1, \mathbf{x}_1) \neq (c_2, \mathbf{x}_2)$ then

$$\begin{aligned} \cap_{i=1}^{i=2} \{G_n(c_i) \cap \{Z_n > c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\}\} &\subseteq G_n(c_i) \cap \{Z_n > c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\} \\ &\subseteq G_n(c_i) \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\} \\ &\subseteq G_n(c_i) \\ &\subseteq \{W_{n-1}^L = b - c_i\} \end{aligned}$$

Similarly,

$$\begin{aligned}
\cap_{i=1}^{i=2}\{G_n(c_i) \cap \{Z_n > c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\}\} &\subseteq G_n(c_i) \cap \{Z_n > c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\} \\
&\subseteq G_n(c_i) \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\} \\
&\subseteq \{\mathbf{X}_{n-1} = \mathbf{x}_i\}
\end{aligned}$$

Thus,

$$\cap_{i=1}^{i=2}\{G_n(c_i) \cap \{Z_n > c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\}\} \subseteq \{W_{n-1}^L = b - c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\}$$

This is true for $i = 1, 2$ hence,

$$\begin{aligned}
\cap_{i=1}^{i=2}\{G_n(c_i) \cap \{Z_n > c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\}\} &\subseteq \cap_{i=1}^{i=2}\{W_{n-1}^L = b - c_i\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}_i\} \\
&= \phi
\end{aligned}$$

Thus we have the result. The union is disjoint. Thus substituting Equation 5.8 into Equation 5.7 we get,

$$1 \geq e^{\gamma b} \sum_{n \geq 0} \sum_{c \geq 0} \sum_{\mathbf{x} \in \mathbf{E}} \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma); G_n(c) \cap \{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}] \quad (5.9)$$

We note here that for those $\mathbf{x}$ for which $\{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}$ is the empty set the resultant terms in the sum 5.9 will be zero. Hence Let $\hat{\mathbf{E}}(c) \subset \mathbf{E}$ be the set of all states $\mathbf{x}$ for which $\{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}$ is **not** the empty set. Then we can rewrite 5.9 as,

$$1 \geq e^{\gamma b} \sum_{n \geq 0} \sum_{c \geq 0} \sum_{\mathbf{x} \in \hat{\mathbf{E}}(c)} \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma); G_n(c) \cap \{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}] \quad (5.10)$$

But conditioned on $\mathbf{X}_{n-1}$, $G_n(c)$ is independent of Z_n and $\mathbf{X}_n$ hence, we can rewrite $\mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma); G_n(c) \cap \{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}]$ as follows,

$$\begin{aligned}
&= \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma) I_{\{G_n(c) \cap \{Z_n > c\} \cap \{\mathbf{X}_{n-1} = \mathbf{x}\}\}}] \\
&= \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma) I_{\{G_n(c)\}} I_{\{Z_n > c\}} I_{\{\mathbf{X}_{n-1} = \mathbf{x}\}\}}] \\
&= \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma) I_{\{G_n(c)\}} I_{\{Z_n > c\}} | \mathbf{X}_{n-1} = \mathbf{x}] \mathbf{P}[\mathbf{X}_{n-1} = \mathbf{x}] \\
&= \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma) I_{\{Z_n > c\}} | \mathbf{X}_{n-1} = \mathbf{x}] \mathbf{E}[I_{\{G_n(c)\}} | \mathbf{X}_{n-1} = \mathbf{x}] \mathbf{P}[\mathbf{X}_{n-1} = \mathbf{x}] \\
&= \mathbf{E}[e^{\gamma(Z_n - c)} \upsilon(\mathbf{X}_n; \gamma) I_{\{Z_n > c\}} | \mathbf{X}_{n-1} = \mathbf{x}] \mathbf{P}[G_n(c) | \mathbf{X}_{n-1} = \mathbf{x}] \mathbf{P}[\mathbf{X}_{n-1} = \mathbf{x}]
\end{aligned}$$

Now we can define,

$$\begin{aligned}\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)|Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}] &= \frac{\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)I_{\{Z_n>c\}}|\mathbf{X}_{n-1}=\mathbf{x}]}{\mathbf{E}[I_{\{Z_n>c\}}|\mathbf{X}_{n-1}=\mathbf{x}]} \\ &= \frac{\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)I_{\{Z_n>c\}}|\mathbf{X}_{n-1}=\mathbf{x}]}{\mathbf{P}[Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}]}\end{aligned}$$

And this is well defined on $\hat{\mathbf{E}}(c)$ Then we can rewrite the right hand side of the last equation as,

$$\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)|Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}]\mathbf{P}[Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}]\mathbf{P}[G_n(c)|\mathbf{X}_{n-1}=\mathbf{x}]\mathbf{P}[\mathbf{X}_{n-1}=\mathbf{x}]$$

But this is just,

$$\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)|Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}]\mathbf{P}[\{Z_n>c\}\cap G_n(c)\cap\{\mathbf{X}_{n-1}=\mathbf{x}\}] \quad (5\ 11)$$

Hence, putting this back into the sum in Equation 5 9 and summing over $\mathbf{x}\in\hat{\mathbf{E}}(c)$ (all other terms being zero) we get,

$$1 \geq e^{\gamma b}\sum_{n,c,\mathbf{x}}\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)|Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}]\mathbf{P}[G_n(c)\cap\{Z_n>c\}\cap\{\mathbf{X}_{n-1}=\mathbf{x}\}]$$

We will write

$$\phi^{-1} = \inf_{c>0,\mathbf{x}\in\hat{\mathbf{E}}(c)}\mathbf{E}[e^{\gamma(Z_n-c)}\upsilon(\mathbf{X}_n,\gamma)|Z_n>c|\mathbf{X}_{n-1}=\mathbf{x}] \quad (5\ 12)$$

This is independent of n for a stationary Markov process $\mathbf{X}$ Thus

$$\phi^{-1} = \inf_{c>0,\mathbf{x}\in\hat{\mathbf{E}}(c)}\mathbf{E}[e^{\gamma(Z_1-c)}\upsilon(\mathbf{X}_1,\gamma)|Z_1>c|\mathbf{X}_0=\mathbf{x}] \quad (5\ 13)$$

Thus,

$$\begin{aligned}1 &\geq e^{\gamma b}\sum_{n\geq0}\sum_{c\geq0}\sum_{\mathbf{x}\in\hat{\mathbf{E}}(c)}\phi^{-1}\mathbf{P}[G_n(c)\cap\{Z_n>c\}\cap\{\mathbf{X}_{n-1}=\mathbf{x}\}]] && (5\ 14)\\ &= e^{\gamma b}\phi^{-1}\sum_{n\geq0}\sum_{c\geq0}\sum_{\mathbf{x}\in\hat{\mathbf{E}}(c)}\mathbf{P}[G_n(c)\cap\{Z_n>c\}\cap\{\mathbf{X}_{n-1}=\mathbf{x}\}]] && (5\ 15)\\ &= e^{\gamma b}\phi^{-1}\sum_{n\geq0}\mathbf{P}[\tau=n]\\ &= e^{\gamma b}\phi^{-1}\mathbf{P}[\tau<\infty]\\ &= e^{\gamma b}\phi^{-1}\mathbf{P}[\sup_{t>0}W_t^L>b]\\ &= e^{\gamma b}\phi^{-1}\mathbf{P}[q>b]\end{aligned}$$

Thus, we have,

$$\mathbf{P}[q > b] \;\leq\; \phi e^{-\gamma b} \tag{5 16}$$

completing the proof

5.4 Calculating The Prefactor

The prefactor is defined by,

$$\begin{aligned}
\phi^{-1} &= \inf_{\mathbf{x}\in\hat{\mathbf{E}}(c),c>0} \mathbf{E}[e^{\gamma(Z_1-c)}\hat{v}(\mathbf{X}_1)|Z_1 - c > 0|\mathbf{X}_0 = \mathbf{x}] \\
&= \inf_{\mathbf{x}\in\hat{\mathbf{E}}(c),c>0} \frac{\mathbf{E}[e^{\gamma(Z_1-c)}\hat{v}(\mathbf{X}_1)I_{\{Z_1-c>0\}}|\mathbf{X}_0 = \mathbf{x}]}{\mathbf{P}[Z_1 - c > 0|\mathbf{X}_0 = \mathbf{x}]} \qquad (5\ 17)
\end{aligned}$$

Now we can write,

$$Z_1 = \sharp\{i \in \{1, \quad , L\}|X_1^i = 0\} - 1 \tag{5 18}$$

$$= \sharp\{i \in \{1, \quad , L\}|X_0^i = 1\} - 1 \tag{5 19}$$

with the first equation due to the fact that arrivals from source i only occur when the Markov chain X_t^i is in state 0 and Z_1 is the total number of arrivals at time $t = 1$ minus the service completed in one tick, that service being 1 The second equation is due to the fact that the Markov chain X_t^i makes transitions to state 0 only from state 1 and from no other state and does so with probability 1 This can be seen from the transition matrix in Equation 4 3 Chapter 4

Let $\mathbf{E}(c) = \{\mathbf{x} \in \mathbf{E}|\sharp\{i|x^i = 1\} - 1 > c\}$ Then, by Equation 5 18 we have,

$$\{\mathbf{X}_0 \in \mathbf{E}(c)\} \subset \{Z_1 - c > 0\}$$

and,

$$\{Z_1 - c > 0\} \subset \{\mathbf{X}_0 \in \mathbf{E}(c)\}$$

Hence,

$$\{Z_1 - c > 0\} = \{\mathbf{X}_0 \in \mathbf{E}(c)\}$$

Thus, $\mathbf{E}(c) = \hat{\mathbf{E}}(c)$ for the cell level model, and,

$$I_{\{Z_1 - c > 0\}} = I_{\{\mathbf{X}_0 \in \mathbf{E}(c)\}}$$

Note also that for $c \geq L - 1$ both sides of this last equation will be 0 We can therefore rewrite Equation 5 17 as,

$$\begin{aligned}
\phi^{-1} &= \inf_{\mathbf{x} \in \mathbf{E}(c), c>0} \frac{\mathbf{E}[e^{\gamma(Z_1 - c)} \hat{v}(\mathbf{X}_1) I_{\{\mathbf{X}_0 \in \mathbf{E}(c)\}} | \mathbf{X}_0 = \mathbf{x}]}{\mathbf{P}[\mathbf{X}_0 \in \mathbf{E}(c) | \mathbf{X}_0 = \mathbf{x}]} && (5\ 20) \\
&= \inf_{\mathbf{x} \in \mathbf{E}(c), c>0} \mathbf{E}[e^{\gamma(Z_1 - c)} \hat{v}(\mathbf{X}_1) | \mathbf{X}_0 = \mathbf{x}] && (5\ 21) \\
&= \inf_{\mathbf{x} \in \mathbf{E}(c), c>0} e^{-\gamma c} \mathbf{E}[e^{\gamma W_1^L} \hat{v}(\mathbf{X}_1) | \mathbf{X}_0 = \mathbf{x}] && (5\ 22) \\
&= \inf_{\mathbf{x} \in \mathbf{E}(c), c>0} e^{-\gamma c} \mathbf{E}[\mathbf{M}_1 | \mathbf{X}_0 = \mathbf{x}] && (5\ 23) \\
&= \inf_{\mathbf{x} \in \mathbf{E}(c), c>0} e^{-c\gamma} \hat{v}(\mathbf{x}) && (5\ 24)
\end{aligned}$$

where $c \in \{1, \quad , L-2\}$ and where Equation 5 22 is due to

$$\begin{aligned}
W_1^L &= W_0^L + Z(\mathbf{X}_1) \\
&= 0 + Z(\mathbf{X}_1) \\
&= Z_1
\end{aligned}$$

And Equation 5 24 is due to an elementary property of Martingales

Let

$$m(\mathbf{x}) = \sharp\{i \in \{1, \quad , L\} | x^i = 1 \quad \text{for} \quad \mathbf{x} \in \mathbf{E}(c)\} \qquad (5\ 25)$$

Then Equation 5 24 is,

$$\phi^{-1} = \inf_{x^i \in E-1, m(\mathbf{x}) > c+1, c>0} e^{-c\gamma} \prod_{i=1}^{L-m(\mathbf{x})} v(x^i) v(1)^{m(\mathbf{x})} \qquad (5\ 26)$$

Note $c \in \{1, \quad , L-2\}$ and $m(\mathbf{x}) \in \{c+2, \quad , L\}$

Now, recall from Chapter 4 that the transformed kernel of the reversed Markov chain for a single line in a homogeneous superposition of L lines for the cell level model, with $\theta = \gamma$ is,

$$\mathbf{P}(\gamma) = e^{-\gamma/L} \begin{pmatrix} 0 & 0 & & \cdot & 0 & \alpha & 1-\alpha \\ e^{\gamma} & 0 & & & 0 & 0 & 0 \\ 0 & 1 & & & 0 & 0 & 0 \\ \cdot & \cdot & & & & \cdot & \\ & & \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ 0 & 0 & & & 1 & 0 & 0 \\ 0 & 0 & & \cdot & 0 & 1-\beta & \beta \end{pmatrix}$$

The equation for v is, subject to normalisation,

$$v = \mathbf{P}(\gamma)v$$

Written in full this then is,

$$\begin{pmatrix} v(0) \\ \\ \\ \cdot \\ \\ \cdot \\ \\ v(s) \end{pmatrix} = e^{-\gamma/L} \begin{pmatrix} 0 & 0 & & & 0 & \alpha & 1-\alpha \\ e^{\gamma} & 0 & & \cdot & 0 & 0 & 0 \\ 0 & 1 & \cdot & \cdot & \cdot\ 0 & 0 & 0 \\ \cdot & & & & & & \\ & & & & & \cdot & \\ 0 & 0 & & & 0 & 0 & 0 \\ 0 & 0 & \cdot & & 1 & 0 & 0 \\ 0 & 0 & & & 0 & 1-\beta & \beta \end{pmatrix} \begin{pmatrix} v(0) \\ \cdot \\ \\ \cdot \\ \\ \\ \\ v(s) \end{pmatrix}$$

From this we obtain the following set of equations for the components of v,

$$e^{-\gamma/L}(\alpha v(s-1) + (1-\alpha)v(s)) = v(0) \tag{5 27}$$

$$e^{-\gamma/L}e^{\gamma}v(0) = v(1) \tag{5 28}$$

$$e^{-\gamma/L}v(1) = v(2) \tag{5 29}$$

$$\cdot$$

$$\vdots$$

$$e^{-\gamma/L}v(s-2) = v(s-1) \tag{5 30}$$

$$e^{-\gamma/L}((1-\beta)v(s-1)+\beta v(s)) = v(s) \tag{5 31}$$

Rearranging Equation 5 31 we have,

$$\frac{v(s-1)}{v(s)} = \frac{e^{\gamma/L}-\beta}{1-\beta} \tag{5 32}$$

$$> 1 \tag{5 33}$$

as $\gamma > 1$ Subtracting equation 5 31 from equation 5 27 we get,

$$e^{-\gamma/L}(\alpha+\beta-1)(v(s-1)-v(s)) = v(0)-v(s)$$

which, from the last inequality 5 33 and the fact that $e^{-\gamma/L} > 0$, implies,

$$v(0) > v(s) \Leftrightarrow \alpha+\beta-1 > 0 \tag{5 34}$$

In the case of bursty traffic we will have $\alpha+\beta-1>0$ Now from equation 5 28 to 5 30 we can see,

$$v(1) > \quad > v(s-1) \tag{5 35}$$

Thus we have,

$$v(1) > \quad \cdots > v(s-1) > v(0) > v(s) \tag{5 36}$$

Now,

$$v(0)+v(1)+ \quad + v(s)(\frac{1-\alpha}{1-\beta}) = s+\frac{1-\alpha}{1-\beta} \tag{5 37}$$

This and the previous inequality imply,

$$v(s) < 1 \tag{5 38}$$

Thus from equation 5 36 we can rewrite equation 5 26 as,

$$\begin{aligned}
\phi^{-1} &= \inf\{e^{-\gamma c}v(s)^{L-m(\mathbf{x})}v(1)^{m(\mathbf{x})} | m(\mathbf{x}) \in \{c+2, \quad , L\}, c \in \{1, \quad , L-2\}\} \\
&= \inf\{e^{-\gamma c}v(s)^{L}(\frac{v(1)}{v(s)})^{m(\mathbf{x})} | m(\mathbf{x}) \in \{c+2, \quad , L\}, c \in \{1, \quad , L-2\}\} \\
&= \inf_{c\in\{1, \quad ,L-2\}} e^{-\gamma c}v(s)^{L}(\frac{v(1)}{v(s)})^{c+2}
\end{aligned} \tag{5 39}$$

Where this last equation is due to 5 36 We can rewrite this equation as,

$$\phi^{-1} = \inf_{c\in\{1,\ldots,L-2\}} v(s)^L\left(\frac{v(1)}{v(s)}\right)^2\left(\frac{e^{-\gamma}v(1)}{v(s)}\right)^c \tag{5 40}$$

Clearly the value of c giving us the infimum only depends on whether,

$$\frac{e^{-\gamma}v(1)}{v(s)} > 1 \tag{5 41}$$

We can prove that this is the case We do this as follows Firstly, by 5 36

$$\frac{e^{-\gamma}v(1)}{v(s)} = e^{\gamma(\sigma-1)}e^{-2\gamma/L}\frac{v(s-1)}{v(s)} \tag{5 42}$$

And by Equation 5 31

$$\frac{v(s-1)}{v(s)} = \frac{e^{\gamma/L}-\beta}{1-\beta} \tag{5 43}$$

$$= \frac{e^{\gamma/L}-1+(1-\beta)}{1-\beta} \tag{5 44}$$

$$= \frac{L(e^{\gamma/L}-1)+\tau}{\tau} \tag{5 45}$$

Hence,

$$\frac{e^{-\gamma}v(1)}{v(s)} = e^{\gamma(\sigma-1)}e^{-2\gamma/L}\frac{L(e^{\gamma/L}-1)+\tau}{\tau} \tag{5 46}$$

Now consider the inverse of the right hand side of this equation,

$$e^{\gamma(1-\sigma)}e^{2\gamma/L}\frac{\tau}{L(e^{\gamma/L}-1)+\tau} \tag{5 47}$$

If we subtract 1 from this and multiply the result by $(1-\alpha)e^{\gamma(\sigma-1)}e^{-2\gamma/L}$ we get,

$$\frac{(1-\alpha)\tau}{L(e^{\gamma/L}-1)+\tau} - (1-\alpha)e^{\gamma(\sigma-1)}e^{-2\gamma/L} \tag{5 48}$$

But this is just,

$$e^{\gamma(\sigma-1)} - \alpha - (1-\alpha)e^{\gamma(\sigma-1)}e^{-2\gamma/L} \tag{5 49}$$

Now consider,

$$e^{\theta(\sigma-1)} - \alpha - (1-\alpha)e^{\theta(\sigma-1)}e^{-2\theta/L} \tag{5 50}$$

At $\theta = 0$ this is 0 and its derivative w r t θ is,

$$(\sigma - 1)e^{\theta(\sigma-1)} - (1-\alpha)(\sigma - 1 - \frac{2}{L})e^{\theta(\sigma-1)}e^{-2\theta/L} \tag{5 51}$$

And this is,

$$e^{\theta(\sigma-1)}((1-\alpha)(1-\sigma+\frac{2}{L})e^{-2\theta/L} - (1-\sigma)) \leq e^{\theta(\sigma-1)}((1-\alpha)(1-\sigma+\tfrac{2}{L}) - (1-\sigma))$$

Now if $\alpha \geq 1/2$ then, $1-\alpha \leq 1/2$ and,

$$\begin{aligned} e^{\theta(\sigma-1)}((1-\alpha)(1-\sigma+\frac{2}{L}) - (1-\sigma)) &\leq e^{\theta(\sigma-1)}((1/2)(1-\sigma+\frac{2}{L}) - (1-\sigma)) \\ &= e^{\theta(\sigma-1)}(\frac{1}{L} - (1/2)(1-\sigma)) \\ &< 0 \end{aligned}$$

if $\sigma < 1 - \frac{2}{L}$ And consequently

$$\frac{v(1)}{v(s)e^{\gamma}} > 1 \tag{5 52}$$

Alternatively,

$$\frac{v(s-1)}{v(s)e^{\gamma}} = \frac{L(e^{\gamma/L}-1)+\tau}{e^{\gamma}\tau} > 1 \tag{5 53}$$

if,

$$\tau < \frac{L(e^{\gamma/L}-1)}{e^{\gamma}-1} \tag{5 54}$$

Recall from Equation 4 4 that,

$$e^{\gamma(\sigma-1)} = \frac{\alpha L(e^{\gamma/L}-1)+\tau}{L(e^{\gamma/L}-1)+\tau} \tag{5 55}$$

Thus,

$$\frac{L(e^{\gamma/L}-1)(e^{\gamma(\sigma-1)}-\alpha)}{1-e^{\gamma(\sigma-1)}} = \tau \tag{5 56}$$

Now consider,

$$\begin{aligned} \frac{L(e^{\gamma/L}-1)}{e^{\gamma}-1} - \tau &= \frac{L(e^{\gamma/L}-1)}{e^{\gamma}-1} - \frac{L(e^{\gamma/L}-1)(e^{\gamma(\sigma-1)}-\alpha)}{1-e^{\gamma(\sigma-1)}} && (5\ 57) \\ &= L(e^{\gamma/L}-1)(\frac{1}{e^{\gamma}-1} - \frac{e^{\gamma(\sigma-1)}-\alpha}{1-e^{\gamma(\sigma-1)}}) && (5\ 58) \\ &= L(e^{\gamma/L}-1)(\frac{1-e^{\gamma(\sigma-1)}-(e^{\gamma}-1)(e^{\gamma(\sigma-1)}-\alpha)}{(e^{\gamma}-1)(1-e^{\gamma(\sigma-1)})}) && (5\ 59) \\ &= L(e^{\gamma/L}-1)(\frac{1-\alpha+\alpha e^{\gamma}-e^{\gamma\sigma}}{(e^{\gamma}-1)(1-e^{\gamma(\sigma-1)})}) && (5\ 60) \end{aligned}$$

This will be positive if $1-\alpha+\alpha e^{\gamma}-e^{\gamma\sigma}>0$ since $\sigma<1$ and $\gamma>0$ and $1>\alpha$ and $L(e^{\gamma/L}-1)>0$ So consider,

$$1-\alpha+\alpha e^{\theta}-e^{\theta\sigma} \tag{5 61}$$

This is 0 at $\theta=0$ and it has derivative with respect to θ,

$$\alpha e^{\theta}-\sigma e^{\theta\sigma}>0 \tag{5 62}$$

on $[0,\infty)$ if $\alpha>\sigma$ Thus,

$$1-\alpha+\alpha e^{\theta}-e^{\theta\sigma}>0 \tag{5 63}$$

on $(0,\infty)$ if $\alpha>\sigma$ and in particular at $\theta=\gamma$ Hence,

$$\frac{L(e^{\gamma/L}-1)}{e^{\gamma}-1} > \tau \tag{5 64}$$

if $\alpha>\sigma$ Hence,

$$\frac{e^{-\gamma}v(s-1)}{v(s)}>1 \tag{5 65}$$

And consequently by 5 42

$$\frac{e^{-\gamma}v(1)}{v(s)}>1 \tag{5 66}$$

if $\alpha>\sigma$

Finally, returning to 5 43

$$\begin{aligned}\frac{L(e^{\theta/L}-1)+\tau}{e^{\theta}\tau} &> \frac{\theta+\tau}{e^{\theta}\tau} && (5\ 67)\\ &= 0 && (5\ 68)\end{aligned}$$

for $\theta=0$, and,

$$\frac{L(e^{\theta/L}-1)+\tau}{e^{\theta}\tau}>\frac{\theta+\tau}{e^{\theta}\tau} \tag{5 69}$$

For all $\theta>0$, and,

$$\frac{\theta+\tau}{e^{\theta}\tau}=\frac{1+\tau}{e\tau}>1 \tag{5 70}$$

For $\theta = 1$ and τ such that,

$$\tau < \frac{1}{e-1} \tag{5 71}$$

Now the derivative of

$$\frac{\theta+\tau}{e^{\theta}\tau} \tag{5 72}$$

with respect to θ is,

$$\frac{e^{\theta}\tau(1-\theta)}{(e^{\theta}\tau)^2} \;\geq\; 0 \tag{5 73}$$

for $0 \leq \theta \leq 1$, and the derivative of

$$\frac{1+\tau}{e\tau} \tag{5 74}$$

with respect to τ is,

$$\frac{-e}{(e\tau)^2} < 0 \tag{5 75}$$

for all τ Thus

$$\begin{aligned} \frac{e^{-\gamma}v(1)}{v(s)} &> \frac{e^{-\gamma}v(s-1)}{v(s)} && (5\ 76)\\ &> 1 && (5\ 77)\end{aligned}$$

for $\tau < 1/(e-1)$ Thus we have,

$$\frac{e^{-\gamma}v(1)}{v(s)} \;>\; 1 \tag{5 78}$$

For any one of $\tau < 1/(e-1)$ or $\sigma < \alpha$ or $\sigma < 1-(2/L)$ if $\alpha \geq 1/2$ all of which are reasonable assumptions, since $\tau < 0\ 5$ in for example the simulations in [9] and α is close to 1 and $\sigma < 1$ and L is large

We need an expression for $v(s)$ in order to find ϕ^{-1} Now from Equation 5 27 to Equation 5 30 we have,

$$\begin{aligned} \sum_{i=0}^{s-1} v(i) &= (e^{(s-1)\gamma/L}e^{-\gamma} + \sum_{i=1}^{s-1} e^{(s-1-i)\gamma/L})v(s-1)\\ &= (e^{(s-1)\gamma/L}e^{-\gamma} + \frac{e^{(s-1)\gamma/L}-1}{e^{\gamma/L}-1})v(s-1)\\ &= (e^{(s-1)\gamma/L}e^{-\gamma} + \frac{L(e^{(s-1)\gamma/L}-1)}{L(e^{\gamma/L}-1)})v(s-1)\\ &= (s-1+e^{\sigma\gamma}e^{-\gamma}e^{-\gamma/L})v(s-1)\\ &= (s-1+e^{\sigma\gamma}e^{-\gamma}e^{-\gamma/L})(1+\frac{\gamma}{\tau})v(s)\end{aligned}$$

From the normalisation of v we now have,

$$
\begin{aligned}
v(s) &= \frac{s+\frac{1-\alpha}{1-\beta}}{(s-1+e^{\sigma\gamma}e^{-\gamma}e^{-\gamma/L})(1+\frac{\gamma}{\tau})+\frac{1-\alpha}{1-\beta}} \\
&= \frac{s/L+\frac{1-\alpha}{L(1-\beta)}}{(\frac{s-1}{L}+\frac{e^{\sigma\gamma}e^{-\gamma}}{e^{\gamma/L}L})(1+\frac{\gamma}{\tau})+\frac{1-\alpha}{1-\beta}} \\
&= \frac{\sigma+\frac{1-\alpha}{\tau}}{(1+\frac{\gamma}{\tau})\sigma+\frac{1-\alpha}{\tau}}
\end{aligned}
$$

Returning to Equation 5 40 we now have,

$$
\begin{aligned}
\phi^{-1} &= v(s)^L(\frac{v(1)}{v(s)})^2(\frac{e^{-\gamma}v(1)}{v(s)})^1 && (5\ 79) \\
&= e^{-\gamma}v(s)^{L-3}v(1)^3 && (5\ 80)
\end{aligned}
$$

Thus we have the following for the prefactor,

$$
\phi^{-1} = e^{-\gamma}(e^{\gamma\sigma}e^{-2\gamma/L}(\frac{L(e^{\gamma/L}-1)}{\tau}+1))^3(\frac{\sigma+\frac{1-\alpha}{\tau}}{(\sigma+\frac{e^{\gamma(\sigma-1-1/L)}-1}{L})(1+\frac{L(e^{\gamma/L}-1)}{\tau})+\frac{1-\alpha}{\tau}})^L
$$

That is,

$$
\phi = e^{\gamma}(e^{-\gamma\sigma}e^{2\gamma/L}\frac{1}{(\frac{L(e^{\gamma/L}-1)}{\tau}+1)})^3(\frac{(\sigma+\frac{e^{\gamma(\sigma-1-1/L)}-1}{L})(1+\frac{L(e^{\gamma/L}-1)}{\tau})+\frac{1-\alpha}{\tau}}{\sigma+\frac{1-\alpha}{\tau}})^L
$$

Now consider,

$$
\begin{aligned}
\lim_{L\to\infty}\frac{1}{L}\log\phi &= \log(\frac{\sigma(1+\frac{\gamma}{\tau})+\frac{1-\alpha}{\tau}}{\sigma+\frac{1-\alpha}{\tau}}) && (5\ 81) \\
&> 0 && (5\ 82)
\end{aligned}
$$

This tells us that for large L the bound,

$$
\mathbf{P}[q^L > b] \le \phi e^{-\gamma b} \qquad (5\ 83)
$$

does not exhibit the economies of scale seen for example in the upper bound obtained using Martingales for the block level model by Buffet and Duffield [3] These economies of scale are seen in the simulations of Corcoran [9] for the rescaled cell level model

5.5 Calculating The Prefactor for the burst level queue

Consider a new Markov chain derived from the Markov chain for the cell level model defined by,

$$\hat{\mathbf{X}}_t = \mathbf{X}_{ts} \tag{5 84}$$

and adjoined to this an additive component defined by,

$$\hat{W}_t = W_{ts} \tag{5 85}$$

with increments defined by,

$$\hat{Z}(\hat{\mathbf{X}}_t) = \hat{W}_t - \hat{W}_{t-1} \tag{5 86}$$

Then,

$$\hat{\mathbf{M}}_t = e^{\gamma \hat{W}_t} \hat{v}(\hat{\mathbf{X}}_t) \tag{5 87}$$

is a Martingale with respect to the canonical filtration $\mathcal{F}$ generated by $(\hat{\mathbf{X}}, \hat{W})$

Proof

Firstly,

$$\hat{W}_t = \hat{W}_{t-1} + \hat{Z}(\hat{\mathbf{X}}_t)$$

Thus,

$$e^{\gamma \hat{W}_t} = e^{\gamma \hat{W}_{t-1}} e^{\gamma \hat{Z}(\hat{\mathbf{X}}_t)}$$

Thus we have,

$$\begin{aligned} \mathbf{E}[\hat{\mathbf{M}}_{t+1}(\gamma)|\mathcal{F}_t] &= \mathbf{E}[e^{\gamma \hat{W}_{t+1}} \hat{v}(\hat{\mathbf{X}}_{t+1}, \gamma)|\mathcal{F}_t] \\ &= e^{\gamma \hat{W}_t} \mathbf{E}[e^{\gamma \hat{Z}(\hat{\mathbf{X}}_{t+1})} \hat{v}(\hat{\mathbf{X}}_{t+1}, \gamma)|\mathcal{F}_t] \end{aligned}$$

Now written in terms of the increments of the workload process for the cell level model,

$$\hat{Z}(\hat{\mathbf{X}}_{t+1}) = Z(\mathbf{X}_{ts+1}) + \ldots + Z(\mathbf{X}_{ts+s-1}) + Z(\mathbf{X}_{ts+s}) \tag{5 88}$$

Thus,

$$e^{\gamma \hat{W}_t}\mathbf{E}[e^{\gamma \hat{Z}(\hat{\mathbf{X}}_{t+1})}\hat{v}(\hat{\mathbf{X}}_{t+1},\gamma)|\mathcal{F}_t] = e^{\gamma \hat{W}_t}\mathbf{E}[e^{\gamma \sum_{i=1}^{s} Z(X_{ts+i})}\hat{v}(\mathbf{X}_{ts+s},\gamma)|\mathcal{F}_t]$$

This is,

$$e^{\gamma \hat{W}_t}\sum_{\mathbf{x}_1\in\mathbf{E}} \cdot \sum_{\mathbf{x}_s\in\mathbf{E}} e^{\gamma \sum_{i=1}^{s} Z(\mathbf{x}_i)}\hat{v}(\mathbf{x}_s,\gamma)\prod_{n=1}^{s}\mathbf{P}(\mathbf{X}_{ts+n}=\mathbf{x}_n|\mathbf{X}_{ts+n-1}=\mathbf{x}_{n-1})$$

which is,

$$\begin{aligned} e^{\gamma \hat{W}_t}\sum_{\mathbf{x}_1\in\mathbf{E}}\cdots\sum_{\mathbf{x}_s\in\mathbf{E}}\hat{v}(\mathbf{x}_s,\gamma)\prod_{n=1}^{s}\mathbf{P}(\gamma)(\mathbf{X}_{ts+n}=\mathbf{x}_n|\mathbf{X}_{ts+n-1}=\mathbf{x}_{n-1}) &= e^{\gamma \hat{W}_t}\hat{v}(\mathbf{x}_0,\gamma) \\ &= \hat{\mathbf{M}}_t \end{aligned}$$

where we had

$$\mathbf{P}[\mathbf{X}_{ts+s}=\mathbf{x}_s, \quad ,\mathbf{X}_{ts+1}=\mathbf{x}_1|\mathbf{X}_{ts}=\mathbf{x}_0]=\prod_{n=1}^{s}\mathbf{P}(\mathbf{X}_n=\mathbf{x}_n|\mathbf{X}_{n-1}=\mathbf{x}_{n-1}) \quad (5\ 89)$$

and,

$$\mathbf{P}(\theta)(\mathbf{x}_n,\mathbf{x}_{n-1})=e^{\theta Z(\mathbf{x}_n)}\mathbf{P}(\mathbf{X}_n=\mathbf{x}_n|\mathbf{X}_{n-1}=\mathbf{x}_{n-1})$$

concluding the proof Note γ and $\hat{v}$ are the same decay rate and eigenvector as those for the full cell level model

Now if we define, the $\mathcal{F}$-stopping time,

$$\tau = \inf\{t|\hat{W}_t > b\}$$

We see that the proof for an upper bound that we used for the cell level model queue gives us an upper bound on q_s^L With the new prefactor defined by,

$$\phi^{-1} = \inf_{c>0,\mathbf{x}\in\hat{\mathbf{E}}(c)}\mathbf{E}[e^{\gamma(\hat{Z}_n-c)}v(\hat{\mathbf{X}}_n,\gamma)|\hat{Z}_n > c|\hat{\mathbf{X}}_{n-1}=\mathbf{x}] \quad (5\ 90)$$

Now $\hat{\mathbf{X}}$ is stationary since $\mathbf{X}$ is stationary It has transition matrix P^s Thus we can rewrite Equation 5 90 as,

$$\phi^{-1} = \inf_{c>0,\mathbf{x}\in\hat{\mathbf{E}}(c)}\mathbf{E}[e^{\gamma(\hat{Z}_1-c)}v(\hat{\mathbf{X}}_1,\gamma)|\hat{Z}_1 > c|\hat{\mathbf{X}}_0=\mathbf{x}] \quad (5\ 91)$$

But,

$$\begin{aligned}\hat{Z}_1 &= Z_1 + \quad + Z_s \\ &= W_s\end{aligned}$$

Thus, we have,

$$\phi^{-1} = \inf_{c>0, \mathbf{x}\in\hat{\mathbf{E}}(c)} \mathbf{E}[e^{\gamma(W_s-c)}\upsilon(\mathbf{X}_s,\gamma)|W_s > c|\mathbf{X}_0 = \mathbf{x}] \tag{5 92}$$

which is,

$$\phi^{-1} = \inf_{c>0, \mathbf{x}\in\hat{\mathbf{E}}(c)} \frac{\mathbf{E}[e^{\gamma(W_s-c)}\upsilon(\mathbf{X}_s,\gamma)I_{\{W_s>c\}}|\mathbf{X}_0 = \mathbf{x}]}{\mathbf{P}[W_s > c|\mathbf{X}_0]} \tag{5 93}$$

Now let $\mathbf{A}(c) = \{\mathbf{x} \in \mathbf{E}|\sharp\{i|x^i \in \{0, \quad , s-1\}\} - s > c\}$ Then,

$$\{\mathbf{X}_1 \in \mathbf{A}(c)\} \subset \{W_s > c\} \tag{5 94}$$

and,

$$\{W_s > c\} \subset \{\mathbf{X}_1 \in \mathbf{A}(c)\} \tag{5 95}$$

thus,

$$\{W_s > c\} = \{\mathbf{X}_1 \in \mathbf{A}(c)\} \tag{5 96}$$

hence,

$$I_{\{W_s>c\}} = I_{\{\mathbf{X}_1\in\mathbf{A}(c)\}} \tag{5 97}$$

Now the numerator of Equation 5 93,

$$\mathbf{E}[e^{\gamma(W_s-c)}\upsilon(\mathbf{X}_s,\gamma)I_{\{W_s>c\}}|\mathbf{X}_0 = \mathbf{x}] \tag{5 98}$$

is,

$$\sum_{\mathbf{y}\in\mathbf{E}} \mathbf{E}[e^{\gamma(W_s-c)}\upsilon(\mathbf{X}_s,\gamma)I_{\{W_s>c\}}|\mathbf{X}_0 = \mathbf{x}, \mathbf{X}_1 = \mathbf{y}]\mathbf{P}[\mathbf{X}_1 = \mathbf{y}|\mathbf{X}_0 = \mathbf{x}] \tag{5 99}$$

by the Law of Total Probability [7] But this is,

$$\sum_{\mathbf{y}\in\mathbf{E}} \mathbf{E}[e^{\gamma(W_s-c)}\upsilon(\mathbf{X}_s;\gamma)I_{\{X_1\in\mathbf{A}(c)\}}|\mathbf{X}_0 = \mathbf{x}, \mathbf{X}_1 = \mathbf{y}]\mathbf{P}[\mathbf{X}_1 = \mathbf{y}|\mathbf{X}_0 = \mathbf{x}] \tag{5 100}$$

by Equation 5 97 which is just,

$$e^{-\gamma c} \sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} \mathbf{E}[e^{\gamma W_s} v(\mathbf{X}_s, \gamma) | \mathbf{X}_0 = \mathbf{x}, \mathbf{X}_1 = \mathbf{y}] \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}] \tag{5 101}$$

which is,

$$e^{-\gamma c} \sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} \mathbf{E}[\mathbf{M}_s | \mathbf{X}_0 = \mathbf{x}, \mathbf{X}_1 = \mathbf{y}] \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}] \tag{5 102}$$

where $\mathbf{M}$ is just the Martingale that we used to get the upper bound for the cell level model queue But then this is just,

$$e^{-\gamma c} \sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} v(\mathbf{y}, \gamma) \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}] \tag{5 103}$$

Next the denominator of 5 93 is,

$$\begin{aligned} \mathbf{P}[W_s > c | \mathbf{X}_0 = \mathbf{x}] &= \mathbf{P}[\mathbf{X}_1 \in \mathbf{A}(c) | \mathbf{X}_0 = \mathbf{x}] & (5\ 104) \\ &= \sum_{\mathbf{y} \in \mathbf{A}(c)} \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}] & (5\ 105) \end{aligned}$$

by Equation 5 96 Thus we can now write,

$$\phi^{-1} = \inf_{c>0, \mathbf{x} \in \hat{\mathbf{E}}(c)} \frac{e^{-\gamma c} \sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} v(\mathbf{y}, \gamma) \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]}{\sum_{\mathbf{y} \in \mathbf{A}(c)} \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]} \tag{5 106}$$

Now we can say the following,

$$\begin{aligned} \frac{e^{-\gamma c} \sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} v(\mathbf{y}, \gamma) \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]}{\sum_{\mathbf{y} \in \mathbf{A}(c)} \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]} &\geq e^{-\gamma c} \inf_{\mathbf{y} \in \mathbf{A}(c)} v(\mathbf{y}, \gamma) \frac{\sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]}{\sum_{\mathbf{y} \in \mathbf{A}(c)} \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]} \\ &= e^{-\gamma c} \inf_{\mathbf{y} \in \mathbf{A}(c)} v(\mathbf{y}, \gamma) & (5\ 107) \end{aligned}$$

Thus,

$$\inf_{c>0, \mathbf{x} \in \hat{\mathbf{E}}(c)} \frac{e^{-\gamma c} \sum_{\mathbf{y} \in \mathbf{A}(\mathbf{c})} v(\mathbf{y}, \gamma) \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]}{\sum_{\mathbf{y} \in \mathbf{A}(c)} \mathbf{P}[\mathbf{X}_1 = \mathbf{y} | \mathbf{X}_0 = \mathbf{x}]} \geq \inf_{c>0, \mathbf{y} \in \mathbf{A}(\mathbf{c})} e^{-\gamma c} v(\mathbf{y}, \gamma)$$

But this is just,

$$\inf_{c>0, \mathbf{y} \in \mathbf{A}(c)} e^{-\gamma c} \prod_{i=1}^{L} v(y^i)$$

Now let $m(\mathbf{y}) = \sharp\{i | y^i \in \{0, \quad , s-1\}\}$ then $\mathbf{y} \in \mathbf{A}(c)$ implies $L \geq m(\mathbf{y}) > s + c$ Then the above becomes,

$$\inf_{c>0, \mathbf{y} \in \mathbf{A}(c), m(\mathbf{y}) > s+c} e^{-\gamma c} \prod_{i=1}^{L-m(\mathbf{y})} v(y^i) v(s-1)^m(\mathbf{y})$$

Then from Equation 5 36 this is,

$$\inf_{c>0, m(\mathbf{y})>s+c} e^{-\gamma c} v(s)^{L-m(\mathbf{y})} v(s-1)^{m}(\mathbf{y})$$

which is, with $\mathbf{y}$ such that each $y^i = s$ or $s-1$ for all i and with $m(\mathbf{y}) = m$ (the choice of $\mathbf{y}$ being superfluous once m is chosen),

$$\inf_{c>0, m>s+c} e^{-\gamma c} v(s)^{L} (\frac{v(s-1)}{v(s)})^{m}$$

which by Equation 5 36 is,

$$\inf_{c>0} e^{-\gamma c} v(s)^{L} (\frac{v(s-1)}{v(s)})^{s+c+1}$$

But we can write this as,

$$\inf_{c>0} (\frac{v(s-1)}{v(s)e^{\gamma}})^{c} v(s)^{L} (\frac{v(s-1)}{v(s)})^{s+1}$$

But by Equation 5 53 to Equation 5 65, for $\alpha > \sigma$ or Equation 5 76 for $\tau < \frac{1}{e-1}$ this is,

$$(\frac{v(s-1)}{v(s)e^{\gamma}}) v(s)^{L} (\frac{v(s-1)}{v(s)})^{s+1}$$

And this can be rewritten as,

$$\begin{aligned}
(\frac{v(s-1)}{v(s)e^{\gamma}})\frac{v(s-1)}{v(s)} v(s)^{L} (\frac{v(s-1)}{v(s)})^{s} &= (\frac{v(s-1)}{v(s)e^{\gamma}})\frac{v(s-1)}{v(s)} v(s)^{L} (\frac{v(s-1)}{v(s)})^{\sigma L} \\
&> ((\frac{v(s-1)}{v(s)})^{\sigma} v(s))^{L}
\end{aligned}$$

Where the last line is by Equations 5 36 and 5 65

5.5.1 Economies of scale

If,

$$(\frac{v(s-1)}{v(s)})^{\sigma} v(s) > 1 \tag{5 108}$$

Then we will have, for all $b > 0$

$$\begin{aligned}
\mathbf{P}[q_s^L > b] &\leq \phi e^{-\gamma b} \\
&\leq (\frac{v(s)e^{\gamma}}{v(s-1)})\frac{v(s)}{v(s-1)}((\frac{v(s)}{v(s-1)})^{\sigma}\frac{1}{v(s)})^{L} e^{-\gamma b} \\
&< e^{-\gamma b}
\end{aligned}$$

Thus our upper bound will be an improvement on the effective bandwidth approximation holding as it does for all $b > 0$ and being less conservative Further if we can show,

$$
\begin{aligned}
(\frac{v(s-1)}{v(s)})^{\sigma} v(s) &> k && (5\ 109)\\
&> 1 && (5\ 110)
\end{aligned}
$$

where k is independent of L, then we will have an upper bound of the form,

$$
\mathbf{P}[q_s^L > b] < \Phi^L e^{-\gamma b} \qquad (5\ 111)
$$

where $\Phi = \frac{1}{k} < 1$ and is independent of L This bound exhibits economies of scale which are seen in simulations for the burst level queue of the cell level model [9]

We will show this to be the case for some values of the model parameters First we will prove the following, useful inequality,

$$
\frac{1}{1 + L(e^{\gamma/L} - 1)(1-\sigma)} > e^{(\sigma-1)\gamma} \qquad (5\ 112)
$$

Proof

Consider,

$$
1 + L(e^{x/L} - 1)(1-\sigma) - e^{(1-\sigma)x} \qquad (5\ 113)
$$

At $x = 0$ this is 0 and its derivative is,

$$
\begin{aligned}
(1-\sigma)(e^{x/L} - e^{(1-\sigma)x}) &= (1-\sigma)(e^{x/L} - e^{(L-s)x/L}) && (5\ 114)\\
&= (1-\sigma)(e^{x/L} - (e^{x/L})^{(L-s)}) && (5\ 115)\\
&< 0 && (5\ 116)
\end{aligned}
$$

for all $x > 0$ Thus, since $\gamma > 0$ we have,

$$
1 + L(e^{\gamma/L} - 1)(1-\sigma) - e^{(1-\sigma)\gamma} < 0 \qquad (5\ 117)
$$

implying,

$$
1 + L(e^{\gamma/L} - 1)(1-\sigma) < e^{(1-\sigma)\gamma} \qquad (5\ 118)
$$

and hence, in conclusion,

$$\frac{1}{1 + L(e^{\gamma/L} - 1)(1 - \sigma)} > e^{(\sigma-1)\gamma} \tag{5 119}$$

We can use this inequality to prove, another useful inequality,

$$\frac{L(e^{\gamma/L} - 1)}{\tau} < \frac{\frac{1-\alpha}{\tau} + \sigma - 1}{\alpha(1-\sigma)} \tag{5 120}$$

$$= \frac{\frac{1}{\rho} - 1}{\alpha(1-\sigma)} \tag{5 121}$$

Proof

$$\frac{1}{1 + L(e^{\gamma/L} - 1)(1 - \sigma)} > e^{(\sigma-1)\gamma} \tag{5 122}$$

$$= \frac{\alpha L(e^{\gamma/L} - 1) + \tau}{L(e^{\gamma/L} - 1) + \tau} \tag{5 123}$$

by Equation 4 4 and Equation 5 112

For convenience write $u = L(e^{\gamma/L} - 1)$ Then the above is,

$$\frac{1}{1 + u(1-\sigma)} > \frac{\alpha u + \tau}{u + \tau} \tag{5 124}$$

Thus rearranging we get,

$$u(u(1-\sigma) - (1 - \alpha - \tau(1-\sigma)) < 0 \tag{5 125}$$

which implies,

$$u(1-\sigma) - (1 - \alpha - \tau(1-\sigma) < 0 \tag{5 126}$$

since $u > 0$ Thus,

$$u < \frac{(1 - \alpha - \tau(1-\sigma)}{(1-\sigma)} \tag{5 127}$$

Thus, replacing u with $L(e^{\gamma/L} - 1)$ and dividing the right hand side above and below by τ and replacing $\frac{1-\alpha}{\tau} + \sigma$ with $\frac{1}{\rho}$ we get in conclusion,

$$\frac{L(e^{\gamma/L} - 1)}{\tau} < \frac{\frac{1-\alpha}{\tau} + \sigma - 1}{\alpha(1-\sigma)} \tag{5 128}$$

$$= \frac{\frac{1}{\rho} - 1}{\alpha(1-\sigma)} \tag{5 129}$$

Returning to Equation 5 46 and Equation 5 79

$$(\frac{v(s-1)}{v(s)})^{\sigma}v(s) = \frac{(\sigma+\frac{1-\alpha}{\tau})(\frac{L(e^{\gamma/L}-1)+\tau}{\tau})^{\sigma}}{(\sigma+\frac{e^{\gamma(\sigma-1-1/L)}-1}{L})(1+\frac{L(e^{\gamma/L}-1)}{\tau})+\frac{1-\alpha}{\tau}}$$

Consider the following We know $\sigma \leq 1$ thus,

$$e^{\gamma(\sigma-1-1/L)} < 1$$

which implies,

$$\sigma+\frac{e^{\gamma(\sigma-1-1/L)}-1}{L} < \sigma$$

Thus,

$$v(s) > \frac{\sigma+\frac{1-\alpha}{\tau}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}$$

and recall that

$$(\frac{v(s-1)}{v(s)}) = 1+\frac{L(e^{\gamma/L}-1)}{\tau} \tag{5 130}$$

Thus,

$$(\frac{v(s-1)}{v(s)})^{\sigma}v(s) > \frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \tag{5 131}$$

Now we want to prove that the right hand side of this equation is greater than 1 We can prove this for $\sigma = 1/2$ and $\rho \leq \alpha$ To do this we first assume,

$$\frac{1-\alpha}{\tau} > \sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma} \tag{5 132}$$

then,

$$\frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma} > 0$$

implying,

$$\frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma} > 0$$

for $\sigma = 1/2$ But since $1+\frac{L(e^{\gamma/L}-1)}{\tau} > 0$ this implies,

$$\begin{aligned}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}(\frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}) &> \frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma} \\ &= \frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}\end{aligned}$$

Hence,

$$(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}(\frac{1-\alpha}{\tau})-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau}) \;>\; \frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}$$

Thus,

$$(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}(\frac{1-\alpha}{\tau})+\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma} \;>\; \frac{1-\alpha}{\tau}+\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})$$

implying,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \;>\; 1 \tag{5 133}$$

for σ under the assumption in Equation 5 132 Now assume the contrary to assumption Equation 5 132 That is, assume

$$\frac{1-\alpha}{\tau} \;\leq\; \sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma} \tag{5 134}$$

and further, assume

$$\rho \;\leq\; \alpha \tag{5 135}$$

We will now prove that both of the assumptions 5 134 and 5 135 cannot hold at the same time This will allow us to imply that if 5 135 is true then Equation 5 132 must be true implying Equation 5 133 for $\sigma = 1/2$ and 5 135 as we claimed Assumption 5 134 implies

$$\frac{1-\alpha}{\tau} \;\leq\; (1-\sigma)(1+\frac{L(e^{\gamma/L}-1)}{\tau}) \tag{5 136}$$

for $\sigma = 1/2$ Thus,

$$\frac{1-\alpha}{\tau}+(\sigma)(1+\frac{L(e^{\gamma/L}-1)}{\tau}) \;\leq\; 1+\frac{L(e^{\gamma/L}-1)}{\tau} \tag{5 137}$$

hence,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \;\geq\; \frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})} \tag{5 138}$$

$$= \;(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma-1} \tag{5 139}$$

$$= \;\frac{1}{\rho}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma-1} \tag{5 140}$$

Thus,

$$(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} > \frac{1}{\rho} \tag{5 141}$$

Now at $\sigma = 1/2$,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \tag{5 142}$$

is equal to,

$$(\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}})^{\sigma}$$

because,

$$(\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}})^{\sigma}$$

is,

$$((\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}})^{2}(1+\frac{L(e^{\gamma/L}-1)}{\tau}))^{\sigma}$$

which is,

$$(\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}})^{2\sigma}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}$$

and when $\sigma = 1/2$ we have $2\sigma = 1$ and the above is then,

$$(\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}$$

But, by 5 141,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}$$

is,

$$\begin{aligned}
&\geq \frac{1}{\rho}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \\
&= \frac{1}{\rho}\frac{\frac{1}{\rho}}{\frac{1}{\rho}+\sigma\frac{L(e^{\gamma/L}-1)}{\tau}} \\
&= \frac{1}{\rho+\rho^{2}\sigma\frac{L(e^{\gamma/L}-1)}{\tau}}
\end{aligned}$$

But by Equation 5 120,

$$
\begin{aligned}
\rho+\rho^2\sigma\frac{L(e^{\gamma/L}-1)}{\tau} &< \rho+\rho^2\sigma\frac{\frac{1}{\rho}-1}{\alpha(1-\sigma)} \\
&= \rho+\rho^2\frac{\frac{1}{\rho}-1}{\alpha} \\
&= \frac{\rho\alpha+\rho-\rho^2}{\alpha}
\end{aligned}
$$

where the second equation is due to $\sigma = 1/2$ Now it is a simple matter to show that the right hand side of the second equation is less than 1 under our assumption 5 135 Subtracting it from 1 we get,

$$
\begin{aligned}
1-\frac{\rho\alpha+\rho-\rho^2}{\alpha} &= \frac{\alpha-\rho\alpha-\rho+\rho^2}{\alpha} \\
&= \frac{(\rho-\alpha)(\rho-1)}{\alpha}
\end{aligned}
$$

Now $\rho < 1$ so this is positive for $\rho \leq \alpha$ which we have assumed in Equation 5 135 This then implies, for $\rho \leq \alpha$ and $\sigma = 1/2$,

$$
\begin{aligned}
\rho+\rho^2\sigma L(e^{\gamma/L}-1) &< \rho+\rho^2\sigma\frac{\frac{1}{\rho}-1}{\alpha(1-\sigma)} \\
&< 1
\end{aligned}
$$

which implies,

$$
\frac{1}{\rho+\rho^2\sigma\frac{L(e^{\gamma/L}-1)}{\tau}} > 1
$$

Thus for $\sigma = 1/2$

$$
\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}
$$

is,

$$
\begin{aligned}
&\geq \frac{1}{\rho}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \\
&> 1
\end{aligned}
$$

and hence, for $\sigma = 1/2$

$$
(\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}
$$

is,

$$
\begin{aligned}
&\geq \frac{1}{\rho}\frac{(\sigma+\frac{1-\alpha}{\tau})}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}})^{\sigma} \\
&> 1
\end{aligned}
$$

Thus, by Equation 5 142, for $\sigma = 1/2$,

$$
\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} > 1 \qquad (5\ 143)
$$

But, returning to our first assumption in Equation 5 134

$$
\frac{1-\alpha}{\tau} \leq \sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}
$$

then,

$$
\frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma} \leq 0
$$

implying,

$$
\frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma} \leq 0
$$

for $\sigma = 1/2$ But since $1+\frac{L(e^{\gamma/L}-1)}{\tau} > 0$ this implies,

$$
\begin{aligned}
(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}(\frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma}) &\leq \frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{1-\sigma} \\
&= \frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}
\end{aligned}
$$

Hence,

$$
(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}(\frac{1-\alpha}{\tau})-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau}) \leq \frac{1-\alpha}{\tau}-\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}
$$

Thus,

$$
(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}(\frac{1-\alpha}{\tau})+\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma} \leq \frac{1-\alpha}{\tau}+\sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})
$$

implying,

$$
\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \leq 1
$$

contradicting Equation 5 143 Hence our two assumptions 5 134 and 5 135 cannot both be true at the same time

Thus we can say the following If we choose $\rho \leq \alpha$ then

$$\frac{1-\alpha}{\tau} > \sigma(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}$$

for $\sigma = 1/2$ And then by Equation 5 133

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} > 1$$

Hence finally by Equation 5 131

$$(\frac{v(s-1)}{v(s)})^{\sigma}v(s) > 1 \quad (5\ 144)$$

for $\sigma = 1/2$ and $\rho \leq \alpha$

This tells us, that the upper bound is an improvement over the effective bandwidth approximation at least for this set of parameter values We would also like to know if,

$$\begin{aligned}(\frac{v(s-1)}{v(s)})^{\sigma}v(s) &> k \quad (5\ 145)\\ &> 1 \quad (5\ 146)\end{aligned}$$

where k is independent of L, as this would tell us something about economies of scale To this end consider the following

Recall from Equations 4 4, 4 6 and 4 5

$$\begin{aligned}e^{(\sigma-1)\gamma(L)} &= \frac{\alpha(e^{\gamma/L}-1)+\frac{\tau}{L}}{(e^{\gamma/L}-1)+\frac{\tau}{L}}\\ &= \frac{\alpha L(e^{\gamma/L}-1)+\tau}{L(e^{\gamma/L}-1)+\tau}\\ &< \frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}\end{aligned}$$

Thus upon rearrangement we have,

$$\sigma < 1+\frac{1}{\gamma(L)}\log(\frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}) \quad (5.147)$$

Now consider the derivative of the right hand side of this equation with respect to $\gamma(L)$ This is,

$$-\frac{1}{\gamma(L)^2}\log(\frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}) - \frac{1}{\gamma(L)}\frac{\gamma(L)+\tau}{\alpha\gamma(L)+\tau}\frac{(1-\alpha)\tau}{(\gamma(L)+\tau)^2}$$

We want to know if this is positive or otherwise Now rearranging and multiplying across by $\alpha\gamma(L)+\tau > 0$ and $\gamma(L)^2 > 0$ and simplifying we get,

$$-(\alpha\gamma(L)+\tau)\log(\frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}) - \gamma(L)\frac{(1-\alpha)\tau}{(\gamma(L)+\tau)}$$

Now consider,

$$-(\alpha u+\tau)\log(\frac{\alpha u+\tau}{+\tau}) - u\frac{(1-\alpha)\tau}{(u+\tau)}$$

at $u=0$ this is zero, and its derivative with respect to u is,

$$-\alpha\log(\frac{\alpha u+\tau}{u+\tau}) + (\alpha u+\tau)\frac{u+\tau}{\alpha u+\tau}\frac{(1-\alpha)\tau}{(u+\tau)^2} - \frac{(1-\alpha)\tau}{(u+\tau)} + u\frac{(1-\alpha)\tau}{(u+\tau)^2}$$

which is,

$$-\alpha\log(\frac{\alpha u+\tau}{u+\tau}) + u\frac{(1-\alpha)\tau}{(u+\tau)^2} \quad > \quad 0$$

for all $u > 0$ since $\frac{\alpha u+\tau}{u+\tau} < 1$ Thus the derivative of the right hand side of equation 5 147 with respect to $\gamma(L)$ is positive for all $\gamma(L) > 0$ Now consider,

$$\sigma \quad = \quad 1+\frac{1}{\gamma}\log(\frac{\alpha\gamma+\tau}{\gamma+\tau}) \tag{5 148}$$

The solution of this is independent of L and the derivative of the right hand side with respect to γ is positive for all $\gamma > 0$ Thus we have,

$$1+\frac{1}{\gamma}\log(\frac{\alpha\gamma+\tau}{\gamma+\tau}) \quad = \quad \sigma \tag{5 149}$$

$$< \quad 1+\frac{1}{\gamma(L)}\log(\frac{\alpha\gamma(L)+\tau}{\gamma(L)+\tau}) \tag{5 150}$$

implying,

$$\gamma(L) > \gamma \tag{5 151}$$

Thus we have,

$$1+\frac{L(e^{\gamma/L}-1)}{\tau} > 1+\frac{\gamma(L)}{\tau} \tag{5.152}$$
$$> 1+\frac{\gamma}{\tau} \tag{5.153}$$

Now consider,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma(L)/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma(L)/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}$$

Replace $1+\frac{L(e^{\gamma(L)/L}-1)}{\tau}$ by x everywhere in this quotient. This becomes,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})x^{\sigma}}{x\sigma+\frac{1-\alpha}{\tau}}$$

Now differentiate this with respect to x. This gives us,

$$\frac{(x\sigma+\frac{1-\alpha}{\tau})\sigma(\sigma+\frac{1-\alpha}{\tau})x^{\sigma-1}-(\sigma+\frac{1-\alpha}{\tau})\sigma x^{\sigma}}{(x\sigma+\frac{1-\alpha}{\tau})^2} = \frac{x^{\sigma-1}\sigma(\sigma+\frac{1-\alpha}{\tau})(\frac{1-\alpha}{\tau}-(1-\sigma)x)}{(x\sigma+\frac{1-\alpha}{\tau})^2}$$

This is positive if,

$$\frac{1-\alpha}{\tau} > (1-\sigma)x \tag{5.154}$$

Now assume,

$$\frac{1-\alpha}{\tau} > (1-\sigma)(1+\frac{L(e^{\gamma(L)/L}-1)}{\tau}) \tag{5.155}$$

Then

$$\frac{1-\alpha}{\tau} > (1-\sigma)x \tag{5.156}$$

for all $x < 1+\frac{L(e^{\gamma(L)/L}-1)}{\tau}$. Thus,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(x)^{\sigma}}{x\sigma+\frac{1-\alpha}{\tau}}$$

is increasing for all $x < 1+\frac{L(e^{\gamma(L)/L}-1)}{\tau}$. Thus by Equation 5.152 we have,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} > \frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{\gamma}{\tau})^{\sigma}}{(1+\frac{\gamma}{\tau})\sigma+\frac{1-\alpha}{\tau}}$$

But if $\sigma = 1/2$ and $\rho \leq \alpha$ then, we have by Equation 5 144

$$\begin{aligned} \frac{1-\alpha}{\tau} &> \sigma(1 + \frac{L(e^{\gamma/L} - 1)}{\tau})^{\sigma} \\ &> \sigma(1 + \frac{\gamma}{\tau})^{\sigma} \end{aligned}$$

and the same argument as that used in Equation 5 132 to Equation 5 133 can be used to show,

$$\frac{(\sigma + \frac{1-\alpha}{\tau})(1 + \frac{\gamma}{\tau})^{\sigma}}{(1 + \frac{\gamma}{\tau})\sigma + \frac{1-\alpha}{\tau}} > 1$$

with the left hand side independent of L Thus we have,

$$\begin{aligned} \frac{(\sigma + \frac{1-\alpha}{\tau})(1 + \frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1 + \frac{L(e^{\gamma/L}-1)}{\tau})\sigma + \frac{1-\alpha}{\tau}} &> \frac{(\sigma + \frac{1-\alpha}{\tau})(1 + \frac{\gamma}{\tau})^{\sigma}}{(1 + \frac{\gamma}{\tau})\sigma + \frac{1-\alpha}{\tau}} \\ &> 1 \end{aligned}$$

Now assume the contrary to Equation 5 155 that is assume,

$$\frac{1-\alpha}{\tau} \leq (1-\sigma)(1 + \frac{L(e^{\gamma(L)/L} - 1)}{\tau}) \tag{5 157}$$

and let $\sigma = 1/2$ and $\rho \leq \alpha$ then the argument in 5 157 to 5 143 gives us,

$$\begin{aligned} \frac{(\sigma + \frac{1-\alpha}{\tau})(1 + \frac{L(e^{\gamma/L}-1)}{\tau})^{\sigma}}{(1 + \frac{L(e^{\gamma/L}-1)}{\tau})\sigma + \frac{1-\alpha}{\tau}} &> \frac{1}{\rho + \rho^2\sigma\frac{\frac{1}{\rho}-1}{\alpha(1-\sigma)}} \\ &> 1 \end{aligned}$$

where the second of the quotients is independent of L Note that we have **not** assumed here that

$$\frac{1-\alpha}{\tau} \leq \sigma(1 + \frac{L(e^{\gamma(L)/L} - 1)}{\tau})^{\sigma} \tag{5 158}$$

so we couldn't have used the contradiction argument used in Equations 5 134 to 5 144 Thus we have proved that if $\sigma = 1$ and $\rho \leq \alpha$ then,

$$(\frac{v(s-1)}{v(s)})^{\sigma} v(s) > k \tag{5 159}$$

$$> 1 \tag{5 160}$$

where k is independent of L

In conclusion then, for $\sigma = 1/2$ and $\rho \leq \alpha$ we, have

$$\mathbf{P}[q_s^L > b] < \Phi^L e^{-\gamma b} \tag{5 161}$$

where $\Phi < 1$ is independent of L We can say that,

$$\Phi = \max\{\rho + \rho^2\sigma\frac{\frac{1}{\rho} - 1}{\alpha(1-\sigma)}, \frac{(1+\frac{\gamma}{\tau})\sigma + \frac{1-\alpha}{\tau}}{(\sigma + \frac{1-\alpha}{\tau})(1+\frac{\gamma}{\tau})^\sigma}\} \tag{5 162}$$

more generally we have,

$$\mathbf{P}[q_s^L > b] \leq (\frac{v(s)e^\gamma}{v(s-1)})\frac{v(s)}{v(s-1)}((\frac{v(s)}{v(s-1)})^\sigma \frac{1}{v(s)})^L e^{-\gamma b} \tag{5 163}$$

$$= e^\gamma(\frac{\tau}{L(e^{\gamma/L}-1)+\tau})^2(\frac{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma + \frac{1-\alpha}{\tau}}{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^\sigma})^L e^{-\gamma b} \tag{5 164}$$

for all parameter values But it has to be determined if this is less than 1 for $\sigma \neq 1/2$ or $\rho > \alpha$ and if

$$(\frac{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma + \frac{1-\alpha}{\tau}}{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^\sigma}) < 1 \tag{5 165}$$

which would imply economies of scale exist for the particular values of σ and α and τ Finally we note that the condition, $\rho > \alpha$, is the same as,

$$\tau < \frac{\alpha(1-\alpha)}{1-\sigma\alpha} \tag{5 166}$$

Chapter 6

Large Deviations Approximations

6.1 Large deviations

The development of large deviation theory began with Cramer who proved the theorem that bears his name for the large deviations of the empirical mean of a sequence of i.i.d. random variables. Chernoff subsequently discovered a general method for calculating the rate function from the cumulant generating function of the i.i.d. random variables. Gartner and Ellis generalised Cramer's theorem for the case when the random variables are not i.i.d. and, from a major result of Varadhan, a generalisation of the method of Chernoff for calculating the rate function was shown to apply to the case of non-i.i.d random variables. This obtains the rate function from the Legendre-Fenchel transform of the scaled cumulant generating function [2]. In this chapter we apply the work of Botvich and Duffield [20] to our new cell level model in order to find (what they called) the Shape Function for this model. We also prove some new results on the shape function in the case of non-negatively associated workload increments that flow from the definition of the shape function and from the sub-additivity theorem (Lemma 6.1.11 of [21]). We begin in this Section with the definitions of a Large Deviation Principle, the rate function and the cumulant generating function. In Section 6.2 we describe the work of Botvich and Duffield from [20]. In Section 6.3 we describe some of the properties of the Legendre-Fenchel Transform giving a number of well known results and some consequences not previously outlined for the shape function We also prove some new theorems on the shape function in the case of non-negatively associated workload increments in subsection 6 3 3 In Section 6 4 we apply the results of Botvich and Duffield [20] to find the shape function for the cell level model We relate this to the simulations of [9] This is also related to the work we carried out in [22] In Section 6 5 we describe the relationship between the shape function and economies of scale, this is from work we carried out in [22] We will now state formally what is meant by Large Deviation Principle, rate function and cumulant generating function

Definition 4 *Let $\{\mathbf{P}_n\}$ be a sequence of probability measures on the real numbers Then $\{\mathbf{P}_n\}$ is said to satisfy a Large Deviation Principle with rate function I and constants V_n if there is a function I $\mathbf{R} \to [0, \infty]$ and a sequence of positive numbers $\{V_n\}$ diverging to $+\infty$ and,*

$$\limsup_{n\to\infty} \frac{1}{V_n} \log \mathbf{P}_n[C] \leq -\inf_{x\in C} I(x) \tag{6 1}$$

for C closed and,

$$\liminf_{n\to\infty} \frac{1}{V_n} \log \mathbf{P}_n[G] \geq -\inf_{x\in G} I(x) \tag{6 2}$$

for G open

Now note that if A is a set with,

$$\begin{aligned} \inf_{x\in A'} I(x) &= \inf_{x\in A} I(x) \\ &= \inf_{x\in \bar{A}} I(x) \end{aligned}$$

where A' is the interior of A and $\bar{A}$ is the closure of A, then

$$\lim_{n\to\infty}\inf \frac{1}{V_n} \log \mathbf{P}_n[A] = -\inf_{x\in A} I(x)$$

and we write,

$$\mathbf{P}_n[A] \approx e^{-V_n \inf_{x\in A} I(x)}$$

For a sequence of random variables, $\{X_t\}$ the cumulant generating function is defined by,

Definition 5 *The cumulant generating function of the sequence of random variables, $\{X_t\}$ is, the real function of real θ*

$$g(\theta) \quad = \quad \log \mathbf{E}[e^{\theta X_t}] \tag{6 3}$$

This can be used to obtain the rate function I and both I and g are convex Thus, for example, if $A = [x_0, \infty]$ then,

$$\begin{aligned} \liminf_{n \to \infty} \frac{1}{V_n} \log \mathbf{P}_n[A] &= - \inf_{x \in [x_0, \infty]} I(x) \\ &= -I(x_0) \end{aligned}$$

6.2 Large deviations and Queues

Let W_t be the workload process for a general single server queue, where t is discrete or real time That is let, A_t be the arrivals to be served in the interval $[-t, 0)$ and let S_t be the service that can be completed in the same time interval Then if the workload has stationary increments and the queue has a FCFS service discipline the queue length will have a unique stationary distribution [14], and the equilibrium queue length will be given by,

$$q = \sup_{t \geq 0} W_t \tag{6 4}$$

For such a general single server queue Glynn and Whitt [12] showed that under very general conditions if the pair $(\frac{W_t}{t}, t)$ with $t \in Z^+$ satisfies a Large Deviation Principle with rate function I, i e ,

$$\mathbf{P}[\frac{W_t}{t} \geq w] \approx e^{-tI(w)} \tag{6 5}$$

then,

$$\mathbf{P}[q \geq b] \approx e^{-\gamma b} \tag{6 6}$$

where,

$$\gamma = \inf_{w} \frac{I(w)}{w} \tag{6 7}$$

This was generalised in [23] for $t \in R$ and more general scaling functions than t.

Now we are interested in the queue q^L in an infinite buffer generated by the L-fold homogeneous superposition of independent sources served at a constant service rate and modelled by the cell level model. Let the superposed workload process be W_t^L and let $W_0^L = 0$. The simulations of Corcoran [9] demonstrate that the broad features of the queue length distribution of this queue remain essentially unchanged when L and the queue length b are jointly scaled. Thus we are led to consider the large deviation properties of the queue length distribution in L.

It was proved in [20] for more general situations than ours, under hypotheses that follow, that

Theorem 3 *For $b \geq 0$,*

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbf{P}[q^L > Lb] = -I(b)$$

Where the function I, termed the shape function, is related to the cumulant generating function of the workload process W_t^L.

The hypotheses under which this result holds are,

Hypothesis 1 *Let, the rescaled cumulant generating function for the workload process W_t^L be,*

$$\lambda_t^L(\theta) \quad = \quad \frac{1}{Lt} \lim_{L\to\infty} \log \mathbf{E}[e^{\theta W_t^L}] \tag{6.8}$$

then,

- *For each real θ the limits,*

$$\begin{aligned} \lambda_t(\theta) &= \lim_{L\to\infty} \lambda_t^L(\theta) \\ \lambda(\theta) &= \lim_{t\to\infty} \lambda_t(\theta) \end{aligned}$$

  exist as extended real numbers. With the first limit existing uniformly for all t sufficiently large.

- *The functions λ and λ_t are both differentiable on the regions where they are finite (effective domain) and $\lim_{n\to\infty} |\lambda'(\theta_n)| = +\infty$ for any sequence $\{\theta_n\}_n$, in the effective domain, which converges to a point on its boundary.*

- *There exists $\theta > 0$ for which $\lambda_t(\theta) < 0$ for all t.*

- *For real t, We define $\hat{W}^L_{t,r} = \sup_{0<r'<r} W^L_{t-r'} - W^L_t$ for $t \geq r \geq 0$. Then for all real θ,*

$$\limsup_{r \to 0} \lim_{L\to\infty} \sup \frac{1}{L} \sup_{t \geq 0} \log \mathbf{E}[e^{\theta \hat{W}^L_{t,r}}] \leq 0$$

The function I called the shape function is,

$$I(b) = \inf_{t>0} t\lambda^*_t(b/t)$$

Comparing these hypotheses to the conditions in the Gartner-Ellis theorem, Duffield and Botvich [20] note that the first two hypotheses mean that for each fixed t the pair (W^L_t, L) satisfies a Large Deviation Principle with rate function $(t\lambda_t)^*$. That is,

$$\begin{aligned}
\lim_{L\to\infty} \sup \frac{1}{L} \log \mathbf{P}[\frac{W^L_t}{L} \in A] &\leq -\inf_{w \in \bar{A}} (t\lambda_t)^*(w) \\
\lim_{L\to\infty} \inf \frac{1}{L} \log \mathbf{P}[\frac{W^L_t}{L} \in A] &\geq -\inf_{w \in A^0} (t\lambda_t)^*(w)
\end{aligned}$$

Now $(t\lambda_t)^*(w) = t(\lambda^*_t)(w/t)$, thus, the third hypothesis means that any root w^*_t of $(t\lambda_t)^*(w)$ is negative, and we then have, for $w \geq w^*_t$,

$$\begin{aligned}
\lim_{L\to\infty} \sup \frac{1}{L} \log \mathbf{P}[\frac{W^L_t}{L} > w] &\leq -(t\lambda_t)^*(w) \\
\inf_{L\to\infty} \frac{1}{L} \log \mathbf{P}[\frac{W^L_t}{L} > w] &\geq -\lim_{t \downarrow w} (t\lambda_t)^*(w)
\end{aligned}$$

It is noted in [20] that the third hypothesis also guarantees the existence of a strictly positive solution γ of the equation $\lambda(\theta) = 0$ which is the asymptotic decay rate of the queue length distribution (recall chapter 4). The fourth hypothesis is a local regularity condition on the sample paths of the workload process.

Note that the result proved in [20] does not assume that the superposition is of i.i.d sources or that the sources are independent at all. We note that in the case of a homogeneous superposition of i.i.d workload processes the $\lambda_t(\theta) = \lim_{L\to\infty} \lambda^L_t(\theta)$ condition is redundant, as then Cramer's Theorem guarantees that the superposed workload process satisfies a Large Deviation Principle with rate function as described above [15].

It was also noted in [20] that the differentiability condition guarantees that there is a unique w_t^* for which $(t\lambda_t)^*(w_t^*) = 0$. This number is such that,

$$\frac{\mathbf{E}[W_t^L]}{L} \xrightarrow{L} w_t^*$$

and in fact, $\frac{W_t^L}{L}$ converges exponentially to w_t^* as L tends to infinity [15],

$$\frac{W_t^L}{L} \overset{\exp}{\to} w_t^*$$

that is, for any $\epsilon > 0$ there exists a number $N > 0$, dependent on ϵ, such that, for all L sufficiently large,

$$\mathbf{P}[|\frac{W_t^L}{L} - w_t^*| \geq \epsilon] \leq e^{-LN}$$

It was noted in [20] that in the case of a homogeneous superposition of independent sources, with single source workload W_t,

$$w_t^* = \mathbf{E}[W_t]$$

and $\frac{W_t^L}{L}$ converges exponentially to $\mathbf{E}[W_t]$ as L tends to infinity. Further, since $\sum_{L=1}^{\infty} e^{-LN}$ is finite, the Borel-Cantelli Lemma implies that $\frac{W_t^L}{L}$ satisfies the Strong Law of Large Numbers, that is,

$$\frac{W_t^L}{L} \to \mathbf{E}[W_t] \quad \text{a.s.}$$

The reason why the result proved in [20] works is roughly the following

$$\begin{aligned} \mathbf{P}[q^L > Lb] &= \mathbf{P}[\sup_{t\geq 0} W_t^L > Lb] \\ &= \mathbf{P}[\bigcup_{t\geq 0}\{W_t^L > Lb\}] \end{aligned}$$

The probability of each event in the union is exponentially small for large L. Thus the probability is dominated by the largest of the probabilities of each of the events in the union, in other words by the probability of the most likely event. This is,

$$\sup_{t\geq 0} \mathbf{P}[W_t^L > Lb]$$

Now if for each fixed t, (W_t^L, L) satisfies a Large Deviation Principle with rate function $(t\lambda_t)^*$ then,

$$\mathbf{P}[W_t^L > Lb] \approx e^{-L(t\lambda_t)^*(b)}$$

In other words we have roughly the following,

$$\begin{aligned}
\mathbf{P}[q^L > Lb] &\approx \sup_{t\geq 0} \mathbf{P}[W_t^L > Lb] \\
&\approx \sup_{t\geq 0} e^{-L(t\lambda_t)^*(b)} \\
&\approx e^{-L \inf_{t\geq 0}(t\lambda_t)^*(b)} \\
&= e^{-LI(b)}
\end{aligned}$$

It was also proved in [20] that the asymptotics of $I(b)$ are,

$$\lim_{b\to\infty} (I(b) - b\gamma) = \nu \tag{6 9}$$

where,

$$\nu = -\lim_{t\to\infty} t\lambda_t(\gamma), \tag{6 10}$$

provided this limit exists, and subject to some regularity requirements in the case of discrete t For large b we can approximate $I(b)$ by [20, 22],

$$I(b) \approx \nu + b\gamma \tag{6 11}$$

Thus for large b and large L we have,

$$\begin{aligned}
\mathbf{P}[q^L > b] &\approx e^{-LI(b/L)} \\
&\approx e^{-(L\nu + b\gamma)} \\
&= e^{-L\nu} e^{-b\gamma}
\end{aligned} \tag{6 12}$$

Thus we can see from 6 12 that in multiplexer models ν determines the economies of scale [22] that can be obtained from statistically multiplexing large numbers of sources Note that, $\nu = 0$ for uncorrelated arrivals as then $\lambda_t(\gamma) = \lambda(\gamma) = 0$ Therefore there are no economies of scale to be obtained from multiplexing large numbers of sources with uncorrelated arrivals If, however the increments of the workloads on disjoint intervals are positively associated then $\nu \geq 0$ [20, 22]

For $t > 0$, if we define $\Lambda_t(\theta) = t\lambda_t(\theta/t)$ and we assume that the workload W_t^L has stationary increments and we define $\Lambda(\theta) = \lim_{t\to 0} \Lambda_t(\theta)$ for real t and assume that the limit exists as an extended real number for all real θ and we make one further assumption, namely, that 0 is in the effective domain of Λ^* Then,

$$I(0) \;=\; \lambda_1(0)$$

for discrete t, and

$$I(0) \;=\; \Lambda^*(0) \tag{6 13}$$

for real t This tells us that for large L (under the conditions given) the workload is most likely to exceed 0 at the smallest times [20]

6.3 The Legendre-Fenchel Transform and the Shape Function

In this section we will define the Legendre-Fenchel Transform f^* of a function f and describe some of its general properties [15] We describe how to calculate the shape function $I(b)$ form the Legendre transform of the cumulant generating function of the workload process W_t^L We show how the derivative of the shape function with respect to b is related to the cumulant generating function In subsection 6 3 2 we prove the new result that the shape function is sub-additive if the increments of the workload are non-negatively associated This has consequences for the shape of the shape function which we demonstrate Under this condition on the workload increments, and assuming $I(0) = 0$, the shape function cannot be convex

6.3.1 The Legendre Transform

Definition 6 *[15] Let f $R \to R$ be a strictly convex function The Legendre-Fenchel transform of f, denoted by f^* is defined by,*

$$f^*(y) = \sup_{x \in R}\{xy - f(x)\} \textit{for } y \in R$$

Lemma 1 *[15] Let f $R \to R$ be a convex function Then,*

1 $(f^*)' = (f')^{-1}$

2 $f(x) = \sup_{y \in R}\{xy - f^*(y)\}$

3 $(f^*)^* = f$ *if f is closed*

4 f^* *is a convex function also known as the conjugate of f*

6.3.2 Some Generalities

Most of the following lemmas and definitions are used either to prove further more useful lemmas or are used in section 6 3 2 Lemma 3 is used to prove Lemma 4 which ın turn ıs used to prove Lemma 5 whıch ıs ıtself used to prove Theorem 4 ın Sectıon 6 3 3 Lemma 6 tells us how to calculate the rate functıon $f_t^*(b)$ from the cumulant generatıng function $f(t,\theta)$ The proofs of well known results are brıef and so are provıded for completeness We begın wıth some definitıons

Definition 7 *Let $\theta_{t,b}$ be the θ at whıch* $\sup_\theta\{b\theta - t\lambda_t(\theta)\}$ *occurs ıf such a θ exısts*

Definition 8 *Let t_b be the unıque t at whıch the* $\inf_{t>0} t\lambda_t^*(b/t)$ *occurs ıf such a t exısts*

Definition 9 *Let $f(t,\theta) = t\lambda_t(\theta)$*

Thıs sımplifies the notatıon slıghtly as we can now wrıte

Definition 10

$$I(b) = \inf_{t>0} f_t^*(b)$$

The followıng Lemma is well known and ıs used later to prove Lemma 6,

Lemma 2 *$f(t,\theta)$ ıs convex as a functıon of θ for $\theta > 0$ ımplyıng $f''(t,\theta) > 0$ for all θ and all $t \geq 0$*

Proof

From Defimtıon6 we have $f(t,\theta) = \log \mathbf{E}[e^{\theta W_t}]$ Now by Schwarz ınequalıty we have,

$$(\mathbf{E}[e^{\theta W_t}])^2 \leq \mathbf{E}[e^{(\theta+\zeta)W_t}]\mathbf{E}[e^{(\theta-\zeta)W_t}]$$

where $0 \leq \zeta \leq \theta$ Thus lettıng $\theta_1 = \theta - \zeta$ and $\theta_2 = \theta + \zeta$ we get

$$\log[\mathbf{E}[e^{\frac{(\theta_1+\theta_2)}{2}W_t}]] \leq \frac{\log[\mathbf{E}[e^{\theta_1 W_t}]] + \log[\mathbf{E}[e^{\theta_2 W_t}]]}{2}$$

provıng that $f(t,\theta)$ ıs convex as a functıon of $\theta \geq 0$ for all $t \geq 0$ hence,

$$f''(t,\theta) \geq 0$$

for $\theta \geq 0$ and $t \geq 0$

We can show the known fact that $f_t^*(b)$ ıs non-negatıve as follows,

Lemma 3 *Let $f(t, \theta)$ be defined as in Definition 9 and $f_t^*(b)$ be the Legendre transform of f Then $f_t^*(b) \geq 0$*

Proof

It is obvious that,

$$b\theta - f(t, \theta) \;=\; 0$$

for $\theta = 0$ for all $t \geq 0$ and all real b

Thus,

$$\sup_{\theta}\{b\theta - f(t, \theta)\} \geq 0$$

We can also prove the known fact,

Lemma 4 *Let $f(t, \theta)$ be defined as in Definition 9 and $f_t^*(b)$ be the Legendre transform of f Then, $f_t^*(\mathbf{E}[W_t]) = 0$*

Proof

For all θ we have by Jensens inequality [8],

$$\begin{aligned} f(t, \theta) &= \log \mathbf{E}[e^{\theta W_t}] \\ &\geq \log e^{\theta \mathbf{E}[W_t]} \\ &= \theta \mathbf{E}[W_t] \end{aligned}$$

Thus,

$$\mathbf{E}[W_t]\theta - f(t, \theta) \;\leq 0$$

But this implies,

$$f_t^*(\mathbf{E}[W_t]) \;\leq\; 0$$

Implying by Lemma 3 that,

$$f_t^*(\mathbf{E}[W_t]) \;=\; 0$$

The next three lemma's are also known results supplied only because they are usedlater. The following Lemma is used in the proof of Theorem 4,

Lemma 5 *Let $f(t,\theta)$ be defined as in Definition 9. Then, for $b \geq \mathbf{E}[W_t]$*

$$\sup_{\theta}\{b\theta - f(t,\theta)\} \;=\; \sup_{\theta\geq 0}\{b\theta - f(t,\theta)\}$$

Proof

For $\theta < 0$ and $b > \mathbf{E}[W_t]$

$$\begin{aligned} b\theta - f(t,\theta) &< \mathbf{E}[W_t]\theta - f(t,\theta) \\ &\leq 0 \end{aligned}$$

But

$$\sup_{\theta}\{b\theta - f(t,\theta)\} \;\geq\; 0$$

Thus,

$$\sup_{\theta}\{b\theta - f(t,\theta)\} \;=\; \sup_{\theta\geq 0}\{b\theta - f(t,\theta)\}$$

The following Lemma tells us how to calculate the rate function $f_t^*(b)$ from the cumulant generating function $f(t,\theta)$.

Lemma 6 *Let $f(t,\theta)$ be defined as in Definition 9 and $f_t^*(b)$ be the Legendre transform of f. Then,*

$$f_t^*(b) \;=\; b\theta_{t,b} - f(t,\theta_{t,b})$$

where $\theta_{t,b}$ is the unique $\theta \geq 0$ such that, for fixed t and b

$$f'(t,\theta) \;=\; b$$

if this equation has a solution for θ.

If on the other hand no such solution exists then, the supremum is only attained at infinity.

Proof

By definition,

$$f_t^*(b) = \sup_\theta \{b\theta - f(t,\theta)\}$$

Then by Lemma 5,

$$f_t^*(b) = \sup_{\theta \geq 0} \{b\theta - f(t,\theta)\}$$

But by Lemma 2 $f(t,\theta)$ is convex as a function of θ Thus $b\theta - f(t,\theta)$ is concave as function of θ Hence, if

$$f'(t,\theta) = b$$

for some $\theta = \theta_{t,b}$ Then,

$$f_t^*(b) = b\theta_{t,b} - f(t,\theta_{t,b})$$

and $\theta_{t,b} \geq 0$

On the other hand if,

$$f'(t,\theta) \neq b$$

for any θ Then,

$$b - f'(t,\theta) \neq 0$$

for any θ Thus as $b\theta - f(t,\theta)$ is concave,

$$b - f'(t,\theta) > 0$$

for all $\theta > 0$ and therefore, since $b\theta - f(t,\theta)$ is strictly increasing,

$$f_t^*(b) = \lim_{\theta \to \infty} (b\theta - f(t,\theta))$$

The next Lemma is used in the proof of Lemma 8,

Lemma 7 *Let $f(t,\theta)$ be defined as in Definition 9 and $f_t^*(b)$ be defined as in Definition 6 and $\theta_{t,b}$ be defined as in Definition 7 Then,*

$$(f_t^*)'(b) = \theta_{t,b}$$

Proof

By Lemma 6,

$$f_t^*(b) = b\theta_{t,b} - f(t,\theta_{t,b})$$

Thus, from Lemma 2 we have,

$$\begin{aligned}
(f_t^*)'(b) &= \theta_{t,b} + b\frac{\partial\theta_{t,b}}{\partial b} - f'(t,\theta)|_{\theta=\theta_{t,b}}\frac{\partial\theta_{t,b}}{\partial b} \\
&= \theta_{t,b} + b\frac{\partial\theta_{t,b}}{\partial b} - b\frac{\partial\theta_{t,b}}{\partial b} \\
&= \theta_{t,b}
\end{aligned}$$

With t_b defined as in Definition 8 we can write the following equation for the shape function $I(b)$,

$$I(b) = f_{t_b}^*(b) \tag{6 14}$$

And we have the following for the slope of $I(b)$ at any b,

Lemma 8 *Let t_b be defined as in Definition 8 and be finite and non-zero Let $\theta_{t,b}$ be defined as in Definition 7 Let $(t_b, f_{t_b}^*(b))$ be a local minimum point for $f_{t_b}^*(b)$ Then,*

$$I'(b) = \theta_{t_b,b}$$

Proof

By Equation 6 5,

$$I(b) = f_{t_b}^*(b)$$

thus ,

$$I'(b) = \frac{df_{t_b}^*(b)}{db}$$

$$
\begin{aligned}
&= f_t^*(b)|_{t_b}\frac{dt_b}{db} + (f_{t_b}^*)'(b) \\
&= 0 + (f_{t_b}^*)'(b) \\
&= (f_{t_b}^*)'(b) \\
&= \theta_{t_b,b}
\end{aligned}
$$

where the third equality follows from the fact that $(t_b, f_{t_b}^*(b))$ is a local minimum point The last equality follows from Lemma 7

The implications of Lemmas 3 to 8 can be summarised by the following diagram, where we have assumed for the purposes of illustration that $I(b)$ is concave (Diagram Over-leaf)

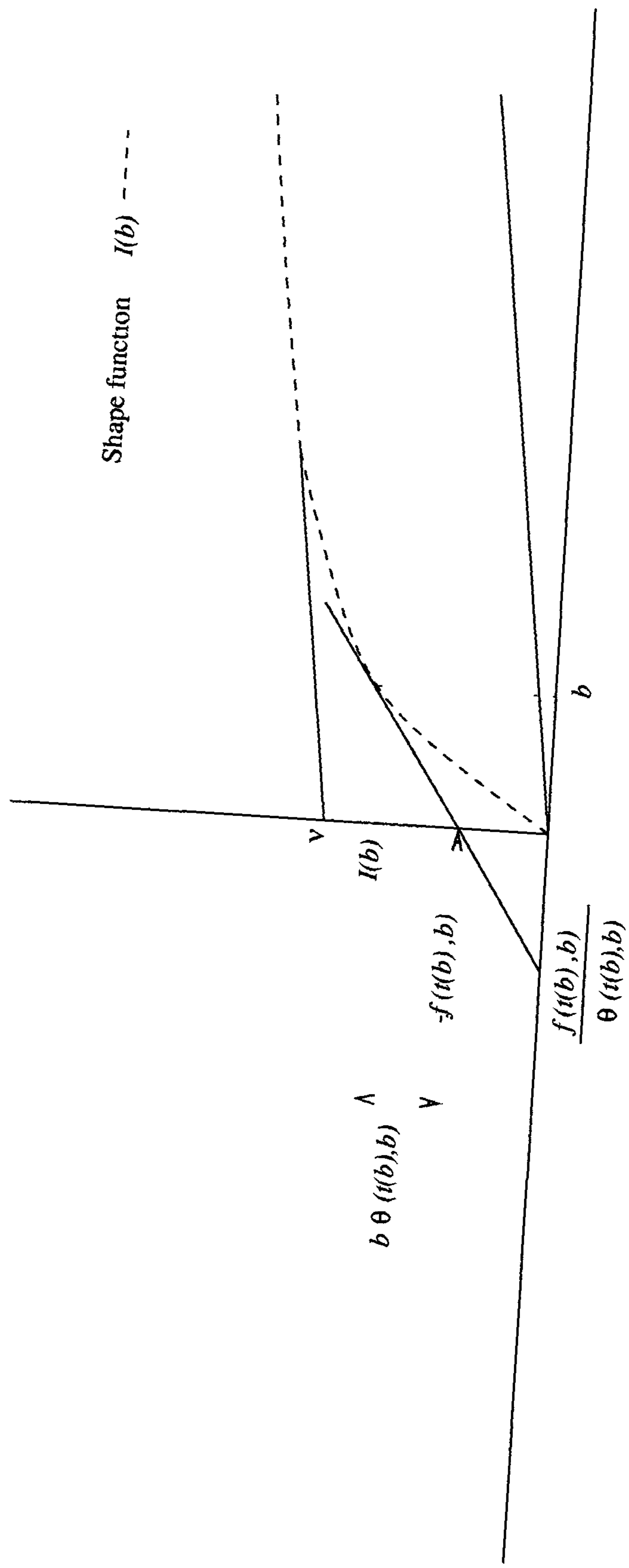

Figure 6-1 *illustration*

The next Lemma tells us that $\frac{\partial\theta_{t,b}}{\partial b} > 0$ for all b When we combine that with Lemma 10 we see that the sign of $I''(b) = \frac{d\theta_{t(b),b}}{db}$ depends on $f'(t(b), \theta_{t,b})\frac{dt(b)}{db}$

Lemma 9 *Let $f(t,\theta)$ be defined as in Definition 9 and $\theta_{t,b}$ be defined as in Definition 7 be the unique finite θ such that $f'(t,\theta) = b$ for all b on some interval Let $f''(t,\theta_{t,b}) > 0$ on this interval Then,*

$$\frac{\partial\theta_{t,b}}{\partial b} = 1/f''(t,\theta_{t,b})$$

Proof

We have $f'(t,\theta_{t,b}) = b$ Thus, differentiating both sides of this equation w r t b gives,

$$f''(t,\theta_{t,b})\frac{\partial\theta_{t,b}}{\partial b} = 1$$

and the result follows

Lemma 10 *Let $f(t,\theta)$ be defined as in Definition 9 and $\theta_{t,b}$ be defined as in Definition 7 Then,*

$$\frac{\partial\theta_{t,b}}{\partial t} = -\frac{f'(t,\theta_{t,b})}{f''(t,\theta_{t,b})}$$

Proof

we prove this as follows,

$$\frac{\partial^2(f^*)(t,b)}{\partial t\partial b} = \frac{\partial\theta_{t,b}}{\partial t}$$

Hence under the assumption,

$$\frac{\partial^2(f^*)(t,b)}{\partial b\partial t} = \frac{\partial\theta_{t,b}}{\partial t}$$

But

$$f_t^*(b) = b\theta_{t,b} - f(t,\theta_{t,b})$$

hence,

$$\begin{aligned}
\frac{\partial(f^*)(t,b)}{\partial t} &= b\frac{\partial\theta_{t,b}}{\partial t} - f'(t,\theta_{t,b})\frac{\partial\theta_{t,b}}{\partial t} - f(t,\theta_{t,b}) \\
&= b\frac{\partial\theta_{t,b}}{\partial t} - b\frac{\partial\theta_{t,b}}{\partial t} - f(t,\theta_{t,b}) \\
&= -f(t,\theta_{t,b})
\end{aligned}$$

Thus

$$\begin{aligned}\frac{\partial^2(f^*)(t,b)}{\partial b\partial t} &= -\frac{\partial f(t,\theta_{t,b})}{\partial b} \\ &= -f'(t,\theta_{t,b})\frac{\partial\theta t, b}{\partial b}\end{aligned}$$

Which by Lemma 9 gives,

$$\frac{\partial^2(f^*)(t,b)}{\partial b\partial t} = -\frac{f'(t,\theta_{t,b})}{f''(t,\theta_{t,b})}$$

Hence we have the result

6.3.3 Theorems

Recall that the Shape function is defined in terms of the Legendre Fenchel transform of the cumulant generating function of the workload process by [20],

$$I(b) = \inf_{t>0} t\lambda_t^*(b/t)$$

Which can be rewritten as,

$$I(b) = \inf_{t>0} f_t^*(b)$$

Where f is defined as in definition 9 We will now prove three new results for the Shape function for the case where the workload process has non-negatively associated and stationary increments These follow very simply from the definition of the Legendre Fenchel transform, the definition of the Shape function, the Sub-additivity Theorem and a simple consequence of the Sub-additivity Theorem We prove that in the case of non-negatively associated and stationary workload increments the Shape function will be sub-additive and as a consequence a certain limit exists and further that the shape function cannot be convex on any interval which contains the origin

We do not use these results again, but they are included here because they follow simply from definitions and are quite general

It was proved in [20] that for workload processes with non-negatively associated and stationary increments $\nu = -\lim_{t\to\infty} t\lambda(t,\theta)$ is non-negative, to this we now add,

Theorem 4 *Let the increments of the workload processes* W_t^L *be non-negatively associated and stationary Let* $f(t,\theta)$ *be defined as in Definition 9 Let* $f_t^*(b)$ *be its Legendre transform Then* $I(b)$ *is sub-additive i e the following conclusion holds*

$$I(b_1 + b_2) \leq I(b_1) + I(b_2)$$

for all $b_1, b_2 \geq 0$

Proof

Firstly

$$f(t_1 + t_2, \theta) \geq f(t_1, \theta) + f(t_2, \theta) \text{ for all } t_1, t_2 \geq 0$$

thus,

$$\begin{aligned} (b_1 + b_2)\theta - f(t_1 + t_2, \theta) &\leq (b_1 + b_2)\theta - f(t_1, \theta) - f(t_2, \theta) \\ &= b_1\theta - f(t_1, \theta) + b_2\theta - f(t_2, \theta) \quad \text{for all } b, \theta > 0 \end{aligned}$$

By Lemma 5,

$$\sup_{\theta > 0}\{b\theta - f(t, \theta)\} = \sup_{\theta}\{b\theta - f(t, \theta)\}$$

thus,

$$\begin{aligned} \sup_{\theta}\{(b_1 + b_2)\theta - f(t_1 + t_2, \theta)\} &\leq \sup_{\theta}\{b_1\theta - f(t_1, \theta) + b_2\theta - f(t_2, \theta)\} \\ &\leq \sup_{\theta}\{b_1\theta - f(t_1, \theta)\} + \sup_{\theta}\{b_2\theta - f(t_2, \theta)\} \end{aligned}$$

and hence,

$$f_{t_1+t_2}^*(b_1 + b_2) \leq f_{t_1}^*(b_1) + f_{t_2}^*(b_2)$$

Then, by Lemma 3

$$\begin{aligned} \inf_{t>0} f_t^*(b_1 + b_2) &= \inf_{\{t_1>0, t_2>0\}} f_{t_1+t_2}^*(b_1 + b_2) \\ &\leq \inf_{\{t_1, t_2>0\}} (f_{t_1}^*(b_1) + f_{t_2}^*(b_2)) \\ &= \inf_{t_1>0} f_{t_1}^*(b_1) + \inf_{t_2>0} f_{t_2}^*(b_1) \end{aligned}$$

with the last equality due to Lemma 3, that is $f_t^*(b) \geq 0$ Thus,

$$I(b_1 + b_2) \leq I(b_1) + I(b_2)$$

This result tells us something about the shape of the Shape Function as Theorem 7 will show. Note that we have equality here for stationary and non-associated arrivals since in that case

$$f(t_1 + t_2, \theta) = f(t_1, \theta) + f(t_2, \theta)$$

But,

$$I(b_1 + b_2) = I(b_1) + I(b_2)$$

Implies,

$$I(b) = bI(1)$$

by a result due to Cauchy. Further

$$f(t_1 + t_2, \theta) = f(t_1, \theta) + f(t_2, \theta)$$

the same result of Cauchy also implies,

$$\begin{aligned} f(t, \theta) &= tf(1, \theta) \\ &= t\lambda_1(\theta) \end{aligned}$$

which implies,

$$\begin{aligned} \lambda_t(\theta) &= \lambda_1(\theta) \\ &= \lambda(\theta) \end{aligned}$$

which in turn implies,

$$\begin{aligned} I(1) &= \inf_{t>0} \lambda^*(1/t) \\ &= \gamma \end{aligned}$$

Thus we can say,

Theorem 5 *Let the increments of the workload processes W_t^L over disjoint time intervals be stationary and non-associated. Then the the following conclusion holds*

$$\begin{aligned} I(b) &= bI(1) \\ &= b\gamma \end{aligned}$$

Theorem 6 *Let the increments of the workload processes W_t^L over disjoint time intervals be stationary and non-negatively associated Then the the following conclusion holds*

$$\lim_{b\to\infty} \frac{I(b)}{b} \text{ exists}$$

and,

$$\lim_{b\to\infty} \frac{I(b)}{b} = \inf_{b>0} \tfrac{I(b)}{b}$$

Proof

By Theorem 4

$$I(b_1 + b_2) \leq I(b_1) + I(b_2)$$

for all $b_1, b_2 \geq 0$ Then the existence of the limit and its equality with the infimum follows from the sub-additivity theorem (Lemma 6 1 11 of [21])

Theorem 7 *Let the increments of the workload processes W_t^L over disjoint time intervals be non-negatively associated and stationary Let $I(b)$ be defined as in Definition 10 Let $I(0) = 0$ Let K be an interval on the real line containing zero Then*

$$I(b) \text{ cannot be convex on } K$$

Proof of Theorem

By Theorem 4 ,

$$I(b_1 + b_2) \leq I(b_1) + I(b_2)$$

for all $b_1, b_2 \geq 0$, with equality if either b_1 or b_2 are zero Hence,

$$I(b_1 + b_2) - I(b_1) \leq I(b_2)$$

for all $b_1, b_2 \geq 0$ Hence, since $I(0) = 0$ we have,

$$I(b_1 + b_2) - I(b_1) \leq I(b_2) - I(0)$$

for all $b_1, b_2 \geq 0$ Thus,

$$\frac{I(b_1 + b_2) - I(b_1)}{b_2} \leq \frac{I(b_2) - I(0)}{b_2}$$

for all $b_1 \geq 0$ and $b_2 > 0$. Thus,

$$\lim_{b_2 \to 0} \frac{I(b_1 + b_2) - I(b_1)}{b_2} \leq \lim_{b_2 \to 0} \frac{I(b_2) - I(0)}{b_2}$$

for all $b_1 \geq 0$ and hence,

$$I'(b_1) \leq I'(0)$$

for all $b_1 \geq 0$.

Hence,

$$I(b) \text{is not convex on } K.$$

This concludes this section devoted to more general discussion of the shape function. In the next section we will return to the shape function for the cell level model.

6.4 The Shape function for the cell level model

We intend, for reasons outlined in Subsection 6.4.1, to look at the behaviour of the cell level model as we scale the number of sources L and the packetization period s while keeping the traffic characteristics constant. In order to do this we rescale the time scale on which the multiplexer operates so that it operates on a scale that is proportional to L. We calculate the cumulant generating function for the time rescaled workload process and from this we plot the Shape function. We then use the Shape function to plot a graph of an approximation to $\log \mathbf{P}[q^L \geq b]$ against b for finite L.

6.4.1 Time rescaling

In applications s, the packetization period, is expected to be very large as the transmission rate of the multiplexer is much faster than the the sampling rate of the individual sources during periods of speech activity. The number of multiplexed sources L will also be large. Thus we are interested in the behaviour of the cell level model for very large s and L. Again we have $s = \sigma L$ for some fixed σ, we have $1 - \beta = \tau / L$ for some fixed τ and we have α independent of L. Then the mean burst period length is,

$$\frac{s}{1-\alpha} = \frac{L\sigma}{1-\alpha}$$

measured in units of the multiplexer transmission period The mean silence length is

$$\frac{1}{1-\beta} = \frac{L}{\tau}$$

also measured in units of the multiplexer transmission period Both the mean burst period length and the mean silence length are thus invariant (independent of L) on a time scale which is proportional to L [22] Rescaling the time in this manner is equivalent to reducing the multiplexer transmission period or scaling the service rate proportional to L If we double the size of the superposition the server operates twice as fast As before the offered load is independent of L The characteristics of the arrivals from each individual source are kept constant modulo discretisation w r t L as is the offered load but the service rate or transmission capacity increases proportional to L For example, in the simulations of Corcoran [9] the actual mean burst period length is maintained at 352ms, the actual mean silence length then depends on the offered load only, and with an offered load of 0 82 the mean silence length is roughly invariant as L is scaled, varying between 712 and 739ms over the range of values of L used in the simulation

We are interested in the the behaviour of the cell level model as we scale L, keeping the traffic constant The ratio of the source sampling period to the multiplexer transmission period, the transmission capacity (server speed) and the size of the superposition are all scaled For the model this means that the single source arrivals process A_t^L is replaced by the time rescaled process A_{Lt}^L which is convergent in distribution to some process A_t as $L \to \infty$ Large L scaling limits were first investigated for modulated fluid processes by Weiss [24] and time rescaled renewal processes were studied by Sriram and Whitt [11] What does Theorem 3 mean in the case of the time rescaled arrival process? Well for a given superposition size L the single source reversed arrival process is A_t^L We define $A_0^L = 0$ The service rate is r which in the case of the cell level model is 1 The superposition of L independent copies of the arrivals generated by each source is denoted by $\sum_L A_t^L$ Then the queue length at time 0 is,

$$q^L = \sup_{t\geq 0}(\sum_L A_t^L - rt)$$

This is invariant under time rescaling where we replace t by Lt, thus,

$$q^L = \sup_{t\geq 0}(\sum_L A_{Lt}^L - rLt)$$

Hence we have,

$$\begin{aligned}
\mathbf{P}[q^L > Lb] &= \mathbf{P}[\sup_{t\geq 0}(\sum_L A_{Lt}^L - rLt) > Lb] \\
&= \mathbf{P}[\bigcup_{t\geq 0}\{\sum_L A_{Lt}^L - rLt > Lb\}]
\end{aligned}$$

For large L the probability of each event in the union becomes exponentially small in L. Hence the probability of the union is dominated by the largest probability among the events of the union. Thus,

$$\mathbf{P}[q^L > Lb] \approx \sup_{t\geq 0}\mathbf{P}[\sum_L A_{Lt}^L - rLt > Lb]$$

Now for any fixed t the single source arrival processes are mutually independent. Thus by Chernoff's theorem [4] we have for large L,

$$\mathbf{P}[\sum_L A_{Lt}^L - rLt > Lb] \approx \inf_{\theta>0} e^{-\theta Lb}\mathbf{E}[e^{\theta(A_{Lt}^L - rt)}]^L$$

Define the cumulant generating function $\lambda_t^L(\theta)$ by,

$$\lambda_t^L(\theta) = \frac{1}{t}\log \mathbf{E}[e^{\theta(A_{Lt}^L - rt)}]$$

Then we can write,

$$\mathbf{P}[\sum_L A_{Lt}^L - rLt > Lb] \approx e^{-L(t\lambda_t^L)^*(b)}$$

Thus,

$$\mathbf{P}[q^L > Lb] \approx e^{-L\inf_{t\geq 0}(t\lambda_t^L)^*(b)}$$

Now A_{Lt}^L approximates A_t for large L hence $\lambda_t^L(\theta)$ approximates $\lambda_t(\theta) = \frac{1}{t}\log \mathbf{E}[e^{\theta(A_t - rt)}]$ for large L. This is made rigourous in Theorem 3 by requiring that,

$$\begin{aligned}
\lambda_t(\theta) &= \lim_{L\to\infty} \lambda_t^L(\theta) \\
\lambda(\theta) &= \lim_{t\to\infty} \lambda_t(\theta)
\end{aligned}$$

This explains the basis of Theorem 3 for the case of the time rescaled process [22].

6.4.2 The Time Rescaled Cell level Model

The limiting reversed arrival process A_t for a single source has bursts of periodic arrivals separated by a fixed period σ The number of arrivals in a burst is geometrically distributed with mean $\frac{1}{1-\alpha}$ Bursts are separated by exponentially distributed silences with mean length $\frac{1}{\tau}$ The arrival process A_t is a function of the continuous time Markov process X_t The process X_t has state space $E = [0,\sigma) \times \{\sigma\}$ The process X_t moves determimstically at unit rate from σ to 0 From 0 it jumps to σ, from where with probability α it moves as before to 0 Alternatively upon reaching σ from 0 it can with probability $1-\alpha$ remain at σ for an exponentially distributed time with mean $\frac{1}{\tau}$ The arrival process A_t is incremented by one arrival each time X_t passes through the state 0 [22] Thus we write,

$$X_t = \min\{\sigma, \text{ time to next arrival}\}$$

We will define $\rho = \frac{1}{\sigma + \frac{1-\alpha}{\tau}}$

The kernel for the limiting rescaled Markov process X_t is,

$$\mathbf{P}_t(x,dy) = \begin{cases} (1-\alpha)e^{-\tau(t-x)}\delta_\sigma(dy) + (1-\alpha)\tau e^{-\tau(t-x-\sigma+y)}dy + \alpha\delta_{\sigma-t+x}(dy) & 0 \le x < t \\ \delta_{x-t}(dy) & \sigma > x \ge t \\ e^{-\tau t}\delta_\sigma(dy) + \tau e^{-\tau(t-\sigma+y)}dy & x = \sigma \end{cases}$$

This has stationary measure,

$$\mathbf{Q}(dx) = \rho(dx + \frac{(1-\alpha)}{\tau}\delta_\sigma(dx))$$

Where dx is the Lebesgue measure on $[0,\sigma)$ and δ_σ is the unit measure at $\{\sigma\}$ This is the unique distribution on $[0,\sigma]$ which governs the steady state to which $\{X_t\}$ tends That this is the stationary distribution is verified by the following The defining equation for the stationary measure is,

$$\int_0^\sigma \mathbf{P}_t(x,dy)\mathbf{Q}(dx) = \mathbf{Q}(dy) \qquad (6\ 15)$$

We will write $U_{\{a,b\}}(dx)$ for the uniform measure on $[a,b]$ with density 1 i e $U(A) = \int_A 1dx$ for $A \subset [a,b]$ Putting our expression for $\mathbf{Q}(dx)$ (up to a multiplicative constant) into this equation 6 15 we have,

$$\begin{aligned}\frac{1}{\rho}\int_x \mathbf{P}_t(x,dy)\mathbf{Q}(dx) &= \int_x \mathbf{P}_t(x,dy)(dx+\frac{(1-\alpha)}{\tau}\delta_\sigma(dx)\\ &= \int_0^\sigma \mathbf{P}_t(x,dy)\,dx + \frac{(1-\alpha)}{\tau}\int_x \mathbf{P}_t(x,dy)\delta_\sigma(dx))\end{aligned}$$

The first term here is,

$$\begin{aligned}\int_0^\sigma \mathbf{P}_t(x,dy)\,dx &= \int_0^t ((1-\alpha)e^{-\tau(t-x)}\delta_\sigma(dy) + (1-\alpha)\tau e^{-\tau(t-x-\sigma+y)}dy + \alpha\delta_{\sigma-t+x}(dy))\,dx\\ &\quad + \int_t^\sigma \delta_{x-t}(dy)\,dx\\ &= \frac{(1-\alpha)}{\tau}e^{-\tau(t-x)}\delta_\sigma(dy)|_{x=0}^t + (1-\alpha)e^{-\tau(t-x-\sigma+y)}dy|_{x=0}^{y-\sigma+t} + \alpha U_{\{\sigma-t,\sigma\}}(dy)\\ &\quad + U_{\{0,\sigma-t\}}(dy)\\ &= \frac{(1-\alpha)}{\tau}\delta_\sigma(dy) - \frac{(1-\alpha)}{\tau}e^{-\tau t}\delta_\sigma(dy) + (1-\alpha)U_{\{\sigma-t,\sigma\}}(dy)\\ &\quad -(1-\alpha)e^{-\tau(t-\sigma+y)}dy\\ &\quad +\alpha U_{\{\sigma-t,\sigma\}}(dy) + U_{\{0,\sigma-t\}}(dy)\end{aligned}$$

The second term is,

$$\frac{(1-\alpha)}{\tau}\int_x \mathbf{P}_t(x,dy)\delta_\sigma(dx)) = \frac{(1-\alpha)}{\tau}e^{-\tau t}\delta_\sigma(dy) + (1-\alpha)e^{-\tau(t-\sigma+y)}dy$$

Combining the two we get,

$$\begin{aligned}\frac{1}{\rho}\int_0^\sigma \mathbf{P}_t(x,dy)\mathbf{Q}(dx) &= \int_0^\sigma \mathbf{P}_t(x,dy)\,dx + \frac{(1-\alpha)}{\tau}\int_x \mathbf{P}_t(x,dy)\delta_\sigma(dx)\\ &= \frac{(1-\alpha)}{\tau}\delta_\sigma(dy) + (1-\alpha)U_{\{\sigma-t,\sigma\}}(dy) + U_{\{0,\sigma-t\}}(dy) + \alpha U_{\{\sigma-t,\sigma\}}(dy)\\ &= dy + \frac{1-\alpha}{\tau}\delta_\sigma(dy)\\ &= \frac{1}{\rho}\mathbf{Q}(dy)\end{aligned}$$

Thus,

$$\int_0^\sigma \mathbf{P}_t(x,dy)\mathbf{Q}(dx) = \mathbf{Q}(dy)$$

as required

6.4.3 Calculating The Cumulant Generating Function

In fact the arrival epochs of the process A_t form a delayed renewal process In other words, let T_0 be the time of the first arrival (renewal) and let $\{S_n\}_{n\geq 1}$ be the times between successive arrivals, let $\{T_n\}_{n\geq 1}$ be the times of successive arrivals Then,

$$T_n = T_0 + \sum_{i=1}^{n} S_i$$

and $\{T_n - T_0 = \sum_{i=1}^{n} S_i\}_{n\geq 1}$ is a Renewal process Renewal theory which we apply here is described in detail in [25] We will give the label G to the distribution function of the time of the first arrival T_0 And we will give the label F to the common distribution of $\{S_n\}_{n\geq 1}$ Then F will be given by,

$$F(dt) = \alpha\delta_\sigma(dt) + (1-\alpha)\int_0^\infty \tau e^{-\tau y}\delta_{\sigma+y}(dt)dy$$

We do not immediately have an expression for G but we can derive it from the distribution of T_0 conditioned on the initial state of the underlying Markov Process X For this distribution which we will call G_x we have,

$$G_x(dt) = \delta_x(dt)\delta_x((0,\sigma)) + \delta_x(\{\sigma\})\int_0^\infty \tau e^{-\tau y}\delta_{\sigma+y}(dt)dy$$

From this we can derive G as follows Recall that the stationary distribution of the Markov process X is,

$$Q(dx) = \rho dx + \rho\frac{(1-\alpha)}{\tau}\delta_\sigma(dx)$$

Thus the distribution of T_0 is,

$$G(dt) = \int_0^\infty G_x(dt)Q(dx)$$

Thus,

$$\begin{aligned} G(dt) &= \rho\int_0^\sigma G_x(dt)dx + \rho\frac{(1-\alpha)}{\tau}\int_0^\sigma G_x(dt)\delta_\sigma(dx) \\ &= \rho\delta_t((0,\sigma))dt + \rho\frac{(1-\alpha)}{\tau}G_\sigma(dt) \end{aligned}$$

We can now derive an expression for the distribution function G as follows,

$$
\begin{aligned}
G(t) &= \int_0^t G(dy) \\
&= \rho \int_0^t \delta_y((0,\sigma))dy + \rho\frac{(1-\alpha)}{\tau}\int_0^t G_\sigma(dy) \\
&= \rho t\delta_t((0,\sigma)) + \rho\sigma\delta_t([\sigma,\infty)) + \rho\frac{(1-\alpha)}{\tau}\int_0^\infty \tau e^{-\tau x}\int_0^t \delta_{\sigma+x}(dy)dx \\
&= \rho t\delta_t((0,\sigma)) + \rho\sigma\delta_t([\sigma,\infty)) + \rho\frac{(1-\alpha)}{\tau}\int_0^\infty \tau e^{-\tau x}\delta_x((0,t-\sigma))\delta_t((\sigma,\infty))dx \\
&= \rho t\delta_t((0,\sigma)) + \rho\sigma\delta_t([\sigma,\infty)) + \rho\frac{(1-\alpha)}{\tau}(1-e^{-\tau(t-\sigma)})\delta_t((\sigma,\infty))
\end{aligned}
$$

Thus, we now have expressions for both G, the distribution of the time T_0 to the first renewal, and F, the common distribution of the inter-renewal times $\{S_n\}_{n\geq 1}$ These are all we need to derive an expression for the cumulant generating function of the workload process W_t

Before that, we can prove that the arrival process A_t has stationary increments and constant renewal rate ρ To do this, we first calculate the mean inter-arrival time as follows,

$$
\begin{aligned}
\mathbf{E}[S_n] &= \int_0^\infty yF(dy) \\
&= \alpha\int_0^\infty y\delta_\sigma(dy) + (1-\alpha)\int_0^\infty\int_0^\infty \tau e^{-\tau t}\delta_{\sigma+t}(dy)dt \\
&= \alpha\sigma + (1-\alpha)\int_0^\infty \tau e^{-\tau t}(\sigma+t)dt \\
&= \alpha\sigma + (1-\alpha)\sigma + \frac{(1-\alpha)}{\tau} \\
&= \sigma + \frac{(1-\alpha)}{\tau} \\
&= \frac{1}{\rho}
\end{aligned}
$$

And we know $\frac{1}{\rho} < \infty$

We want an expression for $\mathbf{E}[A_t]$ We can write, A_t in terms of the renewal epochs $\{T_n\}_{n\geq 0}$ as follows,

$$A_t = \sum_{n=0}^{\infty} I_{[0,t]}(T_n)$$

and hence, we have the following expression for $\mathbf{E}[A_t]$ in terms of G and F,

$$\mathbf{E}[A_t] = \sum_{n=0}^{\infty}\mathbf{P}[T_n \leq t]$$

$$= \sum_{n=0}^{\infty}(G \star F^n(t))$$

But this last expression can be rewritten as,

$$\begin{aligned}
\sum_{n=0}^{\infty}(G \star F^n(t)) &= G + \sum_{n=1}^{\infty}(G \star F^n(t)) \\
&= G + (\sum_{n=1}^{\infty} F^n(t)) \star G \\
&= G + F \star (\sum_{n=0}^{\infty} F^n(t)) \star G \\
&= G + F \star \sum_{n=0}^{\infty}(G \star F^n(t)) \\
&= G + F \star \mathbf{E}[A_t]
\end{aligned}$$

That is, $\mathbf{E}[A_t]$ satisfies, the renewal equation,

$$\mathbf{E}[A_t] = G + F \star \mathbf{E}[A_t]$$

or, on rearranging, we have for G,

$$G = \mathbf{E}[A_t] - F \star \mathbf{E}[A_t]$$

Thus we have $\mathbf{E}[A_t] = \rho t$ iff,

$$G = \rho t - F \star \rho t$$

And in fact this is easily proved as follows,

$$\begin{aligned}
F \star \rho t &= \rho \int_0^t (t-x)F(dx) \\
&= \rho \int_0^t (t-x)\alpha\delta_\sigma(dx) + \rho \int_0^t (t-x)(1-\alpha)\int_0^\infty \tau e^{-\tau y}\delta_{\sigma+y}(dx)dy \\
&= \delta_t((\sigma,\infty))(\rho(t-\sigma)\alpha + \rho(1-\alpha)\int_0^\infty (t-\sigma-y)\tau e^{-\tau y}\delta_y((0,t-\sigma))dy) \\
&= \rho(t-\sigma)\delta_t((\sigma,\infty)) - \rho\delta_t((\sigma,\infty))\frac{(1-\alpha)}{\tau} + \rho\delta_t((\sigma,\infty))\frac{(1-\alpha)}{\tau}e^{-\tau(t-\sigma)}
\end{aligned}$$

Thus we have,

$$\begin{aligned}
\rho t - F \star \rho t &= \rho t - \rho((t-\sigma) - \frac{(1-\alpha)}{\tau}(1-e^{-\tau(t-\sigma)}))\delta_t((\sigma,\infty)) \\
&= \begin{cases} \rho t & \text{for } t \le \sigma \\ \rho\sigma + \rho\frac{(1-\alpha)}{\tau}(1-e^{-\tau(t-\sigma)}) & \text{for } t > \sigma \end{cases} \\
&= G(t)
\end{aligned}$$

Thus we have proved,

$$\mathbf{E}[A_t] \;=\; \rho t \quad \text{for all } t \geq 0 \tag{6 16}$$

Hence A_t has stationary increments and the renewal rate is constant

We return to calculating the cumulant generating function for the workload process In order to calculate the cumulant generating function we need the distribution of the arrivals process We can determine this distribution from the distributions of the renewal epochs First we will write $\{A_t = n\}$ in terms of renewal epoch events

$$\begin{aligned}
\{A_t = n\} &= \{T_{n-1} \leq t, T_n > t\} \\
&= \{T_{n-1} \leq t\} \bigcap \{T_n \leq t\}^c
\end{aligned}$$

But,

$$\{T_n \leq t\} \subset \{T_{n-1} \leq t\}$$

Hence the distribution of the arrivals process is, for $n \geq 1$

$$\begin{aligned}
\mathbf{P}[A_t = n] &= \mathbf{P}[T_{n-1} \leq t] - \mathbf{P}[T_n \leq t] \\
&= G \star F^{n-1}(t) - G \star F^n(t) \\
&= F^{n-1} \star \rho t - 2F^n \star \rho t + F^{n+1} \star \rho t
\end{aligned}$$

and for $n = 0$,

$$\begin{aligned}
\mathbf{P}[A_t = 0] &= \mathbf{P}[T_0 > t] \\
&= 1 - \mathbf{P}[T_0 \leq t] \\
&= 1 - G(t) \\
&= 1 - \rho t + F \star \rho t
\end{aligned}$$

These equations allow us to write the following expression for $\mathbf{E}[e^{\theta A_t}]$

$$\mathbf{E}[e^{\theta A_t}] \;=\; 1 - \rho t + F \star \rho t + \sum_{n=1}^{[t/\sigma]_-} (F^{n-1} \star \rho t - 2F^n \star \rho t + F^{n+1} \star \rho t) e^{\theta n}$$

Where $F^0 \star \rho t$ is defined to be ρt and $[t/\sigma]_- = \max\{m \in \mathbf{Z}^+ | t/\sigma > m\}$

We have, then, the following expression for the cumulant generating function for the time rescaled workload process for the cell level model, with deterministic service rate of one cell per unit time

$$\lambda_t(\theta) = \frac{1}{t}\log[1-\rho t+F\star\rho t+\sum_{n=1}^{[t/\sigma]_-}(F^{n-1}\star\rho t-2F^n\star\rho t+F^{n+1}\star\rho t)e^{\theta n}]-\theta$$

It is then a simple matter to prove that,

$$\lambda_t(\theta) = \frac{1}{t}\log[1+\rho t(e^\theta-1)+\sum_{n=1}^{[t/\sigma]_-}F^n\star\rho t e^{\theta(n-1)}(e^\theta-1)^2]-\theta \qquad (6\ 17)$$

Now we need only calculate $F^n(dx)$ and the convolution with ρt in order to find $\lambda_t(\theta)$ We use Laplace transforms to derive an expression for $F^n(dx)$ We will use the notation $\mathbf{L}[\ ]$ for the Laplace transform

First we find the Laplace transform of the distribution F

$$\begin{aligned}
\mathbf{L}[F] &= \int_0^\infty e^{-\beta x}F(dx)\\
&= \int_0^\infty e^{-\beta x}(\alpha\delta_\sigma(dx)+(1-\alpha)\int_0^\infty \tau e^{-\tau y}\delta_{\sigma+y}(dx)dy\\
&= \alpha\int_0^\infty e^{-\beta x}\delta_\sigma(dx)+(1-\alpha)\int_0^\infty\int_0^\infty e^{-\beta x}\tau e^{-\tau y}\delta_{\sigma+y}(dx)dy
\end{aligned}$$

The first term is,

$$\alpha\int_0^\infty e^{-\beta x}\delta_\sigma(dx) = \alpha e^{-\beta\sigma}$$

The second term is,

$$\begin{aligned}
(1-\alpha)\int_0^\infty\int_0^\infty e^{-\beta x}\tau e^{-\tau y}\delta_{\sigma+y}(dx)dy &= (1-\alpha)\int_0^\infty \tau e^{-\beta(\sigma+y)}e^{-\tau y}dy\\
&= (1-\alpha)e^{-\beta\sigma}\tau\int_0^\infty e^{-(\tau+\beta)y}dy\\
&= (1-\alpha)e^{-\beta\sigma}(\frac{\tau}{\beta+\tau})
\end{aligned}$$

The two combined are,

$$\mathbf{L}[F] = (\alpha+(1-\alpha)(\frac{\tau}{\beta+\tau}))e^{-\beta\sigma}$$

We now use the fact that the Laplace transform of a convolution is the product of the Laplace transforms, that is,

$$\mathbf{L}[F^n] = (\mathbf{L}[F])^n$$

Hence,

$$\mathbf{L}[F^n] = (\alpha + (1-\alpha)(\frac{\tau}{\beta+\tau}))^n e^{-n\beta\sigma}$$

Then expanding binomially we get,

$$\mathbf{L}[F^n] = \sum_{r=0}^{n} \binom{n}{r} \alpha^{n-r}(1-\alpha)^r (\frac{\tau}{\beta+\tau})^r e^{-n\beta\sigma}$$

We can rewrite this as,

$$\mathbf{L}[F^n] = \alpha^n e^{-n\beta\sigma} + \sum_{r=1}^{n} \binom{n}{r} \alpha^{n-r}(1-\alpha)^r (\frac{\tau}{\beta+\tau})^r e^{-n\beta\sigma}$$

Now we need only take the inverse transform of this to find F^n We use the fact that the inverse transform of the sum is the sum of the inverse transforms And we again use the fact that the transform of a convolution is the product of the Laplace transforms to deduce that, the inverse transform of $(\frac{\tau}{\beta+\tau})^r e^{-n\beta\sigma}$ is the convolution of the inverse transform of $(\frac{\tau}{\beta+\tau})^r$ with the inverse transform of $e^{-n\beta\sigma}$ The inverse transform of $(\frac{\tau}{\beta+\tau})^r$ is itself just the r-fold convolution of the inverse transform of $(\frac{\tau}{\beta+\tau})$ with itself The inverse transform of $(\frac{\tau}{\beta+\tau})$ has the density $\tau e^{-\tau x}$ and the r-fold convolution of this with itself has the gamma density $\tau\frac{(\tau x)^{r-1}}{(r-1)!}e^{-\tau x}$ The inverse Laplace transform of $e^{-n\beta\sigma}$ is the measure $\delta_{n\sigma}(dx)$ The convolution of this with $\tau\frac{(\tau x)^{r-1}}{(r-1)!}e^{-\tau x}$ is,

$$\int_0^x \tau\frac{(\tau(x-y))^{r-1}}{(r-1)!}e^{-\tau(x-y)}\delta_{n\sigma}(dy) = \begin{cases} \tau\frac{(\tau(x-n\sigma))^{r-1}}{(r-1)!}e^{-\tau(x-n\sigma)} & \text{for } x \geq n\sigma \\ 0 & \text{otherwise} \end{cases}$$

Thus we have the following for the distribution F^n,

$$F^n(dx) = \delta_x((n\sigma,\infty))\sum_{r=1}^{n}\binom{n}{r}\alpha^{n-r}(1-\alpha)^r\tau\frac{(\tau(x-n\sigma))^{r-1}}{(r-1)!}e^{-\tau(x-n\sigma)}dx + \alpha^n\delta_{n\sigma}(dx)$$

In order to calculate $\lambda_t(\theta)$ we need $F^n \star \rho t$ The convolution with a sum is just the sum of the convolutions, hence we need to get

$$\int_0^t \delta_x((n\sigma,\infty))\tau\frac{(\tau(x-n\sigma))^{r-1}}{(r-1)!}e^{-\tau(x-n\sigma)}\rho(t-x)dx$$

This is,

$$\delta_t((n\sigma,\infty))(\int_{n\sigma}^t \tau\frac{(\tau(x-n\sigma))^{r-1}}{(r-1)!}e^{-\tau(x-n\sigma)}\rho(t-x)dx)$$

Which becomes after integration by parts, and some rearrangement,

$$\delta_t((n\sigma,\infty))\frac{\rho}{\tau}e^{-\tau(t-n\sigma)}\sum_{l=r+1}^{\infty}\frac{(\tau(t-n\sigma))^l}{l!}(l-r)$$

We also need,

$$\int_0^t \rho(t-x)\alpha^n\delta_{n\sigma}(dx)$$

Which is,

$$\delta_t((n\sigma,\infty))\alpha^n\rho(t-n\sigma)$$

Putting the convolutions back into the sum we get,

$$F^n\star\rho t = \delta_t((n\sigma,\infty))(\sum_{r=1}^{n}\binom{n}{r}\alpha^{n-r}(1-\alpha)^r\frac{\rho}{\tau}e^{-\tau(t-n\sigma)}\sum_{l=r+1}^{\infty}\frac{(\tau(t-n\sigma))^l}{l!}(l-r)+\alpha^n\rho(t-n\sigma))$$

Which can be rewritten as,

$$F^n\star\rho t = \delta_t((n\sigma,\infty))(\sum_{r=0}^{n}\binom{n}{r}\alpha^{n-r}(1-\alpha)^r\frac{\rho}{\tau}e^{-\tau(t-n\sigma)}\sum_{l=r+1}^{\infty}\frac{(\tau(t-n\sigma))^l}{l!}(l-r))$$

We will write,

$$b(n,r,\alpha) = \frac{n!}{r!(n-r)!}\alpha^{n-r}(1-\alpha)^r$$

Now we have the following expression for $\lambda_t(\theta)$ the cumulant generating function of the time rescaled workload process, $t\lambda_t(\theta)$ is,

$$\log\left[1+\rho t(e^\theta-1)+\frac{\rho}{\tau}\sum_{n=1}^{[t/\sigma]_-}\sum_{r=0}^{n}b(n,r,\alpha)e^{-\tau(t-n\sigma)}\sum_{l=r+1}^{\infty}\frac{(\tau(t-n\sigma))^l}{l!}(l-r)e^{\theta(n-1)}(e^\theta-1)^2\right]-t\theta$$

6.4.4 The Shape Function For The Time Rescaled Cell Level Model

We begin by using Equation 6 5 [20] to find $I(0)$ for the cell level model For $t\leq\sigma$ we have,

$$t\lambda_t(\theta) = \log[1+\rho t(e^\theta-1)]-\theta t$$

Thus,

$$\begin{aligned}\Lambda_t(\theta) &= t\lambda_t(\theta/t)\\ &= \log[1+\rho t(e^{\theta/t}-1)]-\theta\end{aligned}$$

Thus we have,

$$\begin{aligned}\Lambda(\theta) &= \lim_{t\to 0}\Lambda_t(\theta)\\ &= \begin{cases}-\theta & \text{if } \theta\le 0\\ +\infty & \text{if } \theta>0\end{cases}\end{aligned}$$

Hence we have, the following for $I(0)$

$$\begin{aligned}I(0) &= \Lambda^*(0)\\ &= \sup_\theta\{-\Lambda(\theta)\}\\ &= \sup_{\theta\le 0}\{\theta\}\\ &= 0\end{aligned}$$

Next we will calculate the rate function $f_t^*(b)$ for $t\le\sigma$ For $t\le\sigma$ we have,

$$f(t,\theta) = \log[1+\rho t(e^\theta-1)]-\theta t$$

Recall,

$$f_t^*(b) = \sup_\theta\{b\theta - f(t,\theta)\}$$

Now $f(t,\theta)$ is convex as a function of θ, hence $b\theta - f(t,\theta)$ is concave as a function of θ Thus the supremum occurs for $\theta=\theta_{t,b}$ such that,

$$f'(t,\theta_{t,b}) = b$$

For $t\le\sigma$ this gives us,

$$\frac{\rho t e^{\theta_{t,b}}}{1+\rho t(e^{\theta_{t,b}}-1)} = b+t$$

which implies,

$$\theta_{t,b} = \log[\frac{b+t}{\rho t}]-\log[\frac{1-(b+t)}{1-\rho t}]$$

Then,

$$f_t^*(b) = b\theta_{t,b} - f(t, \theta_{t,b})$$

Hence,

$$f_t^*(b) = (b+t)\log[\frac{b+t}{\rho t}] + (1-(b+t))\log[\frac{1-(b+t)}{1-\rho t}]$$

The derivative of $f_t^*(b)$ w r t t for $t \leq \sigma$ is,

$$f_t^*(b) = \log[\frac{b+t}{\rho t}] - \log[\frac{1-(b+t)}{1-\rho t}] - \frac{b}{t} - \frac{1-\rho(1-b)}{1-\rho t}$$

Thus the second derivative of $f_t^*(b)$ w r t t for $t \leq \sigma$ is,

$$f_t^*(b) = \frac{b}{t^2} - \frac{b}{bt+t^2} + \frac{(1-\rho)^2 + \rho^2 b^2 + 2\rho b(1-\rho)}{(1-(b+t))(1-\rho t)} s$$

For $b+t < 1$ this is positive Thus $f_t^*(b)$ is convex on $(0, \sigma]$ But examples also show that $f_t^*(b)$ is not convex on $(0, t_0)$ as a function of t for $t_0 > \sigma$ This is shown by Figures 6-2 and 6-3

In order to plot the shape function for a given set of parameter values we numerically determine the $\inf_{t>0} f_t^*(b)$ for a range of values of b We do this by first fixing b and $t = \sigma$ then calculating $f_t^*(b)$ for these values of b and t Then we increment t by σ and repeat the procedure until we have found $\inf_{t>0} f_t^*(b)$ for the particular value of b We then increment b and repeat the whole procedure This is carried out by a program written in C The $\inf_{t>0} f_t^*(b)$ values are then plotted against the corresponding values of b giving us the graph of the shape function for these values and the model parameter values chosen The graph of the shape function for $\sigma = 0\ 35696$, $\rho = 0\ 85$, $\alpha = 0\ 995466$ and $\tau = 0\ 00553175$ is shown in Figure 6-4 All graphs were plotted using "gnuplot" The graph of the approximation to $\log \mathbf{P}[q^L > b]$ obtained using the shape function plotted against b for these parameter values and $L = 400$ is shown in Figure 6-5

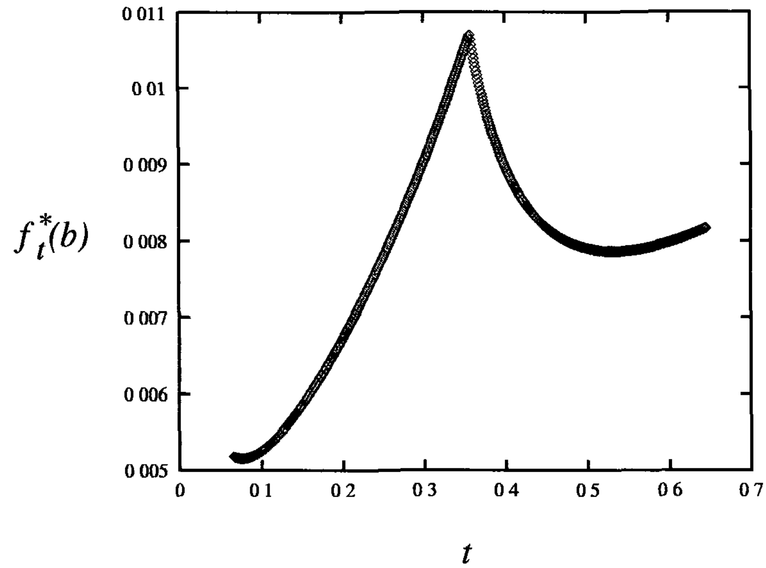

Figure 6-2 *The Legendre-Fenchel transform, $f_t^*(b)$, for $b = 0\ 015$ as a function of t*

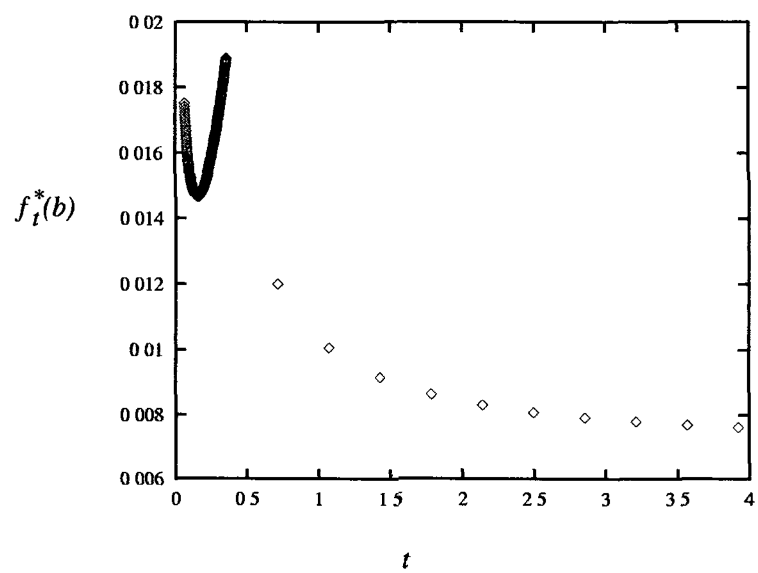

Figure 6-3 *The Legendre-Fenchel transform, $f_t^*(b)$, for $b = 0\ 038$ as a function of t*

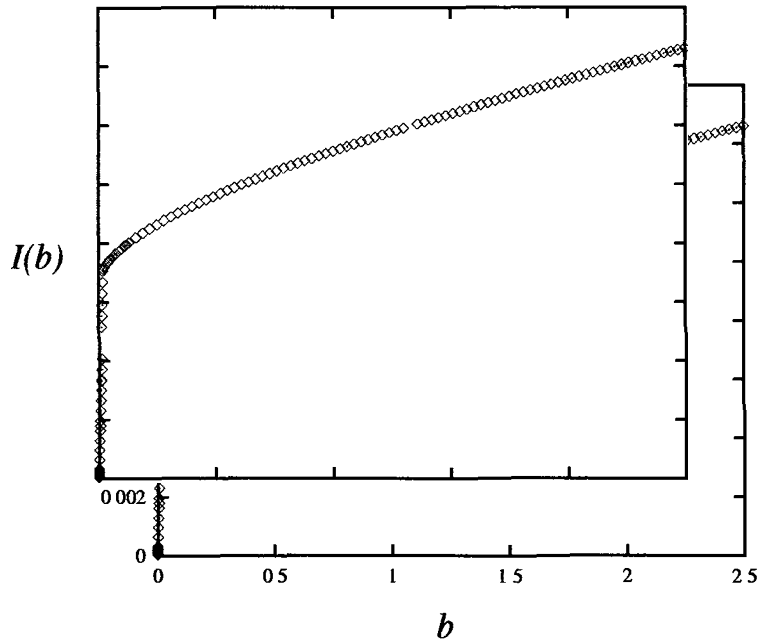

Figure 6-4 *The Shape function, $I(b)$*

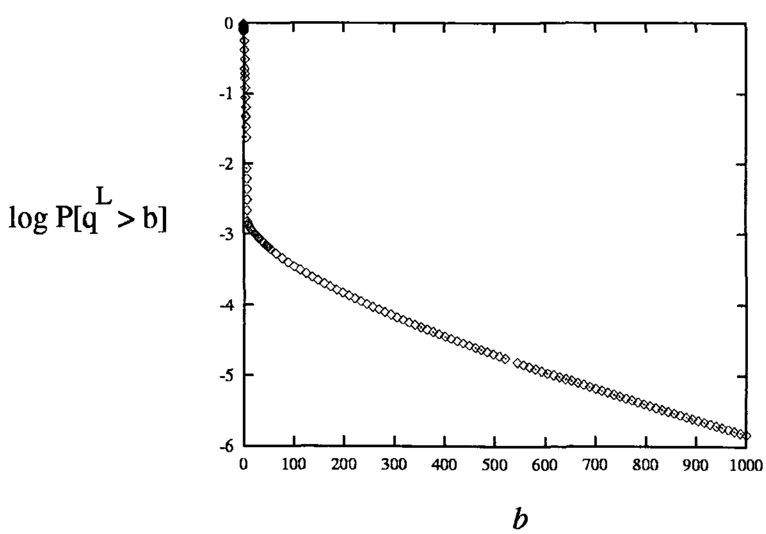

Figure 6-5 *A plot of the approximation to $\log \mathbf{P}[q^L > b]$ obtained using the Shape function against b with $L = 400$*

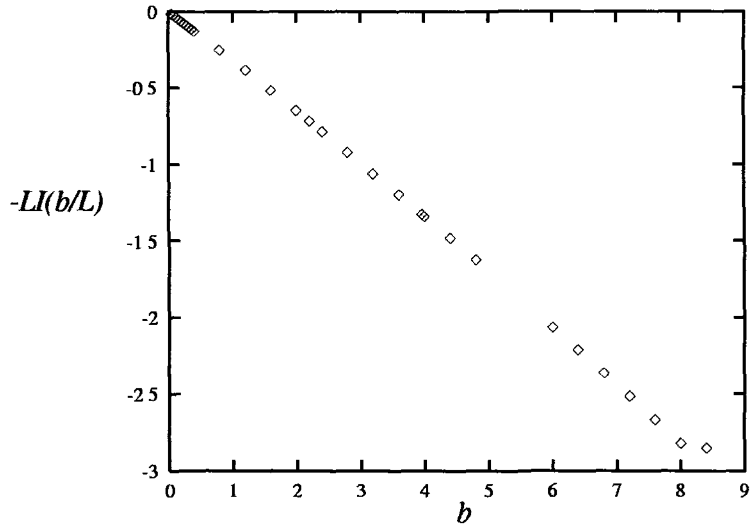

Figure 6-6 *A Plot of The steep portion of fig 6-5*

6.5 Economies of Scale

We mentioned in 6 9-6 12 that for a very large class of models the asymptotic form of the shape function I for large b is [20, 22],

$$I(b) \approx \gamma b + \nu \tag{6 18}$$

and that ν can be seen to determine the economies of scale [22] obtainable by multiplexing large numbers of sources through,

$$\mathbf{P}[q^L > b] \approx e^{-LI(b/L)} \tag{6 19}$$

$$\approx e^{-\gamma b - L\nu} \tag{6 20}$$

For Markov Models it is possible to calculate ν in the following manner [20, 22], Each source is described by a Markov Process X_t, c is the service rate and arrival increments A_t are described by a family of transition kernels $P_t(x, dy \times dz)$, i e

$$P_t(x, Y \times Z) = \mathbf{P}[X_t \in Y, A_t \in Z | X_0 = x] \tag{6 21}$$

The transformed kernel $\hat{P}_t(\theta)$ is given by,

$$\hat{P}_t(x, dy, \theta) = \int_z P_t(x, dy \times dz) e^{\theta z} \tag{6 22}$$

Then γ is the value of $\theta > 0$ for which the largest eigenvalue of $\hat{P}_1(\theta)$ is 1 The value of ν is got from,

$$e^{\nu} = \frac{\int Q(dx)v(x) \int u(dx)}{\int u(dx)v(x)} \tag{6 23}$$

where X_t the underlying Markov chain has stationary distribution $Q(dx)$ and u and v are respectively the left eigenmeasure and right eigenfunction of $\hat{P}_1(\theta)$ with eigenvalue 1 [22]

6.5.1 Calculating ν and γ for the rescaled cell level model

Firstly for the rescaled cell level model the arrival increment A_t for $t \leq \sigma$ is a deterministic function of X_t and X_0 In fact it is a deterministic function of t and X_0 Thus for $t \leq \sigma$ the kernel $P_t(x, Y \times Z)$ is,

$$P_t(x, dy \times dz) = P_t(x, dy)\delta_{h_t(x)}(dz) \tag{6 24}$$

where h is the deterministic function We have,

$$h_t(x) = \begin{cases} 1-t & \text{if } t \geq x \text{ and } x \neq 0 \\ -t & \text{if } t \leq x \\ -t & \text{if } x = \sigma \text{ or } 0 \end{cases} \tag{6 25}$$

Thus the transformed kernel is,

$$\hat{P}_t(x, dy, \theta) = \int_z P_t(x, dy)\delta_{h_t(x)}(dz)e^{\theta z} \tag{6 26}$$

$$= P_t(x, dy)e^{\theta h_t(x)} \tag{6 27}$$

and this is just,

$$\hat{P}_t(x, dy, \theta) = \begin{cases} P_t(x, dy)e^{\theta(1-t)} & \text{if } t \geq x \text{ and } x \neq 0 \\ P_t(x, dy)e^{-\theta t} & \text{if } t \leq x \\ P_t(x, dy)e^{-\theta t} & \text{if } x = \sigma \text{ or } 0 \end{cases} \tag{6 28}$$

Now the right eigenfunction v and left eigenmeasure u with eigenvalue 1 of $P_1(\theta)$ are the right eigenfunction and left eigenmeasure of $P_t(\theta)$ because convolution preserves such right eigenfunctions and left eigenmeasures thus, we need only solve,

$$\int_x u(dx)P_t(x, dy)e^{\theta h_t(x)} = u(dy) \tag{6 29}$$

to find u and,

$$\int_x v(y)P_t(x,dy)e^{\theta h_t(x)} = v(x) \tag{6 30}$$

to find v

So let $u(dx) = e^{\gamma x}dx + k_1\delta_\sigma(dx)$ and let $v(y) = (e^{-\gamma y}, k_2)$ on $[0,\sigma) \times \{\sigma\}$ Then performing the integration and equating the relevant sides we can solve for k_1 and k_2 This gives us, for u,

$$k_1 = e^\gamma \frac{1-\alpha}{\gamma+\tau} \tag{6 31}$$

and, for v,

$$k_2 = \frac{\gamma e^{-\gamma\sigma}}{\gamma+\tau} \tag{6 32}$$

and further we find, in solving for u that,

$$e^{\gamma(1-\sigma)}(\frac{1-\alpha}{\gamma+\tau}+\alpha) = 1 \tag{6 33}$$

which we recall is Equation 4 4 where $L \to \infty$ and $\gamma(L) \to \gamma$ Thus for u we have,

$$u(dx) = e^{\gamma x}dx + e^\gamma \frac{1-\alpha}{\gamma+\tau}\delta_\sigma(dx) \tag{6 34}$$

and for v we have,

$$v(y) = (e^{-\gamma y}, \frac{\gamma e^{-\gamma\sigma}}{\gamma+\tau}) \tag{6 35}$$

Thus for ν we now have,

$$e^{-\nu} = \frac{\int Q(dx)v(x)\int u(dx)}{\int u(dx)v(x)} \tag{6 36}$$

putting u and v into this equation gives us,

$$e^{-\nu} = \frac{1}{\gamma^2}\frac{e^\gamma-1}{\sigma+\frac{1-\alpha}{\tau}}\frac{(\gamma+\tau)(\alpha\gamma+\tau)}{(1-\alpha)\tau+\sigma} \tag{6 37}$$

Chapter 7

Conclusions and Suggestions for Future Study

In this chapter we give our conclusions from the work described in this thesis, and discuss future work that could be undertaken We divide this chapter according to the chapters of the thesis itself

7.1 Conclusions

7.1.1 The Models

We have studied a new model for packetized voice traffic which we have called the cell level model The model consists of the homogeneous superposition of the traffic generated by L independent sources The traffic from each source is modelled by a Markov Chain with a finite state space The states form an irreducible closed set and are recurrent non-null aperiodic (ergodic) having as a result a unique stationary distribution The only assumptions we make about the traffic from a single source is that the duration of talk spurts and in active periods are both exponentially distributed (an assumption made by others [10]) and that as a result bursts and silences are geometrically distributed The model is thus a very accurate representation of packetized voice traffic from a single source We also mention a model that has been studied previously by Buffet and Duffield [3] which is simpler than the cell level model and which is called the block level model and we point out where the connection between the two models breaks down

7.1.2 The Effective Bandwidth Approximation

We calculated approximations to the decay constant γ of the effective bandwidth approximations [2]

$$\mathbf{P}[q \geq b] \approx e^{-\gamma b}$$

for each of the cell level and block level models labeling the two resulting constants γ_{Cell} and γ_{Block} respectively We showed that,

$$\gamma_{\mathrm{Block}} > \gamma_{\mathrm{Cell}} \tag{7 1}$$

and can conclude from this that the upper bound on buffer overflow obtained from the block level effective bandwidth approximation is less conservative than that obtained from the cell level effective bandwidth approximation i e the former upper bound could underestimate the probability of buffer overflow obtained using the latter upper bound

7.1.3 An Upper Bound Via Martingales

We proved an upper bound can be obtained of the form,

$$\mathbf{P}[q \geq b] \leq \phi e^{-b\gamma}$$

on the tail of the queue length distribution for the queue in an infinite buffer When the workload process is a Markov Additive Process (MAP). The cell level model workload process has increments which are controlled by an underlying Markov Process and it is an example of a MAP We calculated the prefactor ϕ for the cell level model and showed that,

$$\phi = e^{\gamma}(e^{-\gamma\sigma}e^{2\gamma/L}\frac{1}{(\frac{L(e^{\gamma/L}-1)}{\tau}+1)})^3(\frac{(\sigma+\frac{e^{\gamma(\sigma-1-1/L)}-1}{L})(1+\frac{L(e^{\gamma/L}-1)}{\tau})+\frac{1-\alpha}{\tau}}{\sigma+\frac{1-\alpha}{\tau}})^L$$

and we showed that asymptotically in L,

$$\lim_{L\to\infty} \frac{1}{L}\log\phi = \log\left(\frac{\sigma(1+\frac{\gamma}{\tau})+\frac{1-\alpha}{\tau}}{\sigma+\frac{1-\alpha}{\tau}}\right) \tag{7.2}$$

$$> 0 \tag{7.3}$$

From which we conclude that for large L the bound,

$$\mathbf{P}[q^L \geq b] \leq \phi e^{-\gamma b} \tag{7.4}$$

does not exhibit the economies of scale seen for example in the upper bound obtained using a different Martingale for the block level model by Buffet and Duffield [3]. These economies of scale are seen in the simulations of Corcoran [9] for the rescaled cell level model. This is not surprising since this upper bound has to bound the full queue including the cell level queue which is due to short term fluctuations in the arrival rate over time periods smaller than the packetization period s. These fluctuations result in short queues and this is exhibited in the graphs of $\log \mathbf{P}[q^L > b]$ versus b by the almost straight steep portion of the graph. The slope and intercept of this part of the graph remain invariant as L is changed in the simulations carried out in [9].

However the upper bound obtained for what we term the burst level queue, the queue resulting from arrivals over longer intervals of time than s, can exhibit these economies of scale since it does not bound the cell level queue. We obtained such a bound and we proved for parameter values $\sigma = 1/2$ and $\rho \leq \alpha$ it does exhibit economies of scale and is an improvement over the effective bandwidth approximation in terms of bounding the queue length distribution of this burst level queue. Note, the condition $\rho \leq \alpha$ can be rewritten as,

$$\frac{1}{\sigma+\frac{1-\alpha}{\tau}} \leq \alpha \tag{7.5}$$

but this is the same as,

$$\tau \leq \frac{\alpha(1-\alpha)}{1-\alpha\sigma} \tag{7.6}$$

which for $\sigma = 1/2$ is,

$$\tau \leq \frac{2\alpha(1-\alpha)}{2-\alpha} \tag{7.7}$$

and since $\sigma > 0$ Equation 7.6 is true, for all σ if,

$$\tau \;<\; \alpha(1-\alpha) \tag{7.8}$$

For $\sigma = 1/2$ and $\rho \leq \alpha$ we, have

$$\mathbf{P}[q_s^L > b] < \Phi^L e^{-\gamma b} \tag{7.9}$$

where $\Phi < 1$ is independent of L. We can say that,

$$\Phi \;=\; \max\{\rho + \rho^2\sigma\frac{\frac{1}{\rho}-1}{\alpha(1-\sigma)}, \frac{(1+\frac{\gamma}{\tau})\sigma + \frac{1-\alpha}{\tau}}{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{\gamma}{\tau})^\sigma}\} \tag{7.10}$$

more generally we have,

$$\begin{aligned}
\mathbf{P}[q_s^L > b] \;&\leq\; (\frac{v(s)e^\gamma}{v(s-1)})\frac{v(s)}{v(s-1)}((\frac{v(s)}{v(s-1)})^\sigma\frac{1}{v(s)})^L e^{-\gamma b} && (7.11)\\
&=\; e^\gamma(\frac{\tau}{L(e^{\gamma/L}-1)+\tau})^2(\frac{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^\sigma})^L e^{-\gamma b} && (7.12)
\end{aligned}$$

for all parameter values. But it has to be determined if this is less than 1 for $\sigma \neq 1/2$ or $\rho > \alpha$ and if

$$(\frac{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}}{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^\sigma}) < 1 \tag{7.13}$$

which would imply economies of scale exist for the particular values of σ, α and τ.

It appears to be extremely difficult to prove (in a manner other than numerically) that the bound exhibits economies of scale for other parameter values. One can for example show that,

$$\frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^\sigma}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} \tag{7.14}$$

is, for fixed $\frac{1-\alpha}{\tau}$ an increasing function of σ for $\sigma \geq \frac{1-\alpha}{\tau}$ but this implies $\sigma > 1/2$, that is it doesn't hold at $\sigma = 1/2$. Now since,

$$\begin{aligned}
h(\sigma) \;&=\; \frac{(\sigma+\frac{1-\alpha}{\tau})(1+\frac{L(e^{\gamma/L}-1)}{\tau})^\sigma}{(1+\frac{L(e^{\gamma/L}-1)}{\tau})\sigma+\frac{1-\alpha}{\tau}} && (7.15)\\
&>\; 1 && (7.16)
\end{aligned}$$

at $\sigma = 1/2$ for $\rho \leq \alpha$ the same must be true for values of σ on some interval centered on $1/2$ but $h(\sigma)$ may be decreasing on this interval i.e. $\frac{1-\alpha}{\tau}$ may not be in this interval. If it were then for all $\sigma \geq 1/2$ we would have $h(\sigma) > 1$

7.1.4 Large Deviations Approximations

Botvich and Duffield proved in [20] that,

Theorem 8 *For $b \geq 0$,*

$$\lim_{L\to\infty} \frac{1}{L} \log \mathbf{P}[q^L > Lb] = -I(b)$$

Where the function I, termed the shape function, is related to the cumulant generating function of the workload process

They also proved that for workload processes with non-negatively associated and stationary increments $\nu = -\lim_{t\to\infty} t\lambda(t,\theta)$ is non-negative (provided the limit exists) We prove,

Theorem 9 *We added the following Let the increments of the workload processes W_t^L be non-negatively associated and stationary Let $f(t,\theta)$ be defined as in Definition 9 Let $f_t^*(b)$ be its Legendre transform Then $I(b)$ is sub-additive i e the following conclusion holds*

$$I(b_1 + b_2) \leq I(b_1) + I(b_2)$$

for all $b_1, b_2 \geq 0$

and as a result,

Theorem 10 *Let the increments of the workload processes W_t^L over disjoint time intervals be stationary and non-associated Then the the following conclusion holds*

$$\begin{aligned} I(b) &= bI(1) \\ &= b\gamma \end{aligned}$$

Note that $I(b) = bI(1)$ means there are no economies of scale to be had from multiplexing large numbers of sources generating a workload process with stationary and non-associated increments over disjoint time intervals We also have,

Theorem 11 *Let the increments of the workload processes W_t^L over disjoint time intervals be stationary and non-negatively associated. Then the the following conclusion holds.*

$$\lim_{b\to\infty} \frac{I(b)}{b} \text{ exists}$$

and,

$$\lim_{b\to\infty} \frac{I(b)}{b} = \inf_{b>0} \frac{I(b)}{b}$$

We make no assumptions here about the existence of ν. We also proved,

Theorem 12 *Let the increments of the workload processes W_t^L over disjoint time intervals be non-negatively associated and stationary. Let $I(b)$ be defined as in Definition 10. Let $I(0) = 0$. Let K be an interval on the real line containing zero. Then*

$$I(b) \text{ cannot be convex on } K$$

These results do not directly bear on the cell level model but follow so simply from the definition of the shape function that we included them.

The simulations of Corcoran [9] demonstrate that the broad features of the queue length distribution remain essentially unchanged when L and the queue length b are jointly scaled. Thus we were led to consider the large deviation properties of the queue length distribution in L. We calculate,

$$\lambda_t(\theta) = \lim_{L\to\infty} \frac{1}{t} \log \mathbf{E}[e^{\theta(A_{Lt}^L - rt)}]$$

for the time rescaled cell level model where A_{Lt}^L is the time rescaled single source arrival process. We then obtained numerically the shape function for the queue generated by the rescaled workload process for a set of model parameter values previously used in the simulations in this was then used to obtain graph of an approximation to $\log \mathbf{P}[q^L \geq b]$ versus b for $L = 400$.

7.2 Future Work

We have only studied here the situation arising from homogeneous superpositions of the cell level model arrivals. Similar techniques could be applied to heterogeneous superpositions of cell level model arrivals For examples with other models see Botvich and Duffield [20] The model itself could be altered to include more than one type of silence, for example

It may be possible to prove an upper bound via Martingales for the burst level queue for the cell level model for which the prefactor is such that it exhibits economies of scale That is,

$$\phi \leq k^L \tag{7 17}$$

where $k < 1$ and independent of L for parameter values other than $\sigma = 1/2$ and $\rho \leq \alpha$ (i e $\tau \leq \frac{2\alpha(1-\alpha)}{2-\alpha}$)

The initial steep portion of the graph in chapter 6 for the shape function approximation to $\log \mathbf{P}[q^L > b]$ which are due to cell level queueing should be similar to that obtained for a queue with Poisson arrivals at the same mean rate It may be possible to substantiate this theoretically and/or by simulation

The theorem on the sub-additivity of the shape function in the case of non-negatively associated workload increments may be of some use in proving that the shape function is always concave in some case/cases

Bibliography

[1] N G Duffield, J T Lewis, N O' Connell, R Russell, and F Toomey The entropy of an arrivals process A tool for estimating qos parameters of atm traffic *Proc of 11th IEE UK Teletraffic Symp Cambridge*, 25 March 1994

[2] J T Lewis and R Russell An introduction to large deviations for teletraffic engineers *Performance, Lausanne*, 1996

[3] E Buffet and N G Duffield Exponential upper bounds via martingales for multiplexers with markovian arrivals *J Applied Prob* , 31 1049–1061, 1994

[4] J Y Hui *Switching and Traffic Theory for Integrated Broadband Networks* Kluwer Academic Publishers, 1991

[5] M De Prycker *Asynchronous Transfer Mode* Prentice Hall, 1995

[6] N G Duffield, J T Lewis, N O' Connell, R Russell, and F Toomey Estimating qos parameters for atm traffic using its entropy *Proc IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Bradford*, 4-6 July 1994

[7] E Cinlar *Introduction to Stochastic Processes* Prentice-Hall, 1975

[8] W Feller *An Introduction to Probability Theory and its Applications, Volume 1* John Wiley and Sons, 1950

[9] T Corcoran Prediction of multiplexer performance by simulation and analysis of a model of packetized voice traffic *M Sc Thesis, Dublin City University*, 1994

[10] J. N. Daigle and J. D. Langford. Models for analysis of packet voice communications systems. *IEEE J. Select. Areas Commun.*, 4 No. 6:847–855, 1986.

[11] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Select. Areas Commun.*, 4 No. 6:833–846, 1986.

[12] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, 31A:131–159, 1994.

[13] N. G. Duffield, M. Huggard, R. Russell, F. Toomey, and C. Walsh. Fast bounds for atm quality of service parameters. *Proc. of 12th IEE UK Teletraffic Symp., Old Windsor*, 15-17 March 1995.

[14] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Proc. Camb. Phil. Soc.*, 58:497–520, 1962.

[15] R. S. Ellis. *Entropy, Large Deviations, and Statistical Mechanics.* Springer-Verlag, 1985.

[16] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, 1975.

[17] D. Williams. *Probability with Martingales.* Cambridge University Press, 1991.

[18] J. F. C. Kingman. A martingale inequality in the theory of queues. *Proc. Camb. Phil. Soc.*, 59:359–361, 1964.

[19] N. G. Duffield. Exponential bounds for queues with markovian arrivals. *Queueing Systems*, 17:413–430, 1994.

[20] D. D. Botvich and N. G. Duffield. Large deviations, economies of scale and the shape of the loss curve in large multiplexers. *Queueing Systems*, 20:293–320, 1995.

[21] A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London, 1993.

[22] D D Botvıch, T J Corcoran, N G Duffield, and P Farrell Economıes of scale in long and short buffers of large multıplexers *Proc 12th IEE UK Teletraffic Symp , Wındsor, IEE London (1995)*, 15-17 March 1995

[23] N G Duffield and N O'Connell Large devıatıons and overflow probabılıtıes for the general sıngle server queue, wıth applıcatıons *Math Proc Camb Phıl Soc* , 118 363–374, 1995

[24] A Weıss A new technıque for analysıng large traffic systems *Adv Appl Prob* , 18 506–532, 1986

[25] W Feller *An Introductıon to Probabılıty Theory and ıts Applıcatıons, Volume 2* John Wıley and Sons, 1966