Self-Similar Traffic Engineering and Applications to Mobile Radio Networks.

Richard Nelson B.E.(Hons), M.E.

A thesis presented for the degree of
Doctor of Philosophy
in
Electrical and Electronic Engineering
at the
University of Canterbury,
Christchurch, New Zealand.

July 1998
ABSTRACT
Mobile networks are experiencing exponential rates of subscriber growth worldwide. In addition they are rapidly developing sophistication and capabilities for delivering multiple service types at widely varying data rates. Despite this very little is known about the traffic characteristics of such networks or the effects of such characteristics on network performance.

There is a growing body of evidence that video and data traffic is not Markovian and cannot be modelled using the conventional models developed for wireline telephone networks. Instead these kinds of traffic are characterised by stochastic self-similarity and are highly bursty in nature. This bursty nature is known to significantly degrade network performance, but few analytic results quantifying this are available.

Assuming that the self-similar nature of packet oriented voice and data traffic will hold when it is carried on mobile networks, this thesis develops the analysis of fundamental network elements, basic queues and random access channels with packet arrivals using a simple mathematical model of a self-similar process.

These results quantify the degradation in network performance through increasing blocking and delay caused by the shift from random Poisson arrivals to self-similar arrivals and the further effects as the self-similarity parameter increases.

Also presented here is the analysis of data sets taken from live New Zealand based cellular telephone networks. This analysis is limited by the data sets available but shows that where handover traffic is high the channel seizure process appears not to be Markovian and may be self-similar.

The thesis concludes with a discussion of the difficulties in obtaining valid results from computer simulation when the output data stream is self-similar. This is caused by the unique statistical nature of such sequences.
ACKNOWLEDGEMENTS
I wish to thank my supervisor Prof. Des Taylor for his help and guidance with my work and in particular his extreme patience with my progress. Sven Ostring, Parag Pruthi and Vern Paxson have also been of great assistance, providing papers and as sounding boards for ideas and by Sven's proof reading of this thesis.

Without the co-operation I have received from people in industry I would never have been able to collect live network data. In particular I would like to thank Peter McFarlane and Alistair Ryan of Telecom Mobile Networks, Rob Sweet of Broadcast Communications Ltd and Graeme Goss, Gary Woolf, Giselle Baker, Arasaratnam Sathyendran and Peter Anderson of BellSouth New Zealand.

The financial support I have received from the University of Canterbury through its Doctoral Scholarship program was crucial to me undertaking this degree.

I want to thank all my paddling buddies, especially Andrew, James, Karyn, Paul, Katherine, Troy, Nick and Al for the fun I have had and the enormous amount I have learned from them over this time. I also owe thanks to many of my flatmates, especially Katharine Holdsworth for the times we shared.

The support of my family has been crucial in helping me complete this work. My parents Martin and Cynthia, have always been extremely supportive and encouraging. My parents-in-law, Geoff and Margaret, have housed and looked after me, allowing me to concentrate on the write up phase. My wife, Sarah has provided enormous encouragement and the motivation required for me to complete the work.

I dedicate this thesis to the memory of my Mother, Cynthia Nelson.
<table>
<thead>
<tr>
<th>CHAPTER 3</th>
<th>INTER-ARRIVAL TIME DISTRIBUTIONS</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>The Exponential Distribution</td>
<td>27</td>
</tr>
<tr>
<td>3.2</td>
<td>The Pareto Distribution</td>
<td>29</td>
</tr>
<tr>
<td>3.3</td>
<td>Pareto Spaced Arrivals</td>
<td>34</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Aggregated Pareto Renewal Processes</td>
<td>35</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Proof of Self-Similarity of Pareto Spaced Arrivals</td>
<td>36</td>
</tr>
<tr>
<td>3.4</td>
<td>Large Variance Gamma Distribution</td>
<td>39</td>
</tr>
<tr>
<td>3.5</td>
<td>Discussion</td>
<td>41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 4</th>
<th>ANALYSIS OF QUEUES WITH PARETO ARRIVALS</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Queueing Models</td>
<td>43</td>
</tr>
<tr>
<td>4.2</td>
<td>Queue Analysis</td>
<td>44</td>
</tr>
<tr>
<td>4.2.1</td>
<td>The Infinite Buffer Queue</td>
<td>45</td>
</tr>
<tr>
<td>4.2.2</td>
<td>The Finite Buffer Queue</td>
<td>49</td>
</tr>
<tr>
<td>4.3</td>
<td>Results</td>
<td>51</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Infinite Buffer Approximation</td>
<td>53</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Effect of Queue Length</td>
<td>56</td>
</tr>
<tr>
<td>4.4</td>
<td>Simulation of P/D/1/N Queues</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Discussion</td>
<td>64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 5</th>
<th>MULTIPLE SERVER QUEUES</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Radio Interface Model</td>
<td>67</td>
</tr>
<tr>
<td>5.2</td>
<td>The G/M/c Queue</td>
<td>68</td>
</tr>
<tr>
<td>5.3</td>
<td>The P/M/c/c Queue</td>
<td>72</td>
</tr>
<tr>
<td>5.4</td>
<td>Results</td>
<td>73</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Effect of Server Aggregation</td>
<td>76</td>
</tr>
<tr>
<td>5.5</td>
<td>Discussion</td>
<td>76</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 6</th>
<th>RANDOM ACCESS PROTOCOLS</th>
<th>81</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Data Traffic</td>
<td>81</td>
</tr>
<tr>
<td>6.2</td>
<td>Pure Aloha</td>
<td>82</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Throughput</td>
<td>83</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Maximum Throughput</td>
<td>87</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Delay</td>
<td>88</td>
</tr>
<tr>
<td>6.3</td>
<td>Slotted Aloha</td>
<td>89</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Throughput</td>
<td>90</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Delay</td>
<td>94</td>
</tr>
<tr>
<td>6.3.3</td>
<td>Stability</td>
<td>94</td>
</tr>
<tr>
<td>6.4</td>
<td>Carrier Sense Multiple Access</td>
<td>98</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Throughput Analysis</td>
<td>98</td>
</tr>
</tbody>
</table>
CONTENTS

6.4.2 Results 104
6.5 Discussion 107

CHAPTER 7 DATA ANALYSIS 111
7.1 Statistical Tests for Self-Similarity 111
  7.1.1 Autocorrelation 112
  7.1.2 Variance Analysis 113
  7.1.3 R/S Analysis 114
7.2 Data from Telecom New Zealand 116
7.3 Data from BellSouth 119
  7.3.1 Channel Occupation Series Results 122
  7.3.2 Arrival Sequence Results 123
7.4 Holding Times 126
7.5 Discussion 126

CHAPTER 8 STOCHASTIC SIMULATIONS WITH LONG RANGE DEPENDENCE 129
8.1 Simulation Studies 129
8.2 Stochastic Simulation Analysis 130
  8.2.1 Batch Means 130
  8.2.2 Spectral Analysis 131
  8.2.3 Independent Replications 132
8.3 Variance of Data Sequences with Long Range Dependence 132
8.4 Central Limit Theorem and Infinite Variance 133
8.5 Discussion 134

CHAPTER 9 CONCLUSION 137
9.1 Mobile Radio Network Traffic 138
9.2 Self Similar Traffic Engineering 138
9.3 Recommendations for Future Work 140
Chapter 1

INTRODUCTION

The research work for this thesis began with a single broad questions to answer:

- How do the traffic characteristics of mobile radio networks affect network performance?

Clearly, in order to answer this question it was first necessary to work out what the characteristics of traffic on mobile radio networks are and how they should be modelled. Upon investigating this, it soon became apparent that, although introducing handovers is widely believed to change telephony traffic characteristics and the rate of handovers is increasing as cells become smaller, the actual characteristics of mobile traffic are not really known at all. Therefore effort has been put into obtaining data from live networks and analysing this to detect differences from the classical Markovian models used for conventional wireline telephony networks.

Looking at the proposed designs for the next generation of mobile radio networks, however, it is apparent that they are increasingly becoming packet oriented systems carrying multiple traffic types, much as current fixed networks are developing. Therefore it seems appropriate to extrapolate the characteristics of current fixed networks carrying mixed voice/data/video traffic to future mobile networks and to use this as a basis of analysis. Recently published research shows that the traffic on such networks is often not able to be described by Markovian models, but is instead self-similar or fractal like in character. Markovian models of network traffic are frequently characterised by tractable analysis such as the standard queueing results for Poisson arrivals [40]. No such elegant results have been published for self-similar models(such as those in section 2.5. Therefore the research presented here has concentrated primarily on finding a simple model that can be applied to fundamental network situations and allow useful results to be obtained.
1.1 CONTRIBUTION OF THE THESIS

This thesis presents for the first time exact analytical results for the blocking probability and expected waiting time in finite buffer queues, both single and multi-server, with a self-similar arrival process. These results are used to quantify the decrease in throughput as the self-similarity parameter of the process increases and also the change in the relationship between buffer size and blocking.

Also presented in the thesis are the exact analytical derivations of the load-throughput relationships for the Aloha, Slotted Aloha and Carrier Sense Multiple Access protocols with a self-similar arrival process for the total load (new arrivals plus re-transmissions). The analysis in chapter 6 shows that Poisson arrivals are not a lower bound on throughput and self-similar arrivals can produce worse performance, unlike the conclusions of at other simulation studies [4, 43]. As well, the delay performance results for Aloha and slotted Aloha are given and the stability of uncontrolled slotted Aloha is investigated.

Finally, presented here is the first analysis of mobile radio traffic within individual cells using heuristic techniques to detect self-similarity. This analysis, although severely limited by the quality of the data sets obtained, shows that although new call arrivals appear Markovian, when handover arrivals start to dominate, the arrival process can appear self-similar.

1.2 THESIS OUTLINE

The thesis is structured as follows:

- Chapter 2 contains background material covering mobile radio networks and self-similar traffic models. The first section covers third generation cellular networks, commonly know as Personal Communications Networks (PCNs). It describes design targets for such networks as well as network and radio interface issues and traffic types and models. The self-similarity section introduces some basic concepts and then describes network studies where the traffic has been shown to be self similar as well as outlining the most common self-similar models.

- Inter-arrival time distributions used for traffic modelling are described in chapter 3. Distributions covered are the exponential distribution of the classical Poisson arrival model, the heavy-tailed Pareto distribution which has been shown to result in a self-similar arrival process and the self-similar-like large-variance Gamma distribution. Inter-arrival time distributions provide
simple models and have been used as the basis of the analytical work on network performance in this thesis.

- The performance of single server queues with self-similar arrivals is analysed in chapter 4. Both the infinite and finite buffer cases are covered and comparison is made to the conventional Poisson model of random arrivals. The effect of the buffer size on the performance of finite buffer queues is investigated. Simulation is used to compare deterministic service queues to the analytical results for exponential service time queues.

- Chapter 5 extends this work to the performance of multi-server queues. In particular results are given for the blocked-calls-lost queue model commonly applied to telephony systems.

- Chapter 6 investigates the performance of the simple random-access protocols, Aloha, Slotted Aloha and Channel Sense-Multiple Access (CSMA), with self-similar traffic. The generalised throughput-load relationship for each protocol is derived and compared to the Poisson arrivals case. Mean delay and some simple stability analysis of the Slotted Aloha protocol is also given.

- Traffic data sequences taken from the Telecom New Zealand and BellSouth New Zealand cellular telephony networks are analysed in chapter 7 using standard heuristic techniques. The particular difficulties of testing for self-similarity, which include sequence length and ensuring stationarity over the measurement period are discussed.

- Chapter 8 discusses the issues of producing statistically valid simulation results from system models with self-similar parameters. The unique statistical properties of such sequences require particular care to obtain useful results.

- Chapter 9 concludes the thesis with a discussion of the results obtained and their implications for network design. Recommendations for future research work are provided.
Chapter 2

BACKGROUND

2.1 MOBILE RADIO NETWORKS

Mobile Radio Networks are one of the significant communications success stories of the 1980's and 90's. The ability to provide seamless telephony irrespective of location or movement has attracted customers worldwide and universally, these networks have seen exponential rates of growth. Further, the ability to provide telephony services without needing to lay cables to every customer's premises has proved very attractive to telecommunications companies, both as a means to compete with established network operators and to provide service to rapidly developing areas with little pre-existing infrastructure.

Due to the high rate of customer demand for mobile communications and the rapidly changing nature of telecommunications services, technological development of mobile networks has been extremely rapid.

Mobile networks are starting to compete with fixed networks in terms of their ability to handle, different services, multiple concurrent traffic types and varying rates. Although high broadband rates are becoming possible, mobile networks are unlikely to ever compete in terms of raw capacity.

The general architecture of such networks is shown in figure 2.1. Mobile terminals communicate by a low-power radio link with their closest base station, which is the radio transmitter/receiver for a local area called a cell. Base stations are connected to and controlled by base station controllers (BSCs) which, in turn, are connected together by Mobile Switching Centers (MSCs). The MSC is usually the point of connection to the general PSTN wireline network. Since Mobile stations may roam freely between cells, location information must be maintained in Location Databases, which must be available to the network as a whole.

The single largest bottleneck mobile networks have is at the radio interface between the mobile terminal and the base station. They are a premium service,
offering features (i.e. mobile communications) unavailable in any other way, so unlike conventional wireline networks their design is not primarily cost driven, but instead is driven by the need to maximise the utilisation of the available radio bandwidth. Unfortunately, while the radio interface is the most important aspect of the design of mobile networks, little is known, even now, about the characteristics of the traffic over this interface and so it presents the greatest engineering challenges.

2.1.1 Personal Communication Networks

Currently, throughout the world a second generation of digital mobile networks is replacing first generation analogue systems. Europe is leading the deployment of these systems with the cellular telephone network Global System for Mobile communications (GSM) [5, 22, 59, 77]. Also being introduced are Digital European Cordless Telecommunications (DECT) [14, 70], European Radio MESSaging System (ERMES) and Trans European Trunked RAdio (TETRA) [20, 34]. In Australia there are three GSM network operators and all analogue services are to be turned off on January 1, 2000. In the the USA a second generation digital TDMA (Time Division Multiple Access) system (IS 54) has been introduced that operates within the original analogue channels of the Advanced Mobile 'Phone
Service (AMPS) system [71]. This is currently being replaced by the IS136 standard and a CDMA (Code Division Multiple Access) system (IS 95) that is also compatible with the analogue standard. The motivating forces driving the introduction of these systems are the shift from analogue to digital radio transmission technology to increase spectral efficiency and transmission quality and the need for greater international standardisation to allow roaming and to reduce equipment costs [59].

However these systems have a number of limitations that will prevent them from meeting future needs. The main limitation is capacity [20]. Mobile services still only serve a small percentage of the population, but demand is growing very quickly and this growth is likely to be sustained while services are improved and costs fall [6]. Forecasts for future service demand predict up to 80% penetration of the total population, creating an enormous user base which the second generation systems are not designed to cope with [1]. For comparison, the mobile phone market penetration as of September 1998 was 23% of the population in the USA [41].

Another limitation is services [15]. Each of these second generation networks provides only a specific range of related services. If users are to take advantage of more mobile network services they will all need to be available from a single terminal. With the boom in multimedia stations and mobile computing, requirements for new services are being created. In particular, Internet based applications such as electronic mail and the world wide web are being deployed at an exponentially increasing rate. Fixed networks are meeting these requirements with high bandwidths and flexibility through Asynchronous Transfer Mode (ATM) transmission based on optical fibres, however mobile users will want access to these services too.

Competition is being introduced into the telecommunications sector. In future it is likely that users will often have multiple mobile service operators available and may handover between them during calls, e.g. from a public operator on the street to a different operator within a building, or will want to be able to select different operators for different calls depending on services and prices [20]. This multi-operator capability is also lacking from most second generation systems. To meet these limitations, researchers are now finalising possible designs for third generation universal mobile networks called Personal Communications Networks (PCNs).

The major aims for these systems are [20, 15]:

- to provide capacity to support large user populations and, particularly, very
high user densities such as occur in the centres of large urban areas.

- to support existing and future voice, data, video and multimedia services at all data rates from very low to as high as practical (current target is 2Mbps).

- to support new services, unique to mobile networks such as location, navigation and traffic information.

- to allow terminals to be used everywhere - in homes and office buildings as well as in public environments and in all areas from rural districts to towns and cities.

- to provide a range of terminals - from low cost portable pocket sized phones which anybody can afford and use, to larger mobile terminals for sophisticated services such as video and high rate data.

- easy inter-operator roaming and complete international standardisation.

The main standardisation efforts are the Future Land Public Mobile Telephone System (FPLMTS) being standardised by the ITU, the Universal Mobile Telephone System (UMTS) under development in Europe and the International Mobile Telecommunications by the year 2000 (IMT-2000) effort in Japan. The aim is to introduce such systems around the year 2000 and to achieve high service penetration by 2010 [11, 21, 34].

2.1.2 Design Considerations

Providing these new capabilities will require new technology. To meet the capacity requirements within the limited available radio bandwidth and to reduce transmitter power levels, PCNs will use microcells with typical coverage of one or two city blocks. Inside buildings, picocells will be used to provide coverage of individual rooms [78]. This means that users will move into new cells much more frequently than in current cellular systems which use macrocells with coverage of several square kilometres. This will rapidly increase the rate of handovers and location updates in proportion to the number of users. Combined with a much larger user population, this will cause the signalling requirements of the system to increase enormously [6].

With these new demands, the current centralised style network infrastructure is likely to be a significant bottleneck. New network designs are therefore being
proposed. In order to meet the processing requirements for location tracking, distributed database based systems are proposed [21, 86, 87]. Distributed processing is also likely for control of call processing and the network and radio resources to increase the system capacity and add reliability through redundancy [37]. These distributed systems imply sophisticated signalling protocols and a high level of communication between network components.

Handovers must occur much more rapidly as they need to be complete before a user may move into yet another cell. This is exacerbated by the radio propagation characteristics of microcells where a user turning a street corner might experience an immediate 20-30dB drop in SNR [23]. Therefore the wireless terminals need to be able to initiate handovers independently and the network must be able to respond and resume the connection with minimal interruption to voice connections and no loss of data. The small cell sizes also increase the demands on the control system which ensures adequate radio resources for each cell to meet fluctuating demand while preventing interference between neighbouring cells [89]. Such Dynamic Channel Assignment provides for increased geographical variation of network traffic yet causes further loading of signalling networks.

PCNs must integrate easily with fixed networks [1]. They will have to provide seamless connectivity to users of other telephony services (the current public switched telephone network, the integrated services digital network and existing mobile systems) to attract users. To provide data services, PCNs will have to
connect to data networks (e.g. X.25). For the delivery of advanced new services, such as multi-media, integration with ATM networks and the Internet will be required [10, 15].

2.1.3 Radio Interface

The radio interface for third generation systems is designed to provide access in all possible radio environments for possibly very dense user populations at access rates up to $2M_{b/s}$ [11]. It will provide both circuit mode and packet mode access and is to integrate with existing radio networks (GSM) and services based on ISDN, ATM and Internet networks. Two $60M_{hz}$ wide radio bands at $1920 - 1980M_{hz}$ and $2110 - 2170M_{hz}$ and a further $20M_{hz}$ wide band at $1900 - 1920M_{hz}$ have been allocated for terrestrial applications.

To attain all these objectives within the bandwidth available, several access modes are to be supported. TDMA and TDMA with spreading modes provide high compatibility with GSM allowing single receiver implementation of dual mode access. DS-CDMA (Direct Sequence -CDMA) provides the most flexible access with different codes allowing for high or low delay spread (i.e. different cell sizes). All three access modes provide access rates varying from a few $kb/s$ to $Mb/s$ in small steps by using flexible slot allocation, variable spreading and multiple spread codes as appropriate. Because of the increasing need to provide international and inter-operator roaming access, there is a great deal of effort being applied to ensure compatibility between the rival standardisation efforts of Europe, Japan and the ITU. Since these projects will provide a compatible follow-on system for the hugely successful GSM system, they are likely to be implemented all over the world, even in the U.S.A..

2.2 MOBILE NETWORK TRAFFIC

2.2.1 Traffic Types

Traditional telephone networks have been oriented towards carrying voice traffic. Modern networks however are also increasingly supporting a range of data applications from real time interactive enquiries to low priority messaging. Also increasing in popularity are graphics and video applications. To support this range of traffic types and indeed to operate at all, mobile networks also need to carry a range of signalling protocols. All these traffic types have their own characteristics and performance requirements.
2.2 MOBILE NETWORK TRAFFIC

Voice transmission requires low delay and delay jitter to maintain acceptable perceived quality of transmission. The ITU allows maximum transmission times of up to 400ms [48], however this is designed to accommodate geosynchronous satellite links which incur long delays. Perceived voice quality deteriorates as delay increases from a few tens of ms. However voice transmission can acceptably incur high error rates ($< 10^{-3}$) and dropped packets. Traditionally, to meet these requirements, voice traffic has been carried on channel oriented services such as FDMA channels or fixed slot allocation TDMA.

In order to conserve radio bandwidth, voice is typically digitally encoded at a low rate for transmission. For example, GSM uses a 13kbit/s voice codec and the continuing improvement of voice codecs will see rates of 8kbit/s and lower become commercially acceptable in the near future [36]. Such low voice codec rates mean that voice packets must be very small if they are not to incur significant inherent delay. Even encoded at 16kbit/s, a 40 octet packet represents 20ms of speech. Halving the codec rate doubles this inherent delay if the packet size remains unchanged. This favours channel oriented service such as fixed time slot allocation TDMA which can multiplex almost arbitrarily small packets with little overhead onto a high transmission rate channel.

However, voice transmissions typically contain speech activity less than 50% of the time [35]. Therefore, speech activity detection (SAD) can be a useful technique to increase capacity. The largest capacity gains are achieved at high gross data rates in small cells [24]. These are precisely the conditions that will be present in the highest subscriber density environments when capacity gains are most needed. This makes SAD very useful. This also indicates that cell channels should be configured as few (ideally one) high capacity channels as possible rather than many low capacity channels. This is in contrast to current generation mobile systems with dedicated circuits for each mobile where groups of small channels are used either to enhance flexibility for channel allocation or to provide backward compatibility. For example GSM provides eight circuits per channel and IS 54 only three [71]. The use of SAD in voice encoders is a reason to favor packetised service for transmission and introduces another level of variability into voice traffic.

Data traffic has the opposite delay/error rate performance requirements for transmission to voice traffic. Delay requirements vary depending on the application, from 100's of ms for interactive applications to the effectively unbounded delay tolerated by store and forward messaging (email). Error rates need to be low ($< 10^{-6}$) and lost packets have to be minimised as they incur retransmissions.
In order to meet the error requirements, data transmission uses higher level protocols for error checking and correction and retransmission. These higher level protocols mean that even continuous transmission applications such as file transfer do not approach 100% channel utilisation. Typically these applications have utilisation rates around 30% and interactive applications may have utilisation below 4%. Therefore the potential capacity gain available from not using fixed slot allocation ranges from 3 up to around 25. Again, the actual capacity gain realised is greatest where the need for it is highest. For these reasons, data transmission strongly favours packet oriented services for transmission.

Data transmissions typically involve either very long or very short packets. Short packets are keystroke information, acknowledgments and other transmission control information. Long packets are parts of transmission sequences and provide bursts of transmission activity. They are often file transfers, screen updates or image transmissions. This produces long periods of low activity interspersed with short periods of high intensity so data traffic is often described as "bursty" in nature. It has been assumed that the bursts from many sources using a single transmission channel tend to average out, but recent evidence shows that this does not necessarily occur [56].

Video transmission requires good synchronisation and thus low delay jitter. If it is interactive video the total delay should also be low as quality degradation similar to voice telephony is experienced when it increases [48]. Error rates for high quality video also should be low; however, it is possible to encode video into high and low priority data. In this case the low priority data can stand higher bit error rates and/or packet losses. Most video encoding schemes with high compression ratios transmit the difference between successive frames and so have highly variable bit rates with large transmission bursts occurring at scene changes and medium or low transmission rates during scenes.

Signalling traffic is inherently bursty with exchanges at the beginning and end of customer connections, but little required in between. To support advanced services and mobility, high volumes of signalling are required. The standard telecommunications network signalling protocol, System 7, is based on a packet network for efficiency and speed. Mobile services have been shown to produce large demands for signalling bandwidth and network processing [57], yet this traffic must be handled quickly in order to ensure good service to customers and high efficiency of the overall network.
2.2 MOBILE NETWORK TRAFFIC

2.2.2 Traffic Aggregation

Traffic aggregation in conventional traffic models reduces variability within the flow. For example, with Poisson arrivals, the standard deviation of the number of arrivals increases only as the square root of the mean. This feature can also be seen by examining Erlang blocking probabilities and can be taken advantage of in telephone networks where aggregating calls from larger areas allows more efficient use of facilities such as trunks. This effect is shown graphically in section 2.3.2. However, aggregating some traffic flows can produce larger flows with the same statistics and the same relative levels of variation. This effect has been demonstrated within real networks (e.g. [56]).

Mobile terminal networks designs are such that calls from within cells are aggregated at switches and all the cells connected to the same switch are from within the same geographical region. Therefore the variations in the traffic flow caused by movements of terminals between the cell areas are exactly cancelled out. Since the number of subscribers in any coverage region is generally proportional to its area and the rate of terminals entering and leaving the area is proportional to its perimeter length then the ratio of terminal movements to total calls falls as the region increases in size. Handovers between cells within the same region have no effect on the aggregated traffic flow.

Calls placed from within small inner city cells can average more than one handover per call so that base stations in these cells receive more handover calls than new calls. So naturally the traffic characteristics are strongly influenced by the flows into and out of the cell. A call placed to a terminal that has roamed to a cell connected to an MSC that is different from its home location uses bandwidth inefficiently, so networks are designed so that this type of call is quite rare. As a result, the traffic flows at these switches have characteristics essentially identical to fixed terminal network switches. This is demonstrated in section 7.2.

2.2.3 Traffic Models

There are many types of mathematical models that can be used to describe telephone traffic. The classical model for wireline telephone networks is a Poisson stream of arrivals, i.e. arrivals with independent, exponentially distributed inter-arrival times. This model describes completely random arrivals (see section 3.1). However, there are reasons to expect that this model will not be adequate for mobile networks [33, 57]. The most obvious factor is that the numbers of traffic sources, as well as varying over time, can vary significantly geographically. As
well as normal day to day fluctuations, events such as large sporting fixtures or simple traffic jams can cause hot-spots, overwhelming the local network resources. Additionally, the assumptions of the Poisson model of an infinite (or sufficiently large) population of independent users do not hold when they are divided into extremely small cells. Future network micro- and picocells may only hold a few users who can easily all be affected by local events as well as each other. Lastly, the traffic in small cells in public areas may be dominated by handovers, rather than new call arrivals. Handover arrival rates are directly related to the flows of vehicles and pedestrians through the cell [50]. Vehicle and pedestrian flows are characterised by platoons (bursts) especially in city centres where they are controlled by traffic lights [66]. Therefore, there are several possible models to consider:

- A conventional Markov or ARIMA (Auto Regressive Integrated Moving Average) model. These models are good at capturing short range dependencies in random behaviour. However due to the number of factors (multiple cells, cell sizes, cell user population, population mobility, etc.) involved in mobile networks they are likely to require a large number of states or state variables making them complex and inefficient in practise.

- A frequency domain model. Many of the fluctuations in mobile populations are periodic in nature: time of day behaviour, traffic light control of traffic, "platooning" of pedestrians. However, these do not necessarily result in regular cycles and there are so many possible influences that a model would be difficult to parameterise.

- Self-similar time series. These models have recently gained exposure for modelling a great variety of physical phenomena including various types of communications traffic. They include behaviour that fluctuates over a wide range of time scales and are inherently bursty in nature. These models can be a simple way of reflecting very complex behaviour, but it is difficult to establish that such models are appropriate.

Markov, ARIMA and frequency domain models are all likely to be highly geography specific, in that the parameters will vary significantly depending upon the local features affecting the user population of the network. Recently published work has shown that characterising traffic explicitly can result in a need to provide models over a range of geographical scales [19].
Because of the apparent difficulties of the other approaches, self-similar models have been investigated in this work so that their applicability and their impact on network performance might be assessed.

2.3 SELF SIMILAR TRAFFIC

Observations of self-similar behaviour in network traffic have been made on extensive and accurate data sets from Ethernet LANs [56], wide area data networks [74], CSN7 signalling networks [18], variable bit rate encoded video samples [9] and an ISDN [68]. Self-similar traffic exhibits bursty behaviour across a wide range of time scales and aggregation of smaller traffic streams does not lessen this burstiness or variability. The impact of this type of traffic behaviour is such that the pioneering work at Bellcore [56] has been described as "arguably the most important networking paper of the decade" [81].

The pioneering work on self-similarity and its application to communications was by Mandelbrot. In 1965 [61] he argued that an inter-arrival time distribution of the form \( P(x) = Pr(\text{interval} \geq x) = Cx^{-\alpha} \) would produce a self-similar arrival process. Here \( 0 < \alpha < 1 \) is a fixed parameter of the distribution and \( X \) is defined between limits that approach 0 and \( \infty \). The distribution is normalised by appropriate choice of the scaling constant \( C \). This distribution has a heavy tail so that arbitrarily long inter-arrival times may occur with finite probability. It also has an asymptote at zero, allowing (nearly) zero inter-arrival times to occur providing bursts of closely spaced arrivals. This means that the arrival count process is invariant when scaled up from even the smallest of initial time intervals. This combination of bursts and spaces is characteristic of self-similar series.

Distributions, such as this which decay more slowly than exponential are referred to as *heavy tailed* and are often associated with self-similar series. The best known heavy tailed distribution is the Pareto distribution which is covered in detail in section 3.2.

### 2.3.1 Long Range Dependence

A property of self-similar sequences is that they demonstrate long range dependence i.e. they have high-lag correlations that, although small, are cumulatively significant. Mathematically, long-range dependence is defined as the autocorrelations of a sequence decaying hyperbolically with increasing lag. That is, a covariance stationary process \( X = (X_t : t = 0, 1, 2...) \), with mean \( \mu \) and variance
$\sigma^2$ is long range dependent if it has an autocorrelation function $r(k), k \geq 0,$ of the form

$$r(k) \sim k^{-\beta} L(t) \quad \text{as} \quad k \to \infty$$

(2.1)

where $0 < \beta < 1$ and $L(t)$ is slowly varying at infinity, i.e. $\lim_{t \to \infty} L(tx)/L(t) = 1$ for all $x > 0$ ($L(t)$ is often assumed to be constant). A demonstration of such long range dependence is given in figure 2.3. The autocorrelations of a Poisson arrival sequence can be seen to be effectively zero for all lags, i.e. the arrivals are independent. However the autocorrelations of an arrival sequence with Pareto distributed inter-arrival times are significant even for large lags. The general hyperbolic shape of the autocorrelation can be seen. Both of the arrival sequences used in this figure were generated by appropriately transforming the same uniformly distributed, pseudo-random number sequence.

This definition of long-range dependence implies that for such a sequence the autocorrelations have no finite sum, i.e.

$$\sum_{k=0}^{\infty} r(k) = \infty$$

(2.2)

Therefore, the existence of a finite sum of the autocorrelation function for any sequence proves that it does not have long range dependence and hence is not self-similar.

One effect of long range dependence is that the power spectrum of a series increases without limit as it approaches zero frequency. The spectral density near the origin of such processes takes the form $\omega^{1-2H}$ where $\omega$ is frequency and $H$ is the self-similarity or Hurst parameter defined in section 2.3.2 and is in the range $1/2 < H < 1$. Since $H$ is frequently observed close to 1 these are often termed 1/f noises, however this term is inaccurate as values of $H$ not close to 1 are also observed [65].

The parameter $\beta$ of the autocorrelation function is related to the Hurst parameter by:

$$H = 1 - \frac{\beta}{2}$$

(2.3)
2.3 SELF SIMILAR TRAFFIC

Traditional traffic models have independent or only short-term autocorrelated traffic values. As they are aggregated, by taking the means of blocks of arrival intervals, they tend towards a sequence of i.i.d. random variables, or covariance stationary white noise. A self-similar sequence, however, either remains statistically indistinguishable from the original after such aggregation (exactly self-similar) or converges to a self-similar sequence (asymptotically self-similar). In either case, the self-similar sequence, before and after aggregation must be distinctly different from white noise.

The currently accepted definition of self-similarity was given by Leland et al [56]. A series, $X = (X_t : t = 0, 1, 2...)$, is assumed to have long range dependence. The aggregated sequence $X^{(m)} = (X^{(m)}_k : k = 1, 2, 3...)$, where $m = 1, 2, 3...$, is a new covariance stationary series obtained by averaging the original series $X$ over non-overlapping blocks of size $m$, i.e. $X^{(m)}_k = (X_{(km-m+1)} + X_{(km-m+2)} +...+ X_{km})/m, k \geq 1$. The autocorrelation function of this series is denoted as $r^{(m)}(k)$. The process $X$ is (exactly) second-order self-similar with self-similarity (Hurst) parameter $H = 1 - \beta/2$ if
for all values of $m > 0$. $X$ is \textit{(asymptotically) second-order self-similar} with self-similarity (Hurst) parameter $H = 1 - \beta/2$ if for all $k$ large enough:

$$r^{(m)}(k) = r(k), \quad k \geq 0 \quad (2.4)$$

Self-similarity can be demonstrated visually, by graphing the traffic on a wide range of time scales. Conventional traffic models converge to white noise as the time scale increases by only a couple of orders of magnitude, whereas self-similar traffic retains similar characteristics over a wide range of time scales, hence the term "self-similar". Self-similar sequences are sometimes referred to as \textit{fractal-like}. Fractals are an example of \textit{deterministic} self-similarity in that they repeat exactly when scaled. However, we are concerned here with \textit{stochastic} self-similarity, that is random sequences that remain statistically the same when scaled. This effect can be seen in figures 2.4 and 2.5 which show a Poisson arrival sequence and an arrival sequence with Pareto distributed inter-arrival times respectively. The lower two traces in each figure show the sequence averaged over ten and one hundred time units and it can be seen that the Poisson sequence quickly tends towards a constant value, whereas the Pareto sequence maintains similar bursty characteristics. The mean number of arrivals per unit time is 10 for the Poisson sequence. The Pareto spaced sequence was generated with a theoretical mean of 10 arrivals per unit time, however due to the high variance of this distribution the actual mean of the sequence shown here is 13.5 arrivals per unit time.

Self-similarity arises in a natural way from the limit theorems for sums of random variables. This was first demonstrated by Lamperti [54, 85]. Lamperti stated that if stochastic sequence $Y_t$ was had continuous probability then the sequence $X_t$ found as the limit in distribution

$$a_n Y(n \cdot) \to_d X(\cdot) \quad \text{as} \quad n \to \infty \quad (2.6)$$

where $a_1, a_2, \ldots$ is a sequence of positive normalising constants such that $\log a_n \to \infty$ and $X(1) \neq 0$ with positive probability. Then there exists an $H > 0$ such that
Figure 2.4 Poisson Arrival Sequence on Logarithmically Spaced Time Scales.

Figure 2.5 Pareto Arrival Sequence on Logarithmically Spaced Time Scales.
for any \( u > 0 \)

\[
\lim_{{n \to \infty}} \frac{a_{nu}}{a_n} = u^H
\]

and \( X_t \) is self-similar with self-similarity parameter \( H \) and has stationary increments. That is, whenever a process is the limit of normalised partial sums of random variables then it is necessarily self-similar. Therefore, self-similar processes play a role among stochastic processes analogous to the role of stable distributions among distributions [8]. Lamperti's theorem also states that all self-similar processes with stationary increments and \( H > 0 \) can be obtained by partial sums in this way.

### 2.3.3 Hurst Effect

A hydrologist, Hurst first noticed this type of effect while he was studying river flows and a range of other natural phenomena [47]. Hurst studied the rescaled adjusted range parameter \( R(d) \) which is defined on a sequence of measured data \( X_i \) of length \( d \). A secondary sequence is found by taking the difference between each measurement and the mean of the entire sequence. For a river system \( X_i \) represents the river flow into a reservoir. The differences between the \( X_i \) and the mean flow over the measured sequence, \( \bar{X}(d) \), represents the outflow required from a reservoir to maintain constant flow downstream. The cumulative sequence \( W_k \) found from these differences gives the total outflow (negative gives inflow) from the reservoir up till that year (year \( k \)). So then the maximum range of \( W_k \) is the maximum volume required of a reservoir to maintain constant flow downstream over the period of the measured sequence. The range of \( W_k \) is calculated, including \( W_0 = 0 \) as the starting point before the initial period. Then \( R(d) \) is the maximum range of the sequence \( W_k \) as given by

\[
R(d) = \max(0, W_1, W_2, \ldots, W_d) - \min(0, W_1, W_2, \ldots, W_d)
\]

\[
W_k = \sum_{j=1}^{k} (X_j - k\bar{X}(d)) \quad (1 \leq k \leq d)
\]

\[
(2.8)
\]

It can be shown that the relationship between \( R(d) \) and the sample standard deviation \( S(d) \) for independent (Markovian) and short term correlated sequences is
However, Hurst found that for the Nile river in particular and indeed most other rivers (the Rhine being the notable exception) that

\[
\frac{R(d)}{S(d)} \sim C_1 d^{\frac{1}{2}} \quad \text{as } d \to \infty \quad (2.9)
\]

This relationship has been shown to be true for a wide range of naturally occurring series with \( H \) commonly being around 0.75. The Hurst parameter, \( H \), is taken to be a measure of the self-similarity of the series. Recent studies have estimated \( H \) for a range of teletraffic types and also almost always found it to be greater than 1/2, typically \( H \approx 0.7 - 0.8 \) (see section 2.4).

### 2.4 NETWORK STUDIES

Self-similarity is only of interest in a networking context because it has been detected in live networks. Studies on a range of network and traffic types have been recently published with findings indicating and sometimes proving self-similar behaviour.

#### 2.4.1 ISDN

The arrival distributions of packets on an ISDN system were characterised in a study by Meier-Hellstern et al [68]. Users of a variety of typical office automation applications generated the studied traffic. Statistics were collected on the length of data packets, mean packet rate and packet inter-arrival time distribution, call lengths and DCE (Data Communications Equipment, the host) response bytes per DTE packet (Data Terminal Equipment -the user terminal).

The transmissions from the DCE consisted almost entirely of short packets; 98% contained three or less data bytes and almost all the remainder contained less than ten data bytes. The DTE packet lengths were bi-modally distributed. Most were very short (< 10 bytes), but a significant fraction were the maximum possible length allowed by the transmission protocol.
The mean call duration was 32 minutes with a standard deviation of 52 minutes. The interesting point is that the standard deviation is so large that the distribution of call durations cannot have been negative exponential.

The DCE response bytes measurement quantifies the host response to DTE stimulus. The mean response was 62 bytes per DTE packet with a standard deviation of 436.

The packet inter-arrival time distribution was modelled using the DTE packet inter-arrivals and the DCE packet inter-arrival time distribution was regarded as supplementary to this. The measured distribution was divided into three states and distributions fitted to each. The three states were:

- State 1 Machine generated packets \([0,0.06]\) seconds gamma distribution (equation 3.32)
- State 2 Active typing \((0.6,0.355]\) seconds gamma distribution (equation 3.32)
- State 3 Pauses \((0.355,\infty]\) seconds Pareto distribution (equation 3.10)

The important result is that state three gives the distribution a much heavier tail than a negative exponential distribution. This heavy tail is typical of arrivals in self-similar processes.

2.4.2 Video

An analysis of traffic generated by variable-bit-rate coding of full motion video was conducted by Beran et al [9]. Twenty different coded video sequences were analysed using R/S (Hurst) analysis, variance-time analysis and periodogram analysis (these tests are described in section 7.1). In this way the results were assured of being independent of the video sequence and of the video coding algorithm used. The large data sets generated by video coding allow analysis over a large range of aggregation scales of the raw data.

It was found that the three techniques produced general agreement in the estimates of the self-similarity parameter, \(H\), of each of the video sequences. The periodogram analysis gave 95% confidence intervals for \(H\). Only one of the real video sequences had a confidence interval which included \(H = 0.5\). The remaining sequences had self-similarity parameters in a wide range, from around 0.6 to nearly 1.0. Analysis of a Markov-chain video sequence simulator performed for comparison purposes produced a confidence interval centred at \(H = 0.5\).

Although the long-range dependence appeared to be a characteristic of all the real video sequences, there appeared little consistency to the \(H\) parameter of the
sequences. The presence of long-range dependence was demonstrated to be a distinguishing factor separating real video sequences from simulated sequences. No conclusions about the impact of long range dependence on network performance were drawn in this paper.

2.4.3 Ethernet Traffic

Ethernet traffic is analysed in the original paper on self-similar network traffic by Leland et al [56]. This paper analyses four long traces of Ethernet traffic on the network at Bellcore. Each trace provided highly accurate time-stamp information on Ethernet packets over measurement periods ranging from 20.86 to 47.91 hours. The traces were taken on four different dates from August 1989 to February 1992, during which time considerable changes occurred in the Bell-core network. Because of this the results do not reflect a single network configuration. Also, different locations were used for recording.

2.4.4 Wide Area Traffic

A variety of different wide-area data protocol streams were analysed in a paper by Paxson and Floyd [74]. It was found that most traffic types did not conform to a conventional Poisson traffic model, but rather showed self-similarity. Some traffic events, however, could be modelled as Poisson streams; these included TELNET session arrivals and RLOGIN traffic. In general it was found that events that involve a new user for each arrival, and hence a theoretically infinite user population, do create memoryless independent arrivals and so form a Poisson arrival stream. However, many types of traffic, including most machine generated data and traffic generated by a user during a session were not memoryless and so not Poisson. Except where there was an underlying regularity to the data stream (eg. hourly news feeds) most of these non-Poisson data streams had Pareto distributed inter-arrival times.

2.5 SELF-SIMILAR TRAFFIC MODELS

There are a number of mathematical models that produce self-similar sequences. These models are all candidates for selection when modelling network traffic that is self-similar, however a quick review demonstrates the mathematical complexity that makes analysis of systems with such input difficult.
2.5.1 Fractional Brownian Motions

Ordinary Brownian motion $B(t, \omega)$ is a real random function in time with independent Gaussian increments. Therefore $B(t_2, \omega) - B(t_1, \omega)$ has zero mean and variance $|t_2 - t_1|$ and the increments in non-overlapping time intervals are independent. Fractional Brownian motion (fBm) of exponent (self-similarity parameter) $H$ has increments with a Gaussian marginal distribution, zero mean and variance of $|t_2 - t_1|^{2H}$. Alternatively an fBm can be viewed as a moving average of $B(t, \omega)$ where past increments are weighted by $(t - s)^{H-1/2}$. This can be expressed by the Holmgren-Riemann-Liouville fractional integral [65]

$$B_H(t, \omega) = \frac{1}{\Gamma(H + \frac{1}{2})} \int_0^t (t-s)^{H-1/2} dB(s, \omega) \quad (2.11)$$

Where $H$ is any positive number. This definition is highly dependent on the origin. A symmetric definition for the increment of the fBm can be written as [65]

$$B_H(t_2, \omega) - B_H(t_1, \omega) = \frac{1}{\Gamma(H + \frac{1}{2})} \left\{ \int_{-\infty}^{t_2} (t_2 - s)^{H-1/2} dB(s, \omega) - \int_{-\infty}^{t_1} (t_1 - s)^{H-1/2} dB(s, \omega) \right\} \quad (2.12)$$

Here, $0 < H < 1$. The denominator $\Gamma(H + \frac{1}{2})$ insures that when $H = 1/2$, the fractional integral becomes an ordinary repeated integral.

2.5.2 Fractional Gaussian Noise

Just as ordinary Brownian motion does not have a true derivative, fBm is not differentiable. However, ways exist of finding generalised differentials of Brownian motions, known as white Gaussian noises. Similarly, generalised differentials of fBms can be found, known as fractional Gaussian noises. These have autocorrelations $r(k)$ which are of the form:

$$r(k) = \frac{1}{2}(|k + 1|^{2H} - 2|k|^{2H} + |k - 1|^{2H}) \quad (2.13)$$

which, for $0.5 < H < 1$ fulfills the requirements of the definition of long-range dependence (equation 2.4).
2.5.3 Fractional ARIMA Models

Fractional ARIMA (Auto Regressive Integrated Moving Average) processes are a generalisation of the Box Jenkins ARIMA \{p, d, q\} model [13] where \( d \) is allowed to take non-integer values. It is found that for \( 0 < d < 1/2 \) then the process is asymptotically second-order self-similar with \( H = d + 1/2 \). An ARIMA \{p, d, q\} process is defined by:

\[
\Phi(B)\nabla^d x_k = \Theta(B)\varepsilon_k
\]  

(2.14)

where \( \Phi(B) \) and \( \Theta(B) \) are polynomials of order \( p \) and \( q \) respectively, in the backward shift operator \( Bx_k = x_{k-1} \) and \( \varepsilon_k \) is a white noise process, i.e. a sequence of i.i.d. random variables with mean zero and variance \( \sigma^2 \). The fractional differencing operator \( \nabla^d \) is defined in terms of a binomial series [46]:

\[
\Delta^d = (1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k
\]

\[= 1 - dB - \frac{1}{2}d(1 - d)B^2 - \frac{1}{3}d(1 - d)(2 - d)B^3 - \ldots \]

(2.15)

The main advantage of fractional ARIMA processes over fractional Gaussian processes is that they allow considerable flexibility in terms of defining the short term correlation behavior of a model, whereas a fractional Gaussian process is essentially fixed in this regard.

2.5.4 Pareto Inter-Arrival Sequences

A count process generated from a sequence of Pareto distributed inter-arrival times, such as the one given in figure 2.5, demonstrate the features expected of self-similar sequences. Models based on this distribution are discussed in depth in section 3.2.

2.6 DISCUSSION

Traffic on cellular systems is likely to be more volatile than on conventional fixed telephone networks [57]. Therefore conventional telephone traffic models may not
be adequate to describe cellular network traffic. This is because the movement behaviour of subscribers and the higher calling rates of mobile telephony users introduce new causes of variation in traffic rates. Such variability is likely to become more exaggerated as traffic rates grow and cell sizes diminish.

The studies described in section 2.4 show that data packet arrival rates, at least, are likely to show self-similarity. This alone is sufficient cause to investigate the effect of this characteristic on radio channel capacity. The ISDN study [68] directly relates human behaviour to the generation of heavy tailed inter-arrival distributions. Poisson arrivals occur when all the events are memoryless. This relies on a theoretically infinite user population, an assumption which may not be good for small cells. The asymptotic behaviour of Pareto distributions are frequently seen in real world data because they are invariant under a wide range of filters, something that is true of no other distribution [64]. Although it would be possible to simply assume that a model with self-similar arrival processes would be the best description of cellular network traffic, it is important to analyse real data to compare the accuracy of such a traffic model with more conventional models.

The analysis of time series for self-similarity is best achieved with very long data series so that features can be examined over a wide range of time scales by aggregation of the original data [7]. Short series only allow the characteristics to be studied over a small range of aggregation making it difficult to determine whether the sequence has converged to its final state.
Chapter 3

INTER-ARRIVAL TIME DISTRIBUTIONS

One of the simplest ways to describe network traffic is to give the probability distribution of the inter-arrival time between transmission units, be they calls, packets or cells. This can be a complete description if inter-arrival times between different packets are independent. The distribution most commonly used in this way is the negative exponential. Self-similar processes are, however, commonly associated with so-called "heavy tailed" inter-arrival distributions. Heavy tailed distributions have the property that

$$P(T \geq t) \sim t^{-\phi} h(t) \quad \text{as} \quad t \to \infty$$

(3.1)

where $0 < \phi < 2$ and $h(t)$ is slowly varying at infinity. The Pareto distribution is the best known and most commonly used heavy tailed distribution. Although the inter-arrival times may be independent, the heavy tail of the distribution leads to high correlations in the count process of arrivals per unit time. Heavy tailed behaviour can be approached by appropriate selection of the parameters of the Gaussian distribution. This approach may be useful for modelling traffic that is "nearly" self-similar.

3.1 THE EXPONENTIAL DISTRIBUTION

The negative exponential distribution in its most common form is given by the probability density function (pdf)

$$p(t) = \lambda e^{-\lambda t}$$

(3.2)
The parameter $\lambda$ is the reciprocal of the mean of the distribution ($E(t) = \frac{1}{\lambda}$) and the distribution variance is given by $\sigma^2(t) = \frac{1}{\lambda^2}$ [80].

This distribution is a special case of the gamma distribution. It has the significant advantage of generally simplifying mathematics in analysis and allowing explicit results to be obtained. For this reason it is often chosen to as a representative distribution even when another would give a more accurate fit.

The significance of the exponential distribution to communications is that as an inter-arrival time it describes random independent arrivals, the Poisson arrival process. This is an accurate model of traffic on conventional wireline telephone networks and has a long history of application in network analysis [80].

To see the random nature of Poisson arrivals, consider a time interval $\delta t$ sufficiently short that the probability of an arrival in that interval is simply proportional to its length, i.e. $P(\text{arrival in } \delta t) = \lambda \delta t$ where $\lambda$ is a constant. Therefore the probability of no arrival in the interval is $1 - \lambda \delta t$. If every short time interval is independent of all others then the probability of no arrivals in any longer time period $t$ is given by

$$P(\text{no arrival in time } t) = (1 - \lambda \delta t)^{t/\delta t} \quad (3.3)$$

By comparing this with the definition of an exponential

$$e^{ax} = \lim_{k \to 0} (1 + ak)^{a/k} \quad (3.4)$$

it can be seen that the limit of this distribution as $\delta t$ approaches 0 becomes

$$P(\text{no arrival in time } t) = \lim_{\delta t \to 0} (1 - \lambda \delta t)^{t/\delta t} = e^{-\lambda t} \quad (3.5)$$

Since the period $t$ is between arrivals, then the Cumulative Probability Function (CDF) of the inter-arrival time becomes

$$Pr[T \leq t] = P(t) = 1 - e^{-\lambda t} \quad (3.6)$$
The derivative of this is the pdf as given in equation 3.2. The mean of this distribution is given by $E[t] = \frac{1}{\lambda}$ and the variance by $\sigma^2 = \frac{1}{\lambda^2}$.

One of the most useful features of this distribution is its so called memoryless (or Markovian) property. At any observation point the distribution of the time until the next arrival time is the same as the inter-arrival time distribution regardless of the elapsed time since the last arrival. This is most simply seen by noting that in the above derivation no restrictions were placed on the start time of the period $t$. Alternatively, if the inter-arrival time is $t$, the distribution of the remaining time $z$ after some period $y$ has already elapsed can be found by conditional probability as

$$p(z) = \frac{P(t = z + y)}{P(t \geq y)} = \frac{\lambda e^{-\lambda(z+y)}}{e^{-\lambda y}} = \lambda e^{-\lambda z}$$

which is identical to the distribution of the inter-arrival period length.

The primary assumption of exponentially distributed inter-arrival times is that all arrivals are completely independent. This is often based on the assumption that every new arrival comes from a new source, otherwise it would be linked to previous arrivals from that source. This is generally referred to as the infinite population assumption. In practice it means that the population needs to be sufficiently large that inter-arrival correlations are negligible.

Another possible source of error occurs when the system involves sources that sometimes fail at their attempt to communicate and then retry until successful. This may be in a telephone system where blocked callers re-attempt to place their call or in a random access channel, where transmissions may be blocked or collide and so the station attempts a re-transmission. These retry attempts clearly are correlated and so normally, their analysis will assume that the delay between each attempt is sufficiently long and randomised that correlations can be ignored [32].

3.2 THE PARETO DISTRIBUTION

The Pareto distribution is named for Vilfredo Pareto, a Swiss professor of economics (born in Italy) [49]. Pareto's original law was published in 1897 and
described the distribution of incomes over a population. It stated that the number of people $N$ with an income greater than $x$ was given by

$$N = Ax^\alpha$$

where $A$ and $\alpha$ are positive constant parameters ($\alpha$ is called either Pareto’s constant or a shape parameter). Pareto felt his law was both universal and inevitable, despite the differing taxation, social and political conditions of the populations to which it was applied. It was controversial and has been refuted several times by economists since. However, many different occurrences of such hyperbolic distributions have been found, notably by Zipf [90], after whom the discrete version of this distribution is named.

The modern Pareto distribution has several different forms. The original law survives as the Pareto distribution of the first kind, given by ($k$ and $\alpha$ are both positive constants)

$$Pr[T \leq t] = 1 - \left(\frac{t}{k}\right)^{-\alpha} \quad t \geq k$$

The Pareto distribution of the second kind is this distribution shifted to the origin. It is given by the pdf

$$p(t) = \frac{\alpha k^\alpha}{(t+k)^{\alpha+1}} \quad t \geq 0$$

The corresponding CDF is

$$Pr[T \leq t] = P(t) = 1 - \left(\frac{t+k}{k}\right)^{-\alpha} \quad t \geq 0$$

Here, $\alpha$ is the distribution shape parameter, $k$ is the location parameter and both are positive real numbers. This is the distribution used throughout this work since non-zero probability of very short inter-arrival times is necessary to describe bursts of arrivals. Additionally, a non-zero minimum inter-arrival time would require some mechanism to achieve it. This is possible for successful
transmissions where arrivals are separated by at least the packet length (service
time) but it is not an appropriate model for arrivals from multiple sources to a
random access system or to a queue.

The \(i\)th moment of the Pareto distribution only exists for \(i \leq \alpha\). The mean
is given by \(E[t] = k/(\alpha - 1)\) for \(\alpha > 1\).

The Pareto distribution is an empirical distribution in that it was developed
to describe observed data rather than derived from an underlying theoretical
situation. This means it lacks the power to explain the phenomena that give rise
to an observed process. It also means that its application to communications
system arrivals makes no assumptions about the source population or their
behaviour. However Mandelbrot in [60] showed that there are a number of filtering
processes such as choosing the maximum from a set, taking a weighted mixture
and aggregation under which asymptotically hyperbolic, and in particular
Paretian, distributions are invariant, apart from scaling. He also shows that, in
order to be invariant under such filtering, a distribution must be asymptotically
hyperbolic. Mandelbrot's conclusion is that a population's distribution will only
be observed clearly if it is hyperbolic in nature as otherwise the occurrence of
such filtering processes means that it will become chaotic.

It is possible to explain a Pareto distribution as a mixture of exponential
distributions [49], where the inverses of the means of these distributions have a
gamma distribution, i.e. if

\[
P(T \leq t \mid \lambda = x) = 1 - e^{-tx}
\]

and the standard (two parameter) form of the distribution is given by [49] (the
gamma function, \(\gamma(\alpha)\) is defined in section 3.4, equation 3.33)

\[
p(\lambda = x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\gamma(\alpha)}
\]

then by conditional probability

\[
P(T \leq t, \lambda = x) = p(T \leq t \mid \lambda = x)p(\lambda = x) = (1 - e^{-tx})\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\gamma(\alpha)}
\]
Integrating over all possible values of $x$ gives

$$P(T \leq t) = \int_0^\infty (1 - e^{-tx}) \frac{x^{\alpha-1}e^{-x\beta}}{\beta^\alpha\gamma(\alpha)} dx
= \frac{1}{\beta^\alpha\gamma(\alpha)} \left[ \int_0^\infty x^{\alpha-1}e^{-x\beta} dx - \int_0^\infty e^{-tx}x^{\alpha-1}e^{-x\beta} dx \right]
= 1 - (\beta t + 1)^{-\alpha}
= 1 - \left( \frac{t - 1/\beta}{1/\beta} \right)^{-\alpha} \tag{3.15}$$

This is a Pareto distribution of the second kind with location parameter $1/\beta$ and shape parameter $\alpha$, both taken from the gamma distribution.

![Figure 3.1 Poisson and Pareto Probability Distribution Functions. Parameters: Poisson $\lambda = 2$, Pareto $\alpha = 1.2, k = 0.4$ and $\alpha = 1.8, k = 1.6$](image)

A comparison of the pdfs of an exponential distribution (equation 3.2) and two Pareto distributions of the second kind (equation 3.10) is shown in figure 3.1. Each of these distributions has a mean of two. This figure shows that the Pareto distribution does have higher probabilities for very short inter-arrival times and so can produce the bursts of arrivals that generally characterise self-similar arrival
processes. The other significant feature of the Pareto distributions is the heavy tail where the probabilities of long inter-arrival times decay to zero much slower than exponentially. This is shown by the plot of the log of the pdf given in figure 3.2. This figure also demonstrates that the Pareto distribution does have a finite value at \( t = 0 \).

![Figure 3.2 Poisson and Pareto Probability Distribution Functions - Log Scale](image)

For Pareto spaced arrivals, the pdf that the next arrival will occur in a further time \( t \) given that a time \( z \) has elapsed since the last arrival is given by

\[
p(t) = \frac{p(t + z)}{p((t + z) > z)} = \frac{\alpha k^\alpha (t + z + k)^{-\alpha - 1}}{(z + k)/k} = \frac{\alpha (k + z)^\alpha}{(t + z + k)^{\alpha + 1}}
\]

This is still a Pareto distribution with the same shape parameter \( \alpha \), but with new location parameter \( z + k \) and so has mean \((k + z)/(\alpha - 1); \alpha > 1\). This is sometimes referred to as the semi-memoryless property of the Pareto distribution [49].
3.3 PARETO SPACED ARRIVALS

The Pareto distribution has been fitted to inter-arrival times observed on live networks in [72] and [68].

An argument is given by Mandelbrot in [61] that a sequence with inter-arrival times given by equation 3.17 is truly self-similar if $0 < \theta < 1$. This argument is based on the invariant shape of the distribution under scaling operations. Since it has both a hyperbolic "heavy" tail and an asymptote at the origin, the infinitely long inter-arrival times possible are balanced by the possibility of infinitely many arrivals in an arbitrarily short period of time. This distribution is a "relative frequency distribution" in that it needs to be divided by the total number of observations, $C$, to give a function which when integrated between zero and infinity equals one as a probability distribution function should. Since the area under the curve given is in fact infinite, the scaled curve is the limit of the Pareto distribution as the location parameter goes to zero, with $\alpha = \theta$.

$$P(u) = u^{-\theta} \quad (3.17)$$

One possible interpretation of the limit of the Pareto distribution as $k \to 0$ is that it is the distribution as the observation interval increases to infinity. This is a possible argument for such a sequence being self-similar. However, there is considerable doubt as to whether such a limit is useful, or even valid [73]. A count process of Pareto spaced arrivals cannot be self-similar when $\alpha$ is less than one [74], since for $\alpha$ in this range, the mean of the Pareto distribution is infinite. With infinite mean inter-arrival time, the expected number of arrivals within any finite time interval is zero. Therefore the autocorrelation function must be zero everywhere and is therefore summable. The summable autocorrelation function implies that the process is not long range dependent and so it cannot be self-similar. The range of greatest interest for such a process is therefore $1 < \alpha < 2$ since for $\alpha > 1$ the mean is defined and finite and for $\alpha < 2$ the function fits the definition of a heavy tailed function (equation 3.1) and the hyperbolic tail is sufficiently heavy that the variance does not exist.

An analysis of the lengths of bursts and lulls of Pareto inter-arrival sequences given in [72] shows that as the bin size used for counting arrivals grows larger (i.e. under aggregation), the expected length (number of bins) of a lull stays constant, but the expected length of a burst grows faster than $\log$ (bin size). The conclusion reached was that the process is not truly self-similar as the process is continually...
changing under aggregation and the limiting sequence is effectively white noise. This analysis used the first form of the Pareto distribution (equation 3.9) and does not cover the second form. In particular the burst size growth will be slower when the minimum inter-arrival time is reduced to zero.

3.3.1 Aggregated Pareto Renewal Processes

The first use of the Pareto distribution to construct a self-similar process was by Mandelbrot [62]. His results were extended by Taqqu and Levy [83] and used by Leland et al [56] in attempting to provide a "phenomenological" explanation for the self-similar behaviour they observed in Ethernet traffic.

This explanation is based on a renewal process, which is a sequence of independent identically distributed random numbers that represent a series of time periods or renewal times, $U_0, U_1, U_2, ...$ [53]. The self-similar process requires that the renewal times are drawn from a distribution with a "heavy tail" such as the Pareto distribution with a shape parameter $1 < \alpha < 2$. A reward sequence $W_0, W_1, W_2, ...$ with zero mean and finite variance represents the value of each renewal in the process. A delayed renewal sequence is derived from the renewal times as

$$S_k = S_0 + \sum_{j=1}^{k} U_j \quad k \geq 0 \quad (3.18)$$

Where $S_0$ is chosen so that the sequence $(S_k)$ is stationary. The renewal reward process $W = (W(t); t = 0, 1, 2, ...)$ is then defined by

$$W(t) = \sum_{k=0}^{t} W_k I_{(S_k-1, S_k]}(t) \quad (3.19)$$

$I_A(.)$ is the indicator function on the set $A$ so that $I_A(x) = 1$ if $x$ is in the set $A$ and $I_A(x) = 0$ otherwise. By aggregating $M$ i.i.d. copies $W^{(1)}, W^{(2)}, ..., W^{(M)}$ of $W$ the process $W^*$ is obtained where

$$W^*(T, M) = \sum_{t=1}^{T} \sum_{m=1}^{M} W^{(m)}(t) \quad (3.20)$$
In [62] and [83] it is shown that for $T$ and $M$ both large and with $T << M$, $W^*$ behaves like fractional Brownian motion and hence its increment process behaves like fractional Gaussian noise.

Leland et al [56] proposed that the rewards took the values $+1$ or $-1$ corresponding to hosts transmitting or not transmitting and the renewal periods corresponded to the lengths of the active and inactive periods of individual hosts. Then the above model could be used to explain their observations of self-similar traffic on an Ethernet network. There is increasing evidence that individual hosts' active/inactive periods do have heavy tailed distributions [76, 17].

### 3.3.2 Proof of Self-Similarity of Pareto Spaced Arrivals

Evidence of the self-similarity of arrivals may be obtained by applying techniques such as Hurst parameter analysis and variance analysis (see chapter 7). However, despite this evidence, proof that such arrivals constitute a self-similar process has not been available until recently.

The proof that a process with Pareto distributed inter-arrival times constitutes a self-similar process is due to Gordon [38]. The sequence of arrival counts can be analysed by using the Laplace Transform, $p_L(s)$, of the Pareto density function, $p(x) = dP(x)/dx$. This is found in terms of the incomplete gamma function $\gamma(a, x)$ which is then expanded in a power series valid for $|x| < \infty$. Thus we obtain

\[
\begin{align*}
    p_L(s) &= \int_0^\infty e^{-st} \frac{\alpha k^\alpha}{(k + t)^{\alpha + 1}} \, dt \\
         &= \alpha(ks)^\alpha e^{ks} \gamma(-\alpha, ks) \\
         &= \alpha \gamma(-\alpha) \left[ (ks)^\alpha e^{ks} - \sum_{n=0}^{\infty} \frac{(ks)^n}{\gamma(n-\alpha+1)} \right] \\
\end{align*}
\]

The correlation density $c(t)$ is defined as the probability that an arrival occurs at time $t$ given that one occurred at time $t = 0$. If $\lambda$ is the mean arrival rate then for a time interval, $dt$, sufficiently small that only one arrival may occur the probability of an arrival in any such interval is $\lambda dt$. Hence from the definition of the normalised autocorrelation function $r(t) = \text{cov}[X(\tau), X(\tau + t)]/\text{var}[X(\tau)]$ it follows that $r(t) = c(t) - \lambda$. Using standard arguments it can be shown that [79]
3.3 PARETO SPACED ARRIVALS

\[ c(t) = \int_0^t c(t-x)p(x)dx + \delta(t) \]  \hspace{1cm} (3.22)

This equation can be solved by taking the Laplace Transform to obtain

\[ c_L(s) = \frac{1}{[1 - p_L(s)]} \quad \text{(3.23)} \]

The power spectrum \( S(f) \) of the sequence of arrival counts is defined as:

\[ S(f) = \pi^{-1} \int_0^\infty c(t)cos(2\pi ft)dt = (2\pi)^{-1} [c_L(2\pi f) + c_L(-2\pi f)] \]  \hspace{1cm} (3.24)

Substituting 3.23 and 3.21 into 3.24 provides the solution that

\[ S(f) \sim \begin{cases} f^{-\alpha} & 0 < \alpha < 1 \\ f^{\alpha-2} & 1 < \alpha < 2 \\ \text{constant} & \alpha > 2 \end{cases} \]  \hspace{1cm} (3.25)

Clearly, for \( 0 < \alpha < 1 \) and \( 1 < \alpha < 2 \), the power spectrum has the form \( \frac{1}{f^k} \), \( 0 < k < 1 \) typical of self-similar series.

The autocorrelation function \( r(t) \) of the arrival count sequence is obtained for the sequence via the inverse Fourier transform of the power spectrum. This autocorrelation function is related to the correlation density by \( r(t) = c(t) - \lambda \), where \( \lambda \) is the mean arrival rate and

\[ r(t) = \int_{-\pi}^{\pi} S(f)e^{2\pi ift}df \]  \hspace{1cm} (3.26)

The value of \( r(t) \) as \( t \to \infty \) is governed by the behaviour of \( S(f) \) when \( f \) is
close to 0. Specifically, by the stationary phase method [75] it can be shown that:

\[
 r(t) \sim \begin{cases} 
 t^{\alpha-1} & 0 < \alpha < 1 \\
 t^{-\alpha} & 1 < \alpha < 2 \\
 t^{-1} & \alpha > 2 
\end{cases} \tag{3.27}
\]

Equation 3.27 shows that for \(0 < \alpha < 1\) and \(1 < \alpha < 2\), the autocorrelation functions decay hyperbolically and are non-summable (c.f. equation 2.1). Therefore, an arrival count process with Pareto distributed inter-arrival times has long range dependence when shape parameter is in these ranges. In order to show that in addition it is self-similar the following result is used [16].

\[
v_m = v \left( \frac{1}{m} + \frac{2}{m^2} \sum_{n=1}^{m} (m - n) r_k \right) \tag{3.28}
\]

This gives the variance, \(v_m\), of the block average of packet counts, \(X_k^{(m)}\), in terms of the variance, \(v\), and autocorrelation coefficients, \(r_k\), of the original packet counts, \(X_k\). The block average is defined as \(X_k^{(m)} = (X_{km} - m + 1 + \ldots + X_{km})/m\) to agree with the definition of self-similarity given in section 2.3.2. Since equation 3.27 shows that for \(0 < \alpha < 1\) and \(1 < \alpha < 2\), \(r_k \sim k^{-\zeta}\), where \(0 < \zeta < 1\), this can be substituted into equation 3.28 to show that \(v_m \sim m^{-\zeta}\) as \(m \to \infty\). This demonstrates that the process has slowly decaying variances for \(\alpha\) in these ranges.

Finally, consider the combining of two consecutive packet count intervals of length \(m\) into a single interval of length \(2m\). The definition of the autocorrelation function gives

\[
r^{(2m)}(k) = \frac{2r^{(m)}(2k) + r^{(m)}(2k - 1) + r^{(m)}(2k + 1)}{2[1 + r^{(m)}(1)]} \tag{3.29}
\]

Similarly, it possible to show from first principles that [38]

\[
v_{2m} = \frac{v_m}{2} \left[ 1 + r^{(m)}(1) \right] \tag{3.30}
\]
Considering that \( v_m \sim m^{-\xi} \) asymptotically, this implies that \( 1 + r^{(m)}(1) \sim 2^{1-\xi} \). This can be substituted into equation 3.29 to show that \( r^{(m)}(k) \sim k^{-\xi} \). Comparing this to equation 3.27, it can be seen that \( r^{(2m)}(k)/r^{(m)}(k) \to 1 \) as \( k \to \infty \), thus satisfying equation 2.5 and so proving that the packet count process is self-similar.

By comparing \( r(t) \) in 3.27 with the definition of long range dependence given in section 2.3.1, the self-similarity parameter for Pareto spaced arrivals with shape parameter \( 1 < \alpha < 2 \) can be found.

\[
H = \frac{3 - \alpha}{2} \tag{3.31}
\]

This is an intuitively pleasing result as when \( \alpha = 2 \), \( H = 1/2 \), which indicates that the process is no longer self-similar. Also, when \( \alpha = 1 \), \( H = 1 \), which is the largest possible value in the self-similar range of \( H \).

### 3.4 LARGE VARIANCE GAMMA DISTRIBUTION

In [12] Bhatnagar analyses the G/M/1 queue using same techniques that Gordon uses in [38] except that instead of using the Pareto distribution for inter-arrival times, Bhatnagar uses the gamma distribution. The gamma distribution is given by the pdf

\[
p(t) = \frac{r \lambda(r \lambda t)^{r-1} e^{-r \lambda t}}{\gamma(r)} \quad t \geq 0 \tag{3.32}
\]

where \( \lambda > 0, r > 0 \) and the gamma function, \( \gamma(r) \) is given by

\[
\gamma(r) = \int_0^\infty y^{r-1} e^{-y} dy \tag{3.33}
\]

The mean of this distribution is given by \( E[t] = 1/\lambda \) and the variance by \( \sigma^2(t) = 1/(r \lambda^2) \). If \( r = 1 \) this distribution simplifies to the negative exponential (equation 3.2). For any given mean inter-arrival time the variance can be controlled independently by the parameter \( r \) and in particular it can be made arbitrarily large by making \( r \) less than one. The gamma distribution pdf (equation 3.32) and the pdfs of an exponential distribution (equation 3.2) and a Pareto
distribution (equation 3.10) are shown in figure 3.3. Note that for \( r < 1 \) this distribution has an asymptote at \( t = 0 \).

Bhatnagar argues that in reality it will be impossible to observe real traffic inter-arrival times with infinite variance and all real observations will be at most almost self-similar or "pseudo-fractal" with a large, but finite variance in inter-arrival times. Therefore, the gamma distribution is, he claims, a good candidate for modelling such traffic.

The decay of the gamma distribution is the product of a hyperbolic and an exponential term and so it does not have the heavy tail commonly associated with self-similar arrivals. This can be seen in figure 3.4 where representative Poisson, Pareto and gamma distributions are plotted on a log scale. This shows that, with \( r \) in this range, the tail of the gamma distribution although much heavier than that of the exponential distribution, still does not have the hyperbolic shape of the Pareto distribution. The exponential term in the gamma distribution pdf (equation 3.32) means that it can never be a heavy tailed distribution as defined in equation 3.1, whatever value \( r \) takes. Since this distribution has an asymptote at \( t = 0 \) for \( r \) in the range of interest, an arrival process that has this inter-arrival time distribution will have a very large number of extremely small inter-arrival
3.5 DISCUSSION

The classical exponential inter-arrival distribution produced truly random arrivals. This represents a theoretical idealisation. Real world arrivals are always correlated in some way. However for sufficiently large populations and sufficiently randomised behaviour, particularly if re-transmissions are involved, the classical model often provides a good fit sufficient for accurate analysis. The exponential distribution also has the useful advantage of good mathematical tractability.

The Pareto distribution has often been associated with self-similarity and network traffic with such inter-arrival times has been recorded from live networks. Only recently has it been proved that an arrival process with independent Pareto distributed inter-arrival times is self-similar. This distribution, however, has the disadvantage that it is not based on any underlying physical model. The first form of the Pareto distribution has been used in network studies of successful arrivals. This will produce very heavy bursts of arrivals. It would be necessary to confirm that this characteristic behaviour is appropriate before use of the large variance gamma distribution could be considered. No evidence linking any real traffic to the Gamma distribution for inter-arrival times has been presented.

Figure 3.4 gamma Probability Distribution Function - Log Scale
transmissions on wireline networks. However it is not proven to be self-similar and analysis strongly indicates that it is not \cite{72}. More importantly, this for has a minimum inter-arrival time which is not appropriate for modelling arrivals to the network from independent sources since a minimum spacing would require a co-ordination mechanism. Therefore the second form of the Pareto distribution has been used throughout this work. The Gaussian distribution can be used to approximate heavy tailed behaviour associated with self-similarity if its parameters are chosen appropriately. This may be useful for modelling nearly self-similar traffic. Other approaches to this situation such as using a truncated Pareto distribution may also be valid, but possibly more difficult to analyse. The large variance range of the gamma distribution also has an asymptote at zero. This will produce heavier bursts of arrivals than the other distributions considered here. No evidence has been presented relating the gamma distribution or the heavy burst characteristic to real network traffic patterns.
Chapter 4

ANALYSIS OF QUEUES WITH PARETO ARRIVALS

4.1 QUEUEING MODELS

Queues are an inherent part of many networks. In packet-switched networks, the packets arriving at any entry or intermediate node have to be stored whenever the appropriate outgoing link is not immediately available. So this storage is modelled as the buffer for a server, which is the transmission link and so the combination becomes a simple queue as in figure 4.1. Analysis of these queueing models is fundamental to the analysis and understanding of network systems.

![Figure 4.1 Single Server Queue Model](image)

Self-similar traffic models are useful because they are able to encapsulate a wide range of time scales in a parsimonious fashion. However, such complex behaviour, often characterised by its remarkably bursty nature, leads to difficulties in analysing the demands for resources such traffic places on networks. This analysis is needed because it is expected that self-similar traffic, because of its bursty nature, will require much higher levels of network resources than is predicted by conventional traffic models in order to provide the same grade of service.

Norris, in [69], studied a continuous storage model with fractional Brownian (self-similar) input. Using this he was able to derive a lower bound for the storage requirements of such a system. Erramilli et al [25] performed some very revealing
queueing simulation experiments to demonstrate the dramatic effect of long range dependence (LRD) on queue delays and queue length. The input processes used were actual measured traces of Ethernet LAN traffic. However, while clearly demonstrating the effect of LRD, their results are difficult to generalise or apply to other networks.

More recently, Gordon [38] established that an arrival process with Pareto distributed inter-arrival times (a Pareto spaced arrival process) could in fact be self-similar. This means that a simple single server queue with Pareto spaced arrivals and exponentially distributed service times forms a simple, and hence attractive, model for analysis. Exponential service times are not an accurate model for many kinds of packet network traffic, but do provide a networking standard that allows direct comparison of the obtained results with conventional models (i.e. the M/M/1 queue) and, more importantly, they do allow useful results to be reached.

Within this chapter, the P(areto)/M/1 queue and the P/M/1/N queues are analysed as semi-Markov processes to find blocking probabilities and expected waiting times. The calculated blocking probabilities are compared with those given by the infinite buffer approximation that can be used for Poisson arrivals and are then used to investigate the effect of increasing queue length. Finally, simulation is used to compare the performance of deterministic service queues to the analytic results for queues with exponentially distributed waiting times.

Queues are commonly described by the notation developed by British statistician D.G. Kendall [80]. The Kendall notation uses a system of the form A/B/C/D, where A describes the arrival distribution, B the service distribution, C denotes the number of servers and D the storage (servers and buffer) of the queue. Where the D is omitted it is taken to be infinity. Commonly used symbols for the arrival and service distributions include M(arkov) for the Poisson distribution, D for the deterministic and G for the General distribution. P is used in this thesis for the Pareto distribution.

4.2 QUEUE ANALYSIS

The common assumption of negative exponential service times has been retained here, primarily because the memoryless property of this distribution enables the analysis of the queue to be approached through the embedded Markov chain technique [40]. This is appropriate since the self-similarity arises due to the population dynamics and so there is no basis for applying such a model to the service time. The negative exponential service time model is a suitable general
model. To model data transmission, it would be better to use a deterministic (e.g. ATM cells) or a bi-modal (e.g. Ethernet packets) service time distribution, but both these distributions lack the memoryless property necessary to produce the renewal points used for Markov chain analysis.

Standard queue analysis notation is used. Packets arrive with a mean rate of \( \lambda \) and are serviced at a mean rate of \( \mu \); both dimensioned as packets/time. The load on the server is defined as \( \rho = \lambda / \mu \).

### 4.2.1 The Infinite Buffer Queue

With Pareto distributed inter-arrival times and exponential service times, we have a specific example of the G/M/c queue. In this queue, the renewal points are the instants immediately before each arrival and the state of the queue is taken to be the number of customers in the queue at this renewal point \([40]\). Assuming that there is one server (i.e. a G/M/1 queue), the probability of \( n \) services during an inter-arrival time, \( b_n \), can be found by integrating the probability of \( n \) exponential departures with mean rate, \( \mu \), over the Pareto arrival time distribution, \( p(t) \)

\[
b_n = \int_{0}^{\infty} \frac{e^{-\mu t} (\mu t)^n}{n!} dP(t) \tag{4.1}
\]

The probability of a transition from state \( i \) to a state \( j \), where \( i \geq j + 1 \) and \( j > 0 \) is \( b_{i \rightarrow j+1} \), (the "+1" accounts for the single arrival during each renewal epoch).

Thus a transition probability matrix \( P \) can be constructed, the columns of which are comprised of the probability of a transition from the corresponding state to all other states \([40]\). For example column 0, corresponding to state 0 contains two non-zero probabilities, the probability of no departures giving a transition to state 1 and the probability of at least one departure leaving the
queue in state 0.

\[
P = \begin{bmatrix}
1 - b_0 & 1 - \sum_{k=0}^{1} b_k & 1 - \sum_{k=0}^{2} b_k & \ldots \\
 b_0 & b_1 & b_2 & \ldots \\
 0 & b_0 & b_1 & \ldots \\
 0 & 0 & b_0 & \ldots \\
 0 & 0 & 0 & \ldots \\
 \vdots & \vdots & \vdots & \ddots \\
 \vdots & \vdots & \vdots & \ddots \\
 \vdots & \vdots & \vdots & \ddots \\
 \end{bmatrix}
\] (4.2)

Since there is only one arrival between successive renewal points, the queue can transition at most one state higher and only the first diagonal below the leading diagonal of \( P \) is non-zero. Since there can be any number of departures (up to the number of customers in the system - equal to the state+1), all positions above the leading diagonal are non-zero.

The steady state probability vector that an arrival to the queue finds \( i \) customers in the system is \( q = \{q_i\} \) and this can be found from the transition probability matrix as the solution to

\[
Pq = q
\] (4.3)

with the normalizing condition that

\[
\sum_{i=0}^{\infty} q_i = 1
\] (4.4)

The solution to this is found by assuming a geometric distribution of the state probabilities. Thus there is constant ratio between state probabilities, denoted as \( \sigma \), so that \( \sigma q_i = q_{i+1} \), and then from equation 4.3, for \( i \geq 1 \) we can write [40]
4.2 QUEUE ANALYSIS

\[ q_i = \sum_{k=0}^{\infty} q_{i+k-1} b_k \]
\[ \sigma q_{i-1} = \sum_{k=0}^{\infty} \sigma^k q_{i-1} b_k \]
\[ \sigma = \sum_{k=0}^{\infty} \sigma^k b_k \]
\[ = \sum_{k=0}^{\infty} \sigma^k \int_{0}^{\infty} \frac{e^{-\mu t} (\mu t)^k}{k!} dP(t) \]
\[ = \int_{0}^{\infty} e^{-\mu t (1 - \sigma)} dP(t) \]  
\[ (4.5) \]

This last integral can be recognised as the Laplace transform of the inter-arrival time probability density distribution evaluated at \( \mu(1 - \sigma) \), denoted as \( p_L(s) \), so we have.

\[ \sigma = p_L(\mu(1 - \sigma)) \]  
\[ (4.6) \]

Since the infinite sum of these probabilities must be unity, it can be found that \( q_0 = 1 - \sigma \), so in general

\[ q_i = (1 - \sigma) \sigma^i \]  
\[ (4.7) \]

This is analogous to the M/M/1 steady state probability solution, given by

\[ p_i = (1 - \rho) \rho^i \]  
\[ (4.8) \]

where \( p_i \) is the general-time (as seen by an independent observer) probability of \( i \) in the system and \( \rho \) is the system load. The analogy can be extended to the analysis of the waiting time distribution by noting that in this case the required probabilities are those seen by the arrivals (i.e., \( q_i \)). For the M/M/1 queue these are equal to the general-time probabilities due to the unique properties of the Poisson arrivals. So the waiting time cumulative distribution for the P/M/1 queue is
and the mean wait is [40]

$$W_q = \frac{\sigma}{\mu(1 - \sigma)}$$  \hspace{1cm} (4.10)

The Laplace transform of the Pareto inter-arrival time distribution is given in equation 3.21 which can be substituted into equation 4.6 to obtain a complete steady state solution for the P/M/1/\infty queue. Plotted waiting time results are shown in figure 4.2 where the mean service time is normalised to one. This figure demonstrates the differences in the performance of the queue caused by the different arrival processes. The self-similar Pareto spaced arrivals experience significantly longer waiting times than Poisson arrivals. Decreasing the shape parameter, $\alpha$ of the Pareto spaced arrival process dramatically increases the waiting time.

Figure 4.2 Waiting Time for Infinite Buffer Queues
4.2 QUEUE ANALYSIS

4.2.2 The Finite Buffer Queue

A more realistic situation is the case of a finite buffer, a P/M/1/N queue. For a finite buffer queue (size N), the transition probability matrix \( P \) becomes finite sized \((N + 1 \times N + 1)\). Because a finite queue rejects new arrivals when it is full, the arrival that occurs immediately after a renewal point does not change the state of the system if it is already in state \( N \). Thus the transition probabilities from state \( N \) (column \( N \)) are the same as those from state \( N - 1 \) (column \( N - 1 \)), yielding

\[
P = \begin{bmatrix}
1 - b_0 & 1 - \sum_{k=0}^1 b_k & 1 - \sum_{k=0}^2 b_k & \ldots & 1 - \sum_{k=0}^{N-1} b_k & 1 - \sum_{k=0}^{N-1} b_k \\
0 & b_0 & b_1 & \ldots & b_{N-1} & b_{N-1} \\
0 & 0 & b_0 & \ldots & b_{N-2} & b_{N-2} \\
0 & 0 & 0 & \ldots & b_{N-3} & b_{N-3} \\
& & & & & \\
& & & & & \\
0 & 0 & 0 & \ldots & b_0 & b_0
\end{bmatrix}
\]

(4.11)

Since the last two columns of the transition probability matrix, \( P \), are identical it is singular and equation 4.3 does not have a unique solution. However, the extra condition given by equation 4.4 provides the extra information required to fully specify a unique solution.

The transition probabilities are found by substituting equation 3.11 into equation 4.1 giving

\[
b_n = \int_0^\infty e^{-\mu t} (\mu t)^n \frac{\alpha k^\alpha}{(k + t)^{\alpha+1}} \, dt
\]

(4.12)

This integral is closely related to the Pareto distribution Laplace transform of equation 3.21. The extra terms are constants, apart from \( t^n \). Multiplying a function by \( t^n \) is equivalent to differentiating its Laplace transform. The differential of the incomplete gamma function is given in [2], and so by applying Leibniz's
Theorem [2] we obtain

\[ b_n = \frac{\alpha \mu^n e^{\mu \beta}}{n!} \left[ (\mu \beta)^{\alpha} \Gamma(-\alpha, \mu \beta) + \binom{n}{1} (\mu \beta)^{\alpha-1} \Gamma(1 - \alpha, \mu \beta) \right. \\
+ \binom{n}{2} (\mu \beta)^{\alpha-2} \Gamma(2 - \alpha, \mu \beta) + \ldots + \binom{n}{r} (\mu \beta)^{\alpha-r} \Gamma(r - \alpha, \mu \beta) \\
+ \ldots + (\mu \beta)^{\alpha-n} \Gamma(n - \alpha, \mu \beta) \right] \]  

(4.13)

The transition probability matrix so found can then be used to obtain the state probabilities, \( q_i \) by solving equations 4.3 and 4.4.

Since the probabilities found by this procedure are the probabilities that the queue is in state \( i \) at the arrival of a new customer, \( q_N \) is the probability that an arriving customer finds the buffer full, i.e. the blocking probability.

As in the infinite buffer case, the cumulative waiting time distribution \( W_q(t) \) can be found by directly using the \( q_i \). Following the derivation in [40] we obtain

\[
W_q(0) = q_0 \\
W_q(t) = \sum_{n=1}^{N-1} \Pr (n \text{ departures in } \leq t \mid \text{ arrival found } n \text{ in system}) + W_q(0)
= \sum_{n=1}^{N-1} q_n \int_0^t \frac{\mu(x)^{n-1}}{(n-1)!} e^{-\mu x} dx + q_0
= \sum_{n=1}^{N-1} q_n \left[ 1 - \int_0^\infty \frac{\mu(x)^{n-1}}{(n-1)!} e^{-\mu x} dx \right] + q_0
= \sum_{n=1}^{N-1} q_n - \sum_{n=1}^{N-1} q_n \sum_{i=n}^{n-1} \frac{(\mu t)^i e^{-\mu t}}{i!} + q_0
= 1 - \sum_{n=1}^{N-1} q_n \sum_{i=n}^{n-1} \frac{(\mu t)^i e^{-\mu t}}{i!}
\]

(4.14)

This is a well behaved function that can be differentiated to give the waiting time probability density function, \( w_q(t) \) as

\[
w_q(t) = -\sum_{n=1}^{N-1} q_n \sum_{i=n}^{n-1} \left[ \frac{i(\mu t)^i e^{-\mu t}}{i!} - \frac{\mu(\mu t)^{i+1} e^{-\mu t}}{i!} \right]
\]

(4.15)
Integrating gives the expected value of this waiting time, \( W_q = E[t] \).

\[
W_q = \int_0^\infty t w_q(t) dt \\
= - \sum_{n=1}^{N-1} q_n \sum_{i=n}^{n-1} \frac{i}{\mu} - \frac{i + 1}{\mu} \\
= \sum_{n=1}^{N-1} \frac{n q_n}{\mu}
\]  

(4.16)

This is the expected result since arrivals find \( n \) customers in the queue with probability \( q_n \) and the mean service time per customer is \( 1/\mu \). Note however that this is the expected mean across all the customers arriving to the queue and since customers that are blocked do not join the queue, they are assessed as having zero waiting time and so have the effect of lowering the mean. The expected waiting time of customers that are not blocked is simply the found by dividing the waiting time experienced by all arrivals by the fraction that actually enter the queue, giving \( W_q/(1 - q_N) \)

4.3 RESULTS

The equations developed in section 4.2 have been used to calculate the blocking probabilities and waiting times for different queue sizes and shape parameters and to compare these with queue performance when Poisson arrivals are assumed. Figures 4.3 shows the blocking probabilities (from equations 4.3 and 4.4) and figure 4.4 show the waiting times (from equation 4.16) for a \( P/M/1 \) queue with a buffer size of five. Curves are given for arrivals with inter-arrival distributions using shape parameters of \( \alpha = 1.2, 1.5, \) and \( 1.8 \). The self-similarity parameters for these arrival processes are \( H = 0.9, 0.75 \) and \( 0.6 \) respectively. For all calculations the exponential mean service rate was kept constant at \( \mu \). The load value is then the reciprocal of the mean inter-arrival time which is given by \( k/(\alpha - 1) \). For each calculation this was adjusted by appropriately selecting the Pareto distribution location parameter, \( k \). The additional curve in each figure provides the results for a \( M/M/1 \) queue of the same size for comparison.

It can be clearly seen that for all three Pareto spaced arrival processes, the blocking performance of the queue is significantly worse than for the Poisson arrival process. Decreasing the value of the inter-arrival distribution shape parameter, \( \alpha \), (increasing the self-similarity parameter, \( H \), of the process) further
significantly increases the blocking at the queue. The expected waiting times follow the same trend at low loads with the low shape parameter Pareto spaced arrival process having the highest waiting time and the Poisson arrival process the lowest. At high loads, however, the waiting times for queues with Pareto spaced arrival processes reduce as a direct consequence of the rapidly increasing fraction of customers which are blocked and therefore have effectively zero waiting time. The dotted lines show the expected waiting times for non-blocked customers. These waiting times continue to increase as expected. Since, for finite sized queues, there is a maximum expected waiting time, equal to $N \mu$, the waiting times of non-blocked customers cannot increase unchecked, but must curve down to meet this "roof".

![Figure 4.3 Blocking Probability for Buffer Size 5.](image)

Figures 4.5 and 4.6 show the blocking probabilities (equations 4.3 and 4.4) and waiting times (equation 4.16) calculated for these arrival processes into single server queues with buffer lengths of 25. The results follow the same pattern with all the Pareto spaced arrival processes resulting in significantly higher blocking and longer expected wait times than the Poisson arrival process. Again, the lower the shape parameter of the processes, the worse the performance. In this case though, only the waiting time curve for $\alpha = 1.2$ is seen to reduce as large numbers
of arrivals are blocked.

The blocking probabilities (equations 4.3 and 4.4) and waiting times (equation 4.16) for a buffer of length 100, single server queue and the same four arrival processes are shown in figures 4.7 and 4.8. The pattern of the results is similar to the buffer size 25 case.

### 4.3.1 Infinite Buffer Approximation

The standard blocking probability formula for $M/M/1/N$ queues is $P_B = (1 - \rho)\rho^N/(1 - \rho^N)$. For $N$ large this can be closely approximated by $P_B \sim (1 - \rho)\rho^N$ which is the probability that an infinite buffer $M/M/1$ queue is in state $N$. We therefore compare the blocking probabilities for the queues of buffer size 100 with the probabilities that an infinite buffer queue with the same arrival process is in state 100 at an arrival time. This comparison is shown in figure 4.9 where the solid lines are the exact results and the broken lines are the infinite queue approximations. As expected, the approximation for the queue with Poisson arrivals is nearly exact however the approximations for the queues with Pareto spaced arrival processes are poor. Moreover the approximation for these queues
Figure 4.5 Blocking Probability for Buffer size 25.

Figure 4.6 Expected Waiting Time for Buffer size 25. solid lines - all customers, broken lines - non-blocked customers only.
4.3 RESULTS

Figure 4.7 Blocking Probability for Buffer length 100.

Figure 4.8 Expected Waiting Time for Buffer length 100, solid lines - all customers, broken lines - non-blocked customers only.
breaks down as the blocking probability rises. The state probabilities of the infinite queue are distributed geometrically (equation 4.7) and this breakdown occurs as the expected value of this distribution, $E[i] = \sigma/(1 - \sigma)$ exceeds the size of the finite buffer.

![Figure 4.9 Infinite Queue Approximation for Blocking Probability. Solid lines - exact results, broken lines - infinite buffer approximation](image)

**4.3.2 Effect of Queue Length**

Figure 4.10 shows the load at which the examined queues reach 0.05 blocking probability as a function of the queue length. This value was chosen to represent a transition from small blocking probabilities to large. Despite using a log scale for the buffer size, the graphs show a distinct curve of decreasing gradient indicating the decreasing gain in blocking performance as the buffer size is increased. The curves also clearly show the relatively worse performance as the shape parameter of the Pareto inter-arrival distribution is decreased.

Another view of the same effect is given in figures 4.11, 4.12 and 4.13 which show the relationship between buffer size and blocking probability at fixed loads of 0.5, 0.7 and 0.9 respectively. These figures show that increasing self-similarity of the arrival process dramatically reduces the gain in performance available through
increasing queue length. It is interesting to note that the relationship between buffer size and blocking probability is asymptotically a straight line when plotted on a log scale for all the input processes. The slopes of these plots are given in table 4.1. The table gives the ratio of the slope of the plot for that input process to the slope of the plot for Poisson arrivals at the same load. These figures show that, as the load increases, the slope values for the plots for all the Pareto spaced arrival processes decrease faster than the slope of the Poisson arrival plots, (the ratios get smaller). In other words, the higher the load the smaller the gain in blocking probability that is attainable, relative to the Poisson standard, from increasing buffer size.

4.4 SIMULATION OF P/D/1/N QUEUES

The analytical results presented here all involve queues with exponentially distributed service times, however, often in networks packets are of a fixed size. This is increasingly true with the growing popularity of Asynchronous Transfer Mode as a core network technology since it uses a fixed 54 octet transmission cell. Many of the traffic types that have been shown to be self-similar are commonly carried
Figure 4.11  Blocking vs Buffer length at Load = 0.5

Figure 4.12  Blocking vs Buffer length at Load = 0.7
Figure 4.13  Blocking vs Buffer length at Load = 0.9

<table>
<thead>
<tr>
<th>Load</th>
<th>Process</th>
<th>Slope</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>Pareto α=1.2</td>
<td>-0.0211</td>
<td>0.070</td>
</tr>
<tr>
<td></td>
<td>Pareto α=1.5</td>
<td>-0.0905</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Pareto α=1.8</td>
<td>-0.1368</td>
<td>0.455</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-0.3010</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>Pareto α=1.2</td>
<td>-0.0064</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>Pareto α=1.5</td>
<td>-0.0304</td>
<td>0.195</td>
</tr>
<tr>
<td></td>
<td>Pareto α=1.8</td>
<td>-0.0546</td>
<td>0.352</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-0.1550</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>Pareto α=1.2</td>
<td>-0.0032</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>Pareto α=1.5</td>
<td>-0.0086</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Pareto α=1.8</td>
<td>-0.0146</td>
<td>0.306</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-0.0478</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.1  Slopes of Blocking Probability/Buffer Size Plots
on ATM networks. Since one of the goals for third generation PCNs is seamless integration with ATM networks it can be expected that mobile networks carrying such fixed transmission cells are likely to become common. ATM cells are fixed in size and hence have a fixed (deterministic) transmission time which is the service time for queues within the network. So the effect of self-similar arrivals on queues with deterministic service times is therefore of particular interest [84].

Since deterministic service times do not possess any memoryless property, they cannot be analysed by the embedded Markov chain technique. The most straightforward way to obtain performance results is therefore by simulation of the queues. Stochastic simulation involving self-similar sequences is discussed in chapter 8. In this case the analysed sequence (the blocking probability) consists simply of zeroes (packet accepted) and ones (packet blocked) and hence always had a finite variance allowing confidence intervals on the results to be calculated.

The simulations were performed using the Akaroa distributed discrete event simulation package [88]. Akaroa automatically runs multiple independent replications of the simulation on different computers. This provides a linear reduction in simulation time and variance reduction when the results from the different runs are combined. The package also allows for automatic termination of the simulations when confidence interval results of the appropriate precision have been obtained [27]. For all the simulation runs, results were produced with relative precision of 0.05 at a confidence level of 0.95.

The results obtained are shown in figures 4.14, 4.15 and 4.16. In these figures the solid lines represent the analytic blocking probability results obtained for queues with exponential service times obtained earlier in this chapter. The horizontal bars crossing the solid lines are the confidence intervals for the simulations performed of the identical system. These confidence intervals almost all contain the true analytical results, thus validating the simulation model. The dotted lines are the mean results for blocking probability for the model when the service time was changed to a fixed value. The horizontal bars crossing the dotted lines are the confidence intervals for these results. Results are shown for Poisson arrivals and Pareto spaced arrivals with shape parameters $\alpha = 1.2$ and $\alpha = 1.8$. The results obtained for $\alpha = 1.5$, although they were simulated and were consistent with the other results have been omitted from the figures for the sake of clarity.

These results show that the queues with deterministic service times have consistently lower blocking probabilities than those with exponential service times irrespective of the arrival distribution. This was expected in line with the known
4.4 SIMULATION OF P/D/1/N QUEUES

Figure 4.14 Simulation results for exponential and deterministic service time queues with buffer size 5 solid lines - exponential holding times theoretical means error bars - simulation confidence intervals, broken lines - deterministic holding times simulation means, error bars - simulation confidence intervals.
Figure 4.15 Simulation results for exponential and deterministic service time queues with buffer size 10 solid lines - exponential holding times theoretical means error bars - simulation confidence intervals, broken lines - deterministic holding times simulation means, error bars - simulation confidence intervals.
4.4 SIMULATION OF P/D/1/N QUEUES

Figure 4.16 Simulation results for exponential and deterministic service time queues with buffer size 25. Solid lines - exponential holding times theoretical means, error bars - simulation confidence intervals. Broken lines - deterministic holding times simulation means, error bars - simulation confidence intervals.
behaviour of queues with Poisson arrivals [40]. However the relative performance between queues with different arrival processes remains unchanged as the service process is altered. Therefore the analytic blocking probability results obtained for queues with exponential service times form an upper bound on the blocking probability of queues with deterministic service times.

A further point demonstrated by these results is the difficulty of obtaining reliable results from computer simulation once the loss probability drops significantly below $10^{-3}$. This demonstrates the value of the analytic results that are only limited by the precision of the computer on which they are calculated.

4.5 DISCUSSION

A queue can be analysed by the embedded Markov chain technique if either the arrival or the service process possesses the memoryless property allowing for renewal points at which transition probabilities can be calculated. By choosing a queue with negative exponentially distributed service times the state probabilities for the self-similar Pareto spaced arrival process can be calculated exactly. These results are useful because they give the blocking probability for the finite queue.

The blocking probability for Pareto spaced arrivals is significantly higher than for conventional Poisson arrivals. This is caused by the bursty nature of these arrivals where a large group of arrivals can occur together. Whenever the number in the burst exceeds the buffer size, all the extra arrivals will be blocked. At high enough loads (greater than one) a queue with Poisson arrivals will reach a stage where it is almost never empty and so the blocking probability tends to $\rho - 1/\rho$ where $\rho$ is the supplied load. With Pareto distributed arrivals, there is a significant probability of arbitrarily long inter-arrival times due to the heavy tail of the distribution, at any load. This inevitably leads to periods when the queue is empty between bursts of higher than average frequency arrivals. Since the bursts lead to overflows, the blocking probability of the queue will always be higher than for conventional arrivals.

The expected queue waiting times show similar relationships to the blocking probabilities with Pareto spaced arrivals. In particular, those from distributions with lower shape parameters demonstrate worse queueing performance. The lowering of the waiting times at high loads is an artefact caused by the high numbers of blocked arrivals that experience "zero" waiting time since they never actually enter the queue. Taking the expected waiting time of only the non-blocked arrivals shows that these do in fact continue to increase towards the maximum set by the queue length.
The Pareto spaced arrival processes all have a similar relationship between increasing buffer size and decreasing blocking probability, however the gains are much smaller and as load is increased they become relatively smaller.

The assumption of exponentially distributed service times results in worse performance (higher blocking probability) than occurs with deterministic service times such as are becoming common in networks with fixed transmission cell sizes. The relative performance differences caused by the different arrival processes to the queues remain the same with deterministic service times. As such the analytic results found in this chapter are an upper bound on the blocking probabilities of such queues, but have the advantage that they are quickly determined even when the probabilities are very low. Many networks are designed to work with very low blocking probabilities (e.g. optical ATM networks and wireline voice networks). Calculation of such low probabilities is important in these networks and will become important in wireless networks as they seek to provide performance close to their fixed network equivalents.
Chapter 5

MULTIPLE SERVER QUEUES

5.1 RADIO INTERFACE MODEL

In any mobile radio system, the most significant capacity limitation is at the radio interface caused primarily by regulatory limits on the bandwidth available to operators. Systems using FDMA or TDMA access techniques have a fixed number of channels available at any base station for users, e.g. GSM, AMPS, IS54 [71]. Systems using CDMA access have a soft limit on the number of users, which varies according to their channel characteristics, power levels and the permissible levels of interference. This situation is more difficult to model as a queue because of these external factors and since they are beyond the scope of this work, only queue models of fixed size are investigated here.

Since we have a fixed number of servers and no buffer beyond the calls in service, this leads to the $G/G/c/c$ queue model, where $c$ is the number of servers. Normally, telephone conversations are assumed to have an exponential holding (service) time and there are no apparent reasons to assume that this might change in the move from fixed to mobile telephony (see section 7.2). Also, as discussed in section 4.2, this assumption allows analysis of the queue as an embedded Markov chain irrespective of the arrival distribution. The service rate of each server is $\mu$, so the maximum service rate of the queue is $c\mu$. We denote the arrival rate per queue as $\lambda$. The queue is shown in figure 5.1.

This model is the same as the model normally used in fixed telephone systems to describe the arrival of calls at network resources, such as switches or trunk groups. This model uses Poisson arrivals and results in the well known Erlang-B call blocking formula [80]
68 CHAPTER 5  MULTIPLE SERVER QUEUES

![Diagram of a queue model with arrivals and departures labeled.](Image)  

Figure 5.1  Queue Model of Radio Interface

\[ P_B = \frac{\rho^c}{c!} \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \]  

where \( \rho = \lambda/c\mu \) is the queue utilisation. This formula has been applied successfully for many decades and has been shown to be effectively independent of the service time distribution. It is, however, dependent on the Poisson arrival distribution. In this chapter the effect of non-Poisson arrivals is investigated. In a mobile radio network, there can be several sources of new arrivals. In addition to new call arrivals, there are re-call attempts and handovers which may be from fast (vehicle based) or slow (pedestrian) moving mobiles or from users moving along the boundary between two cells. The combination of these individual sources is not likely to be a Poisson arrival process. Here the total arrival process is modelled with a Pareto inter-arrival time distribution.

5.2 THE \( G/M/C \) QUEUE

In generalising queue state results from section 4.2 for \( c \) servers, the basic transition matrix relations (equations 4.3 and 4.4) remain the same but the state
transition probabilities change as the service rate changes [40]. If each server has a mean rate of \( \mu \), the mean total system service rate, with \( n \) customers in the system, is either \( n\mu \), when \( n \leq c \), or \( c\mu \), when \( n \geq c \).

Since we are still using the arrival times as our renewal points for Markov chain analysis, there is a maximum increase of one in the state of the system between renewal points. That is, if the probability of transition from state \( i \) (\( i \) customers in system) to state \( j \) (\( j \) customers in system) is \( p_{ij} \), then

\[
p_{ij} = 0 \quad (j > i + 1)
\]

If \( j \) is greater than \( c \) then all the servers will have been busy throughout the renewal interval and hence the service rate will have been \( c\mu \) for this time, therefore

\[
p_{ij} = b_n \quad (i + 1 \geq j \geq c)
\]

Similarly to the single server case, \( b_n \) can be found by integrating the probability of \( n \) exponential departures at mean rate \( c\mu \) over the Pareto arrival time distribution, \( P(t) \),

\[
b_n = \int_0^\infty \frac{e^{-c\mu t}(c\mu)^n}{n!}dP(t)
\]

In the case that \( i \) and \( j \) are both less than \( c \), the service rate changes from \((i + 1)\mu\) to \( j\mu \) during the service interval and therefore the transition probability depends on the state of the system as well as the size of the transition. Since all the servers operate independently, the transition probability is the binomial probability that there are \( i + 1 - j \) individual service completions before the next arrival. The probability of one server completing service before time \( t \) is \( 1 - e^{-\mu t} \), and therefore

\[
p_{ij} = \int_0^\infty \left( \frac{i + 1}{i + 1 - j} \right) (1 - e^{-\mu t})^{i+1-j}e^{-\mu t}dP(t) \quad (j \leq i + 1 \leq c)
\]
The final possible case is when \( i + 1 > c > j \). The system starts with service rate \( cp \) but during the inter-arrival interval \( c - j \) servers become idle. The inter-arrival time is \( T \) and the last arrival goes into service at a time \( V \) after it arrives, \((0 < V < T)\), where the CDF of \( V \) is \( H(v) \). So, similar to the previous case, \( c - j \) service completions must occur in time \( T - V \), which is a binomial probability,

\[
p_{ij} = \int_0^\infty \int_0^t \left( \frac{c}{c-j} \right) (1 - e^{-\mu(t-v)})^{c-j} e^{-\mu(t-v)}j dH(v) dP(t) \quad (i + 1 > c > j)
\]

(5.6)

The time \( V \) is the time for \( i + 1 - c \) customers to be served with all \( c \) servers working. This is, therefore, the \((i + 1 - c)\)-fold convolution of the exponential distribution with rate parameter \( cp \). The convolution of the exponential distribution results in an Erlang type \( k \) distribution [53] and here \( k = i + 1 - c \) so that

\[
h(v) = \frac{dH(v)}{dv} = \frac{(c\mu)^{i+1-c}}{(i-c)!} v^{i-c} e^{-c\mu v}
\]

(5.7)

Substituting for \( H(v) \) in equation 5.6 gives

\[
p_{ij} = \left( \frac{c}{c-j} \right) \frac{(c\mu)^{i+1-c}}{(i-c)!} \int_0^\infty \int_0^t (1 - e^{-\mu(t-v)})^{c-j} e^{-\mu(t-v)}j v^{i-c} e^{-c\mu v} dv dP(t) \quad (i + 1 > c > j)
\]

(5.8)

From the steady state state probability equation (equation 4.3) the probability of any non-empty state \( j \) is

\[
q_j = \sum_{i=0}^{\infty} p_{ij} q_i \quad (j > 0)
\]

(5.9)

When \( j \geq c \) and \( j > 2 \), all the transition probabilities are given by either equation 5.2 or 5.3 so that
\[ q_j = \sum_{i=0}^{j-2} 0q_i + \sum_{i=j-1}^{\infty} b_{i+1-j}q_i \]
\[ = \sum_{k=0}^{\infty} b_k q_{j+k-1} (j \geq c) \quad (5.10) \]

This is identical to the first line of equation 4.5 and so, by comparison with the analysis for \( c = 1 \), it can be seen that

\[ q_j = C\sigma^j \quad (5.11) \]

where

\[ \sigma = p_L(\mu(1 - \sigma)) \quad (5.12) \]

and \( p_L \) is the Laplace transform of the inter-arrival time probability density distribution.

The constant \( C \) has to be determined from the probability normalisation condition (equation 4.4) which becomes

\[ \sum_{j=0}^{\infty} q_j = \sum_{j=0}^{c-1} q_j + \sum_{j=c}^{\infty} C\sigma^j = 1 \quad (5.13) \]

The state probabilities \( q_j \ (j = 0, 1, \ldots, c - 1) \) are found from the first \( c - 1 \) lines of the steady state state probability equation. Since these are all infinite summations they cannot be solved directly. However a recursive relation can be formed since [40].

\[ q_j = \sum_{i=0}^{\infty} p_{ij}q_i \quad (5.14) \]
\[ = \sum_{i=0}^{c-1} p_{ij}q_i + \sum_{i=c}^{\infty} b_{i+1-j}C\sigma^i \]
Combining this with equation 5.2 gives

\[ q_j = \sum_{i=j}^{c-1} p_{ij} q_i + C \sum_{i=c}^{\infty} b_{i+1-j} \sigma_i^0 \quad (1 \leq j \leq c - 1) \]  

(5.15)

Solving this for \( q_{j-1} \), then dividing by \( C \) and substituting \( q_j' = q_j/C \) gives the result

\[ q_{j-1}' = \frac{q_j' - \sum_{i=j}^{c-1} p_{ij} q_i' - \sum_{i=c}^{\infty} b_{i+1-j} \sigma_i^0}{p_{j-1,j}} \quad (1 \leq j \leq c - 1) \]  

(5.16)

\( q_c' \) can be found from equation 5.11 as

\[ q_c' = q_c = \frac{C \sigma^c}{C} = \sigma^c \]  

(5.17)

So by repeated application of equation 5.16, \( q'_{c-1}, q'_{c-2}, ..., q'_0 \) can be found. Finally \( C \) can be found by solving equation 5.13 in terms of the \( q_j' \) giving [40]

\[ C = \left[ \sum_{j=0}^{c-1} q_j' + \frac{\sigma^c}{1-\sigma} \right]^{-1} \]  

(5.18)

In this way, a complete solution for the steady state state probabilities of infinite buffer queues with arbitrary number of servers can be found.

5.3 THE \( P/M/C/C \) QUEUE

This queue is the blocked-calls-lost model for Pareto-spaced call arrivals to a telephone system switch or trunk group, equivalent to the Poisson arrival model from which the classical Erlang-B distribution is derived. In this model \( i \) and \( j \) are always less than, or equal to, \( c \) so the transition probabilities are all given by either equation 5.2 or 5.5. The transition probability matrix is, therefore,
were all the non-zero \( p_{ij} \) are found by substituting the Pareto inter-arrival time CDF (equation 3.11) into equation 5.5, which gives

\[
\begin{bmatrix}
p_{00} & p_{10} & p_{20} & \cdots & p_{e-1,0} & p_{e-1,0} \\
p_{01} & p_{11} & p_{21} & \cdots & p_{e-1,1} & p_{e-1,1} \\
0 & p_{12} & p_{22} & \cdots & p_{e-1,2} & p_{e-1,2} \\
0 & 0 & p_{23} & \cdots & p_{e-1,3} & p_{e-1,3} \\
\ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \cdots & p_{e-1,c} & p_{e-1,c}
\end{bmatrix}
\] (5.19)

Since \( i \) and \( j \) are both integers, the \( e^{-\mu t} \) terms can be expanded and then the integral can be solved by comparison with the Laplace transform of the Pareto pdf given in equation 3.21. For example, if \( i = j = 1 \)

\[
p_{11} = \int_0^\infty \left( \frac{2}{1} \frac{(1 - e^{-\mu t}) e^{-\mu t}}{(k + t)^{\alpha+1}} \right) \frac{\alpha k^{\alpha}}{(k + t)^{\alpha+1}} dt
\] (5.20)

\[
= 2 \left[ \int_0^\infty e^{-\mu t} \frac{\alpha k^{\alpha}}{(k + t)^{\alpha+1}} dt - \int_0^\infty e^{-2\mu t} \frac{\alpha k^{\alpha}}{(k + t)^{\alpha+1}} dt \right]
\]

\[
= 2 \left[ \alpha(k\mu)^\alpha e^{k\mu} \Gamma(-\alpha, k\mu) - \alpha(2k\mu)^\alpha e^{2k\mu} \Gamma(-\alpha, 2k\mu) \right]
\] (5.21)

As before, any arrival to the queue when it is in state \( c \) is blocked and lost so transition probabilities from this state are the same as from state \( c - 1 \). This means that the last two columns of the matrix are equal and the matrix is singular. Therefore the normalisation condition on the state probabilities (equation 4.4) is required to fully specify a unique steady state solution (equation 4.3).

5.4 RESULTS

The blocking results for five, ten, twenty and forty server queues, calculated from equations 4.3 and 4.4 are plotted in figures 5.2, 5.3, 5.4 and 5.5. In each of these
Figure 5.2  Blocking Probabilities for 5 Server Queues

Figure 5.3  Blocking Probabilities for 10 Server Queues
5.4 RESULTS

Figure 5.4  Blocking Probabilities for 20 Server Queues

Figure 5.5  Blocking Probabilities for 40 Server Queues
figures, the load is given per server. This means that a load of 1 implies that each server receives a load equal to its maximum throughput and thus this represents the maximum for the complete system. These graphs show the same pattern established in chapter 4 for single server queues with Poisson arrivals producing the best performance (lowest blocking) and Pareto spaced arrivals with smaller shape parameters resulting in higher blocking than those with larger parameters. Again this effectively demonstrated the debilitating effect the bursty nature of self-similar arrivals can have on network performance.

5.4.1 Effect of Server Aggregation

Despite there being no buffering in any of these queues, those with higher numbers of servers produce lower blocking at all loads than those queues with a lower number of servers. This demonstrates that there are efficiencies to be gained from aggregating traffic that fits the Pareto inter-arrival time model. It is worth remembering here that the self-similarity of this model is asymptotic, not exact. The effect of server numbers on blocking at fixed loads is shown in figures 5.6, 5.7 and 5.8. These plots are all asymptotically straight lines too. Their slopes, estimated using an asymptotic line of best fit are given in table 5.1. These show that as the load is increased the slopes of the plots for the Pareto spaced arrival processes become closer to the slope of the plot for Poisson arrivals at the same load. Also, at the highest load all the arrival processes give absolute blocking levels that differ by less than an order of magnitude. Clearly the blocking probabilities at such loads are much higher than would be acceptable in normal operation of a network, and so this situation represents severe congestion. However, the trend is interesting and shows a smaller penalty for self-similar arrivals relative to Poisson arrivals in this type of queue than in the single server, finite buffer queue.

5.5 DISCUSSION

Multiple server queues are effective models for networks with channel oriented service. The standard model for telephone networks is the blocked-calls-lost model that includes no buffering. With Poisson arrivals and exponential service, analysis of this model produces the classic Erlang-B probability distribution for blocking, which has been widely used.

The embedded Markov chain analysis technique [40] extends readily to multi-server queues, although calculating the state transition probabilities becomes


5.5 DISCUSSION

Figure 5.6 Blocking Probability vs Server Numbers for Load = 0.5

Figure 5.7 Blocking Probability vs Server Numbers for Load = 0.7
more complicated for transitions from states where the number of customers is greater than the number of servers to states that where the number of customers is smaller. The blocked-calls-lost model avoids this problem and the transition probabilities can easily be solved by comparison with the Laplace Transform of the inter-arrival time probability distribution.

As expected, Pareto spaced arrivals result in significantly higher blocking probabilities than Poisson arrivals. Arrival processes that have the highest self-similarity parameters result in the worst performance. The gap between the different processes widens as the number of servers is increased, faster at low loads than at high. This is because at low loads, if the arrivals are not self-similar, the aggregation effect allows lower blocking with higher numbers of servers at the same mean load per server. Pareto-spaced arrivals are asymptotically self-similar, and so the variance reduction from aggregation is not as large as for Poisson arrivals. At high loads, the systems are all congested much of the time and so the blocking probability differences are smaller. This aggregation effect was not seen in the results calculated for single server queues since the arrival rate was held constant as the buffer size varied. The aggregation is based on blocks of arrivals and occurs as the number of servers increases because the arrival rate to
5.5 DISCUSSION

<table>
<thead>
<tr>
<th>Load</th>
<th>Process ( \alpha )</th>
<th>Slope</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>Pareto 1.2</td>
<td>-0.0116</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>Pareto 1.5</td>
<td>-0.0301</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>Pareto 1.8</td>
<td>-0.0432</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-0.0952</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>Pareto 1.2</td>
<td>-0.0061</td>
<td>0.167</td>
</tr>
<tr>
<td></td>
<td>Pareto 1.5</td>
<td>-0.0136</td>
<td>0.373</td>
</tr>
<tr>
<td></td>
<td>Pareto 1.8</td>
<td>-0.0187</td>
<td>0.512</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-0.0365</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>Pareto 1.2</td>
<td>-0.0035</td>
<td>0.302</td>
</tr>
<tr>
<td></td>
<td>Pareto 1.5</td>
<td>-0.0069</td>
<td>0.595</td>
</tr>
<tr>
<td></td>
<td>Pareto 1.8</td>
<td>-0.0088</td>
<td>0.759</td>
</tr>
<tr>
<td></td>
<td>Poisson</td>
<td>-0.0116</td>
<td>1</td>
</tr>
</tbody>
</table>

*Table 5.1* Slopes of Blocking Probability/Server Numbers Plots

...the system is increased so the number of arrivals per service time increases giving a larger aggregated block of arrivals.
Chapter 6

RANDOM ACCESS PROTOCOLS

6.1 DATA TRAFFIC

Increasingly, mobile networks are being designed to carry multiple traffic types such as data and video as well as voice. This additional complexity leads to significant increases in signalling traffic. These types of traffic have low channel utilisation and are not as well suited to the fixed channel multiplexing schemes normally used for voice. Because of this, random channel access schemes are frequently employed especially for data traffic. Such schemes have the advantage that the full channel capacity is made available for each successful transmission. This can significantly reduce the transmission time providing the system is working efficiently.

The studies that have found network traffic to be self-similar have all been on types of data network [9, 18, 56, 68, 74]. In particular, the original Bellcore study found self-similarity on an Ethernet network which uses a protocol developed from the basic random access techniques studied here.

The performance of random access protocols when the packet arrival process is self-similar is, therefore, important because such protocols are commonly used in channel access, especially in digital wireless communication systems. Searching the literature reveals no analytical work studying the performance of these access techniques with any self-similar traffic model.

Harpantidou and Paterakis simulated slotted Aloha systems using Pareto distributed arrivals in [43] to determine whether they could be stable. They used the first form of the Pareto distribution (equation 3.9) under which no inter-arrival time may be smaller than the location parameter and were so able to stabilise the system by ensuring that the location parameter was greater than 0.7 of a slot. The use of the first form of the Pareto distribution to model arrivals to a broadcast channel seems quite remarkable since, the existence of a non-zero minimum arrival time seems to imply co-ordination between stations that the simple
Aloha protocols are designed to avoid. This model was taken from studies of successful transmission on Ethernet networks, where clearly, there is a minimum packet inter-arrival time that is set by the minimum packet length. However, for arrival to a broadcast channel no such minimum exists. An extreme extension of the existence of a minimum inter-arrival time would be to ensure that the packet length was shorter than this minimum which would ensure perfect throughput regardless of load. The use of the arrival distribution location parameter as the controlling variable is interesting too as then the shape parameter (and hence the arrival process self-similarity parameter) varies with the packet arrival rate.

The classical analysis of the performance of these random access protocols assumes that transmitters are from an (effectively) infinite population of independent stations and hence the arrival process of new packets must be Poisson. It further assumes that the retransmission scheme is such that the total arrival stream comprised of new and retransmitted packets is also a Poisson process. This implies that the retransmitted packet stream is also Poisson and therefore the arrival of each of these packets is independent of the arrival of other retransmissions and of all new packet arrivals. Clearly, this is not true, however simulation studies have shown that if the new packet arrival process is Poisson and the retransmission timeout is sufficiently long so as to avoid many re-collisions then the Poisson assumption for the total arrival stream provides reasonable results [42]. As the analysis is based on knowledge of the inter-arrival time distribution, the effect of self-similar traffic can be investigated using the Pareto distribution. This model makes no assumptions about the transmitter population or the retransmission scheme, but merely states that the combined effect of the two is such that the total arrival stream is self-similar and this self-similarity is most easily modelled with the Pareto inter-arrival distribution.

This chapter gives analytical throughput results for the Aloha, slotted Aloha and CSMA random access protocols. Delay results are also given for both of the Aloha protocols as well as stability analysis of uncontrolled slotted Aloha.

6.2 PURE ALOHA

The Aloha random access protocol was first used by Abramson et al at the University of Hawaii in the early 1970s [3]. It is interesting primarily because it was the forerunner of many of the common channel access schemes that have been proposed and/or adopted since. It is attractive for its sheer simplicity. Pure Aloha is often referred to as the broadcast channel since any station that has a packet ready simply transmits it immediately. Thus it possible for two trans-
missions to overlap and therefore, both fail. Consequently the scheme requires a method for detecting the success or failure of each packet transmission, either by a central station designated for this purpose, or by way of positive acknowledgment with timeout. Stations that detect that their transmissions have not been successful must try again, but they must schedule retransmissions randomly so as to avoid further collisions.

6.2.1 Throughput

The analysis of Aloha is based on the assumption that all transmitted packets are the same size and take the same time to transmit, $T_{tr}$. The packet arrival stream consists of new packets and retransmissions and this is the offered load intensity, $G$. The successfully transmitted packets form the throughput intensity, $S$, given by

$$S = G \cdot Pr(\text{Successful Transmission})$$

\[ (6.1) \]

Using the Aloha protocol, a packet transmission is successful if no other station attempts a transmission that overlaps and so interferes with the first, i.e. there is no collision. (Note, this analysis, as do all the analyses in this chapter, assumes that there is no “capture effect” possible, i.e. any collision causes the loss of both packets.) The period during which another station’s transmission would cause a collision is called the vulnerable period. Since a transmitting station does not know whether another station is already transmitting, this includes a period before the transmission starts. This is shown in figure 6.1 which shows that the vulnerable period is twice the packet transmission time, i.e. $= 2T_{tr}$.
If the mean Poisson arrival rate is \( \lambda \) then the probability of \( n \) arrivals in a time period \( t \) is given by

\[
P_n(t) = \frac{e^{-\lambda t} (\lambda t)^n}{n!}
\]  
(6.2)

The classical analysis invokes the memoryless property of Poisson arrivals and finds that the probability of a successful transmission is therefore the probability of no transmissions in a period of \( 2T_{tr} \), or

\[
Pr(\text{Successful transmission}) = P_0(2T_{tr}) = e^{-2\lambda T_{tr}}
\]  
(6.3)

The offered load, \( G \), is an intensity and so is defined as being the packet arrival rate multiplied by the packet transmission time

\[
G = \lambda T_{tr}
\]  
(6.4)

Substituting equations 6.4 and 6.3 into 6.1 gives the classical throughput load relationship for pure Aloha as found in many texts e.g. \([42, 44, 80]\)

\[
S = Ge^{-2G}
\]  
(6.5)

Since Pareto spaced arrivals do not have the memoryless property, it is necessary to analyse the inter-arrival periods before and after the start of transmission separately. For a transmission to be successful, the previous inter-arrival time and the subsequent inter-arrival time must both be greater than the packet transmission time, \( T_{tr} \). The second form of the Pareto inter-arrival probability distribution is given by equation 3.11, which can be rewritten as

\[
Pr[T > t] = \left(\frac{t + k}{k}\right)^{-\alpha}
\]  
(6.6)
where $\alpha$ is the distribution shape parameter and $k$ is its location parameter. This form of the Pareto distribution is chosen because it is valid for any $t \geq 0$. As in previous chapters, in using this distribution, $\alpha$ can be chosen to provide the desired self-similarity parameter and then $k$ adjusted to give the appropriate mean arrival rate. The range of interest is $1 < \alpha < 2$ for which the arrival count process is self-similar and all moments of the distribution higher than the mean do not exist.

Despite the Pareto spaced arrival count process being self-similar and having long-range dependence, individual inter-arrival times are independent identically distributed variables. Therefore the inter-arrival times before and after a transmission are independent events. The probability of a successful transmission is the product of the probabilities that the inter-arrival periods immediately prior and subsequent to the start of transmission are sufficiently long that no collision occurs.

\[
Pr(\text{Successful Transmission}) = (Pr(\text{arrival time} > T_{tr}))^2
= \left[ \left( \frac{T_{tr} + k}{k} \right)^{-\alpha} \right]^2
\]  \hspace{1cm} (6.7)

The mean of the inter-arrival distribution is given by $\bar{t} = k/(\alpha - 1); \alpha > 1$ so the load is related to the packet transmission time by

\[
T_{tr} = \frac{Gk}{(\alpha - 1)} \quad (\alpha > 1)
\]  \hspace{1cm} (6.8)

Substituting equations 6.7 and 6.8 into 6.1 gives the throughput load relationship for Pareto spaced arrivals as

\[
S = \frac{\sqrt{Gk + k}}{\sqrt{\alpha - 1}}^{-2\alpha}
= \frac{G}{\sqrt{\alpha - 1} + 1}^{\frac{-2\alpha}{\alpha - 1}} \quad (\alpha > 1)
\]  \hspace{1cm} (6.9)

This is a useful result because it is dependent on the offered load and the distribution shape parameter only. The length of packet chosen for the system
has no effect on the throughput, despite its reciprocal relationship to the mean inter-arrival time.

Note that applying this technique of analysing the prior and subsequent inter-arrival times to Poisson arrivals results equation 6.5.

![Figure 6.2 Load/Throughput Curves for pure Aloha.](image)

The load/throughput curves for Pareto inter-arrival distributions with shape parameters $\alpha = 1.2$, 1.5 and 1.8, as well as for Poisson arrivals, to a pure Aloha channel are given in figure 6.2. In the plot the packet duration is normalised to 1 and the Pareto shape parameter, $k$ and the Poisson mean $\lambda$ selected to give the correct arrival rate for each load point calculated. As expected, these results show that Pareto spaced arrivals have lower throughput than Poisson arrivals and that as the shape parameter, $\alpha$, decreases (and hence the self-similarity parameter, $H$, increases) the performance decreases further. It is however interesting to note that at loads above the maximum throughput the performance with Poisson arrivals drops away more quickly than with Pareto spaced arrivals. When the load exceeds about twice the channel capacity for pure Aloha the performance with Poisson arrivals becomes worse than with Pareto spaced arrivals. This is because the heavy tail of the Pareto distribution means that, even at very high loads, the probability of inter-arrival times sufficiently long to allow successful
transmissions is significant.

### 6.2.2 Maximum Throughput

The maximum throughput for the system can easily be found by differentiating the load/throughput equation to find the maxima. This is the peak of the curves plotted in figure 6.2. For Poisson arrivals differentiating equation 6.5 gives the results that

\[
\frac{dS}{dG} = 0 = e^{2G}(1 - 2G)
\]

\[
G_{S_{\text{max}}} = \frac{1}{2}
\]

\[
S_{\text{max}} = \frac{1}{2e}
\]  

(6.10)

For Pareto spaced arrivals, differentiating equation 6.9 gives for \((\alpha > 1)\)

\[
\frac{dS}{dG} = 0
\]

\[
= G \cdot \frac{-2\alpha}{\alpha - 1} \left( \frac{G}{\alpha - 1} + 1 \right)^{-1} + \left( \frac{G}{\alpha - 1} + 1 \right)^{-2\alpha}
\]

\[
= \frac{-2G\alpha}{\alpha - 1} \left( \frac{G}{\alpha - 1} + 1 \right)^{-1} + 1
\]

\[
G_{S_{\text{max}}} = \frac{\alpha - 1}{2\alpha - 1}
\]

\[
S_{\text{max}} = (2\alpha)^{-2\alpha}(\alpha - 1)(2\alpha - 1)^{(2\alpha - 1)}
\]  

(6.11)

The maximum throughputs for the different arrival processes that were plotted in figure 6.2 are given in table 6.1. It is interesting to note that the Pareto spaced arrival maximum throughputs, as well as being lower, occur at lower loadings than the Poisson maximum. The vulnerable period for pure Aloha is \(2T_T\) and the maximum throughput for Poisson arrivals occurs at load \(G = 0.5\). The product of these two is 1 which is a tidy result, since this is the maximum available channel throughput. That the Pareto spaced arrival maximums occur at lower loads indicates the inefficiency such processes cause.
6.2.3 Delay

For the operation of an Aloha system to achieve steady state without losing packets from the system, the arrival rate of new packets must equal the departure rate of packets, which is the throughput, $S$, on the channel. Hence if the load intensity of new and retransmitted packets to the system is $G$ the arrival intensity of retransmitted packets must be $G - S$. So the mean number of retransmissions per packet, $N_R$ is given by

$$N_R = \frac{G - S}{S}$$

$$= \frac{G}{S} - 1$$

(6.12)

Each retransmission cycle involves a deterministic period of time required to determine that the previous transmission attempt occurred, a random period of delay so as to stagger the retransmissions (and thus attempt to avoid a further collision) and an additional transmission time for the actual retransmission attempt. For simplicity, however, here the mean time for the entire retransmission cycle is assumed to be $R_D$ times the packet transmission time. Using this, the mean transmission delay, $\bar{D}$, of a packet passing through the system, is the packet transmission time for the first attempt plus the mean number of retransmissions times the mean time per retransmission [42].

$$\bar{D} = T_{tr} + T_{tr}N_RR_D$$

$$= T_{tr} \left[ 1 + R_D \left( \frac{G}{S} - 1 \right) \right]$$

(6.13)
The delay can be normalised by putting $T_{tr} = 1$. It is obviously highly dependent on the mean retransmission cycle time $R_D$, however this is a scaling parameter and does not affect the shape of the curve or the relativity between different curves produced by different load arrival processes. Substituting equations 6.5 (for Poisson arrivals) and 6.9 (for Pareto spaced arrivals) into this result allows mean delay to be plotted against throughput and gives the results shown in figure 6.3. Here the round trip delay, $R_D$, has also been set to 1 for all curves; altering this parameter stretches the curves up the y axis. The delays for each process at their maximum throughputs can be found by substituting the values found in table 6.1 into equation 6.13 which gives the values in table 6.2. These maximum throughputs are grouped in a small range of delays despite the large variation in the throughputs at which they occur. At higher delay the throughputs all decrease dramatically and they become closer together due to the effect of the heavy tails of the Pareto distributions.

6.3 SLOTTED ALOHA

The Aloha protocol is frequently modified by introducing slots. Slots are equal length time periods within which a station must complete any transmission. The
station therefore, once it has a packet ready, waits until the beginning of the next slot period and then transmits. Slots are assumed to be exactly one packet transmission time long. Hence the vulnerable period is equal to the packet transmission time, $T_{tr}$, as is shown in figure 6.4.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{vulnerable_period.png}
\caption{Vulnerable Period of Slotted Aloha Random Access Protocol.}
\end{figure}

\subsection{Throughput}

The arrival process is assumed to refer to the arrival of packets ready for transmission at the station rather than their transmission start times. For Poisson arrivals the probability of a successful transmission is found by applying equation 6.2 to the vulnerable period so that

$$Pr(\text{Successful Transmission}) = P_0(T_{tr}) = e^{-\lambda T_{tr}} \quad (6.14)$$

Substituting this, and equation 6.4, into the load throughput relationship (equation 6.1) gives the classical result [44]

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Arrival Process & $D_{S_{\text{max}}}$ \\
\hline
Poisson & 2.719 \\
Pareto $\alpha = 1.2$ & 3.643 \\
Pareto $\alpha = 1.5$ & 3.374 \\
Pareto $\alpha = 1.8$ & 3.225 \\
\hline
\end{tabular}
\caption{Delay at Maximum Throughput for Pure Aloha}
\end{table}
For Pareto spaced arrivals, it is again necessary to treat the periods before and after the arrival separately. Since the slot periods are independent of the arrivals the arrival can occur anywhere within the slot with probability $1/T_{tr}$. If the arrival occurs at a time $t$ after the beginning of the slot, then the probability that the previous arrival did not also occur after the beginning of that slot is

$$Pr(\text{no previous arrival in slot} \mid \text{arrival occurs at } t) = Pr(\text{inter-arrival} > t) = \left(\frac{t + k}{k}\right)^{-\alpha}$$ (6.16)

The probability that the next arrival is not within the same slot is the probability that the subsequent inter-arrival time is greater than the remaining time in the slot, $(T_{tr} - t)$ which is

$$Pr(\text{no subsequent arrival in slot} \mid \text{arrival occurs at } t) = Pr(\text{inter-arrival} > T_{tr} - t) = \left(\frac{T_{tr} - t + k}{k}\right)^{-\alpha}$$ (6.17)

These two events are independent so that the total probability of a successful transmission for an arrival at time $t$ can be found by multiplying them together. Multiplying this by the probability of such an arrival and integrating over all possible arrivals within the slot (each arrival must occur within a slot) gives the overall probability of an arrival resulting in a successful transmission as

$$Pr(\text{Successful Transmission}) = \int_{t=0}^{T_{tr}} \frac{1}{T_{tr}} \left(\frac{t + k}{k}\right)^{-\alpha} \left(\frac{T_{tr} - t + k}{k}\right)^{-\alpha} dt$$

$$= \frac{k^{2\alpha}}{T_{tr}} \int_{t=0}^{T_{tr}} (t + k)^{-\alpha} (T_{tr} - t + k)^{-\alpha} dt$$ (6.18)
Making the substitutions $z = t + k$ and $K = T_{tr} + 2k$ gives

$$Pr(\text{Successful Transmission}) = \frac{k^{2\alpha}}{T_{tr}} \int_{z=k}^{T_{tr}+k} (z)^{-\alpha}(K-z)^{-\alpha} dz \quad (6.19)$$

A further substitution of $y = z/K$ gives

$$Pr(\text{Successful Transmission}) = \frac{k^{2\alpha}K^{1-2\alpha}}{T_{tr}} \int_{y=k/K}^{(T_{tr}+k)/K} (y)^{-\alpha}(1-y)^{-\alpha} dy \quad (6.20)$$

The integral part of this equation is a generalised incomplete Beta function, which is defined as

$$\beta(z_0, z_1, a, b) = \int_{z_0}^{z_1} (s)^{a-1}(1-s)^{b-1} ds \quad (6.21)$$

Therefore equation 6.20 becomes

$$Pr(\text{Successful Transmission}) = \frac{k^{2\alpha}K^{1-2\alpha}}{T_{tr}} \beta \left( \frac{k}{K}, \frac{T_{tr}+k}{K}, 1-\alpha, 1-\alpha \right)$$

$$= \frac{k^{2\alpha}(T_{tr}+2k)^{1-2\alpha}}{T_{tr}} \beta \left( \frac{k}{T_{tr}+2k}, \frac{T_{tr}+k}{T_{tr}+2k}, 1-\alpha, 1-\alpha \right) \quad (6.22)$$

Substituting this into the load throughput relationship (equation 6.1) gives the throughput/load relationship for slotted Aloha with Pareto spaced arrivals. This can be normalised by setting $T_{tr} = 1$ so that

$$S = Gk^{2\alpha}(1+2k)^{1-2\alpha} \beta \left( \frac{k}{1+2k}, \frac{1+k}{1+2k}, 1-\alpha, 1-\alpha \right) \quad (6.23)$$

where $k = (\alpha - 1)/G$.

Load/throughput curves for slotted Aloha are given in figure 6.5. These show a repeat of the relationships found for pure Aloha with Pareto spaced arrivals.
exhibiting lower throughput than Poisson arrivals and the smaller the shape parameters the worse the throughput becomes. The maximum throughput of the Poisson arrival process is improved relative to pure Aloha by a factor of 2. The maximum throughputs of the Pareto spaced arrival processes all improve by a factor slightly larger than this. This is because Pareto distribution has higher probabilities of short inter-arrival times and so reducing the vulnerable period of the protocol has a greater effect. The Poisson arrivals have maximum throughput when the load equals 1, but the Pareto spaced arrivals have their maximum throughputs at lower loads.

Again at high load levels, greater than about five times the channel capacity, the throughput of Pareto spaced arrivals begins to exceed that of Poisson arrivals due to the heavy tail of the Pareto distribution. Since such high loading implies that the system is operating with a large number of stations that are backed off with retransmissions, the ability of the system to maintain some throughput when highly congested seems to imply greater stability. This is investigated further in section 6.3.3.
6.3.2 Delay

The relationship for the mean number of retransmissions per packet $\bar{N}_R$ found for pure Aloha in equation 6.12 also holds for slotted Aloha. Here too, the mean retransmission cycle delay is denoted as $R_D$. However there is an extra component in the delay, the mean time from the arrival of a packet till the start of the first transmission in the next slot. Since slot boundaries are independent events from packet arrivals, a packet may arrive anywhere within the slot with equal probability and so the mean slot delay is simply half of a slot period, $T_{tr}/2$. Therefore the mean delay is given by

$$D = T_{tr} + \frac{T_{tr}}{2} + T_{tr} \bar{N}_R R_D$$

Substituting the load throughput results for Poisson arrivals (equation 6.15) and Pareto spaced arrivals (equation 6.23) gives the results that are plotted in figure 6.6. Again, the delays at the maximum throughput for each process are closely grouped, even though the throughputs vary widely. The heavy tail effect of the Pareto distribution manifests itself at higher delays when the throughputs of the Pareto spaced arrival processes become higher than the Poisson arrival process.

6.3.3 Stability

The random access system can be modelled as a queue where the channel is a server and backed off users are in the buffer [42]. In this model the arrival rate into the queue is $S/T_{tr}$ and the mean delay through the buffer is the back-off delay. Little's Law [80] can be applied to find the mean number of stations that are backed off awaiting an opportunity to retransmit, $n_b$, as

$$n_b = \frac{S}{T_{tr}} T_{tr} \bar{N}_R R_D$$

$$= R_D (G - S)$$

The throughput result for Poisson arrivals given in equation 6.15 assumes an infinite population and, therefore, is not exact for application to studying the
6.3 SLOTTED ALOHA

Figure 6.6 Delay/Throughput Curves for Slotted Aloha.

Figure 6.7 Throughput v. Backed Off Users for Slotted Aloha.
stability of a system with a finite population. However it can be shown that this result is a lower bound on the performance of systems with finite populations and that it is a good fit for systems with more than twenty users [42]. The result in equation 6.23 makes no assumption about the size of the population, merely requiring that the net load arrival process has Pareto distributed inter-arrival times. Substituting these two results into equation 6.25 gives the results plotted in figure 6.7.

A load line can be defined for a random access system by assuming that the new traffic arrival rate is directly proportional to the number of free users in the system. If the total number of users in the system is \( M \) then the number of free users is \( M - n_b \) and the new traffic arrival rate is

\[
S = \sigma(M - n_b)
\]

where \( \sigma \) is the ratio of new traffic to free users. This equation defines a straight line on the \((n_b, S)\) plane with slope of \(-\sigma\) and which intersects the \( n_b \) and \( S \) axes at \( M \) and \( \sigma \) respectively. The results, as plotted in figure 6.7, are based on maximum throughput so if the load line is imposed on these curves, the intersection defines the operating point for the system. The accepted definition of stability for these systems is due to Kleinrock and Lam [51], who said that a random access system is stable when its load line intersects (non-tangentially) the throughput curve in exactly one place.

Four possible load line/throughput relationships are shown in figure 6.8. In \((a)\) the system is stable since there is only one intersection point of the two lines. In \((b)\) there are multiple intersection points and the system is bistable between the points of least and most load, producing overall instability. The situation in \((c)\) is stable, however the operating point has a high number of backed off users and a low throughput indicating that it is overloaded and \((d)\) represents an infinite population which is unstable.

Comparing these possible situations with the throughput curves shown in figure 6.7 indicates that although the systems with Pareto spaced arrivals have significantly smaller maximum throughput, when they are operated close to their maximum, they are more likely to be stable. This is because the throughput drops off more slowly as the number of backed off users increases. This is due to the effect noted earlier where the system maintains some throughput even when the total arrival rate of new and retransmitted packets is very high. So
Figure 6.8 Possible Load Line Configurations for Channel Stability Analysis. a Stable b Bi(un)-stable c Stable overloaded d Infinite population unstable
systems with Pareto spaced arrivals can be stable with a higher number of users than systems with Poisson arrivals. However the traffic produced by each user must be much lower so that the total traffic is less than the maximum possible throughput. Otherwise the channel will be overloaded.

Unstable systems can be stabilised by increasing the retransmission delay. This stretches the throughput/ backed off users curve along the \( n_b \) axis. Since the systems with Pareto spaced arrivals have a smaller negative slope to this curve, they should need less stretching to achieve this stability and so can use a smaller retransmission delay and hence, the mean delay per packet is not increased as much.

6.4 CARRIER SENSE MULTIPLE ACCESS

Improved random access protocols are possible when the transmitting station is able to sense activity on the line. The simplest of these is non-persistent Carrier Sense Multiple Access (CSMA) [52]. Under this protocol, when a transmitting station has a packet ready, it first listens to the line to check whether there are already any transmissions occurring and only transmits if there are none. It is called non-persistent because, when blocked, a station does not remain listening to the channel to attempt to transmit immediately it becomes free. Instead it immediately backs off when the channel is sensed busy and re-attempts and some randomly selected later time. Because of propagation delay in any real system, collisions are still possible.

6.4.1 Throughput Analysis

A worst case analysis can be obtained by assuming that any station with a packet arriving within \( \tau \) seconds of the first packet onto an unoccupied line will start transmission and cause a collision, where \( \tau \) is the maximum propagation delay within the system. Thus the vulnerable period of this protocol is \( \tau \). The analysis given here follows that given in [44] in which the results for Poisson arrivals can be found (also in [42] and others).

Consider in figure 6.9 that a packet arrives at time \( t_1 \). Then any other packet that arrives before time \( t_1 + \tau \) will also start transmitting. If the last packet in this period arrives at time \( t_1 + Y \) then other packets arriving between times \( t_1 + \tau \) and \( t_1 + Y + T_{tr} + \tau \) will sense a busy channel and be scheduled for retransmission. The period between \( t_1 \) and \( t_1 + Y + T_{tr} + \tau \) is called the transmission period. If no packets arrive during the vulnerable period then the transmission is successful.
and lasts for a time $T_{tr} + \tau$. The interval between two consecutive transmission periods is the idle period and the combination of a transmission period and the subsequent idle period is a cycle. Let $\bar{B}$ be the mean duration of a transmission period and $\bar{I}$ be the mean idle period duration. Then denote $\bar{U}$ as the mean time during a cycle that the channel is used successfully. By renewal theory then the average channel throughput, $S$, is

$$S = \frac{\bar{U}}{\bar{B} + \bar{I}}$$  \hspace{1cm} (6.27)

As there can only be one successful transmission during a cycle the mean successful transmission time is the length of a packet multiplied by the probability that it is successful which is the probability that no other packet arrives in the first $\tau$ seconds after its arrival. For Poisson arrivals, this probability is found from equation 6.2 and the mean transmission time is

$$\bar{U} = T_{tr}e^{-\lambda\tau}$$  \hspace{1cm} (6.28)

For Pareto spaced arrivals the probability is found from equation 6.6 and the mean transmission time is therefore
The mean length of a busy period is \( \bar{B} = T_{tr} + \tau + \bar{Y} \) where \( \bar{Y} \) is the expected duration of \( Y \) the time between the first and last arrival of the busy period. The distribution function for \( Y \) is the probability that no arrival occurs in a period of \( \tau - Y \) and the range is from 0 to \( \tau \). For Poisson arrivals this is given by

\[
P[Y \leq y] = e^{-\lambda(\tau - y)}
\]  

(6.30)

and so the pdf is

\[
p(y) = \frac{dP[Y \leq y]}{dy} = \lambda e^{-\lambda(\tau - y)}
\]  

(6.31)

The mean is then found as

\[
\bar{Y} = \int_{y=0}^{\tau} yp(y)dy \\
= \int_{y=0}^{\tau} y\lambda e^{-\lambda(\tau - y)}dy \\
= \left[ ye^{-\lambda(\tau - y)} - \int e^{-\lambda(\tau - y)}dy \right]_0^\tau \\
= \left[ ye^{-\lambda(\tau - y)} - \frac{1}{\lambda} e^{-\lambda(\tau - y)} \right]_0^\tau \\
= \tau - \frac{1}{\lambda} (1 - e^{-\lambda\tau})
\]  

(6.32)

For Pareto spaced arrivals, the distribution function is

\[
P[Y \leq y] = \left( \frac{\tau - y + k}{k} \right)^{-\alpha}
\]  

(6.33)

and so the pdf is
6.4 CARRIER SENSE MULTIPLE ACCESS

\[ p(y) = \frac{dP[Y \leq y]}{dy} = \alpha k^\alpha (\tau - y + k)^{-(\alpha+1)} \] (6.34)

The mean is then found as

\[
\bar{Y} = \int_{y=0}^{\tau} yp(y)dy = \int_{y=0}^{\tau} y\alpha k^\alpha (\tau - y + k)^{-(\alpha+1)}dy
\]
\[
= \left[yk^\alpha (\tau - y + k)^{-\alpha} - \int k^\alpha (\tau - y + k)dy\right]^\tau_0
\]
\[
= k^\alpha \left[y(\tau - y + k)^{-\alpha} - \frac{(\tau - y + k)^{(1-\alpha)}}{1-\alpha}\right]^\tau_0
\]
\[
= \tau - \frac{k^\alpha}{1-\alpha} \left[(\tau + k)^{(1-\alpha)} - k^{(1-\alpha)}\right] \quad (6.35)
\]

The mean idle period duration, \( \bar{I} \), is the expected time from the end of the transmission period until the next arrival. Due to the memoryless properties of exponential inter-arrival times, for Poisson arrivals this is simply

\[
\bar{I} = \frac{1}{\lambda} \quad (6.36)
\]

If the end of a transmission period occurs at time \( t_2 \), there are two options. a) The last arrival occurred at time \( t_2 - T_{tr} - \tau \) and was the packet that has just finished transmitting. Or b) there were arrivals that found the channel busy and were scheduled for retransmission so that the last arrival occurred between \( t_2 - T_{tr} \) and \( t_2 \). Because it is not possible that the last arrival occurred between \( t_2 - T_{tr} - \tau \) and \( t_2 - T_{tr} \), the two possible ranges are disjoint and their contributions must be analysed separately.

The expected time till the next arrival for option a) is \( (k + T_{tr} + \tau)/(\alpha - 1) \). The probability of option a) is the probability that the last inter-arrival time was longer than \( T_{tr} + \tau \).

For option b), the probability distribution of the time since the last arrival is given by equation 6.6, so the expected time until the next arrival can be found using the semi-memoryless property of the Pareto distribution (equation 3.16) and integrating over the possible interval.
The probability of option b) is the probability that the inter-arrival time following the start of the packet transmission was longer than $\tau$ and shorter than $T_{tr} + \tau$. Because this inter-arrival time cannot have been shorter than $\tau$ the probabilities of both options a) and b) have to be normalised by dividing by the total probability of these two events. They are disjoint events so their probabilities can be summed resulting in a mean idle period given by

$$
E[t] = \int_{y=0}^{T_{tr}} \frac{(k + x)}{(\alpha - 1)} - \alpha k^\alpha (x + k)^{-(\alpha + 1)} dx
$$

$$
= -\frac{\alpha k^\alpha}{(\alpha - 1)} \left[ \frac{(x + k)^{(1-\alpha)}}{1-\alpha} \right]_{0}^{T_{tr}}
$$

$$
= \frac{\alpha k^\alpha}{(\alpha - 1)^2} \left( (T_{tr} + k)^{(1-\alpha)} - (k)^{(1-\alpha)} \right) \tag{6.37}
$$

To find the throughput/load relationship for non-persistent CSMA with Poisson arrivals, the results for mean successful transmission time, busy period length and idle period length (equations 6.28, 6.32 and 6.36) are substituted into equation 6.27. This gives

$$
\bar{I} = \frac{1}{Pr(\text{a or b})} (E[t|\text{a}].Pr(\text{a}) + E[t|\text{b}].Pr(\text{b}))
$$

$$
= \frac{(k + \tau)^{\alpha}}{k} \left[ \left( \frac{k + T_{tr} + \tau}{k} \right)^{-\alpha} \frac{k + T_{tr} + \tau}{\alpha - 1} + \left( \left( \frac{k + T_{tr} + \tau}{k} \right)^{-\alpha} - \left( \frac{k + \tau}{k} \right)^{-\alpha} \right) \frac{\alpha k^\alpha}{(\alpha - 1)^2} \left( (T_{tr} + k)^{(1-\alpha)} - (k)^{(1-\alpha)} \right) \right]
$$

$$
= \frac{1}{(k + \tau)^{-\alpha}(\alpha - 1)} \left[ (k + T_{tr} + \tau)^{(1-\alpha)} + \left( (k + T_{tr} + \tau)^{-\alpha} - (k + \tau)^{-\alpha} \right) \frac{\alpha k^\alpha}{(\alpha - 1)} \left( (T_{tr} + k)^{(1-\alpha)} - (k)^{(1-\alpha)} \right) \right] \tag{6.38}
$$

Substituting in $G = \lambda T_{tr}$ (equation 6.8) and denoting $a = \tau / T_{tr}$ gives

$$
S = \frac{T_{tr} e^{-\lambda \tau}}{T_{tr} + 2\tau - \frac{1}{\lambda}(1 - e^{-\lambda \tau} + \frac{1}{\lambda})} \tag{6.39}
$$
\[ S = \frac{T_{tr}e^{-aG}}{T_{tr} + 2aT_{tr} + \frac{T_{tr}e^{-aG}}{G}} \quad (6.40) \]

Without loss of generality, this can be normalised to \( T_{tr} = 1 \) and rearranged to give the standard result [44]

\[ S = \frac{Ge^{-aG}}{G(1 + 2a) + e^{-aG}} \quad (6.41) \]

The results for the Pareto arrivals are longer, so handling them separately and substituting \( a = \tau/T_{tr} \) and \( T_{tr} = Gk/(\alpha - 1) \) (equation 6.8) into the equations 6.29, 6.35 and 6.38 gives

\begin{align*}
\tilde{U} &= \frac{Gk}{(\alpha - 1)} \left( \frac{aGk/(\alpha - 1) + k}{k} \right)^{-\alpha} \\
\tilde{Y} &= \frac{Gk}{(\alpha - 1)} + \frac{2aGk}{(\alpha - 1)} + \frac{k}{\alpha - 1} \left[ \left( \frac{aGk}{(\alpha - 1)} + k \right)^{(1-\alpha)} - k^{(1-\alpha)} \right] \\
\tilde{I} &= \frac{1}{(k + aGk/(\alpha - 1))^{-\alpha}(\alpha - 1)} \left[ \left( k + \frac{Gk}{(\alpha - 1)} + \frac{aGk}{(\alpha - 1)} \right)^{(1-\alpha)} - \left( k + \frac{aGk}{(\alpha - 1)} \right)^{-\alpha} \right] \\
&\quad - \frac{\alpha k^\alpha}{(\alpha - 1)} \left[ \left( \frac{Gk}{(\alpha - 1)} + k \right)^{(1-\alpha)} - (k)^{(1-\alpha)} \right] \quad (6.42) \\
\end{align*}

Which simplify to

\begin{align*}
\tilde{U} &= \frac{Gk}{(\alpha - 1)} \left( \frac{aG}{(\alpha - 1)} + 1 \right)^{-\alpha} \\
\tilde{Y} &= \frac{Gk}{(\alpha - 1)} + \frac{2aGk}{(\alpha - 1)} + \frac{k}{\alpha - 1} \left[ \left( \frac{aG}{(\alpha - 1)} + 1 \right)^{(1-\alpha)} - 1 \right] \\
\tilde{I} &= \frac{1}{((\alpha - 1)/G + a)^{-\alpha}(\alpha - 1)} \left[ \frac{Gk}{(\alpha - 1)} \left( \frac{\alpha - 1}{G} + 1 + a \right)^{(1-\alpha)} + \right]
\end{align*}
\[
\left( \frac{\alpha - 1}{G} + 1 + a \right)^{-\alpha} - \left( \frac{\alpha - 1}{G} + a \right)^{-\alpha} \\
\frac{\alpha k}{(\alpha - 1)} \left( \left( \frac{G}{(\alpha - 1)} + 1 \right)^{(1-\alpha)} - 1 \right)
\]

These are normalised by putting \( T_r = Gk/(\alpha - 1) = 1 \) and noting this that implies \( k = (\alpha - 1)/G \), so yielding

\[
\bar{U} = (a/k + 1)^{-\alpha} \\
\bar{Y} = 1 + 2a + 1/G \left[ (a/k + 1)^{(1-\alpha)} - 1 \right] \\
\bar{I} = \frac{(k + a)^{\alpha}}{\alpha} \left[ (k + 1 + a)^{(1-\alpha)} + \frac{\alpha}{G} \left( (k + 1 + a)^{-\alpha} - (k + a)^{-\alpha} \right) \left( (1/k + 1)^{(1-\alpha)} - 1 \right) \right]
\]

Substituting these back into equation 6.27 gives the final result that

\[
S = (a/k + 1)^{-\alpha} \cdot \left[ 1 + 2a + 1/G \left[ (a/k + 1)^{(1-\alpha)} - 1 \right] + \frac{(k + a)^{\alpha}}{\alpha} \left[ (k + 1 + a)^{(1-\alpha)} + \frac{\alpha}{G} \left( (k + 1 + a)^{-\alpha} - (k + a)^{-\alpha} \right) \left( (1/k + 1)^{(1-\alpha)} - 1 \right) \right] \right]^{-1}
\]

6.4.2 Results

These results are plotted in figures 6.10 (equation 6.41), 6.11, 6.12 and 6.11 (all equation 6.45) for a number of different values of the propagation delay ratio \( a \). The effect of this parameter can clearly be seen with all the different arrival processes. It is well known that it should be minimised for greatest throughput [44, 42]. A comparison of the results for different arrival processes is shown in figure 6.14. The comparison shows the same pattern as previously observed. Poisson arrivals produce better performance than Pareto spaced arrivals. Pareto spaced arrivals with a higher shape parameter perform better than those with lower \( \alpha \). It also shows that the differences between the arrival processes reduce as \( a \) is reduced towards 0. This is because the vulnerable period is reduced, diminishing the possibility of random arrivals causing collisions.
6.4 CARRIER SENSE MULTIPLE ACCESS

Figure 6.10 Load/Throughput Curves for CSMA with Poisson Arrivals.

Figure 6.11 Load/Throughput Curves for CSMA with Pareto Spaced Arrivals; $\alpha = 1.2$
Figure 6.12 Load/Throughput Curves for CSMA with Pareto Spaced Arrivals; $\alpha = 1.5$

Figure 6.13 Load/Throughput Curves for CSMA with Pareto Spaced Arrivals; $\alpha = 1.8$
The crossover of performance, where the throughput with Pareto spaced arrivals matches and then exceeds that of Poisson arrivals as noted for Aloha and slotted Aloha, at high loads can be seen in the curves for \( \alpha = 0.1 \). It is not seen in the curves for smaller values of \( \alpha \) because their throughputs remains high at all higher load levels for which results have been plotted. The crossover is only seen once the probability of a collision during the vulnerable period is dominated by the heavy tail of the Pareto distribution. The length of this vulnerable period is \( \alpha \). As this becomes smaller, the load must become proportionally heavier for this crossover effect to occur.

6.5 DISCUSSION

The contrast between the mathematics of analysing these protocols with Poisson arrivals and with Pareto arrivals demonstrates both the large analytical advantage of the memoryless property of the negative exponential inter-arrival distribution and the difficulty of obtaining results for even the simplest of self-similar models.

The results found here have several implications for the design of random access systems where the traffic flow is expected to be self similar.
Throughput performance will be worse with self-similar traffic than with traffic that does not display this property. The performance will fall as the self-similarity parameter increases. Delay performance will be similarly affected.

Despite the low maximum throughput, the stability of these systems with finite populations is good. This is because the lulls between traffic bursts provide opportunities for some packets to be transmitted even at very high loads.

With more sophisticated carrier sensing protocols, decreasing the propagation delay reduces the impact of self-similarity by reducing the period in which each transmission can collide with other random arrivals. Reducing the propagation delay does improve the performance for all the arrival processes studied, but the effect is most dramatic for those with the greatest self-similarity. Propagation delay is proportional to the distance between transmitter and receiver. Protocol designers must allow for the maximum delay possible in the system, dependant on the maximum distance between stations. In cellular systems, propagation delay is reduced by reducing cell size.

The causes of these differences can be best seen by plotting the inter-arrival time distributions as in figure 6.15. These are the log-scale plots of the pdfs for the exponential distribution and Pareto distributions with shape parameters $\alpha = 1.2$ and 1.8. The mean of each distribution is 1. The graph clearly shows that the Pareto and exponential distributions cross over twice. The probabilities of very short and very long inter-arrival times are greater for the Pareto distributions. Reducing the vulnerable period of a random access protocol has a more significant effect on the performance with Pareto spaced arrivals, but the probability of a short inter-arrival time and hence a collision will always be greater. Also the probability of long inter-arrival times that will allow successful transmissions through is greater so even at high loads some throughput will occur.

These effects can be expected to hold in general for self-similar traffic regardless of whether the Pareto inter-arrival time model is accurate. This is because all self-similar traffic has a similar bursty nature with long lulls between intense bursts of traffic.
Figure 6.15  Poisson and Exponential Probability Distribution Functions.
Chapter 7

DATA ANALYSIS

In this chapter, data obtained from the cellular telephone networks of Telecom New Zealand and BellSouth New Zealand is analysed for evidence of self-similarity using standard heuristic tests. First, the tests are reviewed and demonstrated using simple synthetic data sequences.

7.1 STATISTICAL TESTS FOR SELF-SIMILARITY

The most commonly used analysis techniques used for detecting long-range dependence and self-similarity are heuristic techniques based on their definitions and basic properties as described in chapter 2. Such heuristic techniques have advantages in that they are straightforward to apply, are reasonably computationally undemanding and provide an easily interpreted graphical output that is generally accurate. The disadvantage of these techniques is that they provide no confidence interval and therefore the results cannot be ascribed any particular level of certainty. More refined techniques are generally based on maximum likelihood estimation (MLE) of the periodogram [8]. In particular, Whittle's approximate MLE has been used for estimation of the self-similarity parameter of network traffic series [56, 9]. However it is reasonable to apply heuristic techniques to discover whether there is any basis for suspecting self-similarity in a data set. The other more comprehensive techniques can be applied subsequently if the heuristic techniques suggest there is significant evidence of self-similarity. The arrival processes used in chapters 4, 5 and 6 for performance comparisons are also used in this chapter to demonstrate these analysis techniques. These are Pareto spaced arrivals with shape parameters of $\alpha = 1.2, 1.5$ and $1.8$ and Poisson arrivals as a standard for comparison. For each arrival process a count sequence of 100000 counts was generated with a mean arrival rate of 5 per count interval. The sequences were generated using the drand48() C psuedo random number generator [82] and transforming the results to produce the appropriate inter-arrival
time sequence. The sequences are compared in table 7.1. This shows that the calculated means of the sample sequences are close to the theoretical means for all the sequences except for Pareto spaced arrivals with \( \alpha = 1.2 \). The variance of this distribution is so high that a sequence of more than 500000 pseudo-random arrival times is not sufficient for convergence to the theoretical mean.

<table>
<thead>
<tr>
<th>Process</th>
<th>Poisson</th>
<th>Pareto spaced</th>
<th>Pareto spaced</th>
<th>Pareto spaced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>( \lambda = 5 )</td>
<td>( \alpha = 1.8 )</td>
<td>( \alpha = 1.5 )</td>
<td>( \alpha = 1.2 )</td>
</tr>
<tr>
<td></td>
<td>( k = 0.16 )</td>
<td>( k = 0.10 )</td>
<td>( k = 0.04 )</td>
<td></td>
</tr>
<tr>
<td>Theoretical Results</td>
<td>Mean</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>( H ) parameter</td>
<td>0.5</td>
<td>0.6</td>
<td>0.75</td>
</tr>
<tr>
<td>Calculated Results</td>
<td>Mean</td>
<td>4.995</td>
<td>4.9876</td>
<td>4.9909</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>5.0153</td>
<td>15.963</td>
<td>22.280</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>17</td>
<td>26</td>
<td>27</td>
</tr>
</tbody>
</table>

Table 7.1 Demonstration Sequences Comparison

### 7.1.1 Autocorrelation

The auto-correlation function of a data series can be easily calculated. Recall that from the definition of long-range dependence, the auto-correlation function of any self-similar series must be of the form

\[
 r(k) \sim k^{-\beta}L(t) \quad \text{as} \quad k \to \infty
\] (7.1)

By taking the logarithm of both sides of equation 7.1, it can be seen that if the slowly varying function \( L(t) \) is assumed to be asymptotically constant, a graph of \( \log(r(k)) \) against \( \log(k) \) should be a straight line with a slope of \( -\beta \). A simple least squares estimate of the slope can therefore be used to estimate \( H \) by using the relationship \( H = 1 - \beta/2 \) (equation 2.3). Figure 7.1 shows the auto-correlation functions for the demonstration arrival processes up to a maximum lag of 100. The plots show that estimates for small auto-correlation values are not as stable as for the larger values. In particular, the calculated auto-correlations...
7.1 STATISTICAL TESTS FOR SELF-SIMILARITY

for the Poisson arrival sequence range over nearly six orders of magnitude. In theory these auto-correlations should all be zero and in practice they are all very small, however their wide range of values makes the least squares calculation of the slope of the plot nearly meaningless.

\[ \log(\text{var}(X^{(m)})) \sim cm^{-\beta} \quad \text{as} \quad m \to \infty \quad (7.2) \]

where \( c \) is some positive finite constant. Therefore a plot of \( \log(\text{var}(X^{(m)})) \) against \( \log(m) \) will have an asymptotic slope of \(-\beta\) which can be simply estimated through a least squares fit to the plot over reasonable values of \( m \). Log/log

\( \text{Pareto} \quad \alpha=1.2, \beta=0.29 \)
\( \alpha=1.5, \beta=0.57 \)
\( \alpha=1.8, \beta=0.83 \)

\( \text{Poisson, } \beta=0.25 \)

Figure 7.1 Log/log Plot of Pareto Arrival Sequence Autocorrelations.

### 7.1.2 Variance Analysis

An alternative method of looking at the behaviour of the auto-correlations is variance analysis. This is derived from the observation that specifying the variance of the aggregated sequence \( \text{var}(X^{(m)} : m \geq 1) \) is equivalent to specifying the auto-correlations \( \{r(k) : k \geq 1\} \) [16]. Therefore, from equation 7.1 (again assuming that \( L(t) \) is effectively constant), we have [9]
plots of the variances of the aggregations of the Poisson arrival sequence and the Pareto inter-arrival sequence are shown in figure 7.2. The block lengths used for aggregation ranged from a minimum of five counts up to one quarter of the total sequence length. For each block length, the means of as many non-overlapping blocks as could be fitted into the sequence length were made and the variance of these means calculated. In each case, high levels of aggregation mean that few non-overlapping blocks can be used for estimating the variance and the estimate is poor. In these cases the slope of the resulting line can be seen to increase at high levels of aggregation, resulting in an over-estimate of $\beta$. Therefore the slopes have been estimated from the first 70% of the block sizes used. This was chosen as a balance between discarding too many data points and including points that were excessively random. In tests with sample sequences the 70% figure produced consistent results.

Figure 7.2 Pareto and Poisson Arrival Aggregated Sequence Variances.

7.1.3 R/S Analysis

The self-similarity or Hurst parameter can also be estimated directly from the definition.
This shows that a plot of $\log(R(d)/S(d))$ against $\log(d)$ should have an asymptotic slope of $H$. There is considerable latitude in the methodology used to ensure accurate determination of the $R$ and $S$ parameters. This procedure is often referred to as rescaled adjusted range or R/S analysis. The log/log plot of $R(d)/S(d)$ against $d$ for the demonstration sequences is shown in figure 7.3. Here the values have been calculated on exponentially increasing block lengths starting from 10 counts and going up to the entire sequence length. For each block length, $R(d)$ and $S(d)$ has been calculated on as many non-overlapping blocks as fit into the entire sequence and the results averaged. The self-similarity parameter is calculated as the least-squares best fit slope to the plotted curve.

This varies slightly from the commonly used technique of plotting all the calculated values for all the non-overlapping blocks of each block size and plotting a best fit line to the resulting scatter diagram. The reason for this change was to allow results from multiple data sets to be plotted on the same figure. The sequences tested here were analysed using this best fit technique and the results were all within 0.01 of the results given here from the averaging technique.

The characteristics of the demonstration sequences and results obtained by applying these techniques are summarised in table 7.2.

<table>
<thead>
<tr>
<th>Process</th>
<th>Poisson spaced</th>
<th>Pareto spaced</th>
<th>Pareto spaced</th>
<th>Pareto spaced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameters</td>
<td>$\lambda = 5$</td>
<td>$\alpha = 1.8$</td>
<td>$\alpha = 1.5$</td>
<td>$\alpha = 1.2$</td>
</tr>
<tr>
<td>$k = 0.16$</td>
<td>$k = 0.10$</td>
<td>$k = 0.04$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Theoretical Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H$ parameter</td>
<td>0.5</td>
<td>0.6</td>
<td>0.75</td>
<td>0.9</td>
</tr>
<tr>
<td>$\beta$ parameter</td>
<td>1.0</td>
<td>0.8</td>
<td>0.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Calculated Results</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>$\beta$</td>
<td>-</td>
<td>0.83</td>
<td>0.57</td>
</tr>
<tr>
<td>$H$</td>
<td>-</td>
<td>0.585</td>
<td>0.715</td>
<td>0.855</td>
</tr>
<tr>
<td>Aggregated</td>
<td>$\beta$</td>
<td>1.01</td>
<td>0.75</td>
<td>0.6</td>
</tr>
<tr>
<td>Variances</td>
<td>$H$</td>
<td>0.505</td>
<td>0.625</td>
<td>0.7</td>
</tr>
<tr>
<td>$H$</td>
<td>0.55</td>
<td>0.66</td>
<td>0.76</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 7.2 Summary of Demonstration Sequences Results
The results in the table show that the heuristic techniques can identify self-similar and Markovian processes consistently. Some of the results, especially for the series with high self-similarity parameters differ slightly from the theoretical values. This is probably because of the highly variable nature of the series, rather than inaccuracy of the applied techniques. Repeating the tests with new sequences generated with different random seeds resulted in a range of values that included the theoretical values. This emphasises the need for large data samples when analysing for the existence of self-similarity.

7.2 DATA FROM TELECOM NEW ZEALAND

Telecom Mobile Networks supplied the complete call records for November 8, 1996 from the billing system for their national cellular telephone network. The data covered 728 cell codes from six regions and included data for 925,238 calls. Checking the data revealed that a large number of these were in fact duplicate (i.e. calls from the same cell at the same time of the same duration). This was confirmed with Telcom NZ. Removal of the duplicate calls left 780,840 call records. Since the data was obtained from the billing system, it only includes
records of new calls. Also, no information on hand-overs was provided.

Because aggregating records from multiple cell sites removes the effects of the geographical movement of subscribers, the data was not analysed in its entirety, but broken down to individual cells. Since Telecom were unprepared to release specific network topology information, data from the Auckland region was chosen to maximise the chances that the cells analysed covered more densely populated urban areas. The Auckland region includes 325 cell codes and produced 296,050 recorded calls with a mean of 911 calls/cell. In order to ensure that a stationary series was used for analysis, data was taken from hours with the same (or nearly the same) mean arrival rate. To ensure consistent characteristics it was desirable that these hours should be the busy hour or close to it.

Two series have been derived from the data. The first, (sequence A), is taken from ten cells that had between 898 and 940 cell arrivals by choosing hours that were within 10% of the busiest hour from all the cells. Twenty five such hour long records were combined to give a series of 2621 call records which was processed to give a sequence of 1500 call-per-minute records. The second series for comparison, (sequence B), was taken from the second and third busiest cells which had 3341 and 3507 calls respectively. Eight hours, having arrival rates within 10% of the busiest hour, were combined to produce a sequence of 3,400 call records. This was processed to give a sequence of 1,440 20 second arrival counts.

These sequences are not particularly long for analysing for the presence of self-similarity, for example the seminal paper by Leland et al [56] analysed sequences of 360000 observations. However the techniques described in section 7.1 were applied to give an indication of whether further data collection should be pursued. The results obtained are shown in figures 7.4, 7.5 and 7.6.

The results from these tests are summarised in table 7.3. The auto-correlation results have been omitted because, similar to the Poisson arrival results in section 7.1, the size of the auto-correlations is small and their relative range is large enough to make the concept of a slope almost meaningless.

Clearly, despite the short length of these sequences, the results of these heuristic analysis techniques strongly counter the hypothesis that the new call arrival process is self-similar at a cellular level. These results do not justify additional analysis by MLE technique.
Figure 7.4 Plot of Telecom Arrival Sequence Autocorrelations.

Figure 7.5 Log/log Plot of Telecom Arrival Aggregated Sequence Variances.
7.3 DATA FROM BELL SOUTH

BellSouth New Zealand supplied data from the traffic management engineering system of their National GSM digital cellular telephone network recorded in late April and May 1998. The GSM system uses radio sub-bands carrying eight TDMA channels and sites are designated as to how many transmitter-receivers (TRX - one per sub-band) they are equipped with. BellSouth supplied data for three different sites: two days from a 1 TRX site with 1 signalling and seven customer traffic channels, three days data from a 2 TRX site with 2 signalling.

### Table 7.3 Summary of Telecom Sequences Results

<table>
<thead>
<tr>
<th>Series</th>
<th>Sequence A</th>
<th>Sequence B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.747</td>
<td>2.361</td>
</tr>
<tr>
<td>Variance</td>
<td>1.906</td>
<td>2.525</td>
</tr>
<tr>
<td>Maximum</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Aggregated $\beta$</td>
<td>1.22</td>
<td>0.94</td>
</tr>
<tr>
<td>Variances $H$</td>
<td>0.39</td>
<td>0.53</td>
</tr>
<tr>
<td>R/S Analysis $H$</td>
<td>0.58</td>
<td>0.58</td>
</tr>
</tbody>
</table>
and 14 traffic channels and two days data from a 3 TRX site with 2 signalling and 22 traffic channels. The data was recorded by the engineering system polling the Base Station Controller for its status in a continuous loop and the data supplied for each day consisted of 2,000 poll samples taken starting at approximately 3pm each day. Polling the 1 TRX site took on average 3.44 sec, the 2 TRX site 5.38 sec and the 3 TRX 5.67 sec. The time-stamp on each record gave the time to the nearest second so the time difference between successive records varies. However it has been assumed for the purposes of the analysis that the poll records were all equally spaced for each site.

The advantage of using data derived from the engineering system is that it records the actual radio channel usage. Data derived from the billing system, such as the Telecom NZ sets, does not record the effect of hand-overs and so although it records the input process to the system it does not include all the population mobility factors.

On examining the data it was quickly apparent that the series were not stationary, in particular each sequence had a traffic peak around 5pm and typically traffic volumes subsided after the peak. This is a particular problem because the techniques for analysing self-similarity are not robust to non-stationarity of the data that can lead to false positive indications. The data from the 1 TRX site was found to be particularly highly varying in its mean input rate and was entirely discarded. The data from the 2 and 3 TRX sites were pruned to give a sequence of reasonably consistent half-hour means. The half hour means for the pruned sequence of 2 TRX site data were 5.24 +/- 0.88 Erlang which gave a total of 3,500 time samples from the three sequences. For the 3 TRX sites the means became 10.76 +/- 1.23 Erlang with a total sequence length of 2,500 samples from the two dates. This process is somewhat arbitrary and involved a trade-off between ensuring stationarity and maintaining adequate sequence length for effective analysis.

A better approach would have been to collect more data from hours entirely within the business day during which the traffic arrivals could be assumed to be stationary. It would also have been desirable to monitor the aggregate traffic from a larger area to see that this was consistently stationary so that fluctuations at the cellular level due to population mobility were not unnecessarily eliminated. Unfortunately, these approaches were not available to us due to the limited time and data that BellSouth were able to provide.
Figure 7.7  Plot of BellSouth Channel Occupation Sequence Autocorrelations.

Figure 7.8  Log/log Plot of BellSouth Channel Occupation Aggregated Sequence Variances.
7.3.1 Channel Occupation Series Results

Since the data sets gave the channel occupation at each record point, this time series was analysed first. The results are shown in figures 7.7, 7.8 and 7.9 and are summarised in table 7.4. The auto-correlation plots show a significant problem with analysis of the channel occupation series, the data is significantly auto-correlated over the period of the channel holding time. After the holding time effect disappears the auto-correlations become insignificant. The same effect appears in the aggregated variance plot where the slope is quite different at low levels of aggregation than at high levels. The results given are the slopes of these plots at the higher levels of aggregation. The R/S analysis plots show constant slopes indicating self-similarity parameters significantly different from the 0.5 of Markovian processes for both cell sites. These plots do not clearly separate the effect of the holding time from the longer-range sequence behaviour. The numbers given are also inconsistent with the results of the aggregated variance results so it appears that the R/S plots have not been able to discriminate against the medium term auto-correlations due to the channel holding time. This is most likely due to inadequate sequence length. The conclusion is therefore that these results do not support a hypothesis that the channel occupancy process is self-similar.
7.3.2 Arrival Sequence Results

<table>
<thead>
<tr>
<th>Series</th>
<th>2TRX Sequence</th>
<th>3TRX Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.355</td>
<td>10.833</td>
</tr>
<tr>
<td>Variance</td>
<td>4.388</td>
<td>8.69</td>
</tr>
<tr>
<td>Maximum</td>
<td>13</td>
<td>21</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Aggregated β</td>
<td>0.87</td>
<td>0.92</td>
</tr>
<tr>
<td>Variances H</td>
<td>0.565</td>
<td>0.54</td>
</tr>
<tr>
<td>R/S Analysis H</td>
<td>0.88</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 7.4 Summary of BellSouth Channel Occupation Sequences Results

The data was processed to produce sequences of the number of arrivals (channel seizures) between each poll record by recording channel transitions from an idle to an occupied state. This process has the potential to miss some arrivals if a channel is released from one connection and seized by another entirely between polling times. This, however, happens infrequently as the channels are assigned to connections in a cyclic pattern. Also it is more likely to happen when the site is busiest and therefore would tend to reduce the size of peaks recorded in the arrival rate and so give conservative results when testing for self-similarity. The results of testing the sequences produced in this manner are shown in figures 7.10, 7.11 and 7.12 with a summary given in table 7.5.

These results consistently show that the arrivals traffic at the 2 TRX site appears to be self-similar with parameter of approximately $H = 0.7$ whereas the arrivals to the 3 TRX site are not self-similar. There are a number of possible explanations. The 2 TRX site carries a smaller volume of traffic and so smaller fluctuations are more significant. It is possible that this series is not sufficiently stationary although this would be expected to result in a more random autocorrelation plot. It is also possible that the 3 TRX site data concealed more channels that release from one connection and are seized by another during a single polling interval since it ran at higher utilisation levels although the additional efficiency of the larger channel group would be expected to prevent this. The data was also analysed for holding times and this showed that the mean channel holding time on the 2 TRX site was 31.5 seconds and the mean on the 3 TRX site was 50 sec. This indicates that the 2 TRX site traffic handed over more quickly and therefore a greater fraction of the arrivals were hand-overs from other sites rather than new calls. These results then give some support to the contention that
Figure 7.10  Plot of BellSouth Channel Arrival Auto-correlations.

Figure 7.11  Log/log Plot of BellSouth Arrival Aggregated Sequence Variances.
mobility effects can introduce self-similar characteristics into cellular telephone traffic.

![R/S Analysis of BellSouth Arrival Sequences](image)

**Figure 7.12** R/S Analysis of BellSouth Arrival Sequences.

<table>
<thead>
<tr>
<th>Series</th>
<th>2TRX Sequence</th>
<th>3TRX Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.886</td>
<td>1.28</td>
</tr>
<tr>
<td>Variance</td>
<td>0.932</td>
<td>1.31</td>
</tr>
<tr>
<td>Maximum</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Auto-correlation</td>
<td>β 0.405</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>H 0.8</td>
<td>-</td>
</tr>
<tr>
<td>Aggregated Variances</td>
<td>β 0.51</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>H 0.745</td>
<td>0.545</td>
</tr>
<tr>
<td>R/S Analysis</td>
<td>H 0.74</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Table 7.5** Summary of BellSouth Arrival Sequences Results
7.4 HOLDING TIMES

The standard holding time model for telephone traffic is an exponential distribution. Clearly mobile calls with hand-overs will have different channel holding times from the un-interrupted call holding time and thus will differ from the standard exponential model. It is, therefore, interesting to plot these to investigate the differences as in figure 7.13. This shows the call holding time distribution from all the cells in the Auckland region of the Telecom cellular network. The decay is approximately linear on the log graph indicating an exponential decay. The other two plots are the channel holding times taken from the 2 TRX and 3 TRX cells on the BellSouth network. These plots have an initial steep slope and then become flatter for longer holding times. The steep slope is the holding time distribution due to calls that are handed over i.e. cellular telephones that are moving during their connection. The flatter part is parallel to the Telecom plot indicating that these are call holding times so must be those calls made from telephones that are stationary during the period of the call. So the channel holding time distribution is composed of two exponentials.

Comparing the holding time distribution of the two BellSouth sites, the call holding sections have a similar decay rate, although the 2 TRX site plot is lower due to the smaller fraction of calls that run to completion within the cell. The hand-over sections however have significantly different decay slopes. This indicates that the hand-overs occur much more quickly at the 2 TRX site. This may be because the site is physically smaller although it could be due to other physical factors such as ratio of vehicular to pedestrian traffic, speed limits etc. Nevertheless this higher speed of hand-over (and hence higher hand-over rate) must be responsible for the appearance of self-similar characteristics in the arrival processes to the site.

7.5 DISCUSSION

The results obtained here have significant limitations imposed by the data sets that were obtained. This work was not a priority for either Telecom Mobile Networks or BellSouth New Zealand and, although individuals within both companies were extremely cooperative, the time they could spend on data collection for this project was tightly constrained. The main drawback has been the lack of opportunity to collect further data sets taking advantage of the experience gained in analysing these ones.
The biggest limitation of the data sets that were obtained was that the consistent stationary sequences that were obtained from them were very short. For comparison the data sets used by Leland et al to determine that Ethernet traffic is self-similar were of the order of a million data points long [56]. This precludes the possibility of firmly concluding that self-similarity exists in any of the traffic sequences tested.

The question of distinguishing between genuine self-similarity and non-stationarity is still very much an open one. It is definitely possible for data from a self-similar process to appear non-stationary even when the underlying physical process is stationary. Equally a non-stationary process may appear to be self-similar. Developing techniques for distinguishing true self-similarity from non-stationarity is beyond the scope of this work. Hence, the simple test of ensuring a limited range of quarter hour traffic rates has been used.

The data set from Telecom was obtained from the network billing system. Although this approach appeared to offer the best possibility for obtaining large data sets, the need to analyse sequences on a cellular level over stationary periods reduced the analysable sequences to a much smaller subset of the total. The billing data also suffered from not containing any hand-over information and
so did not display the full effects of population mobility. Further, no network topology information could be provided for Telecom's own commercial reasons. This prevented the selection of data from the most densely populated cells.

All the data obtained came from within New Zealand. Even New Zealand's largest city, Auckland is very sparsely populated by international standards. Therefore there is great scope for interesting traffic characteristics to develop in smaller densely populated cells in the world's major cities and these might never be seen on New Zealand based networks. Additionally, although penetration of mobile telephones into the general population is now significant, it is still much below the 80% level predicted for future PCNs.

The data obtained indicates that the new call arrival process of mobile networks has no significant differences from that of conventional telephony. Possibly it would be interesting to re-investigate this by concentrating only on small cells with small populations, but the nature of geographical aggregation means that higher network levels will see the conventional arrival process.

The BellSouth data does indicate that smaller cells with higher rates of handovers can cause higher relative fluctuations in the arrival rate to the cell. Using heuristic techniques, this strongly appears to suggest self-similar behaviour. Further research work is necessary to determine whether self-similar models are really appropriate to describe this behaviour or if some other model would be better e.g. a suitably modulated Poisson process. Additionally, it is not clear why such self-similar characteristics in the arrival process do not also appear in the channel occupancy process for the same cell.
Chapter 8

STOCHASTIC SIMULATIONS WITH LONG RANGE DEPENDENCE

8.1 SIMULATION STUDIES

There have recently been several detailed studies of telecommunications network traffic providing clear indication that many traffic types show self-similarity. This has lead to interest in generating synthetic random data sequences with these characteristics in order to perform stochastic computer simulations of queues and complete networks in order to understand the effects of this feature in the traffic load. One important aspect of such simulation experiments, which has not been analysed, is how to correctly interpret the results, given their particular statistical properties.

Leland et al [56] gave consideration to generating synthetic self-similar traces in their paper examining Ethernet traffic. They considered that using exact methods for generating Fractional Gaussian or ARIMA modelled traffic is too slow and therefore only suitable for short traces, not long simulation runs. This is indeed a significant problem, since self-similar traffic inherently needs to be simulated for a long period in order to see the effect of the long range autocorrelations. In addition, the large variances commonly associated with such traffic will require long simulation runs in order to produce useful confidence intervals on the mean.

Several methods of producing large self-similar sequences relatively quickly have been proposed. Mandelbrot proposed a technique called fast Gaussian noise based on a mixture of Markov processes which approximates the autocorrelation function of fractional Gaussian noise [63]. Leland et al [56] investigated a method based on the buffer occupancy of an $M/G/\infty$ queue where the G distribution has infinite variance (e.g. Paretian), since Cox had showed that this process is asymptotically self-similar [16]. They also used a convergence result from Granger showing that aggregating many, suitably chosen, simple AR(1)-processes, the superposition process is asymptotically self-similar [39].
More recently developed methods include a random midpoint displacement algorithm developed by Lau [26], techniques using chaotic maps by Pruthi [76] and various wavelet transform based methods [28, 58].

Simulations based on input data taken from actual traces of Ethernet traffic were performed by Erramilli et al [25] to demonstrate the implications of self-similarity on queueing performance. This approach has the advantage that it is not stochastic and so there is no need to determine confidence intervals on the results. However it also lacks generality and so, although useful for indicating trends and effects, it is difficult to apply the results of this study to other specific situations.

More general queueing simulations, aimed at ATM queues, were described by Mayor and Silvester [67]. They used an improved version of Mandelbrot's fast Gaussian noise technique to generate synthetic traffic traces a million sample points long. The results reported contained no indication of confidence intervals or precision.

8.2 STOCHASTIC SIMULATION ANALYSIS

The normal technique for assessing the usefulness of the results of a statistical sampling experiment is to form a confidence interval around the mean value of the sample data using a multiple of the standard deviation of the sample data found from the student's T distribution, given the desired level of confidence in the results [55]. Accurately assessing the sample variance is straightforward for uncorrelated sample data using the well known formula. However, all queueing simulations produce highly correlated data, since, whatever state one arrival finds the queue in, it is likely that the subsequent arrival finds the queue in a similar state. If the data displays long range dependence then these correlations are significant over all time-scales. Simple analysis of commonly used techniques for dealing with correlated data reveals that most fail to produce the desired results when dealing with long range dependent data.

8.2.1 Batch Means

Batch means is a simple technique whereby the original data sequence is grouped into "batches" and the mean of each batch found [31]. The means form a secondary data sequence which, providing the length of each batch is greater than the maximum correlation lag in the original sequence, will be uncorrelated and so the standard formula for variance can be applied. Unfortunately this technique
fails with long range dependent data because there is effectively no maximum lag of the correlations. Although very small, the extremely long-range correlations become significant when summed. In fact one of the standard ways of describing long range dependence is the scale invariance property which states that the autocorrelation function $r(k)$ remains unchanged under the operation of batching the given data sequence for any size of batch, $m$ [25].

$$r^{(m)}(k) \rightarrow r(k) \text{ as } m \rightarrow \infty$$  \hspace{1cm} (8.1)

Where $r(k)$ holds some significant, non-zero value.

### 8.2.2 Spectral Analysis

A more sophisticated approach to finding the variance of a sequence of observations is based on the fact that it is related to the value of the power spectral density function, $S_f$ at zero frequency. The technique of approximating the power spectral density in order to evaluate the variance is referred to as "spectral analysis" [45] and then

$$\hat{\sigma}^2 [\hat{X}(n)] = \frac{S_f(0)}{n}$$  \hspace{1cm} (8.2)

However, it is known that in the frequency domain, sequences with long range dependence obey a power law at the origin [56]:

$$S_f(\omega) \sim |\omega|^{-\gamma} \text{ as } \omega \rightarrow 0$$  \hspace{1cm} (8.3)

where $0 < \gamma < 1$. This is often referred to as $1/f$ noise since the effect is often observed with $\gamma$ close to one (e.g. flicker noise in oscillators).

The power spectral density then approaches infinity asymptotically as the frequency goes to zero. Therefore any technique for approximating the power spectral density cannot have a valid zero crossing and so the variance cannot be found.
8.2.3 Independent Replications

In the technique of independent replications the simulation experiment is run repeatedly with different, independent random input sequences. In this way the output of each different run of the simulation is independent of all the others and the sequence of results can be analysed. This is the only way of gaining independent observations for a simulation involving long-range dependence and it is often used as the basis for network simulation experiments. Typically, telecommunications traffic can only be regarded as statistically stationary for periods up to one hour, so network simulations are often run as repeated simulations of the busy hour traffic load.

However it is still important to analyse the properties of experimental observations to see if it is valid to form a confidence interval from them.

8.3 VARIANCE OF DATA SEQUENCES WITH LONG RANGE DEPENDENCE

There is an explicit formula to calculate the variance of the sample mean \( \text{var} [\bar{X}_n] \) which does account for the correlations of a data sequence [30], assuming wide sense stationarity:

\[
\text{var} [\bar{X}_n] = \frac{1}{n} \left[ \gamma(0) + 2 \sum_{k=1}^{n-1} \left( 1 - \frac{k}{n} \right) \gamma(k) \right]
\]

\[
\gamma(k) = E [(x_i - \hat{x}(n)) (x_{i+k} - \hat{x}(n))] \\
0 \leq k \leq n - 1
\]

As the formula shows, all the sequence auto-covariances, \( \gamma(k) \) up to the maximum length of the sequence are included in the summation for the true variance. From the theory of long range dependent sequences, we know the shape of the autocorrelation function \( L(t) \) is some slowly varying function, often assumed to be a constant) [25]:

\[
r(k) \to L(t)k^{-\beta} \quad 0 < \beta \leq 1 \\
k \gg 1
\]
Normally we assume that $L(t)$ is effectively a constant, say $C$. Since the autocovariances will also be of this form, it can be seen that the summation part of equation 8.4 converges finitely as $n \to \infty$,

$$
\lim_{n \to \infty} \frac{2}{n} \sum_{k=1}^{n-1} (1 - \frac{k}{n}) \gamma(k) \leq \lim_{n \to \infty} \frac{2C}{n} \sum_{k=1}^{n-1} k^{-\beta} \\
\leq \lim_{n \to \infty} \frac{2C}{n} \int_{0}^{n} k^{-\beta} dk \\
\leq \lim_{n \to \infty} \frac{2C}{n} \left[ (1 - \beta) k^{1-\beta} \right]_{0}^{n} \\
\leq \lim_{n \to \infty} 2C n^{-\beta} (1 - \beta) \\
\leq 0 \quad (8.6)
$$

Therefore the variance of the sample mean will be finite if the zero lag covariance, $\gamma(0)$, is also finite. This is same as the sequence variance for uncorrelated sequences. Unfortunately there are some distributions, which are associated with self-similarity which do not have finite variances. These are the heavy tailed or Paretian distributions [38, 74].

8.4 CENTRAL LIMIT THEOREM AND INFINITE VARIANCE

The Central Limit Theorem is normally used to show that sums of I.I.D. variables converge to a normal distribution. The usual assumptions are that the variable concerned has finite mean and variance. However Feller [29] gives a condition whereby finite variances need not be assumed providing the distribution of the variables lies in the domain of attraction of the normal distribution. The condition for this is that $U(t)$ is slowly varying as $T \to \infty$. Where $U(t)$ is given by:

$$
U(t) = \int_{-t}^{t} x^{2} dF(x) \quad (8.7)
$$

To see how this applies to heavy tailed distributions, it can be applied to the Pareto distribution, as given in equation 3.11:

$$
F(x) = Pr[X \leq x] = 1 - \left( \frac{x + k}{k} \right)^{-\alpha}
$$
\[ dF(x) = f(x)dx = \alpha k^{\alpha}(x + k)^{-(\alpha + 1)}dx \]  

(8.8)

Since \( F(x) \) is only defined for \( x \geq 0 \), \( U(t) \) can be found by parts integration.

\[
U(t) = \int_{x=0}^{t} x^2 \alpha k^{\alpha}(x + k)^{-(\alpha + 1)}dx \\
= k^{\alpha} \left[ x^2(x + k)^{-\alpha} + \frac{2x}{1 - \alpha} (x + k)^{1-\alpha} + \frac{2}{1 - \alpha} \int (x + k)^{1-\alpha} dx \right]_{x=0}^{t} \\
= k^{\alpha} \left[ t^2(t + k)^{-\alpha} + \frac{2t}{1 - \alpha} (t + k)^{1-\alpha} + \frac{2}{(1 - \alpha)(2 - \alpha)} (t + k)^{2-\alpha} \right]_{x=0}^{t} \\
U(t) = k^{\alpha} \left[ t^2(t + k)^{-\alpha} + \frac{2t}{1 - \alpha} (t + k)^{1-\alpha} + \frac{2}{(1 - \alpha)(2 - \alpha)} (t + k)^{2-\alpha} \right] 

(8.9)

Checking each term shows that they will only converge if \( \alpha > 2 \). Therefore for \( U(t) \) to be slowly varying as \( T \to \infty \), \( \alpha \) must be greater than 2. However this is the same condition for the distribution to have a finite variance. So when the distribution has infinite variance, there is no justification for applying the central limit theorem. Note however that, although the count sequence of Pareto spaced arrivals is self-similar (section 3.3.2), the individual members of this series are not from a Pareto distribution.

### 8.5 DISCUSSION

A survey of the techniques available for dealing with correlated simulation output data show that the only one with any possibility of dealing with long range dependent data from studies with self similarity is the method of independent replications. The method of independent replications will only work if the sample variance is finite. Obviously for a finite length sequence, the calculated sample variance will be finite (although possibly extremely large). However, as the sequence is lengthened, it will diverge if the variance of the underlying population is infinite. In practice, it seems likely that such a situation will produce confidence intervals too large to be useful. It is true however that truly infinite variance populations are rare in the real world. For instance, networks that have packet inter-arrival times with heavy tailed distributions often support processes that run on timers and, therefore, there is a maximum possible inter-arrival time. So it will often be possible to ensure that the simulations do have finite variances, yet are still realistic by placing restrictions on the theoretical models used. One
aspect that remains to be investigated is the effect of the traffic being filtered by passing through the system being investigated.

Although, there exists a condition under which the central limit theorem can be applied to distributions that do not have a finite variance, the Pareto distribution does not meet this condition when its variance does not exist. Thus it will not be able to form a confidence interval on the mean of a data sequence from a Pareto distribution with $\alpha < 2$. 
Mobile radio networks are experiencing rapid deployment rates and increasing levels of sophistication. They are popular for both the unique services they provide and the low overheads required for delivery of services to customers. Two major trends are evident in the evolution of such networks. Firstly, cell sizes are shrinking to provide extra capacity within the limited bandwidth available. Since this increases the ratio of cell boundary length to area it inevitably leads to dramatic increases in handover rates as a fraction of connection arrivals to individual base stations. Secondly, carried traffic is moving from being dominated by voice to a mixture of digital voice, video and data traffic requiring a wide mix of data rates. The need to efficiently carry such a range of traffic types while maintaining bandwidth efficiency is driving a move towards packet oriented transmission services. Both these trends have the potential to significantly alter the statistical characteristics of the network traffic.

This thesis has presented analytical derivations of the performance of fundamental network elements with a self-similar traffic arrival model. The blocking and waiting time performance of finite buffer single server queues is given along with the blocking performance of multiple server queues with no waiting room. The results obtained effectively characterise the loss in queue performance with increasing self-similarity of the arrival process and demonstrate the reducing sensitivity of this performance to queue length. Simulation demonstrates that the analytic blocking probability results for exponential service time queues are an upper bound on the performance of deterministic service queues.

Throughput-load relationships have been derived for the Aloha, slotted Aloha and CSMA random access protocols. These results have been used to give the delay performance of Aloha and slotted Aloha as well as the stability characteristics of slotted Aloha. These results also characterise the performance loss with the increasing self-similarity of the arrival process. They also demonstrate how
this performance loss reduces as the vulnerable period of the protocol is reduced.

The thesis also contains a unique analysis of the arrival process at individual live cell sites. This analysis, although limited, does suggest that handover arrivals should be characterised differently from the classical new call arrival model.

9.1 MOBILE RADIO NETWORK TRAFFIC

Self similar traffic will be seen on mobile radio networks. Data, video and signalling traffic have all been shown to be self-similar in a variety of studies and such traffic type will be increasingly carried on mobile radio networks. Packet based radio networks will face the same problems caused by such traffic as are being faced by the current generation of fixed networks. Additional effects may arise through the user population mobility, however the data does not yet support a firm conclusion on this.

New call arrivals are well characterised by the traditional Poisson arrival model, especially when aggregated at network switches. The handover arrival process is not so well understood. The total arrival process at the radio interface in a cell is a combination of the new call and handover arrival processes. It is only as cells become geographically small in busy metropolitan centres that the handover process begins to have a significant effect. The data gathered for this research is far from conclusive but it does indicate that in such situations the Poisson arrival model may not be adequate and that such traffic develops apparently self-similar characteristics.

The effects of handovers is the one call level traffic characteristic that is truly unique to mobile radio systems. As such, accurate characterisation of the effects on traffic of handovers are important to the engineering design of these networks. It is to be expected that handovers will increasingly dominate mobile radio traffic as cell sizes decrease to provide increasing network capacity.

9.2 SELF SIMILAR TRAFFIC ENGINEERING

Self-similar traffic models are parsimonious in that they can describe complex behaviour over a wide range of time scales with relatively simple formulae. However, they are not analytically tractable. The results obtained here are significantly more complex than the classical results based on the Poisson traffic assumption. Nevertheless, such analytical results are valuable. It is quicker to calculate new values with different parameters than to obtain new simulation results. Analyt-
ical values are accurate to much smaller values of blocking probability and they can provide more insight to the effect of process parameters.

Poisson "random" arrivals have often been assumed to be a worst case scenario for network traffic in terms of performance. Clearly this is not true. Self-similar traffic, in particular the Pareto spaced arrivals model results in consistently worse performance with higher losses and longer waiting times. The higher the self-similarity \( (H) \) parameter the worse this performance becomes. This degradation of performance has been demonstrated and quantified analytically here for single and multiple server queues and basic random access protocols.

Many fundamental network components such as switches and transmission lines can be characterised as queues and these results have significant implications for the design of systems using such components to carry self-similar traffic. In particular network blocking can be expected to be much higher than if the traffic is purely random. Increasing queue buffer sizes will not decrease the blocking probability as fast as for Poisson arrivals. At high loads this relative difference becomes larger. Clearly the large increases in buffer size that will be required to achieve reasonable blocking probabilities will result in large increases in the waiting time of packets transiting the network. Deterministic service queues have lower blocking than exponential service queues irrespective of the arrival process. The relative performance of the arrival processes is repeated in the results for multiple server queues with no waiting room. Aggregating servers into larger pools reduces the blocking less for self-similar arrivals than Poisson, although the differences decrease as the load increases.

Random access protocols are central to the operation of packet based multi-user radio networks. These too have significantly lower throughput capacity when carrying self-similar traffic. The option of increased buffering is not possible in such situations. As the protocols increase in sophistication, the vulnerable period in which collisions may occur with each transmission decreases. This decrease in the vulnerable period reduces the degradation in throughput due to the self-similarity of the traffic. Since the increased probability of collisions on the channel is due to the peaks in traffic, designers of systems will need to attempt to design their protocols to filter such peaks and spread them in time. This may be possible using the backoff delay and re-transmission scheme, trading increased delay for increased throughput. The characteristic lulls in self-similar arrivals provide an opportunity for the system to "catch up" and process backed off arrivals. This allows random access channels with self-similar arrivals to provide some throughput even at very high loads and increases the stability of the system.
provided it is operated within its maximum throughput.

9.3 RECOMMENDATIONS FOR FUTURE WORK

The use of the Pareto inter-arrival distribution to model self-similar source behaviour still has open questions. No study yet reported has been able to record the arrival characteristics of packets at the network interface. Rather studies have measured the arrivals of packets successfully transmitted onto the network at which point the arrival process has been filtered by the access protocol. So it would be useful to determine what the arrival process to the network is. Also, since real traffic is often nearly self-similar, it is worth asking whether such traffic is best modelled by an arrival time that is nearly hyperbolic, such as the large-variance Gamma distribution or whether a truncated Pareto distribution would be a more accurate representation.

The filtering effect of queues and multiple access protocols need further investigation. It can be shown that the hyperbolic tail of a distribution remains unchanged after filtering and therefore traffic with inter-arrival times distributed in this way remains self-similar. However network performance is often dominated by the behaviour of traffic peaks, the characteristics of which are dominated by the behaviour of the inter-arrival time distribution at small time intervals and this behaviour is definitely changed as packets pass through a queue or multiple access channel. Designing access protocols to spread arrival peaks in time may be a way to improve network efficiency, although it is not appropriate for all traffic types. Clearly also of interest is the effect of multiplexing multiple sources at downstream network elements. This would be necessary to the queueing results obtained here to the study of networks of queues.

The stability analysis of slotted Aloha shown here demonstrates that self-similar traffic may have some useful stability characteristics despite substantially lower maximum throughput. In practice, random access protocols have to be stabilised with control algorithms to ensure reliable throughput. Validation that these controls still maintain stability with self-similar traffic will be necessary for real use of these protocols. Optimisation of their performance, if possible would also be desirable.

Mobile radio systems using random packet access may also demonstrate the capture effect where two packets collide in time, but one is still successfully transmitted due to the difference in power seen by the receiver. This effect may change the results presented here and there is room to investigate whether such a change is significant and what it might be.
There is clearly still scope for a comprehensive investigation into the characteristics of mobile radio traffic behaviour in small cells where handover arrivals are dominant. Such an investigation would be best carried out by a researcher who is able to directly access the data collection system and use it to make repeated collection runs. As such it will require considerable co-operation from the network operating company.
REFERENCES


REFERENCES


