MODELLING OF PACKET TRAFFIC WITH MATRIX ANALYTIC METHODS

Allan Theodor Andersen

LYNGBY 1995
IMM-PHD-1995-18

IMM

ISSN 0909–3192

Printed by IMM - DTU Technical University of Denmark

Preface

This dissertation has been prepared at the Institute of Mathematical Modelling (IMM) at the Technical University of Denmark as a partial fullfillment of the requirements for obtaining the technical Ph.D. degree.

The dissertation addresses various modelling aspects of packet switched networks with applications in particular in the dimensioning of the future Broadband Integrated Service Data Network (B-ISDN). The emphasis has been put on versatile modelling of arrival processes using Matrix Analytic methods. Currently Public Switched Telephone Networks (PSTN) use the packet switched Common Channel Signalling System No. 7 to command/control the network. In some of the work done in this dissertation Common Channel Signalling System No. 7 has been used as an explicit case study.

The dissertation consists of a summary of 7 research papers written in the period 1993-1995. The research papers consider quite different aspects and engineering problems in packet switched networks. The treatment given of the different subjects has by no means been exhaustive but hopefully illustrates that Matrix Analytic methods are very powerful and versatile tools when solving complex engineering problems in packet switched networks.

Prior knowledge of both packetized communication systems and queueing theory in general, and the Matrix Analytic approach and Signalling System No. 7 in particular, will definitely make this dissertation easier to access.

Lyngby, September 1995

Allan T. Andersen

Acknowledgements

First of all I would like to thank my advisor Associate Professor, Dr. Bo Friis Nielsen for his support, guidance and ideas during the course of this work. Without any doubt he deserves all the credit for my interest in Matrix Analytic methods. Additionally thanks are due to Søren Blaabjerg and Alex Jensen who have been very involved in different aspects of this work and who have provided valuable support during the final phase of this study. Special thanks also to Dr. Dorte Møller with whom I have spent countless hours studying queueing theory in general and Matrix Analytic methods in particular. Of my other colleagues at IMM I would like particularly to thank my office mate through the last three years Mr. Niklas Kohl and also Dr. Jens M. Rygaard for sharing with me their insight in various aspects of life as a Ph.D. student. Furthermore, I would like to thank the faculty and administrative staff at IMM for providing an excellent research environment.

I would also like to acknowledge the financial support from the Danish Technical Research Council under grant 16-5132 which made it possible to carry out this study. Additional financial support from Valdemar Selmer Trane and Wifes Foundation (grant 51/94) , and Otto Mønsteds Foundation (grant 94-81-1077) made it possible for me to visit AT&T Bell Laboratories and Bell Communications Research, New Jersey, U.S.A. 3 months during the fall of 1994, this financial support is deeply appreciated. I wish to express my sincere thanks to, in particular, Dr. Ashok Erramilli for organizing my visit with Bell Communications Research and to Dr. Ronald A. Skoog and Dr. Douglas C. Schmidt for the hospitality extended to me at AT&T Bell Laboratories.

To my mother, Anni Karen Andersen (1942-1990)
In Memoriam

Summary

This dissertation is concerned with modelling various performance aspects pertaining to packet switched telecommunication networks. The emphasis has been put on versatile modelling of the packet arrival process which is a very relevant issue in context with the future Broadband Integrated Service Data Network (B-ISDN). As a case study of packet switched networks Common Channel Signalling System No. 7 (SS7) has been used. SS7 is the packet switched network currently employed to command/control the public switched telephone network i.e. the setup and teardown of trunks and the basis for intelligent network services i.e. 800 and 900 calls and advanced mobile communication services. The Markovian Arrival Process (MAP) has been used as a versatile tool to model the packet arrival process. Applying the MAP facilitates the use of Matrix Analytic methods to obtain performance measures associated with for example the single server queue with a MAP arrival process and a general service time distribution.

Measured SS7 traffic data has been analyzed as a part of this study. Recently there has been expressed concern regarding adverse behaviour of measured SS7 traffic i.e. long range dependence. Our studies did not reveal any adverse behaviour. In fact the observed traffic seemed very close to what would be expected from Poisson traffic. The Changeover/Changeback procedure in SS7, which is used to redirect traffic in case of link failure, has been analyzed. The transient behaviour during a Changeover scenario was modelled using Markovian models. The Ordinary Differential Equations arising from these models were solved numerically. The results obtained seemed very similar to those obtained using a different method in previous work by Akinpelu & Skoog 1985.

Recent measurement studies of packet traffic from Local Area Networks (LAN)

carried out at Bell Communications Research (Bellcore), New Jersey, have revealed behaviour very different from what conventional teletraffic models assume. In fact the observed traffic displays behaviour with (second order) self-similar characteristics over several time-scales. In this study we show that 8-16 state MAPs are able to capture this very variable behaviour over several time-scales. The queueing behaviour of these MAPs has been analyzed with Matrix Analytic methods. The results correspond to those obtained by trace driven simulations of measured LAN traffic. It is shown that two arrival processes with the same first and second order properties of their counting processes can yield substantially different queueing behaviour. This illustrates that one should by very careful when trying to assess queueing behaviour from the first and second order properties of the arrival process.

A heuristic formula for the tail behaviour of a single server queue fed by a superposition of renewal processes has been evaluated. The evaluation was performed by applying Matrix Analytic methods. The heuristic formula has applications in the Call Admission Control (CAC) procedure of the future B-ISDN network. The heuristic formula did not seem to yield substantially better results than already available approximations.

Finally, some results for the finite capacity BMAP/G/1 queue have been obtained. The steady state probability vector of the embedded chain is found by a direct method where a modified **LU** factorisation scheme is employed. The standard **LU** factorisation scheme is modified so that all computations are done cancellation free and in a manner that exploits the special structure of the embedded chain yielding substantial computational savings. The queue-length at an arbitrary time is found by noting the close relationship with the expressions for the corresponding infinite queue. For the special case of a batch Poisson arrival process this observation makes it possible to express the queue length at an arbitrary in terms of the corresponding queue lengths for the infinite case.

Resumé (In Danish)

Denne afhandling omhandler modellering af forskellige performancemæssige aspekter ved pakkekoblede telekommunikations netværk. Vægten er lagt på en alsidig modellering af pakke ankomstprocessen, hvilket er meget relevant i forhold til det fremtidige bredbåndsnet B-ISDN. Fælleskanal Signalerings System nr. 7 (SS7) er blevet undersøgt som et eksempel på et pakkekoblet netværk. SS7 er det pakkekoblede net, der i øjeblikket anvendes til at styre det offentlige telefonnet d.v.s. opsætning og nedtagning af telefonforbindelser og er basis for de intelligente nettjenester som f.eks. 800 og 900 numre samt avanceret mobilkommunikation. Den Markovske Ankomst Proces (MAP) er blevet anvendt som et alsidigt værktøj til modellering af pakke ankomstprocessen. Brugen af MAP muligør anvendelsen af Matrix Analytiske metoder til udregning af performancemål knyttet for eksempel til enkeltbetjenerkøen med en MAP ankomstproces og en generel betjeningstidsfordeling.

Målte SS7 trafikdata er blevet analyseret som del af det udførte arbejde. Fornylig har der været udtrykt bekymring angående at målt SS7 trafik udviste korrelationer over adskillige tidsskalaer. De SS7 trafik datasæt vi analyserede udviste ikke denne form for opførsel, men syntes at opføre sig som Poisson trafik. Changeover/Changeback proceduren i SS7, som bliver brugt til at redirigere trafik i tilfælde af fejl på en signallænke, er blevet analyseret. Den transiente opførsel under et Changeover scenarie blev modelleret med Markov modeller. De sædvanlige differentialligninger som disse modeller gav anledning til blev løst numerisk. De opnåede resultater synes at ligne dem, der blev fundet ved brug af en helt anden metode, i et tidligere arbejde af Akinpelu & Skoog 1985 meget.

Nylige målingsstudier af lokalnetstrafik (LAN) udført hos Bell Communications Research (Bellcore), New Jersey, har afsløret meget forskellig opførsel fra det traditionelle teletrafikmodeller antager. Den observerede trafik udviser opførsel med (anden ordens) selv-similære karakteristika over adskillige tidskalaer. I denne afhandling er det vist at 8-16 tilstands MAP'er er i stand til at modellere denne meget variable opførsel over adskillige tidsskalaer. Disse MAP modellers køopførsel er blevet analyseret med Matrix Analytiske metoder. De opnåde resultater svarer til dem, der er opnået med simulationer drevet af LAN trafikdata. Det er vist at to ankomstprocesser med samme første og anden ordens egenskaber for deres tælleprocesser kan udvise væsentligt forskellig køopførsel. Dette illustrerer at man skal være meget varsom med at forudsige køopførsel udfra første og anden ordens egenskaber for ankomstprocessen.

En heuristisk formel for haleopførslen i en enkeltbetjenerkø med en superposition af renewal processer som ankomstproces blev evalueret. Denne heuristiske formel har anvendelser i forbindelse med acceptering af opkald (eng. Call Admission Control) i det fremtidige B-ISDN netværk. Den heuristiske formel syntes ikke at give væsentligt bedre resultater end allerede kendte approksimationer.

Endelig blev der opnået nogle resultater for den endelige BMAP/G/1 kø. Ligevægtsandsynlighederne for den indlejrede Markovkæde blev fundet med en direkte løsningsmetode, hvor en modificeret **LU** faktorisering blev anvendt. Standard **LU** faktoriseringen er modificeret således at afrundingsfejlene bliver mindsket betydeligt og ligeledes er den specielle struktur af den indlejrede Markovkæde udnyttet, hvilket muliggør væsentlige beregningsmæssige besparelser. Kølængden til et tilfældigt tidspunkt blev fundet ved at bemærke den store lighed i udtrykkene for den tilsvarende uendelige kø. I det special tilfælde hvor ankomstprocessen er batch Poisson leder dette til at man kan udtrykke kølængden til et tilfældigt tidspunkt ved de tilsvarende kølængder i den uendelige kø.

Contents

Chapter 1

Introduction

Here approximately five years before the turn of the century it seems reasonable to assert that the telecommunications industry as a whole, i.e. both network operators and equipment manufacturers, faces challenges that are bigger than perhaps ever before in its history. There are several reasons for this, in the following a few of the more important will be mentioned. Currently the telecommunication market in Europe is being liberalized which means that operators, who used to have monopoly, are forced through major restructuring schemes in order to stay competitive. Alliances between previously independent network operators are being formed. At the same time many engineering problems associated with the change from circuit switched technology to packet switched technology, i.e. related to Broadband Integrated Service Data Networks (B-ISDN), remain largely unsolved even though the explosion in the last couple of years in the use of the Internet reveals that there is public demand for services of a type that one would expect to be main traffic contributors in a B-ISDN network. Also the explosion in mobile communications, for example Global System Mobile (GSM), and Intelligent Net (IN) services in the last couple of years has changed the profile of the traffic in the Public Switched Telephone Network (PSTN). Depending on the versatility of the network the steep increase in Signalling System No. 7 traffic (SS7) [70] due to these services might cause major problems for

some network operators. At the same time most forecasts anticipate an almost explosive economical growth in the telecommunications industry.

In this dissertation some of the engineering problems in dimensioning packet switched networks will be addressed. The use of Matrix Analytic methods can be seen as a common denominator of most of the work carried out in this dissertation even though the aspects studied are quite different in nature. Special attention has been directed towards applying the Batch Markovian Arrival Process (BMAP) [64] or the special case with no batch arrival, i.e. the MAP as a very versatile modelling tool for arrival processes.

This dissertation consists of a summary report and a collection of 7 research papers. The work carried out can naturally be grouped in four different parts with the seven research papers quite unequally distributed among the parts.

The first part concerns modelling and analyzing specific issues in the currently employed Common Channel Signalling System No. 7 (SS7) network which is the packet traffic net used as the command/control net in PSTN and the basis for modern mobile communication and IN services. Here some attention has been devoted to analyzing the packet arrival process of SS7 messages to see whether the traditional Poisson assumption still seemed valid. In particular a lot of short SS7 data traces from KTAS' (Tele Danmark) SS7 network and longer traces from AT&T's network have been analyzed. In paper [A1] a specific SS7 feature is modelled namely the Changeover/Changeback procedure [3],[92] which is invoked for example in case of link failures in the signalling network to ensure that no signalling messages are lost. A transient analysis of a Markov modelled Changeover scenario is carried out in [A1].

Secondly, the interest has been directed towards modelling very variable packet traffic. This work is partly motivated by recent measurement studies carried out at Bell Communications Research (Bellcore) which reveal that LAN traffic [35] and Variable Bit Rate (VBR) Video [17] show variability over several timescales. The measured traffic in fact shows characteristics that often are associated with self-similar behaviour. In the context of B-ISDN it seems quite relevant to be able to model this very variable traffic since for example the interconnection of Local Area Networks (LAN) and Variable Bit Rate (VBR) video are expected to be main contributors to the traffic in B-ISDN. The papers [A2], [A3] and [A4] are concerned with modelling this variable traffic and investigating the performance implications, i.e. queueing behaviour, that this traffic has.

The third part concerns modelling of the tail behaviour of queues fed by a superposition of renewal streams. The motivation for considering this type of arrival process is that a typical source model is a bursty renewal process. The reason for the interest in the tail behaviour of the queue can be contributed to the very strict Quality of Service (QoS) requirements in the future B-ISDN network i.e. loss probabilities $\leq 10^{-9}$. Applying Large Deviations theory [23] and the concept of effective bandwidth [51], [90] knowledge of the tail behaviour of the queue can be used to design a quite simple Connection Admission Control (CAC) scheme which using the effective bandwidth concept decides whether a given connection request should be satisfied or not. Papers [A5] and [A6] propose and evaluate approximations for the tail behaviour of the queue and illustrate how this knowledge can be used in the context of effective bandwidth.

Finally the fourth part is a summary of paper [A7] which is a rather theoretical paper that describes the close relationship between the finite [20] and infinite BMAP/G/1 queue [64], [83] and describes a scheme for the efficient and cancellation free computation of the queuelength distribution in the finite capacity queue.

The rest of this summary report is organized as follows: In section 2 some general comments on the modelling of variable packet traffic are given together with an introduction to the Matrix Analytic methods applied. In this section an attempt is also made to unify the work done within the framework of modelling B-ISDN issues. Sections 3 - 6 form the core of this summary report as they summarize the work carried out ordered in the 4 different parts previously described. Finally in section 7 some conclusions are drawn and suggestion for further work are given.

Chapter 2

On the versatile modelling of variable packet traffic

2.1 The classical communication systems modelling approach

Traditionally the Poisson process has been used extensively as a model for arrival processes in telecommunication systems or more generally in queueing systems. In the current telephone network (PSTN) the dimensioning of trunks is usually done by applying a Poisson arrival process model with intensity parameter defined by the number of calls offered in the busy hour which implies that simple formulas as for example Erlangs B and C formulas [22] suffice for dimensioning purposes. While the Poisson process describes the number of offered calls in a classical telephone network i.e. Plain Old Telephone Service (POTS) quite well, it is often used as a model more due to the analytical simplications which arise from it than how well it actually models the system at hand.

The trend in communication network architectures, as in so many other areas, is increased complexity. With regards to modelling this gradually introduces more variable traffic patterns than what has traditionally been seen in for example

POTS. The increased complexity is due to several factors, most notably probably the versatility introduced in changing from circuit switched technology to packet switched, allowing the individual connections to transmit variable bit rate traffic. However, the effect of increasingly complex layered protocols as traffic sources should not be underestimated either.

Clearly one of the implications of this trend is that there is an increasing need for more versatile tools for modelling traffic patterns than what can be provided by the Poisson process. However, a major obstacle in applying alternative models has been the lack of analytical tools to handle them. The G/G/1 queue for instance, which is the most relevant queueing model in context with the work in this dissertation, has traditionally almost exclusively been analyzed in the special cases where either the inter-arrival times (M/G/1) or service times (GI/M/1) are exponential, that is, in the presence of some Markovian structure (embedded points). For completeness it should be mentioned that some work has been done applying non-Markovian models as for example diffusion models and very recently models for packet traffic with self-similar characteristics have appeared, namely models based on Fractional Brownian Motion [81] and Chaotic Maps [31].

2.2 Matrix Analytical Methods

In the last 15 - 20 years there has been done a considerable amount of work in extending results from the M/G/1 and GI/M/1 queues to much more general models. The idea has been to generalize the Poisson arrival process and exponential service time distribution to more versatile models while maintaining Markovian properties, this has been done, loosely formulated, by generalizing the intensity parameter to matrices of intensity parameters. By selecting these intensity parameters appropriately it is possible to model very general service time distributions and arrival processes including processes with very complex correlation structures. Systems arising from these matrix generalizations are said to be of M/G/1-type respectively GI/M/1-type. Much of this pioneering work is due to M.F. Neuts and co-workers and is described in detail in the two monographs [75] and [77]. Unified these method are called matrix analytic methods. In this dissertation these relatively new tools have been applied with the focus primarily on versatile modelling of the arrival process which has made it natural in particular to consider M/G/1-type queues.

2.3 Signalling System No. 7 as a case

In this dissertation we use Signalling System No. 7 (SS7) as the case study of packet traffic. Signalling System No. 7 is the system that is currently used in the Public Switched Telephone Network (PSTN) to command and control the network. It facilitates the setup and tear-down of trunks in ordinary telephone calls and is the basis for IN services as freephone and also essential in advanced mobile communications for example GSM. For telephone networks SS7 can be percieved as the analog of the central nervous system for humans. In PSTN SS7 is logically separated from the circuit-switched network i.e. the bearer network. By far most of the SS7 traffic seen today can still be contributed to plain old telephone service (POTS). A typical scenario in POTS is the following: Suppose a subscriber A wants to call subscriber B. After subscriber A has finished dialing, a message (called an IAM) is transmitted in the SS7 network towards subscriber B. The purpose of this is to setup a bearer trunk between the two subscribers and of course to alert the called subscriber, (possibly there are several transit exchanges so that trunks have to be set up between each exchange). When the message reaches subscriber B a SS7 message is launched back towards A to indicate that the address has been completed (ACM) and finally a message is sent from subscriber B towards A at the time B (hopefully !) answers. The same type of scenario takes place at the end of a conversation depending on who hangs up first i.e a few SS7 messages are sent to release the trunks. Basically the SS7 traffic created by such a call is a couple of messages sent in each direction during the setup and tear-down phase. Even though the SS7 messages transmitted due to a given call are correlated it appears that since the number of messages created by each call is extremely small compared to both link capacity and total signalling traffic the total signalling traffic observed on a link should be very close to Poisson which indeed also has been shown to be the case in [88]. (Here it deserves mentioning that the signalling traffic is typically transmitted at link speeds of 64 kbit/s or 56kbit/s and that the maximum allowed length of signalling messages is 272 octets).

With the introduction of Integrated Service Data Network (ISDN), IN and GSM

more complex SS7 traffic patterns have been introduced. Basically IN services in todays network introduce additional signalling in the setup and tear-down phase compared to an ordinary call, this typically consists of database queries to get 800 and 900 numbers translated and perhaps billing information. In the IN services today the increase in signalling compared to a simple POTS call is usually very limited, however, it is conceiveable that in the future more complex IN services will involve additional signalling traffic during a call. In GSM the signalling compared to POTS is increased by several factors and for a given call it mainly takes place in the setup and teardown phase. The additional signalling is due to for example locating the mobile subscriber and user authentication. There is, however, also much additional signalling outside calls to keep track of the location of mobile subscribers. In the present implementations of IN and GSM it seems that the problems that have been seen in networks are more due to the vast increase in the volume of especially GSM traffic and perhaps bottleneck problems created for example by very centralized registration of mobile subscribers or exchanges supporting IN services than due to adverse signalling traffic behaviour. The signalling protocol for ISDN supports several enhancements compared to POTS. From a signalling network performance point of view the most interesting is the possibility of sending user-to-user information over the signalling network during a call, this feature could for example be used for passing control information if the ISDN call consists of a data transfer between two computers. In theory the result could be that individual connections send long bursts of user-to-user signalling traffic. Clearly, unrestricted user-to-user signalling could prove fatal for the network and currently the vast majority of networks do not support the transmission of more than a few user-to-user signalling messages if any. Recently a control scheme has been proposed in [55] to prevent user-to-user signalling from degrading network performance.

In the next chapter a short summary is given of some of the work done within this study on characterizing measured SS7 traffic. The purpose of this very superficial description of SS7 is to describe what kind of traffic patterns are predominant in the network more than going into any detail with precisely what messages are transmitted. For a much deeper treatment see [70] which is a good tutorial to SS7 or the ITU SS7 recommendations [92].

2.4 Future communication systems

Currently the signalling and bearer traffic are logically separated i.e. the signalling traffic is either transmitted on dedicated trunks or in dedicated timeslots (typically timeslot 16). With the introduction of B-ISDN this might change since one likely implementation is that both bearer traffic and signalling traffic will be multiplexed onto the same very fast physical connection [49] i.e. they will actually compete for the same bandwidth. In B-ISDN all traffic will be transported in ATM cells which are 53 octets long. The intermediate protocol layers in B-ISDN (ATM Adaption layers [49]) perform the necessary segmentation/reassembly of the payload. On the ATM layer where all information is packed into these 53 octet cells there is only one priority bit to distinguish between the individual calls. This implies that performance of the signalling traffic might become extremely closely related to the performance of the bearer B-ISDN traffic and hence adverse behaviour of the latter could affect the performance of the signalling traffic.

As previously mentioned the next chapter is the only one specifically concerned with modelling Signalling System No. 7. The other work presented here pertains to packet traffic more generally, however, with clear emphasis on modelling traffic characteristics that are very important in the context of B-ISDN, that is, for example modelling of Local Area Network (LAN) with variability over several timescales and evaluating tail behaviour of queues which has applications in Call Admission Control (CAC) for B-ISDN.

In this light it can be argued that general modelling of B-ISDN traffic issues is extremely relevant also in the context of accessing performance of the signalling traffic in B-ISDN.

Chapter 3

Performance aspects in Signalling System No. 7 packet traffic

In this chapter the focus is on two quite different aspects of SS7. First a brief description is given of rather detailed models of the Changeover/Changeback procedure in signalling system No. 7 which have been developed in [A1] and secondly some of the findings in SS7 measurement studies carried out in both Tele Danmarks and AT&T's network will be described.

3.1 Modelling of Changeover/Changback scenario in SS7

The Changeover/Changeback procedures are the SS7 procedures invoked in case of link failure to re-direct signalling traffic in a manner which ensures that no signalling messages are lost. When a signalling link failure is detected in SS7 the Changeover procedure is invoked. The Changeover procedure redirects the

traffic from the failed link to alternative links avoiding loss, missequencing and duplication of messages. A message retrieval procedure moves the messages from the failed link to alternative links. Changeback is the corresponding procedure for putting a link back in service. Both procedures can result in significant queues of messages before message retrieval is initiated. For a Changeover in the transmit buffers associated with the failed link, for a Changeback in the transmit buffer associated with the alternative link. This is due to the time it takes to accomplish the handshaking or timeout procedures before the message retrieval procedure can start. The Changeover scenario has the potential for the largest build-up of queued messages because it also takes time to detect link failure. When a message retrieval procedure is started this prospective queue is propagated through the network. For a Changeover this can result in queues in the transmit buffers associated with the alternative links, and for a Changeback in a queue in the transmit buffer at the link ready to re-enter service.

In order for these procedures to have any meaning careful engineering is needed when implementing the signalling system in practice. The buffers should be large enough to contain all the messages that could arrive during these procedures. The time it takes to perform a Changeover/Changeback should be sufficiently small to avoid unacceptable message delays. The potential risk of having too small buffers could be discarded messages. Another possibility is that a congestion situation is declared due to the buffer fill resulting from the Changeover/Changeback procedures in which case traffic control procedures are invoked. Clearly it is not desirable that a signalling link failure can influence the rest of the signalling network in this manner.

Previously in [3] an analysis of the transient queueing behaviour during a signalling link failure/recovery event has been carried out. A Changeover scenario with one alternative link was modelled. Under normal operating conditions Poisson traffic was assumed for each of the signalling links. The message retrieval procedure was modelled by a fixed bit-rate message transfer from the failed link to the alternative link. During the signalling link failure/recovery event the traffic was found as a superposition of the message retrieval traffic and the normal Poisson traffic. During message retrieval, in a Changeover scenario, this has the following implications: At the transmit buffers associated with the failed link messages arrive according to Poisson and messages are removed at a fixed bit-rate determined by the message retrieval speed. At the transmit buffers associated with the alternative link messages arrive as a superposition of the message retrieval arrivals and the normal Poisson traffic. Messages are removed at link speed. When the transmit buffers associated with the failed link empties the Changeover is completed. The total load destined for these two links is now carried by the alternative link. A normal distribution approximation was made for the number of Poisson traffic messages arrived during the Changeover/Changeback.

We consider two models which, as in [3], assume Poisson traffic under normal operating conditions. The message service (emission) time is modelled using a phase type distribution [75]. The message retrieval traffic is modelled by scaling the phase type distribution for the message service (emission) time. The magnitude of the scaling is determined by the rate at which the message retrieval procedure transfers messages from the failed link to the alternative link. We consider the same Changeover scenario as [3]. Both our models are strictly Markovian. The transient distribution of queue-lengths in the transmission buffers can therefore be calculated by solving the ordinary differential equations arising from the Markov transition matrix numerically [36] and hereby obtaining quantiles of the delay and queue-length distributions. Clearly a practical problem here is the large state space necessary to model a system with realistic buffer sizes. To make a numerical method applicable in practice the sparsity of the Markov transition matrix has to be exploited. We solve the ordinary differential equations numerically by employing an implicit multistep method. The method has variable order and steplength which are selected during the course of the numerical integration.

The first rather approximate model is only concerned with the queue in the transmission buffer at the alternative link during a Changeover scenario. During the message retrieval procedure the arrival process at the alternative links transmission buffer is a superposition of the message retrieval traffic and the ordinary Poisson traffic. When the message retrieval procedure has emptied the transmit queue associated with the failed link the arrival process at the alternative link changes. The arrival process then consists of a superposition of the links ordinary Poisson traffic and the Poisson traffic normally carried by the failed link. That is, a Changeover scenario is modelled using one Markov transition matrix during the message retrieval procedure and another transition matrix upon emptying the transmit buffer associated with the failed link. Numerical results from this first model agree well with the previously obtained results in [3], see [A1] for details.

The second model is a full Markov model of a Changeover scenario including both the transmission buffers at the failed and alternative links. The buffer sizes used at the links determine the order of magnitude the state space has been enlarged compared to the previous model. If B_{max} denotes the maximum number of messages which can be contained in a buffer, then this model has of the order B_{max} more states than the previous model. The initial probability vector here is determined by the queue-lengths at the time of Changeover. An obvious choice would here be to use the stationary distribution for the alternative link and a distribution indicating a large build-up of messages at the failed link. Numerical results pertaining to this model have not been obtained.

Some concluding remarks about the work done on modelling the Changeover/-Changeback procedure in SS7 are:

- Two Markovian models have been proposed for the Changeover/Changeback scenario previously analyzed in [3].
- Numerical results from the first model agree well with previously obtained results in [3] which indicate that the normal approximation employed there is sufficiently accurate.
- No numerical results where obtained for the full Markov Model due to dimensionality problems. It seems that another solution procedure is needed so that numerical integration is avoided. Randomization [48] seems to be a good alternative. Essentially randomization exploits the close relationship between the Markov transition matrix and the corresponding discrete Markov chain so that numerical integration is avoided.
- At the expense of larger state space it easy to analyse more general arrival process models with correlations as long as they still have a Markovian structure as for example Markovian Arrival Processes (MAP)

3.2 Findings in SS7 traffic measurement studies

Since the beginning of the 1990's there has been a growing interest in analyzing measured packet traffic traces from various types of networks. This interest has been triggered by increasing concern with regard to whether commonly held assumptions about the packet traffic actually were valid as for instance the typically held assumption of a Poisson arrival process. That this concern definitely is warranted has recently been illustrated in various publications regarding findings in measurement studies, most notably the publications regarding the ongoing study of Local Area Network traffic (LAN) [33],[35], [61] and [91] carried out since the late 1980's at Bell Communications Research (Bellcore) New Jersey. The Bellcore measurement studies reveal very adverse behaviour. The LAN traffic displays extreme variability over several time-scales, slowly decaying variances and in fact behaviour that is often associated with self-similarity. One implication of this behaviour is that dimensioning after some mean load requirement can prove fatal because it does not take the behavioural dynamics into account and clearly the Poisson arrival process assumption is not valid.

A natural question to ask oneself in this context is of course if this very adverse behaviour extends to other types of packet traffic networks. This was one of the reasons for initiating measurement studies in the Danish SS7 network during the fall of 1993. KTAS supplied measurements and technical expertise for this work that was carried out as a Master's project [15]. The traffic measured in the KTAS net was almost exclusively POTS traffic. At the time of the measurements KTAS introduced IN services in the network, however, these were implemented very centralized with only one IN exchange so that all operations pertaining to IN were actually performed inside this exchange i.e. no external IN signalling traffic was generated. A problem in this measurement study was that the monitoring equipment available only had memory available to monitor relatively short traffic traces around 15 minutes of SS7 traffic. Also KTAS SS7 network is very lightly loaded most links have utilizations that are significantly less than 0.1 Erlang. The main findings where the following.

- The initial SS7 messages in calls, the Initial Adresse Messages (IAM) were reasonably well described by the Poisson process and they did not seem to be correlated.
- A hypothesis of a Poisson arrival process was rejected for the total process of transmitted SS7 messages. Runtests rejected the hypothesis of the total SS7 message process being uncorrelated.

An in depth description of the measurements is given in [15] where also mea-

sured delays through an exchange (Cross office delays) for various SS7 messages were analysed. Our quite modest measurement study revealed another important aspect of performing measurements. Some of the traffic traces revealed an unusual amount of SS7 messages that indicated an error state in the network and the corresponding messages correcting the errors. It turned out that KTAS actually were experiencing some problems at the time regarding the passing along of B-subscriber numbers to a private mobile operator. So monitoring the signalling traffic is a good way to detect potential problems in the network at an early stage. The same benefit of monitoring signalling traffic is mentioned in [29] where calls were detected that did not violate the SS7 protocol but clearly displayed undesirable properties by sending several hundreds of signalling messages during a call.

Our measurement study of SS7 traffic is rather small compared to the one carried out in [29] (at Bellcore) where links where monitored that were relatively heavily loaded up to around 0.30 Erlang during the busiest hour. Also their test equipment facilitated the monitoring of much longer traces than in our case. In [29] it is concluded that the observed signalling traffic displays some of the annomalies seen previously in the Bellcore LAN measurements [33],[35], [61] and [91] that is long range dependence and slowly decaying variances. It is asserted that the measurements indicate that the call holding times are heavy tailed. It is also asserted that it is problematic to define a good reference period for measuring corresponding to the notion of busy hour because the data did not seem to support the idea of a stationary busy period. These findings in the SS7 traffic by Bellcore and made public in 1994 for example in [29] invoked growing concern in ITU's Study Group 2 both with regards to how in general long range dependence would affect performance of the signalling net and how to define measurement periods in a meaningful manner.

During a visit at AT&T Bell Laboratories, Holmdel, New Jersey in the fall of 1994 I had the oppurtunity to analyze SS7 traffic data of comparable quantity and quality as those presented in [29]. A delicate issue when checking the SS7 traffic for self-similarity or long range dependence is that well known trends in the data, i.e. hourly variations of the traffic intensity, have to be removed in order to make the analysis meaningful. So the data should be and was detrended before carrying out any analysis regarding self-similar behaviour. The signalling data was evidently also detrended in [29]. As in [29],[61] and [91] the focus was on second order properties of the traffic, that is, whether or not the traffic displayed second order self-similarity. The statistical tools applied were the variance-time curve and the more heuristic R/S plot. The variance-time curve method is described in the following chapter and the R/S plot is described in [61], [91] and [A2]. In the present chapter only the analyzed SS7 traffic data and the main results will be described. The traffic measured was mostly POTS (in most traces around 90 %) and the rest of the traffic consisted of simple IN queries. There were twelve traffic traces available, some of them more than one hour long and with utilizations between 0.1 and 0.2. The statistical analysis of the second order properties of the data partitioned in each of the described ways did not reveal any significant second order self-similar behaviour. In fact the (detrended) data sets seemed to show behaviour very close to what would be expected from Poisson traffic.

To address the issue of how to select measurement intervals, and the concern expressed in [29] regarding this, the average load over 15 minute intervals for each of the twelve data traces was calculated. For each 15 minute interval the average load over a sliding window of width τ was computed for window placements at intervals $(0, \tau)$, $(\tau, 2\tau)$, etc. For each window size the maximum sliding window average was determined, and the ratio of this maximum sliding window average to the 15 minute average load was calculated. This ratio is often called the peak-to-mean ratio. The window widths examined were 1.0 sec, 5.0 sec, 10.0 sec, 20.0 sec, 40.0 sec, 80.0 sec, 100.0 sec. and 150.0 sec.. This analysis was carried out for both the POTS traffic and IN traffic in isolation as well as for the total traffic and in all cases the traffic was not detrended. The peak to mean ratios were compared to those expected for Poisson traffic and batch Poisson traffic with batch size 2. For the POTS traffic, especially for small window widths (1.0 sec. - 20.0 sec.), the observed traffic had a higher peak-to-mean ratio than Poisson, usually around 5 - 20 % higher. A batch Poisson model with batch size 2 seemed to provide a worst case measure for the observed traffic. Using a batch Poisson model can be justified by the fact that for a given call all SS7 messages related to the setup are transmitted within a quite narrow time window and the same goes for the messages related to teardown. For the IN traffic the peak-to-mean ratio for all examined window sizes was actually smaller than the corresponding ratio for Poisson. This corresponds to the simulation results presented in [13] where it is shown that long messages tend get smoothed as they go along a path (IN related signalling messages are typically at least a factor 2 longer than POTS related signalling messages.). To

summarize the findings in AT&Ts signalling network:

- There was no sign of the signalling traffic displaying significant second self-similar behaviour in any of the twelve traces examined.
- The peak-to-mean ratio of the traffic showed that the POTS traffic displayed more variable behaviour than the Poisson process, however a batch Poisson process with batch size 2 did seem to describe worst case behaviour so that from an engineering point of view there did not seem to be any reason for immediate concern with regard to applying 15 minute measurement intervals and from this extracting information about the behaviour in shorter time intervals.

The results from these measurement studies are presented in [12] and [5].

Chapter 4

Packet traffic with variability over several timescales

The work presented in this chapter is a summary of the work carried out in [A2],[A3] and [A4].

As mentioned in the introduction recent measurement studies of packet traffic from Local Area Networks (LAN) [35], Variable Bit Rate (VBR) Video [17] and Integrated Services Digital Network (ISDN) [57] reveal behaviour very different from what classical traffic models assume. The measurements indicate that these forms of packet traffic have long range dependencies and slowly decaying variances. Characteristics of the observed behaviour have in other contexts been associated with self-similar behaviour. Therefore the measurements motivate the assertion that certain types of packet traffic exhibit self-similar behaviour over several time-scales. The inter-connection of LANs is expected to be a main traffic contributor in the future B-ISDN network which makes LAN traffic behaviour a very relevant issue also in this context.

Packet traffic with self-similar behaviour has serious implications for both system engineering and system performance. It is problematic to design/dimension systems only with respect to mean load requirements obtained by simple models since such an approach totally fails to take into account the behavioural dynamics

associated with self-similarity. The challenge is here to develop models that can describe this adverse behaviour reasonably and are able to quantify new performance measures and dimensioning tools.

This has recently initiated a search for traffic models which are able to describe this behaviour in a relatively simple manner. Here models well known in other contexts have been investigated i.e. Chaotic Maps [31] and Fractional Brownian motions [81].

While arrival processes based on such models can be described using only a few parameters it is analytically very difficult to analyze a queueing system with these processes as input.

The articles [A2],[A3] and [A4] are concerned with modelling these observed phenomena with an analytically tractable and already available versatile point process namely the Markovian Arrival Process (MAP) [64]. The MAP is the special case of the BMAP with batch size one (the BMAP is equivalent with Neuts' versatile point process [74]). MAPs are sufficiently general to serve as approximations for any stationary point process, see for example [11].

One of the main merits in using a MAP model is that a considerable amount of work has already been done on analyzing the queueing behaviour associated with the MAP. Much of this pioneering work is due to M.F. Neuts and co-workers and is described in detail in the two monographs [75] and [77]. The single server queue driven by a BMAP was originally analyzed in [83] by V. Ramaswami using Neuts' versatile point process notation and further in the more transparent BMAP notation by D.M. Lucantoni in [64]. In the context of describing queueing behaviour as a function of the load for a given MAP a very interesting article is [76] where the caudal characteristic curve is introduced. It contains results regarding the asymptotic tail behaviour of queues with a matrix geometric steady state vector [75] as for example a queue driven by a MAP and with a phase-type service time distribution (A special case of the QBD process [75]). To a certain extent this work has very recently been generalized to queues of M/G/1-type [77], see [1] and particularly [26] for details.

This chapter summarizes results from [A2],[A3] and [A4] regarding the use of the MAP as a model for packet traffic with self-similar behaviour spanning several timescales. The analysis carried out on the measured packet traffic with self-similar properties, see for example [61] or [91], has mainly been focused on second order properties of the packet traffic as for example the variance-time curve. Additionally plots of the evolution in time of the number of packets arriving in a given size timeslot have been examined at different aggregation levels (size of the timeslot). In [61] this has been used to assert that the packet traffic data appears to be "visually" self-similar. In [A2],[A3] and [A4] the same approach has been applied when examining MAP models. Where possible analytical results have been used for example for the second order MAP properties but simulation has also been applied . In particular simulation has been used to create artificial "data-traces".

While being the basis for the extreme versatility of the MAP the heavy parameterization is a major practical problem when considering all but the simplest MAPs. For our purposes we have therefore chosen to restrict ourselves to a certain subclass of MAPs and even of the MMPPs (Markov Modulated Poisson Process) which we will call the General Multi Level Process (GMLP). The main advantage of restricting ourselves to the GMLP is indeed that the number of parameters are more managable than in the general case. The GMLP maintains some of the MMPPs versatility regarding modelling rather complex correlation properties i.e. it is not limited to modelling superpositions of simple MMPPs as for example IPPs (Interrupted Poisson Process). In fact some special cases of the GMLP are for example superpositions of IPPs (homogeneous and/or inhomogeneous) and perhaps a Poisson stream. The models examined in this summary will, however, be restricted to these special cases and the interested reader is refered to [A2] for the definition of the general GMLP model and more general GMLP models than the ones examined here. The GMLP model was originally presented in a Masters Thesis [47] by Alex Jensen.

In this summary the time dynamic behaviour of different IPPs is examined. The resulting process obtained by superposing several identical IPPs is also looked into. Next the resulting process obtained by superposing several inhomogeneous IPPs and a Poisson stream is considered. Essentially the same models as in [A4] are used to illustrate the work done in [A2],[A3] and [A4].

In this summary the mean queueing behaviour of the presented MAP models is also considered. As a part of this Ph.D. study the algorithms for the single server queue with a Markovian Arrival Process (MAP) outlined by D.M. Lucantoni in [64] and originally by V. Ramaswami in [83] have been implemented. Queueing analysis of the LAN traffic traces has been initiated in [32] where the queueing

behaviour of a queue driven with LAN traffic traces has been examined. One of the findings was that the queueing behaviour seemed relatively insensitive to marginal properties of the traffic.

Preliminary results of this work was originally presented at the COST 242 MC meeting in Cambridge January 1995 [7]. At the subsequent COST 242 MC in Stockholm May 1995 preliminary results regarding the queueing behaviour of MAP models with self-similar characteristics were presented [8]. At the same meeting related work about the use of Markovian models to model self-similar behaviour was presented [87]. Very recently related work was presented in [42] were a multi-level source model is used to model self-similar behaviour. This model is related to our GMLP model but does not possess a full Markovian structure. Essentially [A2] is an enlarged and revised version of [7] and [A3] is a revised version of [8].

4.1 Measures of Self-Similarity

The short introduction to self-similarity given here is primarily based on [61] and [91]. In the context of packet traffic we are concerned with the process defined by the number of packets arriving in a time-interval. This makes it natural to consider the stationary discrete time stochastic process $X = X_t$ for $t = 1,2,....$ (Number of packets arriving in time-interval 1,2 ....). We also consider the time-series obtained by averaging the discrete time process over non-overlapping blocks. That is, for $m = 1,2,3$ .., we consider the time-series

$X_k^{(m)} = (1/m)(X_{m(k-1)+1} + ... + X_{mk}),\ k = 1,2,3,....$

The stationary discrete time stochastic process X with mean 0 and variance σ^2 is said to be self-similar with self-similarity parameter H if for all $m = 1,2,3....$ $(X_{m(k-1)+1} + ... + X_{mk}) = mX_k^{(m)}$, $k = 1,2,3,...$ has the same finite dimensional distributions as $m^H X$.

If the variance and autocorrelation function of $(X_{m(k-1)+1} + ... + X_{mk}) = mX_k^{(m)}$, $k = 1,2,3,...$ and $m^H X$ are the same for $m = 1,2,3,....$ X is said to be second-order self-similar with self-similarity parameter H. This implies that for all $m = 1,2,3$ $var(X^{(m)}) = \sigma^2 m^{-\beta}$, and that $r^{(m)}(k) = r(k) = \delta^2(|k|^{2-\beta})$ $(k = 1,2,3...)$ where $\beta = 2 - 2H$ and $\delta^2(f)$ denotes the second central difference operator applied to a function, that is, $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$. X is called asymptotically second-order self-similar with self-similarity parameter $H = 1 - \beta/2$ if $r^{(m)}(k) \to \delta^2(|k|^{2-\beta})$, as $m \to \infty$ (k = 1,2,3,.....)

The definition of long-range dependence is that $\sum_k r(k) \to \infty$. For a self-similar process this corresponds to the self-similarity parameter H in the range $0.5 < H < 1.0$. For such a process the averaged time-series $X^{(m)}$ possesses a non-degenerate correlation structure as $m \to \infty$. For all conventional teletraffic models $r_m(k) \to 0$ as $m \to \infty$, (k=1,2,3,...), that is, the averaged process $X^{(m)}$ tends to second order pure noise for large m corresponding to $H = 0.5$ or equivalently $\beta = 1$. In the following a process with $H = 0.5$ will be called a short-range dependent process.

In practice there are several ways to try to detect whether an observed process exhibits self-similar behaviour or not. A simple way to check for any indications of self-similarity is to plot the averaged process $X^{(m)}$ for different values of m, m values differing by an order of magnitude. If the plots "look the same" it is an indication of a self-similar behaviour.

In the context of packet traffic attention has been directed towards three ways of checking for self-similarity. These can be referred to as methods based on the variance-time plot or equivalently the index of dispersion of counts, the pox diagram of the R/S statistic (also called the rescaled adjusted range plot) and finally analysis based on the periodogram (in the frequency-domain). Here only the method based on the variance-time plot is considered. Regarding the statistically more refined methods available when analyzing the periodogram (i.e. the Whittle estimator) see [61] or [91]. A description of the R/S statistic can also be found in these references or in [A2]. The following presentation of the variance-time plot is partly based on [61] and [91].

For an asymptotically second order self-similar process the variance of the averaged process $X^{(m)}$ decreases more slowly than the reciprocal of m, that is, $Var\{X^{(m)}\} \to a_1 m^{-\beta}$ as $m \to \infty$ where $0 < \beta < 1$ and a_1 is a finite positive constant. As mentioned earlier $\beta = 1$ ($H = 0.5$) in traditional teletraffic models. To check for self-similarity $log(Var\{X^{(m)}\})$ is plotted against time ($log(m)$). Ignoring small values of m, a line is fitted through the points and an estimate $(-\hat{\beta})$ of the slope $(-\beta)$ is obtained. If the points reasonably well resemble a straight line with $0 < \beta < 1$ it can be stated that the observed process at least has self-similar behaviour over the observed time-scales. The self-similarity

parameter H is estimated as $\hat{H} = 1 - \hat{\beta}/2$.

When considering the point process of arriving packets in a time-interval, X_t is an integer. In this case it is mathematically equivalent to consider the index of dispersion of counts instead of the variance-time curve when checking for self-similarity and estimating β. Consider a point process and let N_t denote the number of events in the process in the interval $(0, t]$. The index of dispersion of counts for N_t is then defined as IDC(t) = $Var\{N_t\}/E\{N_t\}$. Considering N_t at discretely spaced values of t, i.e. $t = m\Delta t$, where Δt is the time-interval corresponding to one observation X_t and m = 1,2,3..., is equivalent to considering the process $mX^{(m)}$. That is

$$\begin{aligned} IDC(m\Delta t) &= Var\{N_{m\Delta t}\}/E\{N_{m\Delta t}\} \\ &= Var\{mX^{(m)}\}/E\{mX^{(m)}\} \\ &= mVar\{X^{(m)}\}/E\{X^{(m)}\} \end{aligned}$$

In a stationary point process $E\{X^{(m)}\}$ = const. for all m = 1,2,3,... Hence checking the IDC curve of $N_{m\Delta t} = mX^{(m)}$ for self-similarity is mathematically equivalent to checking the variance-time plot of X^m for self-similarity. It is clear from the IDC expression that a self-similar process should have the slope $(1 - \beta)$, $0 < \beta < 1$ when plotting $log(IDC(m\Delta t))$ against $log(m)$.

In our search for MAP models with self-similar behaviour [A2], [A3], [A4] and [47] we have primarily used visual self-similarity as a yardstick to determine whether a given MAP seemed reasonable as a model for self-similar behaviour over several time scales rather than using some sort of stringent control of the Hurst parameter over several time scales using for example the second order methods mentioned above. However, plots of the variance time curve or equivalently the IDC, (index of dispersion of counts) in [A2], [A3], [A4] and [47] clearly show that over the observed time-scales an estimated value of H, $\hat{H}$, satisfies $0.5 < \hat{H} \leq 1$ so that the correlation structure over the observed time scales does not degenerate to a Hurst estimate of $\hat{H}$= 0.5.

MAP models will eventually turn out to be short range dependent in the above described sense. That is, once a given MAP is observed at a time scale larger than the largest time constant in the MAP the correlation structure will yield a Hurst estimate of $\hat{H}$ = 0.5. The objective of our studies in [A2], [A3] and [A4] has been to investigate to what degree it is possible to model self-similar behaviour over several time-scales with some relatively simple special cases of the MAP.

4.2 MAP-model and related issues

4.2.1 General MAP formulas

In this section some characteristics of the Markovian Arrival Process (MAP) used in this summary will be sketched. Further results can be found in [64], [77] and [A2].

For results and derivations regarding the Matrix Analytic approach used when calculating results for the MAP/G/1 queue see [64], [77],[83] or [A3].

The MAP is the point process generated by transition epochs in the Markov renewal process whose transition probability matrix $\mathbf{F}(*)$ is of the form

$$\begin{aligned} \mathbf{F}(x) &= \int_0^x e^{\mathbf{C}u} du\mathbf{D} \\ &= (\mathbf{I} - e^{\mathbf{C}x})(-\mathbf{C}^{-1})\mathbf{D} \end{aligned} \tag{4.1}$$

where the square matrices of order m, $\mathbf{C} = [C_{ij}]$ and $\mathbf{D} = [D_{ij}]$ are respectively a stable matrix and a non-negative matrix whose sum is an irreducible infinitesimal generator. (The matrix exponential is defined by its series i.e. $e^{\mathbf{M}} = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}$, for the properties of the matrix exponential see [16].)

The infinitesimal generator $\mathbf{Q} = \mathbf{C} + \mathbf{D}$ describes the transitions among the transient states $1, ..., m$ in the underlying Markov process. The stationary vector $\vec{\pi}$ of this Markov process satisfies the equations

$$\vec{\pi}(\mathbf{C} + \mathbf{D}) = \vec{\pi}\mathbf{Q} = \vec{0}, \quad \vec{\pi}\vec{e} = 1 \tag{4.2}$$

where $\vec{e}$ denotes a column vector of 1s.

Some well-known processes are (very) special cases of the MAP

- The Poisson process : $\mathbf{C} = (-\lambda)$ and $\mathbf{D} = (\lambda)$ (Poisson process with rate λ)
- The phase-type (PH) renewal process [75] is also a special case of the MAP. It includes both the Erlang and hyper-exponential (equivalent with the Interrupted Poisson Process IPP [59]) and Poisson arrival processes as special cases.
- The Markov-modulated Poisson process (MMPP) is also a special case of the MAP. The MMPP is the doubly stochastic Poisson process whose arrival rate depends on an m-state irreducible Markov process i.e. $\mathbf{C} = \mathbf{Q} - \Lambda, \mathbf{D} = \Lambda$ where the modulating Markov process has infinitesimal generator $\mathbf{Q}$ and $\Lambda = diag(\lambda_1, ..., \lambda_m)$.

A useful property of the class of MAPs is that it is closed under superposition.

For a MAP model the marginal distribution between arrivals is seen to be of phase-type. The probability distribution function for the marginal time t between arrivals is given as

$$P\{T \leq t\} = F(t) = 1 - (\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}e^{\mathbf{C}t}\vec{e} \tag{4.3}$$

The moments of the marginal distribution are given as

$$\mu_i = i!(\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}(\mathbf{C})^{-i}\vec{e} \tag{4.4}$$

4.3 Superpositions of IPPs

Since this summary only considers models that are found by superposition of Interrupted Poisson Processes (IPPs) and perhaps a Poisson source the IPP can be viewed as a model unit for the models examined here. A considerable advantage when examining these rather simple models is that $E\{N_t\}$, $Var\{N_t\}$ and $IDC(t)$ all are scalar functions see for example [A2].

An IPP can in MAP notation be described as follows.

$$\mathbf{C} = \begin{bmatrix} -c_1 - \lambda & c_1 \\ c_2 & -c_2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \tag{4.5}$$

The steady state probability vector is easily seen to be $\vec{\pi} = (\frac{c_2}{c_1+c_2}, \frac{c_1}{c_1+c_2})$.

The mean of the counting process N_t of the IPP is given as

$$E\{N_t\} = \frac{\lambda c_2}{c_1 + c_2} t \tag{4.6}$$

From formulas in [41] for example the scalar expression for $Var\{N_t\}$ can be found as

$$Var\{N_t\} = (\frac{\lambda c_2}{c_1+c_2} + \frac{2\lambda^2 c_1 c_2}{(c_1+c_2)^3})t - \frac{2\lambda^2 c_1 c_2}{(c_1+c_2)^4}(1 - e^{-(c_1+c_2)t}) \tag{4.7}$$

This expression is also very useful when calculating the variance function of processes obtained by superposing several IPPs since the resulting variance can be obtained by summing the variances of the individual processes (The IPP sources are independent). The IDC curve for a superposition of n IPP sources is thus given by

$$IDC(t) = (\sum_{j=1}^{n} Var\{N_t^j\})/(\sum_{j=1}^{n} E\{N_t^j\}) \tag{4.8}$$

In section 4.1 regarding measures of self-similarity it is shown that an exact second order self-similar process has the exact slope $(1 - \beta) = (2H - 1)$, $0 < \beta < 1$ and $0.5 < H < 1.0$ when plotting $log(IDC(t))$ against $log(t)$. Using the substitution $u = log(t)$ the derivative of the $log - log$ plotted $IDC(t)$ curve for a superposition of n IPPs can be calculated as

$$\begin{aligned} \frac{d(log(IDC(t)))}{d(log(t))} &= \frac{d(log(\sum_{j=1}^{n} Var\{N_{e^u}^j\}))}{du} - \frac{d(log(\sum_{j=1}^{n} E\{N_{e^u}^j\}))}{du} \\ &= (\frac{1}{\sum_{j=1}^{n} Var\{N_{e^u}^j\}} \frac{d(\sum_{j=1}^{n} Var\{N_{e^u}^j\})}{du}) - 1 \\ &= (\frac{1}{\sum_{j=1}^{n} Var\{N_{e^u}^j\}} (\sum_{j=1}^{n} \frac{d(Var\{N_{e^u}^j\})}{du})) - 1 \end{aligned} \tag{4.9}$$

The derivative of the log - log plotted IDC curve in the general MAP case is given in [A2].

When superposing homogeneous IPPs the IDC remains constant but the mean arrival rate of course grows proportionately. Examining the expression for the IDC in the special case of the IPP reveals that it is possible to fix both the mean arrival rate and the IDC for different numbers n of superposed homogeneous IPPs. This can be done by fixing the following three constants k_1, k_2 and k_3. (The mean arrival-rate is $\frac{k_3}{k_1}$)

$$\begin{aligned} k_1 &= c_1 + c_2 \\ k_2 &= 2\lambda c_1 \\ k_3 &= n\lambda c_2 \end{aligned} \tag{4.10}$$

Hence it is possible to create superpositions with the exact same first and second order properties of the counting process. In terms of the parameters k_1, k_2 and k_3 and the number of superposed sources n the individual IPP is defined as

$$\begin{aligned} c_1 &= \frac{n k_1 k_2}{n k_2 + 2k_3} \\ c_2 &= \frac{2k_1 k_3}{n k_2 + 2k_3} \\ \lambda &= \frac{n k_2 + 2k_3}{2n k_1} \end{aligned} \tag{4.11}$$

As previously mentioned it is well known that the IPP is stochastically equivalent with the hyper-exponential distribution with two phases [59]. As shown in [A2] this implies that the marginal distribution of a superposition of IPPs (and perhaps a Poisson source) also has a hyper-exponential marginal distribution. In the general case of superposing n inhomogeneous IPPs (and perhaps a Poisson source) the marginal distribution is hyper-exponential with 2^n phases. Superposing n homogeneous IPPs (and perhaps a Poisson source) yields a hyper-exponential marginal distribution with $n+1$ phases.

4.4 Time dynamic behaviour

To illustrate the models examined in [A2],[A3] and [A4] the dynamic behaviour and correlation structure of different special cases of a MAP are examined. As mentioned in the previous section only quite simple models are examined here i.e. superpositions of IPPs (and perhaps a Poisson source) in fact the same models as in [A4]. More general models are considered in [A2] and in [47]. Both the analytical MAP tools outlined above and simulation is used. Simulating MAP processes makes it possible to get a first look at how $X^{(m)}$ behaves at different aggregation levels m. Here it is interesting to examine a realization at different aggregation levels m, differing for example by an order of magnitude, to see how the process behaves over several time-scales. Whenever we in the following visually inspect a realization we look at aggregation levels differing by an order of magnitude. Starting with the largest aggregation level, we successively examine the next lower aggregation level by zooming in on a randomly chosen part of the process at the previous aggregation level. This is similar to the method used in [61] and [62].

4.5 Time dynamic behaviour of IPPS

A very simple MAP yielding correlations in the process X_t is the IPP (4.5). The sum of the transition rates in the modulating process $(c_1 + c_2)$ is called the time constant for the IPP.

We examine the dynamic behaviour of three IPPs with the same λ but with different time constants spanning 4 orders of magnitude. They can be defined

as $\lambda = 4.0$, $c_1 = s\frac{1}{4}$ and $c_2 = s\frac{3}{4}$, where s is either 10^{-2}, 10^{-4} or 10^{-6}. These three IPPs will be denoted respectively IPP$_a$,IPP$_b$ and IPP$_c$.

In figure 4.1 a realization of IPP$_b$ is shown. It is obvious that the on/off type behaviour does not resemble self-similar behaviour. The on/off type behaviour is only visible when viewing the process at aggregation levels corresponding to the number of arrivals in a time interval that is small compared to the reciprocal of the time constant.

Next we consider the index of dispersion of counts (IDC) curve for each of the three processes IPP$_a$, IPP$_b$ and IPP$_c$ in figure 4.2. It is evident in the log-log plot that the index of dispersion of counts for all 3 processes increases approximately linearly over several time-scales, the number of time-scales determined by the IPPs time constant. It is also evident that the slope of the linear growth in the log-log plot corresponds to $\beta \approx 0$ giving a Hurst parameter of $H \approx 1$ in the interval of linear growth. This can more formally be seen from figure 4.3 where the derivatives of three IDC curves in a $log - log$ plot is displayed. This behaviour is not surprising when examining the formulas (4.6) and (4.7). For the IPP it is clear that for $t(c_1+c_2) << 1$ the IDC curve can be represented by its Taylor expansion as

$$IDC(t) = \quad 1 + \frac{\lambda c_1}{(c_1+c_2)} t + o(t), \quad t(c_1 + c_2) << 1 \tag{4.12}$$

It is well known that for a renewal process $IDC(t) \rightarrow (C_{var})^2$, as $t \rightarrow \infty$ ($C_{var} = \frac{\sigma}{\mu_1}$ is the coefficient of variation of the renewal process, μ_1 and σ are respectively the mean and standard deviation of the inter-arrival time distribution). Since the IPP is a renewal process "the price paid" for dependence over many time-scales is therefore a highly dispersed marginal distribution. The three IPPs above have the following coefficients of variation, IPP$_a$: $C_{var} = 14.2$, IPP$_b$: $C_{var} = 141.4$ and IPP$_c$: $C_{var} = 1414.2$.

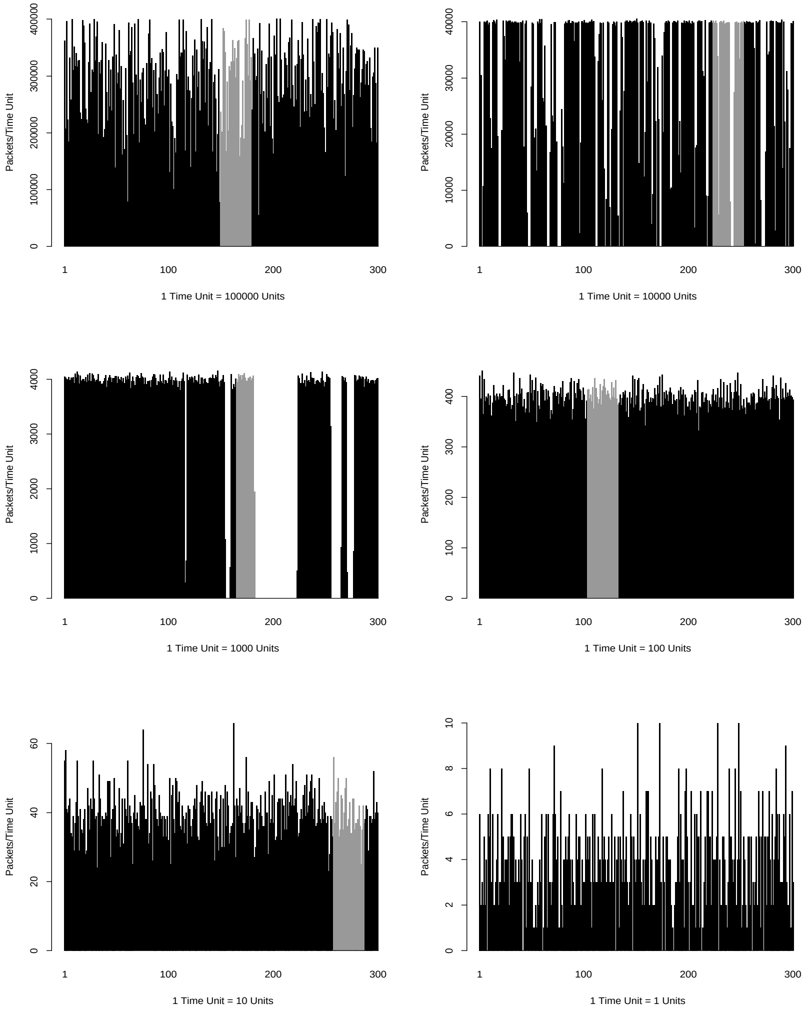

Figure 4.1: **Time dynamic behaviour of an IPP (IPP$_b$) over 6 timescales**

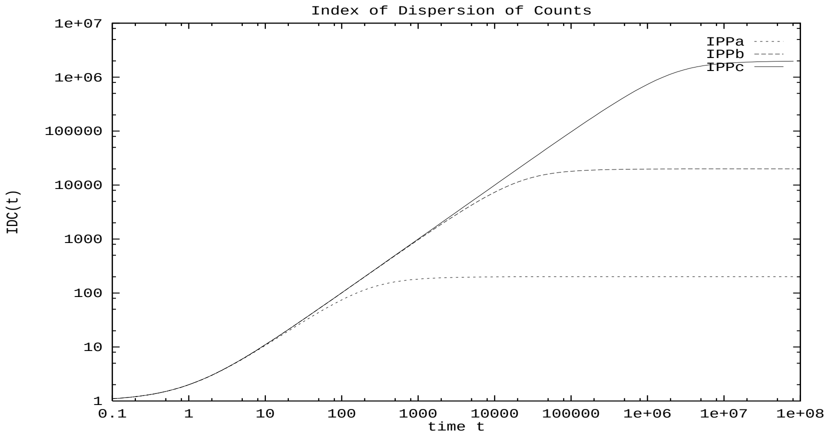

Figure 4.2: **The IDC curves of IPP$_a$, IPP$_b$ and IPP$_c$**

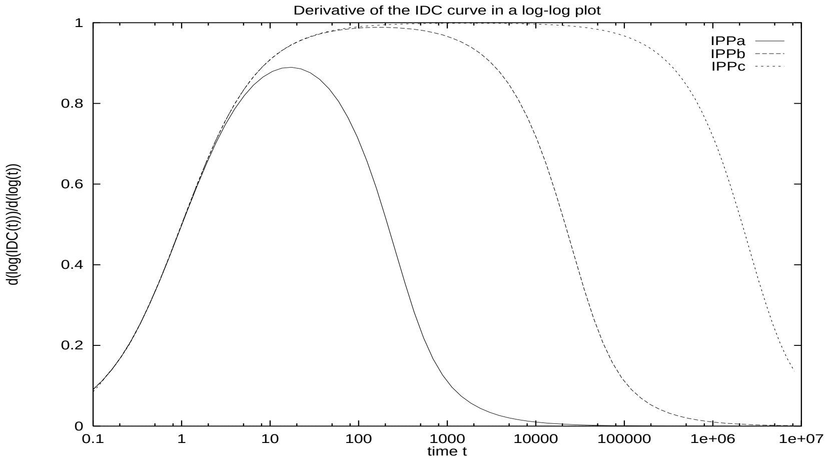

Figure 4.3: **Derivatives of the log-log plotted IDC curves for IPP$_a$, IPP$_b$ and IPP$_c$**

4.6 Time dynamic behaviour of superposed homogeneous IPPs

Next we consider the superposition of several identical IPPs. As described in section 4.3 it is possible to choose a process of superpositions of IPPs to have the same arrival rate and $IDC(t)$ as a single IPP. In this summary we will restrict ourselves to considering a superposition of 9 identical IPPs defined such that the resulting process has the same rate and $IDC(t)$ as IPP$_b$. From (4.11) it can be derived that each of the 9 IPPs to be superposed are defined by $c_1 = 7.5 * 10^{-5}$, $c_2 = 2.5 * 10^{-5}$ and $\lambda = \frac{4}{3}$. Figure 4.4 shows a realization of the superposition of 9 IPPs with the same rate and $IDC(t)$ as IPP$_b$.

While modelling self-similar behaviour significantly better than the single IPP it is still evident from the plots of the different aggregation levels that the plots do not "look the same" over several timescales (at least not for this relatively modest choice of n). An inherent limitation in the use of superposed homogeneous IPPs in the context of modelling self-similarity is that they with regard to $IDC(t)$ are no more general than a single IPP. However, the marginal distribution of the superposed process compared to the single IPP with the same rate and IDC is not as dispersed. For the process of 9 superposed IPPs the coefficient of variation is C_{var} = 44.8. (recall that for IPP$_b$ C_{var} = 141.4).

4.7 Time dynamic behaviour of superposed inhomogenous IPPs

In this summary we only consider an 8 state model, a superposition of three IPPs and a Poisson stream. This particular model has been used in [A4] and [A2]. In [A2] it is the one called ADD$_8$. It is defined by (4.13) and its coefficient of variation is C_{var} = 1.36. The marginal distribution for this model is an 8 phased hyper-exponential distribution and therefore has the same shape as the corresponding exponential distribution for a Poisson arrival process, however, it is obviously much heavier tailed. In figure 4.5 the dynamic behaviour of the superposition (4.13) of the 3 IPPs and a Poisson stream is illustrated.

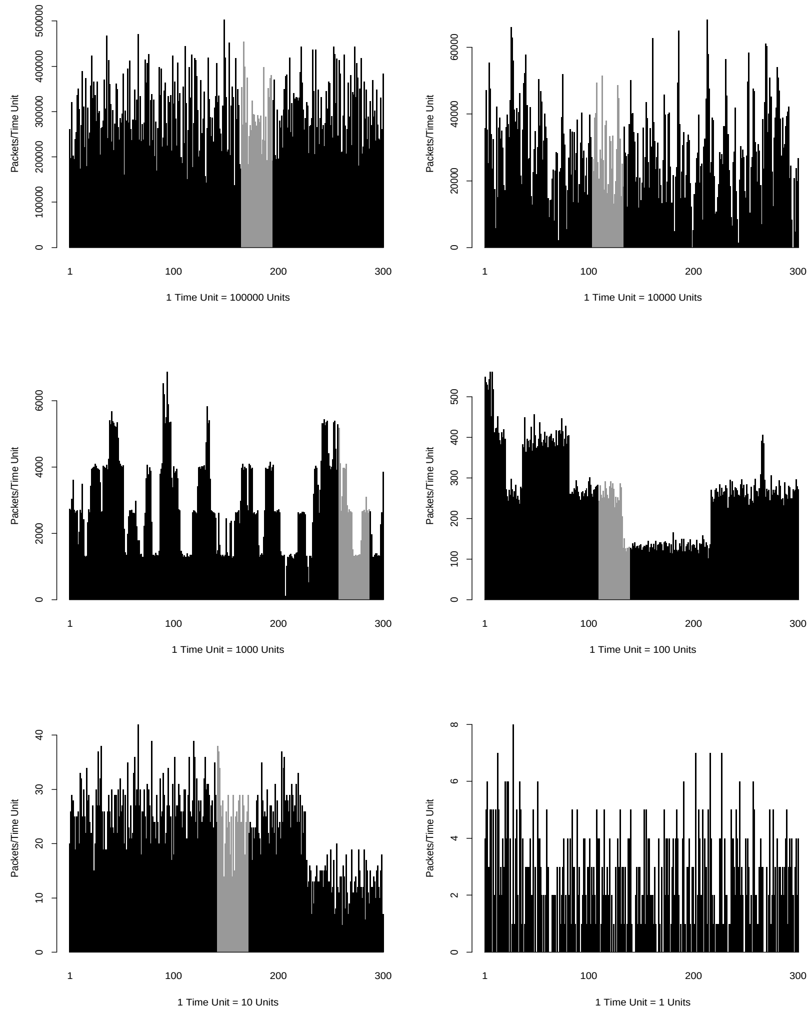

Figure 4.4: **Time dynamic behaviour of a superposition of 9 homogenous IPPs with the same rate and IDC(t) as IPP$_b$ over 6 timescales**

$$\begin{array}{llll}
\lambda_{poisson} & = 0.7 & & \\
\lambda & = 2.0 & , c_1 = 9 * 10^{-3} & \text{and } c_2 = 9 * 10^{-3} \\
\lambda & = 1.5 & , c_1 = 2 * 10^{-4} & \text{and } c_2 = 2 * 10^{-4} \\
\lambda & = 1.1 & , c_1 = 5 * 10^{-5} & \text{and } c_2 = 5 * 10^{-5}
\end{array} \tag{4.13}$$

For this model there is evidently high variability on all the time-scales displayed. Except for the highest and lowest aggregation level it apriori seems difficult to decide which aggregation level the individual plots belong to. It can be asserted that visually they appear to be at least as self-similar as the plots shown in [61]. This can also be asserted about two other models presented in [A2].

To get a more formal measure of the dynamic behaviour of this superposition the IDC curve is calculated in figure 4.6. It clearly indicates that there are correlations spanning several time-scales.

The derivative ($1 - \beta$ for a self-similar process) of the log - log plotted IDC curve (can be found for example from (4.9)) is displayed in figure 4.7.

The derivative definitely lie in a band between 0 and 1 over several time-scales. This indicates a “Hurst parameter” somewhere between $\frac{1}{2}$ and 1 over several time-scales. Hence this superposition of inhomogeneous IPPs and a Poisson stream is, not surprisingly, more versatile in modelling the IDC of a self-similar process than the previous homogeneuos superpositions of IPPs.

4.8 Queueing experiments with MAPs

In the following we will summarize our findings regarding the mean queueing behaviour presented in [A3] and [A4]. As examples the three previously presented MAPs will be analyzed as a function of the load. The same analysis has been presented in [A4]. In figure 4.8 we examine the mean queueing behaviour of the superposition of the 9 IPPs, IPP_b which has the same rate and IDC as the superposition, and the renewal process defined by the marginal distribution of the superposition.

It is clear from figure 4.8 that the process defined by the superposition by far has the most adverse queueing behaviour with a knee around a load of 0.35 i.e.

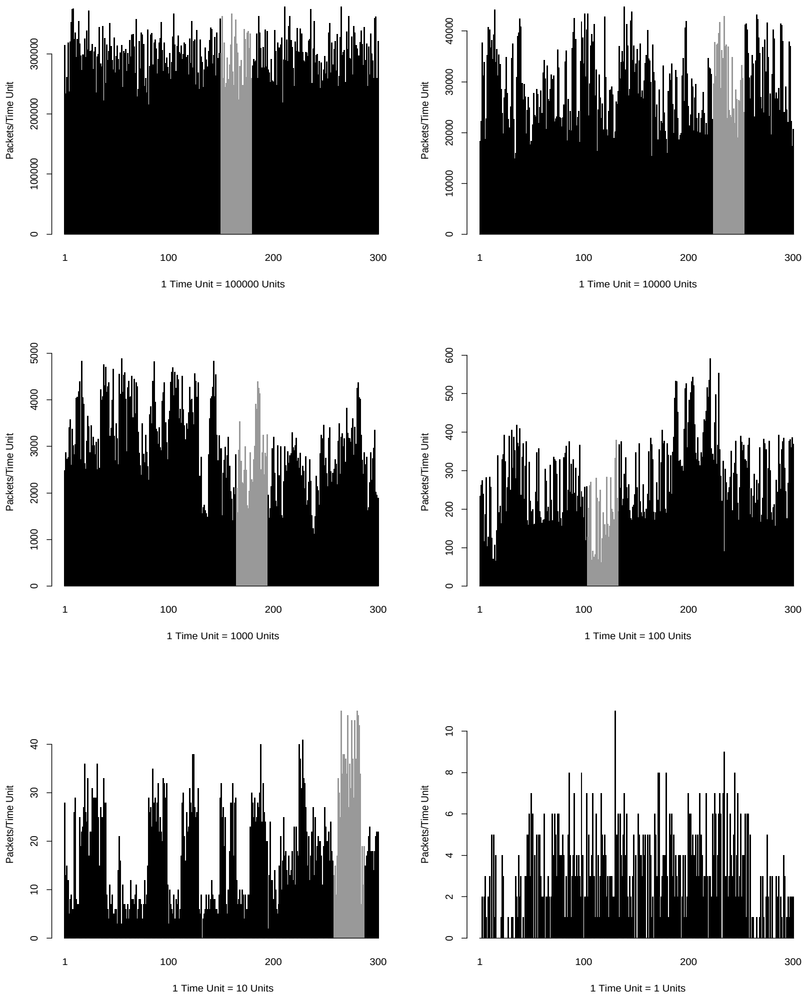

Figure 4.5: **Time dynamic behaviour of the superposition of 3 inhomogenous IPPs and a Poissson stream over 6 time scales**

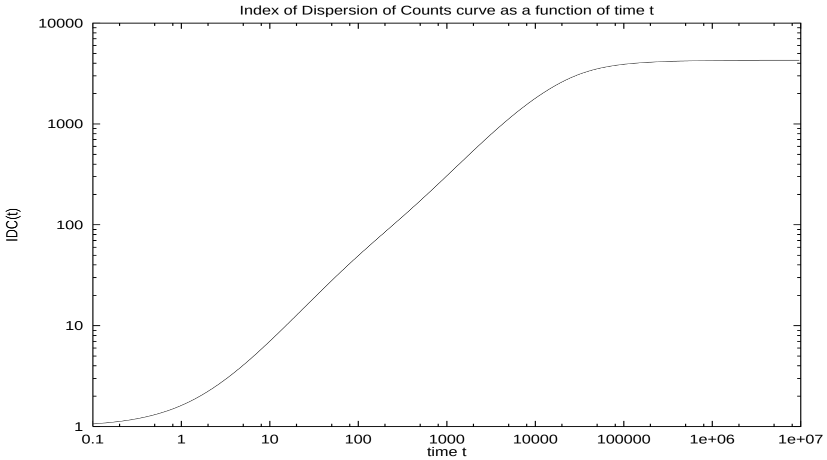

Figure 4.6: **The IDC curve of the superposition of 3 inhomogeneous IPPs and a Poisson stream**

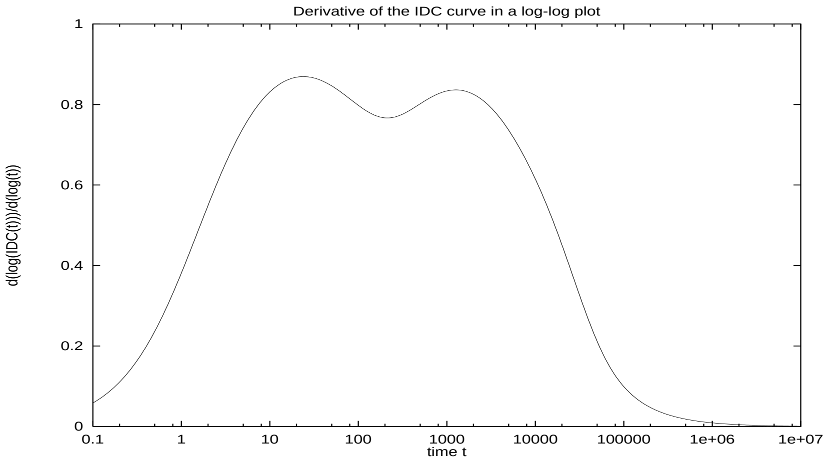

Figure 4.7: **The derivative of the IDC curve of the superposition of 3 inhomogeneous IPPs and a Poisson stream**

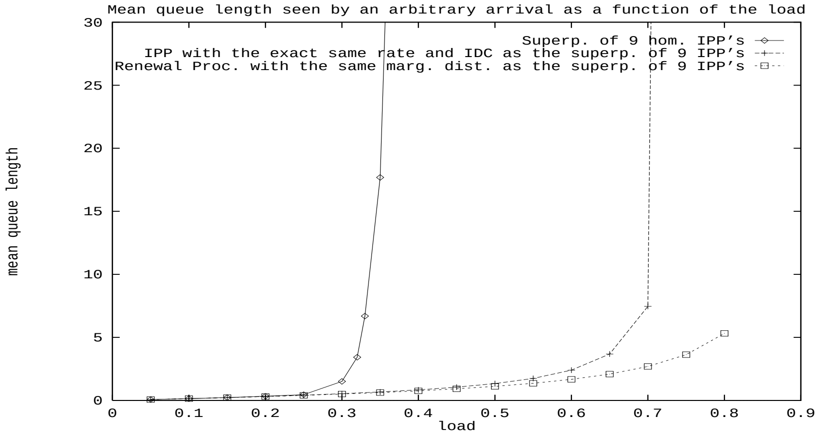

Figure 4.8: **Comparison of the mean queue length seen at arrival epochs a in process defined by a superposition of 9 IPPs , the renewal process defined by its marginal distribution and a single IPP with the same rate and IDC as the superposition**

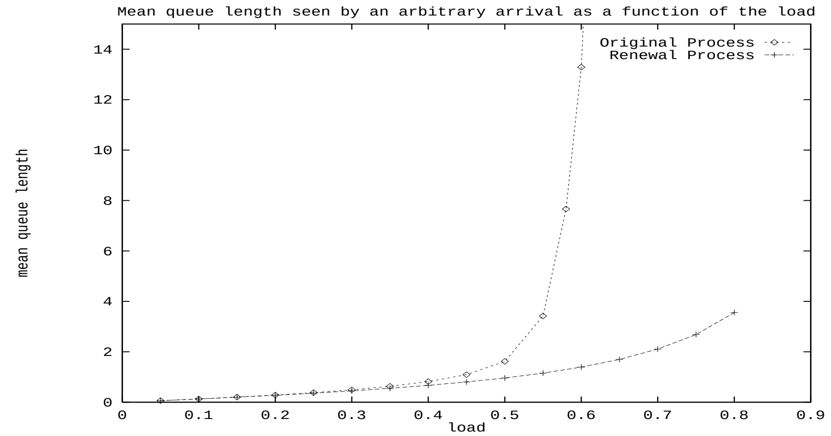

Figure 4.9: **Comparison of the mean queue length seen at arrival epochs in a process defined as a superposition of 3 inhomogenous IPP's and a Poisson stream and the renewal process defined by its marginal distribution**

substantially lower than 1.0 which is the same phenomenon as observed with the LAN traces in [32]. It also seems evident that it is not the marginal properties of the superposition that is responsible for the adverse queueing behaviour. It is very interesting to note that two processes with same rate and $IDC(t)$ can result in such different queueing behaviour as shown in figure 4.8. It might seem somewhat surprising that the IPP has "nicer" queueing behaviour than the superposed process.

In figure 4.9 we have examined the mean queue length behaviour of the superposition of 3 inhomogeneous IPP's and a Poisson source and the renewal process defined by its marginal distribution. The mean queue length shows the same kind of adverse behaviour as described in [32] with a knee around a load of 0.55 - 0.60. This model corresponds both "visually" and in marginal distribution to the LAN data observed in for example [32] or [61]. Also the location of the knee for this curve corresponds very well with those found in [32]. It is striking that the renewal process defined by the marginal distributions shows queueing behaviour that seems harmless and close to M/D/1.

All the models examined in [A2],[A3] and [A4] belong to the class of GMLP models, defined in [A2], which is a special case of the MMPP. In order for the MMPP/G/1 queue to exhibit adverse behaviour it is neccesary for the MMPP to overload the system in at least one of its states and maintain this overload situation long enough for a significant queue to build up. Clearly these observations have a very close relationship with the caudal characteristic curve examined by Neuts in [76] and generalizing recent results in [1] and particularly in [26].

4.9 Conclusions and Future work

Finally some concluding remarks will be given on the results presented in this chapter and more generally on the findings in [A2],[A3] and [A4].

- It is possible to model packet traffic with very variable behaviour with a MAP. Relatively small order MAPs (8-16 states) seem sufficient to model very variable or even self-similar behaviour over several timescales.
- We have applied a rather heuristic approach in finding "self - similar" MAPs i.e. examing realisations at different aggregation levels. In order for "self-similar" MAPs to be of any practical use clearly a more systematic procedure is needed. In [A2] we suggest a curve fitting method where a MAP model is fitted to the desired derivative of the $log - log$ plotted IDC curve i.e. a constrained minimization procedure. Whether this is a good idea or not is for further study. Analytical simplifications are here possible when only considering superpositions of IPPs since the expressions involved in the minimization routine then become scalar.
- The advantage of our model compared to for example Fractional Brownian motion [81] is clearly analytical tractability, but unfortunately the parameters in our models do not have the same immediate physical interpretation as for example those in a Fractional Brownian Motion model. So a problem with the MAP is that altering one parameter usually affects several physical descriptors which makes it more difficult to choose the parameters in the MAP.

- It is dangerous to rely to much on first and second order properties of the arrival process since these might not supply enough information to assess queueing behaviour. Here it might be appropriate to examine how well the observed data are actually described by a second self-similar process since the fractional Brownian motion model (second order self-similar) for example is fully described by its first and second order properties while MAPs clearly are not.
- It is indeed very risky to dimension after some mean load requirements as also described in [32]. Knowledge about the dynamic behaviour of the arrival process is highly desirable before fixing the maximum allowed load, i.e. the maximum load is very dependent on the characteristics of traffic.

Chapter 5

Modelling of the asymptotic queueing behaviour

In this chapter the work in [A5] and [A6] regarding evaluation of a heuristic formula for the tail behaviour of $\sum Gi/Gi/1$ will be summarized. As mentioned in the introduction knowledge about the tail behaviour has applications in the connection admisssion control (CAC) procedure in the future B-ISDN. Here we will briefly describe the large deviation result leading to the asymptotic decay rate and then introduce the heuristic formula and background and finally present some results.

It is known that a large class of queueing systems possesses an exponential tail behaviour. Assuming a large buffer the probability that the queue exceeds a given threshold can be estimated by a simple exponential formula

$$P(Q > r) \approx \alpha e^{-s_\infty r} \tag{5.1}$$

where s_∞ is the exponential asymptotic decay rate at very large occupancy levels (r big) and α is the associated asymptotic constant. Assuming that the asymptotic constant α is 1 leads to a quite simple CAC procedure when considering a total arrival process that is based on superpositions of independent sources.

This has been described in detail in [90]. This approximation has been justified by the fact that QoS requirements in ATM are very strict, around 10^{-9} in cell loss probability, and in this case the relative error of approximating α with 1 is not big. However, as pointed out in [25] the estimate based on the asymptotic slope can be very misleading due to the fact that the asymptotic constant α can be decades smaller than 1 in the case of multiplexing many bursty sources, and it can be decades greater than 1 when multiplexing many smooth sources.

5.1 The Asymptotic Decay Rate and the Asymptotic Constant

The reason for the interest in the asymptotic behaviour is the desire to be able to squeeze as much traffic through an ATM link as possible while maintaining the QoS requirements i.e. the underlying statistical multiplexing problem. We denote the total amount of work arriving in the time interval $]0, t]$ by $A(0, t)$, and assume that it takes one time unit to serve one unit of work, that is the capacity of the server is always 1. The unfinished work at an arbitrary time (also denoted the virtual waiting time is denoted V and from section 5 in [86] we know that $V = \sup_t(A(-t, 0) - t)$. The actual waiting time) is the unfinished work at an arbitrary arrival, and from [1] it is known that the asymptotic decay rate of the virtual waiting time and the waiting time is the same for the general case with Markovian Arrival Processes. Here and in [A5] and [A6] we have mainly considered the virtual waiting time.

5.1.1 The G/D/1 case

The theory of large deviations can be used to prove that for a large class of arrival processes the work load distribution possesses an exponential tail behaviour, see e.g. [28].

Letting $A(0, t)$ be the amount of work arriving to the system in time $(0, t)$, we define the asymptotic decay rate function as

$$\lambda(\theta) = lim_{t \to \infty}[\frac{1}{t} log E(e^{\theta A(0,t)})] \tag{5.2}$$

Let V denote the unfinished work (work load) at an arbitrary time. For "general" arrival processes without long range dependence the following (large deviation) result, corollary 2.3 with $a = 1$ and $v = 1$ [28] remains valid:

$$lim_{r\to\infty}\frac{1}{r}log(P\{V > r\}) = -s_\infty \tag{5.3}$$

where $s_\infty = sup\{\theta > 0|\lambda(\theta) \leq \theta\}$. and the time unit is chosen such that the service rate is 1.

5.1.2 The Markovian Arrival Process

From [74] it is known that the generating function of the number of arrivals in a window of length t can be found as

$$\mathbf{P}(z,t) = \vec{\pi}e^{(\mathbf{C}+z\mathbf{D})t}e \tag{5.4}$$

where the matrices **C** and **D** are matrices defined as described in the previous chapter.

As described in [90] the asymptotic decay rate $\lambda(\theta)$ for this arrival process is the Perron-Frobenius eigenvalue of $\mathbf{C} + e^\theta\mathbf{D}$.

5.1.3 Superposition and thinning of identical renewal processes

Assume that a renewal process with work process $A(0,t)$ gives rize to an asymptotic decay rate s_∞, that is $\lambda_A(s_\infty) = s_\infty$. If this process is thinned by a factor n and n independent processes with work process $A_i(0,t) = A(0,t/n)$ are superposed the asymptotic decay rate function of the superposition is

$$\begin{aligned}lim_{t\to\infty}\frac{1}{t}ln(E[e^{\theta\sum_{i=1}^{n}A_i(0,t)}]) &= lim_{t\to\infty}\frac{1}{nt/n}\sum_{i=1}^{n}ln(E[e^{\theta A(0,t/n)}]) \\ &= lim_{t\to\infty}\frac{1}{t}ln(E[e^{\theta A(0,t)})] = \lambda_A(\theta)\end{aligned} \tag{5.5}$$

from which it is seen that the asymptotic decay rate of the original (single) renewal stream is the same as the asymptotic decay rate of the thinned superposition.

5.1.4 Equivalent Bandwidth from the Asymptotic Slope

Adopting the approximation $P(V > r) \approx e^{-s_\infty r}$, we see that $P(V > r) \leq \epsilon$ (with ϵ being the QoS requirement) is equivalent to $s_\infty \geq \frac{ln(\epsilon)}{r} = \eta$ that is the asymptotic slope should stay above the critical value η. Consider an arrival process described by the asymptotic decay rate function $\lambda(\theta)$ and assume it is an independent sum of component arrival processes described by the rate functions $\lambda_i(\theta)$. Since the asymptotic decay rate function is additive for independent processes we get $\lambda(\theta) = \sum_i \lambda_i(\theta)$. Assume further that the capacity is 1. Define the effective bandwidth of stream i as $e_q(i) = \frac{\lambda_i(s(1))}{s(1)}$, where $s(1)$ solves $\lambda(s(1)) = s(1)$. Now $\sum_i e_q(i) < 1$ if and only if $\frac{\lambda(s(1))}{s(1)} = \sum_i \frac{\lambda_i(s(1))}{s(1)} < 1$ and if and only if $\lambda(s(1)) < s(1)$. This is equivalent to $s(1) > \eta$ due to the convexity of the asymptotic decay rate function $\lambda(\theta)$, see [90] for details on the convexity properties.

5.2 A Heuristic Formula for the tail behaviour of a single server queue

In this section we present a heuristic formula for the complementary distribution of the unfinished work in the nGI/D/1 queue. This formula has its origin in the diffusion analysis of the M/D/1 and nD/D/1 queues. From the heuristic formula it is straightforward to find the (heuristic) asymptotic constant.

In order to obtain a simple closed form approximation for the probability of the virtual waiting time exceeding a given buffer threshold in the nGI/D/1 queue, we look at the nD/D/1 and M/D/1 queues. The classical diffusion approximation of the M/D/1 queue, see e.g. section 6.1.2. in [86], leads to

$$P\{V_{M/D/1} > r\} \approx exp(-\frac{2(1-\rho)}{\rho}r) \tag{5.6}$$

In comparison a Brownian bridge approximation of the nD/D/1 queue, see e.g. section 6.2.2. in [86], leads to

$$P\{V_{nD/D/1} > r\} \approx exp(-\frac{2(1-\rho)}{\rho}r - \frac{2r^2}{n}) \tag{5.7}$$

In the nD/D/1 case we see the same term as in the M/D/1 case but in addition we see a term which on the logarithmic scale decreases with the square of r. By applying the correct exponential term $s_{M/D/1}$ (see [A5]) both approximations can be improved. The improved form states

$$P\{V_{M/D/1} > r\} \approx \rho exp(-s_{M/D/1} r) \text{ and}$$
$$P\{V_{nD/D/1} > r\} \approx \rho exp(-s_{M/D/1} r - \tfrac{2r^2}{n}) \quad (5.8)$$

The multiplication of ρ ensures that the probability of an empty system is correct. Table 1 gives an assessment of the quality of approximation (5.8) in the M/D/1 case with a load of 0.6, and in the 80D/D/1 case with a load of 0.8.

Table 1: Evaluation of the improved M/D/1 and nD/D/1 formulae

buffer level	M/D/1 $\rho = 0.6$			80D/D/1 $\rho = 0.8$		
r	exact	diff.	for. (5.8)	exact	diff.	for. (5.8)
10	5.6 e-5	1.6 e-6	4.6 e-5	1.2 e-3	5.5 e-4	8.8 e-4
20	4.3 e-9	2.6 e-12	3.5 e-9	1.7 e-8	2.1 e-9	6.6 e-9
30	3.3 e-13	4.3 e-18	2.7 e-13	1.7 e-15	5.2 e-17	3.3 e-16

As the table clearly indicates then formula (5.8) represent a significant improvement compared to the usual diffusion approximation.

For the superposition of n renewal streams in which the squared coefficient of variation is c_a^2 we suggest

$$P\{V_{nGI/D/1} > r\} \approx \rho exp(-s_{M/D/1} r - (1-(c_a^2)^\beta)\frac{2r^2}{n + K_a r}) \quad (5.9)$$

In comparison to the expression for the nD/D/1 queue the second term has been modified in two ways.

1. A term $K_a r$ is added to the usual denominator n.
   Thereby we enforce an exponential tail behaviour with rate $s_\infty = s_{M/D/1} + 2(1-(c_a^2)^\beta)/K_a$. The constants K_a and β are to be chosen such that both the correct asymptotic slope s_∞ and the correct mean queue length are obtained.
2. A multiplicative term $(1-(c_a^2)^\beta)$.
   This term disappears in the case with Poisson input making the expression correct in this case. The multiplicative term is 1 in the nD/D/1 case and with $K_a = 0$ our heuristic formula is seen to in accordance with formula (5.8).

It should be noticed that formula (5.9) implies that for $c_a^2 > 1$ the virtual waiting time curve for the nGI/D/1 queue will be above the corresponding M/D/1 curve and for $c_a^2 < 1$ it will be below.

5.3 Obtaining the Asymptotic Constant

With a tail behaviour that is asymptotically exponential with rate s_∞, the asymptotic constant is defined as the limit $\alpha = lim_{r\to\infty} exp(s_\infty r) P(Q > r)$. From formula (5.9) we get

$$log(\alpha_n) = lim_{r\to\infty}[s_\infty r + log(\rho) - s_{M/D/1} r - 2(1-(c_a^2)^\beta)\tfrac{r^2}{n+K_a r}] \quad (5.10)$$

After some calculations (see [A5]) we end up with

$$\alpha_n = \rho exp[\frac{2(1-(c_a^2)^\beta)}{K_a^2} n] = \rho exp[\frac{(s_\infty - s_{M/D/1})^2}{2(1-(c_a^2)^\beta)} n] \quad (5.11)$$

since $s_\infty = s_{M/D/1} + \frac{2(1-(c_a^2)^\beta)}{K_a}$. The prediction that the intercept has an exponential dependence on the number of sources is in nice accordance with the numerical results leading to the conjecture of formula (1.6) in [25].

5.4 Numerical experiments

The arrival processes we have considered in [A5] and [A6] are all superpositions of simple renewal processes with phase-type [75] distributed inter-arrival times i.e. IPPs and Erlang-2s. Since these are all (very) special cases of the MAP and the MAP is closed under superposition it is possible to calculate both the exact tail behaviour and the exact asymptotic tail behaviour using MAP tools to evaluate the heurestic formula.

5.4.1 Comments regarding the applied matrix analytic tools

This authors main contribution to the work in [A5] and [A6] has been in the area of providing analytical tools for accessing the goodness of the heuristic formula, that is matrix analytic tools to calculate the exact tail probabilities in a $MAP/D/1$ queue and also tools for analyzing the exact asymptotic behaviour of the $MAP/D/1$ queue i.e. the calculation of the asymptotic decay rate and the asymptotic constant.

Since details about these implementations are omitted in [A5] and [A6] this seems like an appropriate place to give some comments that are not directly obtainable from the references upon which the implementations are based. The implementation of the $MAP/D/1$ queue is a straightforward implementation of the algorithms outlined in [64] applying the cancellation free scheme proposed in [85]. The algorithm is also quite thoroughly described in [A3]. The asymptotic constant and decay rate for the $BMAP/G/1$ queue are given by theorem 6 p. 119 (formula 17) respectively theorem 8 p. 121 (formula 19) in [1]. The following comments about details in the calculation of the asymptotic constant and decay rate are rather technical so familiarity with [1] will definitely ease the understanding.

Theorem 6 is a BMAP formulation of the large deviation result regarding the asymptotic decay rate. Numerically we have applied Newton - Raphson iterations to solve the two non-linear equations with two unknowns arising from theorem 6 to find the asymptotic decay rate. In our case with the MAP/D/1 queue and MAPs defined as superpositions of simple IPPs and Erlang-2 distributions simplifications in the general formulas of theorem 6 are possible. One of the non-linear equations involves the generating function of the service time distribution. Clearly this function is particularly simple for deterministic service times. The other non-linear equation involves the calculation of $\lambda(\theta)$ which in case of the MAP corresponds to a certain Perron-Frobenius eigenvalue of the tranformed MAP i.e. of $\mathbf{C} + e^{\theta}\mathbf{D}$ as previously outlined. However, since the

superposition is done by the thinning procedure described it suffices to calculate the Perron-Frobenius eigenvalue for the corresponding renewal process which in our examples corresponds to calculating the eigenvalue for a matrix of order two. Clearly this simplifies the numerical implementation considerably. Theorem 8 gives a formula for calculating the asymptotic constant. It involves the calculation of the same matrix $\mathbf{G}$ necessary when solving the BMAP/G/1 queue [64] or [A3]. It also involves finding the left and right eigenvectors of $\mathbf{R}(e^{\theta})$ = $\mathbf{C} + e^{\theta}\mathbf{D}$ corresponding to the Perron-Frobenius eigenvalue. The derivative of this eigenvalue is also needed. The left and right eigenvectors were at first obtained by a brute force method i.e. by applying standard software (EISPACK) to calculate all eigenvectors and then isolating the left and right vector corresponding to the Perron-Frobenius eigenvalue. Clearly as the number of IPPs (or Erlang-2's) superposed increases this method becomes unsuitable and numerical errors start to dominate. However, in the case of superpositions of IPPs this can be avoided since the left and right eigenvectors are given in closed form expressions in [45] where methods devoloped in [10] are used. The derivative of the Perron-Frobenius eigenvalue can be obtained by the same method as applied in the proof of Lemma 1.3.3. p. 17 in [75]. Let $\mathbf{R}(z) = \mathbf{C} + z\mathbf{D}$ and let $\vec{v}(z)$ and $\vec{u}(z)$ denote respectively the right and left eigenvectors of $\mathbf{R}(z)$ corresponding to the Perron-Frobenius eigenvalue $\chi(z)$. The derivative of this eigenvalue is then given by $\chi'(z) = \vec{u}(z)\mathbf{D}\vec{v}(z)$ for $0 < z \leq 1$. The details are described in [75].

5.4.2 Renewal streams examined

The following three interarrival time distributions have been used.

- Erlang-2, which can be seen as a traffic model of a CBR stream which has been jittered as a result of passing through a number of nodes in a peak rate allocated ATM network, see [89],
- A "non bursty"IPP with small mean burst size of 3.375 cells and an activity ratio of 1/3, supposed to represent an on/off source with rapid change between idle and active state.
- A bursty IPP with mean burst size of 60 and an activity ratio of 1/11, also used in [25].

5.4.3 Numerical results

In the section we will describe some of the main results from [A5] and [A6] (for a total list of results see these articles).

The heuristic formula has been evaluated for the three different source types. In each case the resulting queue was examined for the number of superposed sources being respectively 12, 24 and 48.

- For the superposition of Erlang-2 sources it is seen from figures 5.1, 5.2 and 5.3 that the heuristic formula for the tail probability gives reasonably accurate results, overestimating the tail probability somewhat. Unfortunately this overestimation increases with the number of superposed sources. Also the heuristic formula catches the asymptote for the tail probability much later than the exact asymptote does.
- For the “non-bursty” IPP case the results seen in figures 5.4, 5.5 and 5.6 indicate that in this case the heuristic formula seems quite accurate.
- For the bursty IPP source figures 5.7, 5.8 and 5.9 the heuristic formula again gets increasingly inaccurate with the number of superposed sources. While the heuristic formula predicts that the asymptotic behaviour is attained rather quickly the exact models reveal that this is clearly not the case and this behaviour of the heuristic formula seems to cause the inaccuracy when multiplexing many sources.

For all the investigated cases the heuristic approach is an upper bound which is attractive in effective bandwidth applications.

5.5 Conclusions and Future work

- A heuristic formula for the tail probability of a queue has been evaluated. The formula yields additional statistical multiplexing gains as the number of sources increases compared to for example the effective bandwidth approximation in [90].

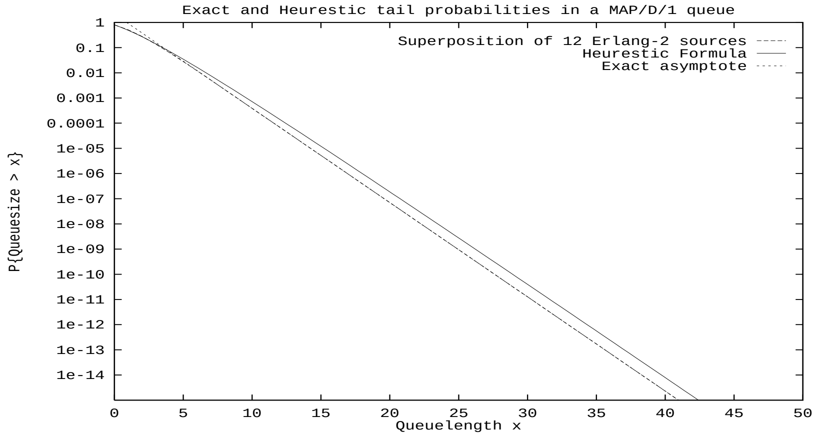

Figure 5.1: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 Erlang-2 sources**

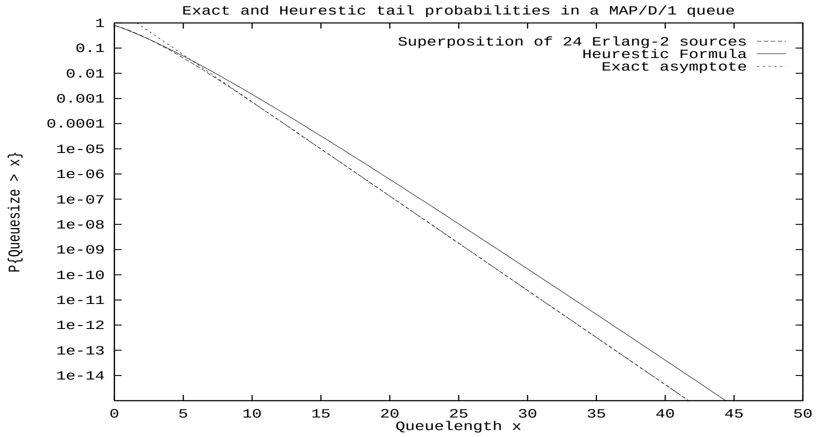

Figure 5.2: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 Erlang-2 sources**

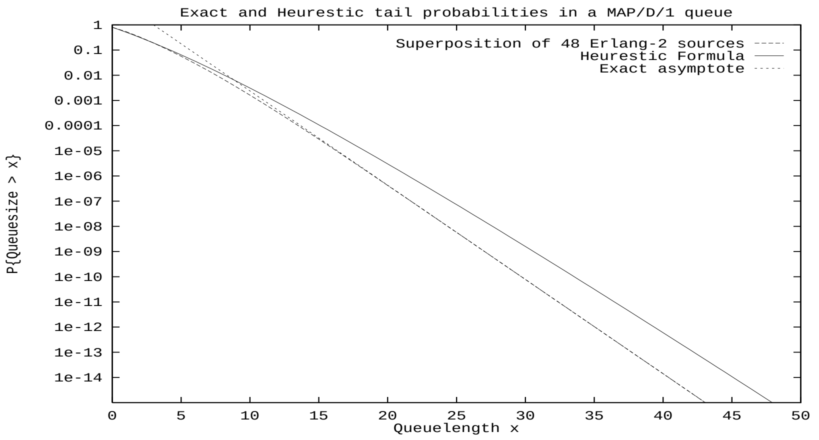

Figure 5.3: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 Erlang-2 sources**

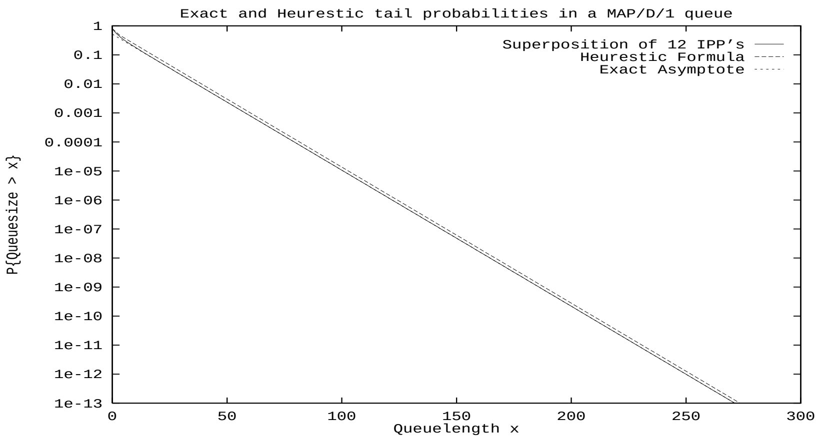

Figure 5.4: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 non bursty IPP sources**

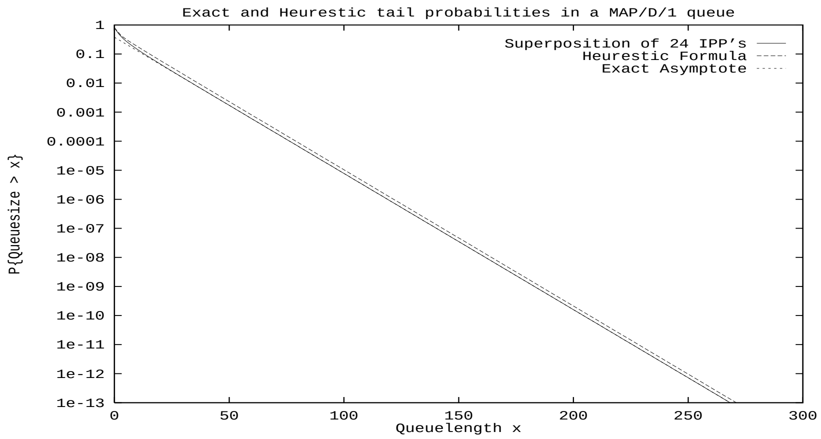

Figure 5.5: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 non bursty IPP sources**

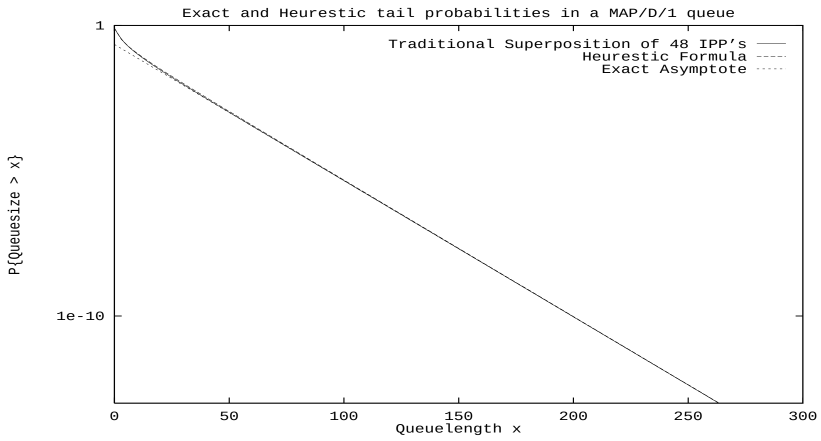

Figure 5.6: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 non bursty IPP sources**

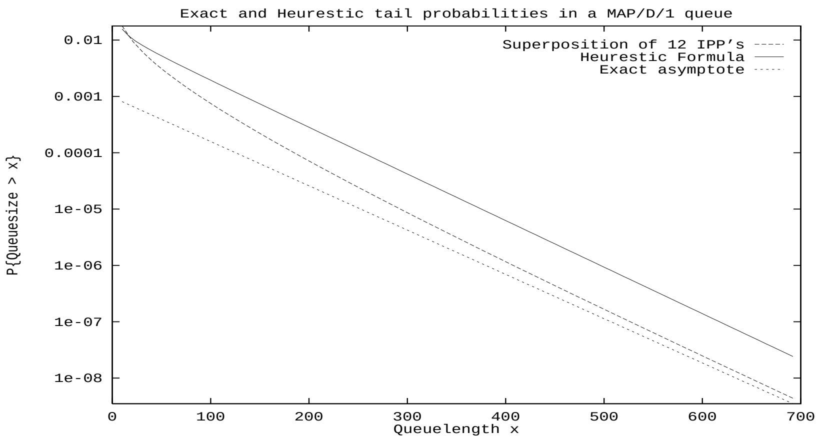

Figure 5.7: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 bursty IPP sources**

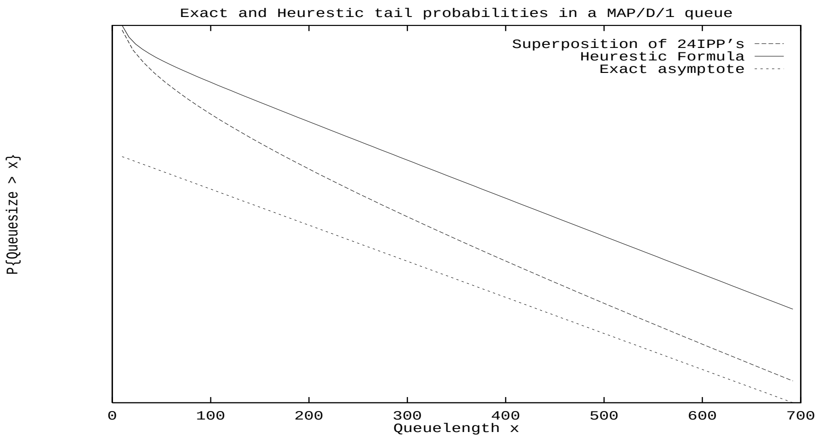

Figure 5.8: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 bursty IPP sources**

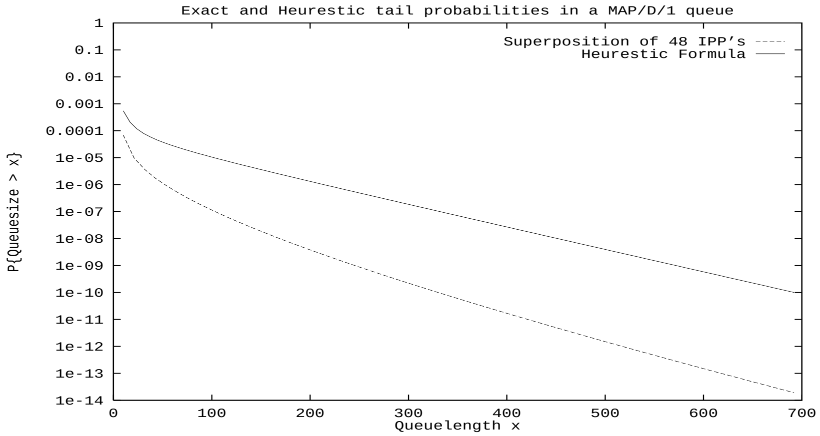

Figure 5.9: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 bursty IPP sources**

- Applying matrix analytic methods made it possible accurately to evaluate the performance of the heuristic formula.
- Suggestions for future work could be for example to try and improve the approximation by perhaps incorporating third order properties of the arrival process.

Chapter 6

Some results regarding the finite capacity BMAP/G/1 queue

In this rather short chapter some results regarding the calculation of the state probabilities in the finite capacity BMAP/G/1 queue will be verbally outlined. For the more specific mathematical details we refer to [A7]. First an attempt will be made to put this rather theoretical article [A7] in context with the rest of the work carried out during the course of this study.

When modelling communication systems the M/G/1 queue has traditionally often been used to dimension buffers, assessing mean and variance of waiting times and so on. This has evidently also been the case when working out recommendations for SS7 [92] where several formulas have been derived from M/G/1 queueing results. The BMAP/G/1 generalizes the M/G/1 results substantially and this can prove valuable when assessing queueing and delay performance as gradually more complex traffic with strong correlation properties is introduced. Especially in the context of modelling B-ISDN traffic the BMAP/G/1 queue seems relevant since B-ISDN traffic is expected to have a quite complicated correlation structure that needs to be taken into account in the performance

evaluation. While a comparison of queueing results obtained by analyzing a finite capacity queue and the corresponding infinite queue might reveal very small differences the finite capacity queue clearly corresponds more to the physical reality when modelling communication systems.

During the course of this study the finite capacity MAP/G/1 queue has been implemented (special case of BMAP with no batch arrivals). We have used the method outlined in [A7]. In this implementation the service time distribution has been chosen as a mixture of deterministic service times (with different lengths), clearly this type of distribution corresponds to the message length distribution in SS7 [92] or [70].

An analysis of the finite capacity BMAP/G/1 queue was carried out by Blondia in [20]. Here expressions for the queue length distribution at departure epochs and at an arbitrary time were derived. Additionally expressions for the virtual waiting time and the mean busy period structure were found.

In [A7] a different approach is applied when calculating the different queuelength distributions. A direct method is used when finding the steady state probability vector for the embedded chain. The queuelength at an arbitrary time is found by noting that the relationship between the embedded chain and the virtual queue is essentially the same as in the infinite capacity case [83] except for the states associated with a full queue. Applying this scheme it is shown in [A7] that in the special case of batch Poisson arrivals the virtual queue length for the finite queue can be given explicitly in terms of the corresponding probabilities in the infinite case.

The direct method applied can be viewed as a slightly modified version of the standard **LU** factorisation scheme used when solving systems of linear equations. One of the modifications is that all computations are done cancellation-free which improves precision considerably. This is achieved by applying the so-called GTH method described in [39]. The complexity of the GTH method is of the same order as the standard **LU** factorisation scheme. Additionally the special structure of the embedded Markov Chain makes it possible to reduce the complexity of the factorisation algorithm from the standard case of order $O(N^3)$ in the number of queueing places N to the order $O(N^2)$. This is achieved by a very slight modification in the standard algorithm. In an implementation the special structure facilitates considerable savings in memory usage compared to the standard case. Due to the way the factorisation scheme works it is also relatively cheap to calculate the queueing behaviour for $N+1$ queueing places if the calculations already have been performed for N queueing places since most of the intermediate results in the factorisation scheme can be reused. Clearly this is a very desirable property when designing a queue for a target blocking probability i.e. in B-ISDN 10^{-9}.

Solving the infinite BMAP/G/1 queue is closely related to finding a certain matrix **G** [77]. The matrix **G** is a solution to a non-linear matrix equation and obtained by applying an iterative scheme. The number of iterations necessary to obtain **G** with sufficient precision becomes extremely large when the queue is conditionally overloaded. By conditional overload we mean that the BMAP has sojourn time(s) in a (set of) state(s) overloading the queue that is long enough for a significant build-up in queue. In this case it might prove faster to solve the finite capacity queue since the performance of the finite capacity algorithm is not dependent on how loaded the queue is. In [A7] it is mentioned that the two algorithms seem to have comparable complexity when the number of iterations needed to obtain **G** is of approximately the same size as the queue length N.

Chapter 7

Conclusions and Further Research

This dissertation has considered various performance aspects of packet switched networks with clear emphasis on the versatile modelling of packet arrival processes with matrix analytic methods. Signalling System No. 7 (SS7) has been used as an explicit case study in some of the work carried out. In the following the main conclusions in the different areas will be presented.

- Traffic data from SS7 have been analyzed and have shown behaviour close to what would be expected for Poisson traffic. A SS7 Changeover/Changeback scenario has been modelled and the results obtained agree well with results previously obtained in [3].
- It has been shown that MAPs of small order, 8-16 states, are able to model self-similar behaviour over several time-scales. Queueing analysis of these MAPs reveal the same adverse behaviour as trace driven simulations of LAN traffic with self-similar characteristics [32]. It has been shown that great care should be taken when predicting queueing behaviour from the first and second order properties of the arrival process since processes with the exact same first and second order properties can yield significantly different queueing behaviour.

- A heuristic formula for the tail behaviour of a queue fed by superpositions of renewal processes has been evaluated. The evaluation was done using exact queueing results for the MAP/D/1 queue. The heuristic formula does not in general seem to out perform already available simplified formulas for the tail behaviour of a single server queue.
- It has been shown that the finite capacity BMAP/G/1 queue can be solved efficiently by applying a cancellation-free scheme yielding greater precision. Also it has been shown that essentially the same approach as with the infinite capacity queue can be applied when calculating the queue length distribution at an arbitrary time. This observation leads to a particularly simple result for the queue length distribution at an arbitrary time in case of a batch Poisson arrival process. In this case the queue length at an arbitrary time can be expressed explicitly by the queue length distributions for the corresponding infinite case.

Finally some suggestions for further work in the different areas will be given.

- The monitoring of packet traffic in a network should be an ongoing effort so clearly analysis of for example measured SS7 traffic should be carried out regularly to reveal characteristics of the current traffic and to avoid being taken by surprise by a change in the traffic profile. Also it might prove valuable to evaluate the Changeover/Changeback procedure for bursty arrival processes.
- A systematic procedure to design MAPs with self-similar characteristics is highly desirable and definitely a necessity if these MAPs with self-similar characteristics are to have more than academic interest i.e. to be applied in practice.
- Call Admission Control (CAC) in B-ISDN is definitely also an area with a lot of research potential. Currently there seems to be far more questions than answers in the literature. A specific initiative could be to try to improve our heuristic formula for example by including third order properties of the arrival process.
- An interesting subject that we currently are beginning to look into is the comparison of the infinite and finite capacity BMAP queues. Here it is of

interest to go deeper into the issue of which of the two types of queues is fastest to solve numerically for different kinds of arrival processes, sizes of load and sizes of queues. It is of independent interest to investigate how large errors are introduced when evaluating a system with known finite capacity by numerically applying a solution method for the infinite capacity queue.

Appendix A

A[1]

A transient queueing study of delay during Changeover
in the Message Transfer Part (MTP)
of Signaling System no. 7 (SS7)

Allan T. Andersen and Bo Friis Nielsen
IMSOR, building 321
Technical University of Denmark
DK-2800 Lyngby

Presented at the Nordic Teletraffic Seminar
NTS11 Stockholm August 1993

A.1 Introduction

This paper is a part of a Ph.D. study concerning common channel signaling systems. The paper presents some preliminary results.

The contribution is concerned with modelling the transient queueing behaviour during signaling link failure/recovery events in the Message Transfer Part (MTP) of CCITT/TSS Signaling System No.7. (SS7).

When a signaling link failure is detected in SS7 Changeover procedures are invoked. The Changeover procedures redirect the traffic from the failed link to alternative links in a manner so that loss, missequencing and duplication of messages is avoided. A message retrieval procedure transfers the messages from the failed link to alternative links.

Changeback is the corresponding procedure for putting a link back in service.

Both Changeover and Changeback procedures can result in a significant queue of messages before message retrieval is initiated. For a Changeover in the transmit buffers associated with the failed link, for a Changeback in the transmit buffers associatied with the alternative link. This is due to the time it takes to accomplish the handshaking or timeout procedures before the message retrieval procedure can start. The case of a failed link (Changeover) has the potential for the largest build up of queued messages because it also takes time to detect link failure.

When a message retrieval procedure is started this prospective queue is propagated through the network. For a Changeover this can result in a queue in the transmit buffers associated with the alternative links, and for a Changeback in a queue in the transmit buffers associated with the link ready to reenter service.

In order for these procedures to have any meaning careful engineering is needed when implementing the signaling system in practice. The buffers should have sufficient size to contain all the messages that could arrive during these procedures. The time it takes to perform a Changeover/Changeback should be sufficiently small to avoid unacceptable message delays.

The potential risk of having a too small buffer could be, depending on the implementation, that once the buffer is full additional messages are discarded. Another possibility is that a congestion situation is declared due to normal buffer fill and traffic control procedures are invoked. In this manner a signaling link

failure can influence the rest of the signaling network. That is, traffic control should only be invoked in case of an overload situation.

An analysis of the transient queueing behaviour during a signaling link failure/recovery event has been carried out in [3]. A Changeover scenario with one alternative link was modelled. Under normal operating conditions Poisson traffic was assumed for each of the signaling links. The message retrieval procedure was modelled by moving messages at a fixed bit-rate from the failed link to an alternative link. During the signaling link failure/recovery events the traffic was found by superposition of the message retrieval traffic and the normal Poisson traffic. During message retrieval, in a Changeover scenario, this has the following implications. At the transmit buffers associated with the failed link messages arrive according to Poisson and messages are removed at a fixed bit-rate determined by the message retrieval speed. At the transmit buffers associated with the alternative link messages arrive as a superposition of the message retrieval arrivals and the normal Poisson traffic. Messages are removed at link speed.

When the transmit buffers associated with the failed link empties the Changeover is completed. The total load destined for these two links is now carried by the alternative link.

A normal approximation was made for the number of arrived Poisson traffic messages during the Changeover/Changeback.

As in [3] our models assume Poisson traffic under normal operating conditions. The message service (emission) time is modelled using a phase type distribution. The message retrieval traffic is modelled by scaling the phase type distribution for the message service (emission) time. The magnitude of the scaling determined by the rate of which the message retrieval procedure moves messages from the failed link to the alternative link.

We consider the same Changeover scenario as [3], but use phase type distributions to characterize the message service (emission) time and the message retrieval traffic.

A benefit of these proposed models are that they are strictly Markovian. The transient distribution of queue-lengths in the transmission buffers can therefore be calculated by solving the ordinary differential equations arising from the

Markov transition matrix numerically and hereby obtaining quantiles of the delay and queue-length distributions.

A problem is here the large number of states necessary to model a system with realistic buffer sizes. To make a numerical method applicable in practice the sparsity of the Markov transition matrix has to be exploited.

The first model is only concerned with the queue in the transmission buffer at the alternative link during a Changeover scenario. During the message retrieval procedure the arrival process to the alternative links transmission buffer is a superposition of the message retrieval traffic and the ordinary Poisson traffic. When the message retrieval procedure has emptied the transmit queue associated with the failed link the arrival process at the alternative link changes. The arrival process then consists of a superposition of the ordinary Poisson traffic and the Poisson traffic normally carried by the failed link.

That is, a Changeover scenario is modelled using one Markov transition matrix during the message retrieval procedure and another transition matrix upon emptying the transmit buffer associated with the failed link.

Numerical results are obtained for this first model of the transmission buffer at the alternative link and compared to those in [3].

The second model is a full Markov model of a Changeover situation including both the transmission buffers at the failed and alternative links.

The buffer sizes used at the links determine the order of magnitude the state space has been enlarged compared to the previous model. In the following B_{max} denotes the maximum number of messages which can be contained in a buffer. This model has in the magnitude B_{max} more states than the previous model.

The initial probability vector here is determined by the queue-lengths at the time of Changeover. An obvious choice would here be to use the stationary distribution for the alternative link and a distribution indicating a large build up of messages at the failed link.

The number of combinations of queue-lengths in this model are $(B_{max} + 1) * (B_{max} + 1)$. In addition several states are generally necessary to describe the state of the arrival and service process for each of these combinations.

The superior transitions in the model are the following. Transitions indicating a message arrival (Poisson). Such an arrival can be to either of the two queues though only to the queue at the failed link if it is non-empty. Transitions indicating that a message has been moved from the failed link to the alternative link. Transitions indicating that a message has been sent on the alternative link.

When a state associated with an empty buffer at the failed link has been reached no states associated with one or more messages in the buffer at the failed link can be reached. This corresponds to the message retrieval procedure being completed. When such a state is reached all Poisson arrivals occurs at the transmit buffer associated with the alternative link.

Numerical results relating to this model has not yet been obtained.

A.2 Relevant procedures in the MTP of SS7

The following is a brief summary of the features of the MTP that has to do with the signaling link operations and the Changeover/Changeback procedures. For further information see [92].

In SS7 each signaling link has full duplex transmission. At link level (level 1 and 2) there are three buffers associated with each end of a signaling link. These are the transmit, the retransmit and the receive buffers. When a message is transmitted it is moved from the transmit buffer to the retransmit buffer. Messages are kept in the retransmission buffer until they are positively acknowledged by the remote end of the link. In order to avoid missequencing each message is given a sequence number.

A positive acknowledgement of a message with a certain sequence number is also a positive acknowledgement of all messages with a lower sequence number. All these messages can therefore be removed from the retransmit buffer. A negative acknowledgement of a message with a given sequence number means that all messages in the retransmit buffer with a sequence number larger or equal to this number has to be retransmitted in sequence. All this is done to avoid message loss and out of sequence messages. The acknowledgement messages are received with the signaling traffic in the opposite direction on the signaling link.

At the receiving end of a link there is signal unit error monitoring to determine the rate of signal unit errors. When there is an excessive failure rate the link is taken out of operation, that is Changeover procedures are initiated. There are several schemes used to detect link failure. One of them is a "leaky bucket" scheme where a counter is incremented every time an errored signal unit is received and decremented for every 256 signal units. If the counter reaches 64 link failure is declared and Changeover procedures are initiated.

Another scenario for failing a link is remote level 2 congestion. That is, excessive fill of the receive buffer at the remote end of the link. This can be caused by errors or overloads other places in signaling network or by higher protocol layers than the MTP. Here timers are used to fail the link if the delay in acknowledging signal units become too excessive.

The main point in the above scenarios for failing a link is that a significant queue of signal units can have formed even before a failure situation is detected.

When one end of a signaling link decides to fail it Changeover procedures are initiated. All received messages over the link are discarded and a Changeover request is sent to the other end of the link. Here of course another signaling path is used. When the far end receives a Changeover request it responds with a Changeover acknowledgement and determines which alternative signaling links to use. It then initiates the message retrieval procedure. This procedure moves all signal units, not already acknowledged, from the transmit and retransmit buffer of the failed link in proper sequence to the alternative links. The signal units that are received during this procedure are queued and sent to the alternative links in proper sequence.

Corresponding Changeback procedures are initiated when a previously failed signaling link is ready to enter service again.

If another signaling path is not available for transmitting the Changeover/Changeback procedures, procedures using timers are invoked to delay messages in order to avoid out of sequence messages. For details regarding these procedures see [92].

The main thing about the above procedure is that no messages are transmitted during handshaking procedures before the message retrieval procedure is initiated.

A.3 Model

The models proposed here require that the ordinary differential equations arising from a Markov transition matrix are solved numerically. That is, the following system is solved as function of time.

$$\frac{d}{dt}\vec{\pi}(t) = \vec{\pi}(t)Q$$

Where Q is the Markov transition matrix and $\vec{\pi}(t)$ denotes state probability vector at time t. The initial condition is defined by $\vec{\pi}(0)$ i.e. the initial distribution of the state probabilities.

Both models are concerned with a Changeover scenario with one alternative link.

As mentioned in the introduction the first model is only concerned with the queueing behaviour in the transmit buffer of the alternative link.

In this model the arrival process during the message retrieval procedure is found as a superposition of the message retrieval traffic and the Poisson traffic.

The service (emission) time is modelled using a phase type distribution [72]. The moments of the phase type distribution are chosen to fit the three first moments of the emission time of a message signal unit. That is, the phase type distribution is determined from the three first moments of the message signal unit length and the link speed. A method for fitting moments in a phase type distribution to a general distribution can be found for example in [4].

This phase type distribution also determines the inter-arrival times of the signal messages that are retrieved from the transmit buffer of the failed queue. The distribution of the inter-arrival time of the retrieved messages is found by scaling the phase type distribution with a scale factor i.e. the ratio between the retrieval speed and the link speed. The distribution of the inter-arrival times of the retrieved signal messages is therefore also of phase type.

During the message retrieval procedure the total arrival process is a superposition of a process with phase type inter-arrival times and exponential inter-arrival times. Such a superposition is a special case of Neuts' versatile Markovian point process [74]. This particular model has been investigated earlier in [58].

When the message retrieval procedure is completed the arrival process concedes of the alternative links normal Poisson traffic and the redirected Poisson traffic from the failed link.

The first model is therefore described by two Markov transition matrices. One describing the queueing behaviour until the message retrieval procedure is completed. The other describing queueing behavior after the message retrieval procedure has been completed.

At the time Changeover is initiated (t=0) the numerical algorithm is started with an initial state probability vector defined by the queue-length in the transmission buffer associated with the alternative link. The Markov transition matrix describing the transient queueing behaviour under the message retrieval procedure is then used for the numerical solution of the differential equation system until the message retrieval procedure ends. When the message retrieval procedure ends the Markov transition matrix describing the pure Poisson arrival process is used. Numerically this implies that the algorithm is restarted at the time message retrieval ends with the new Markov transition matrix and an initial state probability vector equal to the probability vector at the end of the message retrieval procedure.

The Poisson traffic arriving at the failed link and the alternative link is denoted by λ_1 and λ_2. The phase type distribution describing the service time is denoted by $(\vec{\beta}, S)$ and the phase type distribution describing the inter arrival times of retrieved messages is given by $(\vec{\alpha}, T)$. As described before this last phase type distribution describing the message retrieval arrivals is actually a scaled version of the service time distribution that is $\vec{\beta} = \vec{\alpha}$ and $S = \mathrm{a}T$ where a is the ratio between the message retrieval bit-rate and the signal link bit-rate.

The Markov transition matrices described above are of Quasi Birth Death (QBD) type [75]. Using the same notation as in [75] and [72] the two Markov transition matrices can be defined. The general structure of the QBD process is illustrated below.

$$Q = \begin{bmatrix} B_0 & A_0' & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ B_1 & A_1 & A_0 & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & A_2 & A_1 & A_0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & A_2 & A_1 & A_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & A_2 & A_1 & A_0 \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & A_2 & A_1' \end{bmatrix} \tag{A.1}$$

For the period from Changeover to the message retrieval procedure has been completed the matrices defining the QBD process are

$$B_0 = T - \lambda_2 I \tag{A.2}$$
$$B_1 = I \otimes \vec{S^0} \tag{A.3}$$
$$A_0' = (\vec{T^0}\vec{A^0} + \lambda_2 I) \otimes \vec{\beta} \tag{A.4}$$
$$A_0 = (\vec{T^0}\vec{A^0} + \lambda_2 I) \otimes I \tag{A.5}$$
$$A_1 = (T - \lambda_2 I) \otimes I + I \otimes S \tag{A.6}$$
$$A_2 = I \otimes \vec{S^0}\vec{\beta^0} \tag{A.7}$$
$$A_1' = (T + \vec{T^0}\vec{A^0}) \otimes I + I \otimes S \tag{A.8}$$

For the time after the message retrieval procedure has been completed the matrices defining the QBD process have the form

$$B_0 = -\lambda_2 - \lambda_1 \tag{A.9}$$
$$B_1 = \vec{S^0} \tag{A.10}$$
$$A_0' = (\lambda_2 + \lambda_1)\vec{\beta} \tag{A.11}$$
$$A_0 = (\lambda_1 + \lambda_2) I \tag{A.12}$$
$$A_1 = (-\lambda_1 - \lambda_2) I + S \tag{A.13}$$
$$A_2 = \vec{S^0}\vec{\beta^0} \tag{A.14}$$
$$A_1' = S \tag{A.15}$$

It is noted that the last QBD process is the special case of the M/PH/1/K queue.

The initial conditions in the model are determined by the queue-lengths in the transmit buffers of the failed and alternative link at the time Changeover is initiated. The queue-length distribution in the transmit buffer of the alternative link determines the initial probability vector i.e. $\vec{\pi}(0)$. In this model the queue-length in the buffer of the failed link at Changeover time implicitly determines the estimated time it takes to perform the message retrieval procedure.

μ_r is the mean intensity with which messages are moved from the failed link to the alternative link during the retrieval procedure. X_0 is the number of messages in the transfer buffer of the failed link when Changeover is initiated. This can be used to derive a simple fluid flow approximation for time t_r it takes to complete the message retrieval procedure i.e. when the change of Markov transition matrix should occur.

$$\lambda_1 t_r + X_0 = \mu_r t_r \Leftrightarrow t_r = \frac{X_0}{\mu_r - \lambda_1} \tag{A.16}$$

This approximation is used when calculating the time it takes to complete the message retrieval procedure.

The second model is defined by the following superior QBD process.

$$Q = \begin{bmatrix} B_0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ B_1 & A_1 & A_0 & 0 & \dots & 0 & 0 & 0 \\ 0 & A_2 & A_1 & A_0 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & A_2 & A_1 & A_0 \\ 0 & 0 & 0 & 0 & \dots & 0 & A_2 & A_1 \end{bmatrix} \tag{A.17}$$

Here each of the rows of matrices are associated with transitions conditioned on the buffer fill at the failed link.

It is noted that once a state with an empty buffer at the failed link is reached the only other reachable states are associated with an empty transmit buffer at the failed link.

The matrices in **Q** are defined as

$$A_0 = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_1 & 0 & \dots & 0 \\ 0 & 0 & \lambda_1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_1 \end{bmatrix} \tag{A.18}$$

$$A_2 = \begin{bmatrix} 0 & \vec{T}^0\vec{A}^0 \otimes \vec{\beta} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \vec{T}^0\vec{A}^0 \otimes I & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \vec{T}^0\vec{A}^0 \otimes I & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & \vec{T}^0\vec{A}^0 \otimes I \\ 0 & 0 & 0 & 0 & \dots & 0 & \vec{T}^0\vec{A}^0 \otimes I \end{bmatrix} \tag{A.19}$$

$$B_1 = \begin{bmatrix} 0 & \vec{T}^0 \otimes \vec{\beta} & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \vec{T}^0 \otimes I & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & \vec{T}^0 \otimes I & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & \vec{T}^0 \otimes I \\ 0 & 0 & 0 & 0 & \dots & 0 & \vec{T}^0 \otimes I \end{bmatrix} \tag{A.20}$$

$$B_0 = \begin{bmatrix} -(\lambda_2+\lambda_1) & (\lambda_2+\lambda_1)\otimes \vec{\beta} & \dots & 0 & 0 \\ \vec{S}_0 & S-(\lambda_2+\lambda_1)I & \dots & 0 & 0 \\ \vec{0} & \vec{S}^0\vec{\beta}^0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vec{0} & 0 & \dots & S-(\lambda_2+\lambda_1)I & (\lambda_2+\lambda_1)I \\ \vec{0} & 0 & \dots & \vec{S}^0\vec{\beta}^0 & S \end{bmatrix} \tag{A.21}$$

and finally the most complicated

$$A_1 = \begin{bmatrix} B_{i0} & A'_{i0} & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ B_{i1} & A_{i1} & A_{i0} & 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & A_{i2} & A_{i1} & A_{i0} & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & A_{i2} & A_{i1} & A_{i0} & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \dots & A_{i2} & A_{i1} & A_{i0} \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & A_{i2} & A_{i1} \\ \end{bmatrix} \tag{A.22}$$

where

$$B_{i0} = T - (\lambda_2 + \lambda_1)I \tag{A.23}$$
$$B_{i1} = I \otimes \vec{S^0} \tag{A.24}$$
$$A'_{i0} = \lambda_2 I \otimes \vec{\beta} \tag{A.25}$$
$$A_{i0} = \lambda_2 I \otimes I \tag{A.26}$$
$$A_{i1} = (T - (\lambda_2 + \lambda_1)I) \otimes I + I \otimes S \tag{A.27}$$
$$A_{i2} = I \otimes \vec{S^0}\vec{\beta^0} \tag{A.28}$$

A.4 Results

We have some preliminary results regarding the transient queue-lengths in the first model.

The parameter settings we have tried in our model are chosen so that comparison with the previous work in [3] can be done. Unfortunately exact comparison is not yet possible since the form factor of the service time distribution used in [3] was 1.5 while our software does not yet support service time distributions with form factor < 2.

As in [3] the normal load at each link was chosen to be ρ = 0.4, the link and message retrieval speed to be 56 kbs and mean message length was chosen to 28 bytes corresponding to a mean service time of 4 ms. At Changeover time

Figure A.1: **Transient queue-length with service time form factor ϵ = 2.06**

the number of messages in the transmit buffer at the failed link was X_0 = 60, which in our model corresponds to the time t_r = 400 ms to perform the retrieval procedure.

In [3] the above parameters resulted in a maximum queue-length of about 40 messages and a 98 % quantile of about 60 messages during the transient at the alternative link.

Our model is solved with two different values for the form factor of the service time distribution, ϵ = 2.06 and ϵ = 6.25.

These preliminary results indicate that the queue-length is sensitive to the form factor of the service time distribution as would be intuitively expected. The results obtained with ϵ = 2.06 are the most comparable with those obtained in [3] and they do not differ significantly from each other.

Figure A.2: **Transient queue-length with service time form factor ϵ = 6.25**

A.5 Discussion

The results indicate that the distribution of the queue-length is sensitive to changes in the distribution of the signal message length. Such changes can occur when new traffic with different characteristics is introduced in the signaling system. This impact regards both queue-lengths and waiting time.

These observations should be kept in mind when dimensioning the buffer sizes of the signaling links. When the form factor of the distribution of the signal message lengths is increased the number messages allowed in the buffers should also be increased for a given traffic level.

The difference between dimensioning buffers after a maximum number of bytes signal message data or a maximum number of signal message units become more apparent when the form factor of the service time grows.

Another aspect is that in order for the message retrieval procedure to have any meaning the time from a link failure to the retrieval procedure can start has to

be sufficiently small to avoid both excessive delays and an overflow situation at the transmit buffer of the failed link. An implication of this is that strict limits have to be put on the time allowed on the handshaking operations associated with a Changeover.

Further studies would therefore be how the change in the characteristics of the signaling traffic affect the way the signaling system should be designed.

Another area could be the implications of a Changeover on the End to End message signal unit delays.

Appendix B

A[2]

Modelling of apparently self-similar
packet arrival processes with
Markovian Arrival Processes
(MAP).

Allan T. Andersen and Alex Jensen
Bo Friis Nielsen
Institute of Mathematical Modelling
Building 321
Technical University of Denmark
2800 DK-Lyngby

This paper is an enlarged and modified version of COST 242 TD(95)005 which was presented at the COST 242 MC meeting in Cambridge 18-19/1 1995.

B.1 Introduction

Recent measurement studies of packet traffic from Local Area Networks (LAN) [35], Variable Bit Rate (VBR) Video [17] and Integrated Services Digital Network (ISDN) [57] reveal behaviour very different from what classical traffic models assume. The measurements indicate that these forms of packet traffic have long range dependencies and slowly decaying variances. Characteristics of the observed behaviour have in other contexts been associated with self-similar behaviour. Therefore the measurements motivate the assertion that certain types of packet traffic exhibit self-similar behaviour over several time-scales.

Packet traffic with self-similar behaviour has serious implications for both system engineering and system performance. It is problematic to design/dimension systems only with respect to mean load considerations since such an approach totally fails to take into account the behavioural dynamics associated with self-similarity. The challenge is here to develop models that can describe this adverse behaviour reasonably and are able to quantify new performance measures and dimensioning tools.

It is clear that the presence of both short and long term correlations in packet traffic will greatly limit the applicability of traditional renewal traffic models. That is, simple models such as Poisson, Interrupted Poisson and batch Poisson [35] will not be able to model these characteristics.

This has recently initiated a search for traffic models which are able to describe this behaviour in a relatively simple manner. Here models well known in other contexts have been investigated i.e. Chaotic Maps [31] and Fractional Brownian motions [81].

While arrival processes based on such models can be described using only a few parameters it is analytically very difficult to analyze a queueing system with these processes as input.

We believe that further research has to be done in the area of modelling these observed phenomena with already available versatile and analytically tractable point processes. Such an analytical tractable and quite well understood versatile point process is the Markovian Arrival Process (MAP) [64]. The MAP is the special case of the BMAP with batch size one (The BMAP is equivalent with Neuts' versatile point process [74]). MAPs are sufficiently general to serve as

approximations for any stationary point process, see for example [11]. It is clear that the most simple cases of a MAP, the special cases Poisson, Erlang and Interrupted Poisson will not suffice to give a reasonably good description of the observed traffic but it remains to be investigated how complex a MAP model is needed to capture self-similar behaviour over several time-scales. The complexity of the model needed to give a reasonable description of the traffic may make it inapplicable for practical purposes.

One of the main merits in using a MAP model is that a considerable amount of work has already been done on analyzing the queueing behaviour associated with the MAP. Much of this pioneering work is due to M.F. Neuts and co-workers and is described in detail in the two monographs [75] and [77]. The single server queue driven by a BMAP was originally analyzed in [83] by V. Ramaswami using Neuts' versatile point process notation and further in the more transparent BMAP notation by D.M. Lucantoni in [64]. In the context of describing queueing behaviour as a function of the load for a given MAP a very interesting article is [76] where the caudal characteristic curve is introduced. It contains results regarding the asymptotic tail behaviour of queues with a matrix geometric steady state vector [75] as for example a queue driven by a MAP and with a Phase-type service time distribution (A special case of the QBD process [75]). To a certain extent this work has very recently been generalized to queues of M/G/1-type [77], see [1] and particularly [26] for details.

There are several recent publications describing various aspects of the time dynamic behaviour of the MAP (or special cases of the MAP) for example [78], [79] and [80].

This contribution consists of some results regarding the use of the MAP as a model for packet traffic with self-similar behaviour spanning several timescales. The analysis carried out on the measured packet traffic with self-similar properties, see for example [61] or [91], has mainly been focused on second order properties of the packet traffic as for example the variance-time curve. Additionally plots of the evolution in time of the number of packets arriving in a given size timeslot have been examined at different aggregation levels (size of the timeslot). In [61] this has been used to assert that the packet traffic data appears to be "visually" self-similar. In this contribution we apply the same approach when examining our MAP models. Where possible we use analytical results for example for the second order MAP properties but we also use simulation. In particular simulation is used to create artificial "data-traces".

While being the basis for the extreme versatility of the MAP the heavy parameterization is a major practical problem when considering all but the simplest MAPs. This holds true also when considering only MMMPs as a very versatile subclass of the MAPs. In this paper we introduce a subclass of the MMPPs the GMLP with a somehow restricted set of parameter values though still very versatile. The class is introduced as a mean to describe point processes which exhibit self-similar behaviour. Though being a restriction of the MMPP the GMLP still contains many of the usually applied special processes like IPPs and superpositions hereof.

Initially we examine the time dynamic behaviour of different IPPs. Then we look at the resulting process obtained by superposing several identical IPPs. This type of process has been used extensively for modelling purposes since the end of the 1980's. Next we consider the resulting process obtained by superposing several inhomogeneous IPPs and a Poisson stream. Finally we examine some more general GMLP models which can be viewed as generalizations of superposing inhomogeneous IPPs (and a Poisson Stream).

Quite surprisingly it turns that even very simple 8 - 16 state GMLP models as for example superpositions of three to four inhomogeneous IPPs and a Poisson stream seem to be able to model self-similar behaviour over several timescales.

Preliminary results of this work was originally presented at the COST 242 MC meeting in Cambridge January 1995 [7]. At the subsequent COST 242 MC in Stockholm May 1995 related work about the use of Markovian models to model self-similar behaviour was presented [87].

This paper is organized as follows : In section 2 a brief description of self-similarity is given. Section 3 contains a description of the MAP and includes some useful analytical expressions. Here we also introduce the General Multi Level Process (GMLP). In section 4 we look into and discuss the time dynamical behaviour of some special cases of the MAP. Finally, in section 5 we conclude and give directions for further work.

B.2 Measures of Self-Similarity

The short introduction to self-similarity given here is partly based on [61] and [91].

A continuous time stochastic process $X(t)$ with zero mean and stationary increments is said to be self-similar with self-similarity parameter H if $X(at)$ has the same finite dimensional distributions as $a^H X(t)$. H denotes the self-similarity parameter also called the Hurst parameter.

In the context of packet traffic we are concerned with the process defined by the number of packets arriving in a time-interval. This makes it more natural to consider the corresponding stationary discrete time stochastic process $X = X_t$ for t = 1,2,.... (Number of packets arriving in time-interval 1,2 ....).

We also consider the time-series obtained by averaging the discrete time process over non-overlapping blocks. That is, for m = 1,2,3 .., we consider the time-series

$X_k^{(m)} = (1/m)(X_{m(k-1)+1} + \ldots + X_{km})$, k = 1,2,3,....

The stationary discrete time stochastic process X with mean 0 and variance σ^2 is said to be self-similar with self-similarity parameter H if for all m = 1,2,3.... $(X_{m(k-1)+1} + \ldots + X_{km}) = mX_k^{(m)}$, k = 1,2,3,... has the same finite dimensional distributions as $m^H X$.

If the variance and autocorrelation function of $(X_{m(k-1)+1} + \ldots + X_{km}) = mX_k^{(m)}$, k = 1,2,3,... and $m^H X$ are the same for m = 1,2,3,.... X is said to be second-order self-similar with self-similarity parameter H. This implies that for all m = 1,2,3 $var(X^{(m)}) = \sigma^2 m^{-\beta}$, and that $r^{(m)}(k) = r(k) = \delta^2(|k|^{2-\beta})$ (k = 0,1,2,3...) where $\beta = 2 - 2H$ and $\delta^2(f)$ denotes the second central difference operator applied to a function, that is, $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$.

X is called asymptotically second-order self-similar with self-similarity parameter $H = 1 - \beta/2$ if $r^{(m)}(k) \to \delta^2(|k|^{2-\beta})$, as $m \to \infty$ (k = 0,1,2,3,.....)

The definition of long-range dependence is that $\sum_k r(k) \to \infty$. For a self-similar process this corresponds to the self-similarity parameter H in the range $0.5 < H < 1.0$, i.e. for such a process the averaged time-series $X^{(m)}$ possesses a non-degenerate correlation structure as $m \to \infty$. For all conventional teletraffic models $r_m(k) \to 0$ as $m \to \infty$, (k=1,2,3,...), that is, the averaged process

$X^{(m)}$ tends to second order pure noise for large m corresponding to H = 0.5 or equivalently β = 1. In the following a process with H = 0.5 will be called a short-range dependent process.

In practice there are several ways to try to detect whether an observed process exhibits self-similar behaviour or not. A simple way to check for any indications of self-similarity is to plot the averaged process $X^{(m)}$ for different values of m, the values of m differing by an order of magnitude for example. If the plots "look the same" it is an indication of self-similar behaviour.

In the context of packet traffic attention has been directed towards three ways of checking for self-similarity. These can be referred to as methods based on the variance-time plot or equivalently the index of dispersion of counts, the pox diagram of the R/S statistic (also called the rescaled adjusted range plot) and finally analysis based on the periodogram (in the frequency-domain). Here we only consider the methods based on the variance-time plot and the R/S statistic. Regarding the statistically more refined methods available when analyzing the periodogram (i.e. the Whittle estimator) we refer to [61] and [91]. The following presentation of the methods based on the variance-time plot and the R/S statistic is partly based on these references.

First we examine the use of the variance-time plot (and the index of dispersion of counts). For an asymptotically second order self-similar process the variance of the averaged process $X^{(m)}$ decreases more slowly than the reciprocal of m, that is, $Var\{X^{(m)}\} \rightarrow a_1 m^{-\beta}$ as $m \rightarrow \infty$ where $0 < \beta < 1$ and a_1 is a finite positive constant. As mentioned earlier β = 1 (H = 0.5) in traditional teletraffic models. To check for self-similarity $log(Var\{X^{(m)}\})$ is plotted against time ($log(m)$). Ignoring small values of m a line is fitted through the points and an estimate $(-\hat{\beta})$ of the slope $(-\beta)$ is obtained. If the points reasonably well resemble a straight line with $0 < \beta < 1$ it can be stated that the observed process at least has second order self-similar behaviour over the observed time-scales. The self-similarity parameter H is estimated as $\hat{H} = 1 - \hat{\beta}/2$.

When considering the point process of arriving packets in a time-interval, X_t is an integer. In this case it is mathematically equivalent to consider the index of dispersion of counts instead of the variance-time curve when checking for self-similarity and estimating β. Consider a point process and let N_t denote the number of events in the process in the interval $(0, t]$. The index of dispersion of counts for N_t is then defined as $IDC(t) = Var\{N_t\}/E\{N_t\}$. Considering

N_t at discretely spaced values of t, i.e. $t = m\Delta t$, where Δt is the time-interval corresponding to one observation X_t and m = 1,2,3..., is equivalent to considering the process $mX^{(m)}$. That is

$$\begin{aligned} IDC(m\Delta t) = & \; Var\{N_{m\Delta t}\}/E\{N_{m\Delta t}\} \\ = & \; Var\{mX^{(m)}\}/E\{mX^{(m)}\} \\ = & \; mVar\{X^{(m)}\}/E\{X^{(m)}\} \end{aligned}$$

In a stationary point process $E\{X^{(m)}\}$ = const. for all m = 1,2,3,... Hence checking the IDC curve of $N_{m\Delta t} = mX^{(m)}$ for self-similarity is mathematically equivalent to checking the variance-time plot of $X^{(m)}$ for self-similarity. It is clear from the IDC expression that a self-similar process should have the slope $(1-\beta)$, $0 < \beta < 1$ when plotting $log(IDC(m\Delta t))$ against $log(m)$.

Next we consider the rescaled adjusted range of R/S statistic. The rescaled adjusted range or R/S statistic is given by $R(n)/S(n) = 1/S(n)[max(0, W_1, W_2, \ldots W_n)\ min(0, W_1, W_2, \ldots W_n)]$ with $W_k = (X_1 + X_2 + \ldots + X_k) - k\bar{X}(n)$, $k = 1,2,..n$, where $\bar{X}(n)$ is the sample mean and $S^2(n)$ is the sample variance. Historically Hurst, [43] and [44], found naturally occurring time series which appeared to be well represented by the relation $E\{R(n)/S(n)\} \sim cn^H$ as $n \rightarrow \infty$ with Hurst parameter $H > 0.5$ and c a finite positive constant (independent of n). Mandelbrot and Van Ness [67] showed that in a short-range dependent process $E\{R(n)/S(n)\} \sim cn^{0.5}$ as $n \rightarrow \infty$.

In practice R/S analysis is based on a heuristic graphical approach originated by Mandelbrott and Wallis in [68] that tries as fully as possible to exploit the information in a set of observations. Given a sample of N observations $(X_k : k = 1, 2, 3, \ldots, N)$ the whole sample is subdivided into K non-overlapping blocks and the rescaled adjusted range $R(t_i, n)/S(t_i, n)$ is calculated for each of the new "starting points" $t_1 = 1, t_2 = N/K + 1, t_3 = 2N/K + 1, \ldots$ which satisfy $(t_i - 1) + n \leq N$. Then the R/S statistic $R(t_i, n)/S(t_i, n)$ is defined as above with W_k replaced by $W_{t_i+k} - W_{t_i}$ and $S^2(t_i, n)$ is the sample variance of $X_{t_i+1}, X_{t_i+2}, \ldots, X_{t_i+n}$. In this way for a given value of n ("lag"), K R/S samples are obtained for small n and as few as one when n is close to the total sample size N. Taking logarithmically spaced values of n starting with n about 10 and plotting $log(R(t_i, n)/S(t_i, n))$ against $log(n)$ results in the rescaled adjusted range plot (also called the pox diagram of R/S). Typically a rescaled adjusted range plot starts with a transient zone representing the nature of the short-range dependence and then settles down, fluctuating around a straight "street" with a certain slope. The graphical R/S analysis is used to determine whether this linear asymptotic behaviour is supported in the data. If this is the case H can be estimated from the slope.

It is important to keep in mind, even if trivial, that in principle it is impossible to determine if a given data set exhibits self-similar behaviour since the sample size is finite. Given a dataset it is in principle only reasonable to talk about self-similar behaviour over a given number of time-scales determined by the (size of the) dataset.

The merit in very recent packet traffic models considered in the literature for example Fractional Brownian motions [81] and Chaotic maps [31] is that they are able to describe self-similar behaviour with very few parameters. However, as mentioned in the introduction these models are very difficult to handle analytically and to a large extent the tools for calculating performance measures related to these models have not been developed yet. A recent paper about a storage model with self-similar input is [81].

When considering a Markovian Arrival Process (MAP) it is to some extent misleading to talk about self-similar behaviour because it by definition is a short-range dependent process, short-range dependent processes defined as those where the aggregated process $X^{(m)}$ tends to second order pure noise as $m \to \infty$. Even if in theory a MAP can model self-similar behaviour over an arbitrary number of time-scales it will, of course, eventually still turn out to be short-range dependent in the above described sense. The resulting model might also turn out so complex and parameterized that it is inapplicable for practical purposes.

Our objective is to investigate to what degree it is possible to model self-similar behaviour over several time-scales with some relatively simple special cases of the MAP.

B.3 MAP-model and related issues

B.3.1 General MAP formulas

In this section we will briefly sketch some main characteristics of the Markovian Arrival Process (MAP). For further results we refer to to [64] and [77].

The MAP is the point process generated by transition epochs in the Markov renewal process whose transition probability matrix $\mathbf{F}(*)$ is of the form

$$\begin{aligned} \mathbf{F}(x) &= \int_0^x e^{\mathbf{C}u}\,du\mathbf{D} \\ &= (\mathbf{I} - e^{\mathbf{C}x})(-\mathbf{C}^{-1})\mathbf{D} \end{aligned} \tag{B.1}$$

where the square matrices of order m, $\mathbf{C} = [C_{ij}]$ and $\mathbf{D} = [D_{ij}]$ are respectively a stable matrix and a non-negative matrix whose sum is an irreducible infinitesimal generator. (The matrix exponential is defined by its series i.e. $e^{\mathbf{M}} = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}$, for the properties of the matrix exponential see [16].)

The infinitesimal generator $\mathbf{Q} = \mathbf{C} + \mathbf{D}$ describes the transitions among the transient states $1, \ldots, m$ in the underlying Markov process. The stationary vector $\vec{\pi}$ of this Markov process satisfies the equations

$$\vec{\pi}(\mathbf{C} + \mathbf{D}) = \vec{\pi}\mathbf{Q} = \vec{0}, \quad \vec{\pi}\vec{e} = 1 \tag{B.2}$$

where $\vec{e}$ denotes a column vector of 1s.

Some well-known processes are (very) special cases of the MAP

- The Poisson process : $\mathbf{C} = (-\lambda)$ and $\mathbf{D} = (\lambda)$ (Poisson process with rate λ)
- The phase-type (PH) renewal process with representation $(\vec{\alpha}, \mathbf{T})$ is also a special case of the MAP with $\mathbf{C} = \mathbf{T}$, $\mathbf{D} = \vec{T^0}\vec{\alpha}$. It includes both the Erlang and hyper-exponential (equivalent with the Interrupted Poisson Process IPP [59]) and Poisson arrival processes as special cases.

- The Markov-modulated Poisson process (MMPP) is also a special case of the MAP. The MMPP is the doubly stochastic Poisson process whose arrival rate depends on an m-state irreducible Markov process i.e. $\mathbf{C} = \mathbf{Q} - \Lambda, \mathbf{D} = \Lambda$ where the modulating Markov process has infinitesimal generator $\mathbf{Q}$ and $\Lambda = diag(\lambda_1, ..., \lambda_m)$.

A useful property of the class of MAPs is that it is closed under superposition.

The counting function of the MAP is given in the following manner: Let N_t denote the number of arrivals in $(0, t]$ and J_t the state of the Markov Process at time t.

Then the (i, j) entry of the mxm matrices $\mathbf{P}(n, t)$ are given as

$$\mathbf{P}_{ij}(n, t) = P\{N_t = n, J_t = j | N_0 = 0, J_0 = i\} \tag{B.3}$$

The matrices $\mathbf{P}(n, t)$ satisfy the forward Chapman-Kolmogorov equations

$$\begin{aligned} \frac{d}{dt}\mathbf{P}(n, t) = \quad & \mathbf{P}(n, t)\mathbf{C} + \mathbf{P}(n-1, t)\mathbf{D}, n \geq 1, t \geq 0 \\ \mathbf{P}(0, 0) = \quad & \mathbf{I} \end{aligned} \tag{B.4}$$

and the matrix generating function $\mathbf{P}^*(z, t) = \sum_{n=0}^{\infty} \mathbf{P}(n, t) z^n$ is explicitly given by

$$\mathbf{P}^*(z, t) = e^{(\mathbf{C} + z\mathbf{D})t} \quad |z| \leq 1, t \geq 0 \tag{B.5}$$

By differentiating the matrix generating function with respect to z it is possible to derive moments of the number of arrivals N_t at a given time. In particular the mean and variance of N_t, for a given start probability vector of the underlying Markov process, can be found from the first two moments. The variance of N_t is often referred to as the variance time curve. Here we will only consider the mean and variance of N_t in the stationary version of the MAP. Both these quantities are listed below. For further details regarding the detailed derivation of these results we refer to [77].

For the time stationary version of the MAP the mean and variance of the number of arrivals in $(0, t]$ are given by

$$E\{N_t\} = \vec{\pi}\mathbf{D}\vec{e}t \tag{B.6}$$

$$\begin{aligned} Var\{N_t\} = \quad & (1 - 2 * \vec{\pi}\mathbf{D}\vec{e})E\{N_t\} + \\ & 2\vec{\pi}\mathbf{D}(\vec{e}\vec{\pi} - \mathbf{Q})^{-1}\mathbf{D}\vec{e}t - \\ & 2\vec{\pi}\mathbf{D}(\mathbf{I} - e^{\mathbf{Q}t})(\vec{e}\vec{\pi} - \mathbf{Q})^{-2}\mathbf{D}\vec{e} \end{aligned} \tag{B.7}$$

The index of dispersion curve, $IDC(t)$, is defined as the ratio of $Var\{N_t\}$ to $E\{N_t\}$, i.e.

$$\begin{aligned} IDC(t) = \quad & Var\{N_t\}/E\{N_t\} \\ = \quad & (1 - 2 * \vec{\pi}\mathbf{D}\vec{e}) + \\ & 2\vec{\pi}\mathbf{D}(\vec{e}\vec{\pi} - \mathbf{Q})^{-1}\mathbf{D}\vec{e}/(\vec{\pi}\mathbf{D}\vec{e}) - \\ & 2\vec{\pi}\mathbf{D}(\mathbf{I} - e^{\mathbf{Q}t})(\vec{e}\vec{\pi} - \mathbf{Q})^{-2}\mathbf{D}\vec{e}/(\vec{\pi}\mathbf{D}\vec{e}t) \end{aligned} \tag{B.8}$$

Since $e^{\mathbf{Qt}} \rightarrow \vec{e}\vec{\pi}$ as $t \rightarrow \infty$ it is clear from (B.8) that $IDC(t)$ converges towards a fixed value as $t \rightarrow \infty$. The speed of this behaviour depends on the time constants in the system. That is, it is inherent in a MAP - model of traffic with self-similar behaviour that, given the order of the model, there is an upper limit on the time-span over which self-similar behaviour can be modelled.

In section B.2 regarding measures of self-similarity it is shown that an exact second order self-similar process has the exact slope $(1 - \beta) = (2H - 1)$, $0 < \beta < 1$ and $0.5 < H < 1.0$ when plotting $log(IDC(t))$ against $log(t)$. Using the substitution $u = log(t)$ the corresponding derivative for a MAPs $IDC(t)$ curve can be calculated as

$$\frac{d(log(IDC(t)))}{d(log(t))} = \frac{d(log(IDC(e^u)))}{du} = \tag{B.9}$$

$$2\vec{\pi}\mathbf{D}(\mathbf{I}-(\mathbf{I}-\mathbf{Q}e^u)e^{(\mathbf{Q}e^u)})(\vec{e}\vec{\pi}-\mathbf{Q})^{-2}\mathbf{D}\vec{e}e^{-u}/(\vec{\pi}\mathbf{D}\vec{e}IDC(e^u))$$

Formula (B.9) is useful when evaluating how well a given MAP describes a second order self-similar process with a certain Hurst Parameter.

When examining the correlation structure of the MAP the covariance between the number of arrivals in the interval (0,t] and the interval $(t+t_1, t+t_1+t']$ $Cov\{N_{0,t}, N_{t+t_1,t+t_1+t'}\}$ is of great interest. It too can be derived using the matrix generating function. For further details we refer to [77]. The covariance $Cov\{N_{0,t}, N_{t+t_1,t+t_1+t'}\}$ is given as

$$Cov\{N_{0,t}, N_{t+t_1,t+t_1+t'}\} = \vec{\pi}\mathbf{D}(\mathbf{I}-e^{\mathbf{Q}t})e^{\mathbf{Q}t_1}(\mathbf{I}-e^{\mathbf{Q}t'})(\vec{e}\vec{\pi}-\mathrm{Q})^{-2}\mathbf{D}\vec{e} \tag{B.10}$$

For a MAP model the marginal distribution between arrivals is seen to be of phase-type. The probability distribution function for the marginal time t between arrivals is given as

$$P\{T \leq t\} = F(t) = 1-(\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}e^{\mathbf{C}t}\vec{e} \tag{B.11}$$

The moments of the marginal distribution are given as

$$\mu_i = i!(\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}(\mathbf{C})^{-i}\vec{e} \tag{B.12}$$

B.3.2 The special case of superpositions of IPPs

Here we will review some of the basic properties of the Interrupted Poisson Process (IPP) since it can be viewed as a model unit for some of the interesting

GMLP models. More precisely some of the described GMLP models are actually nothing more than superpositions of independent IPP sources and perhaps a Poisson source i.e. superpositions of a number of simple renewal processes. A considerable advantage when applying these models is that $E\{N_t\}$, $Var\{N_t\}$ and $IDC(t)$ are scalar functions.

An IPP can in MAP notation be described as follows.

$$\mathbf{C} = \begin{bmatrix} -c_1-\lambda & c_1 \\ c_2 & -c_2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \tag{B.13}$$

The steady state probability vector is easily seen to be $\vec{\pi} = (\frac{c_2}{c_1+c_2}, \frac{c_1}{c_1+c_2})$.

The mean of the counting process N_t of the IPP is given as

$$E\{N_t\} = \frac{\lambda c_2}{c_1+c_2}t \tag{B.14}$$

From formulas in [41] for example the scalar expression for $Var\{N_t\}$ can be found as

$$Var\{N_t\} = (\frac{\lambda c_2}{c_1+c_2} + \frac{2\lambda^2 c_1 c_2}{(c_1+c_2)^3})t - \frac{2\lambda^2 c_1 c_2}{(c_1+c_2)^4}(1-e^{-(c_1+c_2)t}) \tag{B.15}$$

The covariance function $Cov\{N_{0,t}, N_{t+t_1,t+t_1+t'}\}$ can be calculated as

$$Cov\{N_{0,t}, N_{t+t_1,t+t_1+t'}\} = \tag{B.16}$$

$$\frac{\lambda^2 c_1 c_2 e^{-(c_1+c_2)t_1}(1-e^{-(c_1+c_2)t}-e^{-(c_1+c_2)t'}+e^{-(c_1+c_2)(t'+t)})}{(c_1+c_2)^4}$$

These expressions are also very useful when calculating the variance and covariance functions of processes obtained by superposing several IPPs since the resulting variance can be obtained by summing the variances of the individual processes. This, of course, also applies when summing the individual covariances (The IPP sources are independent). The IDC curve for a superposition of n IPP sources is thus given by

$$IDC(t) = (\sum_{j=1}^{n} Var\{N_t^j\})/(\sum_{j=1}^{n} E\{N_t^j\}) \tag{B.17}$$

Using the substitution $u = log(t)$ the derivative of the log - log plotted IDC curve can be found from

$$\begin{aligned} \frac{d(log(IDC(t)))}{d(log(t))} &= \frac{d(log(\sum_{j=1}^{n} Var\{N_{e^u}^j\}))}{du} - \frac{d(log(\sum_{j=1}^{n} E\{N_{e^u}^j\}))}{du} \\ &= (\frac{1}{\sum_{j=1}^{n} Var\{N_{e^u}^j\}} \frac{d(\sum_{j=1}^{n} Var\{N_{e^u}^j\})}{du}) - 1 \\ &= (\frac{1}{\sum_{j=1}^{n} Var\{N_{e^u}^j\}} (\sum_{j=1}^{n} \frac{d(Var\{N_{e^u}^j\})}{du})) - 1 \end{aligned} \tag{B.18}$$

When superposing homogeneous IPPs the IDC remains constant but the mean arrival rate grows proportionately. Examining the expression for the IDC in the special case of the IPP reveals that it is possible to fix both the mean arrival rate and the IDC for different numbers n of superposed homogeneuos IPPs. This can be done by fixing the following three constants k_1, k_2 and k_3. (The mean arrival-rate is $\frac{k_3}{k_1}$)

$$\begin{aligned} k_1 &= c_1 + c_2 \\ k_2 &= 2\lambda c_1 \\ k_3 &= n\lambda c_2 \end{aligned} \tag{B.19}$$

Hence it is possible to create superpositions with the exact same first and second order properties of the counting process. In terms of the parameters k_1, k_2 and k_3 and the number of superposed sources n the individual IPP is defined as

$$\begin{aligned} c_1 &= \frac{nk_1k_2}{nk_2+2k_3} \\ c_2 &= \frac{2k_1k_3}{nk_2+2k_3} \\ \lambda &= \frac{nk_2+2k_3}{2nk_1} \end{aligned} \tag{B.20}$$

For the IPP it is custom to the define the state where the arrivals occur as the ON state. From (B.20) it is clear that given n superpositions the probability that a particular source is in the ON state is $p_n(\text{ON}) = \frac{2k_3}{2k_3+nk_2}$. Given n sources the probability of n_{ON} sources in the ON state is

$$P\{n_{ON} = j\} = \binom{n}{j} (\frac{2k_3}{2k_3 + nk_2})^j (\frac{nk_2}{2k_3 + nk_2})^{n-j} \tag{B.21}$$

It is straight forward to prove that $np_n(\text{ON}) \to \frac{2k_3}{k_2}$ as $n \to \infty$, $P\{n_{ON} = j\} \to \frac{(\frac{2k_3}{k_2})^j}{j!} e^{-\frac{2k_3}{k_2}}$ as $n \to \infty$ for $j = 0, 1, 2, 3....$ That is, the number of IPPs in the ON state becomes Poisson distributed with parameter $\frac{2k_3}{k_2}$ for $n \to \infty$.

The IPP is stochastically equivalent with the hyper-exponential distribution with two phases [59]. The H_2 distribution function is given in (B.22) with the H_2 parameters expressed in terms of the corresponding MAP parameters [59].

$$\begin{aligned} F(t) &= 1 - pe^{-r_1 t} - (1-p)e^{-r_2 t} \\ r_1 &= \tfrac{1}{2}(\lambda + c_1 + c_2 + \sqrt{(\lambda + c_1 + c_2)^2 - 4\lambda c_2}) \\ r_2 &= \tfrac{1}{2}(\lambda + c_1 + c_2 - \sqrt{(\lambda + c_1 + c_2)^2 - 4\lambda c_2}) \\ p &= \frac{\lambda - r_2}{r_1 + r_2} \end{aligned} \tag{B.22}$$

Since the IPP is a renewal process with H_2 distributed inter-arrival times it is clear that the time to an arrival from a randomly chosen point on the time-axis is also H_2 distributed. Recall that the pdf of the distribution of the time to an arrival from a random point in time, in terms of the inter-arrival time distribution, is given by $f^r(t) = \frac{1-F^i(t)}{m^i_{H_2}}$, where $m^i_{H_2} = \frac{p_i}{r_1} + \frac{1-p_i}{r_2}$. (Parameters with super- or subscript i relate to the inter-arrival time distribution while those with super- or subscript r relate to the distribution of the time to an arrival from a random point in time.)

In terms of the H_2 parameters in the inter-arrival time distribution the H_2 parameters in the distribution of the time to an arrival from a random point in time can easily be found. Clearly only the parameter p differs between the two distributions i.e. $p_r = \frac{r_2 p_i}{r_2 p_i + r_1(1-p_i)}$.

When superposing IPP processes (and perhaps a Poisson source) the above can be used to deduce that the marginal distribution of the superposed process also has a hyper-exponential distribution. This becomes evident when considering the distribution of the minimum of a number, n, of H_2 distributed variables i.e considering $F_{min}(t) = 1 - \prod_{j=1}^{n}(1 - F_j(t))$. In the general case where all the intensity parameters r_{j1} and r_{j2} differ (superpositions of inhomogeneous IPPs) the distribution of $F_{min}(t)$ will be a hyper-exponential distribution with 2^n phases.

Consider the superposed process of a n inhomogeneous IPP processes. The relative amount of arrivals stemming from IPP process j, q_j is given by $q_j = \frac{\lambda_j c_{j2}}{c_{j1}+c_{j2}} / (\sum_{k=1}^{n} \frac{\lambda_k c_{k2}}{c_{k1}+c_{k2}})$, $1 \leq j \leq n$. (Clearly $\sum_{j=1}^{n} q_j = 1$). The probability distribution function for the marginal distribution of the superposition of n IPP processes can be expressed as follows.

$$F_{marg}(t) = 1 - \sum_{j=1}^{n} q_j(1 - F_j^i(t)) \prod_{\substack{k=1 \\ k \neq j}}^{n} (1 - F_k^r(t)) \tag{B.23}$$

It is trivial to find the marginal distribution of the superposition of n IPP processes and a Poisson source from (B.23) since the only modification is that the sum is multiplied with a term $e^{-\lambda_p t}$, that is, $F_{marg_p}(t) = 1 - e^{-\lambda_p t}(1 - F_{marg}(t))$ (λ_p denotes the intensity in the Poisson process). Clearly $F_{marg_p}(t)$ is a hyper-exponential distribution with the same number of phases as $F_{marg}(t)$.

In the case of superposing n homogeneous IPP sources simplifications in the above expression are possible. Instead of the general case of the hyper-exponential

marginal distribution with 2^n phases the one obtained has $n+1$ phases. In this case the marginal distribution is given by $F_{marg}(t) = 1 - (1 - F^i(t))(1 - F^r(t))^{n-1}$. By straightforward calculation the marginal distribution can be found as (B.24)

$$\begin{aligned} F_{marg}(t) = 1 - p_i p_r^{n-1t} e^{-nr_1 t} - (1-p_i)(1-p_r)^{n-1} e^{-nr_2} - \\ \sum_{j=1}^{n-1} \binom{n-1}{j} p_r^j (1-p_r)^{n-1-j} (1-p_i) e^{-(jr_1+(n-j)r_2)t} + \\ \sum_{j=1}^{n-1} \binom{n-1}{j-1} p_r^{j-1} (1-p_r)^{n-j} p_i e^{-(jr_1+(n-j)r_2)t} \end{aligned} \tag{B.24}$$

B.3.3 The General Multi Level Process (GMLP)

In this section we will define a subclass of the Markov Modulated Poisson Process which we will call the General Multi Level Process (GMLP). This process was originally defined in [47]. Interesting special cases of the GMLP will be described.

Let the matrices $\mathbf{F}_{j,k}$ and $\mathbf{G}_{j,k}$ of order $(\prod_{i=1}^{k-1} m_i) \times (\prod_{i=1}^{k-1} m_i)$ have the form

$$\begin{aligned} \mathbf{F}_{j,k} = & \; diag(f_{j,k}, f_{j,k}, ..., f_{j,k}) \\ \mathbf{G}_{j,k} = & \; diag(g_{j,k}, g_{j,k}, ..., g_{j,k}) \end{aligned} \tag{B.25}$$

where $k \geq 1$ and $f_{j,k}, g_{j,k} > 0$

Definition: A $m_d \times m_{d-1} \times \cdots \times m_1$ General Multi Level Process is a MAP with **C** and **D** matrices defined by

$$\begin{aligned} \mathbf{C} = & \; \mathbf{X}_d - \mathbf{D} - diag(r_1, r_2, ...r_{m_1 \cdot m_2 \cdot ..., \cdot m_d}) \\ \mathbf{D} = & \; diag(\lambda_1, \lambda_2, ..., \lambda_{m_1 \cdot m_2 \cdots m_d}), \end{aligned} \tag{B.26}$$

where r_k is the row sum of the kth row in $\mathbf{X}_d$ and $\lambda_k \geq 0$. $\mathbf{X}_d$ is defined by the following recursion formula (applied d times)

$$\mathbf{X}_k = \begin{bmatrix} \mathbf{X}_{k-1} & \mathbf{G}_{1,k} & \mathbf{0}_k & \mathbf{0}_k & \cdots & \mathbf{0}_k & \mathbf{0}_k \\ \mathbf{F}_{1,k} & \mathbf{X}_{k-1} & \mathbf{G}_{2,k} & \mathbf{0}_k & \cdots & \mathbf{0}_k & \mathbf{0}_k \\ \mathbf{0}_k & \mathbf{F}_{2,k} & \mathbf{X}_{k-1} & \mathbf{G}_{3,k} & \cdots & \mathbf{0}_k & \mathbf{0}_k \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k & \mathbf{0}_k & \cdots & \mathbf{F}_{m_k-1,k} & \mathbf{X}_{k-1} \end{bmatrix} \quad \text{(B.27)}$$

starting with $\mathbf{X}_0 = 0$. $\mathbf{0}_k$ denotes the zero matrix of order $(\prod_{i=1}^{k-1} m_i) \times (\prod_{i=1}^{k-1} m_i)$.

□

The individual states in a GMLP model will be denoted as the models base states (at level d). The sets of base states $\{1, 2, .., m_1\}, \{m_1 + 1, .., 2m_1\}$ $..\{(m_d m_{d-1} \cdots m_2 - 1)m_1 + 1, .., m_d m_{d-1} \cdots m_1\}$ will be called the meta states on level $d-1$. By numbering these sets according to their order and dividing them in the sets $\{1, 2, ..m_2\}, \{m_2 + 1, .., 2m_2\}..\{(m_d m_{d-1} \cdots m_3 - 1)m_2 + 1, .., m_d m_{d-1} \cdots m_2\}$ the meta states on level $d-2$ are obtained. This procedure can be continued until level 1 is reached. A model with d levels is said to have depth d.

Before turning to some examples of special cases of the GMLP we will briefly motivate its definition in relation to modelling variability over several time-scales. Clearly the transitions in the modulating process taking place on level k (defined by $\mathbf{F}_{j,k}$ and $\mathbf{G}_{j,k}$ $(j = 1, .., m_k - 1)$) are independent of the transitions taking place on other levels. Let the $\mathbf{F}_{j,k}$'s and $\mathbf{G}_{j,k}$'s be defined in such a manner that the transitions on each level k are associated with distinct time constants. If the transitions on the different levels are associated with different time constants such that these span several time scales it apriori seems feasible to model variability over several time-scales.

Having underlying time constants spanning several time-scales is a necessary, but (unfortunately!) not sufficient requirement when constructing a model with variability over several time-scales. Whether or not a particular model is likely to succeed depends heavily on how the λ's in $\mathbf{D}$ are chosen. The pathological

example of this is that if a $\mathbf{D}$ is chosen with all λ's identical a Poisson Process has been designed (albeit in a complicated manner). Clearly, here the underlying structure of the $\mathbf{F}_{j,k}$'s and $\mathbf{G}_{j,k}$'s has no influence on the process.

While the GMLP limits the number of parameters significantly compared to the general case of the MMPP (or MAP) it is still relatively heavy parameterized. In practice we have therefore restricted ourselves to considering various special cases of the GMLP. Apart from designing the $\mathbf{F}_{j,k}$'s and $\mathbf{G}_{j,k}$'s to yield time constants spanning several time scales the key issue is how systematicly to choose the λ's. Here we have experimented with several strategies. The models we examine in this paper are in fact all special cases of the GMLP.

First we illustrate a very simple case of the GMLP, the superposition of n homogeneous IPP processes. It is well known that the superposed process can be described as a MAP with $n+1$ states i.e..

$$\mathbf{C} = \begin{bmatrix} -n(c_1 + \lambda) & nc_1 & \cdots & 0 & 0 \\ c_2 & -((n-1)(c_1 + \lambda) + c_2) & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & nc_2 & -nc_2 \end{bmatrix} \quad \text{(B.28)}$$

$$\mathbf{D} = diag(n\lambda, (n-1)\lambda, ..., 0)$$

Comparing with (B.26) and (B.27) it is clear that this is a special case of the GMLP process with depth 1 and with the following parameters : $\mathbf{F}_{j,1} = jc_2$, $j = 1, 2, .., n$ $\mathbf{G}_{j,1} = (n+1-j)c_1$, $j = 1, 2, .., n$ and $\lambda_j = (n+1-j)\lambda$, $j = 1, 2, .., n+1$.

In the following we will illustrate two approaches of how to construct GMLP models with variability over several time-scales. In both cases we will restrict ourselves to the symmetric cases where $\mathbf{F}_{l,k} = \mathbf{G}_{m,k}$, $1 \leq l, m \leq \prod_{i=1}^{k-1} m_i$). A main advantage in applying this restriction is that the $\mathbf{Q}$ matrix in the resulting MAP model becomes symmetric and can therefore be diagonalized. This property makes the computation of the matrix exponential particularly simple and is therefore important when examining second order properties of the counting process. For an overview over methods to calculate the matrix exponential we refer to [71]. In [47] more general GMLP models have been examined. For both approaches the outset is to choose a mean arrival rate of the process and the underlying time constants i.e the values $\mathbf{F}_{l,k} = \mathbf{G}_{m,k}$. In the two specific models we consider here a mean arrival rate of 3 per unit is chosen since this seems to correspond well with the Bellcore LAN measurements illustrated in for example [61](1 unit correponds to 10 ms in the Bellcore measurements).

The first method we will denote as the addition method. It is conveniently illustrated graphically in figure B.1. As the figure indicates we will illustrate this method with a specific $2 \times 2 \times 2$ model i.e. a model with depth 3. The underlying transition intensity rates on level 1 are chosen as $5 * 10^{-5}$, on level 2 as $2 * 10^{-4}$ and finally on level 3 as $9 * 10^{-3}$. The scheme for finding the λ's is as follows. Level 0 can be considered a meta state with an associated "arrival rate" 3. The "arrival rates" associated with each of the two meta states on level 1 are found by adding respectively subtracting the same rate, in this specific case 0.75. The "arrival rates" associated with each of the four meta states on level 2 are found by applying the same approach to each of the two meta states on level 1, in this case adding and subtracting the rate 0.55 to the previously found rates. Defining the arrival rates on level 3 the same procedure is applied to each of the four meta states on level 2 where the rate 1.0 is added respectively subtracted. Thus the resulting arrival rates, λ's on level 3 have now been defined.

This special case of the addition method is actually nothing more than a superposition of three inhomogeneous IPPs and a Poisson source. The superposition is defined by :

$$\begin{array}{lll} \lambda_{poisson} & = 0.7 & \\ \lambda & = 2.0 & , c_1 = 9 * 10^{-3} \quad \text{and } c_2 = 9 * 10^{-3} \\ \lambda & = 1.5 & , c_1 = 2 * 10^{-4} \quad \text{and } c_2 = 2 * 10^{-4} \\ \lambda & = 1.1 & , c_1 = 5 * 10^{-5} \quad \text{and } c_2 = 5 * 10^{-5} \end{array} \tag{B.29}$$

The second method will be denoted as the multiplication method. We will illustrate this method with a specific $2 \times 2 \times 2$ model i.e. a model with depth 3. The procedure is graphically outlined in figure B.2. The underlying transition intensity rates on level 1 are chosen as $5 * 10^{-6}$, on level 2 as $5 * 10^{-5}$ and

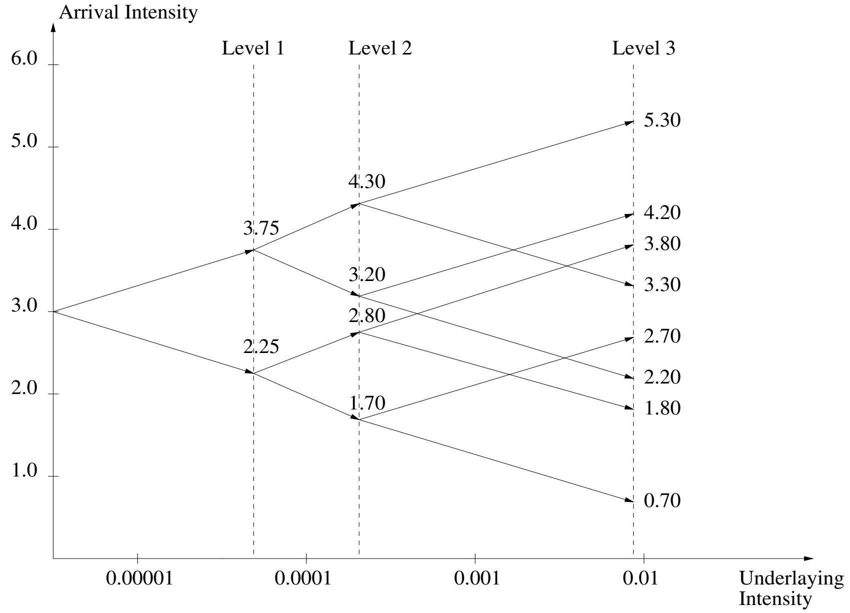

Figure B.1: **Illustration of how the arrival intensities are found with the addition method**

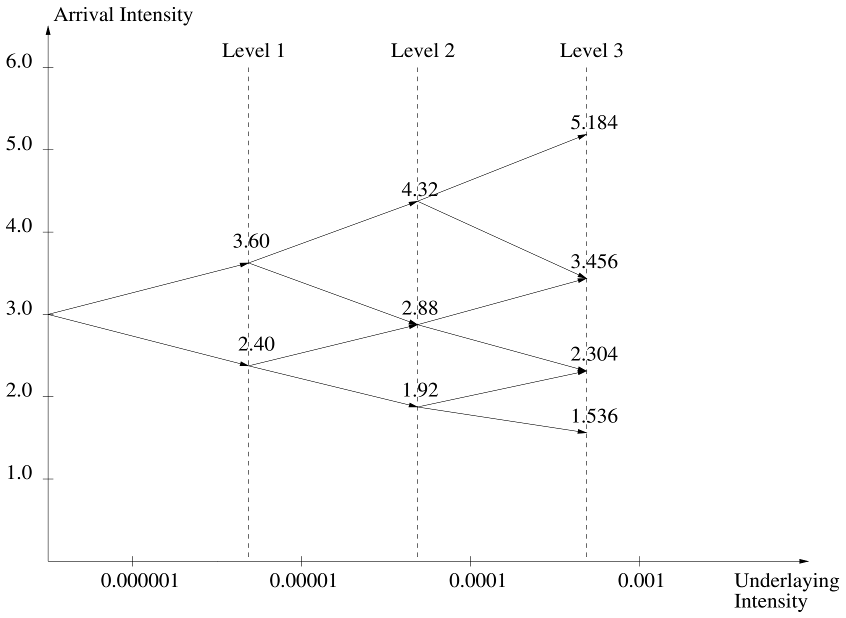

Figure B.2: **Illustration of how the arrival intensities are found with the multiplication method**

finally on level 3 as $5 * 10^{-4}$. The approach in defining the λ's is similar to that of the addition method. Here the outset is also to consider a level 0 with an associated "arrival rate" 3. However, here the "arrival rates" associated with the 2 meta states on level 1 are found as the result of multiplying the factors 0.8 respectively 1.2 with the "arrival rate" on level 0. In the same way as the corresponding procedure for the addition method this multiplication procedure is applied to the meta states on level 1 and level 2 in a recursive manner. The multiplication factors used on each level are 0.8 respectively 1.2.

Since only one transition intensity parameter is associated with each level the "arrival rate" of 3 is clearly preserved through the defining process of both the addition and the multiplication method. Hence implicit in the above procedures for selecting the λ's is that the mean arrival rate is fixed initially. Examining

figure B.2 it becomes evident that using the same multiplication factors on each level implies that with a $2 \times 2 \times \cdots \times 2$ model of depth n only $n + 1$ different λ's are created, hence several states have the same λ. Clearly the number of states with a certain arrival rate can be found as binomial coefficients.

B.4 Time dynamic behaviour of special cases of the MAP arrival process

In this section we examine the dynamic behaviour and correlation structure of different special cases of a MAP. We use both the analytical MAP tools outlined above and simulation. Simulating MAP processes makes it possible to get a first look at how $X^{(m)}$ behaves at different aggregation levels m. Here it is interesting to examine a realization at different aggregation levels m, differing for example by an order of magnitude, to see how the process behaves over several time-scales. Whenever we in the following visually inspect a realization we look at aggregation levels differing by an order of magnitude. Starting with the largest aggregation level, we successively examine the next lower aggregation level by zooming in on a randomly chosen part of the process at the previous aggregation level. This is similar to the method used in [61] and [62]. Simulation of MAPs also makes it possible to calculate the R/S statistic for a given realization.

B.4.1 Time dynamic behaviour of Interrupted Poisson Processes

A very simple MAP yielding correlations in the process X_t is the IPP (B.13). As previously mentioned the IPP is a renewal process with (2 phased) hyper-exponentially distributed inter-arrival times [59] . The sum of the transition rates in the modulating process $(c_1 + c_2)$ is called the time constant for the IPP.

We examine the dynamic behaviour of three IPPs with the same λ but with different time constants spanning 4 orders of magnitude. They can be defined as $\lambda = 4.0$, $c_1 = s\frac{1}{4}$ and $c_2 = s\frac{3}{4}$, where s is either 10^{-2}, 10^{-4} or 10^{-6}. These three IPPs will be denoted respectively IPP_a,IPP_b and IPP_c.

In figure B.3 a realization of IPP$_b$ is shown. It is obvious that the on/off type behaviour does not resemble self-similar behaviour. The on/off type behaviour is only visible when viewing the process at aggregation levels corresponding to the number of arrivals in a time interval that is small compared to the reciprocal of the time constant.

Next we consider the index of dispersion of counts (IDC) curve for each of the three processes IPP$_a$, IPP$_b$ and IPP$_c$ in figure B.4. It is evident in the log-log plot that the index of dispersion of counts for all 3 processes increases approximately linearly over several time-scales, the number of time-scales determined by the IPPs time constant. It is also evident that the slope of the linear growth in the log-log plot corresponds to $\beta \approx 0$ giving a Hurst parameter of $H \approx 1$ in the interval of linear growth. This can more formally be seen from figure B.5 where the derivatives of three IDC curves in a $log - log$ plot is displayed. This behaviour is not surprising when examining the formulas (B.14) and (B.15). For the IPP it is clear that for $t(c_1 + c_2) << 1$ the IDC curve can be represented by its Taylor expansion as

$$IDC(t) = \quad 1 + \frac{\lambda c_1}{(c_1+c_2)} t + o(t), \quad t(c_1 + c_2) << 1 \tag{B.30}$$

This behaviour is, of course, also evident when examining the correlation structure for example via the covariance function given in formula (B.16). Setting $t = t' = \Delta t$ and $t_1 = (k-1)\Delta t$, evaluating the covariance function for k =1,2,3,..... corresponds to calculating the covariance at lag k for the number of arrivals in time-slots of size Δt. Normalizing with the variance ,found from formula (B.15), yields the autocorrelation function and corresponds to the r(k) function in Section B.2. From Section B.2 it is clear that plotting log(r(k)) against log(k) gives a straight line with slope $-\beta$ for a self-similar process.

Plotting r(k) against k for IPP$_b$ with log-scales on both axis shows a nearly linear behaviour over several time-scales see figure B.6. Not surprisingly the slope is very close to 0 corresponding to β = 0 in agreement with the IDC function. At lags corresponding to the time constant in IPP$_b$ the autocorrelations start decaying exponentially also in agreement with the IDC function.

It is well known that for a renewal process $IDC(t) \rightarrow (C_{var})^2$, as $t \rightarrow \infty$ (C_{var} , μ_1 and σ are respectively the coefficient of variation ,the mean and standard

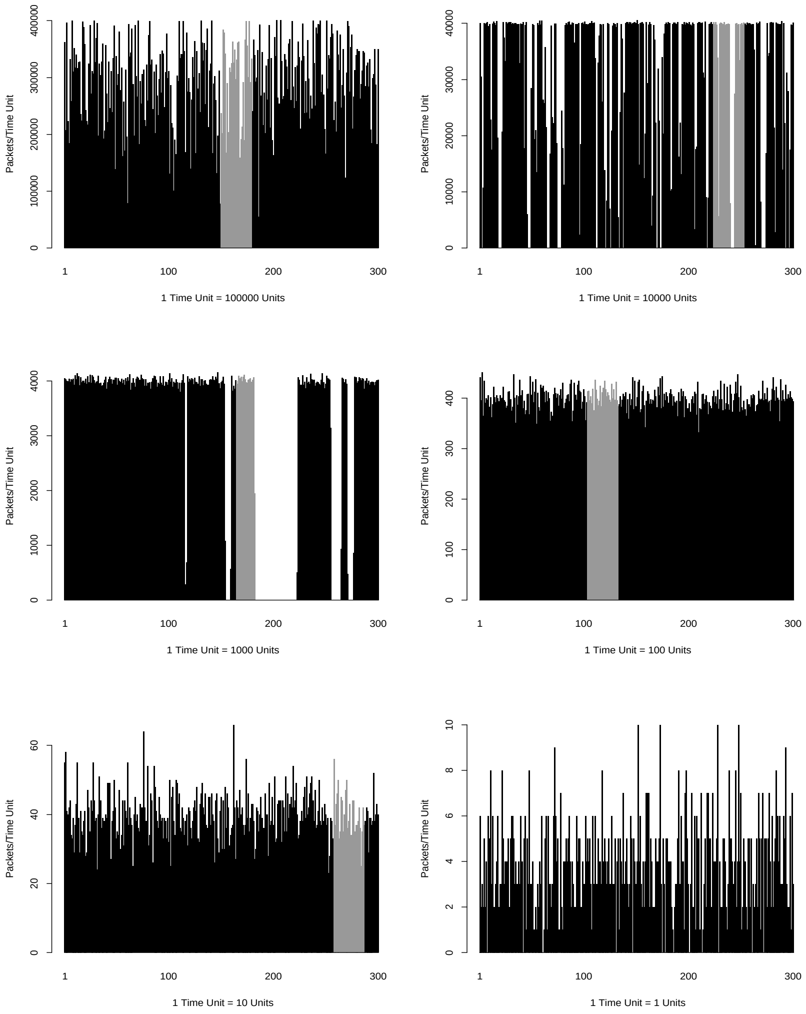

Figure B.3: **Time dynamic behaviour of an IPP (IPP$_b$) over 6 timescales**

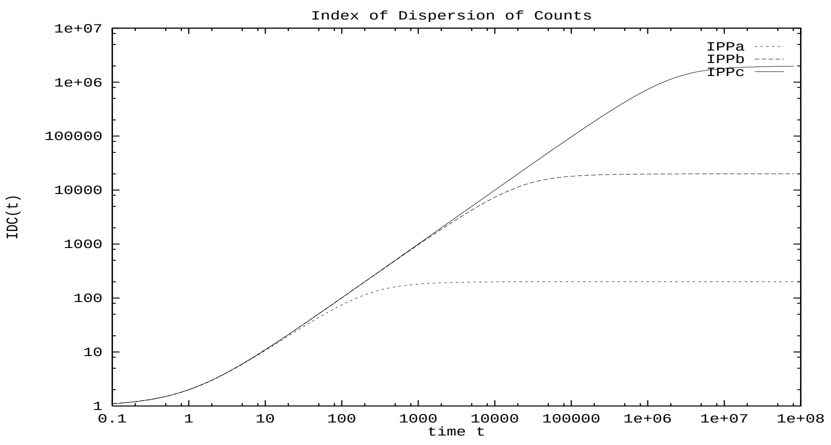

Figure B.4: **The IDC curves of IPP$_a$, IPP$_b$ and IPP$_c$**

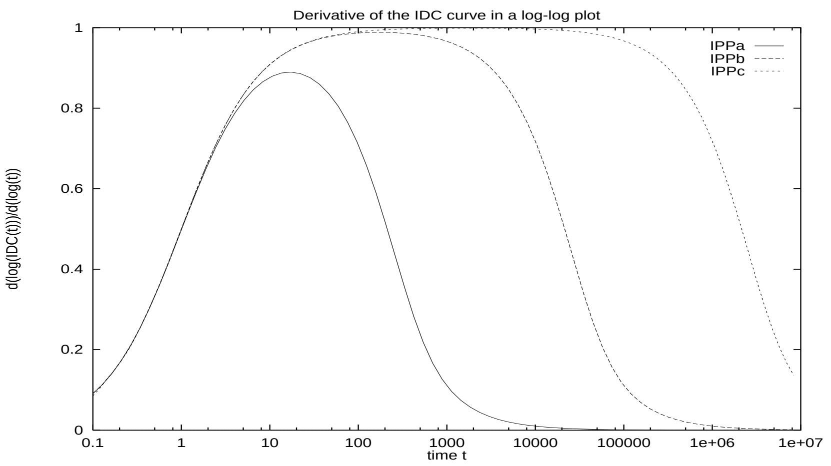

Figure B.5: **Derivatives of the log-log plotted IDC curves for IPP$_a$, IPP$_b$ and IPP$_c$**

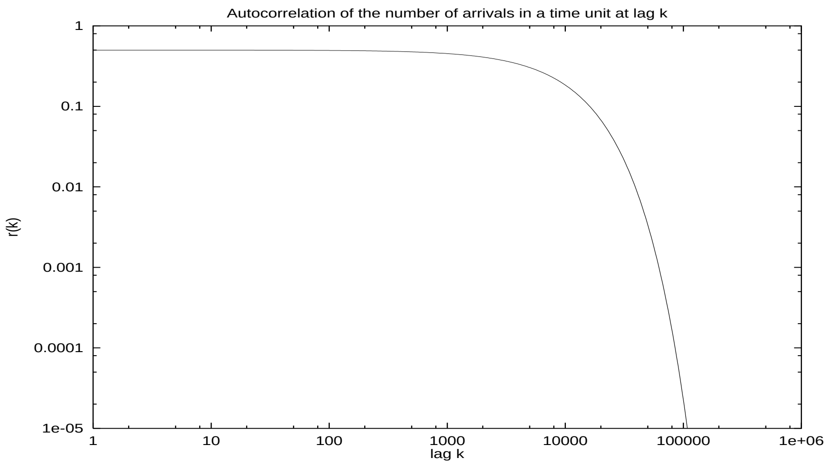

Figure B.6: **The autocorrelation function of IPP$_b$ as a function of the lag**

deviation of the inter-arrival time distribution). Since the IPP is a renewal process "the price paid" for dependence over many time-scales is therefore a highly dispersed marginal distribution. The three IPPs above have the following coefficients of variation, IPP$_a$: C_{var} = 14.2 , IPP$_b$: C_{var} = 141.4 and IPP$_c$: C_{var} = 1414.2.

B.4.2 Time dynamic behaviour of superposed homogeneous Interrupted Poisson Processes

Next we consider the superposition of several identical IPPs. As described in section B.3.2 it is possible to choose a process of superpositions of IPPs to have the same arrival rate and $IDC(t)$ as a single IPP. Here we have chosen to look at two processes, a superposition of 9 identical IPPs and a superposition of 30 identical IPPs, both defined such that the resulting process has the same rate and $IDC(t)$ as IPP$_b$. From (B.20) it can be derived that each of the 9 IPPs to be superposed are defined by $c_1 = 7.5 * 10^{-5}$, $c_2 = 2.5 * 10^{-5}$ and $\lambda = \frac{4}{3}$ and that each of the 30 IPPs to be superposed are defined by $c_1 = \frac{10}{11} * 10^{-4}$, $c_2 = \frac{1}{11} * 10^{-4}$ and $\lambda = 1.1$. Figure B.7 shows a realization of the superposition of 9 IPPs with the same rate and $IDC(t)$ as IPP$_b$ and figure B.8 shows the corresponding realization for the superposition of 30 IPPs.

While modelling self-similar behaviour significantly better than the single IPP it is still evident from the plots of the different aggregation levels that the plots do not “look the same” over several timescales (at least not for these relatively modest choices of n). The superposition of 30 IPP sources seems more realistic in the sense that the levels, i.e number of IPPs in the ON state, are not as easily identified as in the superposition of 9 sources. An inherent limitation in the use of superposed homogeneous IPPs in the context of modelling self-similarity is that they with regards to $IDC(t)$ are no more general than a single IPP. However, the marginal distribution of the superposed process compared to the single IPP with the same rate and IDC is not as dispersed. For the process of 9 superposed IPPs the coefficient of variation is C_{var} = 44.8 and for the process of 30 superposed IPPs the coefficient of variation is C_{var} = 35.5 (recall that for IPP$_b$ C_{var} = 141.4).

B.4.3 Time dynamic behaviour of some multiple level GMLP models

Examples of models created by the addition method

First we consider the 8 state GMLP model created by the addition method which was used as an example in section B.3.3 i.e. the model defined by (B.29). This model can as previously stated be percieved as a superposition of 3 IPPs and a Poisson source. In figure B.9 a realization of the time dynamic behaviour is illustrated. The coefficient of variation of the marginal distribution is C_{var} = 1.36. We will denote this model ADD$_8$.

We also consider the 16 state GMLP model, also defined by the addition method, which can be percieved as the superposition of the 4 IPPs and a Poisson stream in (B.31). For this process the coefficient of variation of the marginal distribution is C_{var} = 1.55. Figure B.10 displays the dynamic behaviour of this model which we will denote ADD$_{16}$.

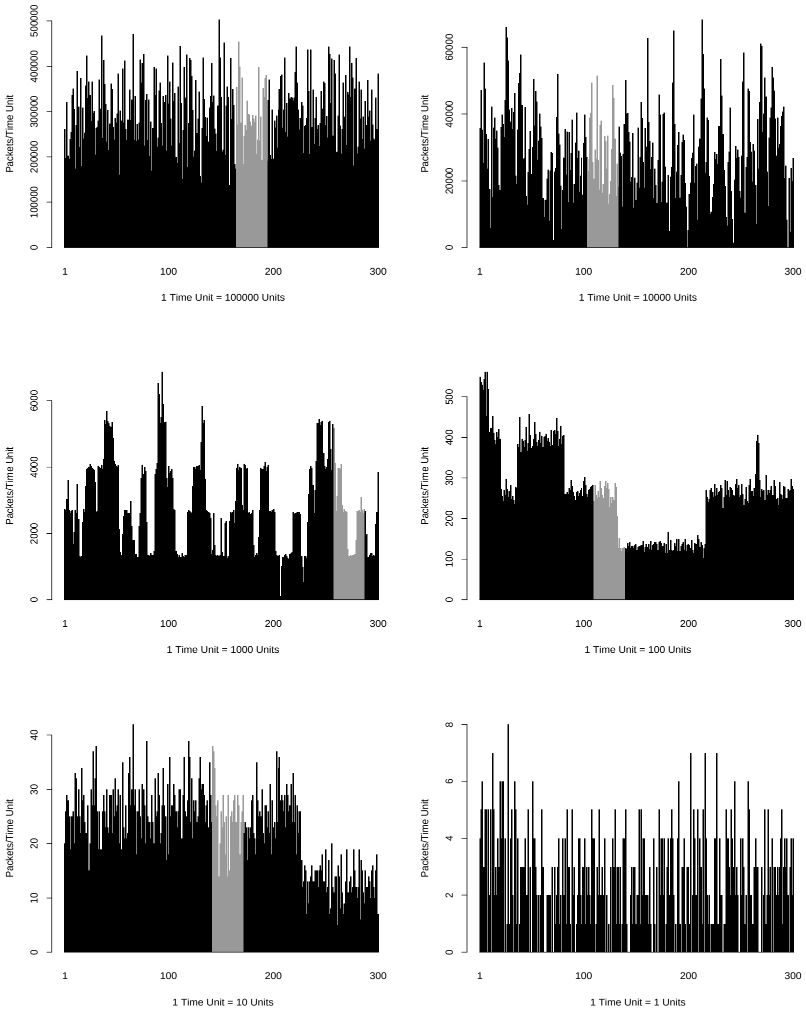

Figure B.7: **Time dynamic behaviour of a superposition of 9 homogenous IPPs with the same rate and IDC(t) as IPP$_b$ over 6 timescales**

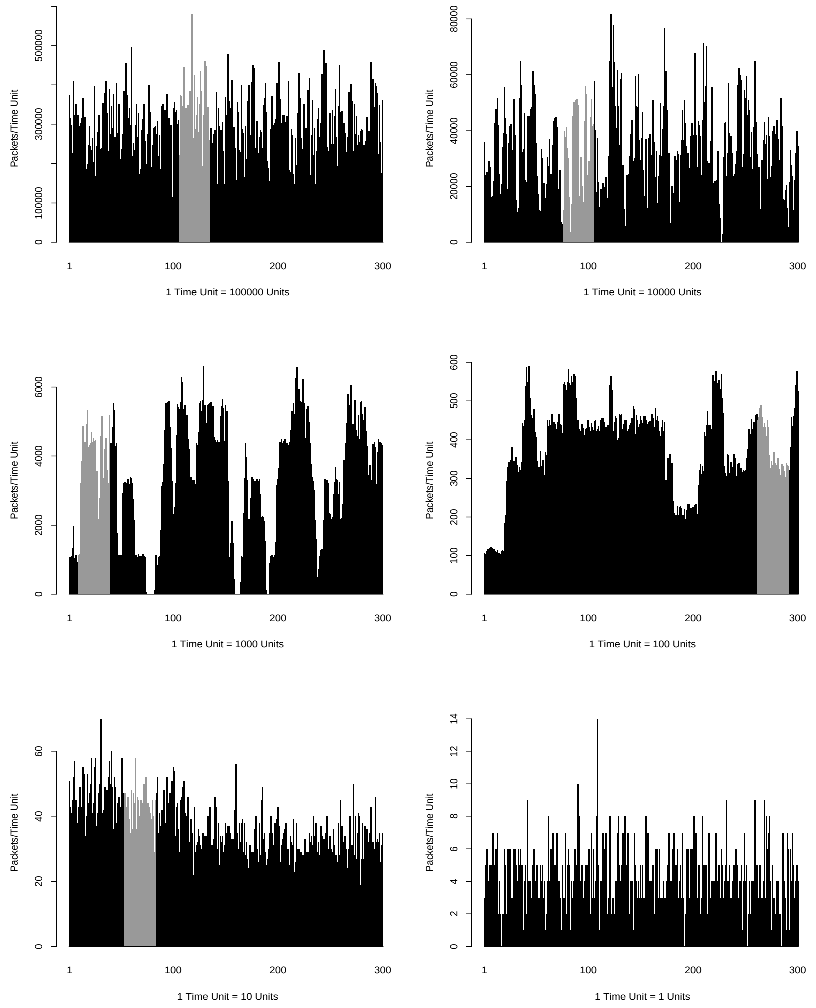

Figure B.8: **Time dynamic behaviour of a superposition of 30 homogenous IPPs with the same rate and IDC(t) as IPP$_b$ over 6 timescales**

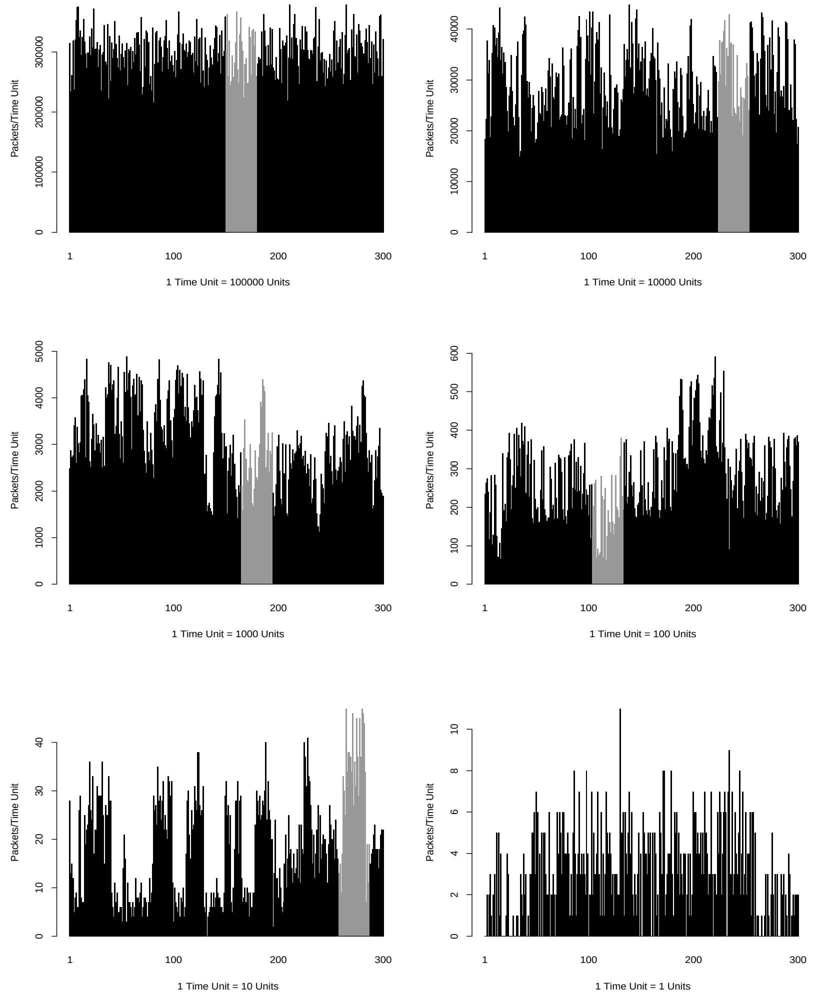

Figure B.9: **Time dynamic behaviour of an 8 state GMLP model created with the addition method ADD$_8$ over 6 time scales**

$$\begin{array}{llll} \lambda_{poisson} & = 0.3 & & \\ \lambda & = 0.8 & , c_1 = 5 * 10^{-3} & \text{and } c_2 = 5 * 10^{-3} \\ \lambda & = 2.0 & , c_1 = 5 * 10^{-4} & \text{and } c_2 = 5 * 10^{-4} \\ \lambda & = 1.1 & , c_1 = 5 * 10^{-5} & \text{and } c_2 = 5 * 10^{-5} \\ \lambda & = 1.5 & , c_1 = 5 * 10^{-6} & \text{and } c_2 = 5 * 10^{-6} \end{array} \tag{B.31}$$

For both models there is evidently high variability on all the time-scales displayed. Except for the highest and lowest aggregation level it apriori seems difficult to decide which aggregation level the individual plots belong to. It can be asserted that visually they appear to be at least as self-similar as the plots shown in [61].

To get more formal measures of the dynamic behaviour of the two models ADD_8 and ADD_{16} the IDC curves are calculated in figure B.11 and figure B.12. They clearly indicate that in both cases there are correlations spanning several time-scales.

The derivatives ($1-\beta$ for a self-similar process) of the log - log plotted IDC curves (can be found for example from (B.9) or (B.18)) are displayed in the figures B.13 and B.14.

The derivatives definitely lie in a band between 0 and 1 over several time-scales. This indicates a “Hurst parameter” somewhere between $\frac{1}{2}$ and 1 over several time-scales. Hence these GMLP models constructed by the addition method are, not surprisingly, more versatile in modelling the IDC of a self-similar process than the previous homogeneuos superpositions of IPPs.

In figure B.15 the autocorrelation function of ADD_8 is shown at different lags. It shows the same signs as figure B.13 of being able to model self-similar behaviour over at least a couple of time-scales with a Hurst parameter between $\frac{1}{2}$ and 1.0.

In figure B.16 the R/S statistic is displayed for ADD_8. Clearly a problem with the R/S plot is that possible variations in the Hurst parameter are not easily detected.

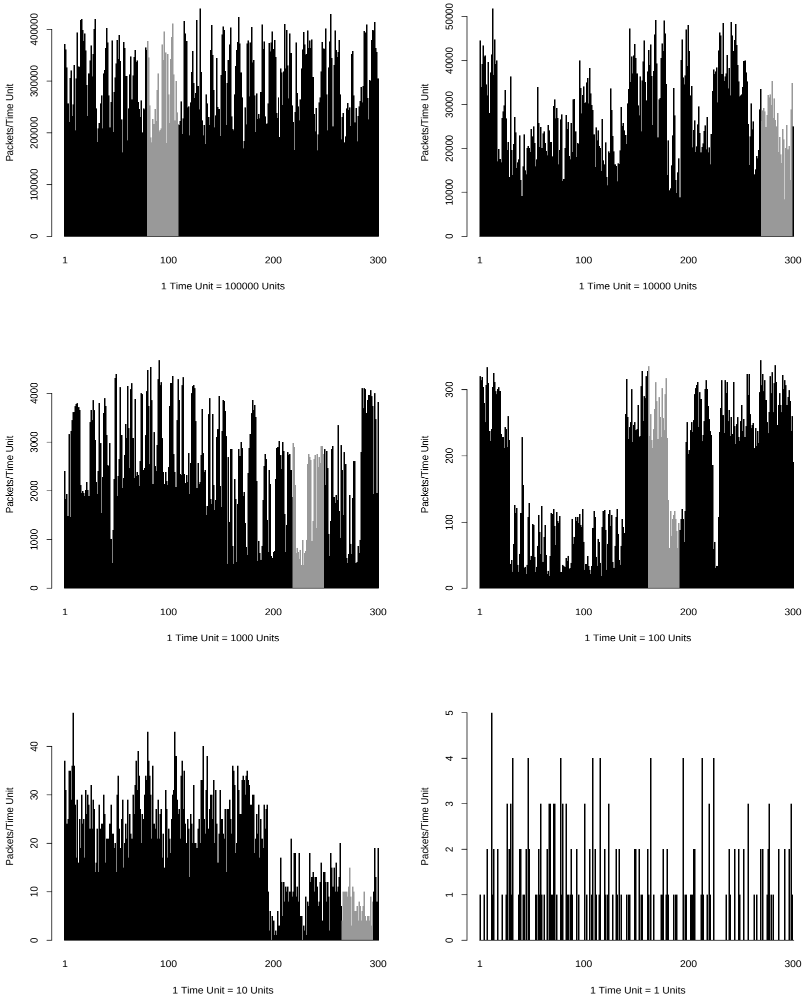

Figure B.10: **Time dynamic behaviour of an 16 state GMLP model created with the addition method ADD_{16} over 6 time scales**

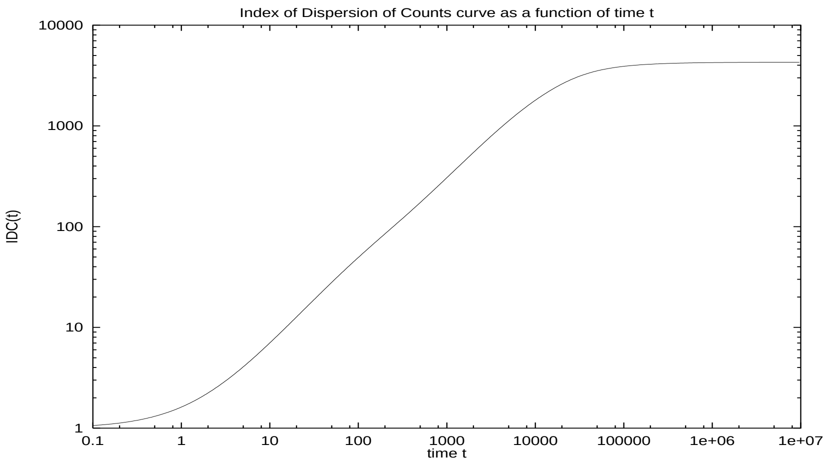

Figure B.11: **The IDC curve for ADD$_8$**

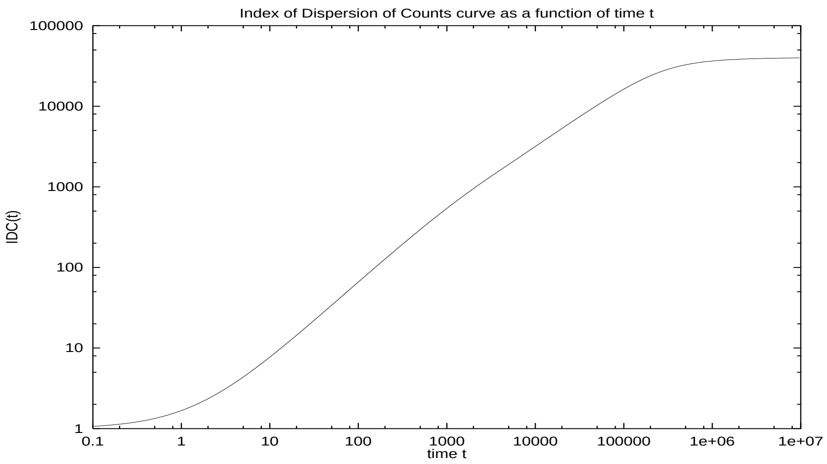

Figure B.12: **The IDC curve for ADD$_{16}$**

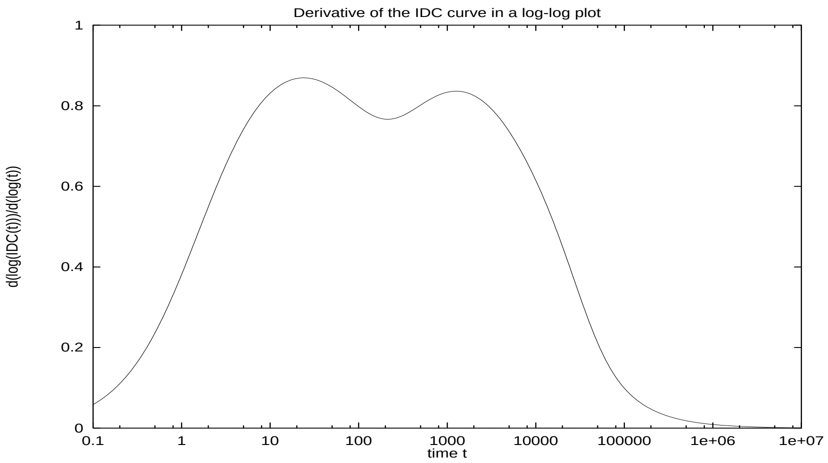

Figure B.13: **The derivative of the IDC curve for ADD$_8$**

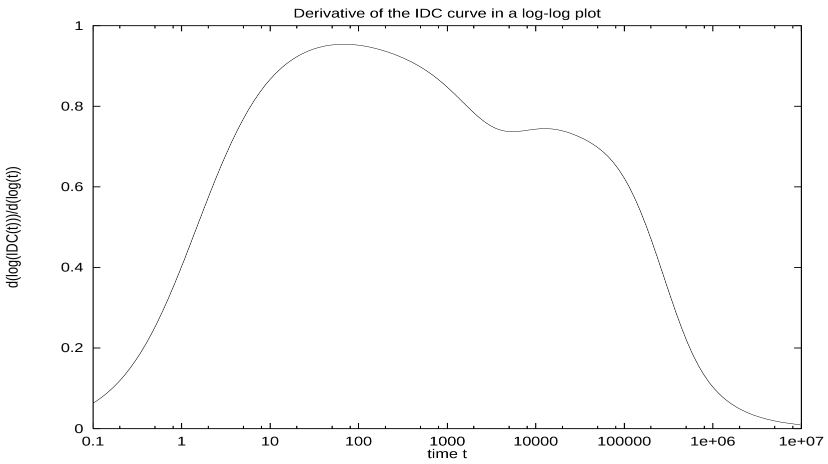

Figure B.14: **The derivative of the IDC curve for ADD$_{16}$**

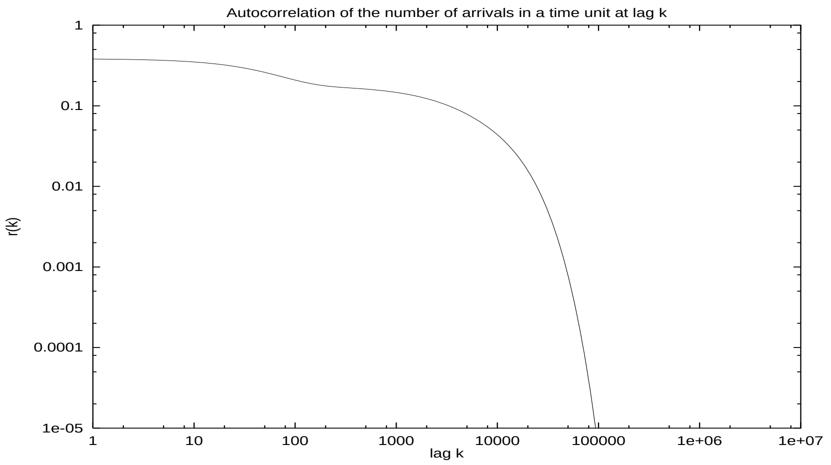

Figure B.15: **The autocorrelation function of the Counting Process for ADD$_8$**

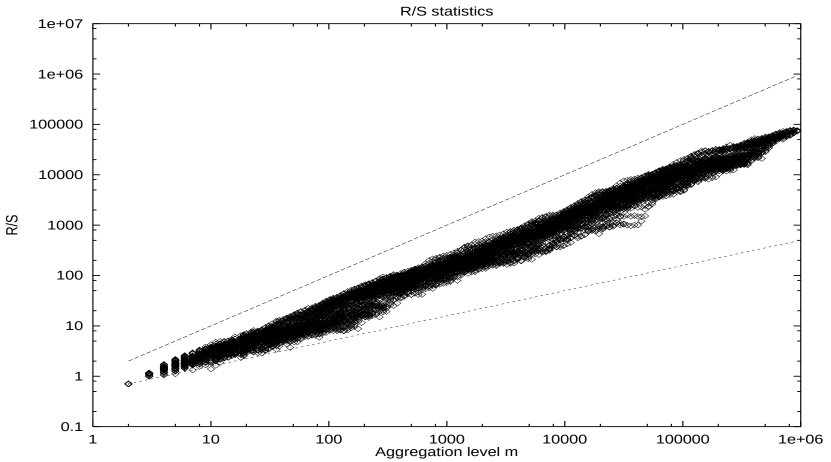

Figure B.16: **R/S statistic for ADD$_8$**

Example of a model created by the multiplication method

Here we consider the 16 state GMLP model created by the multiplication method which can be seen as a generalization of the model used as an example in section B.3.3 and illustrated in figure B.2 to have one more level i.e. to have depth 4. The considered model is a $2 \times 2 \times 2 \times 2$ model. The underlying transition intensity rates on level 1 are chosen as $5 * 10^{-6}$, on level 2 as $5 * 10^{-5}$, on level 3 as $5 * 10^{-4}$ and finally on level 4 as $5 * 10^{-3}$. As in the multiplication example in section B.3.3 we apply the multiplication factors 0.8 respectively 1.2 so the procedure applied here is very analog to the example except here we go one level deeper. The coefficient of variation of the marginal distribution is C_{var} = 1.16. We will denote this model MUL$_{16}$. Figure B.17 displays the dynamic behaviour of this model.

As the models previously created with the addition method MUL$_{16}$ displays high variability on all the time-scales displayed. The IDC curve is calculated in figure B.18 and the derivative of the log - log plotted IDC curve is displayed in figure B.19.

Also in this case the derivative definitely lies in a band between 0 and 1 over several time-scales indicating a “Hurst parameter” somewhere between $\frac{1}{2}$ and 1 over several time-scales. In figure B.20 the autocorrelation function of MUL$_{16}$ is shown at different lags.

In figure B.21 the R/S statistic is displayed for MUL$_{16}$.

B.5 Conclusions and directions for further work

The displayed results from the 3 GMLP models created using the addition or multiplication method all show a Hurst parameter around 0.9 over several timescales. The confidence intervals (95 %) obtained for the Hurst parameter estimates in for example [91] are typically rather wide, at least of the order +- 0.05. It is an open question whether the Hurst parameter is constant or displays small fluctuations in the measured traffic traces of up to an one hour length (see for example [91]). However, in [91] at least in the datasets OCT90 and FEB92 there are statistically significant differences in the Hurst parameters obtained during different 1 hour periods of the same 24 hours.

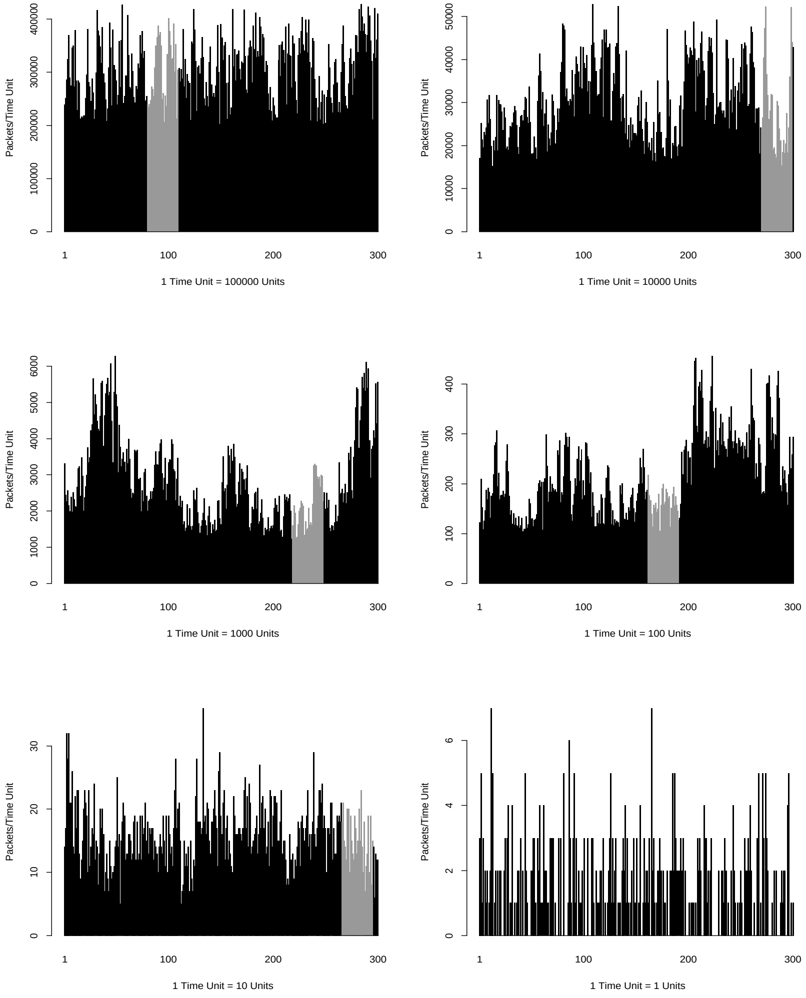

Figure B.17: **Time dynamic behaviour of an 16 state GMLP model created with the multiplication method $\mathbf{MUL_{16}}$ over 6 time scales**

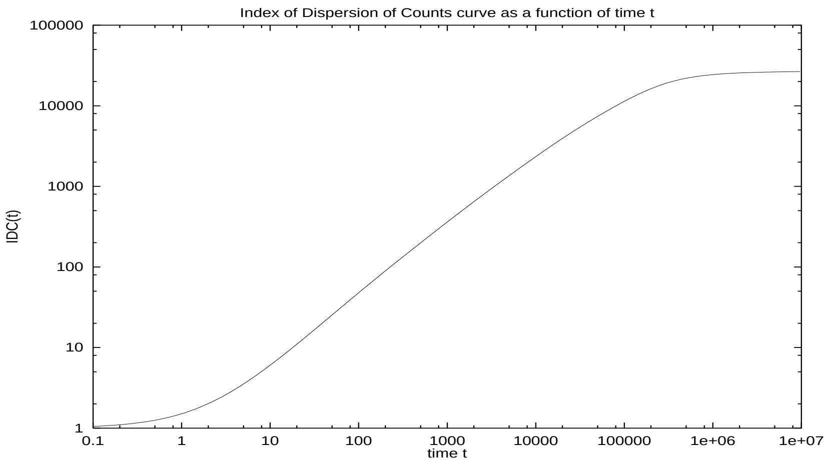

Figure B.18: **The IDC curve for $\mathbf{MUL_{16}}$**

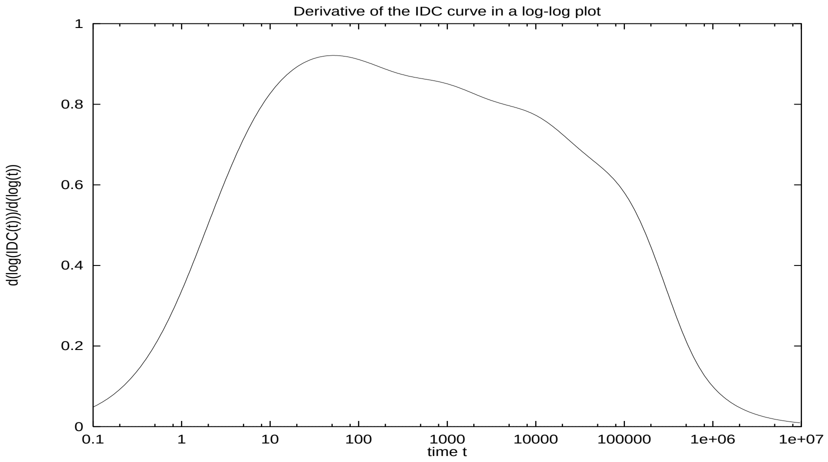

Figure B.19: **The derivative of the IDC curve for $\mathbf{MUL_{16}}$**

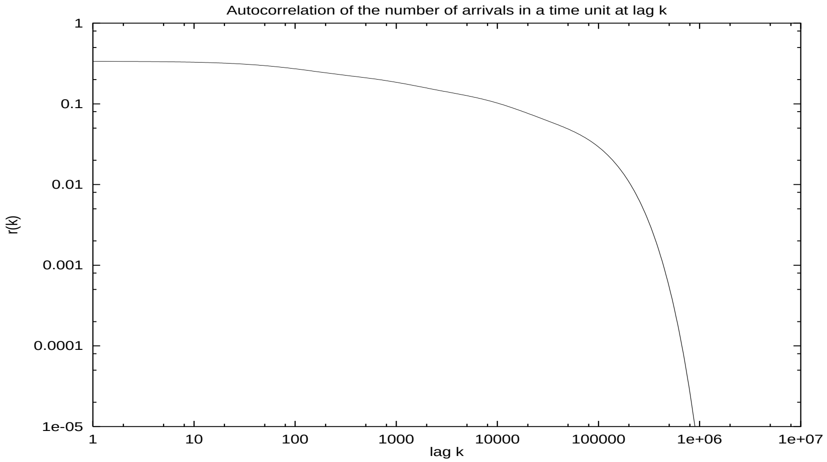

Figure B.20: **The autocorrelation function of the Counting Process for MUL$_{16}$**

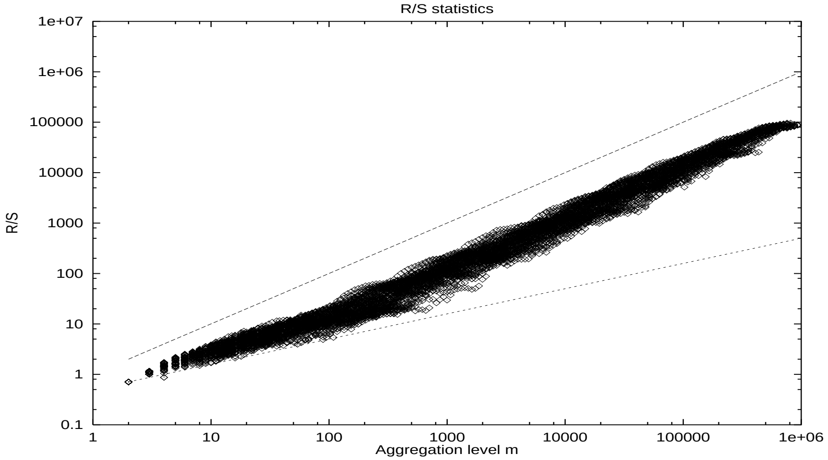

Figure B.21: **R/S statistic for MUL$_{16}$**

Apriori it seems quite surprising that it is possible to create 8-16 state models that display variability over several timescales and at the same time maintain marginal distributions that are not particularly bursty. Recall that for ADD$_8$ C_{var} = 1.36, for ADD$_{16}$ C_{var} = 1.55. and finally for MUL$_{16}$ C_{var} = 1.16. As described in section B.3.2 the 2 models created by the addition method, superpositions of IPPs and a Poisson stream, have hyper exponential marginal distributions (8 phases in ADD$_8$ and 16 phases in ADD$_{16}$). This indicates that the marginal properties of at least the two GMLP models defined by the addition method have marginal properties that are not too different from those observed with a Poisson arrival process in that the frequency functions have approximately the same shape. However, because of the time constants in the system the marginal distribution of the GMLP models of course have heavier tails than the corresponding marginal distribution for a Poisson arrival process.

In this paper our modest goal is to illustrate that it is possible to find MAP models or more precisely GMLP models that can model self-similar behaviour over several timescales. Here we have only presented a few of the GMLP models we have examined, other GMLP models not presented here can be found in [47]. Our approach in defining the GMLP models has been quite heuristic in that we have examined various models defined by the addition and multiplication method. The initial selection of models has mainly been based on how "visually" self-similar they appeared i.e. by examing realizations at different aggregation levels. After the initial selection the analytical properties of the models were examined for example the IDC curve and the derivative of the $log-log$ plotted IDC curve. This approach can also explain why all the models found have a rather high Hurst parameter (around 0.9). We recall from section B.2 that exact second order self-similarity implies that $(X_{m(k-1)+1} + ... + X_{km}) = mX_k^{(m)}$, k = 1,2,3,... and $m^H X$ have the same variance and autocorrelation function. Therefore it is misleading to look at the realizations at different aggregation levels (as we have done !) since the variance and correlation structures are only maintained if H = 1. Clearly it would be better to look at $m^{1-H}X^{(m)}$ since this quantity maintains the same variance and autocorrelation function for a second order self-similar process. In our case of packet traffic with a mean arrival rate ≥ 0 $m^{1-H}X^{(m)}$ does not maintain first order properties. Therefore the expression $m^{1-H}X^{(m)}+(1-m^{1-H})E\{X\}$ seems to be the most interesting since both first and second order properties are maintained for an exact second self-similar process. For H close to 1 observing $m^{1-H}X^{(m)}+(1-m^{1-H})E\{X\}$

corresponds to looking at the aggregate process. In our case the problem with using $m^{1-H}X^{(m)}+(1-m^{1-H})E\{X\}$ instead of $X^{(m)}$ (or actually $mX^{(m)}$) has been that we did not know apriori what Hurst parameter to expect. If the Hurst parameter is known apriori, perhaps approximately, the quantity $m^{1-H}X^{(m)}+(1-m^{1-H})E\{X\}$ definitely seems to be the most relevant.

If the GMLP models are to be used as models for self-similar traffic in practice a more systematic approach is definitely needed. A main advantage of the fractional Brownian model is that if the mean rate, Hurst parameter and peakedness of the traffic [32] is known the model is completely defined. The Hurst parameter determines how the aggregate process smoothes out as the aggregation level increases i.e. it gives the relative amount of smoothing. The peakedness describes how peaked the traffic is on an absolute scale. Clearly both the Hurst parameter and a description of the peakedness of the traffic are needed to give a reasonable description of how variable the traffic is.

In our GMLP models the parameters are not in the same way identifiable physical descriptors. A change in one parameter in our GMLP models will often affect all three physical descriptors i.e. mean rate, Hurst parameter and the peakedness of the traffic. Examining the analytical expressions for the derivative of the $log-log$ plotted IDC curve for example the formulas (B.9) and (B.18) it becomes evident that trying to fit the curve to a fixed Hurst parameter corresponds to fitting a sum of exponentials to a power function. When comparing with the estimated Hurst parameters in the measured traffic and especially the confidence intervals obtained [91] it does not appear to be of essential importance to be able to fix the Hurst parameter exactly over several timescales. It should suffice to keep the Hurst parameter within a reasonably narrow street over several timescales. With this in mind it is tempting to formulate the problem of finding a GMLP model with a certain Hurst parameter as a curve-fitting problem with constraints that assure that the mean rate is fixed and perhaps also constraints that ensure a certain peakedness of the process. (At this point it seems a bit unclear how the peakedness of the GMLP process is most conveniently defined perhaps as some sort of peak mean ratio of the λ's which also takes into account the time-constants in the system.) In order to make the fitting problem easier it might be convenient to fix the underlying time-constants in the model apriori. A good choice seems to be to choose logarithmicly spaced time-constants over the number of time-scales of interest. The parameters in the fitting problem can in this manner be reduced to the λ's. The resulting fitting problem can thus be

defined as for example a least squares fit of the derivative of the IDC curve in a log - log plot to correspond to a certain Hurst parameter at logarithmic spaced points along the time axis with the λ's as parameters . Analytical simplifications of the fitting procedure are possible if restricting the λ's to those that can be defined by the addition method, however, the price paid for this simplification is a great loss of generality in the fitting procedure. How well a fitting scheme as the one just proposed works is at this point quite unclear and therefore a topic for further research. It is also unclear whether some Hurst parameters are easier to model than others as mentioned before all the models we have found with our rather heuristic approach seem to have a relatively high Hurst parameter often around 0.9 on the time-scales of interest . Apriori we suggest to fit to the derivative of the $log-log$ plotted IDC curve rather than the IDC curve itself since this seems to provide more control over the goodness of fit to a certain Hurst parameter than fitting directly to the IDC curve does.

Preliminary results from [6] show that MAPs with the same first and second order properties of the counting process i.e. same rate and variance time curve can yield significantly different queueing behaviour. Clearly this is not possible with a fractional Brownian motion model, since it is fully described by its rate and variance time curve (Hurst parameter and peakedness). This raises the important question of how well the measured traffic is described by an exact second order self-similar model since this might be critical when assessing queueing behaviour.

We have have found that even very simple special cases of the MAP with 8 - 16 states can model self-similar behaviour reasonably well over at least at number of time-scales. In particular we have examined three MAP models with respectively 8, 16 and 16 states that appear visually self-similar over several time-scales and clearly display non-degenerate correlation structures over several time-scales. Two of the examined models are particularly simple since they are merely superpositions of 3 respectively 4 inhomogeneus IPPs and a Poisson source.

These results of very simple cases of the MAP indicate that it would be premature to rule out the MAP as a model for traffic with self-similar characteristics.

Our approach in finding the MAP models with self-similar characteristics has been quite heuristic. It is still an open question how systematicly to find a MAP corresponding to a given rate, Hurst parameter and perhaps burstiness. In section B.5 we have suggested an approach where the parameters of the MAP model are chosen by fitting the derivative of the IDC curve in a log-log plot to imply a certain Hurst parameter under the constraints that the MAP process has a given rate and meausure of burstiness, i.e. the suggested approach is to solve a constrained minimization problem. In general a topic for further research is how systematically to map measures as rate, Hurst parameter and burstiness to a MAP model. More specifically schemes as the one proposed in section B.5 could be investigated.

We have also initiated studies regarding the queueing behavior of MAP models with self-similar characteristics over several time-scales. Preliminary results of these efforts can be found in [6] and [8].

Appendix C

A[3]

An experimental study of queueing performance
of MAPs with
apparently self-similar behaviour

Allan T. Andersen and Bo Friis Nielsen
Institute of Mathematical Modelling
Building 321
Technical University of Denmark
2800 DK-Lyngby

This paper is a slightly modified version of COST 242 TD(95)027 which was presented at the COST 242 MC meeting in Stockholm 10-11/5 1995.

C.1 Introduction

Currently many countries/carriers are on the verge of implementing broadband ISDN (B-ISDN). This raises a number of challenging engineering problems. How should the dimensioning in B-ISDN take place ? Can methods currently used in narrow band ISDN (N-ISDN) be used or do new methods have to be developed ? In other words is it reasonable to expect that the future B-ISDN traffic can be modelled in the same way as the current narrow band traffic.

Recently a number of measurement studies of packet traffic have been initiated, i.e. packet traffic from Local Area Networks (LAN) [35], Variable Bit Rate (VBR) Video [17] and Integrated Services Digital Network (ISDN) [57]. Of these at least the interconnection of LAN's and the transportation of VBR are xpected to be main contributors to traffic in B-ISDN. These studies reveal behaviour very different from what classical traffic models assume. The measurements indicate that these forms of packet traffic have long range dependencies and slowly decaying variances. Characteristics of the observed behaviour have in other contexts been associated with self-similar behaviour. Therefore the measurements motivate the assertion that certain types of packet traffic exhibit self-similar behaviour over several time-scales.

Packet traffic with self-similar behaviour has serious implications for both system engineering and system performance. It is problematic to design/dimension systems only with respect to mean load considerations since such an approach totally fails to take into account the behavioral dynamics associated with self-similarity. The challenge is here to develop models that can describe this adverse behaviour reasonably and are able to quantify new performance measures and dimensioning tools.

This has recently initiated a search for traffic models which are able to describe this behaviour in a relatively simple manner. Here models well known in other contexts have been investigated i.e. Chaotic Maps [31] and Fractional Brownian motions [81].

While arrival processes based on such models can be described using only a few parameters it is analytically very difficult to analyze a queueing system with these processes as input.

In an earlier contribution [7] we have used special cases of the MAP namely the Markov Modulated Poisson Process (MMPP) to model this adverse behaviour.

It was found that a relatively simple MAP - a superposition of four appropriately parameterized Interrupted Poisson Processes (IPP) and a Poisson Stream, i.e. a superposition of 5 renewal streams, was able to model variability, if not self-similar behaviour, over several time-scales. Recently in [47] 8-16 state MMPP's were found which seem at least as "visually" self - similar over several time scales as the packet traffic data presented in [61].

In this contribution we examine the queueing behaviour of some of the MAP models from [7] and [47] with variability/self similar behaviour over several time scales. Here we have implemented some of the algorithms for the single server queue with a Markovian Arrival Process (MAP) outlined by D.M. Lucantoni in [64]. For a detailed analysis of the MAP/G/1 queue see the pioneering work by V. Ramaswami in [83]

Queueing analysis of the LAN traffic traces has been initiated in [32] where the queueing behaviour of a queue driven with LAN traffic traces has been examined. The main findings were that queueing behaviour seemed relatively insensitive to short range correlations, that is correlations on the smallest time scales, but very sensitive to long range correlations, i.e correlations on large time scales.

In this contribution we have also tried (where possible !) to carry out the same sort of queueing experiments done in [32] to determine if the queueing behaviour of our MAP models coincide with the behaviour of real traffic traces.

This paper is organized as follows. Section 2 contains a brief description of self-similarity. In Section 3 a brief introduction to the MAP and the MAP/G/1 queueing system is given. Section 4 contains some queueing experiments that can be utilized to examine which characteristics of the arrival process seem to be most important with regards to queueing behaviour. Section 5 describes aspects of the queueing behaviour of our MAP models and compares them with those found using LAN traffic traces. In Section 6 we conclude and give directions for further work.

C.2 Measures of Self-Similarity

The short introduction to self-similarity given here is primarily based on [61] and [91].

A continuous time stochastic process $X(t)$ with mean zero and stationary increments is said to be self-similar with self-similarity parameter H if $X(at)$ has the same finite dimensional distributions as $a^H X(t)$. H denotes the self-similarity parameter also called the Hurst parameter.

In the context of packet traffic we are concerned with the process defined by the number of packets arriving in a time-interval. This makes it more natural to consider the corresponding stationary discrete time stochastic process $X = X_t$ for t = 1,2,.... (Number of packets arriving in time-interval 1,2 ....).

We also consider the time-series obtained by averaging the discrete time process over non-overlapping blocks. That is, for m = 1,2,3 .., we consider the time-series

$X_k^{(m)} = (1/m)(X_{m(k-1)+1} + \ldots + X_{mk})$, k = 1,2,3,....

The stationary discrete time stochastic process X with mean 0 and variance σ^2 is said to be self-similar with self-similarity parameter H if for all m = 1,2,3.... $(X_{m(k-1)+1} + \ldots + X_{mk}) = mX_k^{(m)}$, k = 1,2,3,... has the same finite dimensional distributions as $m^H X$.

If the variance and autocorrelation function of $(X_{m(k-1)+1} + \ldots + X_{mk}) = mX_k^{(m)}$, k = 1,2,3,... and $m^H X$ are the same for m = 1,2,3,.... X is said to be second-order self-similar with self-similarity parameter H. This implies that for all m = 1,2,3 $var(X^{(m)}) = \sigma^2 m^{-\beta}$, and that $r^{(m)}(k) = r(k) = \delta^2(|k|^{2-\beta})$ (k = 1,2,3...) where $\beta = 2 - 2H$ and $\delta^2(f)$ denotes the second central difference operator applied to a function, that is, $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$.

X is called asymptotically second-order self-similar with self-similarity parameter $H = 1 - \beta/2$ if $r^{(m)}(k) \to \delta^2(|k|^{2-\beta})$, as $m \to \infty$ (k = 1,2,3,.....)

The definition of long-range dependence is that $\sum_k r(k) \to \infty$. For a self-similar process this corresponds to the self-similarity parameter H in the range $0.5 < H < 1.0$. For such a process the averaged time-series $X^{(m)}$ possesses a non-degenerate correlation structure as $m \to \infty$. For all conventional tale-traffic models $r_m(k) \to 0$ as $m \to \infty$, (k=1,2,3,...), that is, the averaged process $X^{(m)}$ tends to second order pure noise for large m corresponding to H = 0.5 or equivalently β = 1. In the following a process with H = 0.5 will be called a short-range dependent process.

In practice there are several ways to try to detect whether an observed process exhibits self-similar behaviour or not. A simple way to check for any indications of self-similarity is to plot the averaged process $X^{(m)}$ for different values of m, m values differing by an order of magnitude. If the plots "look the same" it is an indication of a self-similar behaviour.

In the context of packet traffic attention has been directed towards three ways of checking for self-similarity which all examine second order properties. These can be referred to as methods based on the variance-time plot or equivalently the index of dispersion of counts, the pox diagram of the R/S statistic (also called the rescaled adjusted range plot) and finally analysis based on the periodogram (in the frequency-domain). Further discussion of these three techniques for checking for self-similarity is outside the scope of this work, we refer to [61] and [91].

In our search for MAP models with self-similar behaviour we have primarily used visual self-similarity as a yardstick to determine whether a given MAP seemed reasonable as a model for self-similar behaviour over several time scales rather than using some sort of stringent control of the Hurst parameter over several time scales using for example the second order methods mentioned above. However, plots of the variance time curve or equivalently the IDC, (index of dispersion of counts), curve of the interesting MAP models found in [7] and [47] clearly show that over the observed time-scales an estimated value of H, $\hat{H}$, satisfies $0.5 < \hat{H} \leq 1$ so that the correlation structure over the observed time scales does not degenerate to a Hurst estimate of $\hat{H}$= 0.5.

As stated in [7] our MAP models will eventually turn out to be short range dependent in the above described sense. That is, once you observe a given MAP at a time scale larger than the largest time constant in the MAP the correlation structure will yield a Hurst estimate of $\hat{H}$ = 0.5. However, examining the process on smaller time scales yields Hurst estimates in the range 0.5 - 1.0 which the MAP's in [7] and [47] clearly demonstrate.

C.3 The MAP and the MAP/G/1 queue

In this section we will briefly sketch some main characteristics of the Markovian Arrival Process (MAP). For further results we refer to [64], [77] and [83]. The description of the MAP follows [65] closely. Also we will briefly summarize the Matrix Analytic approach used when solving the MAP/G/1 queue. For an in depth discussion and derivation of formulas we refer to [83] and for the latest developments to [64].

Consider a Markov process on the state space $\{1,2,..,m+1\}$, where $\{1,2,..,m\}$ are transient states and $\{m+1\}$ is absorbing. Absorption from any state is certain. Entering the absorbing state corresponds to an arrival in the arrival process. The Markov process is then instantaneously restarted in a transient state, where the selection of the new state is allowed to depend on the state from which absorption occurred.

More formally the evolution of the arrival process can be defined as follows. Assume the Markov process is in a transient state i, $1 \leq i \leq m$. The sojourn time in this state is exponentially distributed with parameter λ_i. When the sojourn time has elapsed, there are two possibilities. With probability $p_{ij}, 1 \leq i \leq m$, the Markov process enters the absorbing state and is instantaneously restarted in the transient state j. With probability $q_{ij}, 1 \leq i \leq m, j \neq i$, the process immediately enters the transient state j.

Here the following equalities holds for $1 \leq i \leq m$

$$\sum_{\substack{j=1 \\ j \neq i}}^{m} q_{ij} + \sum_{j=1}^{m} p_{ij} = 1 \tag{C.1}$$

An equivalent description is that if we for each i, $1 \leq i \leq m$, define $D_{ij} = \lambda_i p_{ij}$, $1 \leq j \leq m$, $C_{ij} = \lambda_i q_{ij}$, $1 \leq j \leq m$, $i \neq j$ and $C_{ii} = -\lambda_i$, then the elementary probability of an arrival in an infinitesimal interval length dt which leaves the Markov process in state j, given that the Markov process is in state i, is $D_{ij}dt$. Similarly, the elementary probability that the process enters the transient state j (without arrival) in an interval of length dt, given that it is in state i is $C_{ij}dt, i \neq j$. The term 'arrival phase' will in the following be used to denote the state of the underlying Markov process.

The MAP is a Semi-Markovian arrival process whose transition probability matrix $\mathbf{F}(*)$ is of the form

$$\begin{aligned}\mathbf{F}(x) &= \textstyle\int_0^x e^{\mathbf{C}u} du \mathbf{D} \\ &= (\mathbf{I} - e^{\mathbf{C}x})(-\mathbf{C}^{-1})\mathbf{D},\end{aligned} \tag{C.2}$$

where the matrices $\mathbf{C} = [C_{ij}]$ and $\mathbf{D} = [D_{ij}]$ are respectively a stable matrix and a non-negative matrix whose sum is an irreducible infinitesimal generator. (The matrix exponential is defined by its series i.e. $e^{\mathbf{M}} = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}$, for the properties of the matrix exponential see [16]).

The infinitesimal generator $\mathbf{Q} = \mathbf{C} + \mathbf{D}$ describes the transitions among the transient states $i, ..., m$ in the underlying Markov process. The stationary vector $\vec{\pi}$ of this Markov process satisfies the equations

$$\vec{\pi}(\mathbf{C}+\mathbf{D}) = \vec{\pi}\mathbf{Q} = \vec{0}, \quad \vec{\pi}\vec{e} = 1 \tag{C.3}$$

where $\vec{e}$ is, as it is throughout this paper, a column vector of 1's.

The fundamental mean of the transition probability matrix $\mathbf{F}(*)$ is given by $\lambda_a = (\vec{\pi}\mathbf{D}\vec{e})^{-1}$, so that the quantity $(\lambda_a)^{-1}$ is the fundamental rate of the point process i.e. the rate of the stationary point process.

Some well-known processes are (very) special cases of the MAP

- For $\mathbf{C} = (-\lambda)$ and $\mathbf{D} = (\lambda)$ the MAP is the ordinary Poisson process of rate λ
- The phase-type (PH) renewal process is also a special case of the MAP. It was introduced by Neats and he discusses it in detail in [75]. It includes both the Erlang and hyper-exponential (equivalent with the Interrupted Poisson Process IPP [59]) and Poisson arrival processes as special cases. It is defined in an analogous manner to the MAP except that at an arrival, the new state of the Markov process is chosen according to a probability vector $\vec{\alpha}$ which is independent of the state from which absorption occurred (It is a renewal process). This can more formally be stated as $\mathbf{D} = \vec{T^0}\vec{\alpha}$, where $\vec{T^0}$ is a column vector with components $T_i^0, 1 \leq j \leq m$, where T_i^0 is the rate into the absorbing state when in state i. To distinguish the PH-renewal process notationally from the general MAP, we set $\mathbf{C} = \mathbf{T}$.

- The Markov-modulated Poisson process (MMPP) is also a special case of the MAP. The MMPP is the doubly stochastic Poisson process whose arrival rate depends on an m-state irreducible Markov process. The arrival rate therefore takes on only m values $\lambda_1, ..., \lambda_m$ and is equal to λ_j whenever the Markov process is in the state j. If the underlying Markov process has infinitesimal generator $\mathbf{Q}$ and if $\Lambda = diag(\lambda_1, ..., \lambda_m)$, then we have

$$\begin{aligned} \mathbf{C} &= \mathbf{Q} - \Lambda \\ \mathbf{D} &= \Lambda \end{aligned} \tag{C.4}$$

A useful property of the MAP is that it is closed under superposition.

The counting function of the MAP is given in the following way Let N_t denote the number of arrivals in $(0, t]$ and J_t the state of the Markov Process at time t.

Then the (i, j) entry of the mxm matrices $\mathbf{P}(n, t)$ are given as

$$\mathbf{P}_{ij}(n, t) = P\{N_t = n, J_t = j | N_0 = 0, J_0 = i\} \tag{C.5}$$

The matrices $\mathbf{P}(n, t)$ satisfy the forward Chapman-Kolmogorov equations

$$\begin{aligned} \frac{d}{dt}\mathbf{P}(n, t) &= \mathbf{P}(n, t)\mathbf{C} + \mathbf{P}(n-1, t)\mathbf{D}, n \geq 1, t \geq 0 \\ \mathbf{P}(0, 0) &= \mathbf{I} \end{aligned} \tag{C.6}$$

and the matrix generating function $\mathbf{P}^*(z, t) = \sum_{n=0}^{\infty} \mathbf{P}(n, t) z^n$ is explicitly given by

$$\mathbf{P}^*(z, t) = e^{(\mathbf{C}+z\mathbf{D})t} \quad |z| \leq 1, t \geq 0 \tag{C.7}$$

By differentiating the matrix generating function with respect to z it is possible to derive moments of the number of arrivals N_t at a given time. In particular the

mean and variance of N_t, for a given start probability vector of the underlying Markov process, can be found from the first two derivatives. The variance of N_t is often called the variance time curve. Here we will only consider the mean and variance of N_t in the stationary version of the MAP. Both these quantities are listed below. For further details regarding the detailed derivation of these results we refer to [77]. We recall that the index of dispersion of counts (IDC) curve is defined as the ratio of $Var\{N_t\}$ to $E\{N_t\}$.

For the stationary version of the MAP the mean and variance of the number of arrivals from $(0, t]$ are given by

$$E\{N_t\} = \vec{\pi}\mathbf{D}\vec{e}t \tag{C.8}$$

$$\begin{aligned} Var\{N_t\} = & (1 - 2 * \vec{\pi}\mathbf{D}\vec{e})E\{N_t\} + \\ & 2\vec{\pi}\mathbf{D}(\vec{e}\vec{\pi} - \mathbf{Q})^{-1}\mathbf{D}\vec{e}t - \\ & 2\vec{\pi}\mathbf{D}(\mathbf{I} - e^{\mathbf{Q}t})(\vec{e}\vec{\pi} - \mathbf{Q})^{-2}\mathbf{D}\vec{e} \end{aligned} \tag{C.9}$$

(The result for $E\{N_t\}$ is also readily seen from the previously defined fundamental rate of the MAP)

For a MAP model the marginal distribution between arrivals is readily seen to be of phase-type. The probability distribution function for the marginal time t between arrivals is given as

$$P\{T < t\} = F(t) = 1 - (\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}e^{\mathbf{C}t}\vec{e} \tag{C.10}$$

and the corresponding frequency function is

$$f(t) = (\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}e^{\mathbf{C}t}\mathbf{D}\vec{e} \tag{C.11}$$

The moments of the marginal distribution are given as

$$m_i = i!(\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}(\mathbf{C})^{-i}\vec{e} \tag{C.12}$$

In the following we sketch some of the main results regarding the MAP/G/1 queue primarily from [64] and to a certain extent from [83].

The arbitrary service time distribution is denoted by $H(t)$ with finite mean μ_1. We assume subsequent service times to be independent and that the arrival process and service time distributions are independent. Of course the queue is also assumed to be stable i.e. the load $\rho = \frac{\mu_1}{\lambda_a} < 1$.

As standard when solving the M/G/1 queue the approach is here to study the embedded Markov renewal process at departure epochs. Define τ_k to be the epoch of the k-th departure from the queue, with $\tau_0 = 0$, and (ξ_k, J_k) to be respectively the number in system and phase of the arrival process at ${\tau_k}^+$. Then $(\xi_k, J_k, \tau_{k+1} - \tau_k)$ is a semi-Markov process on the state space $\{(i,j) : i \geq 0, 1 \leq j \leq m\}$

The Semi Markov Process is positive recurrent when the traffic intensity $\rho < 1$. The transition probability matrix is given by

$$\mathbf{P}(x) = \begin{bmatrix} \mathbf{B}_0(x) & \mathbf{B}_1(x) & \mathbf{B}_2(x) & \dots & \\ \mathbf{A}_0(x) & \mathbf{A}_1(x) & \mathbf{A}_2(x) & \dots & \\ \mathbf{0} & \mathbf{A}_0(x) & \mathbf{A}_1(x) & \dots & \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_0(x) & \dots & \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}, x \geq 0, \tag{C.13}$$

where for $n \geq 0$, $\mathbf{A}_n(x)$ and $\mathbf{B}_n(x)$ are mxm matrices of mass functions defined by

$[\mathbf{A}_n(x)]_{ij}$ is the probability that given a departure at time 0, which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during that service there were n arrivals.

$[\mathbf{B}_n(x)]_{ij}$ is the probability that given a departure at time 0, which left the system empty and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, leaving n customers in the system.

From the earlier definition of $\mathbf{P}(n,t)$ and the MAP as a Semi-Markovian arrival process i.e. (C.2) and (C.5) it is clear that

$$\mathbf{A}_n(x) = \int_{t=0}^{x} \mathbf{P}(n,t)dH(t) \tag{C.14}$$

and that

$$\mathbf{B}_n(x) = \int_{t=0}^{x} e^{\mathbf{C}t}\mathbf{D}\mathbf{A}_n(x-t)dt \tag{C.15}$$

Also we define $\mathbf{A}_n(\infty) = \mathbf{A}_n$ and $\mathbf{B}_n(\infty) = \mathbf{B}_n = (-\mathbf{C}^{-1})\mathbf{D}\ \mathbf{A}_n$

We define the following transform matrices :

$$\tilde{\mathbf{A}}(z) = \sum_{n=0}^{\infty} \mathbf{A}_n z^n \tag{C.16}$$

$$\tilde{\mathbf{B}}(z) = \sum_{n=0}^{\infty} \mathbf{B}_n z^n = (-\mathbf{C}^{-1})\mathbf{D}\tilde{\mathbf{A}}(z) \tag{C.17}$$

The matrix $\mathbf{A} = \tilde{\mathbf{A}}(1)$ is a stochastic matrix and the stationary vector $\vec{\pi}$ defined in (C.3) also satisfies $\vec{\pi}\mathbf{A} = \vec{\pi}$, $\vec{\pi}\vec{e} = 1$. $\tilde{\mathbf{A}}^{(i)}$ denotes the i-th factorial moment of $\tilde{\mathbf{A}}(z)$. The vector $\vec{\beta}$, whose j-th component is the conditional number of arrivals during a service which starts with the arrival process in phase j is given by

$$\begin{aligned} \vec{\beta} &= (\tfrac{d}{dz}\tilde{\mathbf{A}}(z))_{(z=1)}\vec{e} \\ &= (\tfrac{\mu_1}{\lambda_a})\vec{e} + (\mathbf{A} - \mathbf{I})(\vec{e}\vec{\pi} + \mathbf{Q})^{-1}\mathbf{D}\vec{e} \end{aligned} \tag{C.18}$$

Additionally the traffic intensity ρ is given by $\rho = \vec{\pi}\vec{\beta}$.

It is clear from (C.13) that the stationary probability vector $\vec{x} = [\vec{x}_0, \vec{x}_1, \dots]$, where $\vec{x}_k = [x_{k1}, x_{k2} \dots x_{km}]$, of $\mathbf{P} = \mathbf{P}(\infty)$ can be expressed as in (C.19)

$$\vec{x}_i = \vec{x}_0\mathbf{B}_i + \sum_{l=1}^{i+1} \vec{x}_l\mathbf{A}_{i+1-l}, i \geq 0 \tag{C.19}$$

Setting $\tilde{\vec{X}}(z) = \sum_{i=0}^{\infty} \vec{x}_i z^i$ it can be seen from (C.19) that

$$\tilde{\vec{X}}(z)\left[\mathbf{I} - \tilde{\mathbf{A}}(z)\right] = \vec{x}_0(-\mathbf{C}^{-1})(\mathbf{C} + z\mathbf{D})\tilde{\mathbf{A}}(z) \tag{C.20}$$

so that the generating function is completely determined by $\vec{x}_0$.

$\vec{x}_0$ can be derived from the first passage time distribution for successive returns to the set of states $\{(0,1), ...(0,m)\}$. Here it is convenient to define level i as the set of states $\{(i,1), ...(i,m)\}$. From the structure of $\mathbf{P}$ it is clear that to reach level 0 from level i, $i \geq 1$ each level in between must be visited i.e. the matrix $\mathbf{P}$ is skip-free to the left. Also it is clear from $\mathbf{P}$ that the mechanism governing the first passage time from level $i+1$ to level i is the same for all levels $i \geq 0$.

That is, the first passage time distribution from level $i+1$ to level i, $i \geq 0$ plays a very important role when examining the return time distribution of the level 0.

Define the matrix $\mathbf{G}$(k) with elements $\mathbf{G}_{j\acute{j}}(k)$, $k \geq 1$ as the probability that the first passage from the state $(i+1, j)$ to the state $(i, \acute{j})$, $i \geq 0$,$1 \leq j, \acute{j} \leq m$ occurs in exactly k transitions and that $(i, \acute{j})$ is the first state visited on level i.

By a first passage argument it can be shown [73] that the transform matrix $\tilde{\mathbf{G}}(z)$ defined by

$$\tilde{\mathbf{G}}(z) = \sum_{i=1}^{\infty} \mathbf{G}(i)z^i, |z| \leq 1 \tag{C.21}$$

satisfies

$$\tilde{\mathbf{G}}(z) = z\sum_{l=0}^{\infty} \mathbf{A}_l\tilde{\mathbf{G}}^l(z) \tag{C.22}$$

The matrix $\mathbf{G} = \tilde{\mathbf{G}}(1)$ is stochastic when $\rho \leq 1$. For $\rho < 1$ the invariant probability vector $\vec{g}$ of the positive stochastic matrix $\mathbf{G}$, satisfies

$$\vec{g}\mathbf{G} = \vec{g}, \quad \vec{g}\vec{e} = 1 \tag{C.23}$$

The vector $\vec{\mu}$ is defined by

$$\begin{aligned}\vec{\mu} &= (\tfrac{d}{dz}\tilde{\mathbf{G}}(z))_{(z=1)}\vec{e} \\ &= (\mathbf{I} - \mathbf{G} + \vec{e}\vec{g})(\mathbf{I} - \mathbf{A} - (\vec{e} - \vec{\beta})\vec{g})^{-1}\vec{e}\end{aligned} \tag{C.24}$$

and its j-th component is the expected number of transitions (i.e services) from a state $(i+1, j)$ to level i.

The equality $\vec{g}\vec{\mu} = (1-\rho)^{-1}$ holds.

$\vec{x}_0$ can be found by analyzing the recurrence time of the states x_{0j} $(1 \leq j \leq m)$, more specificly by analyzing the matrix

$$\tilde{\mathbf{K}}(z) = z\sum_{l=0}^{\infty} \mathbf{B}_l\tilde{\mathbf{G}}^l(z) \tag{C.25}$$

Using Markov renewal theory [24] it can be shown that

$$\vec{x}_0 = \lambda_a(1-\rho)\vec{g}(-\mathbf{C}) \tag{C.26}$$

See [64],[65], [77], [83] for details.

Defining $\tilde{\vec{U}}(z) = \vec{x}_0(-\mathbf{C}^{-1})(\mathbf{C} + z\mathbf{D})\tilde{\mathbf{A}}(z)$, the first two factorial moments of $\vec{x}$ can be expressed in the following manner (see [77] or [83] for details regarding the derivations)

The first factorial moment is given by

$$
\begin{aligned}
\tilde{\vec{X}}^{(1)}\vec{e} =& \tfrac{1}{2(1-\rho)}(\tilde{\vec{X}}^{(0)}\tilde{\mathbf{A}}^{(2)}\vec{e} + \tilde{\vec{U}}^{(2)}\vec{e} \\
&+2(\tilde{\vec{U}}^{(1)} - \tilde{\vec{X}}^{(0)}(\mathbf{I} - \tilde{\mathbf{A}}^{(1)}))(\vec{e}\vec{\pi} + \mathbf{I} - \mathbf{A})^{-1}\vec{\beta}) &\text{(C.27)}\\
\tilde{\vec{X}}^{(1)} =& (\tilde{\vec{X}}^{(1)}\vec{e})\vec{\pi} + (\tilde{\vec{U}}^{(1)} - \tilde{\vec{X}}^{(0)}(\mathbf{I} - \tilde{\mathbf{A}}^{(1)}))(\vec{e}\vec{\pi} + \mathbf{I} - \mathbf{A})^{-1} &\text{(C.28)}
\end{aligned}
$$

And the second factorial moment by

$$
\begin{aligned}
\tilde{\vec{X}}^{(2)}\vec{e} =& \tfrac{1}{3(1-\rho)}(3\tilde{\vec{X}}^{(1)}\tilde{\mathbf{A}}^{(2)}\vec{(e)} + \tilde{\vec{X}}^{(0)}\tilde{\mathbf{A}}^{(3)}\vec{e} + \tilde{\vec{U}}^{(3)}\vec{e} \\
&+3(\tilde{\vec{U}}^{(2)} + \tilde{\vec{X}}^{(0)}\tilde{\mathbf{A}}^{(2)} - 2\tilde{\vec{X}}^{(1)}(\mathbf{I} - \tilde{\mathbf{A}}^{(1)}))(\vec{e}\vec{\pi} + \mathbf{I} - \mathbf{A})^{-1}\vec{\beta}) &\text{(C.29)}\\
\tilde{\vec{X}}^{(2)} =& (\tilde{\vec{X}}^{2}\vec{e})\vec{\pi} + (\tilde{\vec{U}}^{(2)} + \tilde{\vec{X}}^{(0)}\tilde{\mathbf{A}}^{(2)} - 2\tilde{\vec{X}}^{(1)}(\mathbf{I} - \tilde{\mathbf{A}}^{(1)}))(\vec{e}\vec{\pi} + \mathbf{I} - \mathbf{A})^{-1} &\text{(C.30)}
\end{aligned}
$$

Using the Key Renewal Theorem for Markov renewal processes [24] it is possible to express the stationary queue length distribution at an arbitrary time in terms of the stationary queue length distribution at departures.

Let ξ_t denote the queue length and J_t denote the arrival process at time t. The time dependent distribution of the joint distribution of queue length and arrival phase is given by the conditional probabilities

$$y(k,j)_t = P\{\xi_t = k, J_t = j | \xi_0 = k_0, J_0 = j_0\} \tag{C.31}$$

for $k \geq 0, 1 \leq j \leq m, t \geq 0$. It can be shown that the limits $y(k,j)_t$, $k \geq 0, 1 \leq j \leq m$, for $t \to \infty$ exist. We denote these limits y_{kj} and define $\vec{y}_k = [y_{k1}, y_{k2} \ldots y_{km}]$.

Using the Key Renewal Theorem for Markov renewal processes [24] it can be shown [83] that the $\vec{y}_0$ is given by

$$\vec{y}_0 = {\lambda_a}^{-1}\vec{x}_0(-\mathbf{C}^{-1}) \tag{C.32}$$

As expected $\vec{y}_0\vec{e} = 1 - \rho$. In [83] it is shown that the generating function $\tilde{\vec{Y}}(z) = \sum_{i=0}^{\infty} \vec{y}_i z^i$ is related to the generating function $\tilde{\vec{X}}(z)$ by

$$
\begin{aligned}
\tilde{\vec{Y}}(z)(\mathbf{C} + z\mathbf{D}) =& {\lambda_a}^{-1}(z-1)\tilde{\vec{X}}(z), |z < 1| &\text{(C.33)}\\
\tilde{\vec{Y}}(1) =& \vec{\pi} &\text{(C.34)}
\end{aligned}
$$

From this it is clear that the vectors $\vec{y}_i$ can be determined as

$$\vec{y}_{i+1} = \left[\vec{y}_i\mathbf{D} - {\lambda_a}^{-1}(\vec{x}_i - \vec{x}_{i+1})\right](-\mathbf{C}^{-1}), i \geq 0 \tag{C.35}$$

Essentially by differentiating in (C.34) and using the generalized inverse $(\vec{e}\vec{\pi} + \mathbf{Q})^{-1}$ of $\mathbf{Q}$ it is possible to derive the factorial moment vectors $\tilde{\vec{Y}}^{(1)}$ and $\tilde{\vec{Y}}^{(2)}$. For details we refer to either [64] or [83].

The first factorial moment is given by

$$
\begin{aligned}
\tilde{\vec{Y}}^{(1)}\vec{e} =& \tilde{\vec{X}}^{(1)}\vec{e} + (\lambda_a\vec{\pi}\mathbf{D} - \tilde{\vec{X}}^{(0)})(\vec{e}\vec{\pi} + \mathbf{Q})^{-1}\mathbf{D}\vec{e} &\text{(C.36)}\\
\tilde{\vec{Y}}^{(1)} =& (\tilde{\vec{Y}}^{(1)}\vec{e})\vec{\pi} + (\lambda_a^{-1}\tilde{\vec{X}}^{(0)} - \vec{\pi}\mathbf{D})(\vec{e}\vec{\pi} + \mathbf{Q})^{-1} &\text{(C.37)}
\end{aligned}
$$

And the second factorial moment by

$$
\begin{aligned}
\tilde{\vec{Y}}^{(2)}\vec{e} =& \tilde{\vec{X}}^{(2)}\vec{e} + 2(\lambda_a\tilde{\vec{Y}}^{(1)}\mathbf{D} - \tilde{\vec{X}}^{(1)})(\vec{e}\vec{\pi} + \mathbf{Q})^{-1}\mathbf{D}\vec{e} &\text{(C.38)}\\
\tilde{\vec{Y}}^{(2)} =& (\tilde{\vec{Y}}^{(2)}\vec{e})\vec{\pi} + 2({\lambda_a}^{-1}\tilde{\vec{X}}^{(1)} - \tilde{\vec{Y}}^{(1)}\mathbf{D})(\vec{e}\vec{\pi} + \mathbf{Q})^{-1} &\text{(C.39)}
\end{aligned}
$$

From the queue length distribution at an arbitrary time it is readily seen that the queue length distribution including phase information just before an arbitrary arrival epoch is given by (C.40).

$$\vec{z_i} = \lambda_a \vec{y_i}\mathbf{D}, i \geq 0 \tag{C.40}$$

The vector $\vec{z} = [\vec{z_0}, \vec{z_1}, \vec{z_2} \ldots]$ denotes the queue length distribution including phase information just before an arbitrary arrival epoch. The factorial moments are given in the same manner i.e. :

$$\tilde{\vec{Z}}^{(1)} = \lambda_a \tilde{\vec{Y}}^{(1)}\mathbf{D} \tag{C.41}$$

$$\tilde{\vec{Z}}^{(2)} = \lambda_a \tilde{\vec{Y}}^{(2)}\mathbf{D} \tag{C.42}$$

For queues without batch arrivals the queuelength distribution seen by an arrival is the same as the queuelength distribution seen by a departure so the previous calculations regarding the queuelength at an arrival are only necessary if detailed phase information is needed.

C.3.1 Notes on implementation details of the MAP/G/1 queue

Our implementation of the MAP/G/1 queue uses a mixture of deterministic services times as the service time distribution. In the context of packet traffic this seems very reasonable since the packet length distribution can be modelled accurately. This may prove very relevant when modelling ATM traffic internally in a switch.

This choice of service time distribution makes it particularly simple to calculate the matrices $\mathbf{A}_n$ using “randomization” so that numerical integration is avoided. Randomization was introduced in [48]. For details regarding the method in the context of matrix-analytic methods we refer to [64] and [66]. Because $\mathbf{B}_n = (-\mathbf{C}^{-1})\mathbf{D}\ \mathbf{A}_n$ it is trivial to calculate $\mathbf{B}_n$ once $\mathbf{A}_n$ is known. Randomization (or uniformization) is also used to calculate the factorial moment matrices $\tilde{\mathbf{A}}^{(i)}$, again we refer to [64] for details.

The $\mathbf{G}$ matrix can be calculated by using the recursion implied by (C.22).

$$\mathbf{G}_{(i+1)} = \sum_{l=0}^{\infty} \mathbf{A}_l \mathbf{G}_{(i)}^l \tag{C.43}$$

When started with $\mathbf{G}_{(0)} = \mathbf{0}$ (C.43) produces a sequence of nonnegative matrices which increase monotonically to a unique solution. For a discussion regarding various schemes for finding $\mathbf{G}$ we refer to [64],[77] and [84] . We have adopted the scheme proposed in [64].

It is strongly recommended NOT to use the equations (C.19) directly when calculating the steady state probability vector $\vec{x}$ since this scheme suffers from cancellation (i.e. subtracting elements of the same magnitude yielding numerical problems). Instead the following scheme due to V. Ramaswami [85], which does not suffer from cancellation problems, is recommended. Here we only list the algorithm.

Given the vector $\vec{x}_0$, the vectors $\vec{x}_i$, $i \geq 1$ can be obtained from the formula (C.44)

$$\vec{x}_i = (\vec{x}_0\bar{\mathbf{B}}_i + \sum_{l=1}^{i-1} \vec{x}_l \bar{\mathbf{A}}_{i+1-l})(1 - \bar{\mathbf{A}}_1)^{-1}, i \geq 1 \tag{C.44}$$

where $\bar{\mathbf{B}}_l = \sum_{i=l}^{\infty} \mathbf{B}_i \mathbf{G}^{i-l}$, and $\bar{\mathbf{A}}_l = \sum_{i=l}^{\infty} \mathbf{A}_i \mathbf{G}^{i-l}$, $l \geq 0$ The matrices $\bar{\mathbf{A}}_l$ and $\bar{\mathbf{B}}_l$ can calculated using a backward recursion scheme. The recursion is started at a value l_{max} defined by the matrices $\sum_{i=l_{max}+1}^{\infty} \mathbf{A}_i$ and $\sum_{i=l_{max}+1}^{\infty} \mathbf{B}_i$ having negligible small elements. In our implementation we have adopted this scheme.

C.4 Queueing experiments

Recently, as briefly mentioned in the introduction, there has been initiated [32] experimental studies regarding the queueing behaviour of LAN traffic using trace driven simulations.

In [32], where LAN traffic traces are considered, one of the approaches used to examine the effects correlations have on queueing behaviour is to carry out various "shuffling" experiments. This is done by considering the inter-arrival times of packets in the data. A given dataset is partitioned in consecutive blocks, each of these blocks containing the same number of inter-arrival times, say M.

Shuffling the inter-arrival times within each of these blocks destroys the correlations on the smallest time scales while maintaining the correlations on larger time scales i.e. between neighboring blocks. This is called internal shuffling. Shuffling the different blocks of the dataset maintains some of the correlations on the smallest time scales while those on larger time scales are destroyed. This is called external shuffling. It is readily seen that performing external shuffling with $M = 1$ corresponds to examining the marginal distribution of the dataset. Equivalently this can be done by performing an internal shuffling where M is the size of the dataset.

For different utilizations the queueing behaviour (for example mean queue length) for different shufflings can be compared to each other and to the original trace.

In principle it is possible to carry out the same sort of shuffling experiments analytically using a MAP to model the arrival process. This is due to the nice closure properties that the MAP possesses.

It is readily seen that external shuffling for MAPs can be expressed in terms of the original process as in (C.45). The resulting MAP, which we denote by $(\mathbf{C}_{es}, \mathbf{D}_{es})$ has a state space of the size Mm. Clearly, the special case of M=1 corresponds to the marginal distribution of the original MAP.

$$\mathbf{C}_{es} = \begin{bmatrix} \mathbf{C} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{C} \end{bmatrix}$$

$$\mathbf{D}_{es} = \begin{bmatrix} \mathbf{0} & \mathbf{D} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{D} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{D} \\ \frac{(\mathbf{D}\vec{e})(\vec{\pi}\mathbf{D})}{(\vec{\pi}\mathbf{D}\vec{e})} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \end{bmatrix} \qquad \text{(C.45)}$$

The internal shuffling is much more complicated to express with a MAP. In earlier a version of this paper [8] we found a quite complicated expression for $M = 2$ which turned out only to be an approximation. The only feasible way

to examine internally shuffled MAPs seems to be simulation i.e. performing internal shuffling on artificially created MAP traffic traces.

From a MAP modelling point of view it is also interesting here to note the clear relationship between internal shuffling and so-called "Local Poissonification" [79], which considers properties of the process obtained be locally rearranging arrivals (from a MAP) according to a Poisson distribution.

C.5 Queueing experiments with MAPs

In the following we will examine the queueing behaviour of three MAP's with variability over several time scales. Two of them have previously been presented in [7]. The last MAP seems at least as visually self-similar as the LAN data trace displayed in [61] over a number of time scales (3-4 perhaps). All of the MAPs considered in this contribution are the special cases found by superposing IPP processes and a Poisson stream.

$$\mathbf{C} = \begin{bmatrix} -(c_1 - \lambda) & c_1 \\ c_2 & -c_2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{(C.46)}$$

First we consider the MAP defined be superposing 9 homogeneous IPP where each IPP in MAP notation is described by $c_1 = 7.5 * 10^{-5}$, $c_2 = 2.5 * 10^{-5}$ and $\lambda = \frac{4}{3}$ We will denote this MAP : model 1. The coefficient of variation for the marginal distribution of this MAP is $C_{var} = 44.8$. Secondly we examine the MAP denoted model 2, see (C.47), which is a superposition of 4 inhomogeneous IPPs and a Poisson stream. The coefficient of variation for the marginal distribution of this MAP is $C_{var} = 1.55$.

$$\begin{array}{lll} \lambda_{poisson} & = 0.3 & \\ \lambda & = 0.8 & , c_1 = 5 * 10^{-3} \quad \text{and } c_2 = 5 * 10^{-3} \\ \lambda & = 2.0 & , c_1 = 5 * 10^{-4} \quad \text{and } c_2 = 5 * 10^{-4} \\ \lambda & = 1.1 & , c_1 = 5 * 10^{-5} \quad \text{and } c_2 = 5 * 10^{-5} \\ \lambda & = 1.5 & , c_1 = 5 * 10^{-6} \quad \text{and } c_2 = 5 * 10^{-6} \end{array} \qquad \text{(C.47)}$$

Finally we consider an 8 state MAP, see (C.48), called model 3, found by superposition of 3 inhomogeneous IPP's and a Poisson stream. The marginal distribution using this model has the coefficient of variation C_{var} = 1.36.

$$
\begin{array}{lll}
\lambda_{poisson} & = 0.7 & \\
\lambda & = 2.0 & , c_1 = 9 * 10^{-3} \text{ and } c_2 = 9 * 10^{-3} \\
\lambda & = 1.5 & , c_1 = 2 * 10^{-4} \text{ and } c_2 = 2 * 10^{-4} \\
\lambda & = 1.1 & , c_1 = 5 * 10^{-5} \text{ and } c_2 = 5 * 10^{-5}
\end{array}
\tag{C.48}
$$

In figure C.1 we illustrate the dynamic behaviour of the superposition (C.48) of 3 IPP's and a Poisson stream in the same manner as in [61]

First we try the simplest of the shuffling experiments carried out in [32] in that we examine the mean queue length behaviour as a function of the load for our three models and their respective marginal distributions.

It is clear that in all three cases the mean queue length show the same kind of adverse behaviour as described in [32] i.e. have knees at loads that are substantially lower than 1. The worst behaviour is clearly seen for model 1 where the knee is located at an utilization of about 0.35. For the other models, 2 and 3, the knee is located at utilization levels of about 0.55 - 0.60. These last two models correspond better both "visually" and in marginal distribution to the LAN data observed in for example [61] or [32]. Also the location of the knees for these two curves correspond very well with those found in [32]. It is also striking that in all three cases the corresponding marginal distributions show queueing behaviour that seems harmless and close to M/D/1.

Clearly, the (mean) queueing behaviour of these MAPs have little to do with their marginal distribution.

A lesson learned from examining the queueing behaviour is that with our MAPs we often see a factor 2 in difference between the mean queue lengths at an arbitrary time and the mean queue lengths experienced by arrivals. There is therefore good reason to be careful not to emphazise to much on the queue length at an arbitrary time, since the queue length experienced by arrivals is the actual issue.

To go further into what properties of a MAP process governs its queueing behaviour we compared each of the three models above with an IPP designed

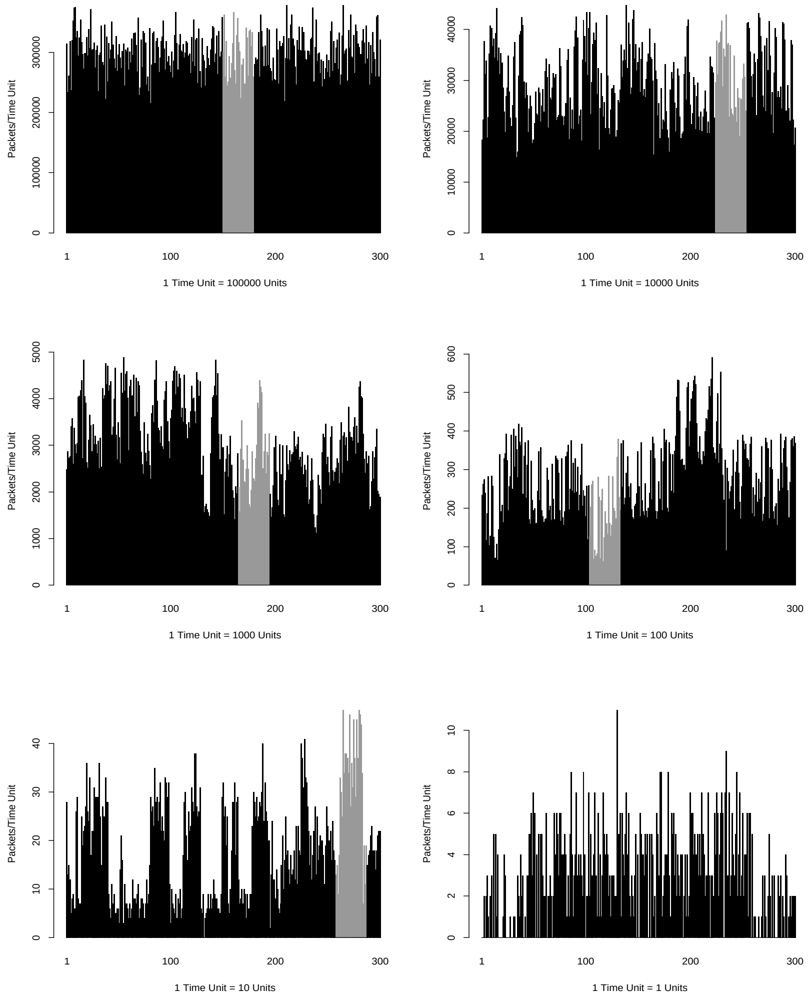

Figure C.1: **Dynamic behaviour of an 8-state MMPP over 6 time scales**

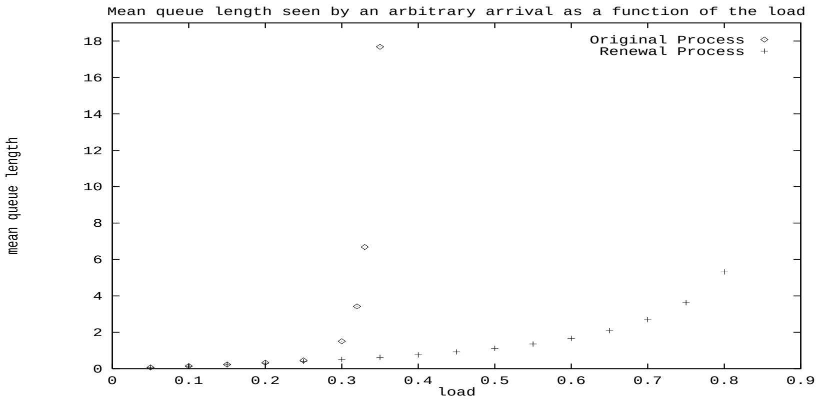

Figure C.2: **Comparison of the mean queue length seen at arrival epochs in a 10 state MAP process (model 1) and the renewal process defined by its marginal distribution**

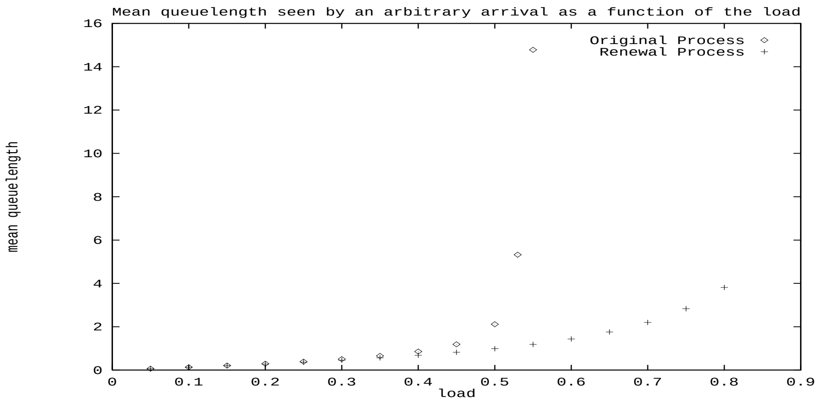

Figure C.3: **Comparison of the mean queue length seen at arrival epochs in a 16 state MAP process (model 2) and the renewal process defined by its marginal distribution**

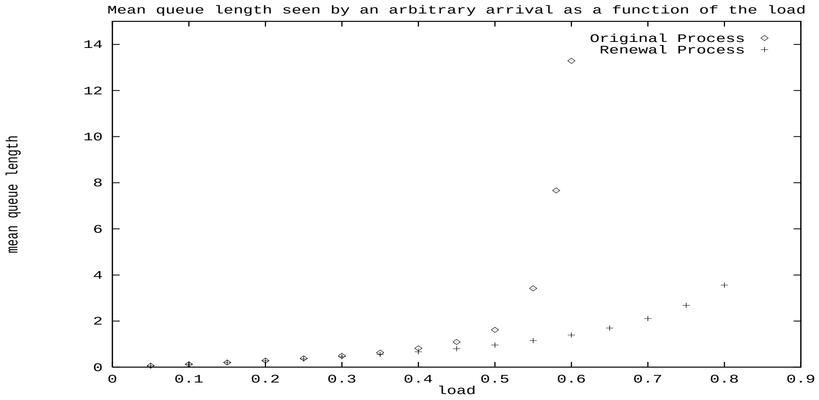

Figure C.4: **Comparison of the mean queue length seen at arrival epochs in a 8 state MAP process (model 3) and the renewal process defined by its marginal distribution**

to follow the original processes Index of Dispersion of Counts (IDC) curve as closely as possible. The three parameters available in the IPP (C.46) were found by matching the rate of the process, the asymptotic value of the IDC curve and finally the time constant which is defined by the point in time where the IDC "breaks". The three IPP's found had the following parameters shown in (C.49)

$$\begin{array}{llll} IPP1: & \lambda = 1.50 & , c_1 = 1.667 * 10^{-5} & \text{and } c_2 = 3.333 * 10^{-5} \\ IPP2: & \lambda = 1.11 & , c_1 = 4.545 * 10^{-7} & \text{and } c_2 = 4.5455 * 10^{-6} \\ IPP3: & \lambda = 1.11 & , c_1 = 4.545 * 10^{-6} & \text{and } c_2 = 4.5455 * 10^{-5} \end{array} \tag{C.49}$$

The IDC curves of the original model and the IPP fitted to its IDC curve are shown in the figures C.5, C.6 and C.7.

When comparing the queueing behaviour of each of these IPPs with the MAP defining them it is found that in all three cases the defining MAP's show a much more critical queueing behaviour than the corresponding IPP. IPP1 defined by model 1, shows knee behaviour at a load of $\rho = 0.6$. From figure C.2 it seen that

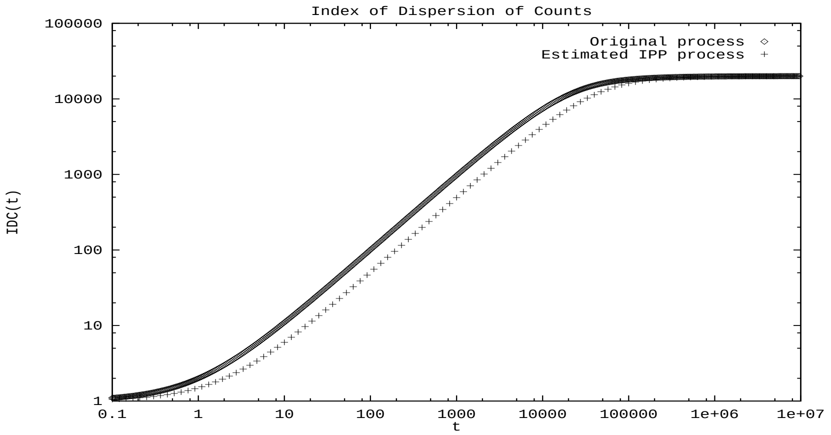

Figure C.5: **The IDC curve of model 1 and the corresponding fitted IPP**

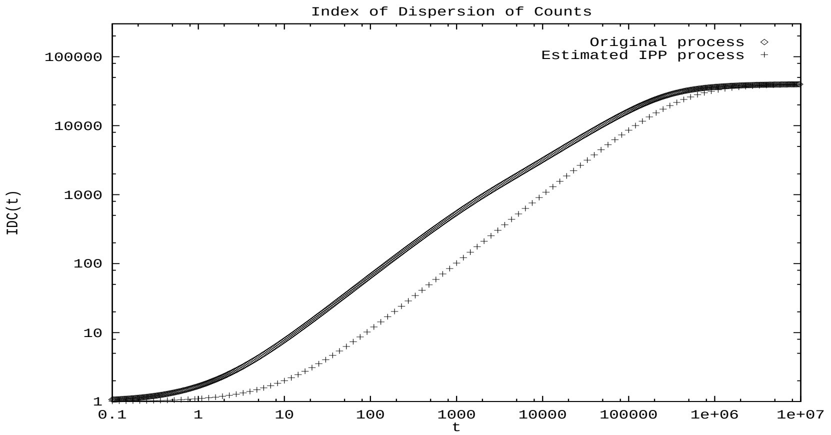

Figure C.6: **The IDC curve of model 2 and the corresponding fitted IPP**

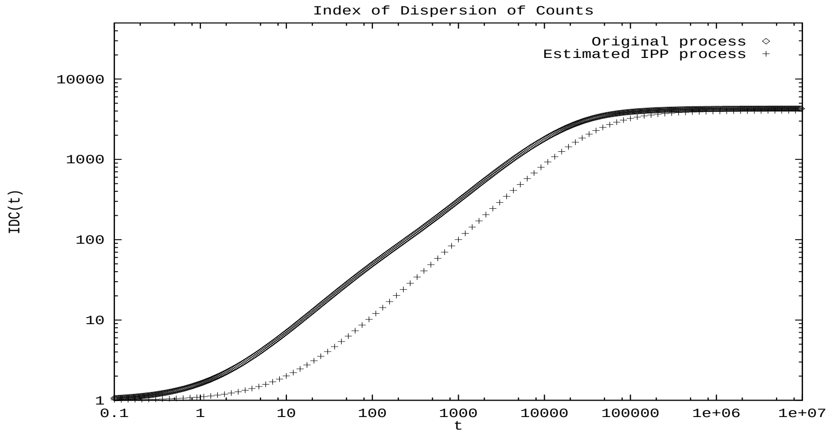

Figure C.7: **The IDC curve of model 3 and the corresponding fitted IPP**

in model 1 the knee is located around $\rho = 0.6$. For IPP2 and IPP3 corresponding respectively to model 2 and 3 no knee behaviour is seen for $\rho < 0.8$ whereas both model 2 and model 3 show critical behaviour around $\rho = 0.6$.

It is striking here to note how different queueing behaviour can be obtained with processes with very close IDC curves. It is note worthy that in our three examples the fitted IPP's yield some sort of "best case" queueing behaviour.

All the models we have considered in this contribution are of the special case MMPP of the MAP. It is clear from the above that in order for a MMPP/G/1 to show critical behaviour ,the MMPP must overload the system in at least one of its states. Another requirement is that it stays in this (set of) state(s) maintaining overload long for a build up in the queue to occur. These observations have a very close relationship with the caudal characteristic curve examined by M.F. Neuts in [76].

C.6 Conclusions and Future work

The above clearly demonstrates that dimensioning after some mean load requirements is very risky business. The load at which a given system should be driven is clearly very dependent on the characteristics of the traffic it experiences. The caudal characteristic curve [76] can prove to be a valuable tool when trying to quantify what load is reasonable for a given arrival process. Currently we are investigating the caudal characteristic curve associated with MAPs with variability over several time scales.

Recently we have begun studies regarding the implications of this highly variable traffic for finite queueing systems. A finite queueing system seems more realistic when modelling ATM. Implementing the finite MAP/G/1 queue makes it possible to quantify cell loss probabilities as a function of load for a given MAP. This is of great interest given the very strict cell loss requirements in ATM. The finite MAP/G/1 queue has a clear advantage over the infinite when examining systems which are conditionally overloaded ,i.e in the critical region around the knee, in that iterations for the **G** matrix are avoided. When examining queues in the critical load region we have observed that 10000 - 100000 iterations often are needed to obtain **G** with sufficient precision.

Finally, it can be noted the model 2 and model 3 seem to show queueing behaviour very similar to what trace driven simulation of LAN traffic have shown [32]. These findings indicate that it seems feasible to use the MAP to model very variable LAN traffic even though it is an unresolved and open question how systematicly to design a MAP to describe the variable traffic reasonably over several time scales.

Appendix D

A[4]

Modelling and performance study of packet-traffic
with self-similar characteristics
over several time-scales with
Markovian Arrival Processes (MAP).

Allan T. Andersen and Alex Jensen
Bo Friis Nielsen
Institute of Mathematical Modelling
Building 321
Technical University of Denmark
2800 DK-Lyngby

Presented at the Nordic Teletraffic Seminar
NTS12 Espoo, Finland August 1995
VTT Symposium 154 pp. 269 - 283

D.1 Introduction

Currently many countries/carriers are on the verge of implementing broadband ISDN (B-ISDN). This raises a number of challenging engineering problems. How should the dimensioning in B-ISDN take place ? Can methods currently used when modelling POTS be used or do new methods have to be developed ? In other words is it reasonable to expect that the future B-ISDN traffic can be modelled in the same way as the current narrow band traffic.

Recently a number of measurement studies of packet traffic have been initiated, i.e. packet traffic from Local Area Networks (LAN) [35], Variable Bit Rate (VBR) Video [17] and Integrated Services Digital Network (ISDN) [57]. Of these at least the interconnection of LAN's and the transportation of VBR are expected to be main contributors to traffic in B-ISDN. These studies reveal behaviour very different from what classical traffic models assume. The measurements indicate that these forms of packet traffic have long range dependencies and slowly decaying variances. Characteristics of the observed behaviour have in other contexts been associated with self-similar behaviour. Therefore the measurements motivate the assertion that certain types of packet traffic exhibit self-similar behaviour over several time-scales.

Packet traffic with self-similar behaviour has serious implications for both system engineering and system performance. It is problematic to design/dimension systems only with respect to mean load considerations since such an approach totally fails to take into account the behavioral dynamics associated with self-similarity. The challenge is here to develop models that can describe this adverse behaviour reasonably and are able to quantify new performance measures and dimensioning tools.

This has recently initiated a search for traffic models which are able to describe this behaviour in a relatively simple manner. Here models well known in other contexts have been investigated i.e. Chaotic Maps [31] and Fractional Brownian motion [81].

While arrival processes based on such models can be described using only a few parameters it is analytically very difficult to analyze a queueing system with these processes as input.

Here we examine the dynamic behaviour of three special cases of the Markovian Arrival Process (MAP). All three cases are the very special cases obtained by

superponing Interrupted Poisson Processes (IPP's) . A relatively simple 8 state MAP - a superposition of three appropriately parameterized Interrupted Poisson Processes (IPP) and a Poisson Stream, i.e. a superposition of 4 renewal streams is shown to yield variability, if not self-similar behaviour, over several time-scales. Other 8-16 state MAP's presented in [7] and [47] also seem at least as "visually" self - similar over several time scales as the packet traffic data presented in [61].

In this contribution we also examine the mean queueing behaviour of the presented MAP models. Here we have implemented some of the algorithms for the single server queue with a Markovian Arrival Process (MAP) outlined by D.M. Lucantoni in [64] and originally by V. Ramaswami in [83].

Queueing analysis of the LAN traffic traces has been initiated in [32] where the queueing behaviour of a queue driven with LAN traffic traces has been examined. One of the conclusions in [32] was that short term correlations in the arrival process is of less importance regarding queueing behaviour.

This paper is organized as follows. Section 2 contains a brief description of self-similarity. In Section 3 a very brief introduction to the MAP is given. Section 4 contains a description of the three MAP models and an examination of some aspects of their dynamic behaviour. Section 5 describes aspects of the mean queueing behaviour of the three MAP models. In Section 6 we conclude and give directions for further work.

D.2 Measures of Self-Similarity

The short introduction to self-similarity given here is primarily based on [61] and [91]. In the context of packet traffic we are concerned with the process defined by the number of packets arriving in a time-interval. This makes it natural to consider the stationary discrete time stochastic process $X = X_t$ for $t = 1,2,....$ (Number of packets arriving in time-interval 1,2 ....). We also consider the time-series obtained by averaging the discrete time process over non-overlapping blocks. That is, for $m = 1,2,3$ .., we consider the time-series

$X_k^{(m)} = (1/m)(X_{m(k-1)+1} + ... + X_{mk}),\ k = 1,2,3,....$

The stationary discrete time stochastic process X with mean 0 and variance σ^2 is said to be self-similar with self-similarity parameter H if for all m = 1,2,3.... $(X_{m(k-1)+1} + \ldots + X_{mk}) = mX_k^{(m)}$, k = 1,2,3,... has the same finite dimensional distributions as $m^H X$.

If the variance and autocorrelation function of $(X_{m(k-1)+1}+\ldots+X_{mk}) = mX_k^{(m)}$, k = 1,2,3,... and $m^H X$ are the same for m = 1,2,3,.... X is said to be second-order self-similar with self-similarity parameter H. This implies that for all m = 1,2,3 $var(X^{(m)}) = \sigma^2 m^{-\beta}$, and that $r^{(m)}(k) = r(k) = \delta^2(|k|^{2-\beta})$ (k = 1,2,3...) where $\beta = 2 - 2H$ and $\delta^2(f)$ denotes the second central difference operator applied to a function, that is, $\delta^2(f(k)) = f(k+1) - 2f(k) + f(k-1)$. X is called asymptotically second-order self-similar with self-similarity parameter $H = 1 - \beta/2$ if $r^{(m)}(k) \to \delta^2(|k|^{2-\beta})$, as $m \to \infty$ (k = 1,2,3,.....)

The definition of long-range dependence is that $\sum_k r(k) \to \infty$. For a self-similar process this corresponds to the self-similarity parameter H in the range $0.5 < H < 1.0$. For such a process the averaged time-series $X^{(m)}$ possesses a non-degenerate correlation structure as $m \to \infty$. For all conventional tele-traffic models $r_m(k) \to 0$ as $m \to \infty$, (k=1,2,3,...), that is, the averaged process $X^{(m)}$ tends to second order pure noise for large m corresponding to $H = 0.5$ or equivalently $\beta = 1$. In the following a process with $H = 0.5$ will be called a short-range dependent process.

In practice there are several ways to try to detect whether an observed process exhibits self-similar behaviour or not. A simple way to check for any indications of self-similarity is to plot the averaged process $X^{(m)}$ for different values of m, m values differing by an order of magnitude. If the plots “look the same“ it is an indication of a self-similar behaviour.

In the context of packet traffic attention has been directed towards three ways of checking for self-similarity which all examine second order properties. These can be referred to as methods based on the variance-time plot or equivalently the index of dispersion of counts (IDC), the pox diagram of the R/S statistic (also called the rescaled adjusted range plot) and finally analysis based on the periodogram (in the frequency-domain). Further discussion of these three techniques for checking for self-similarity is outside the scope of this work, we refer to [61] and [91].

In our search for MAP models with self-similar behaviour we have primarily used visual self-similarity as a yardstick to determine whether a given MAP

seemed reasonable as a model for self-similar behaviour over several time scales rather than using some sort of stringent control of the Hurst parameter over several time scales using for example the second order methods mentioned above. However, plots of the variance time curve or equivalently the IDC curve of the interesting MAP models found here and in [7] and [47] clearly show that over the observed time-scales an estimated value of H, $\hat{H}$, satisfies $0.5 < \hat{H} \leq 1$ so that the correlation structure over the observed time scales does not degenerate to a Hurst estimate of $\hat{H}$= 0.5.

MAP models will eventually turn out to be short range dependent in the above described sense. That is, once you observe a given MAP at a time scale larger than the largest time constant in the MAP the correlation structure will yield a Hurst estimate of $\hat{H}$ = 0.5. However, examining the process on smaller time scales yields Hurst estimates in the range 0.5 - 1.0 which one of the MAP's presented here (and others in [7] and [47]) clearly demonstrate.

D.3 Some results regarding the MAP

In this section we will very briefly sketch some main characteristics of the Markovian Arrival Process (MAP). For further results we refer to [64], [77] and [83]. For results and derivations regarding the Matrix Analytic approach used when solving the MAP/G/1 queue we refer to the same three references mentioned above.

The MAP is the point process generated by the transition epochs of Markov renewal process whose transition probability matrix $\mathbf{F}(*)$ is of the form

$$\mathbf{F}(x) = \int_0^x e^{\mathbf{C}u} du \mathbf{D} = (\mathbf{I} - e^{\mathbf{C}x})(-\mathbf{C}^{-1})\mathbf{D}, \tag{D.1}$$

where the square matrices of order m, $\mathbf{C} = [C_{ij}]$ and $\mathbf{D} = [D_{ij}]$ are respectively a stable matrix and a non-negative matrix whose sum is an irreducible infinitesimal generator. (The matrix exponential is defined by its series i.e. $e^{\mathbf{M}} = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}$, for the properties of the matrix exponential see [16]).

The infinitesimal generator $\mathbf{Q} = \mathbf{C} + \mathbf{D}$ describes the transitions among the transient states $i, ..., m$ in the underlying Markov process. The stationary vector $\vec{\pi}$ of this Markov process satisfies the equations

$$\vec{\pi}(\mathbf{C} + \mathbf{D}) = \vec{\pi}\mathbf{Q} = \vec{0}, \quad \vec{\pi}\vec{e} = 1 \tag{D.2}$$

where $\vec{e}$ denotes a column vector of 1's.

Some well-known processes are (very) special cases of the MAP

- The Poisson process : $\mathbf{C} = (-\lambda)$ and $\mathbf{D} = (\lambda)$ (Poisson process with rate λ)
- The phase-type (PH) renewal process is also a special case of the MAP. It includes both the Erlang and hyper-exponential (equivalent with the Interrupted Poisson Process IPP [59]) and Poisson arrival processes as special cases.
- The Markov-modulated Poisson process (MMPP) is also a special case of the MAP. The MMPP is the doubly stochastic Poisson process whose arrival rate depends on an m-state irreducible Markov process i.e $\mathbf{C} = \mathbf{Q} - \Lambda, \mathbf{D} = \Lambda$ where the underlying Markov process has infinitesimal generator $\mathbf{Q}$ and $\Lambda = diag(\lambda_1, ..., \lambda_m)$.

For a MAP model the marginal distribution between arrivals is readily seen to be of phase-type. The probability distribution function for the marginal time t between arrivals is given as

$$P\{T \leq t\} = F(t) = 1 - (\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}e^{\mathbf{C}t}\vec{e} \tag{D.3}$$

The moments of the marginal distribution are given as

$$m_i = i!(\vec{\pi}\mathbf{D}\vec{e})^{-1}\vec{\pi}\mathbf{D}(\mathbf{C})^{-i}\vec{e} \tag{D.4}$$

Since we in this paper use the IPP (Special case of the two state MMPP with $\lambda_2 = 0$) as a sort of "building stone" for more general processes it is convenient to list some expressions regarding the IPP.

The IPP can be expressed in MAP notation in the following manner

$$\mathbf{C} = \begin{bmatrix} -(c_1 + \lambda) & c_1 \\ c_2 & -c_2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \tag{D.5}$$

The steady state probability vector is easily seen to be $\vec{\pi} = (\frac{c_2}{c_1+c_2}, \frac{c_1}{c_1+c_2})$.

The mean of the counting process N_t for the time stationary IPP is given as

$$E\{N_t\} = \frac{\lambda c_2}{c_1 + c_2}t \tag{D.6}$$

The expression for the Index of Dispersion of counts curve , $IDC(t)$ ($= Var\{N_t\}/E\{N_t\}$) for the time stationary IPP is known to be (see for example [41]):

$$IDC(t) = \quad 1 + \frac{2\lambda c_1}{(c_1+c_2)^2} - \frac{2\lambda c_1}{t(c_1+c_2)^3}(1 - e^{-(c_1+c_2)t}) \tag{D.7}$$

It is clear from the definition of $IDC(t)$ that the process defined by a superposition of identical independent processes maintains the same $IDC(t)$ curve as a single one of the processes. This is, of course, obtained at the expense of a growing mean arrival rate of the resulting process.

Examining the expression for the IDC in the special case of an IPP (D.7) reveals that it is possible to fix both the mean arrival rate and the IDC when investigating the process of superponed homogeneuos IPP's. This can be done by fixing the following three constants k_1, k_2 and k_3 and examining the resulting process for different N (N is the number of IPP's superponed).

$$\begin{aligned} k_1 = & \quad c_1 + c_2 \\ k_2 = & \quad \lambda c_1 \\ k_3 = & \quad N\lambda c_2 \end{aligned} \tag{D.8}$$

(The mean arrival-rate is $\frac{k_3}{k_1}$)

D.4 Dynamic behaviour of special cases of the MAP

In this section we examine the dynamic behaviour and correlation structure of different special cases of a MAP. We use analytical MAP tools and simulation. Simulating MAP processes makes it possible to get a first look at how $X^{(m)}$ behaves at different aggregation levels m. Here it is interesting to examine a realization at aggregation levels m differing for example by an order of magnitude to see how the process behaves over several time-scales. Whenever we in the following visually inspect a realization we look at aggregation levels differing by an order of magnitude. Starting with the largest aggregation level, we successively examine the next lower aggregation level by zooming in on a randomly chosen part of the process at the previous aggregation level. This is similar to the method used in [61] and [62].

A very simple MAP yielding correlations in the process X_t is the IPP (D.5). Some important remarks are here that the IPP is a renewal process and that the inter-arrival time distribution is hyper-exponential [59] . The sum of the transition rates in the modulating process $(c_1 + c_2)$ is called the time constant for the IPP.

We examine the dynamic behaviour of three IPP's with the same λ but with different time constants spanning 4 orders of magnitude. They can be defined as $\lambda = 4.0$, $c_1 = s\frac{1}{4}$ and $c_2 = s\frac{3}{4}$, where s is either 10^{-2}, 10^{-4} or 10^{-6}. These three IPP's will be denoted respectively IPP$_a$,IPP$_b$ and IPP$_c$.

In figure D.1 a realization of IPP$_b$ is shown. It is obvious that the on/off type behaviour does not resemble self-similar behaviour. The on/off type behaviour is only visible when viewing the process at aggregation levels corresponding to a time interval that is small compared to the reciprocal of the time constant.

Next we consider the index of dispersion of counts (IDC) curve for each of the three processes IPP$_a$, IPP$_b$ and IPP$_c$ in figure D.2. It is evident in the log-log plot that the index of dispersion of counts for all 3 processes increases approximately linearly over several time-scales, the number of time-scales determined by the IPP's time constant. It is also evident that the slope of the linear growth in the log-log plot corresponds to $\beta \approx 0$ giving a Hurst parameter of $H \approx 1$ in the interval of linear growth. This is not surprising when examining the formula (D.7). It is clear that for $t(c_1 + c_2) << 1$ the IDC curve can be represented by its Taylor expansion as

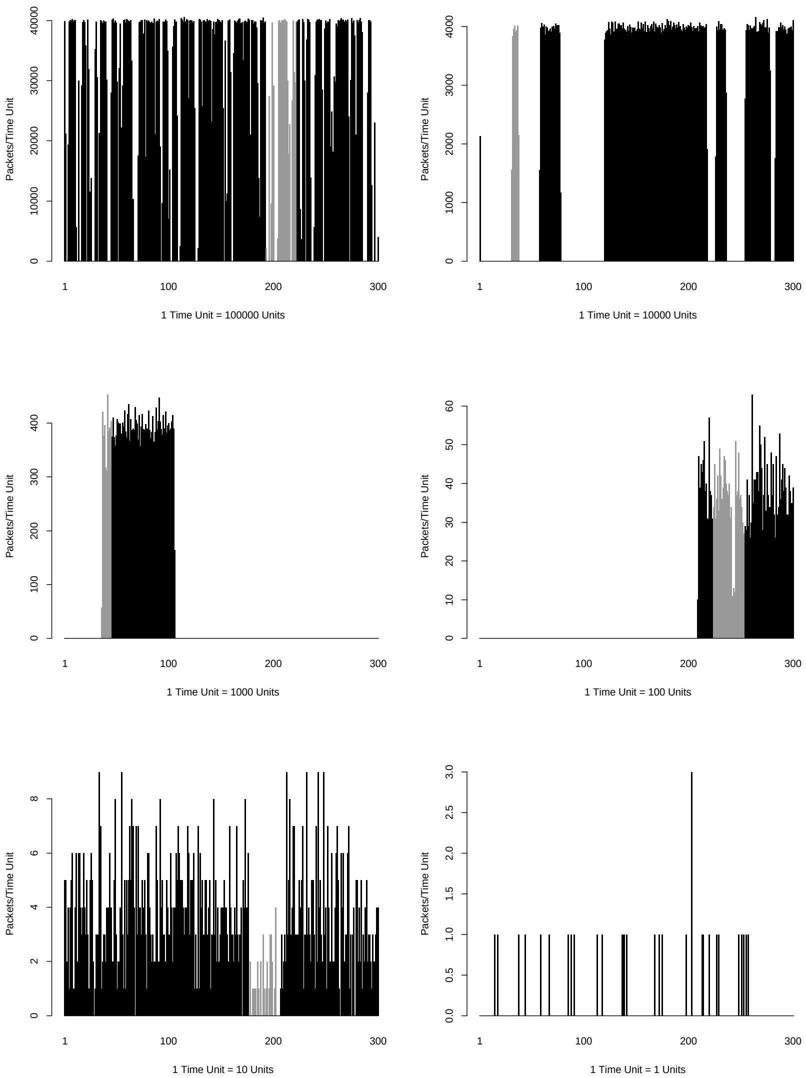

Figure D.1: **Dynamic behaviour of an IPP (IPP$_b$) over 6 timescales**

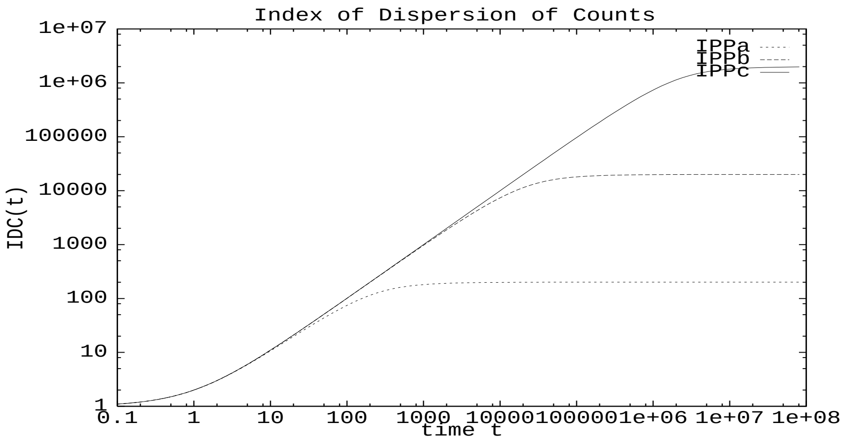

Figure D.2: **The IDC curves of the three IPP's, IPP$_a$, IPP$_b$ and IPP$_c$**

$$IDC(t) = \quad 1 + \frac{\lambda c_1}{(c_1+c_2)} t + o(t), \quad t(c_1 + c_2) << 1 \qquad \text{(D.9)}$$

It is well known that for a renewal process $IDC(t) \rightarrow (C_{var})^2$, as $t \rightarrow \infty$. ($C_{var} = \frac{\sigma}{\mu_1}$ is the coefficient of variation of the renewal process, μ_1 and σ being the mean and standard deviation of the inter-arrival time distribution). Since the IPP is a renewal process "the price paid" for dependence over many time-scales is therefore a highly dispersed marginal distribution. The three IPPs above have the following coefficients of variation, IPP$_a$: C_{var} = 14.2 , IPP$_b$: C_{var} = 141.4 and IPP$_c$: C_{var} = 1414.2.

Next we consider the superposition of several identical IPP's. As mentioned earlier it is possible to choose a process of superpositions of IPP's to have the same arrival rate and $IDC(t)$ as a single IPP. Here we have chosen the process defined by a superposition of 9 identical IPP's so that the resulting process has the same rate and $IDC(t)$ as IPP$_b$. From (D.8) it can be derived that each of the 9 IPP's to be superponed are defined by $c_1 = 7.5 * 10^{-5}$, $c_2 = 2.5 * 10^{-5}$ and $\lambda = \frac{4}{3}$. Figure D.3 shows a realisation of the superposition of 9 IPP's with the same rate and $IDC(t)$ as IPP$_b$.

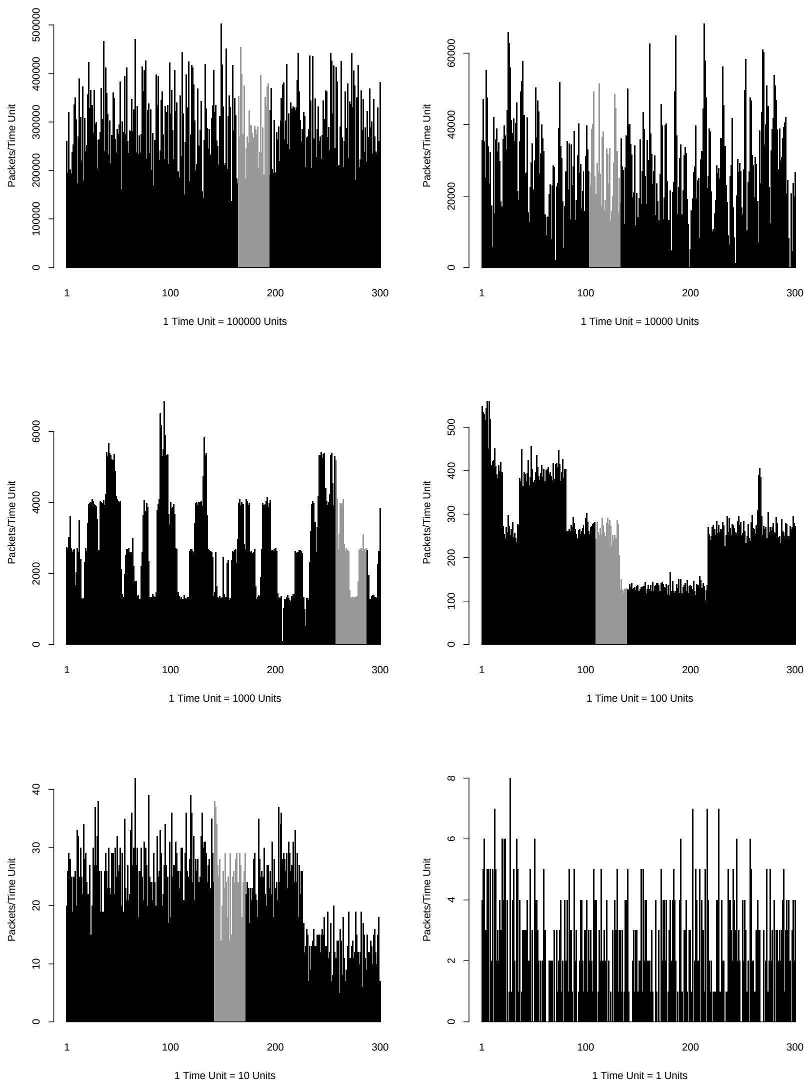

Figure D.3: **Dynamic behaviour of a superposition of 9 homogenous IPP's with same rate and IDC(t) as IPP$_b$ over 6 timescales**

While modelling self-similar behaviour significantly better than the single IPP it is still evident from the plots of the different aggregation levels that the plots do not "look the same" over several timescales (at least not for this relatively modest choice of N). An inherent limitation in the use of superponed homogeneous IPP's in the context of modelling self-similarity is that they, of course, exactly as in the case of a single IPP only are able to model traffic with $H \approx 1$. However, the marginal distribution of the superponed process compared to the single IPP with same rate and IDC is not as dispersed. For this process of 9 superponed IPP's the coefficient of variation is C_{var} = 44.8 (recall that for IPP$_b$ C_{var} = 141.4).

Another approach is to investigate a process of superponed inhomogeneous IPP's. Here we consider an 8 state MAP, see (D.10), found by superposition of 3 inhomogeneous IPP's and a Poisson stream. The marginal distribution for this model has the coefficient of variation C_{var} = 1.36. In figure D.4 we illustrate the dynamic behaviour of the superposition (D.10) of the 3 IPP's and the Poisson stream.

$$\begin{array}{lllll} \lambda_{poisson} & = 0.7 & & & \\ \lambda & = 2.0 & , c_1 = 9 * 10^{-3} & \text{and} & c_2 = 9 * 10^{-3} \\ \lambda & = 1.5 & , c_1 = 2 * 10^{-4} & \text{and} & c_2 = 2 * 10^{-4} \\ \lambda & = 1.1 & , c_1 = 5 * 10^{-5} & \text{and} & c_2 = 5 * 10^{-5} \end{array} \qquad \text{(D.10)}$$

There is evidently high variability on all the time-scales displayed. It is apriori very difficult to decide which aggregation level the individual plots belong to. It can be asserted that visually they appear to be at least as self-similar as the plots shown in [61].

To get a more formal measure of the dynamic behaviour of the superposition of 3 inhomogenous IPP's and a Poisson stream the IDC curve is calculated in figure D.5.

A value of the derivative ($1 - \beta$ for a self-similar process) in the log - log plotted IDC curve can be obtained by a finite difference approximation. Using the logarithmically spaced points from the calculation of the IDC curve the derivative is calculated in figure D.6.

The derivative definitely lies in a band between 0 and 1 over several time-scales. This indicates a "Hurst parameter" somewhere between $\frac{1}{2}$ and 1 over several

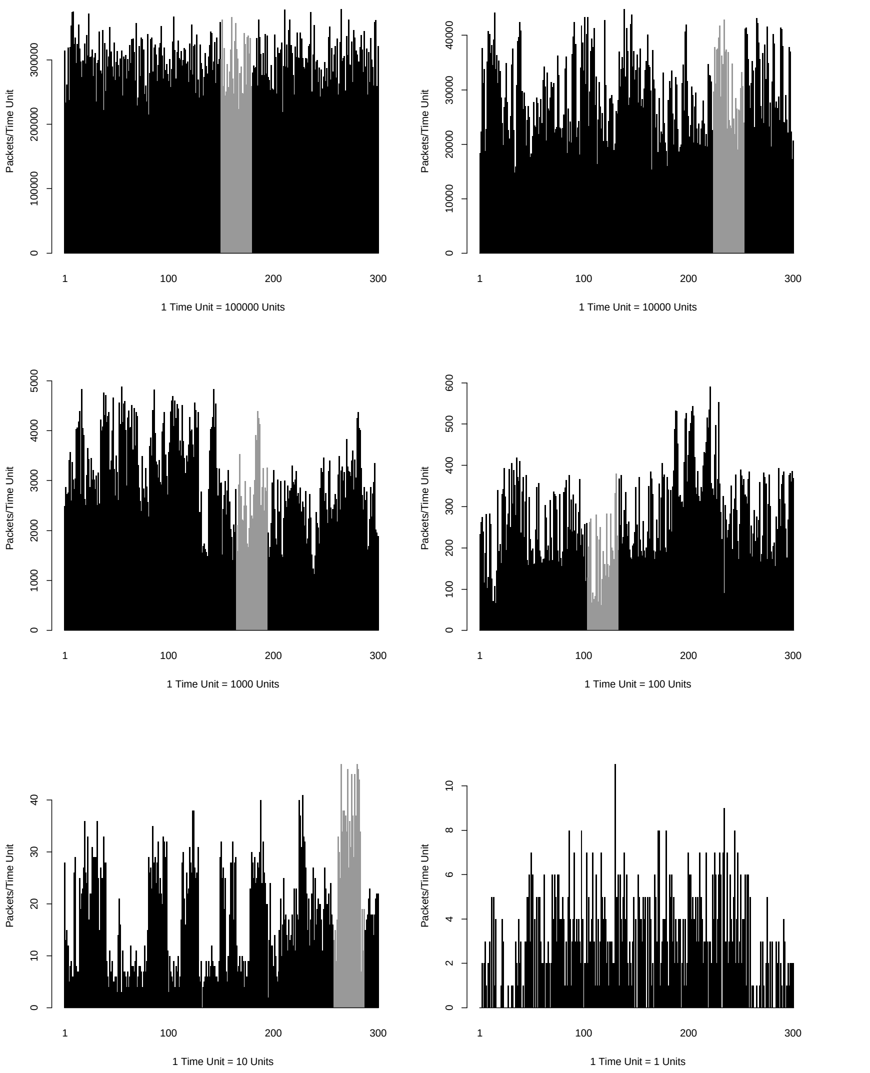

Figure D.4: **Dynamic behaviour of the superposition of 3 inhomogenous IPP's and a Poisson stream over 6 time scales**

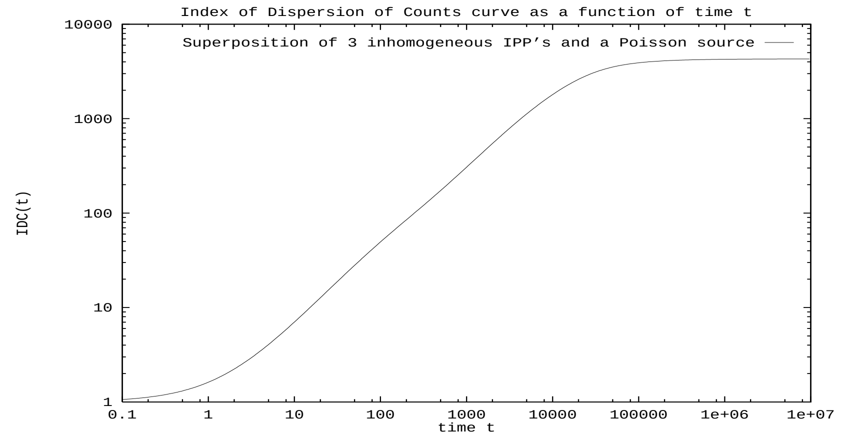

Figure D.5: **The IDC curve of the superposition of 3 inhomogenous IPP's and a Poisson stream**

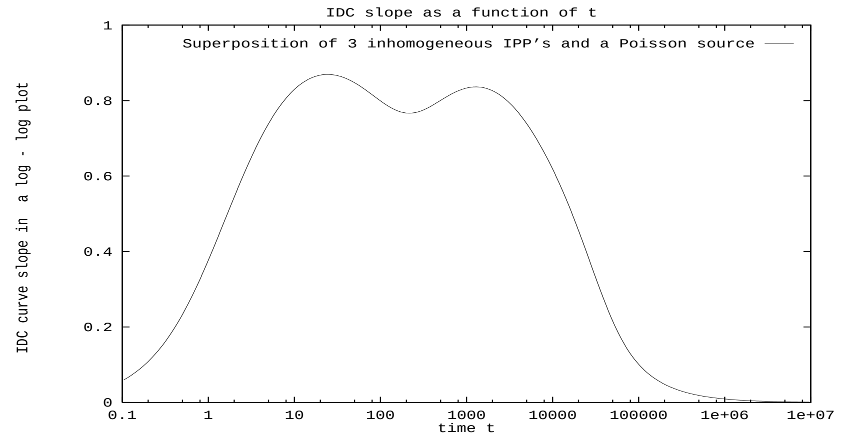

Figure D.6: **The calculated derivative of the IDC curve for a superposition of 3 inhomogenous IPP's and a Poisson stream**

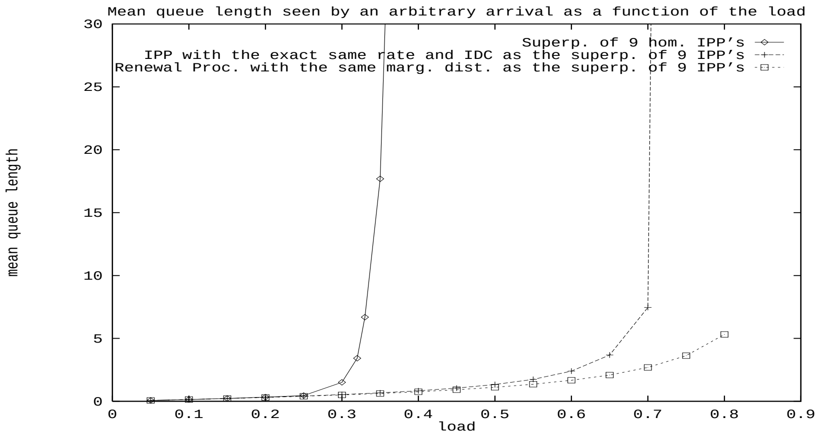

Figure D.7: **Comparison of the mean queue length seen at arrival epochs in process defined by a superposition of 9 IPP's , the renewal process defined by its marginal distribution and a single IPP with same rate and IDC as the superposition**

time-scales. Hence this superponed process is more versatile in modelling the IDC of a self-similar process than the previous homogeneuos superpositions.

D.5 Queueing experiments with MAPs

In the following we will examine the mean queueing behaviour as a function of the load for the three previously presented MAP's. In figure D.7 we examine the mean queueing behaviour of the superposition of the 9 IPP's, IPP$_b$, which has the same rate and IDC as the superposition, and the renewal proces defined by the marginal distribution of the superposition.

It is clear from figure D.7 that the process defined by the superposition by far has the most adverse queueing behaviour with a knee around a load of 0.35 i.e. substantially lower than 1.0 which is the same phenomenon as observed with

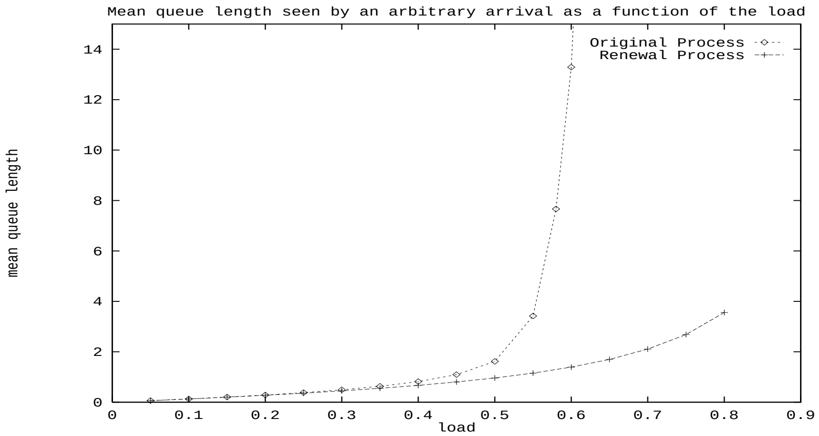

Figure D.8: **Comparison of the mean queue length seen at arrival epochs in a process defined as a superposition of 3 inhomogenous IPP's and a Poisson stream and the renewal process defined by its marginal distribution**

the LAN traces in [32]. It is very interesting to note that two processes with same rate and $IDC(t)$ can result in such different queueing behaviour as shown in figure D.7. It might seem some what surprising that the IPP has "nicer" queueing behaviour than the superponed process. In this light, it is interesting to note that the coefficient of variation for the inter arrival time distribution is significantly smaller for the superposition than for the corresponding IPP.

In figure D.8 we have examined the mean queue length behaviour of the superposition of 3 inhomogeneous IPP's and a Poisson source and the renewal process defined by its marginal distribution. The mean queue length shows the same kind of adverse behaviour as described in [32] with a knee around a load of 0.55 - 0.60. This model corresponds both "visually" and in marginal distribution to the LAN data observed in for example [61] or [32]. Also the location of the knee for this curve corresponds very well with those found in [32]. It is striking that the renewal process defined by the marginal distributions shows queueing behaviour that seems harmless and close to M/D/1.

D.6 Conclusions and Future work

It seems that even very simple MAPs are able to model self-similar behaviour reasonably over at least a number of time-scales. A MAP created by superponing 3 inhomogeneous IPP's and a Poisson stream seems to be visually self-similar over several time-scales and has a non-degenerate correlation structure over several time-scales.

The queueing examples clearly demonstrates that dimensioning after some mean load requirements is very risky business. The load at which a given system should be driven is clearly very dependent on the characteristics of the traffic it experiences.

It has been shown that two processes with same rate and $IDC(t)$ can lead to substantially different queueing behaviour which raises the question of what characteristics of an arrival process determines queueing behaviour.

It can be noted that especially the presented superposition of inhomogeneous IPP's seems to show queueing behaviour very similar to what trace driven simulation of LAN traffic have shown [32]. These findings indicate that it seems feasible to use the MAP to model very variable LAN traffic even though it is an unresolved and open question how systematicly to design a MAP to describe the variable traffic reasonably over several time scales.

Appendix E

A[5]

Intercept and Asymptotic Behaviour
of the nGI/G/1 Queue

S. Blaabjerg, Allan T. Andersen
Bo Friis Nielsen
Technical University of Denmark
DK-2800 Lyngby

Presented at the Nordic Teletraffic Seminar
NTS12 Espoo, Finland August 1995
VTT Symposium 154 pp. 167 - 180

E.1 Introduction

The exact evaluation of the general single server queue is surprisingly intractable. Only when either the arrival process is Poisson or the service time distribution is exponential it is easy. With the introduction of the Markovian Arrival Process (MAP) introduced by Neuts in [74] it has become manageable by considerable computational power to handle quite general single server queues and even the modelling of apparently self-similar traffic appears to be feasible with the MAP tool [7].

It is, however, inquestionable that the performance evaluation remains complex and it is difficult both to quantify what properties of the arrival process are really important and to use the MAP tool directly in real time applications like equivalent bandwidth in ATM networks.

In this paper we consider the single server queue, we restrict the arrival process to a superposition of renewal streams, and present a very simple approximation which;

- utilizes that the asymptotic behaviour for large queue lengths is exponential with an easy computable decay rate,
- and which takes into account the intercept which from recent studies e.g. [25] is known occasionally to be significant.

The usual way to study superposition of arrival processes is simply to add them and to thin them out proportionally. However, an alternative way which keeps the index of dispersions for counts fixed also exists. Both kinds of superpositions are studied here.

E.2 Superposition of Arrival Processes

In this section we consider the superposition of a number of independent homogeneous ON-OFF sources, each represented by an Interrupted Poisson Process (IPP). The IPP is a special case of the MAP, with the representation:

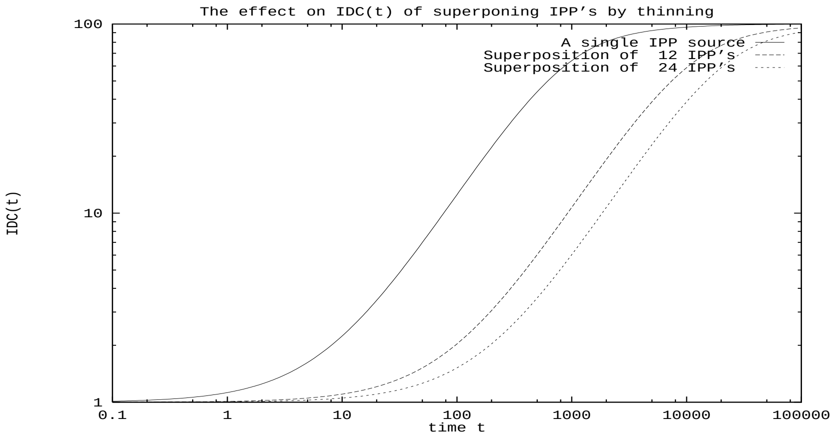

Figure E.1: **The effect on the resulting IDC curve for the traditional way of superponing processes by thinning them out proportionally**

$$\mathbf{C} = \begin{bmatrix} -(c_1+\lambda) & c_1 \\ c_2 & -c_2 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix} \tag{E.1}$$

The steady state probability vector is easily seen to be $\vec{\pi} = (\frac{c_2}{c_1+c_2}, \frac{c_1}{c_1+c_2})$.

E.2.1 The Traditional Way

The traditional way to superpose such sources is to scale the parameters $\lambda^{(n)}$, $c_1^{(n)}$, $c_2^{(n)}$ to maintain the average load. Thus $\lambda^{(n)}=\frac{\lambda}{n}$, $c_1^{(n)}=\frac{c_1}{n}$, $c_2^{(n)}=\frac{c_2}{n}$. This corresponds to a scaling of the time axis and is illustrated in figure E.1.

E.2.2 Keeping the Index of Dispersion for Counts fixed

Another approach exists where besides the rate, the IDC of the aggregated process is unaltered. For a detailed description see [6]. In table E.1 we show

Number of sources	λ	c_1	c_2
1	0.13750	$2.2917\ 10^{-3}$	$2.2917\ 10^{-4}$
12	0.12604	$2.5000\ 10^{-3}$	$2.0833\ 10^{-5}$
24	0.12552	$2.5104\ 10^{-3}$	$1.0460\ 10^{-5}$
48	0.12526	$2.5156\ 10^{-3}$	$5.2408\ 10^{-6}$

Table E.1: IPP Parameters fixing the IDC curve for the process defined as a superposition of IPP's

parameters that fix the IDC curve for the resulting process when superponing a different number of IPP sources. The single IPP source in table E.1 is the same as the one studied in [25].

E.3 Asymptotic Tail Behaviour of Queueing Systems

For ease of understanding we have chosen to present how the tail behaviour of a single server queueing system can be computed by considering a sequence of examples with increasing generality and increasing mathematical suffitication.

E.3.1 The M/G/1 case:

The waiting time for this queueing system is known to have the following Laplace Transform

$$W^*(s) = \frac{s(1-\rho)}{s-\lambda+\lambda B^*(s)} \tag{E.2}$$

The tail behaviour is determined by the dominant root of $W^*(s)$ and it is given by the solution to

$$(-s) - \lambda + \lambda B^*(-s) = 0 \text{ which is equivalent to } \frac{\lambda}{\lambda+s}B^*(-s) = 1 \tag{E.3}$$

E.3.2 The GI/G/1 case:

For this system with a general renewal input, the Kingman bound see e.g. [54], applies and the asymptotic slope of the tail s_0 is determined by

$$s_0 = sup\{s > 0 | A^*(s)B^*(-s) = 1\} \tag{E.4}$$

where $A^*(s)$ is the Laplace Transform of the interarrival time distribution a and $B^*(s)$ is the Laplace Transform of the service time distribution b. In case of Poisson input $A^*(s) = \frac{\lambda}{\lambda+s}$ from which we see that the Kingman way of computing s_0 is an extension of the approach from the M/G/1 case.

E.3.3 The G/G/1 case

The theory of large deviations can be used to prove for that for a large class of arrival processes the queue length distribution possesses an exponential tail behaviour, see e.g. [28].

Let $W(0,t)$ be the amount of work arriving to the system in time $(0,t)$, and define

$$\lambda(\theta) = lim_{t\to\infty}[\frac{1}{t}logE(e^{\theta W(0,t)})] \tag{E.5}$$

In the case where the arrival process is Poisson with intensity λ, and the service time is deterministic we get $logE(e^{\theta W(0,t)}) = log[exp(\lambda t(e^{\theta}-1)] = \lambda t(e^{\theta}-1)$ from which we see $lim_{t\to\infty}[\frac{1}{t}logE(e^{\theta W(0,t)})] = \lambda(e^{\theta}-1)$. This computation should take away a little of the lack of intuition which the definition of $\lambda(\theta)$ is burdened with.

Let V denote the virtual waiting time. For "general" arrival processes without long range dependence the following (large deviation) result, [28] remain valid

$$lim_{B\to\infty}\frac{1}{B}log(P\{V > B\}) = -s_0 \tag{E.6}$$

where $s_0 = sup\{\theta > 0 | \lambda(\theta) \leq \theta\}$ and the time unit is chosen such that it takes one time unit to serve one unit of work.

To illustrate the generality of the Large Deviation result we consider two special cases.

The Markov Modulated Poisson Process

From [74] it is known that the generating function of the number of arrivals in a window of length t can be found as

$$P(z,t) = \vec{\pi} exp[(Q - \Lambda + z\Lambda)t]\vec{e} \quad \text{(E.7)}$$

where Λ is the rate matrix and Q is the generator of the underlying Markov Process. From this expression it can be seen that in the case where the generator matrix can be diagonalized and the service time is deterministic we have that $\lambda(\theta)$ is the dominant eigenvalue of $Q - \Lambda + e^{\theta}\Lambda$.

Superposition of identical renewal processes

In this case $W_n(0,t) = \sum_{i=1}^{n} W_i(0,t)$. Put $\lambda(\theta) = lim_{t\to\infty} log E(e^{\theta W_i(0,t)})$. Then
$\lambda_n(\theta) = lim_{t\to\infty} log E(e^{\theta W_n(0,t)}) = lim_{t\to\infty} log E(e^{\theta \sum_{i=1}^{n} W_i(0,t)})$
$= lim_{t\to\infty} log E(\Pi_{i=1}^{n} e^{\theta W_i(0,t)}) = lim_{t\to\infty} \sum_{i=1}^{n} log E(e^{\theta W_i(0,t)} = n\lambda(\theta)$.

From this computation we see that if s_0 solves $\lambda(\theta) = \theta$ then s_0 *also solves* $\lambda_n(\theta) = n\lambda(\theta) = n\theta$ from which we see that the asymptotic slope remains the same if we superpose a number n of identical streams and increase the server capacity by the same factor n. (This corresponds to the traditional way of superposition)

E.4 Intercept Behaviour in the nG/D/1 queue

In order to obtain a simple closed form approximation for the probability of the queue length exceeding a given buffer threshold in the nGI/D/1 queue, we look at the nD/D/1 and M/D/1 queues.

The classical diffusion approximation of the M/D/1 queue, see e.g. section 6.1.2. in [86], leads to

$$P\{V_{M/D/1} > r\} \approx exp(-\frac{2(1-\rho)}{\rho} r) \quad \text{(E.8)}$$

In comparison a Brownian bridge approximation of the nD/D/1 queue, see e.g. section 6.2.2. in [86], leads to

$$P\{V_{nD/D/1} > r\} \approx exp(-\frac{2(1-\rho)}{\rho} r - \frac{2r^2}{n}) \quad \text{(E.9)}$$

In the nD/D/1 case we see the same term as in the M/D/1 case but in addition we see a term which on the logaritmic scale decrease with the square of r. By applying the correct exponential term $s_{M/D/1}$ from solving (E.3) with $B^*(s) = e^s$ both approximations can be improved. The improved form states

$$\begin{aligned} P\{V_{M/D/1} > r\} &\approx \rho exp(-s_{M/D/1} r) \text{ and} \\ P\{V_{nD/D/1} > r\} &\approx \rho exp(-s_{M/D/1} r - \tfrac{2r^2}{n}) \end{aligned} \quad \text{(E.10)}$$

where $s_{M/D/1}$ solves $s - \lambda + \lambda e^s = 0$. The multiplication of ρ ensures that the probability of an empty system is correct.

For the superposition of n renewal streams in which the squared coefficients of variations are c_a^2 we suggest

$$P\{V_{nGI/D/1} > r\} \approx \rho exp(-s_{M/D/1} r - (1 - (c_a^2)^{\beta}) \frac{2r^2}{n + K_a r}) \quad \text{(E.11)}$$

In comparison to the expression for the nD/D/1 queue the second term has been modified in two ways.

1. A term $K_a r$ is added to the usual denominator n.
   Thereby we enforce an exponential tail behaviour with rate $s_0 = s_{M/D/1} + 2(1-(c_a^2)^{\beta})/K_a$. The constants K_a and β are to be chosen such that both the correct asymptotic slope s_0 as well as the correct mean queue length are obtained.

2. A multiplicative term $(1 - (c_a^2)^{\beta})$.
   This term vanish in the case with Poisson input making the expression correct in this case. The multiplicative term is 1 in the nD/D/1 case and with $K_a = 0$ our heuristic formula is seen to in accordance with formula (E.10).

It should be noticed that formula (E.11) implies that for $c_a^2 > 1$ the nGI/D/1 curve will be above the M/D/1 curve and for $c_a^2 < 1$ it will be below.

With a tail behaviour that is asymptotically exponential, the *intercept* is defined as the limit $I_c = lim_{r\to\infty} exp(s_0 r)P(V > r)$. With the heuristic formula (E.11) we get

$$\begin{aligned} log(I_c) = & \quad lim_{r\to\infty}[s_0 r + log(\rho) - s_{M/D/1}r - 2(1-(c_a^2)^\beta)\frac{r^2}{n+K_a r}] \\ = & \quad lim_{r\to\infty}[log(\rho) + x(s_0 - s_{M/D/1}r - \frac{2(1-(c_a^2)^\beta)}{K_a}\frac{1}{1+\frac{n}{K_a r}} - 1 + 1)] \\ = & \quad lim_{r\to\infty}[log(\rho) + x(\frac{2(1-(c_a^2)^\beta)}{K_a}\frac{n/(K_a r)}{1+n/(K_a r)}] \\ = & \quad log(\rho) + \frac{2(1-(c_a^2)^\beta)}{K_a^2}n \end{aligned}$$

Therefore we end up with

$$I_c(n) = \rho exp[\frac{2(1-(c_a^2)^\beta)}{K_a^2}n] = \rho exp[\frac{(s_0 - s_{M/D/1})^2}{2(1-(c_a^2)^\beta)}n] \tag{E.12}$$

since $s_0 = s_{M/D/1} + \frac{2(1-(c_a^2)^\beta)}{K_a}$. The prediction that the intercept has an exponential dependence on the number of sources is in nice accordance with the numerical results leading to the conjecture of formula (1.6) in [25].

E.5 Application: Equivalent Bandwidth of a Renewal Stream

In this short section we indicate how the result of the former section can be used to assign an equivalent bandwidth with a renewal stream.

E.5.1 Without Intercept

Assume that our renewal stream with rate λ is offered to a buffer of length B and that the QoS is given by a quantile $P(V > B) \le e^{-\delta}$. The equivalent bandwidth e_q of the renewal stream is now

$$e_q(\delta, B) = \inf\{C > 0 | e^{-s_0 B} \le e^{-\delta}\} \tag{E.13}$$

which correspond to finding the smallest link capacity C for which $s_0(C) \ge \frac{\delta}{B}$. We see that s_0 as a function of C becomes the function of prime importance and that for C approaching λ from above makes s_0 tend to zero from above.

Consider the Kingman method. Here $s_0(C)$ is found as the positive root to $A^*(s)e^{s/C} = 1$. Viewing $F(C,s) = A^*(s)e^{s/C}$ as a function of C, and s with s being a function of C the Implicit Function Theorem says $\frac{ds_0(C)}{dC} = -\frac{\partial F/\partial C}{\partial F/\partial s}$

Since $\frac{\partial F}{\partial C} = -s(C)A^*(s)e^{s/C}/C^2$ and since $\frac{\partial F}{\partial s} = (A^*(s)/C + dA^*(s)/ds)e^{s/C}$ we get

$$s'(C) = \frac{s(C)A^*(s(C))/C^2}{A^*(s)/C + dA^*(s)/ds} \tag{E.14}$$

Since both $s(C) \to 0$ and $A^*(s)/C + dA^*(s)/ds \to 0$ when $C \to \lambda$ we need to apply the rule of L'Hospital's and after isolation of s' we end up with

$$\frac{ds_0(\lambda)}{dC} = \frac{2}{c_a^2} \tag{E.15}$$

Approximating s_0 by its first order Taylor expansion, this is reasonable for loads close to 1, and inserting in (E.13) we finally end up with

$$e_q(\lambda) = \lambda + c_a^2\frac{\delta}{2B} \tag{E.16}$$

.

E.5.2 Taking the Intercept into account

The equation $s_0(C) \ge \frac{\delta}{B}$ is to be substituted with $s_0(C) \ge \frac{\delta}{B} + \frac{(s_0(C)+s_{M/D/1}(C))^2}{2B(1-(c_a^2)^\beta)}$ when the intercept is to be taken into account. Approximating s_0 and $s_{M/D/1}$ with there 1 order expansions gives when inserting rise to the following second order equation in $C - \lambda$.

$$\frac{2(c_a^2-1)^2 n}{B(1-(c_a^2)^\beta)}(C-\lambda)^2 - 2c_a^2(C-\lambda) + \frac{\delta}{B}c_a^4 = 0 \tag{E.17}$$

Choosing the physically right solution we get for $c_a^2 < 1$

$$C - \lambda = c_a^2 \frac{B(1-(c_a^2)^\beta)}{2(c_a^2-1)^2 n}\left[1 - \sqrt{1 - \frac{2(c_a^2-1)^2 n}{B(1-(c_a^2)^\beta)}\frac{\delta}{B}}\right] \tag{E.18}$$

and Taylor expanding the square root to second order gives

$$e_q(C) = \lambda + c_a^2 \frac{\delta}{2B}\left(1 + \frac{(1-c_a^2)^2 n}{2B(1-(c_a^2)^\beta)}\frac{\delta}{B}\right) \tag{E.19}$$

while for $c_a^2 > 1$ we get

$$C - \lambda = c_a^2 \frac{B((c_a^2)^\beta - 1)}{2(c_a^2-1)^2 n}\left[\sqrt{1 + \frac{2(c_a^2-1)^2 n}{B((c_a^2)^\beta - 1)}\frac{\delta}{B}} - 1\right] \tag{E.20}$$

and Taylor expanding the square root to second order gives

$$e_q(C) = \lambda + c_a^2 \frac{\delta}{2B}\left(1 - \frac{(c_a^2-1)^2 n}{2B((c_a^2)^\beta - 1)}\frac{\delta}{B}\right) \tag{E.21}$$

As expected we see that for $c_a^2 < 1$ the equivalent bandwidth is increasing with n while for $c_a^2 > 1$ it is decreasing.

E.6 Numerical Results

In table E.2 we show the parameters used in the heurestic approximation for the 6 processes examined here.

E.6.1 Performance of the approximation for traditional superposition

In figure E.2-E.7 we evaluate the accuracy of the heuristic formula (E.11). The exact results are obtained using the solution approach for the MAP/G/1 [64]. The figure indicates that the accuracy of the formula detoriates as the number of sources increases.

System	ρ	$s_{M/D/1}$	s_0	β	${c_a}^{2\beta}$	K_a
12 IPP/D/1	0.30	2.06457	0.0181046	0.474	8.8789	7.69996
24 IPP/D/1	0.30	2.06457	0.0181046	0.490	9.5757	8.38096
48 IPP/D/1	0.30	2.06457	0.0181046	0.574	12.5433	11.2813
12 E_2/D/1	0.80	0.430842	0.861684	0.83	0.56253	2.03077
24 E_2/D/1	0.80	0.430842	0.861684	0.86	0.55095	2.08451
48 E_2/D/1	0.80	0.430842	0.861684	0.79	0.57834	1.95736

Table E.2: Parameters used in the heuristic approximation

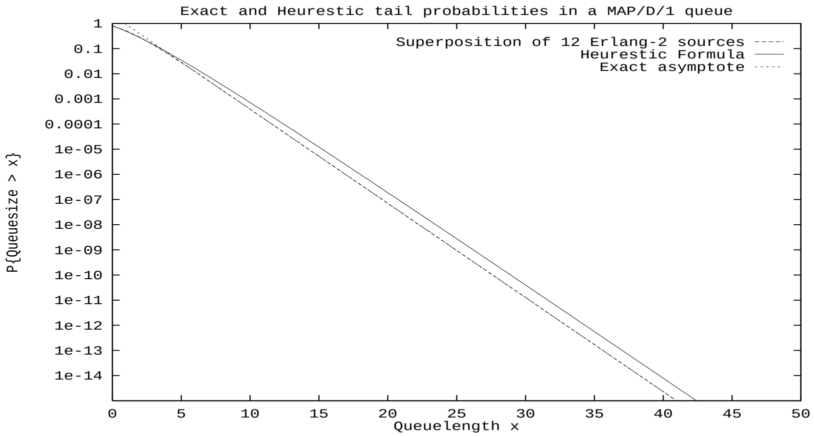

Figure E.2: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 Erlang-2 sources**

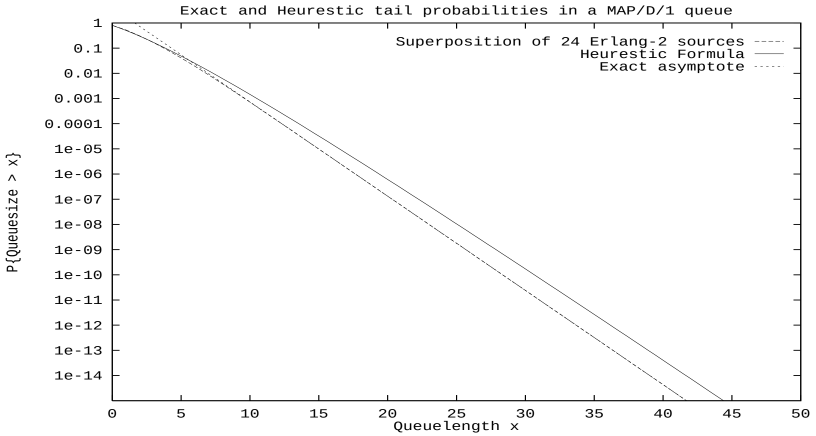

Figure E.3: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 Erlang-2 sources**

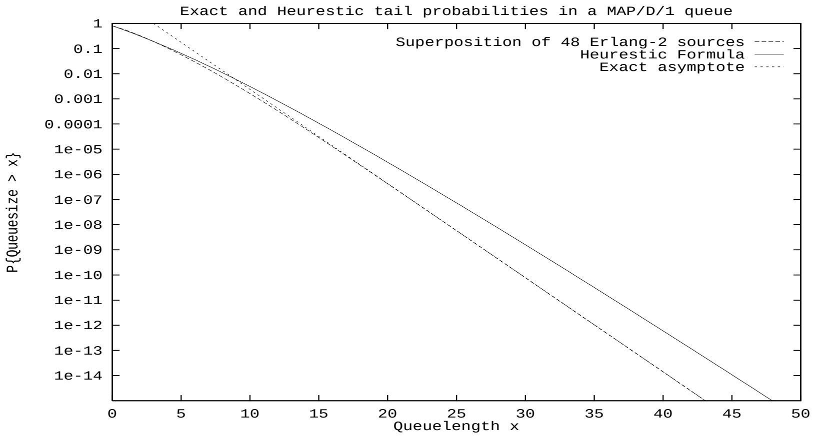

Figure E.4: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 Erlang-2 sources**

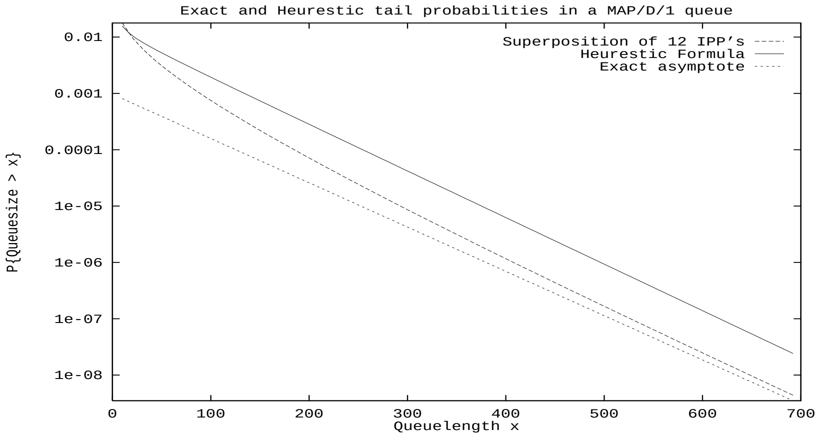

Figure E.5: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 IPP sources**

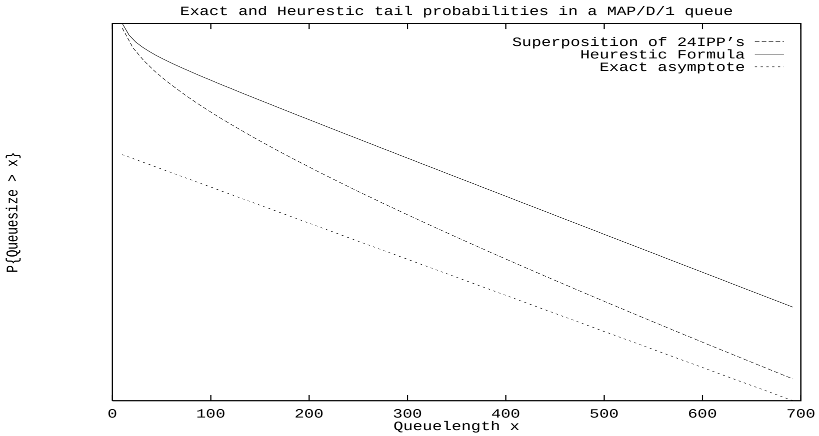

Figure E.6: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 IPP sources**

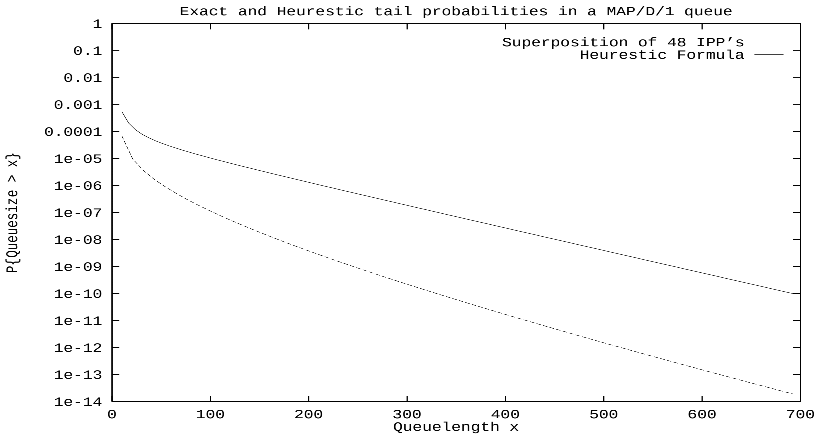

Figure E.7: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 IPP sources**

E.6.2 Queueing behaviour for the two different approaches of doing the superposition

The queueing behaviour for different superpositions using the traditional way of thinning the resulting process is shown in figure E.8.

The queueing behaviour for different superpositions found by fixing both the rate and IDC curve of the resulting process is shown in figure E.9.

E.7 Conclusion

We have compared two different approaches of doing superpositions and shown that for the traditional way of superposing there is a multiplexing gain, while the alternative approach leads to worse behaviour when multiplexing many sources. Further, we have suggested a very simple heuristic formula for the complementary queue length distribution with applications to equivalent bandwidth. In our experiments this heuristic turns out to be an upper bound.

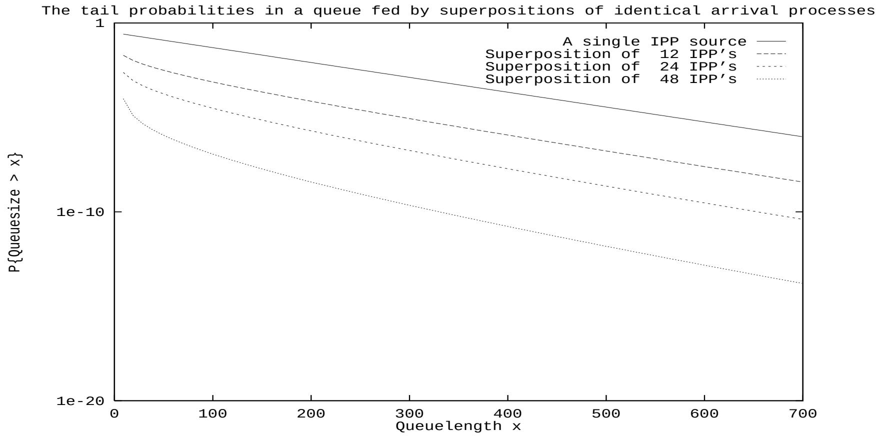

Figure E.8: **Tail probabilities in queues driven by different superpositions of the same IPP. Rate fixed in each case by scaling the resulting process**

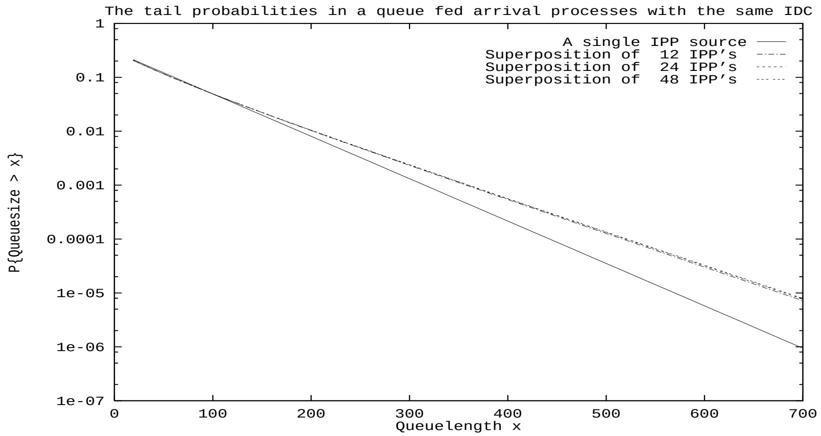

Figure E.9: **Tail probabilities in queues driven by different superpositions where the rate and IDC curve for the resulting processe is fixed**

The continuation of this study will be to investigate the possibilities for keeping the IDC fixed for superpositions of other types of sources. Since we find that the heuristic is promising within the field of equivalent bandwidth we will continue our studies by investigating the upper bound properties and try to improve on the accuracy of the heuristic.

E.8 Appendix: Kingman bound as special case of Large Deviation

We will now indicate why the Large Deviation way of computing s_0 is consistent with the Kingman bound in the case where the arrival process is a single renewal stream. From [27] it is shown that the double Transform of the number of arrivals in in time interval $(0, t)$, $N(0, t)$ (first Z-transform with respect to number of arrivals and then Laplace transform with respect to s) is given by

$$G_N^{**}(s, z) = \frac{1}{s} + \frac{z-1}{m_a s^2}\frac{1-A^*(s)}{1-zA^*(s)} \tag{E.22}$$

A small rewriting gives

$$G_N^{**}(s, z) = \frac{1}{s}\frac{1-zA^*(s)+(z-1)A_0^*(s)}{1-zA^*(s)} \tag{E.23}$$

where $A_0^* = \frac{1-A^*(s)}{m_a s}$ is the Laplace Transform of the residual interarrival time. By assuming simple poles and rational Laplace transform of A we can invert with respect to s and get an expression of the form

$$G_N^{*}(t, z) = \sum_{i=1}^{n} p_j(z)e^{r_j(z)t} \tag{E.24}$$

where $r_j(z)$ are the poles and $p_j(z)$ are the associated residues.

To get from $N(0, t)$ to $W(0, t)$ in the transform domain we simply use the fact that a random sum of identically distributed and independent random variables is given by the composition of the transforms. This yields

$$G_W^{*}(t, \theta) = \sum_{i=1}^{n} p_j(B^*(-\theta))e^{r_j(B^*(-\theta))t} \tag{E.25}$$

Now $lim_{t\to\infty}\frac{1}{t}log(G_W(t, \theta)) = r_n(B^*(-\theta))$ when r_n is the pole with largest real part, and it is necessarily real. Since it is real it must be the solution to the equation $1 = zA^*(s)$ which is the same as $A^*(s) = 1/B^*(-\theta)$. On the real line A^* is monotone decreasing and we may invert to get $r_n = (A^*)^{-1}(1/B^*(-\theta)) = \lambda(\theta)$. Since the condition is $\lambda(\theta) \leq \theta$ which is equivalent to $1/B^*(-\theta) \geq A^*(\theta)$ implying $A^*(\theta)B^*(-\theta) \leq 1$.

Appendix F

A[6]

Superposition and Equivalent Bandwidth of Renewal Streams

S. Blaabjerg
Institute of Telecommunications, Building 343, DTU
DK-2800 Lyngby, Denmark
e-mail sbb@it.dtu.dk
A. T. Andersen
Institute of Mathematical Modelling, Building, 321, DTU
DK-2800 Lyngby, Denmark
e-mail ata@imm.dtu.dk
B. F. Nielsen
Institute of Mathematical Modelling, Building, 321, DTU
DK-2800 Lyngby, Denmark
e-mail ata@imm.dtu.dk

Submitted to INFOCOM 1996 San Fransisco, Calif

Abstract :
This paper considers the superposition of renewal streams with applications to effective bandwidth in ATM networks. From a computation of the exact asymptotic decay rate in the nGI/D/1 queue and by utilizing characteristics of the corresponding formula for the nD/D/1 queue in the Brownian Bridge approximation a closed form approximation of the virtual waiting time in the nGI/D/1 queue is derived. From this approximation it is straigtforward to get an estimate of the asymptotic constant, from which illuminating formulas for the effective bandwidth of a renewal stream is obtained which takes into account the statistical multiplexing gain. Numerical examples are provided to evaluate the accuracy of the approximations.

Keywords : ATM, effective bandwidth, asymptotic decay rate, asymptotic constant, renewal stream.

F.1 Introduction

The procedure of connection admission control in ATM networks can be greatly simplified if it is possible to associate to any type of ATM connection a single bandwidth parameter, based on burstiness characteristica and QoS requirements, to be used for bandwidth allocation at connection set up. Such an approach would enable the application of the well known theory for multi-rate circuit switched network to call scale modelling of ATM networks, see [37], where performance parameters can be obtained in known exact or approximate ways, see e.g. [46] and [69]. Over the last couple of years there has therefore been an intensive study of the feasibility of the concept of effective bandwidth as well as numerous suggestions for how much bandwidth to allocate to a given type of connection.

The approaches seen so far to assign effective bandwidths can be partitioned into two schools of thoughts which are destinguished by the assumption on the buffer size in the underlying statistical multiplexer problem.

In the first approach a small buffer which can only handle congestion on cell scale is assumed. On burst scale, a loss system appears, and from intensive numerical experience [63] or by applying the Chernoff bound to estimate the probability that the input rate exceeds the link capacity, the first method in [51] manage to assign effective bandwidth to the individual connections.

In the second approach, a large buffer is assumed and the probability of the queue length exceeding a given threshold can be estimated by a simple exponential formula

$$P(Q > r) \approx \alpha e^{-s_\infty r} \tag{F.1}$$

where s_∞ is the exponential asymptotic decay rate at very large occupancy levels (r big) which a large class of queueing systems is known to possess and α is the associated asymptotic constant. In order to obtain af sufficient degree of simplicity the asymptotic constant α is then put to 1. This has been justified with the fact that QoS requirement in ATM is very strict (around 10^{-9} in cell loss) and in this case the relative error of approximating α with 1 is not big.

In [38] the exponential asymptotic decay rate is used to assign effective bandwidth to ATM sources when the statistical multiplexer model is described by the fluid flow queueing model, [30] generalises their results to the case in which

the sources are general Markovian sources, while [52] uses the theory of Large Deviations to prove that for a very large class of sources, it is possible to assign an effective bandwidth from this exponential tail behaviour. [51] gives a very illustrative description of how it is possible to assign effective bandwidth to both the burst scale loss system as well as the burst scale delay system with either tail behaviour constraints or constraints on the mean queue length. [40] is probably the first paper in which a combination of burst scale loss and burst scale delay systems is used to derive a single effective bandwidth formula to be applied independently of the buffer size.

The main disadvantage with the burst scale loss approach is that the amount of effective bandwidth allocated can be much too large due to the fundamental assumption on very small switch buffers which may not hold true. On the other hand the approach of very large buffers may also not hold true. With reasonable sized buffers the obtainable utilization is often well below 1, and for this case it is pointed out in [25] that the estimate based on the asymptotic slope can be very misleading due to the fact that the asymptotic constant α can be decades smaller than 1 in the case of multiplexing many bursty sources, and it can be decades greater than 1 when multiplexing many smooth sources.

The objective of this paper is to point out that in the case where the source can be modelled by a renewal process (which covers the on/off source with exponential on- and general distributed off period), it is possible by a heuristic closed form approximation of the virtual waiting time distribution of the nGI/D/1 queue to

- assign effective bandwidth to any renewal stream including those with squared coefficient of variation smaller than 1,
- to get an estimate of the asymptotic constant, and its dependence on the number of multiplexed sources
- to take advantage of the statistical multiplexing gain in the sense that the number of multiplexed sources is a parameter in the formula for the effective bandwidth.

In section 2 we shortly review relevant methods for obtaining the exponential decay rate as well as the asymptotic constant. In section 3 we first introduce a closed form heuristic approximation for the virtual waiting time in the nGI/D/1 queue, and show how this enables a very simple estimate of the asymptotic constant as well as simple approximate formulas for the effective bandwidth which takes into acount the potential multiplexing gain in contrast to the usual large buffer approach. In section 4 we present numerical results to evaluate the heuristic formula, the estimate of the asymptotic constant as well as the formulas for effective bandwidth.

F.2 The Asymptotic Decay Rate and the Asymptotic Constant

In the statistical multiplexing problem which is the basis for the whole analysis we denote the amount of work arriving in the time interval $]0, t]$ by $A(0, t)$, and assume that *the time unit is chosen such that it takes one time unit to serve one unit of work*, that is the capacity of the server is always 1. The unfinished work at an arbitrary time (also denoted the virtual waiting time) is denoted V and from section 5 in [86] we know that $V = \sup_{t>0}(A(-t, 0) - t)$. The actual waiting time is the unfinished work at an arbitrary arrival, and from [1] it is known that the asymptotic decay rate of the virtual waiting time and the actual waiting time is the same for the case with Markovian Arrival Processes. Here we shall mainly work with the virtual waiting time.

F.2.1 Asymptotic Decay Rate

Since the renewal process has a central role in this paper and since a classical result (Kingman bound) for the single server queue with renewal input exist, we consider this case first, and after that the more general method based on large deviation is shortly described.

The GI/G/1 case

For the single server queue with a general renewal arrival process, the Kingman bound see e.g. [54], and [53] applies, and the asymptotic decay rate s_∞ of

the waiting time distribution is determined as the strictly positive real root of $A^*(s)B^*(-s) = 1$ i.e.

$$s_\infty = \inf\{s > 0 | A^*(s)B^*(-s) = 1\} \tag{F.2}$$

where $A^*(s)$ is the Laplace Transform of the interarrival time distribution and $B^*(s)$ is the Laplace Transform of the service time distribution.

The G/D/1 case

The theory of large deviations can be used to prove that for a large class of arrival processes the virtual waiting time distribution possesses an exponential tail behaviour, see e.g. [28].

With $A(0, t)$ being the amount of work arriving to the system in time $(0, t)$, we define the *asymptotic decay rate function*[1] as

$$\lambda(\theta) = lim_{t\to\infty}[\frac{1}{t} log E(e^{\theta A(0,t)})] \tag{F.3}$$

With V denoting the virtual waiting time, the following (large deviation) result, corollary 2.3 with $a = 1$ and $v = 1$, ($a = 1$ and $v = 1$ implies that no long range dependence is allowed), [28] states that:

$$lim_{r\to\infty} \frac{1}{r} log(P\{V > r\}) = -s_\infty \tag{F.4}$$

where $s_\infty = sup\{\theta > 0 | \lambda(\theta) \le \theta\}$
and the time unit is chosen such that the service rate is 1.

In [19] it is indicated by an inverse Laplace transformation of the double transform of the workload (formula (3.2.6) in [27]) in the time variable, that the s_∞ from the Kingman bound is indeed in agreement with the positive θ which solves $\lambda(\theta) = \theta$.

[1] we adopt the name from [90]

Superposition and thinning of identical renewal processes

Assume that a renewal process with work process $A(0,t)$ gives rize to an asymptotic decay rate s_∞, that is $\lambda_A(s_\infty) = s_\infty$. If this process is thinned by a factor n and n independent processes with work process $A_i(0,t) = A(0,t/n)$ are superposed the asymptotic decay rate function of the superposition is

$$lim_{t\to\infty} \frac{1}{t} ln(E[e^{\theta \sum_{i=1}^{n} A_i(0,t)}]) = lim_{t\to\infty} \frac{1}{nt/n} \sum_{i=1}^{n} ln(E[e^{\theta A(0,t/n)}])$$
$$= lim_{t\to\infty} \frac{1}{t} ln(E[e^{\theta A(0,t)}] = \lambda_A(\theta) \quad (F.5)$$

from which it is seen that the asymptotic decay rate of the original (single) renewal stream is the same as the asymptotic decay rate of the thinned superposition.

Equivalent Bandwidth from the Asymptotic Slope

Adopting the approximation $P(V > r) \approx e^{-s_\infty r}$, we see that $P(V > r) \leq \epsilon$ (with ϵ being the QoS requirement) is equivalent to $s_\infty \geq \frac{ln(\epsilon)}{r} =: \eta$ that is the asymptotic slope should stay above the critical value η. Consider an arrival process described by the asymptotic decay rate function $\lambda(\theta)$ and asume it is an independent sum of component arrival processes described by the rate functions $\lambda_i(\theta)$. Since the asymptotic decay rate function is additive for independent processes we get $\lambda(\theta) = \sum_i \lambda_i(\theta)$. Assume further that the capacity is 1. Define the effective bandwidth of stream i as $e_q(i) = \frac{\lambda_i(s(1))}{s(1)}$. where $s(1)$ solves $\lambda(s(1)) = s(1)$ Now $\sum_i e_q(i) < 1$ if and only if $\frac{\lambda(s(1))}{s(1)} = \sum_i \frac{\lambda_i(s(1))}{s(1)} < 1$ if and only if $\lambda(s(1)) < s(1)$ and this is equivalent to $s(1) > \eta$ due to the convexity of the asymptotic decay rate function $\lambda(\theta)$, see [90] for details on the convexity properties.

F.2.2 The Asymptotic Constant

For the M/D/1 queue where the Laplace transform $W^*(s)$ of the waiting time distribution is available, the asymptotic constant can be found as the residue

associated with the dominant real and simple pole of $W^*(s)$. The same methods can be applied for a set of other single server queues in which it is possible to obtain the transform of the waiting time distribution and the dominant pole is real and simple.

For the GI/D/1 case, it is proven in [53] that the asymptotic constant is upper and lower bounded by

$$e^{-s_\infty} \leq \alpha \leq 1 \quad (F.6)$$

which is valid for determinsitic service time 1. From this expression and from the fact that $s_\infty \to 0$ as $\rho \to 1$ it is clear that the asymptotic constant will approach 1 in heavy load. For more general arrival processes the same tendency is seen in [25].

For the general case α can be difficult to compute, but within the framework of Markovian Arrival Processes it is, by some effort, possible, see [1] for details.

In the case with traditional on/off ATM sources it is possible to apply the well known decomposition of the queueing process in cell scale and burst scale, see [56], [60], and [9]. As an approximation for the asymptotic constant use the probability that the input rate from the sources exceed the capacity of the link. In the case with identical on/off sources this gives

$$\alpha_n \approx \sum_{i > N_{max}} \frac{n!}{i!(n-i)!} a^i (1-a)^{n-i} \quad (F.7)$$

where N_{max} is the largest number of simultaneous active connections that the link can carry simultaneous, and a is the fraction of time a source is active. However, to get a more transparent expression, the Chernoff bound can be applied to obtain

$$\alpha_n \approx exp[-n(\gamma ln(\frac{\gamma}{a}) + (1-\gamma) ln(\frac{1-\gamma}{1-a}))] \quad (F.8)$$

where γ is the capacity divided by the total number of sources.

Finally, it should also be mentioned that it has been suggested as an approximation for the asymptotic constant to use

$$\alpha_n \approx s_\infty E(V) \quad (F.9)$$

This approximation follows if we assume $P(V > r) = \alpha e^{s_\infty r}$ and require the mean occupancy to be correct.

F.3 Equivalent bandwidth of a Renewal Stream

In this section we first present a heuristic formula for the complementary distribution of the unfinished work in the nGI/D/1 queue. This formula has its origin in the diffusion analysis of the M/D/1 and nD/D/1 queues. From the heuristic formula it is straigthforward to find the (heuristic) asymptotic constant, and its dependence on the number of multiplexed sources turns out to show an exponential behaviour in line with the conjecture presented as formula (1.6) in [25]. Finally, we derive closed form expressions for the effective bandwidth of a renewal stream both when it is considered alone and the asymptotic constant is put to 1, and when it is multiplexed with a number of other renewal streams with same characteristics and the asymptotic constant is computed from the heuristic method.

F.3.1 A Heuristic Formula for the Complementary Unfinished Work Distribution

In order to obtain a simple closed form approximation for the probability of the virtual waiting time exceeding a given buffer threshold in the nGI/D/1 queue, we look at the nD/D/1 and M/D/1 queues.

The classical diffusion approximation of the M/D/1 queue, see e.g. section 6.1.2. in [86], leads to

$$P\{V_{M/D/1} > r\} \approx exp(-\frac{2(1-\rho)}{\rho} r) \qquad \text{(F.10)}$$

In comparison a Brownian bridge approximation of the nD/D/1 queue, see e.g. section 6.2.2. in [86], leads to

$$P\{V_{nD/D/1} > r\} \approx exp(-\frac{2(1-\rho)}{\rho} r - \frac{2r^2}{n}) \qquad \text{(F.11)}$$

In the nD/D/1 case we see the same term as in the M/D/1 case but in addition we see a term which on the logaritmic scale decrease with the square of r. By applying the correct exponential term $s_{M/D/1}$ from solving (F.2) with $A^*(s) = \rho/(\rho+s)$ and $B^*(s) = e^s$ both approximations can be improved. The improved

form states

$$\begin{aligned} P\{V_{M/D/1} > r\} &\approx \rho exp(-s_{M/D/1} r) \quad \text{and} \\ P\{V_{nD/D/1} > r\} &\approx \rho exp(-s_{M/D/1} r - \tfrac{2r^2}{n}) \end{aligned} \qquad \text{(F.12)}$$

The multiplication of ρ ensures that the probability of an empty system is correct. Table 1 gives an assessment of the quality of approximation (F.12) in the M/D/1 case with a load of 0.6, and in the 80D/D/1 case with a load of 0.8.

Table 1: Evaluation of the improved M/D/1 and nD/D/1 formulae

buffer level	M/D/1 $\rho = 0.6$			80D/D/1 $\rho = 0.8$		
r	exact	diff.	form. (F.12)	exact	diff.	form. (F.12)
10	5.6 e-5	1.6 e-6	4.6 e-5	1.2 e-3	5.5 e-4	8.8 e-4
20	4.3 e-9	2.6 e-12	3.5 e-9	1.7 e-8	2.1 e-9	6.6 e-9
30	3.3 e-13	4.3 e-18	2.7 e-13	1.7 e-15	5.2 e-17	3.3 e-16

As the table clearly indicates then formula (F.12) represent a significant improvement compared to the usual diffusion approximation.

For the superposition of n renewal streams in which the squared coefficients of variations is c_a^2 we suggest

$$P\{V_{nGI/D/1} > r\} \approx \rho exp(-s_{M/D/1} r - (1-(c_a^2)^\beta)\frac{2r^2}{n + K_a r}) \qquad \text{(F.13)}$$

In comparison to the expression for the nD/D/1 queue the second term has been modified in two ways.

1. A term $K_a r$ is added to the usual denominator n.
   Thereby we enforce an exponential tail behaviour with rate $s_\infty = s_{M/D/1} + 2(1-(c_a^2)^\beta)/K_a$. The constants K_a and β are to be chosen such that both the correct asymptotic slope s_∞ as well as the correct mean queue length are obtained.

2. A multiplicative term $(1-(c_a^2)^\beta)$.
   This term vanish in the case with Poisson input making the expression correct in this case. The multiplicative term is 1 in the nD/D/1 case and with $K_a = 0$ our heuristic formula is seen to in accordance with formula (F.12).

It should be noticed that formula (F.13) implies that for $c_a^2 > 1$ the virtual waiting time curve for the nGI/D/1 queue will be above the corresponding M/D/1 curve and for $c_a^2 < 1$ it will be below.

F.3.2 Obtaining the Asymptotic Constant

With a tail behaviour that is asymptotically exponential with rate s_∞, the asymptotic constant is defined as the limit $\alpha = lim_{r\to\infty} exp(s_\infty r)P(Q > r)$. From formula (F.13) we get

$$\begin{aligned}
log(\alpha_n) = & \; lim_{r\to\infty}[s_\infty r + log(\rho) - s_{M/D/1}r - 2(1-(c_a^2)^\beta)\tfrac{r^2}{n+K_a r}] \\
= & \; lim_{r\to\infty}[log(\rho) + r(s_\infty - s_{M/D/1}r - \tfrac{2(1-(c_a^2)^\beta)}{K_a}\tfrac{1}{1+\frac{n}{K_a r}})] \\
= & \; lim_{r\to\infty}[log(\rho) + r(\tfrac{2(1-(c_a^2)^\beta)}{K_a}\tfrac{n/(K_a r)}{1+n/(K_a r)})] \\
= & \; log(\rho) + \tfrac{2(1-(c_a^2)^\beta)}{K_a^2}n
\end{aligned}$$

Therefore we end up with

$$\alpha_n = \rho exp[\frac{2(1-(c_a^2)^\beta)}{K_a^2}n] = \rho exp[\frac{(s_\infty - s_{M/D/1})^2}{2(1-(c_a^2)^\beta)}n] \tag{F.14}$$

since $s_\infty = s_{M/D/1} + \frac{2(1-(c_a^2)^\beta)}{K_a}$. The prediction that the intercept has an exponential dependence on the number of sources is in nice accordance with the numerical results leading to the conjecture of formula (1.6) in [25].

F.3.3 Equivalent Bandwidth of a Renewal Stream

In this section we illustrate how the result of the former section can be used to assign an effective bandwidth to a renewal stream.

Assuming the Asymptotic Constant is 1

When the asymptotic constant is put to one and the QoS requirement is $P(V > r) \le \epsilon$ an application of (F.1) yields $s_\infty \ge \frac{ln(1/\epsilon)}{r}$. In [2] an expansion of s_∞ in $1-\rho$ is given. The result for renewal input is

$$s_\infty(1-\rho) = \frac{2}{c_a^2}(1-\rho) + a_2(1-\rho)^2 + O((1-\rho)^3) \tag{F.15}$$

where a_2 is a constant involving the first three moments of the interarrival (and service) time distribution(s). Taking into acount only the first order term, corresponding to a heavy limit approximation, we obtain from $s_\infty = \frac{ln(1/\epsilon)}{r}$

$$1 = \frac{\rho}{1 - c_a^2\frac{ln(1/\epsilon)}{2r}} \approx \rho(1 + c_a^2\frac{ln(1/\epsilon)}{2r}(1 + c_a^2\frac{ln(1/\epsilon)}{2r})) \tag{F.16}$$

This is an illuminating result. The amount of non accessible capacity is seen to be proportional to the variability measure c_a^2 and the QoS requirement $ln(1/\epsilon)$ and inverse proportional to the buffer size. A complete analog result in the case of on/off sources has been derived in appendix C of [82].

Taking into Account the Asymptotic Constant

The equation $s_{\infty(1)} = \frac{ln(1/\epsilon)}{r}$ is to be substituted with $s_\infty \ge \frac{ln(1/\epsilon)}{r} + \frac{(s_\infty + s_{M/D/1})^2}{2r(1-(c_a^2)^\beta)}$ when the asymptotic constant α_n is to be taken into account. Approximating s_∞ and $s_{M/D/1}$ with their second order expansions in $1-\rho$ gives after removing terms of order three and higher rise to the following second order equation in $1-\rho$.

$$[\frac{2(c_a^2-1)^2 n}{r(1-(c_a^2)^\beta)} - a_2 c_a^4](1-\rho)^2 - 2c_a^2(1-\rho) + \frac{ln(1/\epsilon)}{r}c_a^4 = 0 \tag{F.17}$$

The solution is straightforward to obtain but difficult to use to obtain an illuminating closed form formula for the effective bandwidth. However, if we only expand s_∞ and $s_{M/D/1}$ to first order it is easier. Choosing the physically right solution we get for $c_a^2 < 1$

$$1-\rho = c_a^2\frac{r(1-(c_a^2)^\beta)}{2(c_a^2-1)^2 n}[1 - \sqrt{1 - \frac{2(c_a^2-1)^2 n}{r(1-(c_a^2)^\beta)}\frac{ln(1/\epsilon)}{r}}] \tag{F.18}$$

and applying the second order expansions $\sqrt{1+x} \approx 1 + x/2 - x^2/8$ valid for x small and $\frac{1}{1-y} \approx 1 + y + y^2$ valid for y small we get

$$1 \approx \rho[1 + c_a^2 \frac{ln(1/\epsilon)}{2r}\{1 + c_a^2 \frac{ln(1/\epsilon)}{2r} + \frac{(1-c_a^2)^2 n}{4r(1-c_a^2)^\beta)} \frac{ln(1/\epsilon)}{r}\}] \tag{F.19}$$

while for $c_a^2 > 1$ we get

$$1 - \rho = c_a^2 \frac{r((c_a^2)^\beta - 1)}{2(c_a^2-1)^2 n}[\sqrt{1 + \frac{2(c_a^2-1)^2 n}{r((c_a^2)^\beta - 1)} \frac{ln(1/\epsilon)}{r}} - 1] \tag{F.20}$$

and doing the same expansions we get

$$1 \approx \rho[1 + c_a^2 \frac{ln(1/\epsilon)}{2r}\{1 + c_a^2 \frac{ln(1/\epsilon)}{2r} - \frac{(c_a^2-1)^2 n}{4r((c_a^2)^\beta - 1)} \frac{ln(1/\epsilon)}{r}\}] \tag{F.21}$$

A comparison of formula (F.16) (F.19) and (F.21) shows that the only difference is the term $\frac{(1-c_a^2)^2 n}{4r(1-c_a^2)^\beta} \frac{ln(1/\epsilon)}{r}$ which appears with positive sign for $c_a^2 < 1$ and negative sign for $c_a^2 > 1$ illustrating how the effective bandwidth is increasing with n in case of smooth traffic while it is decreasing in n for bursty traffic.

F.4 Numerical Results

In this section we first evaluate the accuracy of the heuristic formula by comparing the results with exact results of the superposition of a number of renewal streams which can be obtained by the standard solution method for the MAP/D/1 queue when the interarrival times are of phase type.

The following three interarrival time distributions have been used.

- Erlang-2, which can be seen as a traffic model of a CBR stream which has been jittered by passing through a number of nodes in a peak rate allocated ATM network, see [89],
- a “non bursty” Interupted Poisson Process (IPP) with small mean burst size of 3.375 cells and an activity ratio of 1/3, supposed to represent an on/off source with rapid change between idle and active state.

- a bursty IPP with mean burst size of 60 and an activity ratio of 1/11, also used in [25].

F.4.1 The Heuristic Formula

We first evaluate the heuristic formula in the case of Erlang-2 renewal streams in the case with 12, 24 and 48 sources. The results are displayed in figure 1, 2 and 3.

The heuristic formula gives reasonable accurate results but an increasing overestimation in the number of sources is seen to take place. In the exact model the asymptote is reach for small buffer levels, while the heuristic model predicts a “capture” of the asymptote for much larger buffer levels.

We then evaluate the heuristic formula in the "non-bursty" IPP case, and the results are displayed in figure 4, 5 and 6.

Here the heuristic formula is quite accurate.

Finally, we evaluate the heuristic formula in the case of a bursty on/off source, and the results are displayed in figure 7, 8 and 9.

Here the heuristic formula again gets inaccurate when multiplexing many sources. Here the heuristic formula predicts the asymptote to be correct already at moderate buffer levels while the exact model tells that this will not be the case, and this causes the inaccuracy for a large number of multiplexed sources.

For all the investigated cases the heuristic approach is an upper bound, which is attractive for applications to effective bandwidth.

F.4.2 The Asymptotic Constant

Here we compare 3 different approximate ways of computing the asymptotic constant and compare it with the exact value for the superpositions of the two different IPP’s. The case with the non-bursty IPP’s and load of 0.8.

Table 2: Asymptotic constant for the non-bursty IPP

no. of sources	exact	heuristic	Fluid argument	$s_\infty E(V)$
12 IPP	0.613	0.661	0.834	0.728
24 IPP	0.447	0.502	0.696	0.615
48 IPP	0.262	0.221	0.485	0.452

We see that the heuristic approch outperforms the two other approximations.

The case with the bursty IPP's and load of 0.3.

Table 3: Asymptotic constant for the bursty IPP

no. of sources	exact	heuristic	Fluid argument	$s_\infty E(V)$
12 IPP	3.20 e-3	1.23 e-2	1.16 e-1	2.04 e-2
24 IPP	1.43 e-5	8.45 e-4	1.34 e-2	9.01 e-3
48 IPP	-	4.96 e-5	1.80 e-4	7.24 e-3

Here the heuristic approach is again better than the two others, however, for a large number of multiplexed sources it gives too large values.

F.4.3 Equivalent Bandwidth

In this section we evaluate the accuracy of different effective bandwidth formulae. The table below shows the maximum number of homogeneous sources which can be multiplexed on the same link without violating $P(V > r) \leq 10^{-9}$. The following methods are included

1. Exact method. Performs an exact computation of $P(V > r) \leq 10^{-9}$.
2. Our method taken into account the asymptotic constant and an exact evaluation of s_∞
3. Our method by using formula (F.21)
4. The method of [40], i.e. formula (2), (10) and (11) of [40]
5. The traditional big buffer method where the asymptotic constant is put to 1.

These five methods have been tested on the two different IPP's which are also used above for different bufferlevels and obtainable utilization.

Table 4: Evaluation of different effective bandwidth formulae

method	non-bursty (r=76)	non-bursty (r=180)	bursty (r=1200)	bursty (r=1660)
(1)	24	48	12	12
(2)	21	47	11	10
(3)	22	47	9	12
(4)	27	50	9	10
(5)	21	47	9	10

From the table we see that the different methods give similar results, however, it is interesting to note that the classical method of [40] underestimate the required bandwidth. We believe this is due to the fact that the fluid flow queueing model, which their method is based on (the statistical multiplexing gain not sufficient) gives a lower bound in case of interupted Poisson Processes. The differences between our own two methods in the two bursty cases originate from the low obtainable utilization and for this case a linear approximation of s_∞ as well as the expansions to obtain (F.21) are somewhat inaccurate.

F.5 Conclusion

In this paper we have, based on a heuristic formula for the virtual waiting time in the nGI/D/1 queue, derived simple illuminating formulas for the effective bandwidth of renewal streams. The virtual waiting time formula enables an estimate of the asymptotic constant from which the effective bandwidth formula was derived. Our formula gives in some cases more accurate values than the traditional methods due to the fact that the potential statistical multiplexing gain is taken into account by our method.

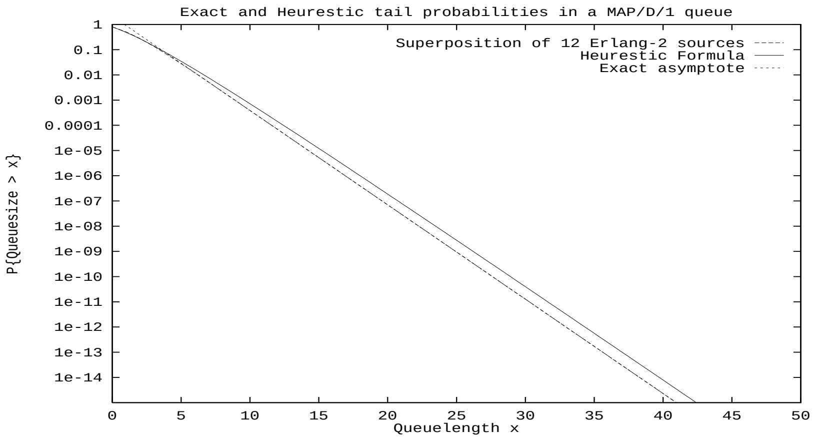

Figure F.1: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 Erlang-2 sources**

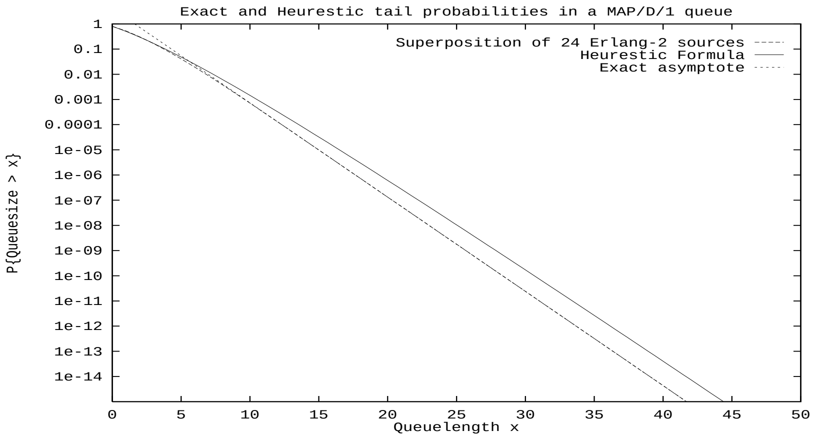

Figure F.2: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 Erlang-2 sources**

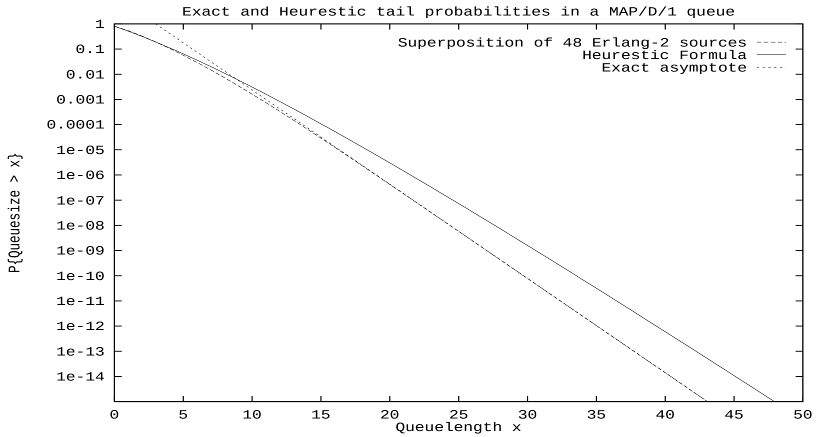

Figure F.3: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 Erlang-2 sources**

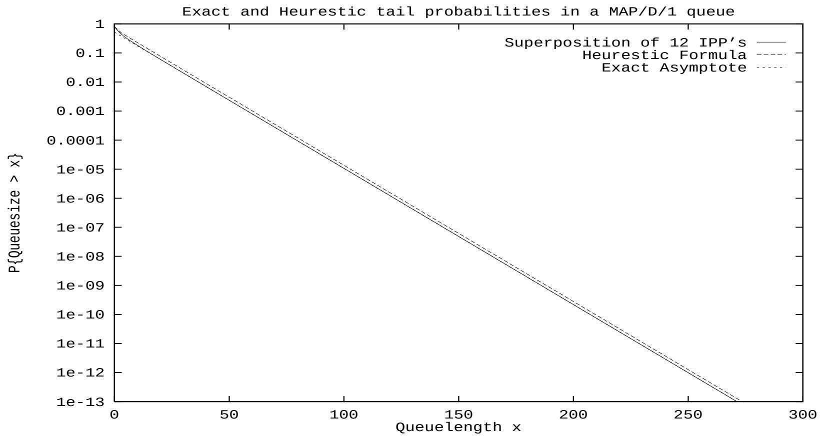

Figure F.4: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 non bursty IPP sources**

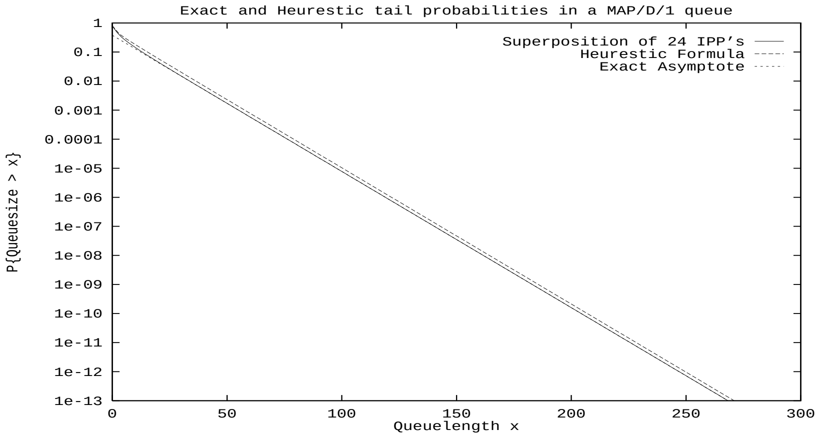

Figure F.5: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 non bursty IPP sources**

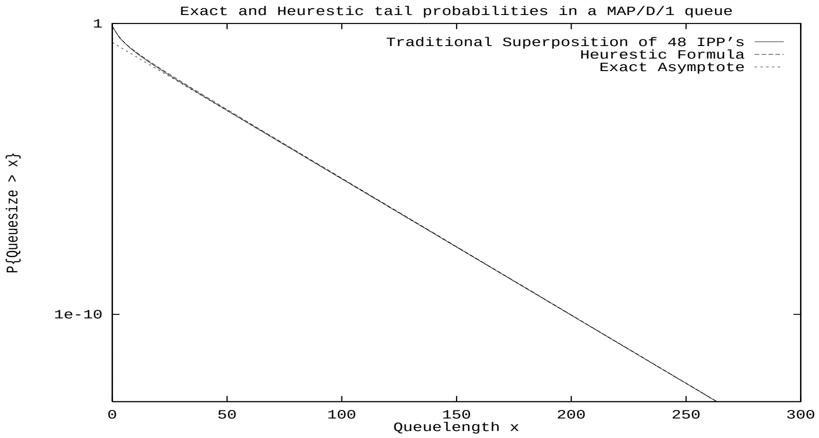

Figure F.6: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 non bursty IPP sources**

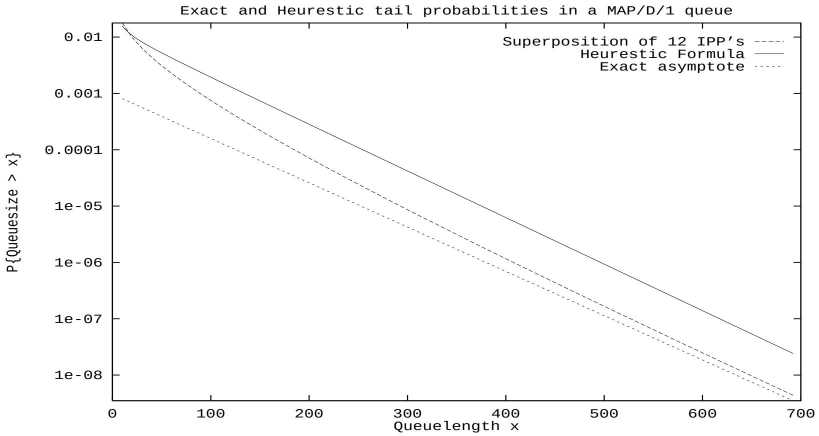

Figure F.7: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 12 bursty IPP sources**

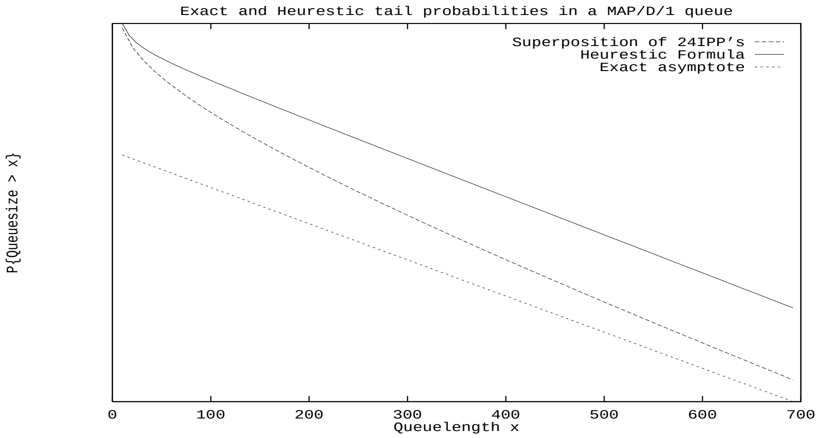

Figure F.8: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 24 bursty IPP sources**

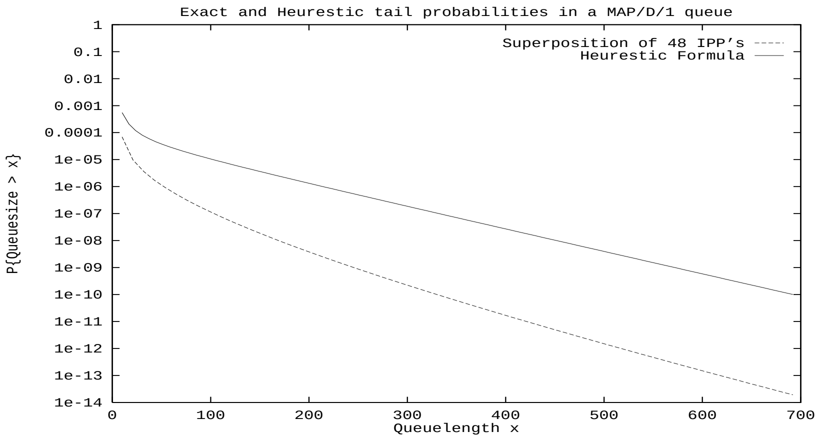

Figure F.9: **Heurestic approximation compared with the exact tail probabilities in the queue fed by a superposition of 48 bursty IPP sources**

Appendix G

A[7]

A note on the Calculation of Steady State Probabilities
in the Finite Capacity BMAP/G/1 Queue

Allan T. Andersen
Institute of Mathematical Modelling
Building 321
Technical University of Denmark
2800 DK-Lyngby

G.1 Introduction

In this work we present methods for calculating the steady state queue length distribution in the finite capacity BMAP/G/1 queue both at departure epochs, at an arbitrary time and at arrival epochs. We also present various blocking probabilities of potential interest. Finally we examine the queue length distributions in the special case of batch Poisson arrivals.

A detailed analysis of the corresponding infinite capacity queue was carried out by V. Ramaswami in [83]. In this reference the arrival process is denoted Neuts' process, N' (or the Versatile Markovian Point process) previously defined by M.F. Neuts in [74]. The Versatile Markovian Point Process or Neuts' process has later been shown to be equivalent with the Batch Markovian Arrival Process, BMAP, which has a much simpler notation [64].

A previous analysis of the finite capacity BMAP/G/1 queue (Using the Versatile Markovian Point Process notation) was carried out previously by Blondia in [20]. Here expressions for the queue length distribution at departure epochs and at an arbitrary time were derived. Additionally expressions for the virtual waiting time and the mean busy period structure were found.

This contribution consists of an alternative approach to calculate the different queue length distributions. We employ a direct method for the embedded Markov chain which is cancellation free and very easy to implement. The queue length at an arbitrary time is calculated essentially using the same approach as in the infinite case, [83], with a modification at the states associated with a full queue.

The steady state queue length distribution for the embedded Markov chain is found by a direct method. This method is a slightly modified version of the standard **LU** factorization scheme used when solving systems of linear equations. The modifications consist of exploiting the special structure of the embedded Markov chain so that the complexity of the algorithm in terms of the queue length N is of the order O(N^2) compared to the standard case of O(N^3) and additionally the standard algorithm is modified to do all operations cancellation free. This is done by employing a scheme proposed in [39]. A cancellation free scheme improves precision considerably and can be of crucial importance when reliable information about queueing behaviour is needed. In the case of the infinite BMAP/G/1 queue the cancellation free scheme proposed by Ramaswami

in [85] is considered to be a major step forward in the efficient computation of the queue length distribution for the embedded chain. Our direct approach for calculating the embedded queue length distribution avoids certain restrictions on the BMAP/G/1 queue assumed in [20]. These will later be pointed out.

The queue length distribution at an arbitrary time is calculated in the same manner as in the infinite case [83] i.e. using Markov renewal theory [24]. It turns out that the convolution in the infinite case which expresses the queuelength at an arbitrary time in terms of the queue length at embedded points [83] still holds for all but the states associated with a full queue in the corresponding finite queue. Here we also make some elementary observations regarding various blocking probabilities which may be of interest when assessing the implications of a given batch arrival process.

Finally we show that in the special case of batch Poisson arrivals it is possible to express the queue length distribution explicitly in terms of the corresponding distributions in the infinite case.

G.2 The Batch Markovian Arrival Process (BMAP) and the BMAP/G/1/N queue

In this section we will briefly sketch some main characteristics of the Batch Markovian Arrival Process (BMAP). For further results we refer to [64], [77] and [83]. The description of the BMAP follows parts of [64] closely.

The Batch Markovian Arrival Process can be constructed as a generalization of the batch Poisson process to allow for non-exponential times between arrivals of batches while still maintaining the underlying Markovian structure. $N(t)$ denotes the number of arrivals in $(0, t]$. Consider a 2-dimensional Markov process $\{N(t), J(t)\}$ on the state space $\{(i, j) : i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator **Q** having the structure

$$\mathbf{Q} = \begin{bmatrix} \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \mathbf{D}_3 & \cdots \\ & \mathbf{D}_0 & \mathbf{D}_1 & \mathbf{D}_2 & \cdots \\ & & \mathbf{D}_0 & \mathbf{D}_1 & \cdots \\ & & & \mathbf{D}_0 & \cdots \\ & & & & \ddots \end{bmatrix} \tag{G.1}$$

where $\mathbf{D}_k$, $k \geq 0$ are $m\text{x}m$ matrices, $\mathbf{D}_0$ has negative diagonal elements and nonnegative off-diagonal elements , $\mathbf{D}_k$, $k \geq 1$, are nonnegative and $\mathbf{D}$ defined by (G.2) is an irreducible infinitesimal generator. To avoid pathological processes we assume that $\sum_{k=1}^{\infty} \mathbf{D}_k \neq \mathbf{0}$.

$$\mathbf{D} = \sum_{k=0}^{\infty} \mathbf{D}_k \tag{G.2}$$

If $N(t)$ represents a counting variable and $J(t)$ a phase variable then the above Markov process (G.1) defines a batch arrival process where transitions from a state (i, j) to state $(i+k, l)$, correspond to batch arrivals of size k and hence the batch size can depend on j and l. The matrix $\mathbf{D}_0$ is a stable matrix which implies that it is non-singular and the sojourn time in the set of states $(i, j) : 1 \leq j \leq m$ is finite with probability 1.

It is readily seen that the above defined process indeed is a generalization of the batch Poisson process by examining the special case of $m = 1$. In this case all $\mathbf{D}_k$,$k \geq 0$ become scalar quantities and can be expressed in the form $\mathbf{D}_0 = -\lambda$ and $\mathbf{D}_j = \lambda p_j$, $j \geq 1$ where $\sum_{j=1}^{\infty} p_j = 1$.

At this point it is convenient to define the matrix generating function

$$\tilde{\mathbf{D}}(z) = \sum_{k=0}^{\infty} \mathbf{D}_k z^k, \, |z| \leq 1 \tag{G.3}$$

Let $\vec{\pi}$ be the stationary probability vector of the Markov process with generator $\mathbf{D}$

$$\vec{\pi}\mathbf{D} = \vec{0}, \quad \vec{\pi}\vec{e} = 1 \tag{G.4}$$

where $\vec{e}$ is, as it is throughout this paper, a column vector of 1's. The fundamental arrival rate for the arrival process, λ_a, is given by

$$\lambda_a = \vec{\pi} \textstyle\sum_{k=1}^{\infty} k\mathbf{D}_k \vec{e} = \vec{\pi}\vec{d} \tag{G.5}$$

where $\vec{d} = \sum_{k=1}^{\infty} k\mathbf{D}_k \vec{e}$

A useful property of the class of BMAPs is that it is closed under superposition.

The counting function of the BMAP is given in the following way. Let N_t denote the number of arrivals in $(0, t]$ and J_t the state of the Markov Process at time t.

Then the (i, j) entry of the $m\text{x}m$ matrices $\mathbf{P}(n, t)$ are given as

$$\mathbf{P}_{ij}(n, t) = P\{N_t = n, J_t = j | N_0 = 0, J_0 = i\} \tag{G.6}$$

The matrices $\mathbf{P}(n, t)$ satisfy the forward Chapman-Kolmogorov equations

$$\frac{d}{dt}\mathbf{P}(n, t) = \textstyle\sum_{j=0}^{n} \mathbf{P}(j, t)\mathbf{D}_{n-j}, \; n \geq 1, t \geq 0 \tag{G.7}$$

$$\mathbf{P}(0, 0) = \mathbf{I} \tag{G.8}$$

and the matrix generating function $\mathbf{P}^*(z, t) = \sum_{n=0}^{\infty} \mathbf{P}(n, t)z^n$ is explicitly given by

$$\tilde{\mathbf{P}}^*(z, t) = e^{\tilde{\mathbf{D}}(z)t} \quad |z| \leq 1, t \geq 0 \tag{G.9}$$

(The matrix exponential is defined by its series i.e. $e^{\mathbf{M}} = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}$, for properties of the matrix exponential see [16]).

Some well-known processes that are special cases of the BMAP are : the Markovian Arrival Process (MAP) (structure is BMAP but no batch arrivals) [65], the Markov Modulated Arrival Process (MMPP) [34]and [41] ,the phase-type renewal process [75], the Erlang arrival process and the Interrupted Poisson Process IPP [59].

In the following we sketch the standard embedded approach used when examining the BMAP/G/1/N queue. We follow [64], where the corresponding infinite queue is examined, closely. We also draw upon [83].

The arbitrary service time distribution is denoted by $H(t)$ with finite mean μ_1. We assume subsequent service times to be independent and that the arrival process and service time distributions are independent.

As standard when solving the M/G/1 queue the approach is here to study the embedded Markov renewal process at departure epochs. Define τ_k to be the epoch of the k-th departure from the queue, with $\tau_0 = 0$, and (ξ_k, J_k) to be respectively the number in system and the phase of the arrival process at $\tau_k{}^+$. Then $(\xi_k, J_k, \tau_{k+1} - \tau_k)$ is a semi-Markov process on the state space $\{(i,j) : N-1 \geq i \geq 0, 1 \leq j \leq m\}$ (It is trivial here to note that at departure epochs you NEVER have a full queue i.e. N customers)

The transition probability matrix is given by

$$\mathbf{P}(x) = \begin{bmatrix} \mathbf{B}_0(x) & \mathbf{B}_1(x) & \mathbf{B}_2(x) & \dots & \mathbf{B}_{N-2}(x) & \sum_{k=N-1}^{\infty} \mathbf{B}_k(x) \\ \mathbf{A}_0(x) & \mathbf{A}_1(x) & \mathbf{A}_2(x) & \dots & \mathbf{A}_{N-2}(x) & \sum_{k=N-1}^{\infty} \mathbf{A}_k(x) \\ \mathbf{0} & \mathbf{A}_0(x) & \mathbf{A}_1(x) & \dots & \mathbf{A}_{N-3}(x) & \sum_{k=N-2}^{\infty} \mathbf{A}_k(x) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{A}_0(x) & \sum_{k=1}^{\infty} \mathbf{A}_k(x) \end{bmatrix} \tag{G.10}$$

where , $x \geq 0$, $n \geq 0$, $\mathbf{A}_n(x)$ and $\mathbf{B}_n(x)$ are mxm matrices of mass functions defined by

$[\mathbf{A}_n(x)]_{ij}$ is the probability that given a departure at time 0, which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during that service there were n arrivals.

$[\mathbf{B}_n(x)]_{ij}$ is the probability that given a departure at time 0, which left the system empty and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during the service period of the departing customer there were n arrivals.

From the earlier definition of $\mathbf{P}(n,t)$ it is clear that

$$\mathbf{A}_n(x) = \int_{t=0}^{x} \mathbf{P}(n,t) dH(t) \tag{G.11}$$

and that

$$\mathbf{B}_n(x) = \sum_{k=1}^{n+1} \int_{t=0}^{x} e^{\mathbf{D}_0 t} \mathbf{D}_k \mathbf{A}_{n-k+1}(x-t) dt \tag{G.12}$$

Also we define $\mathbf{A}_n(\infty) = \mathbf{A}_n$ and $\mathbf{B}_n(\infty) = \mathbf{B}_n$.

We define the following transform matrices :

$$\tilde{\mathbf{A}}(z) = \sum_{n=0}^{\infty} \mathbf{A}_n z^n \tag{G.13}$$

(G.14)

$$\tilde{\mathbf{B}}(z) = \sum_{n=0}^{\infty} \mathbf{B}_n z^n = z^{-1}(-\mathbf{D}_0{}^{-1})(\tilde{\mathbf{D}}(z) - \mathbf{D}_0)\tilde{\mathbf{A}}(z) \tag{G.15}$$

The matrix $\mathbf{A} = \tilde{\mathbf{A}}(1)$ is a stochastic matrix and the stationary vector $\vec{\pi}$ defined in (G.4) also satisfies $\vec{\pi}\mathbf{A} = \vec{\pi}$, $\vec{\pi}\vec{e} = 1$.

G.3 Steady state queuelength distribution at departures

It is clear from (G.10) that the stationary probability vector $\vec{x} = [\vec{x}_0, \vec{x}_1, ..., \vec{x}_{N-1}]$, where $\vec{x}_k = [x_{k1}, x_{k2}....x_{km}]$, of $\mathbf{P} = \mathbf{P}(\infty)$ satisfies

$$\vec{x}\mathbf{P} = \vec{x} \Leftrightarrow \vec{x}(\mathbf{P} - \mathbf{I}) = \vec{x}\mathbf{R} = \vec{0}, \quad \vec{x}\vec{e} = 1 \tag{G.16}$$

or equivalently

$$\mathbf{V}\vec{u} = \vec{0} \tag{G.17}$$

where $\mathbf{V} = \mathbf{R}^T(\mathbf{R}_{diag})^{-1} = (\mathbf{P} - \mathbf{I})^T(\mathbf{R}_{diag})^{-1}$, $\vec{u} = \mathbf{R}_{diag}\vec{x}^T$ and $\mathbf{R}_{diag} = diag(\mathbf{R}_{11}, \mathbf{R}_{22}, .., \mathbf{R}_{(Nm)(Nm)})$.

Here it can be noted that

$$\vec{x}_i = \vec{x}_0\mathbf{B}_i + \sum_{l=1}^{i+1} \vec{x}_l\mathbf{A}_{i+1-l} \tag{G.18}$$

holds for $0 \leq i \leq N - 2$. In the corresponding infinite case this holds for all $i \geq 0$.

It is easily seen that in $\mathbf{V}$ all diagonal elements are 1 i.e. $\mathbf{V}_{ii} = 1$ and that all off diagonal elements satisfy $\mathbf{V}_{ij} \leq 0$. It is also readily seen that all column sums in $\mathbf{V}$ are 0 or more formally $\sum_{i=1}^{Nm} \mathbf{V}_{ij} = 0$ for $1 \leq j \leq Nm$. It is well known that an irreducible matrix possessing these properties has a **LU** factorization i.e. there exists a unique pair of matrices **L** and **U**, where **L** is lower triangular and **U** is upper triangular, such that $\mathbf{V} = \mathbf{LU}$ (See for example [18]). Additionally **L** is a non-singular matrix and **U** is a singular matrix with elements $\mathbf{U}_{ii} \geq 0, i = 1, Nm-1$ and $\mathbf{U}_{(Nm)(Nm)} = 0$. The **L** and **U** factorizations of **V** are obtained by straight forward Gaussian elimination without pivoting.

We have that

$$\mathbf{V}\vec{u} = \vec{0} \Leftrightarrow \mathbf{LU}\vec{u} = \vec{0} \Leftrightarrow \mathbf{U}\vec{u} = \vec{0} \tag{G.19}$$

so given the **LU** factorization it is only necessary to solve the upper triangular system. This is done by setting $\vec{u}_{Nm} = 1$ and calculating the elements in $\vec{u}$ recursively using "backward substitution". The vector $\vec{u}$ is then determined up to a multiplicative constant. From this $\vec{x}$ is easily obtained.

The special structure of **V** makes it possible to do the Gaussian elimination procedure cancellation free. To see this we first examine the straight forward Gaussian elimination procedure. We use a pseudo Pascal syntax.

```
for k:=1 to Nm-1 do
    for i:=k + 1 to Nm do
        c := V[i][k]/V[k][k]
        V[i][k] := c
        for j := k + 1 to Nm do
            V[i][j] := V[i][j] − cV[k][j]
```

As mentioned earlier the column sums in **V** are zero. When examining the Gaussian elimination procedure it can be noted that at the beginning of the k-th step, $k = 1, 2.., Nm - 1$ this property still holds i.e. at the beginning of the k-th step the matrix stored in the lower right $(Nm - k + 1)$x$(Nm - k + 1)$ corner of **V** has column sums equal to zero.

Because of the sign distribution in **V** it is clear that cancellations in the Gaussian elimination procedure only occur for diagonal elements. However, due to the fact that the column sums are zero at the beginning of the k-th step, the diagonal elements can be calculated without cancellation. This is done by inserting the following piece code of at the beginning of the k-th loop.

```
c:=0
    for i:=k + 1 to Nm do
        c := c + V[i][k]
    V[k][k] := −c
```

Hence by a simple alteration of the standard Gaussian elimination procedure the computations can be done cancellation free.

This cancellation free scheme for calculating steady state probabilities in a Markov Chain was proposed in [39].

In the above description of the general approach for determining the steady state probabilities for the embedded chain we have followed the description given in [14].

The special structure of **V** can be exploited to reduce both the computational complexity and storage requirements significantly. From (G.10) it easily seen that **V** has the following form.

$$\mathbf{V} = \begin{bmatrix} \mathbf{F}_0 & \mathbf{E}_0 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{F}_1 & \mathbf{E}_1 & \mathbf{E}_0 & \dots & \mathbf{0} \\ \mathbf{F}_2 & \mathbf{E}_2 & \mathbf{E}_1 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{F}_{N-2} & \mathbf{E}_{N-2} & \mathbf{E}_{N-3} & \dots & \mathbf{E}_0 \\ \mathbf{F}_{N-1} & \mathbf{G}_{N-1} & \mathbf{G}_{N-2} & \dots & \mathbf{G}_1 \end{bmatrix} \tag{G.20}$$

Taking advantage of all the zeros above the diagonal can be done by restricting the inner j loop in the Gaussian elimination procedure only to take place when $\mathbf{V}[k][j] \neq 0$. That **U** has a band structure can also be exploited when doing the "backward substitution". The standard complexity when solving a system of linear equations is of the order O(m^3N^3) but exploiting the zeros in (G.20) immediately reduces the complexity to the order O(m^3N^2), i.e. quadratic in the queue size. In the case of the infinite BMAP/G/1 queue it is standard to set $\mathbf{A}_i$ and $\mathbf{B}_i$ to $\mathbf{0}$ when $\sum_{k=i}^{\infty} \mathbf{A}_i$ and $\sum_{k=i}^{\infty} \mathbf{B}_i$ have negligible small components. In (G.20) this corresponds to setting $\mathbf{F}_i$, $\mathbf{E}_i$ and $\mathbf{G}_i$ to $\mathbf{0}$ for all $i \geq k$. For a (very!) large queue this may yield a band structure in (G.20) in which case the complexity grows linearly in terms of the queuelength. Implementationwise the special structure of the embedded chain can be used to reduce the storage requirements for the sketched **LU** factorization scheme to approximately $4Nm^2$ compared to standard N^2m^2.

Another very important aspect of using this factorisation scheme is that if that factorisation is known for queue size N it is extremely cheap to calculate the factorisation for queuelength $N+1$ since most of the factorisation already obtained can be reused. This observation can yield substantial computational savings

when for example trying to dimension a buffer for a given target probability. This observation has also been made in related work by Blondia & Casal [21].

The approach used in [20] to find the steady state probability vector of the embedded chain requires the matrix $\mathbf{A}_0$ to be non-singular. Our approach does not put this assumption on $\mathbf{A}_0$.

G.4 Steady state queuelength at an arbitrary time

Applying Markov renewal theory and the Key Renewal Theorem [24] we use the same approach as in [83] to express the stationary queue length distribution at an arbitrary time in terms of the stationary queue length distribution at departures. We use this approach for the queue lengths $0, N-1$. The probabilites of being in each of the states associated with queue length N are readily determined from known probabilites associated with the queue lengths $0, N-1$ and the stationary probability vector $\vec{\pi}$)

Let ξ_t denote the queue length and J_t denote the phase of the arrival process at time t. The time dependent distribution of the joint distribution of queue length and arrival phase is given by the conditional probabilities

$$y(k,j)_t = P\{\xi_t = k, J_t = j | \xi_0 = k_0, J_0 = j_0\} \tag{G.21}$$

for $N \geq k \geq 0, 1 \leq j \leq m, t \geq 0$. It can be shown that the limits $y(k,j)_t$, $N \geq k \geq 0, 1 \leq j \leq m$, for $t \to \infty$ exist. We denote these limits y_{kj} and define $\vec{y}_k = [y_{k1}, y_{k2} \dots y_{km}]$.

To apply the Key Renewal Theorem it is essential to know the mean recurrence time of each state.

First the sojoun times in each of Nm states in the embedded chain are easily seen to be

$$\vec{\delta}_0 = \int_{t=0}^{\infty} t \sum_{k=0}^{\infty} \mathbf{B}_k(t) dH(t) = (-\mathbf{D}_0)^{-1}\vec{e} + \mu_1 \vec{e} \tag{G.22}$$

$$\vec{\delta}_i = \int_{t=0}^{\infty} t \sum_{k=0}^{\infty} \mathbf{A}_k(t) dH(t) = \mu_1 \vec{e} \qquad 0 < i \leq N-1 \tag{G.23}$$

Where $\vec{\delta}_k = [\delta_{k1}, \delta_{k2}, ..., \delta_{km}]'$ is the vector of mean sojourn times in state k. Additionally we define $\vec{\delta} = \left[\vec{\delta}_0, \vec{\delta}_1, ..., \vec{\delta}_{N-1}\right]'$.

Using limiting properties of the Markov renewal equation (see for example [24] p.152 - 157) we have that the inverse of the mean reccurrence times m_{ij}, $1 \leq j \leq m$ and $0 \leq i \leq N-1$ can be expressed as

$$m_{ij}^{-1} = (\vec{x}\vec{\delta})^{-1} x_{ij} = (\vec{x}_0(-\mathbf{D}_0)^{-1}\vec{e} + \mu_1)^{-1} x_{ij} = \mu_{sj}^{-1} x_{ij} \tag{G.24}$$

μ_{sj} can interpreted as the mean time between transitions in the embedded Markov chain.

Applying the Key Renewal Theorem p. 153 in [24] in exactly the same manner as in [83] to the states associated with an empty system gives

$$\vec{y}_0 = \mu_{sj}^{-1}\vec{x}_0 \int_{t=0}^{\infty} e^{\mathbf{D}_0 t} dt = \mu_{sj}^{-1}\vec{x}_0(-\mathbf{D}_0)^{-1} \tag{G.25}$$

Before applying the Key Renewal Theorem to the states associated with $1, 2, ..N-1$ customers in the system we make the following definitions to be somewhat consistent with [83].

$$\mathbf{U}_n(x) = \int_{t=0}^{x} e^{\mathbf{D}_0 t} dt \mathbf{D}_n \tag{G.26}$$

and obviously $\mathbf{U}_n = \mathbf{U}_n(\infty) = (-\mathbf{D}_0)^{-1}\mathbf{D}_n$

Inserting (G.26) in (G.12) gives

$$\mathbf{B}_n(x) = \sum_{k=1}^{n+1} (\mathbf{U}_k * \mathbf{A}_{n-k+1})(x) \tag{G.27}$$

where * denotes the convolution operator.

Using the Key Renewal Theorem exactly as in [83] we have for $1 \leq i \leq N-1$

$$\vec{y}_i = \mu_{sj}^{-1} \sum_{k=1}^{i} (\vec{x}_0\mathbf{U}_k + \vec{x}_k) \int_{t=0-}^{\infty} (1 - H(t))\mathbf{P}(i-k, t) dt \tag{G.28}$$

From (G.9),(G.11) and (G.13) it is readily seen that

$$\tilde{\mathbf{A}}(z) = \int_{t=0}^{\infty} e^{\tilde{\mathbf{D}}(z)t} dH(t) \quad |z| \leq 1 \tag{G.29}$$

Using partial integration it can be derived that for $|z| < 1$ the following holds

$$(\tilde{\mathbf{A}}(z) - \mathbf{I})(\tilde{\mathbf{D}}(z))^{-1} = \int_{t=0-}^{\infty} (1 - H(t)) e^{\tilde{\mathbf{D}}(z)t} dt \tag{G.30}$$

The different terms in the z transform yield

$$\mathbf{A}_0 - \mathbf{I} = \int_{t=0-}^{\infty} (1 - H(t))\mathbf{P}(0, t) dt \mathbf{D}_0 \tag{G.31}$$

and for all $i > 0$

$$\mathbf{A}_i = \sum_{k=0}^{i} \int_{t=0-}^{\infty} (1 - H(t))\mathbf{P}(k, t) dt \mathbf{D}_{i-k} \tag{G.32}$$

Using (G.30),(G.32), (G.28) and (G.18) the sum $\sum_{j=1}^{i} \vec{y}_{i-j}\mathbf{D}_j$, $1 \leq i \leq N-1$, can be found as

$$\sum_{j=1}^{i} \vec{y}_j \mathbf{D}_{\,i-j}$$

$$\begin{aligned}
&= \mu_{sj}^{-1} \textstyle\sum_{j=1}^{i} \sum_{k=1}^{j} (\vec{x}_0 \mathbf{U}_k + \vec{x}_k) \int_{t=0-}^{\infty} (1 - H(t)) \mathbf{P}(j-k,t) dt \mathbf{D}_{\,i-j} \\
&= \mu_{sj}^{-1} \textstyle\sum_{k=1}^{i} \sum_{j=k}^{i} (\vec{x}_0 \mathbf{U}_k + \vec{x}_k) \int_{t=0-}^{\infty} (1 - H(t)) \mathbf{P}(j-k,t) dt \mathbf{D}_{\,i-j} \quad \text{(G.33)} \\
&= \mu_{sj}^{-1} \textstyle\sum_{k=1}^{i} (\vec{x}_0 \mathbf{U}_k + \vec{x}_k) (\sum_{j=0}^{i-k} \int_{t=0-}^{\infty} (1 - H(t)) \mathbf{P}(j,t) dt \mathbf{D}_{\,i-j-k}) \\
&= \mu_{sj}^{-1} (-\vec{x}_0 \mathbf{U}_i - \vec{x}_i + \textstyle\sum_{k=1}^{i} (\vec{x}_0 \mathbf{U}_k + \vec{x}_k) \mathbf{A}_{i-k}) \\
&= \mu_{sj}^{-1} (\vec{x}_{i-1} - \vec{x}_i - \vec{x}_0 \mathbf{U}_i)
\end{aligned}$$

Hence using (G.25) it is clear that

$$\sum_{j=0}^{i} \vec{y}_j \mathbf{D}_{\,i-j} = \mu_{sj}^{-1} (\vec{x}_{i-1} - \vec{x}_i), \qquad 1 \le i \le N-1 \qquad \text{(G.34)}$$

which exactly the same type of convolution as in [83]. Using the stationary probability vector of $\mathbf{D}$ the probability vector $\vec{y}$ can be found recursively

$$\begin{aligned}
\vec{y}_0 &= \mu_{sj}^{-1} \vec{x}_0 (-\mathbf{D}_0)^{-1} \\
\vec{y}_i &= ((\textstyle\sum_{j=0}^{i-1} \vec{y}_j \mathbf{D}_{\,i-j}) - \mu_{sj}^{-1} (\vec{x}_{i-1} - \vec{x}_i))(-\mathbf{D}_0)^{-1}, 1 \le i \le N-1 \quad \text{(G.35)} \\
\vec{y}_N &= \vec{\pi} - \textstyle\sum_{j=0}^{N-1} \vec{y}_j
\end{aligned}$$

As mentioned earlier the method for obtaining $\vec{y}_i$ for $0 \le i \le N-1$ is the same for the infinite case. However, in the infinite case μ_{sj} is equal to $1/\lambda_a$. This is not surprising since μ_{sj} denotes the mean time between departures of the queue which of course in the infinite case is the inverse of the fundamental rate. In the finite case the same argument gives that μ_{sj} is the inverse of the fundamental rate of the carried traffic. From this the rejected traffic can easily be found. More precisely we find the percentage of blocked arrivals to be.

$$p_{block} = \lambda_a^{-1} (\lambda_a - \mu_{sj}^{-1}) = 1 - \lambda_a^{-1} (\vec{x}_0 (-\mathbf{D}_0)^{-1} \vec{e} + \mu_1)^{-1} \qquad \text{(G.36)}$$

We have now found expressions for the queue length distribution at an arbitrary time. However, the queue length distribution seen by arrivals seems to be of greater interest. Having a batch arrival process it might be of interest both to know the queue length distribution seen by an arbitrary arrival in a batch and the queue length distribution seen by an arbitrary batch.

Let the vector $\vec{z}^s = \left[\vec{z}_0^s, \vec{z}_1^s, \vec{z}_2^s \ldots, \vec{z}_N^s\right]$ denote the queue length distribution seen by an arbitrary arrival. (In this context batch arrivals can be perceived as parallel arrivals).

$$\vec{z}_i^s = (\lambda_a)^{-1} \vec{y}_i \textstyle\sum_{k=1}^{\infty} k \mathbf{D}_k, \quad 0 \le i \le N \qquad \text{(G.37)}$$

This approach provides a (more complicated) formula for computing the blocking probability previously found in (G.36).

$$z_{block}^s = p_{block} = (\lambda_a)^{-1} \sum_{i=0}^{N} \vec{y}_i (\sum_{k=N+1-i}^{\infty} (k - (N-i)) \mathbf{D}_k) \vec{e} \qquad \text{(G.38)}$$

which for the special case of batch size 1, i.e. the MAP, gives $z_{block}^s = (\lambda_a)^{-1} \vec{y}_N \mathbf{D}_1 \vec{e}$

Assuming that the batches are ordered the conditional blocking $z_{cond\,block}^s(k)$ conditioned on the position k in a batch can for the non-trivial cases be found as

$$z_{cond\,block}^s(k) = (\vec{\pi} \bar{\mathbf{D}}_k \vec{e})^{-1} (\sum_{i=N-k+1}^{N} \vec{y}_i) \bar{\mathbf{D}}_k \vec{e}, 1 \le k \le N, \bar{\mathbf{D}}_k \ne \mathbf{0} \qquad \text{(G.39)}$$

where $\bar{\mathbf{D}}_k = \sum_{i=k}^{\infty} \mathbf{D}_k$

The blocking probability $z^s_{block}(k)$ for the position k in a batch can for the non-trivial cases be found as

$$z^s_{block}(k) = (\textstyle\sum_{i=N-k+1}^{N} \vec{y}_i)\bar{\mathbf{D}}_k\vec{e}, \quad 1 \le k \le N, \bar{\mathbf{D}}_k \neq \mathbf{0} \tag{G.40}$$

Let the vector $\vec{z}^b = \left[\vec{z}^b_0, \vec{z}^b_1, \vec{z}^b_2 \ldots, \vec{z}^b_N\right]$ denote the queue length distribution seen by an arbitrary batch of arrivals.

$$\vec{z}^b_i = (\vec{\pi} \textstyle\sum_{k=1}^{\infty} \mathbf{D}_k\vec{e})^{-1}\vec{y}_i \sum_{k=1}^{\infty} \mathbf{D}_k, \quad 0 \le i \le N \tag{G.41}$$

G.5 The queue length for the special case of the M^x/G/1/N queue

The special case of batch Poisson arrivals has a particularly simple structure which makes it possible to calculate the queue length distributions for the finite case very easily if they are known for the corresponding infinite case. (Here it is of course only of interest to look at queues with load $\rho < 1$).

A key observation is that the embedded chain for batch Poisson arrivals has the exact same structure as for Poisson arrivals. It is well known [50] that for Poisson arrivals the stationary probability vector of the embedded chain for the finite queue can be found by scaling the corresponding probabilities from the infinite queue. Since the structure is the same this also holds for batch Poisson arrivals i.e

$$x_i = \frac{x_i^{\infty}}{(\sum_{k=0}^{N-1} x_k^{\infty})} \quad 0 \le i \le N-1 \tag{G.42}$$

where x_i^{∞} denote the probabilites in the corresponding infinite queue.

By inserting the batch Poisson arrival process $\mathbf{D}_0 = -\lambda$ and $\mathbf{D}_j = \lambda p_j, j \ge 1$ where $\sum_{j=1}^{\infty} p_j = 1$ in (G.36) we obtain.

$$\begin{aligned} y_0 &= \mu_{sj}^{-1} x_0(\lambda)^{-1} \\ y_i &= (\textstyle\sum_{j=0}^{i-1} y_j p_{i-j}) - \mu_{sj}^{-1}(x_{i-1} - x_i)(\lambda)^{-1}, \quad 1 \le i \le N-1 \\ y_N &= 1 - \textstyle\sum_{j=0}^{N-1} y_j \end{aligned} \tag{G.43}$$

Where $\mu_{sj} = \mu_1 + x_0/\lambda$. As before μ_1 denotes the mean service time. The fundamental arrival rate of the batch Poisson process is readily seen to be $\lambda_a = \lambda \sum_{j=1}^{\infty} j p_j$.

For the corresponding infinite case we have (see [64] or [83])

$$\begin{aligned} y_0^{\infty} &= \lambda_a x_0^{\infty}(\lambda)^{-1} \\ y_i^{\infty} &= (\textstyle\sum_{j=0}^{i-1} y_j^{\infty} p_{i-j}) - \lambda_a(x_{i-1}^{\infty} - x_i^{\infty})(\lambda)^{-1}, \quad 1 \le i \end{aligned} \tag{G.44}$$

Hence it is clear that the queue length probabilities $y_0, y_1 \ldots y_{N-1}$ can be obtained from the corresponding infinite queue. In the infinite case $\rho = \lambda_a\mu_1$ which implies that $y_0 = 1 - \lambda_a\mu_1$. In the finite case the carried traffic is $(\mu_{sj})^{-1}\mu_1$

It easily verified that

$$\begin{aligned} y_0 &= K y_0^{\infty} \\ y_i &= K y_i^{\infty} \quad 1 \le i \le N-1 \\ y_N &= 1 - \textstyle\sum_{j=0}^{N-1} y_j \end{aligned} \tag{G.45}$$

where $K = 1/(\lambda_a(x_0^{\infty}/\lambda + \mu_1 \sum_{k=0}^{N-1} x_k^{\infty}))$

G.6 Concluding remarks

As mentioned in the introduction Blondia also finds expressions for the virtual waiting time in [20]. It is therefore tempting to see if applying Markov renewal theory and the Key Renewal Theorem in this context leads to simplifications.

Trying this approach , however, we seem to end up with the same type of expressions as in [20].

It is of independent interest to compare the complexity of the algorithms for calculating the queue length distributions for the finite and infinite queue. In the infinite case the method for finding the stationary probability vector of the embedded chain involves a number of iterations to obtain a certain matrix, **G**, with sufficient precision. (For further details about **G** we refer to [64],[77] or [83]) A rough estimate is that when the number of iterations needed to obtain G with sufficient precision is approximately the same as the queue size N the two algorithms have similar complexity. Fewer iterations tend to favour the algorithm, for the infinite queue, more iterations tend to favour the algorithm for the finite queue. In practise it is well known that the iterative schemes currently available for calculating the **G** matrix require many iterations when the queue is heavy loaded or conditionally overloaded. In these case the number of iterations needed can easily be of the order 10^3 , 10^4 or more. By conditional overload we mean that the BMAP arrival process overloads the queue in some of its phases and has sojourn times in these phases that are long enough to yield a substantial build up in the queue. This has a very close relationship with the caudal curve of queues examined by M.F. Neuts in [76]. In these cases, where the queueing behaviour is adverse, the finite BMAP queue is a feasible alternative to the infinite queue. In lightly loaded queues without these conditional overload phenomenas the algorithm for the infinite case clearly performs better than the corresponding for the finite case.

Due to the very close relationship between the expressions for the queue length at an arbitrary time for the finite and infinite queue it is a relatively small task to modify an existing piece of computer code for the infinite queue to handle the finite case. The main change is the inclusion of a piece of code that can find the embedded chains stationary probability vector. Very simple modifications of the standard **LU** decomposition scheme yield a cancellation free code that exploits the structure of the embedded chain so this should be no major difficulty.

In the context of modelling Broadband ISDN (B-ISDN) with very strict loss requirements (cell loss probabilities must be less then 10^{-9}) analyzing the finite capacity queue seems to be more relevant than the corresponding infinite queue.

Bibliography

[1] Joseph Abate, Gagan L. Choudhury, and Ward Whitt. Asymptotics for Steady-state Tail Probabilities in Structured Markov Queueing Models. *Commun. Statist. -Stochastic Models*, 10(1):99–143, 1994.

[2] Joseph Abate and Ward Whitt. A heavy-traffic expansion for asymptotic decay rates of tail probabilities in multichannel queues. *Operations Research Letters*, 15:223–230, 1994.

[3] Jacqueline Akinpelu and Ronald Skoog. Controlling Transients and Overloads in Common Channel Signalling Networks. In Minoru Akiyama (Editor), editor, *Teletraffic Issues in an Advanced Information Society, ITC-11*, pages 996–1002. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1985.

[4] Tayfur Altiok. On the Phase-Type Approximations of General Distributions. *IIE Transactions*, 17(2):110–116, 1985.

[5] Allan T. Andersen. Measurement studies in AT&T's Signalling System nr 7. network. Technical report (to appear), Institute of Mathematical Modelling, Technical University of Denmark, DK - 2800, Lyngby, 1995.

[6] Allan T. Andersen, Alex Jensen, and Bo Friis Nielsen. Modelling and performance study of packet-traffic with self-similar characteristics over several timescales with Markovian Arrival Processes (MAP). In Ilkka Norros & Jorma Virtamo, editor, *Nordic Teletraffic Seminar, NTS 12, VTT Symposium 154*, pages 269–283, Espoo, Finland, aug 1995. NTS, VTT.

[7] Allan T. Andersen, Alex Jensen, and Bo Friis Nielsen. Modelling of apparantly self-similar packet arrival processes with Markovian Arrival Processes (MAP). In *COST 242 TD(95)005*, pages 1–21. COST, jan 1995.

[8] Allan T. Andersen and Bo Friis Nielsen. An experimental study of Queueing Performance of MAP's with apparently self-similar behaviour. In *COST 242 TD(95)027*, pages 1–23. COST, may 1995.

[9] C. Rasmussen anf J. Sørensen, K. Kvols, and S.B. Jacobsen. Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks. *IEEE Journal On Selected Areas In Communications*, 9(3):351–358, apr 1991.

[10] D. Anick, D. Mitra, and M.M. Sondhi. Stochastic Theory of a Data-Handling System with Multiple Sources. *The Bell System Technical Journal*, 61(8):1871–1894, Jan 1982.

[11] Søren Asmussen and Ger Koole. Marked Point Processes as Limits of Markovian Arrival Streams. *J. Appl. Prob.*, 30:365–372, 1993.

[12] AT&T (Associate Rapporteur for Question 12/2). *Measurements of Signalling Traffic Variability and Self-Similarity*. ITU, Geneva, Switzerland, Dec. 1994. Study Group 2, Temporary Document 3/2-111-E.

[13] AT&T (Associate Rapporteur for Question 12/2). *A Simulation Study of the Effect of Long Messages on the Queueing Delays in a Traffic Stream*. ITU, Geneva, Switzerland, Jun. 1994. Study Group 2, Delayed Contribution EDIN TD06, Edinburgh Q12/2.

[14] V.A. Barker. Numerical Solution of Sparse Singular Systems of Equations Arising From Ergodic Markov Chains. *Commun. Statist.-Stochastic Models*, 5(3), 1989.

[15] Jens Barnow. Analysis of SS7 Traffic measurements i KTAS' network (in Danish). IMSOR-EKS-1994-14 (Masters Thesis), jan 1994.

[16] R. Bellman. *Introduction to Matrix Analysis*. McGraw-Hill, New York, 1960.

[17] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Variable-Bit-Rate Video Traffic and Long-Range Dependence. *IEEE Transactions on Communications (subject to revisions)*, 1993.

[18] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

[19] Søren Blaabjerg, Allan T. Andersen, and Bo Friis Nielsen. Intercept and Asymptotic Behaviour of the nGI/G/I queue. In Ilkka Norros & Jorma Virtamo, editor, *Nordic Teletraffic Seminar, NTS 12, VTT Symposium 154*, pages 167–180, Espoo, Finland, aug 1995. NTS, VTT.

[20] C. Blondia. The $N/G/1$ Finite Capacity Queue. *Commun. Statist.-Stochastic Models*, 5(2):273–294, 1989.

[21] C. Blondia and O. Casals. Statistical multiplexing of VBR sources: A matrix-analytic approach. *Performance Evaluation*, 16:5–20, 1992.

[22] E. Brockmeyer, H.L. Halstrøm, and Arne Jensen. *The Life and Works of A.K. Erlang*. Academy of Technical Sciences, Copenhagen, 1948.

[23] J.A. Bucklew. *Large Deviation Techniques in Decision, Simulation and Estimation*. John Wiley & Sons, New York, 1990.

[24] Erhan Çinlar. Markov Renewal Theory. *Adv.Appl.Prob.*, 1:123–187, 1969.

[25] Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt. On the Effectiveness of Effective Bandwidths for Admission Control in ATM networks. In J. Labetoulle and J. W. Roberts (Editors), editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, ITC-14*, pages 411–420. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1994.

[26] Gagan L. Choudhury and Ward Whitt. Heavy-Traffic Asymptotic Expansions for the Asymptotic Decay Rate in the $BMAP/G/1$ Queue. *Commun. Statist. -Stochastic Models*, 10(2):453–498, 1994.

[27] D.R. Cox. *Renewal Theory*. Methuen, London, 1962.

[28] N.G. Duffield and Neil O'Connell. Large Deviations and overflow probabilities for the general single-server queue, with applications. Technical Report DIAS-STP-93-30, Dublin Institute for Advanced Studies, Ireland, 1993.

[29] D.E. Duffy, A.A. McIntosh, M. Rosenstein, and W. Willinger. Statistical Analysis of CCSN/SS7 Traffic Data from Working CCS Subnetworks. *IEEE Journal On Selected Areas In Communications*, 12(3):544–551, april 1994.

[30] A.I. Elwalid and D. Mitra. Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. *IEEE/ACM Transactions on Networking*, 1(3):329–343, 1993.

[31] A. Erramilli, R.P. Singh, and P. Pruthi. Chaotic Maps as Models of Packet Traffic. In J. Labetoulle and J. W. Roberts (Editors), editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, ITC-14*, pages 329–338. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1994.

[32] Ashok Erramilli, Onuttom Narayan, and Walter Willinger. Experimental Queueing Analysis with Long-Range Dependent Packet Traffic. *Submitted to IEEE/ACM Transactions on Networking*, 1994.

[33] Ashok Erramilli and Walter Willinger. Fractal Properties in Packet Traffic Measurements. In A. Koucherayvy B. Goldstein and M. Shneps-Shneppe, editors, *Proceedings from ITC Seminar Digital Communication Network Management, St. Petersburg*, pages 144–158, jun 1993.

[34] Wolfgang Fischer and Kathleen Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.

[35] Henry J. Fowler and Will E. Leland. Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management. *IEEE Journal On Selected Areas In Communications*, 9(7):1139–1149, sep 1991.

[36] C.W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[37] COST 242 (Eds. P. Tran Gia). Multi-rate Models for Dimensioning and Performance Evaluation of ATM Networks. Interim report, COST, june 1994.

[38] R.J. Gibbens and P.J. Hunt. Effective Bandwidth for Multi-Type UAS Channels. *Queueing Systems*, 9(3):17–28, 1991.

[39] Winfried K. Grassmann, Michael I. Taksar, and Daniel P. Heyman. Regenerative Analysis and Steady State Distributions for Markov Chains. *Operations Research*, 33(5):1107–1116, Sep-Oct 1985.

[40] R. Guerin, H. Ahmadi, and M. Naghshineh. Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks. *IEEE Journal On Selected Areas In Communications*, 9(7):968–981, 1991.

[41] H. Heffes and D.M. Lucantoni. A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. *IEEE Journal On Selected Areas In Communications*, 4(6):856–868, sep 1986.

[42] Bjarne Helvik and Lars Hofseth. Self-similar traffic and multilevel source models. In Ilkka Norros & Jorma Virtamo, editor, *Nordic Teletraffic Seminar, NTS 12, VTT Symposium 154*, pages 285–298, Espoo, Finland, aug 1995. NTS, VTT.

[43] H.E. Hurst. Long-Term Storage Capacity of Reservoirs. *Trans. Amer. Soc. Civil Engineers*, (116):770–799, 1951.

[44] H.E. Hurst. Methods of Using Long-Term Storage in Reservoirs. *Proc. Institution Civil Engineers Part I*, (116):770–799, 1956.

[45] Ichiro Ide. Superposition of Interrupted Poisson Processes and Its Application to Packetized Voice Multiplexers. In M. Bonatti (Editor), editor, *Teletraffic Science for New Cost Effective Systems, Networks and Services, ITC-12*, pages 1399 – 1405. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1989.

[46] Villy Bæk Iversen. The Exact Evaluation of Multi-Service Loss Systems with Access Control. *Teleteknik (English Edition)*, 31(2):56–61, 1987.

[47] Alex Jensen. Modelling of Self-similar behaviour with the Markovian Arrival Process,MAP (In Danish). IMM-EKS-1995-7 (Masters Thesis), jan 1995.

[48] Arne Jensen. Markoff chains as an aid in the study of Markoff processes. *Skand. Aktuarie tidskrift*, 36:87–91, 1953.

[49] Masatoshi Kawarasaki and Bijan Jabbari. B-ISDN Architecture and Protocol. *IEEE Journal On Selected Areas In Communications*, 9(9):1405–1415, dec 1991.

[50] J. Keilson. The ergodic queuelength distribution for queueing systems with finite capacity. *J. Royal Statistical Soc. Series B*, 28:190–201, 1966.

[51] F. P. Kelly. Effective Bandwidths at Multi-class Queues. *Queueing Systems*, 9:5–16, 1989.

[52] G. Kesidis, J. Walrand, and C.S. Chang. Effective Bandwidth for Multiclass Markov Fluids and Other ATM Sources. *IEEE/ACM Transactions on Networking*, 1(4):424–428, 1993.

[53] J.F.C. Kingman. Inequalities in the Theory of Queues. *Journal of the Royal Statistical Society, Series B*, 32(1):102–110, 1970.

[54] Leonhard Kleinrock. *Queueing Systems. Volume II Computer Systems.* John Wiley & Sons, 1975.

[55] Haluk Kosal and Ronald A. Skoog. A Control Mechanism to Prevent Correlated Message Arrivals from Degrading Signaling No. 7 Network Performance. *IEEE Journal On Selected Areas In Communications*, 12(3):439–445, april 1994.

[56] H. Kröner. Statistical Multiplezing of Sporadic Sources. In A. Jensen and V. B. Iversen (Editors), editors, *Teletraffic and Datatraffic in a Period of Change, ITC-13*, pages 787–793. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1991.

[57] K.S.Meier-Hellstern, P. E. Wirth, Y. Yan, and D.A. Hoeflin. Traffic Models for ISDN Data Users: Office Automation Application. In A. Jensen and V. B. Iversen (Editors), editors, *Teletraffic and Datatraffic in a Period of*

Change, ITC-13, pages 167–172. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1991.

[58] A. Kuczura. Queues with mixed renewal and Poisson inputs. *Bell Systems Technical Journal*, 51:1305–1326, 1972.

[59] A. Kuczura. The Interrupted Poisson Process An Overflow Process. *Bell Systems Technical Journal*, 52(3):437–448, mar 1973.

[60] K. Kvols and S. Blaabjerg. Analysis of the Periodic On/Off Queue with Applications to ATM Traffic Control. In *Proceedings from INFOCOM'92, Florence, Italy*, 1992.

[61] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, 2:1–15, february 1994.

[62] Will E. Leland, Walter Willinger, Murad S. Taqqu, and Daniel V. Wilson. On the Self-Similar Nature of Ethernet Traffic. In *SIGCOMM '93 Conference Proceedings*, pages 183–194. ACM Press, sep. 1993.

[63] K. Lindberger. Dimensioning and Design Methods for Integrated ATM networks. In J. Labetoulle and J. W. Roberts (Editors), editors, *The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, ITC-14*, pages 897–906. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1994.

[64] David M. Lucantoni. New Results on the Single Server Queue with a Batch Markovian Arrival Process. *Commun. Statist. -Stochastic Models*, 7(1):1–46, 1991.

[65] David M. Lucantoni, Kathleen S. Meier-Hellstern, and Marcel F. Neuts. A Single-Server Queue with Server Vacations and a Class of Non-Renewal Arrival Processes. *Adv.Appl.Prob*, 22:676–705, 1990.

[66] David M. Lucantoni and V. Ramaswami. Efficient Algorithms for Solving The Non-linear Matrix Equations Arising In Phase Type Queues. *Commun. Statist. -Stochastic Models*, 1:29–51, 1985.

[67] Benoit B. Mandelbrot and John W. Van Ness. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review*, 10(4):422–437, Oct. 1968.

[68] B.B. Mandelbrott and J.R. Wallis. Some Long-Run Properties of Geophysical Records. *Water Resources Research*, 5:321–340, 1969.

[69] D. Mitra and J.A. Morrison. Erlang Capacity and Uniform Approximations for Shared Unbuffered Resources. *IEEE/ACM Transactions on Networking*, 2(6):558–570, 1994.

[70] Abdi R. Modaressi and Ronald A. Skoog. Signaling System No. 7:A Tutorial. *IEEE Communications Magazine*, 28(7):19–35, jul 1990.

[71] Cleve Moler and Charles Van Loan. Nineteen Dubious Ways to Compute the Exponential of a Matrix. *SIAM Review*, 20(4):801–836, oct 1978.

[72] M. F. Neuts. Probability distrubutions of phase type. In *Liber Amicorum Professor Emeritus H. Florin*, pages 173–206, Department of Mathematics, University of Louvian, Belgium, 1975.

[73] Marcel F. Neuts. Moment Formulas for the Markov Renewal Branching Process. *Adv.Appl.Prob.*, 8:690–711, 1976.

[74] Marcel F. Neuts. A Versatile Markovian Point Process. *J.Appl.Prob.*, 16:764–779, 1979.

[75] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models an Algorithmic Approach*, volume 2 of *John Hopkins Series in Mathematical Sciences*. The John Hopkins University Press, 1981.

[76] Marcel F. Neuts. The Caudal Characteristic Curve of Queues. *Adv. Appl. Prob.*, 18:221–254, 1986.

[77] Marcel F. Neuts. *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications*, volume 5 of *Probability: Pure and Applied*. Marcel Dekker, Inc, 1989.

[78] Marcel F. Neuts. The Burstiness of Point Processes. *Commun. Statist. -Stochastic Models*, 9(3):445–466, 1993.

[79] Marcel F. Neuts, Dan Liu, and Surya Narayana. Local Poissonification of the Markovian Arrival Process. *Commun. Statist. -Stochastic Models*, 8(1):87–129, 1992.

[80] Marcel F. Neuts and H. Sitaraman. The Square-Wave Spectral Density of a Stationary Renewal Process. *Journal of the Applied Mathematics and Simulation*, 2:117–130, 1989.

[81] Ilkka Norros. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.

[82] (Ed. I. Norros) RACE Combine Deliverable D10. Handbook on Performance Analysis of Heterogeneous Networks. R2030/kta/wp1/ds/p/010/b1.

[83] V. Ramaswami. The $N/G/1$ Queue and its Detailed Analysis. *Adv.Appl.Prob*, 12:222–261, 1980.

[84] V. Ramaswami. Nonlinear Matrix Equations in Applied Probability - Solution Techniques and Open Problems. *SIAM Review*, 30(2):256–263, jun 1988.

[85] V. Ramaswami. A Stable Recursion for the Steady State Vector in Markov Chains of $M/G/1$ Type. *Commun. Statist. -Stochastic Models*, 4(1):183–188, 1988.

[86] COST 224 (Eds. J. Robert). Performance evaluation and design of multiservice networks. Final report, COST, 1991.

[87] Stephan Robert and Jean-Yves Le Boudec. Can self-similar traffic be modeled by Markovian processes ? Technical Document COST242TD(95)26, COST 242, 1995.

[88] Ronald Skoog. Study of Clustered Arrival Processes and Signaling Link Delays. In A. Jensen and V. B. Iversen (Editors), editors, *Teletraffic and Datatraffic in a Period of Change, ITC-13*, pages 61–66. ITC, Elsevier-Science Publishers B.V. (North-Holland), 1991.

[89] J.L. van den Berg and J.A.C. Resing. Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. *Submitted to IEEE/ACM Transactions on Networking*, 1995.

[90] Ward Whitt. Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues. *Telecommunication Systems*, 2:71–107, 1993.

[91] Walter Willinger, Murad S. Taqqu, Will E. Leland, and Daniel V. Wilson. Self-Similarity in High-Speed Packet Traffic: Analysis and Modeling of Ethernet Traffic Measurements. *Statistical Science*, 10(1):67–85, 1995.

[92] CCITT Study Group XI. *Specifications of Signaling System No. 7. Blue book, vol. VI, Fascicle VI.7*. Geneva, Switzerland, 1989.

Ph. D. theses from IMM

1. **Larsen, Rasmus.** (1994). *Estimation of visual motion in image sequences. xiv* + 143 pp.

2. **Rygard, Jens Moberg.** (1994). *Design and optimization of flexible manufacturing systems. xiii* + 232 pp.

3. **Lassen, Niels Christian Krieger.** (1994). *Automated determination of crystal orientations from electron backscattering patterns. xv* + 136 pp.

4. **Melgaard, Henrik.** (1994). *Identification of physical models. xvii* + 246 pp.

5. **Wang, Chunyan.** (1994). *Stochastic differential equations and a biological system. xxii* + 153 pp.

6. **Nielsen, Allan Aasbjerg.** (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data. xxiv* + 213 pp.

7. **Ersbøll, Annette Kjær.** (1994). *On the spatial and temporal correlations in experimentation with agricultural applications. xviii* + 345 pp.

8. **Møller, Dorte.** (1994). *Methods for analysis and design of heterogeneous telecommunication networks.* Volume 1-2, *xxxviii* + 282 pp., 283-569 pp.

9. **Jensen, Jens Christian.** (1995). *Teoretiske og eksperimentelle dynamiske undersøgelser af jernbanekøretøjer.* ATV Erhvervsforskerprojekt EF 435. *viii* + 174 pp.

10. **Kuhlmann, Lionel.** (1995). *On automatic visual inspection of reflective surfaces.* ATV Erhvervsforskerprojekt EF 385. Volume 1, *xviii* + 220 pp., (Volume 2, *vi* + 54 pp., fortrolig).

11. **Lazarides, Nikolaos.** (1995). *Nonlinearity in superconductivity and Josephson Junctions. iv* + 154 pp.

12. **Rostgaard, Morten.** (1995). *Modelling, estimation and control of fast sampled dynamical systems. xiv* + 348 pp.

13. **Schultz, Nette.** (1995). *Segmentation and classification of biological objects. xiv* + 194 pp.

14. **Jørgensen, Michael Finn.** (1995). *Nonlinear Hamiltonian systems. xiv* + 120 pp.

15. **Balle, Susanne M.** (1995). *Distributed-memory matrix computations. iii* + 101 pp.

16. **Kohl, Niklas.** (1995). *Exact methods for time constrained routing and related scheduling problems. xviii* + 234 pp.

17. **Rogon, Thomas.** (1995). *Porous media: Analysis, reconstruction and percolation. xiv* + 165 pp.

18. **Andersen, Allan Theodor.** (1995). *Modelling of packet traffic with matrix analytic methods. xvi* + 242 pp.