



UNIVERSITY
OF
JOHANNESBURG

COPYRIGHT AND CITATION CONSIDERATIONS FOR THIS THESIS/ DISSERTATION

This copy has been supplied on the understanding that it is copyrighted and that no quotation from the thesis may be published without proper acknowledgement.



Please include the following information in your citation:

Name of author

Year of publication, in brackets

Title of thesis, in italics

Type of degree (e.g. D. Phil.; Ph.D.; M.Sc.; M.A. or M.Ed. ...etc.)

Name of the University

Website

Date, accessed

Example

Surname, Initial(s). (2012) Title of the thesis or dissertation. PhD. (Chemistry)/ M.Sc. (Physics)/ M.A. (Philosophy)/M.Com. (Finance) etc. [Unpublished]: [University of Johannesburg](http://UniversityofJohannesburg). Retrieved from: <https://ujdigispace.uj.ac.za> (Accessed: Date).

OPTIMIZATION OF RESOURCES ALLOCATION FOR H.323 ENDPOINTS AND TERMINALS OVER VoIP NETWORKS

BY

Samuel Nlend

Submitted for the fulfilment of the requirements for the degree



ELECTRICAL AND ELECTRONIC ENGINEERING SCIENCE

IN

THE FACULTY OF ENGINEERING AND BUILT ENVIRONMENT

AT

UNIVERSITY OF JOHANNESBURG

SUPERVISOR: Dr. Theo Swart

CO-SUPERVISOR: Prof. Willem Clarke

ABSTRACT

Without any doubt, the entire range of voice and TV signals will migrate to the packet network. The universal addressable mode of Internet protocol (IP) and the interfacing framing structure of Ethernet are the main reasons behind the success of TCP/IP and Ethernet as a packet network and network access scheme mechanisms. Unfortunately, the success of the Internet has been the problem for real-time traffic such as voice, leading to more studies in the domain of Teletraffic Engineering; and the lack of a resource reservation mechanism in Ethernet, which constitutes a huge problem as switching system mechanism, have raised enough challenges for such a migration. In that context, ITU-T has released a series of Recommendation under the umbrella of H.323 to guarantee the required Quality of Service (QoS) for such services.

Although the “utilisation” is not a good parameter in terms of traffic and QoS, we are here in proposing a multiplexing scheme with a queuing solution that takes into account the positive correlations of the packet arrival process experienced at the multiplexer input with the aim to optimize the utilisation of the buffer and bandwidth on the one hand; and the ITU-T H.323 Endpoints and Terminals configuration that can sustain such a multiplexing scheme on the other hand.

We take into account the solution of the models from the M/M/1 up to G/G/1 queues based on Kolmogorov's analysis as our solution to provide a better justification of our approach.

This solution, the Diffusion approximation, is the limit of the Fluid process that has not been used enough as queuing solution in the domain of networking. Driven by the results of the Fluid method, and the resulting Gaussian distribution from the Diffusion approximation, the application of the asymptotic properties of the Maximum Likelihood Estimation (MLE) as the central limit theorem allowed capturing the fluctuations and therefore filtering out the positive correlations in the queue system.

This has resulted in a queue system able to serve 1 erlang (100% of transmission link capacity) of traffic intensity without any extra delay and a queue length which is 60% of buffer utilization when compared to the ordinary Poisson queue length.

ACKNOWLEDGEMENTS

First I say thank to God for the strength He granted me to achieve this research.

I want to thank my supervisor Dr. Theo Swart and co-supervisor Prof. W. Clarke for all their supports, comments and suggestions that made this project happen.

I thank particularly the Telecommunication Research Group and the University of Johannesburg for having put me within an environment that made this project possible.

I would like to thank specially Mannie Cahn and his family for their patience and supports.

Thanks to Raymond Bloch, Teuns Moolman, Rahmo Mehovic and all those close to me for the support each one brought to me in South Africa.

I dedicate this Thesis:

To my late mother Ngo Nlend Augustine and father Mooh Joseph who surely would have liked to see this earlier;

To my wife Aurelie and son Patrick Toussaints for the difficulties they endured during this period.

To my whole family in Cameroon.



ACRONYMS

ATM Asynchronous Transfer Module

ACF Acknowledge Confirm

ARQ Acknowledge Request

ARJ Acknowledge Reject

B-ISDN Broad band Integrated Service Digital Network

CNAME Canonical Name

CTMC Continuous Time Markov Chain

DSCP Differentiated Service Code Point

DTMC Discrete Time Markov Chain

DWDM Dense Wave Division Multiplexing

QoS Quality Of Service

MC Multipoint Controller

MCU Multipoint Control Unit

MMDP Markov Modulated Discrete Process

MMPP Markov Modulated Poisson Process

MP Multipoint Processor

LAN Local Area Network

N-ISDN Narrow band Integrated Service Digital Network

FCFS First Come First Serve

FIFO First In First Out

FDM Frequency Division Multiplexing

FR Frame Relay

IEEE Institute of Electrical and Electronic Engineering



IPTV Internet Protocol Television

LUP Look-Up Table

MAC Media Access Control

MPLS Multiprotocol Label Switching

PDH Asynchronous Digital Hierarchy

PSTN Public Switched Telecommunication Network

PLMN Public Land Mobile Network

RAS Registration Admission and Status

RFC Request For Comment

RSVP Resources Reservation Protocol

RTP/RTCP Real Time Protocol/ RTP Control Protocol

RTT Round Trip Time

SCM Selected Communication Mode

SCN Switched Circuit Network

SDH Synchronous Digital Hierarchy

SMP Semi Markov Process

TDM Time Division Multiplexing

TCP/IP Transport Control Protocol/ Internet Protocol

ToS Type Of Service

TSAP Transport Service Access Point

UDP Packet Data Unit

VAD Voice Activity Detection

VoIP Voice Over Internet Protocol



TABLE OF CONTENTS

CHAPTER 1.....	
INTRODUCTION.....	1
1-1) GETTING STARTED.....	1
1-2) PROBLEM STATEMENT.....	2
1-3) RESEARCH QUESTION.....	3
1-4) THE PROJECT'S OBJECTIVE AND SCOPE.....	4
1-5) THE PROJECT OVERVIEW.....	5
CHAPTER 2.....	
ETHERNET AND H.323 STANDARD BACKGROUNDS.....	6
2-1) INTRODUCTION.....	6
2-2) THE ETHERNET AND QoS ISSUES.....	6
2-2-1) Quality of Service (QoS) in the packet network.....	7
2-2-2) Switching Ethernet QoS issues	7
2-2-3) IP QoS Techniques.....	8
2-3) ETHERNET AND TRANSPORT ISSUES.....	9
2-3-1) Voice transport quality issues in an IP network.....	9
2-4) BACKGROUND OF THE H323 STANDARD.....	11
2-4-1) The H.323 protocols suite.....	11
2-4-2) H.323 Equipments.....	14
2-4-3) Definitions and Terminals Characteristics.....	15
2-5) CONCLUSION.....	21
CHAPTER 3.....	
STATISTICAL MULTIPLEXING BACKGROUND.....	23
3-1) – INTRODUCTION.....	23
3-2) TRAFFIC MODELS.....	23
3-2-1) ON-OFF Sources.....	24
3-2-2) Markovian sources.....	25
3-3) TRAFFIC DESCRIPTORS.....	26
3-3-1) Stochastic process.....	26
3-3-2) Traffic descriptors: Arrival process and burstiness characteristics.....	32
3-4) QUEUING SYSTEMS BACKGROUND.....	33
3-4-1) History of Queuing Theory.....	33
3-4-2) Queuing Techniques background.....	34
3-4-3) Queuing models background.....	34
3-4-4) Queuing model solutions.....	35

3-4-5) Performance tools of statistical multiplexer.....	41
3-4-6) Voice packets statistical multiplexing.....	42
3-5) CONCLUSION.....	44
CHAPTER 4	
QUEUE MODELLING THEORY.....	46
4-1) INTRODUCTION.....	46
4-2) DIFFERENT QUEUE MODELS.....	46
4-2-1) The M/M/1 queue.....	46
4-2-2) G/M/1 Queue.....	48
4-2-3) M/G/1 Queue.....	51
4-2-4) G/G/1 Queue.....	52
4-3) CONCLUSION.....	57
CHAPTER 5	
REVIEW OF SIMILAR WORK.....	58
5-1) INTRODUCTION.....	58
5-2) M/M/1 QUEUE.....	58
5-2-1) The steady-State Solution of Adam and Resing.....	58
5-3) THE G/M/1 QUEUE.....	60
5-3-1) Approximating the arrival Process in the G/M/1 Queue.....	60
5-4) – THE M/G/1 QUEUE.....	63
5-5) – THE G/G/1 QUEUE.....	65
5-5-1) Moment queues.....	65
5-5-2) Bounds solutions.....	69
5-6) PREVIOUS WORKS SELECTED RESULTS.....	70
5-7) CONCLUSION.....	72
CHAPTER 6.....	
PROPOSED SOLUTION.....	73
6-1) INTRODUCTION	73
6-2) OUR SOLUTION AND PREVIOUS WORKS.....	73
6-2-1) The bounds solution.....	73
6-2-2) The moments solution.....	74
6-3) OUR APPROACH MOTIVATIONS.....	74
6-4) NETWORK DESIGN MOTIVATIONS.....	75
6-5) NETWORK ARCHITECTURE DESIGN.....	76
6-6) LEVEL 1 MULTIPLEXING DESIGN.....	78
6-6-1) The univariate Diffusion Approximation.....	78
6-6-2) The univariate Gaussian distribution properties.....	80

6-6-3) Diffusion Approximation for ON-OFF voice packet sources.....	82
6-6-4) The queue Analysis.....	84
6-7) LEVEL 2 MULTIPLEXING DESIGN.....	85
6-7-1) The traffic model and scale.....	86
6-7-2) Multivariate Diffusion model.....	87
6-7-3) Multivariate Gaussian properties.....	89
6-7-4) Queue Analysis.....	93
6-8) CONCLUSION.....	96
CHAPTER 7.....	96
EXPERIMENTS AND RESULTS.....	97
7-1) INTRODUCTION.....	97
7-2) EXPERIMENTAL DESIGN.....	97
7-2-1) Experiment 1: level 1 multiplexer.....	98
7-2-2) Experiment 2: Level 2 multiplexer.....	102
7-2-3) The experimental results.....	103
7-3) QUEUE IMPLEMENTATION.....	104
7-3-1) Queue Implementation algorithm in MATLAB.....	104
7-3-2) Diffusion queue simulation and comparison with MM1 queue.....	105
7-3-3) A workable solution.....	107
7-4) PROJECT RESULTS.....	109
7-5) CRITICAL ANALYSIS.....	111
7-6) CONCLUSIONS.....	112
CHAPTER 8.....	
CONCLUSION AND FUTURE WORK.....	113
8-1) INTRODUCTION.....	113
8-2) RESEARCH QUESTION ANSWER.....	113
8-3) OUR CONTRIBUTION.....	114
8-4) FUTURE WORK.....	114
APPENDIX A.....	116
APPENDIX B.....	130
APPENDIX C.....	132
REFERENCES.....	146

CHAPTER 1

INTRODUCTION

1-1) GETTING STARTED

The transport of voice over Internet Protocol (IP) always raises the questions of why and how: the common answer is the New Generation Network, a network platform supporting multimedia services. This attempt started with N-ISDN [1] during the 80's and B-ISDN [2] in the late 90's, whereby the telecommunication industry was conceptualizing how to build a broadband network which could support a multitude of services. ATM [3] came as a best compromise for a single switching infrastructure.

Meanwhile, the evolution of the Internet has led to the economics of scale, which has driven practically all end users data networks to Ethernet LAN architecture.

While Ethernet has taken over the data network, the voice counterpart evolved at a much slower pace. We are still having the same TDM voice switching on 64 kbits/s and that state of the art time-slot will still be in the service for a long time to come due to huge investments that have mainly been made to support the telephone service.

The data traffic however, with its exponential growth, shows according to [4] that by 2015, voice will represent a fraction of data traffic. But despite its high volume, the revenues generated by voice are by far greater than those generated by data, driving the "data specialist" to include voice into packet services. Some facts have contributed in this process:

- Technological progress: Gordon Moore's statement that processing power doubles every 18 months [5], more than 30 years ago still holds, making control intelligence, signal processing, data storage cheaper and cheaper. Therefore, DWDM [6] systems in optical technology double their bandwidth almost every 15 months and electronic SDH every 2 years.
- Data explosion: consequence of the increase of computer processing power and storage capacity, the availability of high speed transmission links and the release of standards both in computer communication protocols and software operating systems and applications.

However, such an operation of migrating voice over Internet is not happening without challenges, particularly when it comes to providing the required acceptable and commercial

level of service as experienced in the old legacy network. Based on this requirement, we would like to bring in our contribution by tackling the Quality of Service (QoS) problems resulting from the impairments experienced by the transport of voice over IP (VoIP) within an H.323 environment.

1-2) PROBLEM STATEMENT

Taking into account the history of Telecommunications, the market behaviour and the impact of the Internet, it is obvious that the prime service, the telephone is considered as an application to transport over internet. This will not happen overnight as the telecommunication landscape invested in too many fixed assets which still need to show a return on investment, but which with hindsight, has already paved the way for such a migration:

- Most of the radio communications are using PDH [7] or SDH [8] transport systems. The latter systems with the introduction of virtual concatenation, statistical multiplexing and Ethernet switching functionality became flexible and cost effective, thus enhanced to carry Ethernet.
- The distribution of the intelligence (Figure 1.1) leads to a network structured in two levels of switching: edge and core, which corresponds roughly to a concentration and distribution level. Full meshing of the core nodes is nowadays economically possible due to the feasibility of very big nodes and the decreasing prices of transmission optics.
- The distribution of the network intelligence out of the transport part: this was already initiated with the intelligent network and has been continued in the form of a server where the network services are implemented. We can derive the very important role the access network and its constituent have to play by providing all the required services.

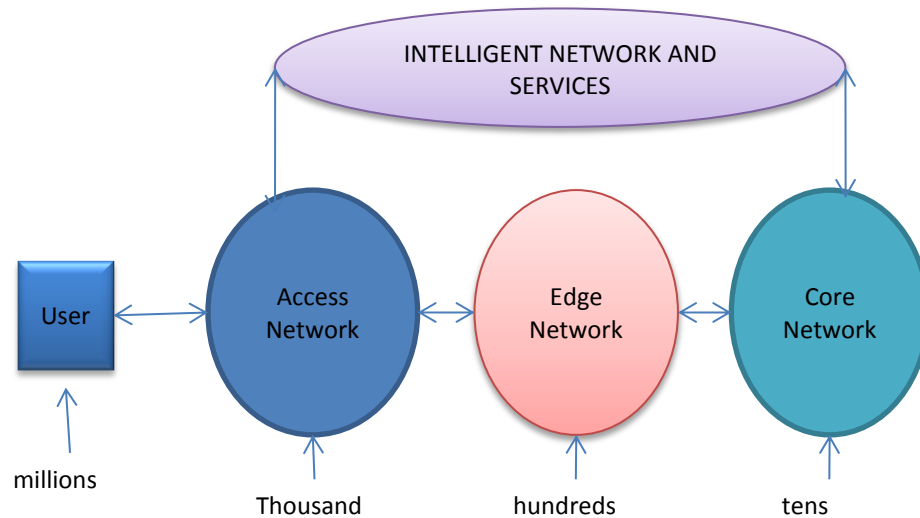


Figure 1.1: Network paradigm by 2015

The telecommunication companies dealing with such assets can then exploit their backbone to provide telecommunication services to millions of users with less to do in terms of investments.

But if to provide a service is good, then to provide a good quality service is better. Indeed, to provide voice over IP through the LANs which have no guarantee of quality, such as Ethernet, raises many challenges which ITU-T tried to overcome through a series of H.323 recommendations. Despite this, today when using voice over IP, the parameters obtained after four hours of traffic observation at a non busy time on average are [9]:

- Answer rate: 44%
- Call duration: 2.2 min
- Post Dial Duration: 15 sec

These low values are due to the impairments the network is still facing, all of them basically resulting from packet loss, jitter, and transfer delay. These (QoS) network parameters depend on the availability of resources such as buffer and bandwidth, hence our topic: ***“The optimization of resource allocation on H.323 Endpoints and Terminals over VoIP networks”***.

The optimization of resources in our sense is to accommodate multiple On-Off sources with different bit rates in a single LAN architecture while managing efficiently the resources dedicated to their Internet access through two multiplexing levels within an ITU H.323 environment.

However, the study conducted by Sriram and Whitt [89] in characterizing the aggregate arrival process from one voice source in the multiplexer has shown that the positive correlations, or to

use a better concept; the time dependence in the queuing system is the cause of the high delay at the multiplexer output, consequently of packet loss, therefore reducing the efficient utilization of network resources. Introducing variable bit rates at the multiplexer input whose output is already plagued by autocorrelations add as much challenges as questions.

1-3) RESEARCH QUESTION

Taking our approach, that consists of statistically multiplexing different voice type sources while better managing resources, raises the following questions:

- Can we allocate resources efficiently with multiple types of voice sources, despite the positive correlation in the queue system already experienced by one source?
- Can the Diffusion process, chosen as queuing solution method, be able to filter out the correlations in the queuing system?
- Does the nullification of the correlations in the aggregate arrival process at the multiplexer input lead to reducing the transfer delay in the queue system resulting from that effect?
- What can the best fit be in terms of configuration and characteristics of H.323 Endpoints and Terminals to handle this new situation?

1-4) THE PROJECT'S OBJECTIVE AND SCOPE

The objective of this project is to propose a two-level statistical multiplexing scheme whose queuing solution takes into account the positive correlations within a suitable H.323 network Endpoints and Terminals configuration.

The main objective will be divided into two sub-objectives in view to keep the required or acceptable level of Quality of Service (QoS) of our interest, which is the transfer delay. They are:

- To introduce for both statistical multiplexing levels, the Diffusion process as queuing solution with the aim to manage multiple sources with different bit rates on the one hand; and
- To characterize the H.323 Endpoints and Terminals which are suitable for such an operation or configuration on the other hand.

The ITU H.323 is meant for LAN without a guaranteed QoS that the current voice networks (PSTN/PLMN) provide. This concerns the IEEE 802.X access series, however; in this case we will focus on the Ethernet [10], particularly its QoS and voice transport issues. We will therefore

follow the requirements as prescribed to better handle the above mentioned access networks through a series of ITU-T H.323 Recommendations [11] in compliance with the statistical multiplexing we are here bringing to the fore.

Statistical multiplexing by definition directs our study to queuing theory. Queuing theory in this case, can only be addressed in terms of absolute QoS through the traffic models which characterize the sources. This characterization, derived from analysis based on stochastic processes, leads us to adopt the Markov Process modelling where Kolmogorov's equation matters. Poisson Process has been used extensively in queuing systems. For multiple types of traffic, Modulated Poisson and Fluid processes have been proposed. The problems experienced by different queue solutions lead us to the limit of the Fluid model, a Gaussian Markov process called the Diffusion process. Note that the traffic we are mentioning here is any real-time service traffic, but for purposes of simplicity we will focus on voice traffic as the topic implies; particularly on voice source encoders associated with a Voice Activity Detector (VAD): On-Off sources.

1-5) THE PROJECT OVERVIEW

In the following development, we present the remainder of this dissertation into 7 chapters:

In Chapter 2, we present the H.323 Standard and environment as well as the Quality of Service requirements of the related network which match with our goal: "the optimization of resources allocations".

In Chapter 3, we complement Chapter 2 by providing the second part description of this topic: the statistical multiplexing background.

In Chapter 4, we give the theoretical review and analysis of queuing models.

In Chapter 5, we give results obtained in previous works through selected queue models, with the aim to justify our approach.

In Chapter 6, we give the detailed analysis of our project's approach.

In Chapter 7, we present the research methodology and the relevant tests for the performance of the statistical multiplexer we are proposing; the required means to implement the project; and present through simulations the results and the explanations that sustained our approach.

In Chapter 8, we conclude the project with the overview of the results as well as discussing elements not taken into account in this project, leaving it as possibilities for future work.

CHAPTER 2

ETHERNET AND H.323 STANDARD BACKGROUNDS

2-1) INTRODUCTION

It is clear that in the near future the majority of telecommunication traffic will be generated by applications which are running on top of the IP protocol. In short, voice will entirely migrate to the packet network. It makes sense to choose IP as the convenient packet network. Also, the trend of the market, the operational simplicity (the framing structure readily operational for the IEEE 802.x access series) and the resulting cost savings have been the most compelling arguments for the use of Ethernet switching as an Internet access mechanism. Therefore, the QoS issues of this networking scheme need to be presented as well as the problems occurred for conveying voice over IP in the public environment.

As previously mentioned, the purpose of the H.323 protocol suites is to provide the non guaranteed QoS networks with an ISDN QoS-like. We will then follow up by presenting the ITU H.323 standard and analyzing the characteristics required for H.323 Endpoints and Terminals with the aim of providing an acceptable quality of service level the subscribers are used to.

2-2) THE ETHERNET AND QoS ISSUES

The most common used data network interface and switching technology is Ethernet, an IEEE standard. The first widely employed version was 10 Mbits/s media [12], the total combined available bandwidth to all devices connected. The problem was when the end stations try to access the media simultaneously, we experience collision of packets, and these packets needed to be retransmitted. The collision property made the throughput of Ethernet unpredictable as it was difficult to predict the probability of collision and retransmission.

Then came full duplex 100 Mbits/s Ethernet [13] switching technology with dedicated media for each subscriber, putting the collision in the past and the packet loss no worse than any other packet switch. The only place whereby the packet loss can be observed is the output queue of the switch due to the congestion on the output link. In addition, 1 Gbits/s is also now commonly available [14].

The most important reason behind Ethernet's popularity is the fact that its devices are compatible with each other and are plug-and-play technologies.

2-2-1) Quality of Service (QoS) in the packet network

Two types of quality of service can be supported in the packet switching network:

- Relative QoS, often called Class of Service (CoS),
- Absolute QoS.

Relative QoS assigns different priorities to packets by marking them. This type of CoS implementation can be based on the "Differentiated Service" mechanism [15].

Absolute QoS in the packet network can be characterized by bandwidth, delay, delay variation and packet loss. It is applicable to a certain point-to-point traffic stream between two end interfaces.

2-2-2) Switching Ethernet QoS issues

a) Ethernet QoS

Unlike ATM or FR (Frame Relay), standard Ethernet does not provide any mechanism for bandwidth reservation inside the network. The typical philosophy in the Ethernet network is such that the network's administrator simply assumes that there is "enough bandwidth" so that there is no way to worry about it. The availability of the guaranteed bandwidth is assured by policing the user traffic according to a traffic contract at the edge of the network. This lack of resource reservation could imply that switched Ethernet could not be used for real-time traffic and particularly for VoIP.

Inside the network, some CoS oriented mechanisms need to be used to allocate a sufficient amount of bandwidth on all the links that are traversed by the traffic.

b) Ethernet CoS

The Ethernet network provides a very good "differentiated service" relative QoS that gives priority to packets. In fact the standard CoS mechanism of Ethernet networks (802.1P part of 802.1D) provides very good support for differentiated services relative QoS. The 802.1D supports 8 different CoS markings on packets, which can be used by queuing systems to make the packet handling decision.

- IEEE 802.1P [16] defines traffic class and dynamic multicast filtering. It provides a means for implementing QoS at the MAC level through Eight classes of services using 3-bit user priority field in an IEEE 802.1Q [17] header added to the frame.

- IEEE 802.1D [18] is the IEEE MAC Bridges standard that specifies through Bridging, Spanning Tree how to link many of the other 802 projects including the widely deployed 802.3 (Ethernet), 802.11 (Wi-Fi) [19] and 802.16 (WiMax) [20] standards.

Figure 2.1 shows an Ethernet frame which is basically made up of destination address (6-bytes); source address (6-bytes); frame type/length (2-bytes); payload (*n*-bytes) and the 1-byte flag. With the 802.1P/Q field, there is 4-bytes tag control information added to 2 bytes frame type. This field can be presented as follows.

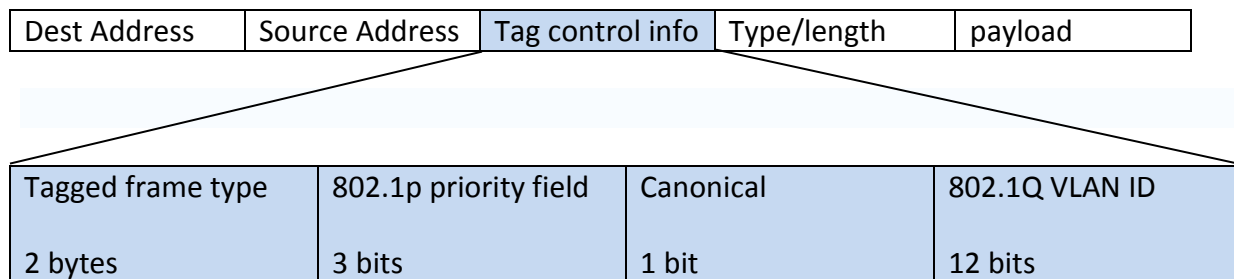


Figure 2.1: Ethernet frame with 802.1p/Q

The QoS of the lower layer as Ethernet is more effective if the upper layer provides its best in that regard. Most QoS mechanisms are found at layer 3 of the TCP/IP suites.

2-2-3) IP QoS Techniques

The layer 3 of the TCP/IP suites, also called the IP layer, provides a wide range of QoS mechanisms which ensure a good management of the lower layers of the protocol suites.

a) TOS/IP Precedence/DiffServ

Over time, the header of an IP frame had 1 unused byte, whereby all the bits are set to a “0” value which according to the version 4 specification, is called the Type of Service (TOS) byte. As a primary attempt to provide IP related QoS, TOS/IP Precedence were defined as follows:

The 4-bits of the TOS in RFC 1349 [21] allows 4 classes of service: minimize delay; maximize throughput; maximize reliability and minimize cost.

According to RFC 791 [22] and RFC 1812 [23], IP precedence uses the first 3 bits in the TOS byte to allow 8 different type of classes of service.

DiffServ [24] is the latest attempt for providing QoS using 6 bits of the TOS byte, referred to as Differentiated Services Code Point (DSCP). This allows 2^6 different classes of service.

The routers have to look at bits in the IP header all the times to understand the required service. See Figure 2.2.

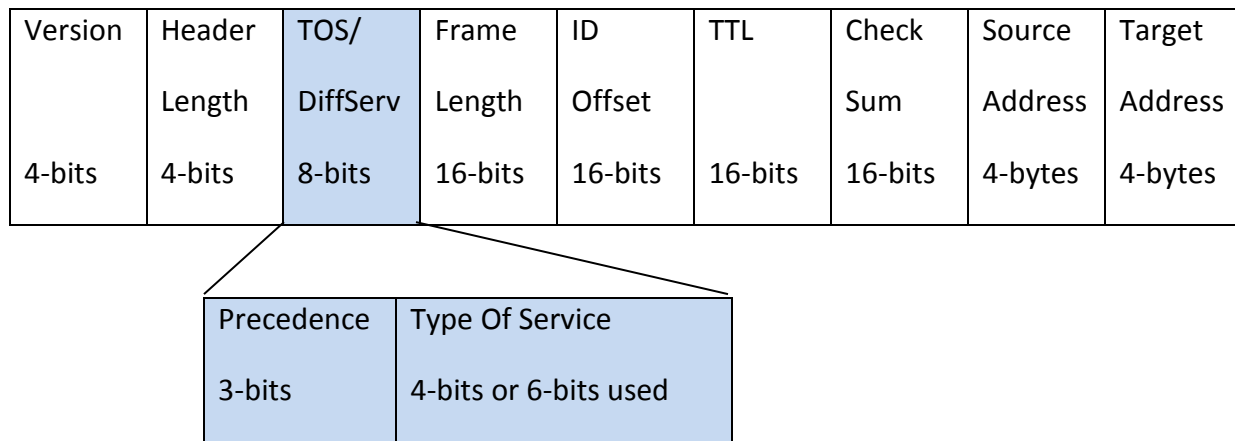


Figure 2.2: IP header

b) Resource Reservation Protocol (RSVP)

The main goal of RSVP [25] is to reserve resources with the aim to satisfy the required bandwidth, jitter and delay for a selected network path’s routers. According to RFC 2205 [26], also termed Integrated Service, the RSVP utilizes traffic control from one end to another instructing routers to reserve some of their resources for the forthcoming application.

c) Multiprotocol Label Switching (MPLS)

Being more than a QoS technique; MPLS [27] is tagged at the start of the IP header with 2-bytes for each source and destination labels which are used by MPLS aware routers, also called LSRs, to forward packets through the network in replacement of the traditional IP header addresses. Different paths through the network, called Label Switch Paths (LSP), are configurable according to the variety of label values, providing operators with a way to offer different classes of service.

However, the QoS techniques mentioned above need a good transport system which can be summarized in a good signal level and less transfer delay as the following section presents.

2-3) ETHERNET AND TRANSPORT ISSUES

2-3-1) Voice transport quality issues in an IP network

Voice quality has always been subject of discussions. However, its measurements have been standardized by a set of subjective tests. The transmission rating factor according to E-model

[28] defined in ITU-T Rec. G.107 [29] and G.109 [30] is related to 5 speech transmission qualities as in Table B.1 in Appendix B.

E-model [28] is a tool for transmission planning. It establishes a relation between the perceived speech qualities to the combined effects of all the transmission impairments:

$$R = R_0 - I_s - I_e - I_d + A,$$

where:

- R_0 is the basic signal-to-noise ratio (S/N), taking into account the power addition of all noise sources.
- I_s is the sum of probable impairments related to voice transmission such as low signal level, non-optimal side tones and quantization distortion.
- I_e is the equipment impairment factor associated with the use of low bit rate codecs. It also reflects according to Recommendation G113 [31], in the packet network the effect of packet loss.
- I_d is the sum of impairments related to delay of voice signals which include talker and listener echo as well as loss of interactivity due to excessive delay. The latter parameter is the most important cause when we deal with voice over packet network, as it is always bigger in VoIP than in PSTN. It is made up of:
 - * Encoding-Decoding delay: that is coding and decoding time intervals of audio codec samples in use.
 - * Packetization delay: that is the required encapsulation time.
 - * Propagation delay: that is the time needed by the signal to propagate from one point to another.
 - * Queuing and transmission delay: that is the time needed to transmit the packet through the network.
 - * Jittering Delay: that is the delay introduced by the use of a buffer at the destination to absorb delay variation in the packet network.
- A is an advantage factor reflecting the degradation the user is prepared to suffer for gaining in other features such as: access convenience, mobility and lower tariffs.

Tables B.2 and B.3 in Appendix B outline the correlation of the parameters above and the signal level.

To reach this type of QoS, the ITU-T has released the ITU H.323 recommendation series which we will investigate in the following section. The key issue is to analyse how the H.323 requirements help to achieve a good voice signal level and an acceptable delay.

2-4) BACKGROUND OF THE H.323 STANDARD

The H.323 standard [14] defines means for transporting audio; video and data across IP based networks to enhance multimedia communications over LANs that do not provide a guaranteed QoS.

2-4-1) The H.323 protocols suite

The H.323 protocols suite is a set of 27 collaborative LAN based protocols for multimedia communications can be mapped in the TCP/IP architecture [32], see Figure 2.3.

With the scope of this project, we will focus on the ones related to real time services which include H.225.0, R.A.S, Q.931, H.245, RTP/RTPC, audio codec (G.711, G.723, G.728,...) and video codec (H.261, H.263), particularly the one related to audio service for simplicity.

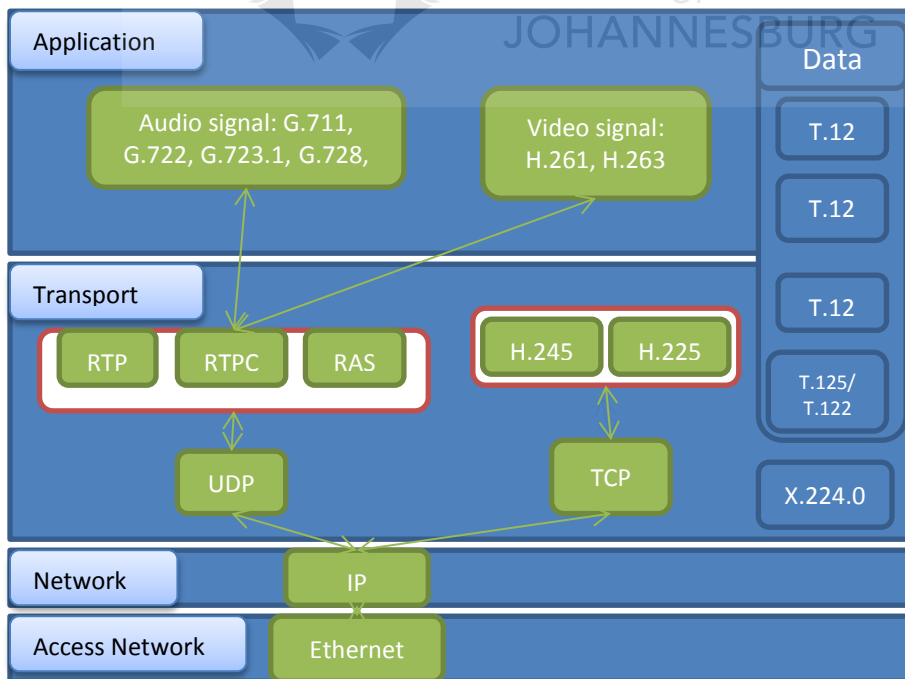


Figure 2.3: The H323 protocol suite [11]

a) The Audio Codec

Located at the application layer, they are audio signals coded with the voice algorithm that complies with the ITU standard's name they bear. This range comprises G.711 [33], G.722 [34], G.723 [35], G.728 [36], G.729 [37], etc.

b) The H.225.0 standard

H.225.0 is a standard [38] which covers narrow-band visual telephone services defined in H.200/AV.120-Series Recommendations [39] describing how audio, video, data and control information through ITU Rec. Q.931 [40] on a packet based network, can be managed to provide conversational services in H.323 environment.

c) The H.245 Protocol

According to [41], H.245 is line transmission of non-telephone signals, that is: receiving and transmitting messages in the ASN.1 syntax [42] for capabilities as well as mode preference from the receiving end, logical channel signalling, Control and Indication. Acknowledged signalling procedures are specified to ensure reliable audio visual and data communication. These procedures can be found in the ITU-T Rec. H.245.

d) The RAS Protocol

As its name implies, the Registration, Admission and Status (RAS) [43] controls registration, and provides rights and information related to the status of Endpoints in the network where the Gatekeeper is present.

e) The RTP/RTPC Protocol**- The RTP**

The RTP (Real Time Protocol) [44] is a Real Time Transfer Protocol that ensures the transport of real time services over the Internet. In its frame structure shown in Figure 2.4, besides the usual identification and counting fields, the RTP packet defines 2 important fields:

- Payload Type (PT, (7 bits) associated with its Marker (1 bit)) specifies the media type and its interpretation by the applications.
- Timestamp (32 bits): First octet of the structure, this field defines the sampling instants which are incremented monotonically and linearly by a clock for synchronization and jitter computation purposes. For details about jitter computation, see [44].

The RTP header extension is provisioned for personal implementations as new payload functions that require additional information to be carried in the RTP packet header.

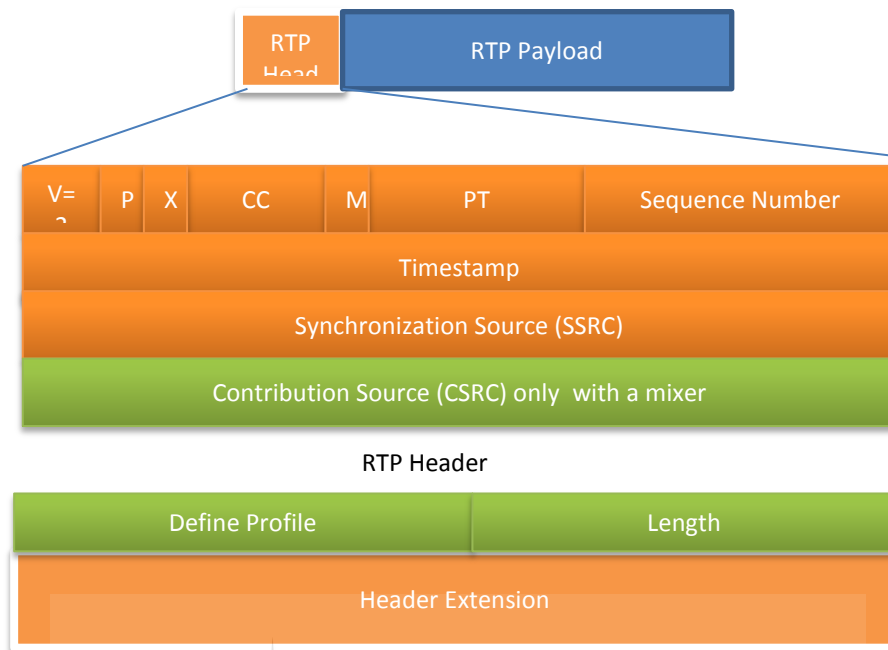


Figure 2.4: RTP frame structure

- The RTPC Protocol

Based on the same distribution mechanism as RTP data packet, The RTP Control protocol (RTCP) [45] transmits control packets to all participants involved in the session. Provided that the RTCP header carries a variety of information, its packet header depends on the packet type which can be according to [45]:

- SR packet: Sender Report packet for transmission and receiving of statistics from all participants.
- RR packet: Receiver Report for source identification. It includes CNAME (Canonical Name) for source identification and media association.
- BYE packet: End of participation.
- APP packet: for Application specific function.

Each of these RTPC packets begins with the fixed part of the RTP packet followed by the elements of the packet type. For more details, the reader is referred to RFC 1889 [45].

- **RTP processing layer.**

There is an integrated layer processing design principle [45] that allows multiplexing, mixing and translation of different media.

The multiplexing is defined by the destination transport address but the number of multiplexing points must be minimized for the efficiency of the processing protocol.

The mixing is another intermediary RTP level used for the low-bandwidth link for the multicast application. The mixer is placed between the high-bandwidth path and the low-bandwidth path of the link to present to the next mixer a new processed signal.

At this level, there is also a function called translator: this is used for firewall applications.

2-4-2) H.323 Equipments

We are presenting here the architecture of an ITU H.323 network. The terminals and Endpoints can be used for multipoint configurations and can operate through a gateway with H.310 [46] and H.321 [47] equipments of B-ISDN, H.320 and telephone set, and H.322 [48] terminal (LAN with guarantee QoS) of N-ISDN, H.324 [49] terminals, V.70 terminals [50] and telephone set of GSTN. The aim of this section is to describe the components of a H.323 system such as: Terminals, Gateways, Gatekeepers, Multipoint Control Units (MCU) and Multipoint Controllers (MC).

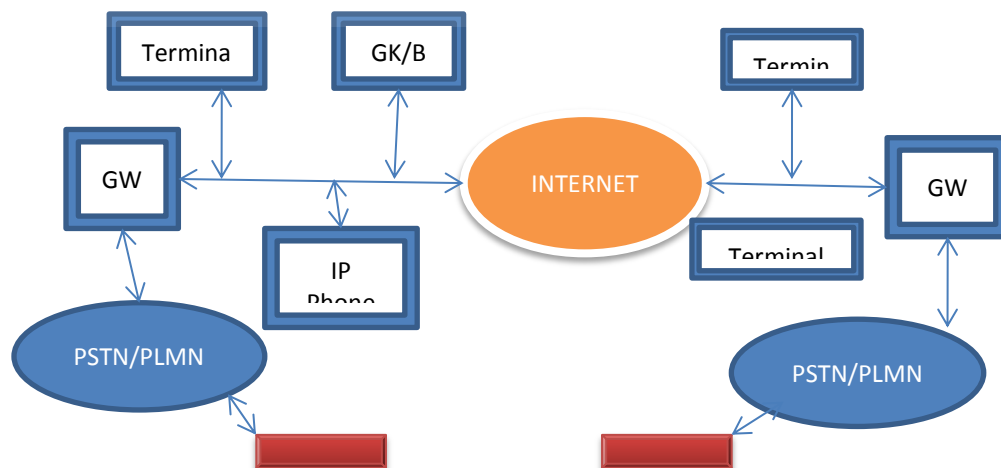


Figure 2.5: H.323 Architecture

The components are described as:

- BE: The border element.

- GW: The gateway provides conversion to and from standard multimedia to IP. As an Endpoint the GW must support all the functionalities required in view to allow bidirectional communications.
- GK: The gatekeeper performs call control function and registration and offers bandwidth management for all the endpoints in the LAN.
- Terminal: any H.323 Terminal for instance MCU.
- T: They are H.310 and H.321 equipments of B-ISDN, H.320 terminals, H.322 terminal (LAN with guarantee QoS), and telephone set of N-ISDN, H.324 terminals, (V.70) terminals, and telephone set of GSTN.

The above architecture according to ITU-T Rec. H.323 is called zone. The T (Terminals) equipments depend on the end user's application and it is not in the focus of this topic.

2-4-3) Definitions and Terminals Characteristics

We are concerned here in determining the characteristics of the H.323 Terminals that comply with the traffic of interest in this project.

a) ITU-T H.323 Definitions

Multipoint Control Unit: Also known as Conference Bridge, it is an Endpoint of the LAN that allows three or more Terminals together with their Gateway to participate in the multipoint conference. The H.323 conference bridge, a communication server, operates as the H.231 [51] conference of the switched circuit network (SCN). A multipoint conference comprises two main parties:

- A Multipoint Controller that is mandatory; and
- Multipoint processors, which are optional.

Multipoint Controller: The H.323 entity within the LAN ensuring the management of at least three Endpoints participating in the conference multipoint. It allows the negotiation with all Terminals, the required resources for the peer-to-peer communications, control those resources in view to determine for example the diffusion mode. It does not mix nor switch signals.

Multipoint processor: The H.323 entity within the LAN that centralizes and processes signals within the multipoint conference. It can therefore mix and switch signals.

Multipoint Conference: Conference between three or more terminals.

Endpoint: The Endpoint is a Terminal, Gateway or Conference Bridge that can be called or call.

Gateway: The H.323 Gateway is an Endpoint of the LAN ensuring real time bidirectional communications between Terminals of the LAN and other Terminals of the WAN.

Gatekeeper: This is a H.323 entity of the LAN that converts addresses, control access to the LAN for Terminals, Gateways and Conference Bridges, manages bandwidth and locates the Gateways.

Call: Point-to-point multimedia communication between two H.323 Endpoints. The call begins with the call set-up procedure and ends with the call termination procedure. It is thus a collection of reliable and unreliable channels between the Endpoints.

Callable: Capable of being called. Terminals, MCUs and Gateways are callable, but Gatekeepers and MCs are not.

b) Terminal Characteristics.

The Terminal characteristics as shown in Figure 2.6 are:

- I/O video: equipment meant for camera and display unit ensuring the processing of video signals.
- I/O audio: equipment allowing for speech detection such as speakers and telephone set, and echo suppression.
- Data applications and associated user interfaces using T.120 [52] data services. It assures data transfer, database access, audio graphic conference, etc.
- RAS (Registration-Admission-Status), controls access to the network in terms of rights, and status of the Endpoints and Terminals.
- System management for user and operations to manage the Endpoints.
- The video codec: encoding video signal coming from a camera for transmission and decoding the video signal received that is sent to the display unit.
- The audio codec encodes the audio signal for transmission, decodes the received audio signal and that is restored to the speaker.
- The module system management assures the required signalling, the connection commands and capacity exchanges.

- As a LAN interface, the H.225.0 layer formats signals to transmit and extracts the incoming messages to and from the LAN, assures the sequence numbering, error detection and correction; and other services as described in the Rec. ITU H.225.0. This requires an end to end reliable service (TCP, SPX) for command, data and signalling and the logical channel H.245; and an end to end unreliable service (UDP, IPX) for audio and video.

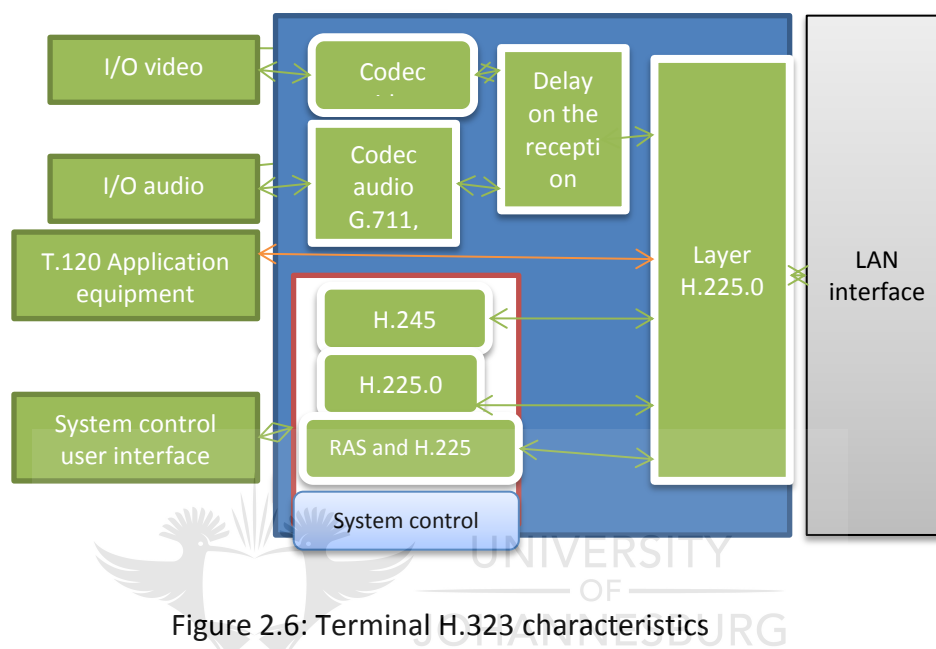


Figure 2.6: Terminal H.323 characteristics

- Terminals and Audio Codec

According to [11] all H.323 terminals shall comply with the capability of encoding/decoding G.711; transmitting/receiving in A-law and μ -law [53]; and optionally encoding/decoding G.722, G.728, G.729, MPEG 1 audio [54], and G.723. The considered audio algorithm shall derive from H.245 capability exchange negotiations that should result to an asymmetric operation for all audio capabilities the Terminal has declared within its capability set; or may result to more than one simultaneous audio channels (to allow two languages to be conveyed).

The resulting audio packets should be delivered no later than 5 ms to the TSAP after the whole multiple of the audio algorithm frame time interval clocked from the delivery of the first related audio frame. Audio codec may limit the audio delay jitter to allow receivers to reduce if required, their jitter delay buffers.

There are others audio processing operations according to ITU H.323 that can be found in the related reference.

- Terminals and H.245 Control Function

It is a set of protocol entities, each specified by its syntax, semantics and a set of procedures which s prescribed by H.245 are: capabilities exchange, logical channel signalling, opening and closing of logical channels, mode preference requests, flow control messages, and general commands and indications.

Capabilities exchange: It is a Terminal's capability describing its ability to communicate in various combinations of modes simultaneously as presented in Table 2.1.

Table 2.1: Example of voice Terminal Capability

Capability Table	Mode No	AlternativeCapabilitySet	SimultaneousCapability	CapabilityDescriptor No	CapabilityDescriptor
G.711 A	1	{1,2,3}	{{1,2,3},{2,3,4}}	13	{{1,2,3},{2,3,4},13}
G.711 μ	2	{1,3,5}	{{1,2,3,4},{1,3,5}}	42	{{1,2,3,4},{1,3,5},42}
G.723	3	{2,3,4}			
G.728	4	{1,2,3,4}			
G.729	5				

The transmitting Terminal assigns each individual mode the terminal is capable of operating in a number in a **CapabilityTable**, which are grouped into **AlternativeCapabilitySet** structures, indicating exactly one operating mode for the terminal. These **AlternativeCapabilitySet** structures are grouped into **SimultaneousCapability** structures to indicate a set of modes the terminal is capable of using simultaneously. The Terminal's total capabilities are stored in a set of **CapabilityDescriptor** structures, each being a single **SimultaneousCapability** structure represented by a **CapabilityDescriptorNumber**. If more than one **CapabilityDescriptor** is sent, the Terminal may precise operating modes which it can simultaneously use (See Mode Preference in [41]).

Logical Channel Signalling: Established between two Endpoints; an Endpoint and an MC, or an Endpoint and a Gatekeeper, the H.245 logical channel signalling is made up of:

- The control channel, established exactly one on logical channel 0, it is permanently opened for the duration of the call, for each call and as many as the number of calls that the Endpoint is participating in, using the messages and procedures H.245.
- The logical channel carries RTP packet from a transmitter to one or more receivers. It is identified by a unique logical number in the range [0, 65535] for each direction if

unidirectional. A logical channel is opened/closed accordingly using the OpenLogicalChannel/CloseLogicalChannel messages. Each opened Channel message describing the content of the logical channel: media type, algorithm in use, any option or other information needed for its interpretation in the receiver. The reaction from the latter through OpenLogicalChannelAck includes the MediaTransportChannel for RTP transport address and the MediaControlChannel for forwards RTPC transport address useful for the sender.

The flow control: it commands the bit rate limit such that a logical channel's bandwidth shall have an upper bound specified by the minimum of the Endpoints transmit/receive capability. This determines the maximum bit rate of the logical channel for the transmission of the information stream, not including RTP headers, RTP payload headers and LAN headers and other overhead, based on the FlowControlCommand message of H.245. For more information, see [41].

- **RAS Signalling Function**

The RAS Signalling Channel is established before any other channel procedures between H.323 Endpoints which are Call Signalling Channel; the H.245 Control Channel; and H.245 open logical channel procedures.

- **Call Signalling Function**

The call signalling function connects two H.323 Endpoints if no Gatekeeper, or between the Endpoint and the Gatekeeper, or between the Endpoints themselves as chosen by the Gatekeeper using H.225.0 call signalling procedures, therefore independent from the other channel procedures. It is set up before the establishment of the H.245 Channel and any other logical channels between H.323 Endpoints.

- **H225.0 Layer**

It is the H.323 network's interface to all networks with no guaranteed QoS. It therefore formats signals to transmit and extracts the incoming messages to and from the LAN, assures the sequence numbering, error detection and correction. It ensures a reliable end to end transfer of H.245 signalling channel (TCP) and an unreliable end-to-end transfer of logical channels (UDP).

c) Gateway Characteristics

The Gateway shall convert the transmission format H.225.0 to/from H.221 [55], the communication control signals and procedures H.245 from/to H.242 [56], and the call set-up and clearing signals and procedures between the H.225.0 Call Signalling and the SCN signalling system when required.

The Gateway shall pass through all unknown signalling received from SCN Endpoints to the LAN Endpoints and vice versa.

There are 4 different functionalities of a Gateway as described in the ITU-T H.323. But if as in our case the Gateway includes an MCU function on the LAN side, that function shall be an H.323 MCU on the LAN side. Or if it is the case on the SCN side, it may be an H.231/H.243 [51]/ [57] MCU or a MCU for H.310 or H.324 systems on the SCN side. This is the configuration required for our topic and can be described as in Figure 2.7.

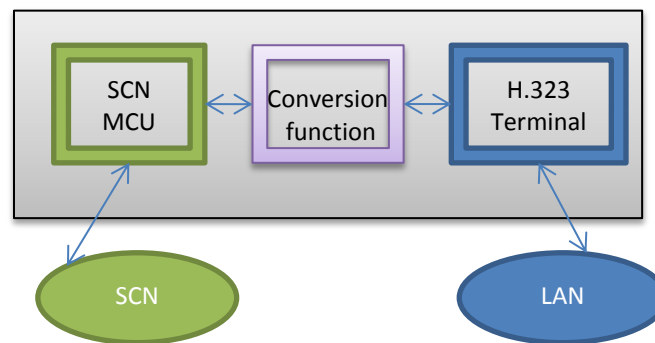


Figure 2.7: Gateway configuration requirement

d) Gatekeeper Characteristics

The Gatekeeper, optional in an H.323 system, ensures call control services to the H.323 Endpoints. It is logically separated from the Endpoints; but may coexist with a terminal, MCU, Gateway, MC, or other non-H.323 LAN device. When it is present in a system, the Gatekeeper as a zone manager shall translate address, control admission using ARQ/ACF/ARJ H.225.0 messages, and control bandwidth through BRQ/BRJ/BCF messages. See ITU-T H.225.0 [38].

e) Multipoint Controller (MC) Characteristics

The MC responsible of control functions that support conferences between three or more Endpoints ensures capabilities exchange; revises capacities according to the circumstances; and selects transmitting mode as prescribed in H.245 messages in a multipoint conference. The Selected Communication Mode (SCM) for the conference may be common for all Endpoints in the conference or may differ for some Endpoints in compliance with the capability of the Endpoints or the MC. As part of multipoint conference set-up, an Endpoint connects to an MC on its H.245 Control Channel explicitly with an MCU or implicitly with the MC within a Gatekeeper or another Terminal, Gateway where it may be located as in Figure 2.8.

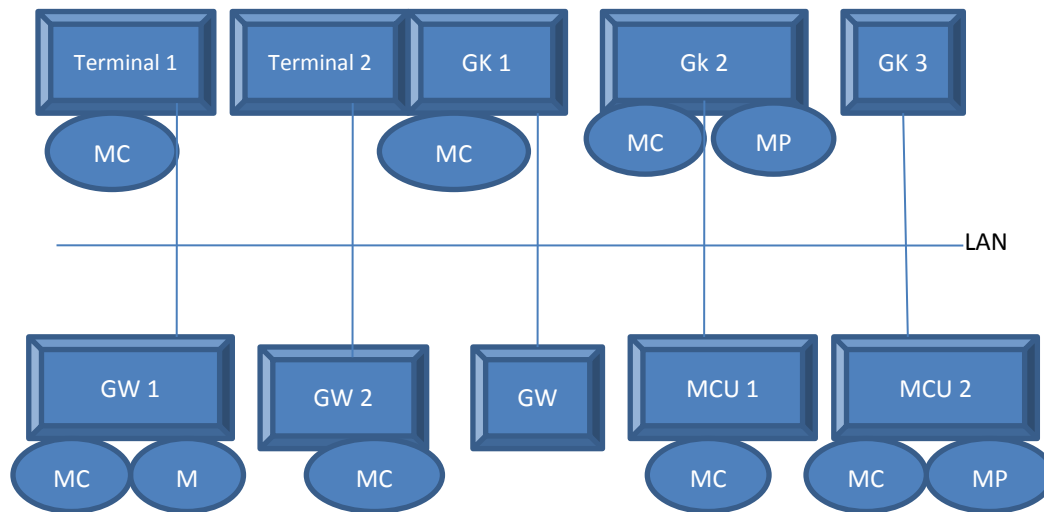


Figure 2.8: Possible location of MC and MP in an H.323 system

f) Multipoint Processor (MP) Characteristics

The MP processes audio, video and/or data streams from the endpoints involved in a centralized or hybrid multipoint conference by multiplexing, mixing or combining both, M -audio inputs to N -audio outputs in an appropriate format of the related conference.

The MP as arrival and departure of media channels may filter out unwanted signals or reduce noise from signal. The MP is not callable; the MCU in which it is settled is callable.

g) Multipoint Control Unit (MCU) Characteristics

The MCU is an Endpoint ensuring function of multipoint conferences using H.245 messages and procedures to implement features similar to those found in Recommendation H.243 [57]. A typical MCU that supports centralized multipoint conferences consists of an MC and an audio, video and data MP. The latter media only supports the Recommendation T.120 in case of decentralized conference. The MCU shall be callable by other Endpoints using the procedures in [41]. It allows the multipoint capability to the Endpoints and Terminals, which can be centralized (mandatory), decentralized or hybrid.

- Bandwidth Changes

Call bandwidth is initially established and approved by the Gatekeeper during the admission exchange. An Endpoint aggregates the bit rate of all transmitted and received audio /video excluding RTP header, LAN header and other overhead, data and control channel. Any request of bandwidth changes occurs if and only if the Endpoint needs to increase its bit rate or to use a reduced bandwidth for an extended period of time using the message BRQ (Bandwidth Request) and wait for a response from the Gatekeeper which can be BRJ (Bandwidth Reject) or BCF (Bandwidth Confirm). See ITU-T H.323 Recommendation series for more details.

2-5) CONCLUSION

According to ITU-T H.323, the resources allocation is realised and controlled at the transport layer through the H.245 protocol upon the Endpoints capability negotiations under the control of the Gatekeeper. The Gatekeeper that controls the network access allows only calls that comply with the aggregate bandwidth communicated to it by the Endpoints.

From this overview, one can point out the key role the network configuration plays in the H.323 network in terms of relative QoS. This configuration requires a Centralized Multipoint Capability at the edges of the network, ready for multiplexing to achieve better resources management. The above requirements can be completed with those standardized in the following ITU-T Recommendations [58] and [59] to yield the network configuration shown in Figure 2.9.

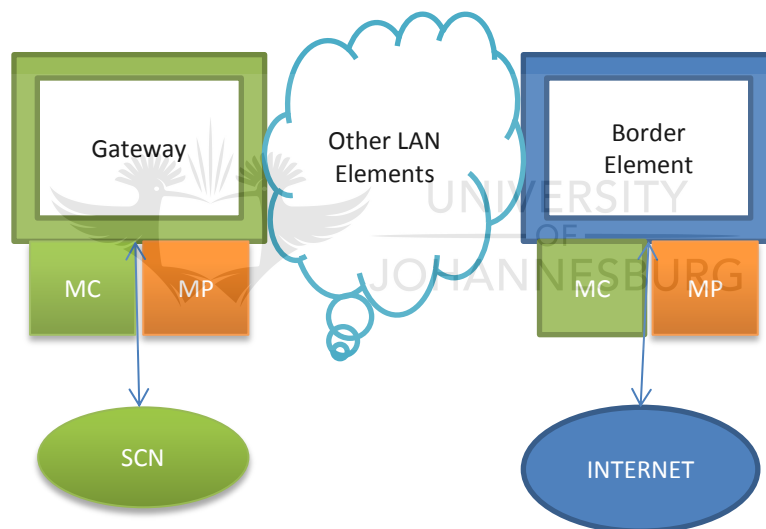


Figure 2.9: The H.323 Endpoints configuration.

CHAPTER 3

STATISTICAL MULTIPLEXING BACKGROUND

3-1) INTRODUCTION

In the last chapter, we have seen that the network configuration and particularly the edge-network characteristics play a key role in terms of resource allocations. There are two options when it comes to allocating resources in a packet network [60]:

- The aggregate peak demand; and
- Statistical multiplexing.

Statistical multiplexing, our choice for this case, allows the efficient use of network resources, which are buffers and bandwidth. Unlike the circuit switched multiplexing techniques that allocate each source a capacity equal to its peak rate, statistical multiplexing allocates a capacity that lies between the average and peak rates and buffers the traffic during periods when demand exceeds channel capacity. But this implies greater packet delay variation and packet transfer delay. These delays are often the result of correlations or time dependency experienced by the system's queue due to the variations observed in different point processes that occur at a point in time [61].

In the present chapter, we describe through the stochastic processes the traffic models and their network parameters that allow us to implement such multiplexing schemes, as well as the characteristics to evaluate their related queue performance.

3-2) TRAFFIC MODELS

Modelling the traffic allows one to set the first characteristics of the queuing system: the arrival process, also called multiplexer input or simply source model. Many studies have been done in that regard. For Frost and Melamed [62], the traffic is composed of samples called packets arriving in the sequence of random arrival times associated with random workloads which can represent packet size or batch size. The solution of the traffic model which is a "Renewal processes" in such a case is obtained when the independent random variables "arrival time" and "workload" are independent of each other.

The studies regarding Ethernet traffic showed according to Gusela [63] that it is non-stationary and characterized by a “long-tailed” interarrival time distribution while Leland [64], on his studies of Ethernet traffic over several timescales proposed a “Self-Similar process” as a model based on dependence features obtained in the traffic measurements.

Thus, the packet’s time dependence features observed in the traffic measurements have to be captured accurately by traffic models in their arrival process. In this regard, the relevance of time scale on one hand and the statistical parameters, such as first and second order moments, on the other hand have been proposed as characteristics to approximate the arrival process at the multiplexer input. It is therefore important to study the source generator of the traffic.

3-2-1) ON-OFF Sources

The ON-OFF source [65] is a bursty source modelling traffic characterized by a succession of active “ON” periods separated by silent “OFF” periods (Figure 3.1). This traffic model analysis relies on the emission of packets and the statistical properties of the different periods:

- The nature of packet emission:

* A periodic packet emission leads to a deterministic analysis.

* A continuous-time process is applicable if the packets are sent in respect with a Bernoulli or Poisson process.

- The statistical properties of the “ON” and “OFF” periods yield several possible cases:

* The source may switch from “ON” to “OFF” according to a Continuous Time Markov Chain (CTMC) [66] simulated to a Poisson emission process: the two-states Markov Modulated Poisson Process (MMPP). Because the sojourn time of a CTMC is exponentially distributed, the silence periods are by corollary exponentially distributed.

* A discrete form of this type of source [67] is determined by conducting a Discrete Time Markov Chain (DTMC) analysis with two-states whose transitions occur at instants $t_n = n\Delta t$, with $n = 0,1,2,3,\dots$ being the different instants and Δt the transition period. The steady state is then geometrically distributed.

* There is a hybrid case whereby a CTMC and a deterministic emission of information process are considered and is called a Modulated Markov Deterministic Process (MMDP) [68]; or a more general analysis based on Semi-Markov Process (SMP) [69] is proposed and we will detail the latter in the next section.

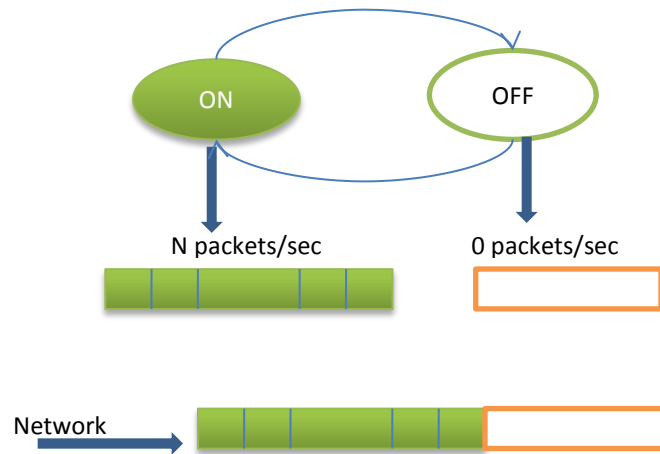


Figure 3.1: On-Off sources

3-2-2) Markovian sources

a) Markov Modulated sources

Markov modulated source models are extensively used in broadband network's performance evaluation [69]. It allows a general analysis of a multi-rate traffic or multi-type sources. When the system is in state i , the source emits packets at rate R_i . After a certain time in that state (sojourn time), the process switches to a state j emitting packets at rate R_j . The arrival process is therefore set to be a Markov Modulated Rate Process (MMRP). The embedded process consisting of changes of state is assumed to be a Markov Chain in continuous or discrete time (See Figure 3.2).

Provided that the packets are generated according to a Markov Chain, a Markov modulated source is modelled by the following conditions called "degrees of freedom" [69]:

- The number of states (also referred to as phases) of the Markov Chain;
- The transition probability which is the modulator structure, considered herein as birth death process;
- The process in each state which models sources depending on the parameter "time" considered. If the time is continuous, the packets are generated according to a Poisson process and we have a MMPP. If the time is discrete, the source is then modelled as a Bernoulli process (MMBP). A periodic time can be set for packet emission, which is then a Deterministic Process (MMDP); and
- For Continuous Time Markov Chain (CTMC) characterized by its time evolution, we have an exponentially distributed sojourn time. For discrete time Markov chain (DTMC),

the sojourn time is geometrically distributed if the time is slotted. Otherwise, the sojourn time is generally distributed and the phase process is then a Semi-Markov Process (SMP).

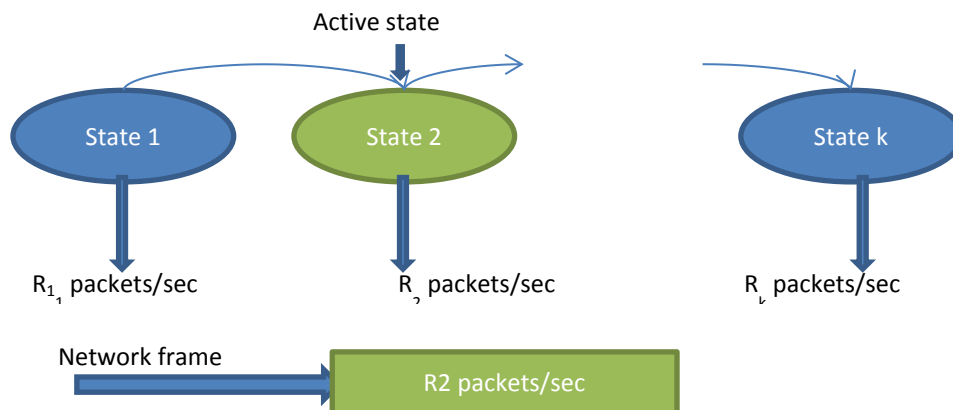


Figure 3.2: Markov Chain source

b) Other Markovian sources

The definition of other source models different to those mentioned above requires other approaches for analysing associated multiplexers. They present a more universal model encompassing many special cases in such a way that particular solution methods are needed. Many solutions have been proposed: Neuts [70] has initiated researches in that led to the wide class of arrival process called Markov Arrival Process (MAP) [71]. Other forms such as a type of renewal process called phase process (PH process) [72], up to the Discrete MAP (DMAP) [73] or batch arrival (BMAP) [74] which are equivalent to Neuts versatile MAP [70] were studied. Blondia [75] has given a discrete time version of the batch arrival, the D-BMAP which is very applied for slotted time since it can yield a Bernoulli source, a Markov Modulated Bernoulli process source or the latter form with a correlation in the batch arrival.

3-3) TRAFFIC DESCRIPTORS

The traffic descriptors are network parameters that are obtained by characterizing the communication network through traffic models. They are derived from the study of stochastic processes.

3-3-1) Stochastic process

This term in the traffic analysis results from the asynchronous property that does not allocate a reserved time slot to a connection. As packets are transmitted as they appear, each connection is considered as a stochastic process.

By definition, a stochastic process is a parameterized family of random variables $\{X(t), t \leq T\}$. The parameter t belong to the index set T , which if countable gives rise to a discrete-parameter process (time for example), hence the notation $\{X_n\}$; and in the other case yields a continuous-parameter process [76]. The state space is the set of all possible values the random variable $X(t)$ may take.

a) Characterization of the process

Characterizing a stochastic process turns out to capturing its properties by entirely specifying its multivariate probability distribution function (pdf) $f_{x_{n_1}, x_{n_2}, \dots, x_{n_m}}(x_1, x_2, \dots, x_m)$ for all m -tuples (n_1, n_2, \dots, n_m) and all positive integers m , which is not feasible in practice. One resorts then to approximation through a small number of moments, basically mean and variance of the distribution leaving us with an incomplete knowledge of the process.

To overcome this difficulty, one possibility is to sample interarrival values: this is referred to as a "Renewal process" [77]. Having the same interarrival time yields the deterministic process in which the same sequence of information is sent periodically.

This characterization of the process is made possible by using the second order statistics which outline relations between the occurrences at two time instants. They are Autocorrelation and Autocovariance [78] and [79], which can be represented in frequency domain. This complete characterization of the process can be found in [80] and [81].

b) Markov Process

Since the knowledge of the multivariate pdf is required to fully characterize the stochastic process, which is difficult to achieve in practice, the Markov process with its property [82], is more practicable for mathematical analysis and gives rise to many applications in the domain that is the scope of this project.

Let $\{N(t), t \geq 0\}$ be a time function considered as the size of the population at time t , the classical Markov property is imposed as a restriction on the process $N(t)$, i.e. given the value of $N(s)$, the values of $N(s+t)$ for $t > 0$ are not influenced by the values of $N(u)$ for $u > s$. In other words, the way in which the entire past history affects the future of the process is completely determined by the current state of the process. This is written mathematically as follows [83]:

A stochastic process $\{N(t), t \geq 0\}$ with set T and discrete state space is said to be a continuous-time Markov Chain if:

$$\Pr[N(t_{n+1}) = n_{n+1} / N(t_n) = n_n, \dots, N(t_1) = n_1] = \Pr[N(t_{n+1}) = n_{n+1} / N(t_n) = n_n], n \geq 1 \quad (3.1)$$

Equation (3.1) represents the probability that the process makes a transition from state $n_n = i$ at time $t = t_n$ to the state $n_{n+1} = j$ at time $t_{n+1} = t + h$ for $h > 0$.

Such a probability denoted $p_{ij}(t)$ is referred to as a state transition probability for the Markov process and its discrete state space makes it a Markov Chain.

- Discrete Time Markov Chain (DTMC)

A DTMC is a process characterized by the transition probabilities from one state i to another j , possibly to the same, at well determined instant t_n denoted $P_{ij} = \Pr[X_{n+1} = j | X_n = i]$. These one-step probabilities are arranged in the one-step probability transition matrix $P = \{P_{ij}\}$ with

the normalization condition $\sum_{j=0}^{\infty} P_{ij} = 1$, for all i .

The transition matrix P is such that if the initial state vector $(p_0^0, p_1^0, \dots, p_n^0)$ is known (at time t_0), we can get the state vector at time t_1 by applying the one-step transition matrix P as follows:

$$[p_0^1, p_1^1, \dots, p_n^1] = [p_0^0, p_1^0, \dots, p_n^0]P. \quad (3.2)$$

The probabilities at the equilibrium are defined by:

$$\pi_i = \lim_{n \rightarrow \infty} \Pr[X_n = i]. \quad (3.3)$$

The system being invariant under this condition, one should have:

$$\pi P = \pi, \text{ with } \pi = [\pi_0, \pi_1, \dots, \pi_n] \text{ and } \sum_{i=0}^{\infty} \pi_i = 1. \quad (3.4)$$

- Continuous Time Markov Chain (CTMC)

The continuous-time Markov Chain is characterized by transition as in DTMC but occurring in continuous time; this requiring therefore additional analysis elements.

Assume $X(t)$ is a time homogenous Markov chain; the probability to move from state i at time t to state j at time $t + h$, $h > 0$ is defined by:

$$p_{ij}(t) = \Pr[X(t+h) = j | X(t) = i], t > 0, h \geq 0. \quad (3.5)$$

The normalization conditions suggest: $0 \leq p_{ij}(t) \leq 1$, $\sum_{j=0}^{\infty} p_{ij}(t) = 1$, and $p_{ij}(0) = \delta_{ij}$.

The probability matrix is defined by: $[P(t)]_{ij} = p_{ij}(t)$ and it follows that $P(0) = 1$ where 1 is the unit matrix.

From the probability density, the generator matrix of the Markov Chain is defined as:

$$Q = \lim_{h \rightarrow 0} \frac{P(h) - 1}{h},$$

whose elements $Q = \{q_{ij}\}$ can be interpreted as follows:

For a small time Δt , the transition probability from state i to state j during that interval is approximately equal to: $p_{ij} = \Delta t q_{ij}$. Further, if we let $q_i = -q_{ii}$, the infinitesimal generator is:

$$Q = \begin{pmatrix} -q_0 & q_{01} & q_{02} & \cdots & q_{0m} \\ q_{10} & -q_1 & q_{12} & \cdots & q_{1m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_{m0} & q_{m1} & \cdots & \cdots & -q_m \end{pmatrix} \quad (3.6)$$

The infinitesimal generator allows the restoration of the transition matrix $P(t)$ which may be

$$\text{written as } P(t) = e^{Qt} \text{ or by using its series: } P(t) = 1 + \sum_1^{\infty} \frac{Q^i t^i}{i!} \quad (3.7)$$

Given that the process starts in state i at time t , it has made transition to state j at time $t + \Delta t$ with probability:

$$\Pr[X(t + \Delta t) = j | X(t) = i] = q_{ij} \Delta t + O(\Delta t), \quad i \neq j \text{ and } \Pr[X(t + \Delta t) = i | X(t) = i] = 1 - q_{ii} \Delta t + O(\Delta t), \quad i = j,$$

where the approximate value $p_{ij} = \Delta t q_{ij}$ as $O(\Delta t) \rightarrow 0$.

For these probabilities to be conserved, that is to add up to 1, the off-diagonal elements of Q must be non-negative while the diagonal elements satisfy the condition $q_{ii} = \sum_j q_{ij}$.

If $p_i = \Pr[X(t) = i]$, the evolution of the Markov Chain is given by the first-order differential equation [84]:

$$\frac{\partial p_i}{\partial t} = p_i Q. \quad (3.8)$$

The probability that no transition happens is then:

$$\Pr[X(t + \Delta t) = i \mid X(t) = i] = e^{-q_{ii}\Delta t}.$$

The sojourn time in the Markov continuous-time chain is exponentially distributed with mean $1/q_{ii}$. At equilibrium, the continuous time probabilities are such that $\pi Q = 0$.

c) Renewal Process

The Poisson process finds its generalization in the Renewal process [85]. In fact, the Poisson process is a continuous-time Markov process on the positive integers (including 0 usually) which has independent and identical distributed holding time (exponential) at each integer i before advancing with probability 1 to the next integer $i + 1$. The same idea characterizes the Renewal process except that the holding time distribution is general.

Let $\{S_i\}$ be a sequence of independent identically distributed random variables, where S_i is the i -th holding time (sojourn time).

Define for $n \geq 0$ the n -th jump time as $j_n = \sum_{i=1}^n S_i$ and the intervals $[j_n, j_{n+1}]$ as renewal

intervals, then the random variable $X_{t \geq 0}$ given by: $X_t = \sum_{n=1}^{\infty} I_{\{j_n \leq t\}} = \sup \{n : j_n \leq t\}$ is called the

“Renewal Process”.

Y4

d) Semi-Markov Process

By combining a Markov process and a Renewal process, one obtains a Semi-Markov Process (SMP). In the DTMC, transitions take place at well determined times t_i while in CTMC with the supplementary elements, the transition times are exponentially distributed. For an SMP, the sojourn time $f_{ij}(\tau)$ depends on the initial state i and the final state j . It comes with the SMP that the time is a continuous process and the state space is a discrete process. Also, it cannot be considered as a Markov process since the present is not sufficient to determine the future.

Let a stochastic process $\{X_n : n = 0, 1, 2, \dots\}$ whose values are taken in an set $S = \{0, 1, 2, \dots\}$. Define the transition times T_0, T_1, T_2, \dots such that $0 \leq T_0 \leq T_1 \leq \dots$, the two-dimensional process $(X, T) = \{X_n, T_n : n = 0, 1, 2, \dots\}$ is called the Markov Renewal process if it has the following property [86]:

$$\Pr[X_{n+1} = j, T_{n+1} - T_n \leq t | X_n], j \in S, t \geq 0. \quad (3.9)$$

Assuming the Markov Renewal Process to be time homogeneous, define:

$$q_{ij}(t) = \Pr[X_{n+1} = j, T_{n+1} - T_n \leq t | X_n = i]. \quad (3.10)$$

$$p_{ij} = \lim_{t \rightarrow \infty} q_{ij}(t) \text{ and } F_{ij}(t) = \frac{q_{ij}(t)}{p_{ij}}, \quad (3.11)$$

where p_{ij} is the transition probability of the embedded Markov Chain $\{X_n : n = 0, 1, 2, \dots\}$ and $F_{ij}(t)$ is the cumulative distribution function of the sojourn time of the process in state i and is defined by:

$$F_{ij}(t) = \Pr[T_{n+1} - T_n \leq t | X_{n+1} = j, X_n = i]. \quad (3.12)$$

Let $N_j(t)$ be the number of visits of the process to state j during the interval $[0, t]$, and define the counting process $N(t) = \sum_{j \in S} N_j(t)$.

Let $Y(t) = Y_{N(t)}$ be the state of the Markov Renewal process (X, T) then, the process $\{Y(t)\}$ defines the Semi-Markov process [86].

A Markov Renewal process therefore characterizes the two-dimension transition time- system state of the process at that epoch. A Semi-Markov Process is therefore the combination of the system state at an arbitrary time point according to the Markov Renewal process, and the number of times the process complete all the states of the chain up to time t following the Markov Renewal counting process.

e) Self-Similar and Long-range dependence process

The long-range dependence is exactly the process we would like to prevent in the traffic model of this topic in view to provide an acceptable delay in the queuing system. Therefore, its analysis is found in Appendix A-1.

The stochastic characterization of the sources and therefore the characterization of traffic models with network parameters or descriptors give the basis requirement of the statistical multiplexer input.

3-3-2) Traffic descriptors: Arrival process and burstiness characteristics

The traffic bursts are structures characterized by a successive occurrence of different interarrival times, first order descriptors such as the ratio of peak to average rate are well placed for the analysis. The high value of the peak rate/mean rate indicates a high variability of the arrival pattern [86].

This requires a more relevant approach to capture the variability of the packet arrival. Therefore, the index of dispersion has been proven more significant than the simpler indexes such as coefficient of variation to quantify the burstiness [87].

The index of dispersion of interarrival times $I_\tau(t)$ is defined by:

$$I_\tau(t) = \frac{\text{Var}[\tau]}{nE^2[\tau]}. \quad (3.13)$$

Alternately, the index of dispersion for intervals $I_N(t)$ of the counting process [87] can be calculated as the previous index (3.13) and is given by:

$$I_N(t) = \frac{\text{Var}[N(t)]}{E[N(t)]}. \quad (3.14)$$

The magnitude and rate of increase in these traffic descriptors show the degree of correlation in the arrival process. The correlation among successive interarrival times in the aggregate packet arrival process with the related index of dispersion [88] for k successive arrivals is then given by:

$$I_\tau(t) = C_1^2 + \frac{2 \sum_{j=1}^n \sum_{k=1}^j \text{Cov}[X_j, X_{j+k}]}{nE^2[X]}, \quad (3.15)$$

which for a stationary arrival process, the limiting values as n and t tend to infinity are shown to be related to the normalized autocorrelation coefficients $\rho_\tau(j)$ for $j = 1, 2, \dots, n$ as follows [88]:

$$I_N = I_\tau = C_\tau^2 \left[1 + 2 \sum_{j=1}^n \rho_\tau(j) \right]. \quad (3.16)$$

This cumulative covariance is very important for the multiplexer application since the large packet delays mentioned above is due neither to the single time interval C_1^2 nor the covariance, but to the cumulative effects of many small individual covariances as proved by Sriram and Whitt [89], and Livny [90]. These indexes of dispersion are valuable tools for characterizing the

arrival processes in the queue but have limited application for deriving an explicit measure of queue performance.

3-4) QUEUING SYSTEMS BACKGROUND

In a packet switching system, we normally encounter some nodes where the arriving packet needs to be processed (analysis of routing information, rate adaptation, multiplexing) before continuing on its way. Often, the incoming traffic may be bigger than the processing capacity of the node in the short term, but this is crucial for the quality requirements. The solution is then to store the arriving packets into a buffer accordingly and process them one by one in the compliance with the queue order of packets. This is called statistical multiplexing (Figure 3.3). Statistical multiplexers are therefore modelled as queuing systems with finite buffer space, served (on a service discipline basis) by one or more transmission links of fixed or varying capacity [60].

A queuing system is made up of queuing techniques that are QoS oriented, and queuing models that are packet processing related. In the following sections, we will describe both although we are focusing on queuing models.

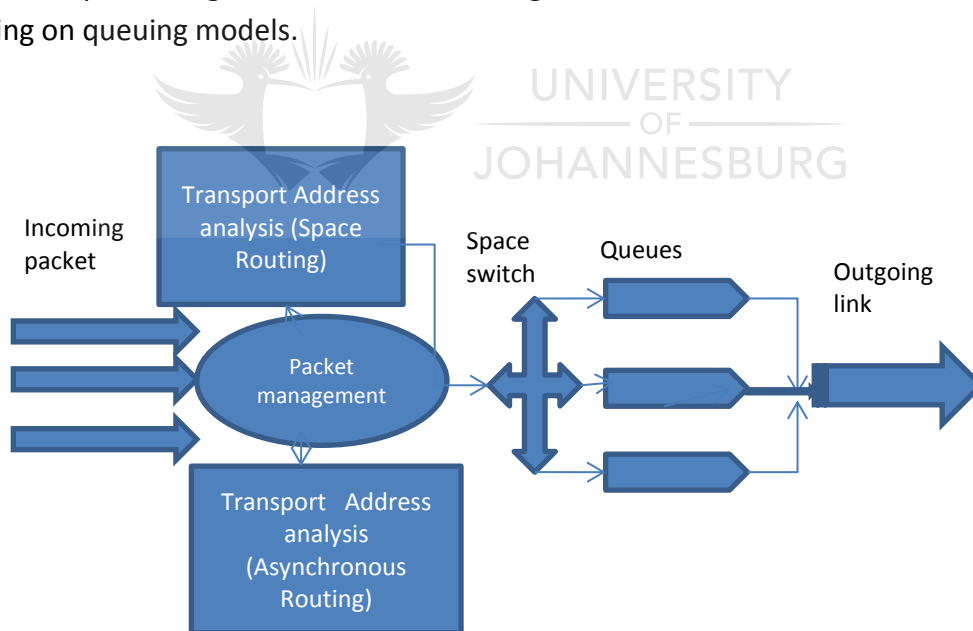


Figure 3.3: Statistical multiplexer

3-4-1) History of Queuing Theory

The pioneer is Erlang [91] whose name gave the measure of the traffic intensity by introducing the Poisson process to congestion theory between 1909 and 1929. The statistical equilibrium was defined by balance state equations called Chapman-Kolmogorov equations. Pollazeck [92]

however came with a non-equilibrium study using finite intervals while Bailey [93] leaning on Takacs [94] generating functions, introduced the time dependent solution. From this foundation, a huge amount of studies have followed. Kendall [95] studied the Non-Markovian queues by introducing the method of Embedded Markov Chains while Cox [118] came with a technique known as supplementary variable technique to provide queue solutions. This trend of stochastic processes associated with the theory of queues is carried on up to the time we are writing this thesis, which leans on the work done by Qiang and Kobayashi [96], [97]; and Reiser and Kobayashi [98]. Let us start by mentioning the queuing techniques, since their impact in the queuing systems is relevant.

3-4-2) Queuing Techniques background

The queuing technique is a group of relative QoS techniques. Also known as schedulers, they provide different queue levels and therefore a way of handling different classes of services.

a) Weighted Fair Queuing (WFQ)

The WFQ [99] is a flow-based technique where the arriving packets are classified according to the same source address/port and destination address/port combination. These Different traffic flows are queued to prevent lack of bandwidth: hence, "fair". A weight is assigned to prioritize those flows according to some schemes, usually another QoS technique that can be IP precedence or DiffServ.

b) Class Based Weighted Fair Queuing (CBWFQ)

The CBWFQ [100] is the recent form of the WFQ technique that defines user traffic classes in compliance with the protocol, the port, the access control, input queue or DiffServ bits. The bandwidth and the queue limit can be assigned to the classes which are managed accordingly, causing the queue to drop packets when it fills up. The CBWFQ may be used with a feature called Low Latency Queuing (LLQ) [101] which offers delay sensitive traffic priority handling over other traffic types.

c) Weighted Random Early Detection (WRED)

Instead of trying to deal with congestion after it occurs like the previous schemes, the WRED [102] anticipates it by randomly dropping packets from a selected traffic type before the queue fills up. Consequently, the technique is not effective for protocols with the retransmit feature. It relies therefore on IP precedence bits to help decide which packet to drop.

3-4-3) Queuing models background

The queuing model is a group of absolute QoS, defining networks parameters that characterize bandwidth usage, transfer delay and delay variation based on the queue analysis.

a) Queue model description and function

Definition of Queue. The queue is defined according to the specification of characteristics which describe the system: A/S/N/D/C/P [103].

- **A** is the arrival process: it is the arrival process distribution of the traffic as modelled at the source involving a degree of uncertainty in the interarrival time and number characterizing the arriving variable.
- **S** is the service time process: it is the processing of the arriving variable featured by its time process distribution.
- **N** is the number of servers: it is the number of servers required for processing the arriving variables
- **D** is the queue discipline: it defines how to manage the arrivals before the service occurs.
- **C** is the queue capacity that can be finite or infinite.
- **P** specifies the population size from which the sample is drawn.

In this thesis, provided the randomness of most of the variables in telecommunication, we shall limit ourselves to A/S/N/D/C. Since we are concerned with RTP traffic, we assume in the first place that D is on a First Come First Served (FCFS) queuing discipline. This reduces our queue description to A/S/N/C. All along the queue models, we will consider an infinite queue with one server; thus reducing our study to A/S/1. The model reached therefore needs to be sorted out.

3-4-4) Queuing model solutions

It is always desirable to have at least a method (whether analytical or numerical) that allows for the solving of queuing models. There are three solutions given up to now:

- The matrix method;
- The moment generating function method; and
- The fluid method.

a) The matrix method

The traffic model is composed of packets from a Markov Modulated source [104] implying an analysis of the statistical multiplexer in a discrete time. The resulting matrix solution is obtained from solution of a system of linear equations at the equilibrium state buffer occupancy. Taking

into account the queue size and the number of Markov states in the chain, solving accurately this system is difficult hence forth some assumptions [104]:

- The time is assumed to be slotted: the transition matrix determines the probability the emitting modelled source may send packets per time slot.
- The service time is deterministically measured to one time slot.
- The system is a queue with k packets size capacity.

The probability of k packets to be generated is assumed known per time slot and independent from one slot to another: $\Pr[x = k / \text{timeslot}] = a_k$ such that $\sum_{k=1}^{\infty} a_k = 1$.

The queue occupancy at the well determined instants is a random variable whose evolution is given by the following transition matrix M :

$$M = \begin{pmatrix} a_0 & a_1 & a_2 & \dots & a_{k-1} & \sum_{i=k}^K a_i \\ a_0 & a_1 & a_2 & \dots & a_{k-1} & \sum_{i=k}^K a_i \\ 0 & a_0 & a_1 & \dots & \dots & \vdots \\ 0 & 0 & a_0 & \dots & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_0 & \sum_{i=k}^K a_i \end{pmatrix} \tag{3.17}$$

This matrix is derived as follows in Figure: 3.4:

		<i>0</i>	<i>1</i>	<i>2</i>	<i>...</i>	a_{k-1}	K
$M =$	<i>0 in the queue</i>	a_0	a_1	a_2	\dots	a_{k-1}	$\sum_{i=k}^K a_i$
	<i>1 in the queue</i>	a_0	a_1	a_2	\dots	a_{k-1}	$\sum_{i=k}^K a_i$
	<i>2 in the queue</i>	0	a_0	a_1	\dots	a_{k-2}	
	<i>3 in the queue</i>	0	0	a_0	\dots	a_{k-3}	
	\vdots					\vdots	
	\vdots					\vdots	
	<i>K in the queue</i>	0	0	0	\dots	a_0	$\sum_{i=k}^K a_i$

Figure 3.4: Matrix determination

The entry (i, j) of M represents the probability transition from i packets in the queue at slot t to j packet in the queue at slot $t + 1$. In the first entry, the entry $(0,0)$ is the probability from 0 in the queue at slot t to 0 packet in the queue at slot $t + 1$. Since there is no arrival, this entry is a_0 . By continuing in that fashion, if we keep to row 0, the transitions $(0,1)$ requires 1 arrival, hence a_1 , $(0,n)$ requires n arrival where $n \leq k$. Row 1 is for $(1, x)$ transition. As one packet is assumed to be served per slot, with the entry $(1,0)$, there is 1 packet in the queue at slot t and 0 packet at slot $t + 1$, which means no arrival at all, hence a_0 . That makes this row identical to row 1. Row 2 corresponds to the entry $(2, x)$. As one packet is assumed to be served at the time, it is impossible to have 0 packet in the queue, therefore 0 for column 0 and a_0 for column 1 provided the 1 serviced is only replaced: no transition in the queue, while $(2,2)$ will probably transitioned to a_1 .

In the same way, the matrix is filled up and which in practical use is made of "0" below the diagonal "1", gives rise to the *upper Hessenberg* matrix [105].

As mentioned above, the probability depends on the modulator state also known as phase. Let the modulator be in i -th phase of a set of S phases. A transition $i \rightarrow j$ still requires $k = j - i + 1$ packets. The probability of k arrival packets is given by the state of the modulator which is likely to change in the last time slot, variation that drives one to consider a bivariate process: queue content $X(t)$ and the phase of the modulator $Y(t)$. The knowledge of $(X(t), Y(t))$ is sufficient to determine the statistic of the process for $t > t_0$.

Define then a state vector for a single phase modulator as $p = (p_0, p_1, \dots, p_{k-1})$ of the matrix, for S phases we have:

$$P_S = \begin{pmatrix} P_{0,0} & P_{0,1} & \dots & P_{0,s-1} \\ P_{1,0} & P_{1,1} & & \\ \vdots & \vdots & \ddots & \vdots \\ P_{k,0} & P_{k,1} & \dots & P_{k,s-1} \end{pmatrix}$$

where $p_{k,s}$ defines the probability of having k packets in the buffer with the modulator in phase s . Changing the state vector a_i into the transition matrix A_i of $S \times S$.

$$(A_i)_{mn} = \Pr[\text{transition } m \rightarrow n \text{ and packets are generated in state } n \text{ per slot}].$$

Define the transition matrix T , when the generator is in phase n the probability of generating

i packet is $a_i(n)$, then:

$$(A_i)_{mn} = T_{mn} a_i(n). \tag{3.18}$$

The resultant matrix as previously achieved in (3.17) is extended to the multidimensional matrix as follows also derived as the previous matrix:

$$M = \begin{pmatrix} A_0 & A_1 & A_2 & \cdots & A_{k-1} & \sum_{i=k}^K A_i \\ A_0 & A_1 & A_2 & \cdots & A_{k-1} & \sum_{i=k}^K A_i \\ 0 & A_0 & A_1 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & & & \vdots & & \vdots \\ & & & \cdots & A_0 & \sum_{i=k}^K A_i \end{pmatrix} \tag{3.19}$$

Another analysis in this category is given to us by Miller [106] exploiting Neuts [107]: the matrix geometric analysis method. Such a solution proposed in [108] is presented in Appendix A-2.

b) The Probability Generating Function (PGF) method

The main goal of the probability generating function (PGF) as a solution of the queue computation is the introduction of the complex plane to widen the analysis to many complex applications [109].

It is based on the Laplace-Stieltjes transform which for a continuous time function is denoted:

$$\tilde{f}(s) = \int_0^\infty e^{-st} dF(t),$$

where $F(t)$ is the probability distribution function; or on z-transform for the discrete time function as:

$$P(z) = \sum_{k=0}^\infty p_k z^k \text{ with } P(z) \text{ being the probability mass function.} \tag{3.19}$$

For a discrete probability distribution $p_k, k = 0,1,2,\dots$ the n order probability is given by

$$\frac{1}{n!} \frac{d^n P(z)}{dz^n} \Big|_{z=0} = p_n \text{ and the mean of a random variable } X \text{ is } \frac{dP(z)}{dz} \Big|_{z=1} = \bar{X}.$$

Applying Laplace-Stieltjes transform to model the general interarrival (GI) process with a single server queuing where the service time is deterministic equal to 1 time slot, we have:

If S_n is the size of the queue at time slot n when there are A_n arrivals, the next slot is given by:

$$S_{n+1} = [S_n - 1]^+ + A_n, \text{ and } [x]^+ = \max(0, x). \text{ Therefore one can deduce:}$$

$$\Pr[S = k] = \Pr\{([S - 1]^+ + A) = k\}.$$

Let $p_k = \Pr[S = k]$, $a_i = \Pr[A = i]$ and $b_i = \Pr[B = i]$ with $B = [S - 1]^+$.

Assuming the two random variables A and B are independent provided the GI assumption, the distribution of their sum using the z-transform property is the convolution as follows:

$$\left(\sum_{k=0}^{\infty} p_k z^k\right) = \left(\sum_{k=0}^{\infty} b_k z^k\right) \left(\sum_{k=0}^{\infty} a_k z^k\right) \text{ denoted } P(z) = B(z)A(z).$$

According to $B = [S - 1]^+$, the shift can be materialized in z-transform by:

$$B(z) = \frac{s_0(z-1) + S(z)}{z},$$

yielding:

$$S(z) = \frac{A(z)(1-z)}{A(z)-z} \Pr[S = 0]. \quad (3.21)$$

Since $\Pr[S = 0]$ is the probability of the empty system, one can write $\Pr[S = 0] = 1 - \rho$ where ρ is the system load given by $\frac{dA(z)}{dz} \Big|_{z=1} = \rho$.

Assume $A(z)$ geometrical arrival processes per time slot with parameter α , $a_n = \Pr[A = n] = \alpha(1-\alpha)^n$ with $n=0, 1, \dots$ and mean load $\rho = \frac{1-\alpha}{\alpha}$.

$A(z) = \sum_{k=0}^{\infty} a_k z^k = \frac{\alpha}{1-(1-\alpha)z}$ with (3.20) leading to the generating function:

$$S(z) = (1-\rho) \sum_{n=0}^{\infty} (\rho z)^n. \quad (3.22)$$

By identification with the definition given by $P(z) = \sum_{n=0}^{\infty} p_n z^n$, it comes that:

$$p_n = (1-\rho)\rho^n.$$

c) The Fluid method

The idea of the fluid approximation comes from the fact that the discrete arrival is insignificant to cause congestion in the queue, as the rain drops are for a flood [110]. Hence, a leaky bucket

queue model whose input is the continuous “fluid” information of varying intensity and output is the leak rate at a constant service time.

Let $\{Y(t), t \geq 0\}$ be the Continuous Time Markov Chain that takes values $\{0, 1, 2, \dots, N\}$, and let the infinitesimal generator be the matrix $Q = \{q_{ij}\}$. When in state j , fluid arrives at the rate a_j , served at a constant rate c . Assuming the buffer empty, the net rate change in the queue is $r_j = a_j - c$, which is an underload state when negative; and positive in the case of an overload state.

Let a continuous random variable $X(t)$ be the queue size at time t satisfying $0 \leq X(t) \leq K$, where K is the buffer size. We are interested in the solution of the equilibrium overflow which probability is given by:

$$G(x) = \lim_{x \rightarrow \infty} \Pr[X(t) > x]. \quad (3.23)$$

Since the process is bivariate, i.e. depending on the stochastic process $(X(t), Y(t))$, the system is in the equilibrium overflow if and only if:

$$F_j(x) = \lim_{t \rightarrow \infty} F_j(x, t) = \Pr[X \leq x, Y = j]. \quad (3.24)$$

For more accuracy, define the $(N + 1)$ -dimensional vector $F(x) = [F_0(x), F_1(x), \dots, F_N(x)]$ as the probability function at the equilibrium. Clearly:

$$G(x) = 1 - F(x), \text{ with } 1 = [1, 1, 1, \dots, 1]^T.$$

The time evolution of $F_j(x, t)$ is governed by the following equation [111]:

$$F_j(x, t + \Delta t) = \sum_{i, i \neq j} q_{ij} \Delta t F_i(x - r_{ij} \Delta t, t) + (1 - \sum_{i, i \neq j} q_{ij} \Delta t) F_j(x - r_j \Delta t, t) + O(\Delta t). \quad (3.25)$$

One can say at infinitesimal time $t + \Delta t$ that the probability of having the buffer filled at most x while the modulator is in state j consists of two terms.

The first applies to cases where the modulator was in one of the states i other than j , ($i \neq j$) at time t to end up in state j at $t + \Delta t$. There should then be a probability $q_{ij} \Delta t$ to make a transition from i to j ($O(\Delta t)$ within). Meanwhile, somewhere during Δt in the buffer, the content will change, hence the net rate of change r_{ij} which ends up with $X \leq x$ at $t + \Delta t$ so that one should have at t , $X \leq x - r_{ij} \Delta t$.

The second term probability excludes only the transition of the modulator.

Knowing that $O(\Delta t)$ goes faster to 0 as Δt tends to 0, subtracting $F_j(x, t)$ from both sides, we get the differential equation:

$$\frac{\partial F_j(x, t)}{\partial t} + r_j \frac{\partial F_j(x, t)}{\partial x} = \sum_i q_{ij} F_i(x, t) \quad (3.26)$$

which solution at the equilibrium is given by: $r_j \frac{\partial F_j(x)}{\partial x} = \sum_i q_{ij} F_i(x)$, $j = 0, 1, \dots, N$.

Introducing the rate matrix $R = \text{diag}\{r_j\}$ and the generator matrix $Q = \{q_{ij}\}$, the solution of the equation is given by $F(x) = \phi e^{QR^{-1}x}$ where ϕ is a vector of constants.

In conclusion, computing a fluid queue comes from finding the eigenvalues z_i of the matrix QR^{-1} so that we can have an exponential form $e^{z_i x}$.

The above queue solutions and the traffic descriptors allow analysing the multiplexer performance and modelling the queue.

3-4-5) Performance tools of statistical multiplexer

The statistical multiplexing gain (SMG) [112], a parameter that quantifies the performance of the multiplexer, is defined as the ratio of the number of variable bit rate (VBR) sources that can be multiplexed on a fixed capacity link under the specified delay or loss constraint to the number of sources that can be supported on the basis of peak rate allocation.

To determine and maximize the SMG, conditions that rationalize traffic characteristics to performance constraints and relate traffic model to system parameters have to be set.

Eckberg in [113] and [114] estimated index of dispersion of the queue size using the peakedness functional through the second-order traffic descriptors.

The peakedness of the queue [115] represents the ratio of the variance and the expectation of the number of busy servers in an infinite server system driven by a stationary traffic process.

The peakedness functional is defined as:

$$Z(F) = \frac{E[L(t)]}{\text{Var}[L(t)]}, \quad (3.27)$$

determines the teletraffic engineering parameters in planning system resources such as to estimate the blocking probability of finite server systems as presented by Fredericks [116]

where $Z(F)$ is an approximation. For example, the Erlang’s blocking formula to estimate the additional servers imposes $Z > 1$.

The characterization of the sources has been generalized. We can now do the same within the focus of our topic: the voice source.

3-4-6) Voice packets statistical multiplexing

a) Packet arrival process

In a FCFS discipline, if the arriving packet finds the queue empty with probability $1 - \rho$ where ρ is the traffic intensity, the waiting time is zero.

If the arriving packet finds n other packets in front, it will experience a random waiting time that is the sum of n independent exponentially distributed random variables with probability density function (pdf) [118]:

$$f_w(t) = (1 - \rho)\delta(t) + \mu(1 - \rho)e^{-\mu(1-\rho)t}, \quad t \geq 0, \tag{3.28}$$

where $f_w(t)$ is an exponential distribution with jump at the origin, μ is the service rate.

b) Statistical characteristic of packetized voice process

The voice packet sequence from a single voice source with a voice active detector is characterized by a fixed time interval T ms which is its packetization time during activities, and no packet arrivals at all during silence periods as shown in Figure 3.5.

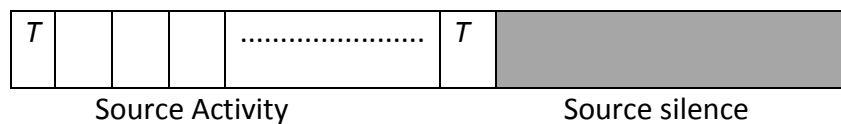


Figure 3.5: Voice source structure

The active source time is characterized by a mean time α^{-1} ms and a number N of packets with mean $E[N]$ and the silence intervals have as mean time β^{-1} ms.

Each packet in the stream, since each source is homogeneous, is characterized by a length T and a probability $(N - 1) / N$ during activity in which we add the packetization period T of its next occurrence that will happen during silence with probability $1 / N$.

As characterized, the voice stream at the multiplexer input from one source is geometrically distributed with arrival mean $\frac{1}{\alpha T}$. But the aggregate traffic from many sources exhibits some correlation at the buffer discarding the process of being renewal. Some discrepancies such as

packet loss and delay are experienced due to the variability. This variability is due to the fact that the number of sources in activity varies. We are suggesting hence forth, the diffusion approximation process as solution of the queue system.

The source being modelled as a Renewal process, the interval time distribution is as in (3.28) and [117]:

$$F(t) = [(1 - \alpha T) + \alpha T(1 - e^{-\beta(t-T)})]U(t - T),$$

where $U(t)$ is the unit step function.

This corresponds as in the previous results to a geometric distribution in terms of number of voice packets with mean $1/\alpha T$ with an exponential distribution talking periods of mean $1/\alpha$, and silence period with mean $1/\beta$, T being the packetization time interval.

As corollary, the renewal theory results can be used to evaluate the mean and the variance of the number of arrival within an interval time for its characterization.

Defining Laplace-Stieltjes transform $L_F(s) = \int_0^{\infty} e^{-st} dF(t) = [1 - \alpha T + \alpha T\beta/(s + \beta)]e^{-sT}$, we have

the mean packet arrival $\lambda = -\frac{1}{L_F(0)} = 1/(T + \alpha T/\beta)$.

Let $N(0, t)$ represent the number of arrivals of a stationary renewal process in the interval $[0, t]$.

Define the i -th moment $m_i(t) = E[N(0, t)]$ and its Laplace Transform $\hat{m}_i(s) = L[m_i(t)]$.

From the results of Takacs [125] we can write: $\hat{m}_1(s) = \lambda/s^2$ and $\hat{m}_2(s) = \frac{\lambda}{s^2} \left(\frac{1 + L_F(s)}{1 - L_F(s)} \right)$.

The inverse Laplace transform gives $m_1(t) = \lambda t$.

The indexes of dispersion of interval time and of count are related according to the following:

$$\lim_{t \rightarrow \infty} I(t) = \lim_{t \rightarrow \infty} \frac{\text{Var}[N(0, t)]}{m_1(t)} = \frac{\text{Var}[X]}{E^2[X]},$$

where $\{X(t)\}$ is the interarrival time process.

Exploiting these results, one can write:

$$\lim_{t \rightarrow \infty} \frac{\text{Var}[N(0, t)]}{m_1(t)} = \frac{1 - (1 - \alpha T)^2}{(\alpha T + \beta T)^2}.$$

Considering the superposition of n identical independent voice packet sources, the number of arrivals during $[0, t]$ from the i -th stream $N_i(0, t)$ of a stationary process, the aggregate gives:

$$N_n(0, t) = \sum_{i=1}^n N_i(0, t),$$

therefore, the mean of the superposition is given by: $M_1(t) = nm_1(t)$.

The index of dispersion is written as: $\frac{\text{Var}[N_n(0, t)]}{E[N_n(0, t)]} = \frac{\text{Var}[N(0, t)]}{E[N(0, t)]}$.

Thus the superposition keeps the indexes of dispersion identical, and this characteristic may be exploited in the approximation.

b) Approximating the superposition of voice processes

To approximate the superposition of voice processes, three propositions are given in [119] which can be summarized as follows:

- For a Poisson Process, the index of dispersion is $I(t) = C_k^2 = 1$ for all t and k .
- For a renewal process $C_k^2 = C_1^2$ for the index of time interval, then $\text{Cov}[X_i, X_j] = 0$.
- For a superposition of n independent and identically distributed renewal processes, $I(\infty; n) = \lim_{t \rightarrow \infty} I(t; n) = C_{\infty n}^2 = \lim_{k \rightarrow \infty} C_{kn}^2 = C_{11}^2$. Here C_{11}^2 is the squared coefficient of variation of a single interval ($(k=1)$ in one of the renewal processes ($n=1$) being superposed.
- The superposition of n independent and identically distributed renewal processes, each with rate λ/n tends to a Poisson process with rate λ as $n \rightarrow \infty$.

3-5) CONCLUSION

The identification of time scale in the superposition of voice packet traffic is crucial in determining the right buffer size. This results to a small queue size which derives basically from congestion due to a cumulative rate greater the service capacity when different sources start emitting in the same time-slot.

Contrary, the lack of time scale identification gives chances to the traffic burstiness and correlation that influence the queue according to the complementary probability

$G(X) = \Pr[X > x]$ where X is the random variable of the buffer occupancy in steady state. This necessitates larger queue size since sustained transmission of a number of sources at the peak rate leads to a build-up in the queue size for time durations that are functions of the burst state statistics.

However, to reach the complementary probability or to determine the queue size, one need to use one of the queue solutions as presented in this chapter which results in the A/S/N/D/C/P queue: the operation is called Queue modelling. In the next chapter we will present the theory behind queuing models.



CHAPTER 4

QUEUE MODELLING THEORY

4-1) INTRODUCTION

In the previous chapter, we saw that modelling a queue results in providing its parameters based on one of the queue solutions through the arrival and service time processes. The queuing model is thus a group of absolute QoS, defining the network parameters that characterize bandwidth usage, transfer delay and delay variation. It is therefore necessary to recall how to determine those network parameters for the different queue models. In the following sections, we focus on the queue models of the form A/S/1 for simplicity reasons in the FCFS mechanism.

4-2) DIFFERENT QUEUE MODELS

4-2-1) The M/M/1 queue

The M/M/1 queuing is modelled as a single-server with infinite capacity queuing systems having packets arriving according to a Poisson process (M) with an arrival rate λ packet per unit time and exponential (M) service time distribution with service rate μ packet per unit time. The memoryless property of the exponential distribution, here referred to as M, allows the Birth-Death process as a modelling tool for queue (Figure 4.1).

When a packet enters an empty system, its service starts at once. If we are dealing with a non empty system, the incoming packet is queued; and will enter the service facility once at the occurrence of the service completion of the packet ahead of it.

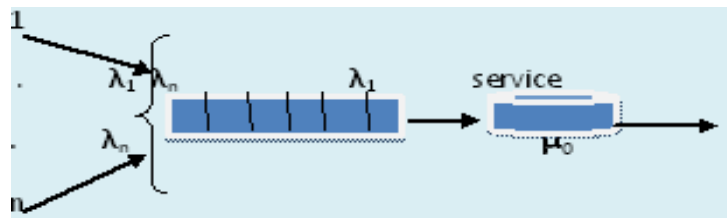


Figure 4.1: Queue (FCFS) facility

a) The Birth-Death Process

The Birth-Death process (Figure 4.2), a Markov Chain with time-homogeneous transition probabilities, is characterized by the fact that the discrete state variable changes by most 1 if it changes, during an infinitesimal time interval. The process thus satisfies the following probability distributions governing the number of birth-death in a specific time interval which depend on the length of the interval and not on its starting point [120]:

- The probability of exactly one birth occurring in an infinitesimal time interval Δt for a given population size at time t is $\lambda_n \Delta t + 0(\Delta t)$ where λ_n is a constant.
- The probability of exactly one death occurring in an infinitesimal time interval Δt for a given population size at time t is $\mu_n \Delta t + 0(\Delta t)$ where μ_n is a constant.
- The probability of more than one birth and the probability that more than one death in that infinitesimal time interval are both $0(\Delta t)$ provided that the process increases or decreases by 1.

This can be mathematically rewritten as follows:

$$\Pr[N(t + \Delta t) = n + 1 / N(t) = n] = \lambda_n \Delta t + 0(\Delta t), n \geq 0; \quad (4.1)$$

$$\Pr[N(t + \Delta t) = n - 1 / N(t) = n] = \mu_n \Delta t + 0(\Delta t), n \geq 1; \quad (4.2)$$

$$\Pr[N(t + \Delta t) = n / N(t) = n] = 1 - (\lambda_n + \mu_n) \Delta t + 0(\Delta t), n \geq 1 \quad (4.3)$$

$$\Pr[N(t + \Delta t) = k / N(t) = n] = 0(\Delta t), |k - n| \geq 2 \quad (4.4)$$

where, the quantity $0(\Delta t)$ is such that $\lim_{\Delta t \rightarrow \infty} 0(\Delta t) = 0$.

The Chapman-Kolmogorov's equation [121] is given by:

$$p_{ij}(t + h) = \sum_{k=0}^{\infty} p_{ik}(h) p_{kj}(t). \quad (4.5)$$

This means that in order to move from state i to state j in time $t + h$, the queue size process $N(t)$ moves to some intermediate state k in time h and then from k to j in the remaining time t .

By combining the Chapman-Kolmogorov equation (4.5) with the transition probability equations above (4.1), (4.2) and (4.3); by introducing Bayes' Formula [122], given by:

$$\begin{aligned}
 P_n(t) &= \Pr[N(t) = n] = \sum_{k=0}^{\infty} \Pr[N(t + \Delta t) = n / N(t) = k] \Pr[N(t) = k] \\
 &= P_n(t)[1 - (\lambda_n + \mu_n)\Delta t + o(\Delta t)] + P_{n-1}(t)[\lambda_{n-1}\Delta t + o(\Delta t)] + P_{n+1}(t)[\mu_{n+1}\Delta t + o(\Delta t)] + o(\Delta t); \quad (4.6)
 \end{aligned}$$

and by rearranging (4.6), it comes that:

$$\frac{P_n(t + \Delta t) - P_n(t)}{\Delta t} = -(\lambda_n + \mu_n)P_n(t) + \mu_{n+1}P_{n+1}(t) + \lambda_{n-1}P_{n-1}(t) \quad (4.7)$$

This tends to $\frac{dP_n(t)}{dt}$ as $\Delta t \rightarrow 0$.

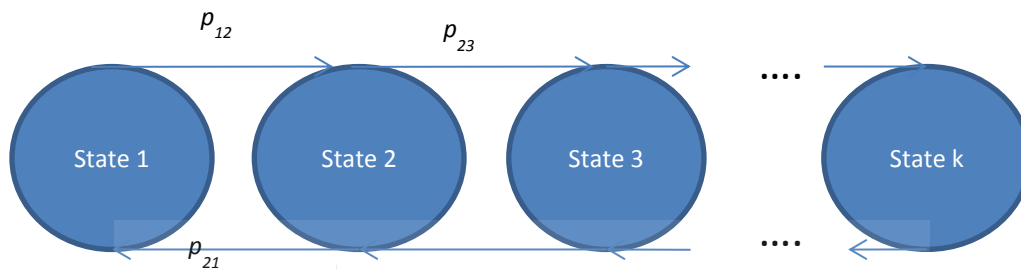


Figure 4.2: Markov Birth-Death Chain

The solution of the M/M/1 queue is based on the matrix method which defines the transition probabilities. Based on the solution obtained at the equilibrium state, the mean waiting time is:

$$W = \frac{\rho}{\mu(1 - \rho)}.$$

As one can see with this waiting time, the system loses its stability for the resource utilization close to 100%.

4-2-2) G/M/1 Queue

The G/M/1 queue results from the difficulties of achieving an acceptable arrival process in the M/M/1 queue. Its analysis leans on the Markov Embedded process that yields the matrix solution combined with the Probability Generating Function (PGF).

a) The matrix solution

The queue parameters are obtained through the Markov Chain embedded on arrival instants. This can be described by i number of packets in the system just before an arrival or better the system state to reach is given by the number in the system immediately before the arrival instants. This process is shown in the Figure 4.3.

Consider the system immediately before the i -th arrival having n_i number of packets in the system just before this arrival instant and service S_i .

If the system has served S_{i+1} packets between the i -th and the $(i+1)$ -th arrival, one can see that:

$$n_{i+1} = n_i + 1 - S_{i+1} \quad n_i = 0, 1, 2, \dots, \infty \quad \text{and} \quad S_{i+1} \leq n_i + 1 .$$

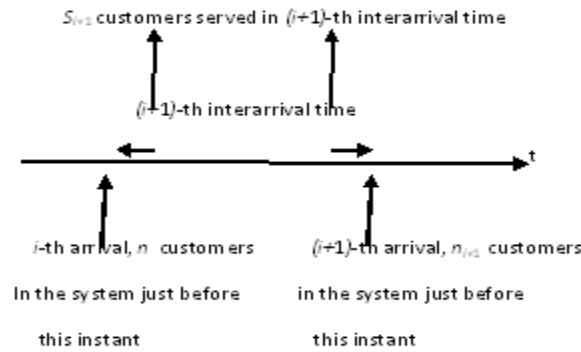


Figure 4.3: Embedded Markov Chain for G/M/1 queue

The equilibrium state of this embedded Markov Chain is reached when $i \rightarrow \infty$. Under this condition, let $n_{i+1} = k$ and $n_i = j$; the one-step probability transition of the chain is given by:

$$\begin{aligned} p_{jk} &= P[n_{i+1} = k \mid n_i = j], k > j + 1 \\ p_{jk} &= 0, \quad \text{otherwise,} \end{aligned} \tag{4.8}$$

where p_{jk} is the probability that $(k - (j + 1))$ packets get served between interarrival times. The equilibrium condition values are consequently obtained for the one-step transition probabilities from where the equilibrium state probability is derived and given by:

$$p_k = \sum_{j=0}^{\infty} p_j p_{jk} \quad \text{for } k = 0, 1, 2, \dots,$$

with the normalization condition $\sum_{k=0}^{\infty} p_k = 1$, and p_j the probability to find j packets in the system state considered but with an arbitrary arrival instant for $j = 0, 1, 2, \dots, \infty$.

To specify the transition probabilities, one introduces the probabilities α_n defined as the probability that n packets are served during an interval time with the system state number of at least n packets. This probability is conditioned by the interarrival time as follows:

$$\alpha_n = \int_{t=0}^{\infty} \frac{(\mu t)^n}{n!} e^{-\mu t} a(t) dt, \quad n = 0, 1, 2, \dots, \infty. \quad (4.9)$$

The balance equations [123] can be written as follows:

$$p_0 = \sum_{k=0}^{\infty} \alpha_{k+1} p_k, \quad p_j = \alpha_0 p_{j-1} + \sum_{k=0}^{\infty} \alpha_{k+1} p_{j+k}, \quad j = 1, 2, 3, \dots, \infty. \quad (4.10)$$

Define system boundaries if more than n packets are served during an interarrival time as the related probability b_n so that one can write $b_n = \sum_{k>n} a_k$, the transition probability matrix P takes the following form:

$$P = \begin{pmatrix} b_0 & a_0 & 0 & 0 & 0 & \dots \\ b_1 & a_1 & a_0 & 0 & 0 & \dots \\ b_2 & a_2 & a_1 & a_0 & 0 & \dots \\ b_3 & a_3 & a_2 & a_1 & a_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \quad (4.11)$$

b) Moment generating function solution

The moment generation solution is obtained by considering the probability expression α_n (4.9) as the coefficients of the z^n in the Laplace-Stieltjes transform expression $L_A(\mu - \mu z)$ of the probability density function (pdf) of the interarrival times [124]. This implies:

$$\sum_{j=0}^{\infty} \alpha_j z^j = \int_0^{\infty} e^{-\mu t(i-z)} a(t) dt = L_A(\mu - \mu z). \quad (4.12)$$

Then, the series expansion of $L_A(\mu - \mu z)$ yield the coefficients α_j of the z^j , which are used to give a solution to the balance equations (4.10) thereafter, the probabilities p_j , $j = 0, 1, 2, \dots, \infty$.

Let σ be the unique root of the equation $\sigma = L_A(\mu - \mu\sigma)$, the solution for the equilibrium state probabilities is given by [125]:

$$p_j = (1 - \sigma)\sigma^j, \quad j = 0, 1, 2, \dots, \infty. \quad (4.13)$$

As long as $0 < \sigma < 1$, the number of packets in the system at the embedded arrival instants is therefore a geometric distribution with parameter σ .

As a result, if we consider an arrival to a G/M/1 queue with waiting time W in the queue system, which operates in the FCFS queuing discipline. This yields as probability density function for the waiting time (3.28) as follows:

$$f_W(t) = (1 - \sigma)\delta(t) + \mu(1 - \sigma)e^{-\mu(1-\sigma)t}, \quad t \geq 0. \quad (4.14)$$

Exploiting the memoryless property of the distribution, we obtain the mean waiting time W as:

$$W = \sum_{n=1}^{\infty} \frac{n}{\mu} (1 - \sigma)\sigma^n = \frac{\sigma}{\mu(1 - \sigma)}. \quad (4.15)$$

The results of this queuing solution can then be derived from the results obtained from the M/M/1 by replacing ρ by σ . But, given the parameters ρ , σ and μ , or their estimates, approximating arrival process with the arrival rate $\lambda = \rho\mu$ can be obtained as follows.

If $\sigma > \rho$, the interarrival time has a hyper-exponential distribution.

If $\sigma = \rho$, the interarrival time has an exponential distribution.

If $\sigma < \rho$, the interarrival can be generalized with either an Erlang or shifted exponential distribution.

One notices the introduction of another parameter in the queuing system resulting from Cruz bounds [137], commonly denoted $R(\rho, \sigma)$.

4-2-3) M/G/1 Queue

The M/G/1 queue is the dual of the G/M/1 queue [126]. Where the embedded Markov points were the arrival instants in the G/M/1 queue, in the M/G/1 queue the embedded points are shifted to the departure instants. Since the departures are of interest, the distribution of interest is the service time distribution which is general and therefore can undergo the same approximations as the G/M/1 queue. This is observed even in the transition matrix P of the G/M/1 queue when compared to that of M/G/1:

$$P_{MG1} = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & b_n \\ a_0 & a_1 & a_2 & \vdots & \vdots \\ 0 & a_0 & a_1 & a_2 & b_2 \\ \cdots & 0 & a_0 & a_1 & b_1 \\ \cdots & 0 & 0 & a_0 & b_0 \end{pmatrix} \quad \text{and} \quad P_{GM1} = \begin{pmatrix} b_0 & a_0 & 0 & 0 & 0 & \cdots \\ b_1 & a_1 & a_0 & 0 & 0 & \cdots \\ b_2 & a_2 & a_1 & a_0 & 0 & \cdots \\ b_3 & a_3 & a_2 & a_1 & a_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \end{pmatrix}.$$

As in the previous section of this chapter, provided that in one step transition the evolution in the queue can only be +1 or -1 of its state according to Kleinrock's principle [127], then the probability distribution obtained at the arrival instants (4.12) is applicable also at the departure instants of packets from the system and so are both queue solutions.

The G/M/1 or M/G/1 queue then according to the result on the waiting time, show that they belong to the group of Cruz bounds $R(\rho, \sigma)$.

4-2-4) G/G/1 Queue

The G/G/1 queue is modelled as a single queue server, where packets arrive according to an interarrival time with general distribution characterized by a pdf $A(t)$ with rate $E[t_n]$ and are served according to a general service time distribution characterized by a pdf $B(t)$ with mean service time $E[X_n]$. Clearly we have:

$$E[t_n] = \frac{1}{\lambda} \text{ and } \sigma_\tau^2 = E[t_n^2] - E[t_n]^2 \text{ for the interarrival distribution,}$$

$$E[X_n] = \bar{x} \text{ and } \sigma_x^2 = E[X_n^2] - E[X_n]^2 \text{ for the service time distribution.}$$

For the stability condition, one can write $\lambda\bar{x} < 1$.

The solution of this queue is based on moment generating function (Chapter 3), Lindley's equations and the spectral solution.

a) Background of Lindley's Equations

Given waiting time w_n for an n -th packet, the waiting time for the $(n+1)$ -th packet is:

$$w_{n+1} = \begin{cases} w_n + x_n - t_{n+1}, & w_n + x_n - t_{n+1} \geq 0, \\ 0, & w_n + x_n - t_{n+1} < 0. \end{cases} \quad (4.15)$$

Define the variable $u_n = x_n - t_{n+1}$, clearly, the equilibrium condition imposes $\lim_{n \rightarrow \infty} E[u_n] < 0$. This can be shown as follows:

$$\lim_{n \rightarrow \infty} E[u_n] = \lim_{n \rightarrow \infty} (E[x_n] - E[t_{n+1}]) = \bar{x} - \bar{t} = \bar{t}(\rho - 1). \quad (4.16)$$

The waiting time of the $(n+1)$ -th packet is:

$$w_{n+1} = \begin{cases} w_n + u_n, & w_n + u_n \geq 0, \\ 0, & w_n + u_n < 0. \end{cases} \quad (4.17)$$

This means that $w_{n+1} = \sup\{0, w_n + u_n\} = (w_n + u_n)^+$. The waiting time distribution at the equilibrium $W(y) = \lim_{n \rightarrow \infty} \Pr[w_n \leq y]$ is required.

To derive this waiting time distribution, Liu [128] first defines and derives $C_n(u)$, the n -th packet at time u as the pdf for u_n :

$$C_n(u) = \Pr[u_n = x_n - t_{n+1} \leq u] = \int_{t=0}^{\infty} \Pr[x_n \leq u + t \mid t_{n+1} = t] dA(t) = \int_{t=0}^{\infty} B(t+u) dA(t) = C(u). \quad (4.18)$$

Similarly, for $W(y)$, when $y \geq 0$ it comes that:

$$W_{n+1}(y) = \Pr[w_n + u_n \leq y] = \int_{w=0^-}^{\infty} \Pr[u_n \leq y - w \mid w_n = w] dW_n(w). \quad (4.19)$$

Since u_n is independent of w_n , and taking into account the value of $C_n(u)$, we have:

$$W_{n+1}(y) = \int_{w=0^-}^{\infty} C_n(y-w) dW_n(w), \quad y \geq 0, \quad (4.20)$$

which for large n gives rise to the *Lindley's Integral Equation*, also referred to as first form of that *family's equation*:

$$W(y) = \int_{0^-}^{\infty} C(y-w) dW(w), \quad y \geq 0. \quad (4.21)$$

It is clear that $W(y) = 0$ when $y < 0$. Thus, by integrating Lindley's equation (4.21),

$$\begin{aligned} W(y) &= [C(y-w)W(w)]_{w=0^-}^{\infty} - \int_{0^-}^{\infty} W(w) dC(y-w), \\ W(y) &= \lim_{w \rightarrow \infty} C(y-w)W(w) - C(y)W(0^-) - \int_{0^-}^{\infty} W(w) dC(y-w). \end{aligned} \quad (4.22)$$

Provided that $C(y-w) = 0, w \rightarrow \infty$ and $W(0^-) = 0$, the waiting time is finally given by:

$$W(y) = \begin{cases} - \int_{0^-}^{\infty} W(w) dC(y-w), & y \geq 0, \\ 0, & y < 0. \end{cases} \quad (4.23)$$

This equation represents *the second form of Lindley's equation*.

Having reached the second form leads to the third one. The latter is obtained by variable change $u = y - w$. Bearing in mind that $C(u)$ is a distribution, henceforth $c(u)$ the density function is given by:

$$W(y) = \begin{cases} \int_{-\infty}^y W(y-u)dC(u), & y \leq 0, \\ 0, & y < 0. \end{cases} \quad (4.24)$$

Taking into account these equations, two methods were exploited to approximate the G/G/1 queue model: the spectral and the bounds methods.

b) The spectral solution

Liu [128] characterized this queue using the spectral method from the Laplace transform of the waiting time Lindley's equation extended to a full plane as follows:

$$\begin{aligned} L[W(y) + W^c(y)] &= L\left[\int_{-\infty}^y W(y-u)c(u)du\right], \\ \therefore \phi(s) + \phi^c(s) &= \phi(s)C^*(s) = \phi(s)A^*(-s)B^*(s), \end{aligned}$$

where $L[c(u)] = L[a(-u)]L[b(u)]$ or $C^*(s) = A^*(-s)B^*(s)$ and the probability function $W(y)$ is bound with its density function $w(y)$ in Laplace domain by $s\phi(s) = L[w(y)] = W^*(s)$.

Liu considered furthermore a queuing system for which $A^*(s)$ and $B^*(s)$ are rational functions in the Laplace domain, yielding $A^*(-s)B^*(s) - 1 = \frac{\varphi_+(s)}{\varphi_-(s)}$ where $\lim_{|s| \rightarrow \infty} \varphi_+(s) = s$ for $\text{Re}\{s\} > 0$, and $\lim_{|s| \rightarrow \infty} \varphi_-(s) = -s$ for $\text{Re}\{s\} < \theta$.

Applying the theorem of Liouville [129] that requires a constant for any bounded analytic function within all finite values of its support,

$$\begin{aligned} \phi^c(s)\varphi_-(s) &= \phi(s)\varphi_+(s) = K, \\ \therefore \phi(s) &= \frac{K}{\varphi_+(s)}, \end{aligned}$$

where K is the constant to be determined through $s\phi(s) = W^*(s) = \int_0^\infty e^{-sy}dW(y)$.

When $s \rightarrow 0$, one reaches:

$$\lim_{s \rightarrow 0} \phi(s) = \lim_{s \rightarrow 0} \int_{y=0^-}^\infty e^{-sy}dW(y) = 1 = \lim_{s \rightarrow 0} \frac{sK}{\varphi_+(s)}. \quad (4.25)$$

This finally leads to $K = \lim_{s \rightarrow 0} \frac{\varphi_+(s)}{s}$, i.e. $W^*(s) = s\phi(s)$ for the waiting time.

c) Kingman Queue Bounds

- Exponential Bounds

Kobayashi [130] carries on with the result (4.25) $K = \lim_{s \rightarrow 0} \frac{\varphi_+(s)}{s}$ where $K = W(0) = \frac{1-\rho}{\lambda} \varphi_-(0)$ to derive an upper bound on the complementary waiting time $W_n^c(t) = \Pr[w_n > t]$, where W_n is the waiting time of the n -th packet in a busy cycle.

Let C_n be the n -th arriving packet in the queue and $t_n = \tau_n - \tau_{n-1}$ the time interval between C_n and C_{n-1} . Define x_n and w_n as service time and waiting time respectively in the queue for C_n , define the relation between 2 consecutive packet waiting times by:

$$w_{n+1} = \begin{cases} w_n + x_n - t_{n+1}, & w_n + x_n - t_{n+1} \geq 0, \\ 0, & w_n + x_n - t_{n+1} < 0, \end{cases}$$

where for any variable $u_n = x_n - t_{n+1}$. Clearly, the equilibrium condition imposes $\lim_{n \rightarrow \infty} E[u_n] < 0$; and the random variables of sequence $\{u_n : n \geq 1\}$ are independent and identically distributed.

When packet C_{n+1} arrives, it finds a non-empty system if and only if $w_n + u_n > 0$. If this is true, it therefore holds that:

$$w_{n+1} = \max(w_n + u_n, 0) = (w_n + u_n)^+. \quad (4.26)$$

By recursion on w_n and applying:

- Kolmogorov's Inequality [131] which is the generalized form of Chebyshev's Inequality [132];
- the notion of martingales and semi-martingale [130] (or sub-martingale) variables extended to the inequality as proposed by Feller [133]; and
- Laplace-Stieltjes transform of $C(u)$,

one obtains the tighter complementary upper bound at the steady state for small n and θ a positive real number satisfying $e^{\theta w_n} = \max(y_0, y_1, \dots, y_n)$.

One reaches:

$$W_n^c(t) \leq e^{-\theta_n t + n g(\theta_n)}, \quad (4.27)$$

where $\theta_m < \theta_0$ and for large n from (4.26), one obtains the upper bound of the tail of the complementary waiting time given by:

$$W_n^c(t) \leq e^{-\theta_0 t} . \quad (4.28)$$

It is important to notice that the rational function $\varphi_+(s)$ might have zeros and poles. The zero that is closest to the origin ($s \rightarrow 0$) satisfies $s = -\theta_0$.

- Moment Bounds

Kingman [134] has given a more tractable method on the mean tail probability for the waiting time in a G/G/1 queue system.

Define the independent random variables w, u which satisfy equality (4.26), $w_n + u_n$ has the same distribution as w_n , therefore, for large n , one can write in terms of distribution $w = (w + u)^+$. In particular, $E[w] = E[(w + u)^+]$ and $E[w^2] = E[((w + u)^+)^2]$.

Define $(w + u)^- = -\min(0, w + u)$, then it turns out that $w + u = (w + u)^+ - (w + u)^-$.

Therefore, $E[(w + u)] = E[(w + u)^+ - (w + u)^-] \Rightarrow E[u] = -E[(w + u)^-]$.

Also, $E[(w + u)^2] = E[((w + u)^+ - (w + u)^-)^2]$ and provided that $0 = (w + u)^+ (w + u)^-$, one gets

$$E[w] = \frac{E[u^2] - E[((w + u)^-)^2]}{-2E[u]} = \frac{\sigma_u^2 - \sigma_{(w+u)^-}^2}{-2E[u]}, \quad (4.29)$$

where $u_n = x_n - t_{n+1}$, then $\sigma_u^2 = \sigma_x^2 + \sigma_t^2$ and $\sigma_{(w+u)^-} \geq 0$.

The proposition of Kingman's moment bound of the mean waiting time in FCFS queuing discipline suggests that:

$$W \leq \frac{\lambda(\sigma_x^2 + \sigma_t^2)}{2(1 - \rho)} .$$

This is the upper bound and is most useful.

Another interesting bound is given in [135] for the special case where the interarrival time t_n satisfies the property [136]: $E[t_n - t | t_n > t] \leq \frac{1}{\lambda}$ (mean arrival rate), for all $t \geq 0$:

$$\frac{\lambda(\sigma_x^2 + \sigma_t^2)}{2(1-\rho)} - \frac{1+\rho}{2\lambda} \leq W \leq \frac{\lambda(\sigma_x^2 + \sigma_t^2)}{2(1-\rho)}. \quad (5.12)$$

Hence, if the mean and variance of the interarrival time, as well as the service time are known, the following bounds hold for the waiting time in queue of any G/G/1 queue model:

$$\frac{\lambda\sigma_x^2 - \bar{x}(2-\rho)}{2(1-\rho)} \leq W_q \leq \frac{\lambda(\sigma_x^2 + \sigma_t^2)}{2(1-\rho)}. \quad (5.13)$$

4-3) CONCLUSION

Based on these theories, one can say that there are two ways of modelling a queue from the general distribution perspective:

- The moment based model yielding the queue in the form A/S/1 as we reduced the model; and
- The bounds method resulting from the lack of moments. Among those bounds, the Cruz bounds generally denoted by $R(\rho, \sigma)$, the Kingman moment bounds which can be moment bounds or exponential bounds.

Based on these results from the queuing model theory, let us review some of the works done in that regard.

CHAPTER 5

REVIEW OF SIMILAR WORK

5-1) INTRODUCTION

Poisson process with its index of dispersion equal to 1 has been used extensively to characterize queuing models. However, the related queues experience high variability that leads to large packet delays in the multiplexer under heavy load [89] which according to [119], is due to the bursty nature of the voice packet arrival from one source comprising intervals of talk spurts and intervals of silence. It is therefore important to generalize the interarrival time or the service time distributions and approximate the choice done by a convenient Renewal process. This thus yields a G/M/1 or M/G/1 queue model depending on the system state. For a better concept, this will lead to consider a G/G/1 queue.

5-2) M/M/1 QUEUE

There were many studies conducted to characterize the M/M/1 queue using the matrix solution. In this project, we are interested in the simple M/M/1 queue in an FCFS discipline considering infinite waiting room (open networks) as proposed by Adam and Resing [138].

5-2-1) The steady-State Solution of Adam and Resing

Adam and Resing in [138] take into account the state probabilities $P_n(t)$ of (4.6)

$P_n = \lim_{t \rightarrow \infty} P_n(t)$, $n = 0, 1, 2, \dots$, and $\frac{dP_n(t)}{dt} = 0$ and the resulting infinitesimal generator matrix

$$P' = P \begin{bmatrix} -\lambda & \lambda & 0 & \dots & 0 \\ \mu & -(\lambda + \mu) & \lambda & \dots & 0 \\ 0 & \mu & -(\lambda + \mu) & \lambda & \vdots \\ \vdots & \vdots & \mu & \dots & \lambda \\ 0 & 0 & 0 & \dots & -(\lambda + \mu) \end{bmatrix}, \quad (5.1)$$

where P is a row vector of P_n .

By considering that as $t \rightarrow \infty$, $P'_n(t) \rightarrow 0$ and $P_n(t) \rightarrow P_n$, yields the queue length distribution:

$$P_n = (1 - \rho)\rho^n, \quad n > 0, \quad \rho < 1. \quad (5.2)$$

The mean queue length is $E[N] = \sum_{n=0}^{\infty} nP_n = (1 - \rho) \sum_{n=0}^{\infty} n\rho^n = \frac{\rho}{1 - \rho}$.

By using Little's theorem [139] the mean waiting time:

$$E[W] = \frac{\rho}{\mu(1 - \rho)}. \quad (5.3)$$

Miller [106] and Zhang and Shi [140] considered the M/M/1 queue with the pre-emptive and non pre-emptive priority queues with two classes of packets (Appendix A-3). By considering the higher priority between two types of packets arriving with rates λ_1 and λ_2 , all served exponentially with the same mean service time $1/\mu$; and using the quasi-birth-and-death process as operator, the equilibrium for this circumstance is:

$$\rho_1 = \frac{\lambda_1}{\mu}, \quad \rho_2 = \frac{\lambda_2}{\mu}, \quad \rho = \rho_1 + \rho_2 < 1.$$

Gevros [141] developed the Early Random Detection (RED) M/M/1 scheme to improve the congestion avoidance of TCP/IP traffic. This technique has been developed for the best effort service model and Chandrayana [142] studied the configuration of the RED parameters.

But the RED approach as developed by Pitts, Wang, Yang and Schormans [143] gives an explicit formula for load, configuration and performance based on excess rate of the queue for real-time traffic as VoIP. The distribution of packet arrival in a packet service time is given by:

$$s[k] = \int_0^{\infty} s(\mu, t) a(\lambda, k, t) dt = \left(\frac{\rho}{1 + \rho} \right)^k \frac{1}{1 + \rho},$$

where $s(\mu, t)$ and $a(\lambda, k, t)$ are service time and k arrival distribution functions.

Bera [144] studied the M/M/1/N overflow and packet loss queuing system to finally derive the expected waiting time per customer in the queue as previously (see Appendix A-3):

$$E[W_N] = \frac{E[Q_N]}{\lambda_{\text{eff}}} = \frac{\lambda[(\mu^N - \lambda^N) - N\lambda^{N-1}(\mu - \lambda)]}{\mu(\mu - \lambda)(\mu^N - \lambda^N)}.$$

5-3) THE G/M/1 QUEUE MODEL

In the previous section, we presented the M/M/1 queue. With its index of dispersion of 1 its network parameters can be taken as baseline. But this model does not provide a lot of knowledge when it comes to capturing variations in the interarrival process, particularly when the traffic is aggregated. Thus to get insights into variations in the interarrival process, we resort to a more general approach resulting in the G/M/1 queue. But this implies many ways of analysis of the general distribution which lean on approximation methods.

5-3-1) Approximating the arrival Process in the G/M/1 Queue

As we stated earlier, the aggregate traffic exhibits variability that discard the process from Poisson process. Approximating the arrival process therefore becomes a requirement for a better result. This drove almost all the previous works in this queue category.

a) Indirect approach: the moment solution

Pioneered by Ward [145] the indirect approach presents an approximation procedure that can be used to fit in changes brought in by assumptions. It aims to describe the level of congestion in the queue that will prevail if the same arrival process becomes the input of the service mechanism, which has been changed purposely. This is done through parameter estimators as proposed by Cooper [146] and Kuczura [147]. This indirect method is also equivalent to the random approach followed by Wallstrom [148] and at the least measure by Wilkinson [149].

Let $Q(t)$ denote the number of packets in the queue at time t and let $q_i(t) = E[Q(t)^i]$, $t \geq 0$ be the i -th moment of $Q(t)$. Assuming a non-periodic time, then $\lim_{t \rightarrow \infty} Q(t) \rightarrow Q$ and $\lim_{t \rightarrow \infty} q_i(t) \rightarrow q_i$.

The first two moments also characterize the asymptotic behaviour and they are given by:

$q_1 = \frac{\rho}{1-\sigma}$ and $q_2 = \frac{\rho(1+\sigma)}{(1-\sigma)^2}$ (geometric distribution) from which one can get ρ and σ :

$$\sigma = \frac{q_2 - q_1}{q_2 + q_1}, \text{ and } \rho = \frac{2q_1^2}{q_2 + q_1}. \quad (5.4)$$

Let $Q(t_k)$ be the queue content congestion level, the two moment estimates are given by:

$$\hat{q}_{1n} = \frac{\sum_{k=1}^n Q(t_k)}{n}, \quad n \geq 1 \quad \text{and} \quad \hat{q}_{2n} = \frac{\sum_{k=1}^n Q^2(t_k)}{n}, \quad n \geq 1,$$

which by similarity with the result in (5.4), can be written as:

$$\hat{\sigma}_n = \frac{\hat{q}_{2n} - \hat{q}_{1n}}{\hat{q}_{2n} + \hat{q}_{1n}} \quad \text{and} \quad \hat{\rho}_n = \frac{2\hat{q}_{1n}}{\hat{q}_{2n} + \hat{q}_{1n}}.$$

The sequence $Q(t_k)$, $k \geq 1$ taken at arrival epochs initiating busy periods are independent therefore, the estimators $\{q_{in}\}$ are consistent: $\lim_{n \rightarrow \infty} q_{in} = q_i$ which furthermore are unbiased if the system is taken at the steady state resulting in $E[q_{in}] = q_i$ for all n .

The application is carried on the $H_2^b/M/1$ queue with the probability distribution of the aggregated two types of traffic given by:

$$f(t) = p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t}, \quad t \geq 0, \quad 0 \leq p \leq 1.$$

Iglehart [150] provides a method to sort out these estimators for small samples, while Kuehn [151] and Morse [152] over balance mean reduced the number of parameters of an hyper-exponential $H_2^b/M/1$ queue to balance mean rate given by:

$$\frac{1}{\lambda} = \frac{2p}{\lambda_1} = \frac{2(1-p)}{\lambda_2}.$$

The solution results in the Cruz bounds solution [137] as the theory implies.

Adam, Boxma, and Perry [123] presented a revisited form of the G/M/1 queue by introducing the actual and virtual waiting times based on set-up times at the start of each busy period.

b) Markov Modulated Poisson Process (MMPP) approximation approach

In [153], Choi et al. studied the priority of this model. In the queue analysis, 2 classes of customers are adopted. Class 1 with arrival rate λ_{i1} , $i = 1, 2$ and the class 2 traffic with arrival rate λ_{i2} . With sojourn times r_i , a non-preemptive HOL (Head Of Line) priority is adopted for a delay constraint. When a service is completed, the next customer to be served is selected from class 1 if any. Note that the service is exponentially distributed for both classes. The following transition matrix is reached:

$$M = \begin{bmatrix} \lambda_{11} + r_1 + \lambda_2 & -r_2 \\ -r_1 & \lambda_{12} + r_2 + \lambda_2 \end{bmatrix}.$$

Abate and Whitt [154] have carried out studies for calculating the time dependent performance of the MMPP/M/1 queue to derive bounds of $\Sigma M/M/1$ queue as long as the arrival rate λ_i is constant.

Romano, Ciciani, Santoro and Quaglia [155] provided an approximate solution for the response time distribution of the model through two-state Markov Modulated Poisson Process. They show the relevance of identifying the lower bound of the cumulative distribution function. The steady state was solved by introducing the modified Bessel function of the first kind having argument $2\mu\sqrt{\rho_j}t$.

With Robert and Leboudec in [167], the multiplexer is modelled being an MMPP arrival process into a single-server queue operating in an FCFS discipline. That is MMPP/M/1 Queue. They consider two-states Markov Chain, the process is modulated as in Figure 5.1:

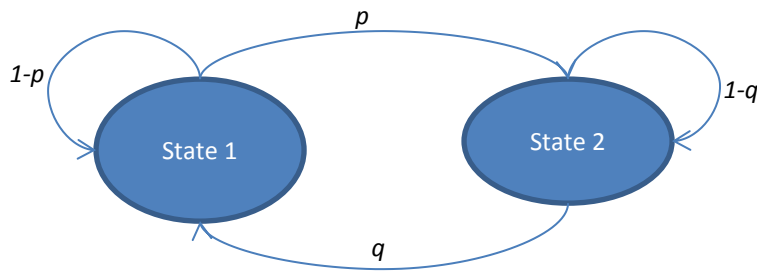


Figure 5.1: A Two-State Markov Chain

When the arrival is state 1, the process is Poisson with rate λ_1 . The system moves from state 1 to state 2 with transition probability p . When in state 2, the Poisson arrival has an arrival rate of λ_2 and moves from state 2 to 1 with transition probability q . The transition matrix is given by:

$$T = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}.$$

The equilibrium distribution $\pi = (\pi_1, \pi_2)$ obeys the following probabilities when $p + q \neq 0$:

$$\Pr[\text{state } 1] = \pi_1 = \frac{q}{p+q} \text{ and } \Pr[\text{state } 2] = \pi_2 = \frac{p}{p+q}. \tag{5.5}$$

The resulting probability is given by $p_n = (1-p)^{n-1} p$. Then the state 1 average time is:

$$\bar{T}_1 = \sum_{n=1}^{\infty} n p_n = \frac{1}{p},$$

similarly for $\bar{T}_2 = \frac{1}{q}$. (5.6)

Brady [156] exploited these results to characterize On-Off sources by referring to state 1 as On and state 2 as Off periods and exploiting the results in (5.5) and (5.6).

Since $1/\lambda_1$ is the interarrival mean time, the average load of the server is reduced to the On load that is λ_1 . The total load produced by such a source is given by:

$$\frac{\bar{T}_1}{\bar{T}_1 + \bar{T}_2} \lambda_1 = \frac{q}{p + q} \lambda_1 = \pi_1 \lambda_1 ,$$

where the intensity of traffic for stability reason impose $\pi_1 \lambda_1 < 1$.

This load is the intensity of traffic in the queuing system which according to Little's theorem is given by $\lambda_1 E[S]$. It turns out that the deterministic mean service time is given by:

$$E[S] = \frac{1}{\mu} = \frac{q}{p + q} .$$

By applying Little's theorem, the waiting time of voice packets in the queue system is:

$$W = \frac{\rho^2}{\mu(1 - \rho)} .$$

5-4) – THE M/G/1 QUEUE

The M/G/1 model can be analyzed as a G/M/1 queue according to Kleinrock's principle. But the most innovative work in this model is done by Sinclair [157], leaning on the Pollaczek-Khinchin mean value formulae [158]. This results from the fact that the model admits a general approximation based on average values and application of Little's theorem yielding the required analysis of packets in the queue. Hence, the following considerations as in [157]:

- The server is busy if the queue is non-empty
- No packet exits the queue without having completed service
- The service discipline is independent on information about packet service times.

The steady state is reached when the i -th packet arrives in the queue. Define then:

- N_i the number of packets in the queue when packet i arrives $\bar{N}_i = E[N_i]$ and $\lim_{i \rightarrow \infty} \bar{N}_i = N_Q$.
- X_i the service time of packet i , $\bar{X} = E[X_i]$ its mean and its second moment $\bar{X}^2 = E[X^2]$.

- W_i the waiting time packet i spends in the queue $\bar{W}_i = E[W_i]$ and $\lim_{i \rightarrow \infty} \bar{W}_i = W$.
- R_i the residual time seen by packet i , $\bar{R}_i = E[R_i]$ and $\lim_{i \rightarrow \infty} \bar{R}_i = R$.
- λ the arrival packet rate, ρ the intensity of traffic or server utilization according to Little's theorem is $\rho = \lambda \bar{X}$.

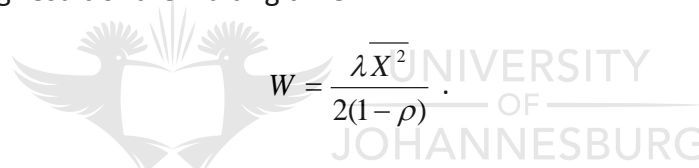
The waiting time in the system is $W_i = R_i + \sum_{k=i-N_i}^{i-1} X_k$, which in the mean value becomes:

$$\bar{W}_i = \bar{R}_i + E[\sum (E[X_k / N_i])] = \bar{R}_i + \bar{X}N_i.$$

At the steady state this yields the waiting time

$$W = \frac{R}{1 - \rho}.$$

Sinclair [157] computed the Residual time by a series of triangle of slope -1 referred to as slots and got the following result of the waiting time:



$$W = \frac{\lambda \bar{X}^2}{2(1 - \rho)}.$$

The advantage of this result is the introduction of queue with vacation, useful for time slotted traffic. For any deterministic service time M/D/1 queue with waiting time W_{MD1} , its slotted counterpart W_{SMD1} is given by (as computed in [159]):

$$W_{SMD1} = \frac{\lambda / \mu^2}{2(1 - 1/\mu)} + \frac{1/\mu^2}{2/\mu} = \frac{\rho}{2\mu(1 - \rho)} + \frac{\bar{V}}{2} = W_{MD1} + \frac{\bar{X}}{2}.$$

This has been useful for TDM and FDM slotted systems exploiting the second term which is the vacation period.

Schormans and Pitts [143] presented an excess rate (ER) solution based on Poisson process adjustment to produce a geometric series for the tail of the distribution: a geometrically distributed number of ER arrivals per service time, called the Geometrically Approximated Poisson process (GAPP). This ER arrival distribution for a probability of k arrivals in a service time $s[k]$ is given by:

$$p[k] = (1 - \rho) \left[\frac{q + (1 - q)((1 - s[0] - s[1]) / (1 - s[1]))}{(s[0] / (1 - s[1]))} \right]^k,$$

where q is the probability of having another ER arrival in the same service period as the previous one.

5-5) THE G/G/1 QUEUE

The G/G/1 model comes as the importance of generalizing both the interarrival time and the service time distributions and their fitting solution approaches. It is a single server queue with a general interarrival time and a general service time distributions. There has been a huge amount of works done, we will present selected results.

A part of it is basically intended to approximating the general process into a Renewal process to achieve a moment based queue model.

Another part however, bases the ongoing studies on the buffer occupancy with the aim to model the queue through the bounds solutions.

5-5-1) Moment queues

Liu [128] used the spectral method to derive the M/M/1 queue by considering arrival time $a(t)$ and service time $b(t)$ exponential distributions, their Laplace transforms, and the zeros and poles of the analytical functions as follows:

$$A^*(s) = \frac{\lambda}{s + \lambda} \text{ and } B^*(s) = \frac{\mu}{s + \mu}, \text{ then}$$

$$A^*(-s)B^*(s) - 1 = \frac{s^2 + s(\mu - \lambda)}{(\lambda - s)(\mu + s)} = \frac{\varphi_+(s)}{\varphi_-(s)}.$$

The waiting time distribution is given by:

$$W(y) = 1 - \rho e^{\mu(1-\rho)y}, \quad y \geq 0.$$

The result shows that the distribution is a function of the drift in the queue system.

Heffes and Lucantoni [160] used the bivariate matrix method to derive the M/G/1 queue: the MMPP/G/1. Although other approximations such as Shahram [161] exist, the advantage here is its wide range of applications. The aim being to determine the virtual waiting time, which is given in the following.

Let the vector $W(x)$ have components $W_j(x)$ which are given by the probability distribution (see Appendix A-5):

$$W_j(x) = \lim_{t \rightarrow \infty} \Pr[V(t) \leq x, J(t) = j | X(0) = i, J(0) = l]. \quad (5.7)$$

Denote by $W(s)$ the Laplace-Stieltjes transform of $W(x)$. Applying the renewal theorem yields the virtual waiting time of the MMPP/G/1 given by [160]:

$$W(s) = \begin{cases} sy_0[sI + R - \Lambda F_s(s)]^{-1}, & s > 0, \\ \pi, & s = 0. \end{cases}$$

Sriram and Whitt [119] involved the probability generating function of the number of arrivals in an interval and is given by:

$$g(z, t) = \pi \exp\{[R + (z-1)\Lambda]t\}, \quad (5.8)$$

where, for 2-state MMPP, the equilibrium probability vector π is given by $\pi = \frac{1}{r_1 + r_2}(r_2, r_1)$,

$e = (1, 1)^T$, $R = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}$ and $\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$, λ_1 and λ_2 are the arrival rates of the Poisson processes, r_1 and r_2 are the sojourn times of the states 1 and 2 of the underlying Markov chain confined in the transition matrix R , and Λ is the diagonal matrix compounded of arrival rates.

Schwartz [162] and Luhanga [163] studied voice and data queues using the Fluid process approximation by considering $(N+1)$ -state Markov Chain of N asynchronous sources. Both consider two On-Off sources with the same characteristics as in Chapter 3.

In [162], Schwartz considers a pre-emptive priority for voice over data in an FCFS queue discipline, but by reserving a part of transmission link capacity to data traffic.

In his approach to determining the normalization coefficients of the differential fluid equation, (see Appendix A-6). Luhanga [163] dedicated a link capacity C and queue length Q for each type of traffic (C_1, x_1 for voice) and applied Kolmogorov's forward equation, we have:

$$r_k \frac{dp_k(x_1)}{dx_1} = (N+k-1)\alpha p_{k-1}(x_1) + (k+1)\beta p_{k+1}(x_1) - [(N-k)\alpha + k\beta]p_k(x_1), \quad (5.9)$$

where $k = 0, 1, 2, \dots, N$, $r_k = R_k - C$ is the net rate of change of the buffer content, R_k being the aggregate arrival rate of k sources, and $p_k(x_1) = 0$ for $k < 0$ and $k > N$. This differential equation becomes: $D \frac{dP(x_1)}{dx_1} = MP(x_1)$, and stability condition $\rho = \frac{\alpha N}{(\alpha + \beta)C} < 1$ for voice,

where $P(x_1) = [p_0(x_1), p_1(x_1), \dots, p_N(x_1)]'$, ' denotes the transpose of P and D the diagonal rate matrix: $D = \text{diag}[-C, (R_1 - C), (R_2 - C), \dots, (R_N - C)]$.

The probability matrix is given by:

$$M = \begin{bmatrix} -\alpha N & \beta & 0 & \dots & 0 \\ N\alpha & -[(N-1)\alpha + \beta] & \dots & & \vdots \\ 0 & (N-1)\alpha & & \vdots & \\ \vdots & \vdots & & -[(\alpha + (N-1)\beta)] & N\beta \\ 0 & 0 & \dots & \alpha & -N\beta \end{bmatrix}.$$

Under certain assumptions and by using Little's theorem, the average waiting time is given by:

$$W_d = Q_d \lambda_2^{-1},$$

where $Q_d = \sum_0^{n_e-1} \frac{K_n^d}{S_n^d} (e \phi_n^d)$, with $e = [1, 1, \dots, 1]'$ and ϕ_n^d is the column eigenvector (Appendix A-6).

The Diffusion approximation solution

The diffusion approximation is characterized by the following differential equation:

$$dx(t) = \beta dt + z(t) \sqrt{\alpha} dt$$

where $dx(t)$ represents the incremental changes of the continuous path of the process $x(t)$; β is the drift; α is the variance; and $z(t)$ is the white Gaussian Process which can become Brownian motion. Then, $x(t)$ is a Brownian motion with drift satisfying the probability:

$$\frac{\partial f(x, t)}{\partial t} = -\beta \frac{\partial f(x, t)}{\partial x} + \frac{\alpha}{2} \frac{\partial^2 f(x, t)}{\partial x^2}.$$

The solution of this diffusion approximation equation is obtained under the boundary conditions and discretization [169]:

The reflecting Barrier condition is characterized by the following equations at the equilibrium

$$\begin{cases} f(x) = \lim_{t \rightarrow 0} f(x, t), \\ \lim_{x \rightarrow 0} \left[-\alpha f(x) + \frac{\beta df(x)}{2dx} \right] = 0. \end{cases}$$

The resulting solution is given by $f(x) = -\frac{2\alpha}{\beta} e^{\frac{2\alpha}{\beta}x}$.

The elementary Return Boundary condition that characterizes this type of diffusion approximation is the following:

$$\lim_{x \rightarrow 0} \left[-\alpha f + \frac{\beta}{2} \frac{df}{dx} + \lambda \pi_0(t) \frac{dB(x)}{dx} \right] = 0.$$

At the steady state, the solution is given by:

$$y = e^{\frac{2\alpha}{\beta}x} \left(\int_0^x \frac{2}{\beta} \lambda_i \pi_0 - \lambda_i \pi_0 (1 - B(t)) e^{-\frac{2\alpha}{\beta}t} dt + C \right) = f(x),$$

where α and β are the diffusion parameters, π_0 and B are probability of an idle system and service time function respectively.

The applications of the Diffusion approximation have been discussed by Cox and Miller [164], Newell [165], Gaver and Shelder [166] and Kobayashi [168]. Yu and Shinya [169] evaluated the heavy traffic download on the server access operation for Web traffic. But the most significant works done on G/G/1 queue Diffusion solution were carried by Reiser and Kobayashi [98] which gives accurate results for the classical queue model, and by Qiang-Kobayashi [96] and [97] which give rise to exponential bounds methods.

The accuracy is given by Reiser and Kobayashi [98]. By introducing the appropriate boundary conditions, the type 1 solution was obtained in the form $f(x) = \frac{2\alpha}{\beta} e^{-\frac{2\alpha}{\beta}x}$ which for small queue sizes, is the exponential process. With this resulting continuous process, we need to infer its discrete original process. The discrete process is then interpreted as a geometrical distribution of the queue size variable n with the same decrement factor $\hat{\rho} = e^{-\frac{2|\beta|}{\alpha}}$.

By applying the central limit theorem for small queue size, Reiser and Kobayashi characterized the change of queue between t and $t + \Delta t$ for a sufficiently large interval by a normal distribution with mean $E[Q] = (\lambda - \mu)\Delta t = \beta\Delta t$ and variance $Var[Q(t)] = (C_a\lambda + C_s\mu)\Delta t$ where λ is the arrival rate, μ is the processing rate, C_a and C_s are squared coefficients of variation of the interarrival and service times respectively. According to the results obtained, the geometrical distribution is written as:

$$\hat{p}(n) = \begin{cases} 1 - \rho, & n = 0, \\ \rho(1 - \hat{\rho})\hat{\rho}^{n-1}, & n \geq 1. \end{cases} \quad (5.10)$$

The queue size is then given by $Nq = \frac{\rho}{1 - \hat{\rho}}$.

5-5-2) Bounds solutions

Ren and Kobayashi [96] determines the queue bounds fed by On-Off sources for univariate Ornstein-Uhlenbeck (O-U), and ATM traffic for multivariate O-U Diffusion approximation using the exponential bounds.

For the Univariate queue analysis, the buffer behavior is carried out over the steady state traffic, i.e. $t \rightarrow \infty$. The resulting probability density function is given by $\lim_{t \rightarrow \infty} f(y, x, t) = f(y, x)$.

Using the differential equation of the Markov Birth-Death Chain of k -type On-Off sources, in which variables are separated as $f(y, x) = F(x).g(y)$ to reduce $\frac{F'(x)}{F(x)} = \frac{L.g(y)}{\sum R_k y_k - C} = u$

where u is a real constant to determine, $F(x)$ obeys the general solution e^{ux} , and $g(y)$ is obtained from the equation in the form of O-U Diffusion approximation

$$\left(\sum (R_k y_k - C)\right) \frac{\partial g(y, x)}{\partial x} = \sum \left[(\alpha_k + \beta_k) \frac{\partial [(y_k - y_k^*)g(y, x)]}{\partial x} + \left(\frac{N_k \alpha_k \beta_k}{\alpha_k + \beta_k}\right) \frac{\partial^2 g(y, x)}{\partial x^2} \right], \quad (5.11)$$

where R_k, α_k and β_k , are arrival rate, active and silence periods respectively of the k -type source.

The resolution of the O-U diffusion differential equation yields a quadratic equation derived from Weber's equation which gave 2 roots whereby the negative one is retained. Considering the positive probability of the queue $\Pr(Q > 0)$ equal to the probability of having the arrival rate R higher than C_1 , $\Pr(R > C_1)$ computed by Roberts [170] for large deviation, the lower bound queue length is derived:

$$F(Q > x) = \frac{1}{\mu \sqrt{2\pi \sigma_R^2}} \exp\left(-\frac{1}{2} \sigma_R^2 \mu\right) \exp(u_0^- x) \quad \text{for large } x.$$

For the multivariate O-U Diffusion Process viewed by Ren and Kobayashi [97], the equilibrium state of the multidimensional stochastic differential equation is:

$$dX(t) = B(X(t) - x^*)dt + \sqrt{A}dW(t). \quad (5.12)$$

This equation is the multivariate **Ornstein-Uhlenbeck** process which satisfies the probability density function differential equation:

$$\frac{\partial f(x,t)}{\partial t} = -\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \beta_{ij} \frac{\partial}{\partial x_i} [(x_i - x_i^*)f(x,t)] + \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \frac{1}{2} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f(x,t), \quad (5.13)$$

where β_{ij} and a_{ij} are the (i,j) -th entries of $K \times M$ matrices B and A .

Exploiting the Gaussian process property and the approximation from Roberts [170] for $\Pr(R > C)$, they obtained the complementary queue content distribution at that equilibrium as follows:

$$F(Q > x) = \frac{\mu_{\tilde{R}} - n}{C - n} \exp\left\{-2 \frac{(C - \mu_R)}{a} x\right\} \text{ and } P(\tilde{R} > C) \approx \frac{1}{z^* \sqrt{2\pi}} \exp\left\{-\frac{z^{*2}}{2}\right\}, z^* = \frac{C - \mu_{\tilde{R}}}{\sigma_{\tilde{R}}},$$

which are the upper bound and the lower bound of the queue content.

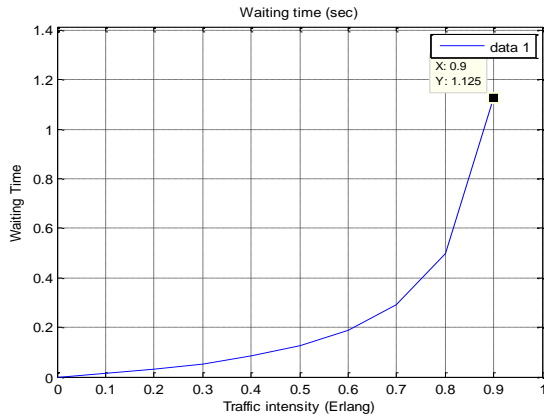


5-6) PREVIOUS WORKS SELECTED RESULTS

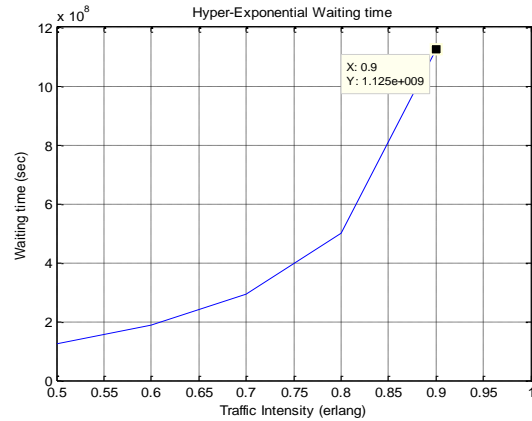
We will in this section present some results to motivate our work.

Figure 5.1 (a) shows the delay of the M/M/1 queue as developed in [138]. It exhibits a heavy tail traffic that increases with the intensity of traffic showing then the variability in the system.

Figure 5.1 (b) is the waiting time of a $H^2/M/1$ queue taken as G/M/1 queue for $\sigma \geq \rho$, using Takacs' solution approximation [119] as proposed by Whitt [145]. It is a bounds method solution showing bigger delay in a heavy traffic (hyper-exponential distribution) than in Figure 5.1 (a) (exponential distribution).



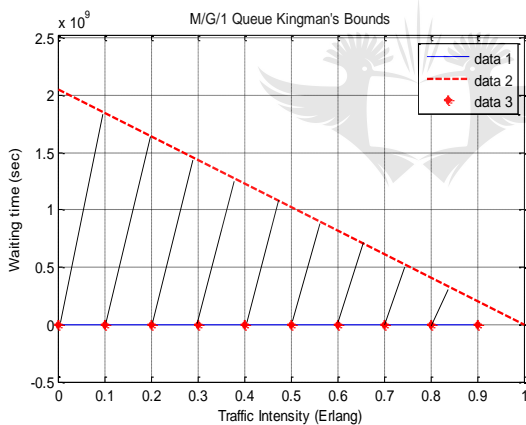
(a)



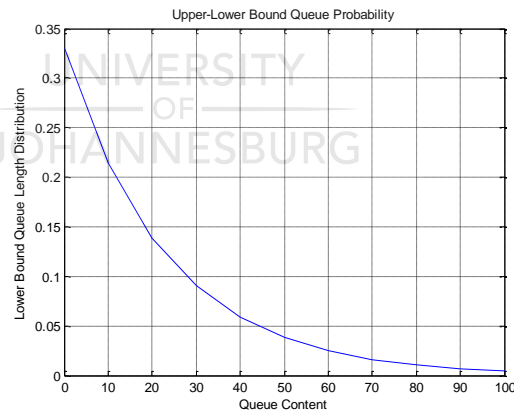
(b)

Figure 5.1: Queue waiting time with the traffic intensity (a) M/M/1 and (b) H²/M/1

Figure 5.2 shows the Kingman bounds moments based of the arrival time and service time; and the exponential bounds as presented by Qiang and Kobayashi Diffusion approximation. The latter exhibits a faster decay in transition which implies a lesser blocking probability.



(a)



(b)

Figure 5.2: (a) Moment Bounds waiting time and (b) Exponential bounds waiting time (sec)

Figure 5.3 represents the Diffusion approximation when approximated with the exponential decrements as Qiang and Reiser; and the queue solution based on Pollaczek-Khinchin mean value for small size samples. This is carried out over the exponential distribution and the hyper-exponential distribution for the heavy traffic perspective. The plots show the quasi-equivalence of those two queue solutions.

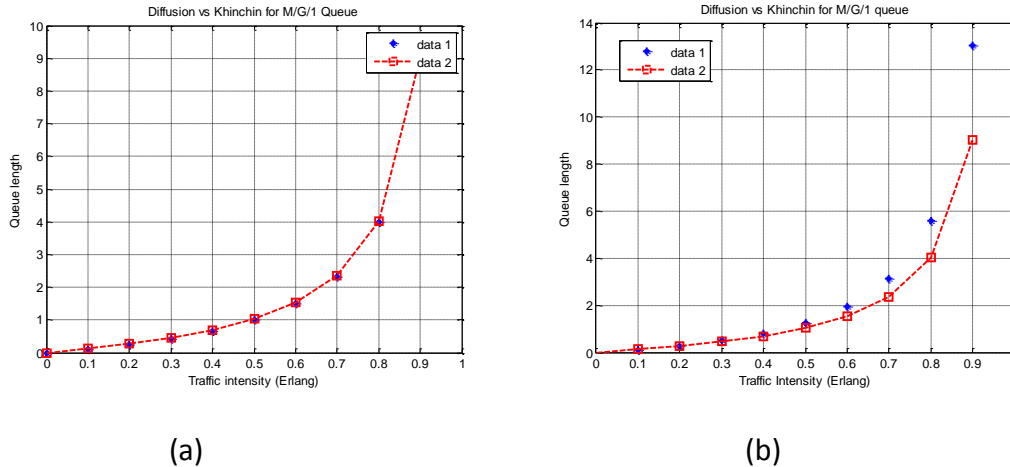


Figure 5.3: Diffusion solution compared to Pollaczek-Khinchin mean value: (a) hyper-exponential $H^2/M/1$ and (b) exponential distribution $M/M/1$

In Figure 5.3, we represent the following data:

Data 1: Diffusion approximation Queue length;

Data 2: Pollaczek-Khinchin mean-formulae Queue length.

Figure 5.4 shows the 3 states possible from a 2-state On-Off source with the Fluid Process when applying LLN (Law of Large Number). The traffic greater than 1 increases to infinity, data 1 and data 2 with positive eigenvalues, the traffic intensity less than 1 erlang (data 3) decreases to 0. We chose as a normalization factor value 1 for simplicity so that all traffic levels have the same initial probability value. Actually, only the queue with traffic intensity less than 1 characterized by the decreasing slope is advisable.

The Fluid approximation as an alternative shows changes in the queue that need to be captured. Based on this, a limit solution of the Fluid method is proposed: the Diffusion approximation.

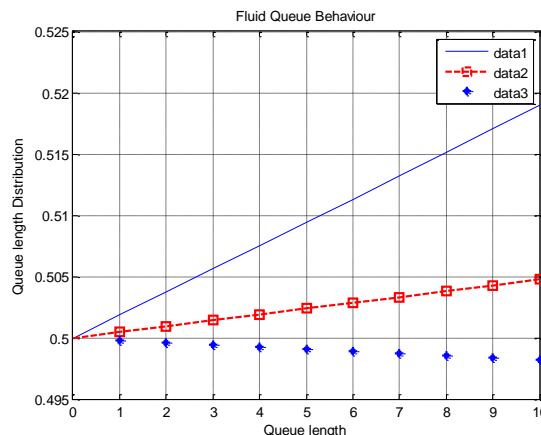


Figure 5.4: Fluid Diffusion Queue Behaviour with LLN

5-6) CONCLUSION

We have discussed some of the most important works done in queue modelling; we put our focus on those whose analysis follows Kolmogorov's equation. These selected results were reviewed in this chapter for a better clarification for the motivation behind our queue modelling solution, which is proposed in the following chapter.



CHAPTER 6

PROPOSED SOLUTION

6-1) INTRODUCTION

In the previous chapters, we have outlined the main aspects that paved our way to answering our research questions. In fact in Chapter 2, we have discussed about the voice transportation issues over a packet network. The E-model defined the transport requirements; and ITU-H.323 prescribed quality guarantee requirements in those packet networks. We have shown that the configuration of H.323 Endpoints and Terminals is crucial in terms of QoS, particularly when dealing with real time traffic. In Chapter 5, queuing models were presented from selected previous works with the aim to justify our approach. In this chapter, we will bring to the fore this approach through related motivations and a proposed solution with the aim to answer the questions raised in transporting voice over the Internet.

6-2) OUR SOLUTION AND PREVIOUS WORKS

The prime motivation of our approach derives from the previous works reviewed in Chapter 5. In the M/M/1 queue as presented by Adam and Resing [138], one would like to have the arrival and service rates proportional. Unfortunately, above a certain range of traffic intensity, the proportionality is no longer guaranteed. One observes then a heavy tail traffic (see Appendix A-1) resulting from the long range dependence, which Sriram and Whitt [89] proved is the cause of delay experienced in the queue.

Therefore, a more general approach of the queue model is needed in which one will have, as the theory implies, 2 types of solution.

6-2-1) The bounds solution

The indirect method of Whitt [119] using Takacs solution based on Laplace Stieltjes transform has a double condition as Cruz $R(\sigma, \rho)$ bounds which restrict the range of the traffic intensity, reducing it to half for some distributions. With a delay that entails a large buffer size, this restriction is proven to be a waste of resources.

The Kingman moments bounds queue solution is not far from the Cruz bounds, showing a reduced delay but still needs a large buffer size provided its slow decay of probability excess.

In contrast, the exponential bounds show a faster decay than the moment bounds therefore a smaller blocking probability. We recall that the Diffusion approximation achieves the exponential server, whether with the bounds solution or moments solution, suggesting then its use as a queuing solution.

6-2-2) The moments solutions

The Pollaczek-Khinchin mean value, the MMPP solution and the Fluid method have been selected in this category of queue modelling. As a result, we get:

- The quasi-equivalence between the Pollaczek-Khinchin mean value which presents asymptotically better results than the MMPP and the Diffusion approximation of Reiser-Kobayashi. However, Pollaczek-Khinchin mean value solution is better exploited in synchronous or plesiosynchronous (almost synchronous) networks while the Diffusion approximation of Reiser-Kobayashi entails small queue size. This suggests the use of the Diffusion approximation as queue solution.
- The Fluid Process as a proposal to take into account the variability in the queue, captures the drift of buffer occupancy, showing therefore piecewise linear behaviours in the queue length. This comes from the fluid scale which captures changes in the queue in n -time units for n -state order. The changes observed in the fluid scale have led us to consider instead, rescaling the fluid queue at its limit. This is called the Diffusion approximation.

6-3) OUR APPROACH MOTIVATIONS

Other factors besides the ones mentioned above have motivated our approach:

- *The statistical motivation* behind this choice is to capture the variability in the queue.
 - The central limit theorem is chosen through the Maximum Likelihood Estimator (MLE). The latter has essentially no optimal properties for finite samples; however, it possesses a number of attractive asymptotic properties for many problems that include:
 - * Asymptotic normality as the sample size increases,
 - * Consistency: the estimator converges in probability to the value being estimated,
 - * Efficiency: there is no asymptotically unbiased estimator that has a lower asymptotic mean square error than the MLE.

- The Diffusion process is the limit of the fluid method. The latter captures the drift of buffer occupancy to achieve a limit that is linear to its duration, showing piecewise linear behaviour of the queue lengths [189].
- *The traffic model:* Markov Modulated Poisson Process (MMPP) and the Markov Modulated Rate Process (MMRP) are related with their first identical moments and by their moment generating functions as follows [171]:

$$g_{MMRP}(z, t) = g_{MMPP}(1 + \log z, t).$$

- *The commercial motivation:* this proposal is an attempt to maximize the Operational Expenditure (OPEX) provided the pricing fluctuations in the market during the financial year, of ISP (Internet Service Provider) backbones in providing a more convenient QoS in terms of average call duration (ACD) and answer rate (ASR).

6-4) NETWORK DESIGN MOTIVATIONS

The motivations of the design we are proposing comes from the E-model transmission tool planning of Section 2-3. We can recall that the E-model formula is given by:

$$R = R_0 - I_s - I_e - I_d + A,$$

with R_0 being the basic signal-to-noise ratio (S/N).

From the network requirements and according to the E-model [30] relation above, we can deduce:

- Provided that we want to provide a voice service at the same standard quality level of the old legacy network, this removes the degradation the user is ready to suffer: then one can set $A = 0$.
- As seen in the Tables B.2 and B.3 in Appendix B, some voice codecs provide efficient use of bandwidth. Avoiding transcoding; and dimensioning the network with enough bandwidth and an adequate priority over packet (CoS) will reduce I_e , which is the equipment impairment factor associated with the use of low bit rate codecs.
- Today systems are built in such a way to avoid quantizing distortion or outputting the appropriate level of signal. Thus I_s , which is the sum of impairments such as low signal level, non-optimal side tones and quantization distortion, becomes negligible.

- We saw that I_d is the most important parameter in the E-model equation since it involves all the impairments causing the delay experienced by the voice signal, which are:

- i) The impairments due to talker listener echo; and
- ii) The loss of connectivity due to excessive delay.

The H.323 Endpoints such as Gateways are built with echo suppression allowing the removal of the first impairment. The excessive delay comes therefore from the following:

- Encoding-Decoding delay: that is coding and decoding time intervals of audio codec samples in use.
- Packetization delay: that is the required encapsulation time.
- Propagation delay: that is the time needed by the signal to propagate from the source to the destination.
- Queuing and transmission delay: that is the time needed to transmit the packet through the network.
- Jittering Delay: that is the system delay introduced by the use of a buffer at the destination to absorb delay variation in the packet network.

The packetization and jitter delays have been standardized and managed accordingly with the different voice algorithms, H.323, and RTP protocols and they are acceptable.

The propagation and queuing system delays remain. For the mouth-to-ear delay to be bounded in an acceptable delay we need:

To find a good compromise between buffer delay and bandwidth for the queuing delay; and voice packet to get Head of Line (HoL) (priority): this is achieved by the CoS defining priorities and the network topology set in such a way that only the edge Endpoints route the Egress and Ingress traffics. This constitutes the motivation of our network configuration design.

6-5) NETWORK ARCHITECTURE DESIGN

Our aim of introducing a 2-level multiplexing scheme relies entirely on the good structural configuration of the H.323 environment as described in Figure 6.1. These Endpoints and Terminals are connected to various external sources of different types of traffic with different

operational bit rates, in compliance with the H.245 capability structure requirements of Chapter 2, a set of voice algorithms (in our case) in which they are able to transmit and receive.

The definition of the queuing system and the network configuration requirements of ITU-H.323 can now lead us to consider the multiplexing scheme we are proposing. The gateway requires the implementation of MC and MP functionalities for the voice backbone network of the Service Provider bridging the Switched Circuit Network (SCN). Meanwhile, the other bridge of the network must comply with the local Egress traffic by implementing the same functionalities. The bridge can be a router or an H.323 bridge Endpoint.

Figure 6.1 below shows the two-level multiplexing that will be used in the following sections.

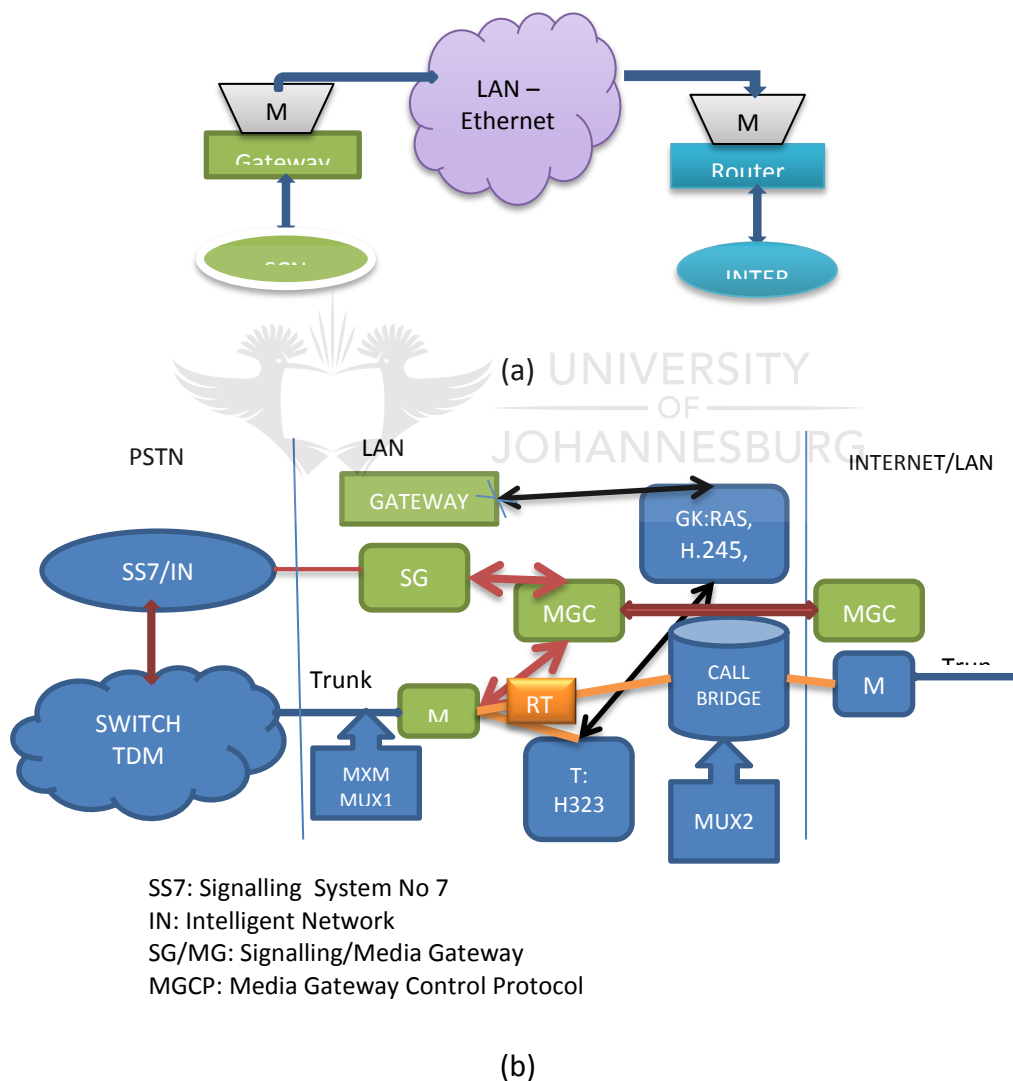


Figure 6.1: The Proposed VoIP LAN with (a) H.323 network access design and (b) Network function distribution design.

6-6) LEVEL 1 MULTIPLEXING DESIGN

We are accommodating multiple voice codec from different sources (assumed to be On-Off) in the statistical multiplexer, herein referred to as level 1 multiplexing (Figure 6.2). It is shown that the multiplexer input with the Diffusion parameters approximates the process to a Markov Gaussian Process, hence the queue structure in this chapter denoted as M/G/1 queue to specify Markov interarrival time, general service time and 1 server.

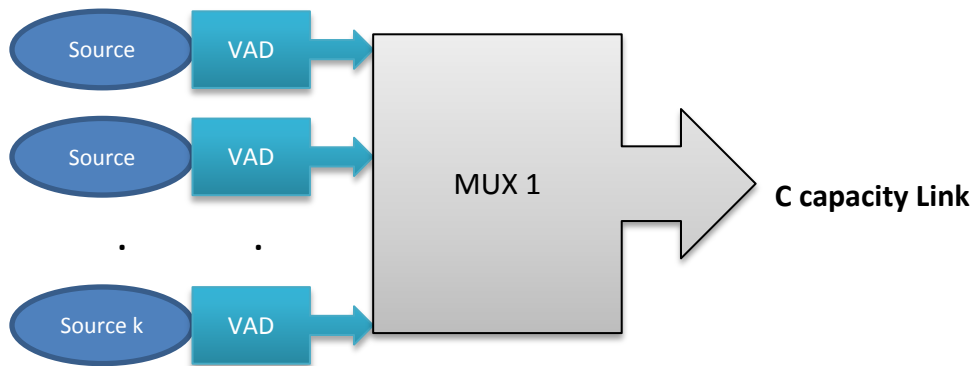


Figure 6.2: The proposed level 1 Multiplexing, Source with Voice Activity Detector

6-6-1) The univariate Diffusion approximation

The univariate diffusion approximation is characterized by the following differential equation:

$$\frac{\partial f(x,t)}{\partial t} = -\beta \frac{\partial f(x,t)}{\partial x} + \frac{\alpha}{2} \frac{\partial^2 f(x,t)}{\partial x^2},$$

This is composed of three factors, necessary conditions for its solution [169]:

- The diffusion parameters which are infinitesimal mean and variance,
- The boundary conditions,
- The discretization.

a) Reflecting Barrier Boundary condition

The reflecting Barrier condition is characterized by the following equations at the equilibrium:

$$\begin{cases} f(x) = \lim_{t \rightarrow 0} f(x,t), \\ \lim_{x \rightarrow 0} \left[-\alpha f(x) + \frac{\beta df(x)}{2dx} \right] = 0. \end{cases}$$

The resulting solution is given by:

$$f(x) = -\frac{2\alpha}{\beta} e^{\frac{2\alpha}{\beta}x}. \quad (6.1)$$

The amount of the variable from the diffusion approximation of the reflecting barrier 0 is:

$$E[X(t)] = \int_0^{\infty} xf(x)dx = -\frac{\beta}{2\alpha} \quad (6.2)$$

b) Elementary Return Boundary condition

The following equation characterizes this type of diffusion approximation at the steady state [169]:

$$\lim_{x \rightarrow 0} [-\alpha f + \frac{\beta}{2} \frac{df}{dx} + \lambda \pi_0(t) \frac{dB(x)}{dx}] = 0.$$

At the steady state, one imposes:

$$\frac{\partial \pi_0(t)}{\partial t} = -\lambda \pi_0(t) + [-\alpha f + \frac{\beta}{2} \frac{df}{dx}]_{x=0},$$

$\lim_{t \rightarrow \infty} f(x, t) = \lim_{t \rightarrow 0} f(x, t) = 0$ where $f(x)$ is given by the following diffusion equation:

$$-\alpha f + \frac{\beta}{2} \frac{df}{dx} + \lambda \pi_0(t) \frac{dB(x)}{dx} = 0.$$

By integration of this equation and letting $f(x) = y$, we can write:

$$\frac{\beta}{2} y' - \alpha y = \lambda_i \pi_0 - \lambda_i \pi_0 B(x).$$

Therefore:

$$y' - \frac{2\alpha}{\beta} y = \frac{2}{\beta} \lambda_i \pi_0 - \lambda_i \pi_0 (1 - B(x)).$$

The resulting differential equation is of the form $y' + p(x)y = q(x)$, whose solution is given by

$$y = e^{-\int_0^x p(x)dx} \left(\int_0^x q(x) e^{\int_0^x p(x)dx} dx + C \right).$$

One can then have the following solution:

$$y = e^{\frac{2\alpha}{\beta}x} \left(\int_0^x \frac{2}{\beta} \lambda_i \pi_0 - \lambda_i \pi_0 (1 - B(t)) e^{-\frac{2\alpha}{\beta}t} dt + C \right) = f(x), \quad (6.3)$$

where α and β are the diffusion parameters, π_0 and B are probability of an idle system and service time function respectively.

c) The univariate Ornstein-Uhlenbeck Diffusion approximation

The Ornstein-Uhlenbeck (O-U) process approximation is a Diffusion process characterized by the diffusion differential equation in the form:

$$\frac{df(x,t)}{dt} = \beta \frac{df(x,t)}{dx} + \sigma^2 \frac{d^2f(x,t)}{dx^2}, \quad (6.4)$$

where β and σ^2 are the mean and variance of the O-U univariate process. The solution of the conditional probability of this differential equation is a Gaussian distribution at the equilibrium state ($t \rightarrow \infty$) [172]:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(x-\beta)^2}{2\sigma^2}\right\}. \quad (6.5)$$

Let $f(x)$ and $F(x)$ be the density and probability distribution functions defined as (6.5). We have: $F(x) = \int_0^x f(u)du$ which is close to 1 for large x and its complementary is given by

$1 - F(x) = \int_x^1 f(u)du$ which is close to 0 for the same condition.

6-6-2) The univariate Gaussian distribution properties

The properties are outlined here to expand the possibilities of analyzing the queue system from the bounds methods up to the moment based methods.

a) Bounds of the distribution

The bounds methods allow determining the overflow of the buffer occupancy when we lack statistical parameters. The lower and upper bounds can be determined as follows [173]:

$$\frac{x}{1+x^2} f(x) < 1 - F(x) < \frac{f(x)}{x}. \quad (6.6)$$

Furthermore, we can exploit the following boundaries of the erfc function defined as:

$$\operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-x^2} dx,$$

where the integral follows the inequalities:

$$\frac{e^{-z^2}}{z + \sqrt{z^2 + 2}} < \int_z^{\infty} e^{-x^2} dx < \frac{e^{-z^2}}{z + \sqrt{z^2 + 4/\pi}}. \quad (6.7)$$

b) Maximum likelihood Estimation

Another property of the Gaussian distribution is that it can be maximized by the likelihood of the variable of interest through its estimates. These estimates are derived from samples when the moments of the distribution are not known. They are: sample mean and sample variance-covariance for the multivariate process.

Suppose X_1, X_2, \dots, X_n are each independent normally distributed with mean μ and variance $\sigma^2 > 0$. Let estimate the sample mean and variance mean of this sample.

Using the Maximum Likelihood, the values of μ and σ that maximize the Likelihood function are such that:

- The sample mean: $\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, and

-The sample variance is given for $\sigma > 0$ by: $\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$. (6.8)

These results need to be consistent. The consistency can be achieved using Fisher Information which is defined in [174].

Fisher information of a random variable X with probability distribution $f(X/\theta_0)$, where θ_0 is the maximizer of the Log-Likelihood of the distribution, and is defined by:

$$I(\theta_0) = E\left[\left(\frac{\partial}{\partial \theta} \operatorname{Log} f(X/\theta)\right)\Big|_{\theta=\theta_0}\right]^2,$$

which by [175] yields:

$$-I(\theta_0) = E\left[\frac{\partial^2}{\partial \theta^2} \operatorname{Log} f(X/\theta_0)\right].$$

This tool tells us how good our estimates are regarding the properties of the MLE.

c) The central limit theorem

Define the sequence of random variables $\bar{X}_n = (X_{1n}, X_{2n}, \dots, X_{kn})$ then:

$$\sqrt{n}(\bar{X}_n - \mu) \rightarrow N(0, \sigma) \text{ as } n \rightarrow \infty.$$

6-6-3) Diffusion Approximation for ON-OFF voice packet sources

The traffic in the multiplexer is the superposition of packet streams from many voice sources as defined in the H.245 capability set. Since some sources are OFF when others are ON, the input of the multiplexer queue is variable. The statistical multiplexer we are proposing here is labelled Level 1 multiplexing scheme within the H.323 environment (Figure 6.2).

Let N_k be the number of sources of type k with a rate of R_k packets/sec, and A_k be the number of active sources with mean time α^{-1} ms, or $N_k - A_k$ the number of OFF sources with mean silence time β^{-1} ms, the multiplexer input is given by:

$$R(t) = \sum_{k=1}^N R_k A_k(t). \quad (6.9)$$

Define $Q(t)$ as the queue size of the buffer and C packets/sec the constant capacity of the transmission link. By definition [60], the statistical multiplexer allocates capacity that lies between the average and the peak rates and buffers the traffic when the load exceeds the capacity. Therefore the changes in the multiplexer can be captured by the following differential equation:

$$\frac{dQ(t)}{dt} = \begin{cases} R(t) - C, & R(t) > C, Q(t) > 0, \\ 0, & \text{otherwise.} \end{cases} \quad (6.10)$$

Since neither $R(t)$ nor $Q(t)$ are Markovian, this motivates us to step towards the multivariate process $(A(t), Q(t))$ such that $(A(t), Q(t)) = \{A_k, 1 \leq k \leq K; Q(t)\}$ is a Markov process.

Define a vector $A = (a_1, a_2, \dots, a_k)$, 1_k the identity matrix of size K ; one can define the probability function of the process as follows: $P(A, x, t) = \Pr[A_k = a_k, 1 \leq k \leq K; Q(t) \leq x]$

which satisfies the stochastic differential equation:

$$\begin{aligned} \frac{\partial p(A, x, t)}{\partial t} + \left(\sum_{k=1}^K R_k A_k(t) - C \right) \frac{\partial p(A, x, t)}{\partial x} = \\ - \sum_{k=1}^K [(N_k - A_k)\beta_k + A_k\alpha_k] p(A, x, t) + \sum_{k=1}^K [N_k - A_k + 1]\beta_k p(A_k - 1, x, t) + \sum_{k=1}^K (A_k + 1)\alpha_k p(A_k + 1, x, t). \end{aligned} \quad (6.11)$$

As $x \rightarrow \infty$, the equation (6.11) above becomes a Markov birth-death process which is known to result in a Bernoulli solution:

$$P(A, t) = \lim_{k \rightarrow \infty} P(A, x, t) = \prod_{k=1}^K P(A_{kj}, t), \text{ where } P(A_k, t) = \binom{N_k}{A_k} q_k^{A_k} (1 - q_k)^{N_k - A_k} (t). \quad (6.12)$$

It is also known that the binomial distribution tends to a Gaussian distribution for large number; Let that continuous process to be $\{Y(t)\}$, the probability density function of the couple $(Y(t), Q(t))$ approximation is given by: $f(y, x, t) = \Pr[y_k \leq y \leq y_k + dy_k; Q(t) \leq x]$ which differential equation in the second order Taylor's series representation is given by:

$$\begin{aligned} \frac{\partial f(y, x, t)}{\partial t} + \sum_{k=1}^K (R_k y_k - C) \frac{\partial f(y, x, t)}{\partial x} = - \sum_{k=1}^K [(N_k - y_k)\beta_k + y_k\alpha_k] f(y, x, t) \\ + \sum_{k=1}^K (N_k - y_k + 1)\beta_k \left[f(y, x, t) - \frac{\partial f(y, x, t)}{\partial x} \right] + \frac{1}{2} \sum_{k=1}^K (N_k - y_k)\beta_k \frac{\partial^2 f(y, x, t)}{\partial x^2} \\ + \sum_{k=1}^K (y_k + 1)\alpha_k \left[f(y, x, t) + \frac{\partial f(y, x, t)}{\partial x} \right] + \frac{1}{2} \sum_{k=1}^K y_k\alpha_k \frac{\partial^2 f(y, x, t)}{\partial x^2} \\ = - \sum_{k=1}^K \frac{\partial}{\partial x} [(N_k\beta_k - (\alpha_k + \beta_k)y_k) f(y, x, t)] + \sum_{k=1}^K \frac{1}{2} \frac{\partial^2}{\partial x^2} [(N_k\beta_k - (\beta_k - \alpha_k)y_k) f(y, x, t)]. \end{aligned} \quad (6.13)$$

By analogy, one defines the infinitesimal statistical properties of the process:

$$\text{- The infinitesimal mean: } m_k = N_k\beta_k - (\alpha_k + \beta_k)y_k. \quad (6.14)$$

$$\text{- The infinitesimal variance: } v_k = N_k\beta_k - (\alpha_k - \beta_k)y_k. \quad (6.15)$$

Considering the infinitesimal arrival process only of mean zero, we have:

$$f(y, t) = \lim_{x \rightarrow \infty} f(y, x, t) \text{ and } y_k^* = \frac{N_k\beta_k}{\alpha_k + \beta_k}.$$

Therefore, the mean can be written $m_k(y) = -(\alpha_k + \beta_k)(y - y_k^*)$, and the related variance becomes:

$$v_k(y_k^*) = \frac{2N_k \beta_k \alpha_k}{\alpha_k + \beta_k}. \quad (6.16)$$

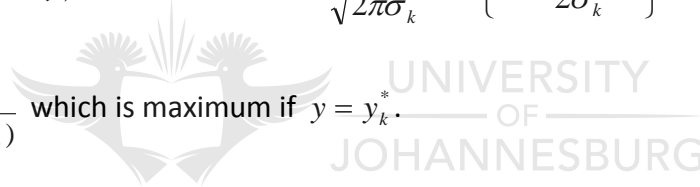
The diffusion equation above taken for individual source types can now be written as:

$$\frac{\partial f_k(y,t)}{\partial t} = (\alpha_k + \beta_k) \frac{\partial}{\partial y} [(y - y_k^*) f_k(y,t)] + \frac{N_k \alpha_k \beta_k}{\alpha_k + \beta_k} \frac{\partial^2}{\partial y^2} f_k(y,t). \quad (6.17)$$

The diffusion process that is characterised by (6.17) is called an Ornstein-Uhlenbeck (O-U) process [96]. At the equilibrium-state ($t \rightarrow \infty$), the density function, solution of the equation truncated at the reflecting boundaries $y = 0$ and $y = N_k$, yields a Gaussian distribution defined by:

$$\lim_{t \rightarrow \infty} f_k(y,t) = f_k(y) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{(y - y_k^*)^2}{2\sigma_k^2}\right\},$$

where $\sigma_k^2 = \frac{v_k(y_k^*)}{2(\alpha_k + \beta_k)}$ which is maximum if $y = y_k^*$.



Our methodology starts here by leaning on these results. We are exploiting the properties of the Gaussian distribution: Maximum Likelihood Estimator (MLE) to carry out our project.

6-6-4) The queue Analysis

The queue changes (ΔQ) are given by the differential equation in (6.13) completed by (6.17) and by identification with the accuracy method of Reiser and Kobayashi we get:

$$\left(\sum_{k=1}^K (R_k y_k - C)\right) \frac{\partial f(y,x)}{\partial x} = \sum_{k=1}^K [(\alpha_k + \beta_k) \frac{\partial [(y_k - y_k^*) f(y,x)]}{\partial x} + \left(\frac{N_k \alpha_k \beta_k}{\alpha_k + \beta_k}\right) \frac{\partial^2 f(y,x)}{\partial x^2}].$$

By rearranging and applying the central limit theorem on the number of arrival y on the differential equation above, such that the solution yields the standard form we got in Chapter

5, which becomes $\left(\sum_{k=1}^K (R_k y_k - C)\right) \frac{\partial f(y,x)}{\partial x} = \sum_{k=1}^K \frac{N_k \alpha_k \beta_k}{\alpha_k + \beta_k} \frac{\partial^2 f(y,x)}{\partial x^2}$. We can then deduce that:

The mean queue length is given by $E[Q] = (\widehat{R}_k - C)\Delta t$, where $\widehat{R}_k - C$ is the drift of the Diffusion process representing the variation of the mean and C the capacity transmission link.

The variance in the queue is given by $Var[Q] = (C_a \widehat{R}_k + C_s C)\Delta t = \sum_{k=1}^K \frac{2N_k \alpha_k \beta_k}{\alpha_k + \beta_k} \Delta t$, where C_a and

C_s are the indexes of dispersion of the arrival and service time processes respectively,

$C_a \widehat{R}_k + C_s C = \sum_{k=1}^K \frac{2N_k \alpha_k \beta_k}{\alpha_k + \beta_k}$ is the volatility which is the variation of the variance of the

diffusion process .

In Chapter 4, we saw that those two parameters can be accurately taken as decrement factor denoted $\hat{\rho} = \exp\{-\frac{2\beta}{\alpha}\}$, where β is the drift and α is the volatility of the process.

Therefore, the probability distribution of the queue size is geometric and given by:

$$\hat{p}_n = \rho(1 - \hat{\rho})\hat{\rho}^{n-1}, \quad n \geq 1,$$

where the traffic intensity is given by $\rho = \widehat{R}_k / C < 1$.

6-7) LEVEL 2 MULTIPLEXING DESIGN

Having multiplexed On-Off sources at the one edge of the H.323 environment, we are interested in doing so at the other edge of the network. It is worthy to recall that here we are facing multi-state sources from the Level 1 multiplexing. The arrival process therefore needs to be redefined accordingly. In this case, we have chosen a Markov Modulated Rate Process (MMRP) rather than the Markov Modulated Poisson Process; just for the fact that the latter always shifts to the next state in the chain. Hence the MMRP/G/1 queue.

This section is intended to propose the second multiplexing scheme designated as Level 2 multiplexing as shown in Figure 6.3. Before carrying out this objective, let us outline the basics of this traffic model and scaling.

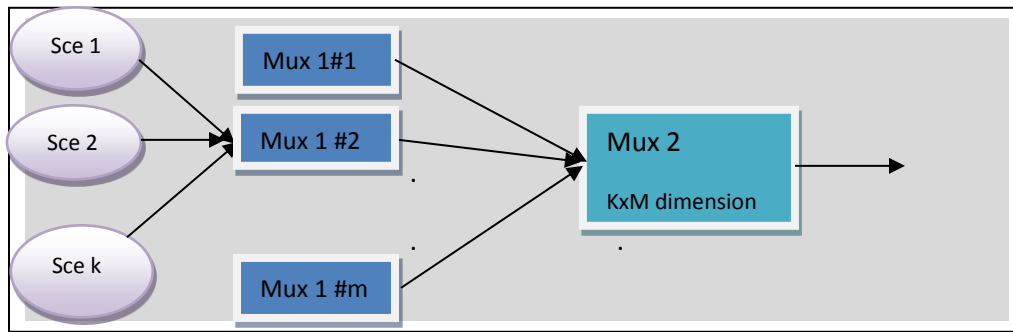


Figure 6.3: Level 2 Multiplexing scheme

6-7-1) The traffic model and scale

The multiplexing of On-Off traffic from the Switched Circuit Network (SCN) at the Gateway, leads us to find a suitable parametric traffic model that is compatible with the network of interest: Ethernet. Ethernet is one of the link layer technologies most used in the LAN networks, without a well defined traffic model. During recent years, studies have shown that traffic exhibits self similar properties [176] while others found it multifractal [177]. In our approach, we have modelled sources as Markov Modulated Rate sources. Since the traffic from the multiplexer is of multiple types, it is reasonable to rescale the aggregated traffic over CSMA/CD, which governs Ethernet access.

In fact, the aggregation of the traffic of different types from different sources in the LAN before the second multiplexer input raises a concern of time rescaling to avoid error of bits which may need another bit correction mechanism: adding therefore another delay. This is done through bit estimates moments that can be found in Appendix A-7. The traffic is modelled according to the behaviour of K superposed sources, each one governed by an M states Markov Chain with probability transition matrix $P = \{p_{ij}\}$, where $(i, j) = 0, 1, 2, \dots, M - 1$.

When a source is in state i , it generates packets at R_i packets/sec. After a holding time generally distributed with mean α_i^{-1} ms and variance σ_i^2 , it leaves from state i to state j with probability transition p_{ij} . This allows representing the state transition of the $K \times M$ states MMRP sources by a closed queuing network with M nodes and a total of K customers. Each source in state i is considered as a customer at node i served by one of the K parallel servers. It comes that the arrival process at node i is the sum of those departures from other nodes routing their calls to that node. The formulation of the process is summarized in Figure 6.4.

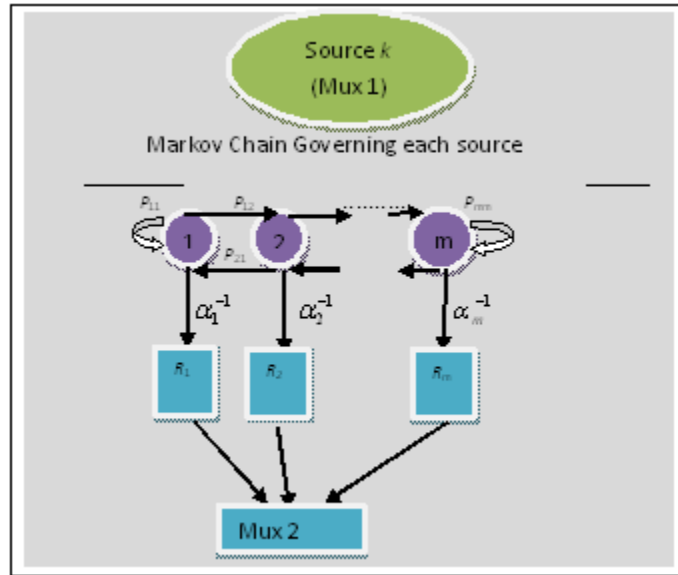


Figure 6.4: A level 2 Statistical Multiplexer input model

6-7-2) Multivariate Diffusion model

We have reached after the first level multiplexing, an arrival process which is the superposition of traffic from Level 1 multiplexers. This happens particularly in the H.323 network whereby Multipoint Control Units bridge the LAN. It seems therefore compelling for our study to introduce a Level 2 multiplexing for the requirement of QoS. Here, the endpoints serving as multipoint control units in Figure 6.3 are sources allowing a multiple-state Markov Modulated Rate Process, instead of 2 states as Level 1. The analysis of the queue here is leaning on the diffusion process as in Level 1.

The formulation of the Diffusion model is the system composed of statistical multiplexer and M independent sources. Each source is characterised by k -state Markov Chain (k -type sources of Level 1) yielding a K -dimensional process. Let $N(t)$ be that process, it is defined by the transposed matrix: $N(t) = [N_0(t), N_1(t), \dots, N_{k-1}(t)]'$, where $N_k(t)$ is the number of sources in the state k at time t . When a source is in state k , it generates packets at the rate R_k packets/sec and it will move from state j to state k with a probability transition p_{jk} . The holding time of this state has as mean α_k^{-1} and variance σ_k^2 . The mean departure rate from node j can be given by $\alpha_j N_{jk}$ and the counting arrival process at node k is equal to the aggregation of those departures from node j . The superposed traffic at the multiplexer input is given by:

$$R(t) = \sum_{k=0}^{K-1} R_k N_k(t).$$

Given $N(t)$, and in compliance with the model as described, we define its Diffusion approximation $X(t) = [X_0(t), X_1(t), \dots, X_{k-1}(t)]'$, according to Proposition 1 of the Halfin and Whitt asymptotic regime in [178], $X(t)$ follows the stochastic differential equation:

$$dX(t) = b(X(t))dt + DdW(t), \quad \sum_k X_k(t) = M, \quad X(0) = x_0, \quad (6.18)$$

where $W(t)$ is the Brownian motion and $b(x)$ is the infinitesimal mean pointing out the drift of the process compared to the long term equilibrium. The higher b is, the faster is the speed of the drift: the variation of the mean.

D represents the randomness of the process outlining the volatility of the process. The higher the value of D is, the larger is the magnitude of the volatility of the system or variation of the variance.

From (6.18), we define the conditional probability density function

$$f(x, t : x_0, 0) = \Pr[x_k \leq X_k(t) \leq x_k + dx_k \mid X(0) = x_0],$$

which satisfies the multidimensional differential equation as previously defined in (6.3):

$$\frac{\partial f(x, t)}{\partial t} = - \sum_{i=0}^{K-1} \frac{\partial}{\partial x_i} [b_i(x) f(x, t)] + \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \frac{1}{2} \frac{\partial^2}{\partial x_i \partial x_j} [a_{ij}(x) f(x, t)]. \quad (6.19)$$

We define therefore the multivariate mean $b_i(x)$ and the covariance $a_{ij}(x)$.

$b(x) = Bx$, where B is the infinitesimal mean matrix with $B = \{\beta_{ij}\}$ and $x = [x_0, x_1, \dots, x_{k-1}]'$ associated to $X(t)$.

$A(x)$ is the infinitesimal covariance matrix such that $D = \sqrt{A(x)}$ with $A = \{a_{ij}\}$.

The Halfin-Whitt differential equation (6.17) becomes:

$$dX(t) = BX(t)dt + \sqrt{A(x)}dW(t),$$

$$\text{with } \sum_{k=1}^K X_k(t) = K. \quad (6.20)$$

Let $x^* = [x_0, x_1, \dots, x_{k-1}]$ be the asymptotic equilibrium state of the process $X(t)$, such that the process is stable with a constant reverting drift, it comes that $b(x) = B(x^* - x)$ and at this particular narrow region, the infinitesimal covariance is constant. At the equilibrium state (6.20) becomes:

$$dX(t) = B(x^* - X(t))dt + \sqrt{A}dW(t). \quad (6.21)$$

This equation is the multivariate Ornstein-Uhlenbeck process which satisfies the probability density function differential equation:

$$\frac{\partial f(x,t)}{\partial t} = -\sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \beta_{ij} \frac{\partial}{\partial x_i} [(x_i - x_i^*) f(x,t)] + \sum_{i=0}^{K-1} \sum_{j=0}^{K-1} \frac{1}{2} a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f(x,t) \quad (6.22)$$

where β_{ij} and a_{ij} are the (i,j) -th entries of $K \times M$ matrices B and A respectively.

Our analysis starts with this Fokker-Plank equation (6.22) [179]. The objective is to implement our methodology which is to exploit the properties of the multivariate Gaussian distribution we have reached here.

6-7-3) Multivariate Gaussian properties

In the multivariate Diffusion process, the differential equation takes the following form:

$$\frac{df(x,t)}{dt} = -\sum_i \frac{d}{dx_i} [\mu_i(x) f(x,t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} [\sigma_{ij}(x) f(x,t)]. \quad (6.23)$$

The maximum likelihood estimation has been a problem [180] for multivariate Gaussian distribution since:

- All the univariate diffusion processes are reducible but not all multivariate.
- The boundary conditions imposed by the Brownian motion which is the volatility of the process (Section 6-6-1).
- The covariance matrix is not necessarily singular.

However, for estimation strategies of the continuous-time variable relying on discrete sampled data, the inference of the implications brought by the infinitesimal time evolution of the process is needed for the finite sampling time intervals. The transition function is therefore the key vector, but unfortunately which is difficult to obtain in the infinitesimal level.

Motivated by the results of the Level 1 multiplexing proposed herein we would like to keep on using the Maximum Likelihood Estimator. Because the log-likelihood plays the same role,

Sahalia [181] proposed a closed-form of the transition log-likelihood which we carry out in this part of the project. This closed form is based on the series expansion of Hermite and the change of variable in the Jacobian form. This overcomes the difficulty of having the transition function by defining the log-transition function, which in turn makes log-likelihood inference possible. It is noteworthy to mention that this is encouraging the more so as in [98], Reiser and Kobayasi have found the eigenvalues of their development in the form of Hermite coefficients.

The strategy of [181] is to determine the Hermite series expansion of the transition function p_Y which is $N(0,1)$ distribution, the reduced form of the transition function p_X derived from the diffusion process $dX_t = \mu(X_t)dt + \sigma(X_t)dW_t$.

The Hermite series approximation is in the form:

$$\tilde{p}_y^j(y/y_0, \Delta) = \Delta^{-m/2} \phi\left(\frac{y-y_0}{\Delta^{1/2}}\right) \sum_{tr|h| \leq j} \eta_h(\Delta, y_0) H_h\left(\frac{y-y_0}{\Delta^{1/2}}\right), \tag{6.24}$$

where Δ is the sampling interval, $\phi(x)$ is the density of a normal distribution with mean zero and identity covariance matrix, since we choose to take into account the sum of arrival distribution of mean zero and variance σ^2 . H_h is the Hermite polynomials associated with the vector $h = [h_1, h_2, \dots, h_m]^T$ such that

$$H_h(x) = \frac{(-1)^{tr[h]}}{\phi(x) \frac{\partial^{tr[h]}}{\partial x_1^{h_1} \dots \partial x_m^{h_m}}},$$

with respect of its orthogonality. $\eta_h(\Delta, y_0)$ are Hermite coefficients by expansion in Δ and are given by:

$$\eta_h(\Delta, y_0) = \frac{1}{h_1! \dots h_m!} E[H_h\left(\frac{y-y_0}{\Delta^{1/2}}\right) / Y_t = y_0], \tag{6.25}$$

where the expectation entity is evaluated by the Taylor expansion and is given by:

$$E[f(Y_\Delta, Y_0, \Delta) | Y_0 = y_0] = \sum \frac{\Delta^k}{k!} A_Y^k \cdot f(y, y_0, \delta) |_{y=y_0, \delta=0} + O(\Delta^{k+1}). \tag{6.26}$$

A_Y^k is the infinitesimal generator of the process Y , which by applying to f yields the solution of the diffusion differential equation (6.23) [182].

For the Log-transition, for any given j where the convergence of the Hermite polynomials is verified as $j \rightarrow \infty$, the resulting log-expansion has the form:

$$l_Y^k(y | y_0, \Delta) = -\frac{m}{2} \ln(2\pi\Delta) + \frac{C_Y^{-1}(y | y_0)}{\Delta} + \sum_{k=0}^K C_Y^k(y | y_0) \frac{\Delta^k}{k!}, \quad (6.27)$$

whose coefficients C_Y^k for $k = -1, 0, 1, 2, \dots, K$ are combination of the coefficients identified in the Hermite series approximation (6.24) and $l_y = \ln p_y$.

Their estimation strategies based on discretely sampled data is to find the inference of the infinitesimal time evolution of the process for the finite time interval at which the process is actually sampled Δ . The transition function plays a key role in that context. Unfortunately, that transition function is unknown. This leads to a closed form expansion of the log-transition of a large class of multivariate diffusion; the latter making feasible the quasi-likelihood inference.

We are proposing herein an alternative method which determines the coefficients series expansion satisfying Kolmogorov's equations that describe the evolution of the process. It is noteworthy to mention that when the diffusion is not reducible, the coefficient of expansion involves a double series in time and state variables. But in this case, the queuing network is independent of state provided that the Markov Chain is applicable at the source levels which in turn are independent.

Consider the forward Kolmogorov equation [183]:

$$\frac{\partial p_y(y | y_0, \Delta)}{\partial \Delta} = -\sum_{i=1}^m \frac{\partial [\mu_i(y) p_y(y | y_0, \Delta)]}{\partial y_i} + \frac{1}{2} \sum_{i=1}^m \frac{\partial^2 p_y(y | y_0, \Delta)}{\partial y_i^2}, \quad (6.28)$$

from which the equivalent form for the log-likelihood is given by:

$$\frac{\partial l_y(y | y_0, \Delta)}{\partial \Delta} = -\sum_{i=1}^m \frac{\partial \mu_{y_i}(y)}{\partial y_i} - \sum_{i=1}^m \mu_{y_i}(y) \frac{\partial l_y(y | y_0, \Delta)}{\partial y_i} + \frac{1}{2} \sum_{i=1}^m \frac{\partial^2 l_y(y | y_0, \Delta)}{\partial y_i^2} + \frac{1}{2} \sum_{i=1}^m \left(\frac{\partial l_y(y | y_0, \Delta)}{\partial y_i} \right)^2.$$

Exploiting the log-function above and substituting it for the log-likelihood function, we get the following:

$$\left\{ \begin{array}{l} \frac{\partial l_y^{(k)}(y/y_0, \Delta)}{\partial \Delta} = -\frac{C_y^{(-1)}(y/y_0)}{\Delta^2} - \frac{m}{2\Delta} + \sum_{k=1}^{K-1} C_y^{(k)}(y/y_0) \frac{\Delta^{k-1}}{(k-1)!}, \\ \frac{\partial l_y^{(k)}(y/y_0, \Delta)}{\partial y_i} = \frac{1}{\Delta} \frac{\partial C_y^{(-1)}(y/y_0)}{\partial y_i} + \sum_{k=0}^K \frac{\partial C_y^{(-1)}(y/y_0)}{\partial y_i} \frac{\Delta^k}{k!}, \\ \frac{\partial^2 l_y^{(k)}(y/y_0, \Delta)}{\partial y_i^2} = \frac{1}{\Delta} \frac{\partial^2 (C_y^{(-1)}(y/y_0))}{\partial y_i^2} + \sum_{k=0}^K \frac{\partial^2 C_y^{(-1)}(y/y_0)}{\partial y_i^2} \frac{\Delta^k}{k!}. \end{array} \right.$$

By equating the coefficient of second order of Δ , we get the leading coefficient given by:

$$C_Y^{(-1)}(y | y_0) = -\frac{1}{2} \left(\frac{\partial C_Y^{(-1)}(y | y_0)}{\partial y_i} \right)^T \left(\frac{\partial C_Y^{(-1)}(y | y_0)}{\partial y_i} \right).$$

By satisfying the condition that the density must approximate a Gaussian density as $\Delta \rightarrow 0$, the approximate solution is given by [184]:

$$\Delta C_Y^{(-1)}(y | y_0) = -\frac{1}{2} \|y - y_0\|^2 = -\frac{1}{2} m \sum_{i=1}^m (y - y_{0i})^2. \quad (6.29)$$

By next equating the terms of the first order, we get:

$$\sum_{i=1}^m \frac{\partial C_Y^{(0)}(y | y_0)}{\partial y_i} (y - y_{0i}) = \sum_{i=1}^m \mu_{Y_i}(y) (y - y_0),$$

this yields by integration:

$$C_Y^{(0)}(y | y_0) = \sum_{i=1}^m (y - y_0) \int_0^1 \mu_{Y_i}(y_0 + u(y - y_0)) du.$$

In the same way the higher coefficients are determined, hence the Theorem 1 of Sahalia [184], determining the recurrence for the higher order:

$$C_Y^{(k)}(y | y_0) = k \int_0^1 G_Y^{(k)}(y_0 + u(y - y_0) | y_0) u^{k-1} du, \quad k \geq 1, \quad (6.30)$$

where $G_Y^{(k)}$ are given by:

$$G_Y^{(k)}(y | y_0) = -\sum_{i=1}^m \frac{\partial \mu_{Y_i}(y)}{\partial y_i} - \sum_{i=1}^m \mu_{Y_i}(y) \frac{\partial C_Y^{(0)}(y | y_0)}{\partial y_i} + \sum_{i=1}^m \left(\frac{\partial^2 C_Y^{(0)}(y | y_0)}{\partial y_i^2} + \left[\frac{\partial C_Y^{(0)}(y | y_0)}{\partial y_i} \right]^2 \right);$$

and for $k \geq 2$,

$$G_Y^k(y|y_0) = -\sum_{i=1}^m \mu_{Y_i}(y) \frac{\partial C_Y^{k-1}(y|y_0)}{\partial y_i} + \frac{1}{2} \sum_{i=1}^m \frac{\partial^2 C_Y^{k-1}(y|y_0)}{\partial y_i^2} + \frac{1}{2} \sum_{i=1}^m \sum_{h=0}^{K-1} \binom{K-1}{h} \frac{\partial C_Y^{(h)}(y|y_0) \partial C_Y^{(k-1-h)}(y|y_0)}{\partial y_i \partial y_i}. \quad (6.31)$$

The change of variable based on the Jacobian for a differentiable function $\gamma(x)$ is given by:

$\nabla \gamma(x) = \sigma^{-1}(x)$, then we can write:

$$Y_t = \gamma(X_t) = \int_0^x \frac{du}{\sigma(u)}, \quad (6.32)$$

with Y_t satisfying:

$$dY_t = \mu_Y(Y_t)dt + dW_t. \quad (6.33)$$

This change of variable is intended to revert to log-likelihood of the process $\{X_t\}$.

Noting that the covariance of the multivariate process is given by $v(x) = \sigma(x)\sigma^T(x)$ and letting

$D_v(x) = \frac{1}{2} \ln[\det[v(x)]]$, we can transform Y_t to X_t according to (6.32). This change of variable is called a Lamperti transform.

Therefore, the log-likelihood of X denoted $l_X = \ln p_X$ is given by:

$$l_X(x|x_0, \Delta) = -\frac{1}{2} \ln(\text{Det}[v(x)]) + l_Y(\Delta, \gamma(x) | \gamma(x_0)) = -D_v(x) + l_Y(\Delta, \gamma(x) | \gamma(x_0)). \quad (6.34)$$

From (6.34), we can express the approximation of order K in Δ and reached the following [184]:

$$l_X^K(x|x_0, \Delta) = -\frac{m}{2} \ln(2\pi\Delta) - D_v(x) + \frac{C_Y^{(-1)}(\gamma(x) | \gamma(x_0))}{\Delta} + \sum_{k=0}^K C_Y^{(k)}(\gamma(x) | \gamma(x_0)) \frac{\Delta^k}{k}. \quad (6.35)$$

6-7-4) Queue Analysis

The variation of the traffic in the queue is given by the differential stochastic equation:

$dX(t) = B(x^* - X(t))dt + \sqrt{A}dW(t)$ which satisfies (6.21).

X being reducible and applying the change of variable $\gamma(x) = \sigma^{-1}x$, the resulting reducible process differential equation is given by:

$$dY_t = (\sigma^{-1}Bx^* - \sigma^{-1}Bx^*Y_t)dt + dW_t = k(\eta - Y_t)dt + dW_t,$$

where $\eta = \sigma^{-1}x^* = [\eta_i]_{i=1,2}$, $k = \sigma^{-1}B\sigma = [k_{i,j}]_{i,j=1,2}$; and $\sigma = \sqrt{A}$.

The coefficients of the series expansion are obtained as follows:

$$C_y^{(-1)}(y/y_0) = -\frac{1}{2}(y_1 - y_{01})^2 - \frac{1}{2}(y_2 - y_{02})^2,$$

$$C_y^{(0)}(y/y_0) = -\frac{1}{2}(y_1 - y_{01})[(y_1 + y_{01} - 2\gamma_1)k_{11} + (y_2 + y_{02} - 2\gamma_2)k_{12}] \\ - \frac{1}{2}(y_2 - y_{02})[(y_1 + y_{01} - 2\gamma_1)k_{21} + (y_2 + y_{02} - 2\gamma_2)k_{22}],$$

$$C_y^{(1)}(y/y_0) = \frac{1}{2}(k_{11} - [(y_{01} - \eta_1)k_{11} + (y_{02} - \eta_2)k_{12}]^2) + \frac{1}{2}(k_{22} - [(y_{01} - \eta_1)k_{21} + (y_{02} - \eta_2)k_{22}]^2) \\ - \frac{1}{2}(y_1 - y_{01})[(y_{01} - \eta_1)(k_{11}^2 + k_{21}^2) + (y_{02} - \eta_2)(k_{11}k_{12} + k_{21}k_{22})] \\ + \frac{1}{24}(y_1 - y_{01})^2(-4k_{11}^2 + k_{12}^2 - 2k_{12}k_{21} - 3k_{21}^2) - \frac{1}{2}(y_2 - y_{02})[(y_{01} - \eta_1)(k_{11}k_{12} + k_{21}k_{22}) + (y_{02} - \eta_2)(k_{12}^2 + k_{22}^2)] \\ + \frac{1}{24}(y_2 - y_{02})^2(-4k_{22}^2 + k_{21}^2 - 2k_{12}k_{21} - 3k_{12}^2) - \frac{1}{3}(y_1 - y_{01})(y_2 - y_{02})(k_{11}k_{12} + k_{21}k_{22}),$$

$$C_y^{(2)}(y | y_0) = \frac{1}{12}(2k_{11}^2 + 2k_{22}^2 + (k_{12} + k_{21})^2) \\ + \frac{1}{6}(y_1 - y_{01})(k_{12} - k_{21})[(y_{01} - \eta_1)(k_{11}k_{12} + k_{21}k_{22}) + (y_{02} - \eta_2)(k_{12}^2 + k_{22}^2)] \\ + \frac{1}{12}(y_1 - y_{01})^2(k_{12} - k_{21})(k_{11}k_{12} + k_{21}k_{22}) + \frac{1}{12}(y_2 - y_{02})^2(k_{21} - k_{12})(k_{11}k_{12} + k_{21}k_{22}) \\ + \frac{1}{6}(y_2 - y_{02})(k_{21} - k_{12})[(y_{01} - \eta_1)(k_{11}^2 + k_{21}^2) + (y_{02} - \eta_2)(k_{11}k_{12} + k_{21}k_{22})] \\ + \frac{1}{12}(y_1 - y_{01})(y_2 - y_{01})(k_{12} - k_{21})(k_{22}^2 + k_{12}^2 - k_{11}^2 + k_{21}^2).$$

To ensure no interference as stated above, the non-diagonal term of the matrix $K = \{k_{ij}\}$ must be zero, making the infinitesimal mean matrix to have real eigenvalues. The conditions $k_{11} > 0$ and $k_{22} > 0$ are necessary to impose on the process to be stationary, so that the standard asymptotic gives the asymptotic distribution of the maximum likelihood estimators.

Since we are interested in MLE, we define $X_e = [X_{e1}, X_{e2}]'$ as the asymptotical estimate value, therefore the drift of the Diffusion process is reverting and the diffusion equation process becomes:

$$dX(t) = B(X_e - X(t))dt + \sqrt{A}dW_t.$$

Also, since the process is reducible, Ito's lemma of the variable γ satisfies:

$$\nabla\gamma(x) = \sigma^{-1}(x).$$

Each element of m -dimension diffusion is reducible by means of the simple transformation known as Lamperti Transform:

$$Y_t \equiv \gamma(X_t) = \int_{x_t} \frac{du}{\sigma(u)}.$$

Therefore $X_t = \exp(Y_t) = \gamma^{-1}Y_t$ to revert to the process X .

The test of Log-transition probability is then done in MATLAB and the result is shown in the figure below.

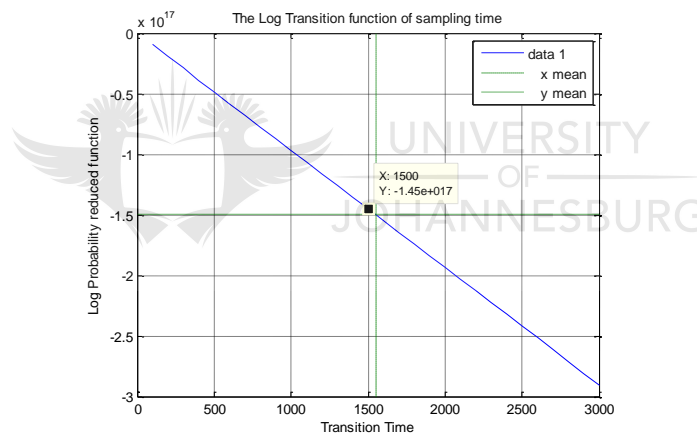


Figure 6.5: Log-transition probability

The plot the Figure 6.5 shows that the log-transition probability is a straight line with a negative slope. Therefore, the probability of the reduced function is an exponential distribution with a rate given by the forward Kolmogorov's equation (6.28) as follows:

$$\frac{\partial L_y}{\partial \Delta} = \sum_i \frac{\partial \mu_y}{\partial y_i}.$$

Since the infinitesimal mean is $d\mu_y = K(\eta - y)dt$, it comes that this rate given by the drift of the diffusion process is $-K$. This implies that the transition probability is in the form $P_y(t) \propto \exp(-K\Delta)$.

As the Diffusion approximation's purpose is to overcome the exponential server by considering the mean and the variance of the service time distribution, K being a function of the mean and the variance $K = \sigma^{-1}\beta\sigma$ is suitable.

Since the change of variable is independent of Δ the sampling interval, and according to Kolmogorov's equation related to the Diffusion process X , we can deduce that the probability transition of the Diffusion process $X(t)$ is given as the previous result in Chapter 5 by:

$$P_x(t) = \psi \exp(-Kt),$$

where ψ is the normalization coefficient; and K is the diffusion coefficient of drift and volatility of the Diffusion process. We saw that the inference of these exponential increments can be written as in Chapter 5 yielding the geometric distribution:

$$\hat{p}_n = \begin{cases} 1 - \rho, & n = 0, \\ \rho(1 - \hat{\rho})\hat{\rho}^{n-1}, & n \geq 1. \end{cases}$$

where $\hat{\rho} = \exp\{-kt\}$.



6-8) CONCLUSION

This chapter has proved that the On-Off sources can be modelled as Gaussian sources. The MLE has been chosen in our methodology, through the sample mean and the sample variance to accommodate the variability due to the fact that the number of active sources and the silence duration are variable by applying the central limit theorem. As a result, we reach a Geometric queue length distribution.

At level 2 multiplexing, we obtained the Log-Likelihood that suggests also the Gaussian properties. Because the Diffusion process is a Markov process, we obtain here a log-likelihood function sampled at finite time intervals. The observation of such a process is reduced to the sum of log-transition of successive pairs of observations, allowing us to characterize the model in compliance with the sampling interval of the scaling requirement mentioned in this chapter. By inference, this resulted in the same Geometric distribution.

Having analyzed the MLE for both multiplexing levels, the doors are now opened to carry out the implementation of our main objective : Level 1 and Level 2 multiplexing schemes. In the following chapter, we will carry out some tests to give an answer to our research question.

CHAPTER 7

EXPERIMENTS AND RESULTS

7-1) INTRODUCTION

The implementation of this project is based on the diffusion approximation of O-U as described in Chapter 6. To achieve the optimization, we introduced the maximum likelihood estimation (MLE) which allows us to exploit numerous properties of the Gaussian distribution from the resulting equations of the O-U Diffusion approximation. The Diffusion Process is an attempt to overcome the exponential server.

Our goal in this project is to better utilize the resources within an H.323 based network. This network provides some advantages as the “callability” and “RTP multiplexing” of its Endpoints make it feasible for the Egress (Outbound) traffic to be multiplexed as we are proposing here: a two level statistical multiplexing scheme. The first is intended to multiplex the On-Off sources, which in turn are multiplexed within a closed queuing network (see Figure 6.3).

Level 1 multiplexer (Mux 1) handles a sample of each active On-Off source, each with its own emitting rate. The Gaussian distribution is derived and the resulting queue is MG/G/1, standing for Markov Gaussian arrival (MG), General Service time (G) and 1 server.

Level 2 multiplexer (Mux 2) accommodates as source the level 1 multiplexer which is governed by a Markov Chain. Each source of k -states emits R_j packets/sec when in state j and moves to another state k with a non-zero probability in a reducible Markov Process. Hence also, the MMRP/G/1 queue: standing for Markov Modulated Rate Process arrival (MMRP), General Service time, and 1 server.

For both multiplexing levels M/G/1 (M for Markov) is retained for simplicity, we will rely on the queue length and/or queue transfer delay (provided by Little’s theorem) as test and MATLAB as programming tool to investigate our proposed solution.

7-2) EXPERIMENTAL DESIGN

In our experiment, we need for each multiplexer a test that shows that the positive correlations in the queue system are filtered out, as well as the reduction of the delay.

7-2-1) Experiment 1: Level 1 multiplexer

We are proposing in this project the Diffusion process as the underlying stochastic process of the statistical multiplexing. The approximation used herein is the Diffusion approximation of O-U as in Chapter 6. This characterizes the Level 1 multiplexing.

The aim of this experiment is:

- to show that the distribution function obtained from the On-Off sources is a Gaussian distribution, and
- By applying the Central limit theorem to evaluate the waiting time or the queue length.

Test Procedure and setup.

Let N_k be the total number of k -type sources with rate R_k ; let a_k be the number of active k -type sources with mean active period β_k^{-1} ; and let $N_k - a_k$ be the number of silence period of k -type sources with mean value α_k^{-1} .

Since the aggregate rate depends on the number of active k -type sources,

$$R(t) = \sum_k R_k a_k(t).$$

Since the number varies over time, let $E[N_k(t)]$ be the mean number of N_k sources, we can write $\beta_k^{-1} = E[a_k(t)]T$, where T is the packetization duration of the voice algorithm.

Test 1 of the Gaussian distribution

The distribution function of the number of k -type arrivals of the diffusion process at the equilibrium state is given by:

$$f_k(y) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{1}{2} \frac{(y - y^*)^2}{\sigma_k^2}\right\}$$

with mean $y^* = \frac{\alpha_k N_k}{\alpha_k + \beta_k}$ and variance $\sigma_k^2 = \frac{\alpha_k \beta_k N_k}{\alpha_k + \beta_k}$.

The maximum likelihood is obtained by the probability function:

$$P(y) = \frac{1}{\sigma_R^2 \sqrt{2\pi}} \exp\left\{-\frac{1}{2} \frac{(y - \hat{y})^2}{\sigma_R^2}\right\}, \text{ where } \hat{y} = \frac{1}{n} \sum_{k=1}^n y_k.$$

The probability of the queue to not be empty is given by:

$$P(Q > 0) = \frac{1}{\mu\sqrt{2\pi}\sigma_R^2} \exp\left\{-\frac{1}{2}\sigma_R^2\mu\right\} \text{ where } \mu = \frac{C_1 - m_R}{\sigma_R^2}.$$

According to Little's theorem the number of jobs in the queue is given by: $N_Q = m_R W$.

The maximum number of arrival packets is given by the maximum likelihood which is combined with the positive queue probability to yield N_Q so that the waiting time is given by:

$$W(y) = \frac{1}{m_R\mu\sqrt{2\pi}\sigma_R^2} \exp\left\{-\frac{1}{2}\sigma_R^2\mu\right\} \frac{1}{\sigma_R^2\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(y - \hat{y})^2}{\sigma_R^2}\right\}.$$

We carry out these two tests on MATLAB with two different sources namely G.711 A as source 1 and ADPCM as source 2. Our objective is to filter out the variability observed in the queue waiting time or length.

Source 1: G.711 A algorithm.

The parameters for this source are:

- $T_1=1\text{ms}$: standard value of packetization delay.
- $R_1=64\text{ kbits/s}$: standard bit rate.
- $N_1=30$ standard E_1 , Number of sources.
- $a_1=22$, number of active sources.
- $\beta_1^{-1}=a_1 \cdot T_1$, the mean active time.
- $\alpha_1^{-1}=100\text{ ms}$, the mean silence time:

Source 2: ADPCM algorithm

The parameters for this source are:

- $T_2=16\text{ ms}$: standard value of packetization delay.
- $R_2=32\text{ kbits/s}$: standard bit rate.
- $N_2=24$, number of sources.
- $a_2=16$, the number of active sources.
- $\beta_2^{-1}=a_2 \cdot T_2$, the mean active time.
- $\alpha_2^{-1}=\alpha_1^{-1}$, the mean silence time.

The results are analyzed using the MATLAB programs which are found in Appendix C. The results are shown in Figure 7.1.

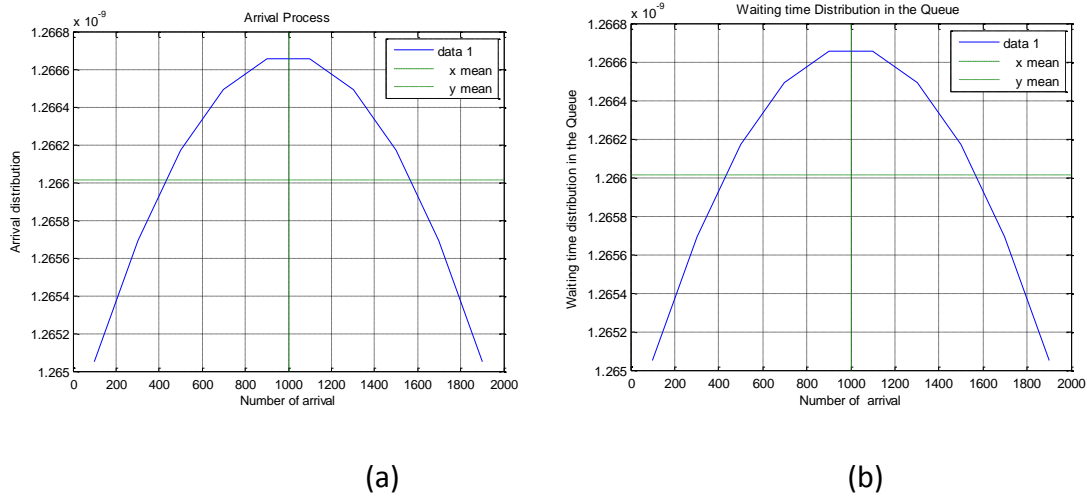


Figure 7.1: (a) The arrival process of On-Off sources using the Diffusion Approximation and (b) the waiting time Distribution in the Diffusion Approximation Queue

The test shows that the diffusion approximation of On-Off sources can be approximated to a Gaussian distribution with mean equal to the mean sample. This slope is intended to justify the use of Gaussian properties, particularly the Maximum Likelihood Estimators in all of our next developments.

Test 2: Queue length distribution and Waiting time in the queue

The queue changes (ΔQ) are captured by (6.5) provided the normal distribution. The stochastic process of the queue length is derived from the stochastic differential equation (6.5) and by rearranging and applying the central limit theorem on the number of arrival y on the following differential equation:

$$\left(\sum_{k=1}^K (R_k y_k^* - C)\right) \frac{\partial f(y, x)}{\partial x} = \sum_{k=1}^K \left[\left(\frac{N_k \alpha_k \beta_k}{\alpha_k + \beta_k}\right) \frac{\partial^2 f(y, x)}{\partial x^2} \right],$$

such that the solution yields the standard form we got in Chapter 5, we deduce the mean queue length is given by:

$$E[Q] = (\widehat{R}_k - C)\Delta t = \beta \Delta t,$$

where $\widehat{R}_k - C$ is the drift of the diffusion process and C the capacity transmission link.

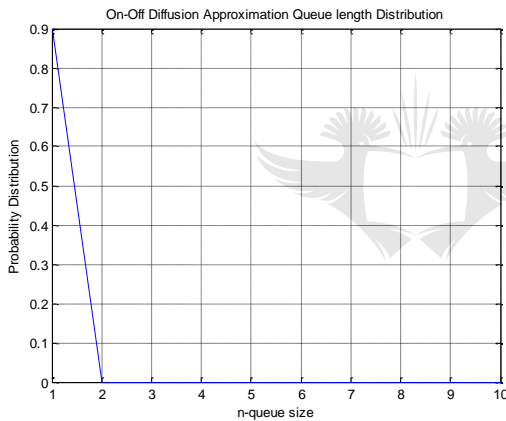
The variance in the queue is given by:

$$Var[Q] = (C_a \widehat{R}_k + C_s C) \Delta t = \sum_{k=1}^K \frac{2N_k \alpha_k \beta_k}{\alpha_k + \beta_k} \Delta t = \alpha \Delta t,$$

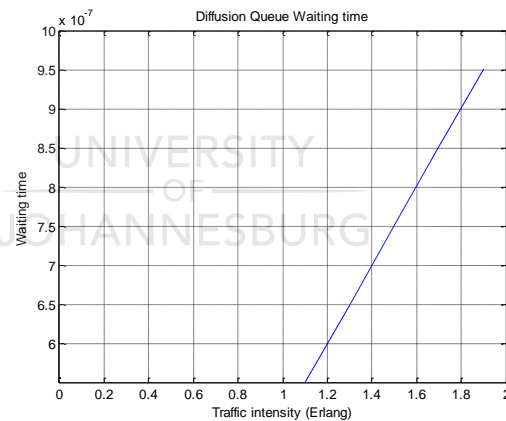
where C_a and C_s are the indexes of dispersion of the arrival and service time processes respectively, β and α are the drift and the volatility of the Diffusion process respectively. The

decrements are given by $\widehat{\rho} = \exp\{-\frac{2\beta}{\alpha}\}$ factor.

The probability distribution is $\widehat{p}_n = \rho(1 - \widehat{\rho})\widehat{\rho}^{n-1}$, $n \geq 1$.



(a)



(b)

Figure 7.2: (a) Diffusion queue length distribution overflow On-Off sources and (b) Waiting time distribution in the On-Off sources queue.

Figure 7.2 (a) shows that the probability excess $\Pr[X > x]$ presents a linear decay up to 20% of the number of packet in the queue.

Figure 7.2 (b) shows that the delay becomes linearly significant for traffic intensity bigger than 1 erlang. With that linearity, it comes that there no variation in time of mean and variance. It turns out that the diffusion approximation of O-U presents no correlations (time dependency) at all and it is suitable for heavy load traffic.

7-2-2) Experiment 2: Level 2 multiplexer

We keep in mind that the aggregate input of the Level 2 multiplexer has a Gaussian distribution with mean zero and the variance is positive. In this test, we are interested only in voice packet traffic, which assumes homogeneous time sources. As mentioned earlier on, the maximum likelihood is our solution method.

The aim of this test is to analyze the queue performance when using such multivariate Gaussian MLE, which actually can only be found in a closed form.

In this test we are interested in the queue size distribution or the queue waiting time filtered out from the variability. In test 1, we saw that the diffusion probability can accommodate +20% of margin in the queue size without apparent extra delay or variability. What can the results be with this new approximation?

To answer this question, we proceed with the MATLAB set up that can be found in Appendix C.

The mean queue change for a large sample time is given by $E[\Delta Q] = (\lambda - \mu)\Delta = \beta\Delta$ where λ is the arrival rate and μ is the processing rate.

The variance of this change process is given by $Var[\Delta Q] = (C_a + C_s)\Delta = \alpha\Delta$, where C_a and C_s are the squared coefficients of variation for the interarrival and service time respectively.

Each source is governed by a Markov Chain and a non-zero routing probability.

The transition arrival matrix is $P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$ and the diffusion process $X = [X_1, X_2]'$.

The infinitesimal mean $b = \begin{pmatrix} p_{11}\alpha_1 & p_{12}\alpha_1 \\ p_{21}\alpha_2 & p_{22}\alpha_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and the infinitesimal variance of the diffusion

approximation is $A = \begin{pmatrix} ((C_1 - 1)p_{11} + 1)p_{11}\alpha_1 x_1 & ((C_1 - 1)p_{12} + 1)p_{12}\alpha_1 x_1 \\ ((C_2 - 1)p_{21} + 1)p_{21}\alpha_2 x_2 & ((C_2 - 1)p_{22} + 1)p_{22}\alpha_2 x_2 \end{pmatrix}$.

Also, since we are interested in MLE, define $X_e = [X_{e1}, X_{e2}]'$.

The parameters then are:

- $C_1=3; C_2=2$; the squared coefficients of the two types of heavy traffic,
- $P_{11}=0.6; P_{12}=0.4; P_{21}=0.5; P_{22}=0.5$; the probability transition matrix,
- $\alpha_1=1/200 \cdot 10^{-3}; \alpha_2=1/100 \cdot 10^{-3}$; the holding time of the two states,

- $X_{01}=1000$; $X_{02}=1200$; the number in the queue at $t=0$,
- $X_1=1400$; $X_2=1600$; the number for the two types of sources,
- $C=100$ Mbits/s the transmission capacity link used for Ethernet,
- $X_{e1}=(1/5)C$; $X_{e2}=(1/10)C$ the transmission rate available for each source.

The results are the following plots:

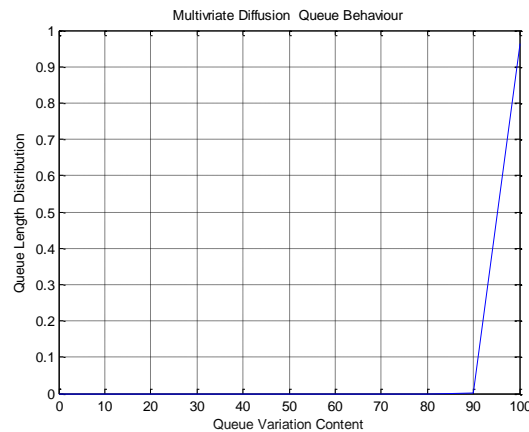


Figure 7.3: Multivariate Diffusion Approximation Queue Behaviour

Figure 7.3: The queue content becomes significant for high traffic intensities (0.98 erlang) in compliance with the result obtained at Level 1 multiplexing. By considering the probability of excess ($P[X>x] = 0.02$), and applying the Little's theorem in the scale of 10, it comes that -20% of excess is sensitive to the delay. Also, the indexes of dispersion are constant (no time dependency) regarding the linearity obtained as in the previous multiplexing level.

7-2-3) Experimental results

We are expecting in both cases, having a queue distribution and or a delay system capable of sustaining traffic intensity in the vicinity of 1 erlang without any correlation.

From Figure 7.1, one can see that the On-Off sources are modelled as Gaussian Markov Process which has led to the exploitation of the Gaussian properties.

This maximization of the arrival has given a simplified Kolmogorov's equation in probability for the queue content which result is shown in Figure 7.3.

Here, the waiting time or delay impacts the queue for an extremely heavy load, the traffic intensity more than 1 erlang. This is already shown in Figure 7.2, where the probability distribution has a maximum of the traffic load taken at the steady state, which is asymptotic for the present case. The changes in the queue experience a linear decrease up to 20% of the

queue size. The meaning of this is that the diffusion approximation can still accommodate $\pm 20\%$ of the size queue without adding any delay on the one required by the delay system. Actually, for the traffic intensity above 1 erlang, (1.1 erlang for this case), the Diffusion process presents a linear growth as the Fluid Process. Contrary to the fluid model, the drift and the volatility of the approximating process has been taken into account. The linear output is a proof that the variability, cause of the transfer delay is now the matter of the past. This result also shows that the delay is bounded by the asymptotical approximation through the central limit theorem exploiting the MLE, avoiding the condition of the explosion of the system.

7-3) QUEUE IMPLEMENTATION

7-3-1) Queue implementation algorithm in MATLAB

We consider a statistical multiplexer whose inputs are made up of two incoming links with rate r_1 and r_2 . We want to determine the Maximum Likelihood Estimates (MLE) from the resulting Gaussian distribution derived from the arrival process as we are expecting a large number of sources. These parameters, the sample mean (sam) and variance (sav) as the Diffusion approximation resulted to a Geometric distribution allow to calculate the decrement factor r_i of the process in view to analyze the fluctuations in the queue from the traffic intensity r_0 .

The related queue behaviour is given by the plots of the multiplexer output first as an histogram showing the different output levels; and the stair figure to capture the transition epochs. These plots need to be compared to their exponential counterpart as we stated earlier. To get the results, the following mathematical is needed.

The geometric distribution characterizing the queue behaviour according to Chapter 6 is given by:

$$p(n) = r_i^{n-1} (1 - r_0).$$

The decrement factor is obtained through the expression $r_i = \exp\{-2 \frac{sam}{sav}\}$, $0 < r_i < 1$.

The sample mean and the sample variance are given by:

$$sam = \frac{1}{n} \sum_{i=1}^n x_i \text{ and } sav = \frac{1}{n} \sum_{i=1}^n (x_i - sam)^2,$$

sampled from the diffusion process $\{X(t)\}$ defined by $dx(t) = \beta dt + \alpha dw$ where w is Brownian.

These parameters are taken into account in the equations of the two main developments of this thesis which are:

- Level 1 statistical multiplexer:
- Level 2 statistical multiplexer:

The exploitation of the equations above mentioned requires the following inputs:

- n_1, r_1 the number and rate of type 1 sources;
- n_2, r_2 the number and rate of type 2 sources;
- and C the transmission link capacity in bit/sec.
- N the number of samples

The processing phase outputs:

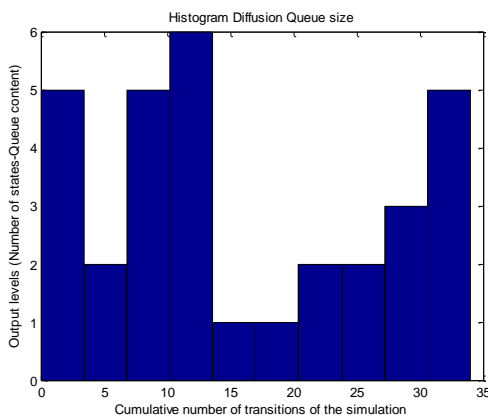
- The number and the processing rate in the queue.

The structure plan to solve this problem is given in Appendix-C as a MATLAB program where we choose to express the unbiased MLE to avoid having the expected values different of the parameters being estimated; and to take into account the reverting Diffusion process from its asymptotical stationary state. The following subsection presents the results obtained.

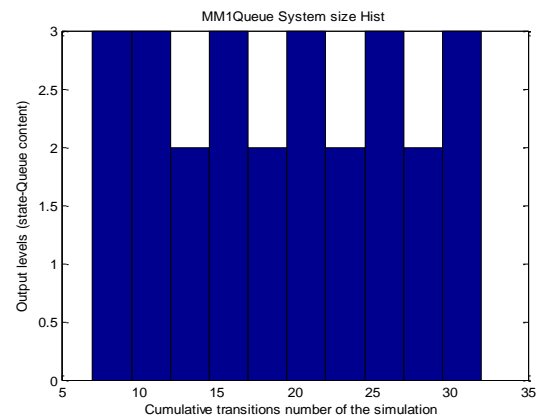
7-3-2) Diffusion Queue simulation and Comparison with M/M/1 queue

a) MATLAB simulation

The simulations of the Diffusion queue as prescribed above in MATLAB have given the following results, which are compared with the exponential Queue.



(a)



(b)

Figure 7.5: (a) The Diffusion and (b) the M/M/1 queue contents

X-axis represents the cumulative number of transitions of the simulation while Y-axis is the number of output levels (state-queue content).

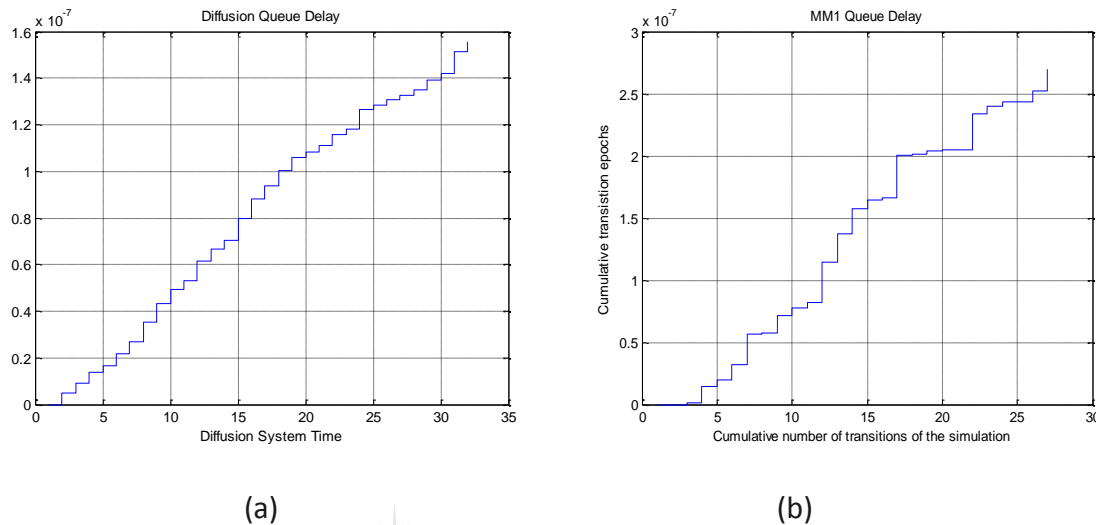


Figure 7.6: (a) Diffusion queue and (b) M/M/1 queue transition times

X-axis represents the cumulative number of transitions of the simulation while Y-axis is the required time in stairs for each transition epochs.

b) Simulation results

The simulation of the M/G/1 queue using the Diffusion Approximation is now directly compared to the M/M/1 queue.

The M/M/1 queue shows through the histogram (Figure 7.1 (b)) two levels, which comply roughly with the definition of the statistical multiplexer stating that the resource allocation lies between the average and the peak value. But when we get inside the transition epochs through the stair graphs (Figure 7.2 (b)), one can observe variabilities in the M/M/1 queue, which the previous studies have shown are the cause of an extra delay in the transfer delay. This is responsible for packet loss and of course for low average call duration.

The histogram (Figure 7.1(a)) of the M/G/1 queue when using the Diffusion approximation however shows multiple levels of the queue content. One actually would expect more variations in the transition times or processing times, which is not the case as it is shown through the stairs graph (Figure 7.2 (a)). This can be explained by the fact that the Diffusion Approximation Process exploiting the Gaussian property of MLE has resulted in reducing the

correlated non-renewal process into a Markov renewal process in the sequence $(X(t), Q(t))$ of the phase of the Markov process $X(t)$ and the queue content $Q(t)$ at the departure times $t > 0$. It is also important to mention that the variables $(X(t), Q(t))$ satisfy $N + 1$ levels as in the Fluid process.

The transfer delay of the Diffusion process according to the results above is about 0.6 times the transfer delay of the exponential queue, resulting in smaller buffer size. As in the simulation conducted regarding the M/G/1 Diffusion queue and the M/M/1 queue, if the LAN card of 100 Mb/s provides 250 ms of buffering, the buffer size is $250 \cdot 10^{-3} \times 100 \cdot 10^6 = 3$ Mbytes minimum, while 5 Mbytes is needed for the Poisson queue (8 kHz sampling rate). This result implies, in the new network paradigm introduced in Chapter 1, the possibility to process more voice packets at the edges of the LAN without any extra delay, therefore to provide service to millions of users.

7-3-3) A workable solution

A workable implementation of this design requires besides the control algorithm above, the following elements:

- Stream or Flow control protocols;
- Process control; and
- Multiplexer design.



Flow and Process Control

We have seen in Chapter 2 that H.323 is a collection of protocol suites built to provide a guaranteed Quality of Service (QoS). This protocol suite ensures the flow control through the H.245 protocol. The H.245 messages are managed by the Multipoint Controller (MC) which implements the features required by the control functions.

Also, to process signals, H.323 provides an entity called the Multipoint Processor (MP). The MP is capable of receiving M -audio inputs from where it can output N -audio signals in any combination between M to N as the controller requires in the above proposed algorithm.

Both entities can be found in the Endpoint called the Multipoint Control Unit (MCU) which in turn must be callable. This leads to the following characteristics:

SCN Gateway Characteristics:

Besides of the functions related to the Gateway that are presented in Chapter 2, we can add:

- H.323 Gateway enhanced with MCU (MC+MP),
- E-model Requirements,
- Multiplexing M input, N output,
- Processor speed = sum of SCN line cards speed.

LAN Internet Bridge Characteristics

- MCU bridge (MC+MP),
- LAN router,
- E-model Requirements,
- Multiplexing M input, N output,
- Processor speed = LAN line card speed.

Multiplexer design

We are proposing a FIFO queue implemented in the Shift Register (SR) associated with a Look-Up Table (LUT): SRL16, assuming a transport of a 16-bit coding of the voice packet.

- Shift Register of depth-16 in one 4-LTU cell

This combination improves the performance; leads to significant cost savings; enables hardware design where delay or latency compensation is required and is useful for FIFO application.

- LUT 16:1 multiplexer with:

4 input ports address, data input, clock and clock enabled as in the synchronised RAM.

One output provided by the last indexed flip-flop which normally is 15 called Q15 (Q_0 - Q_{15}) from the library primitives or called MC15 if one uses FPGA program;

- Dynamic length adjustment: the length of the Shift Register address can be changed dynamically. This emulates n -bits Shift Register providing an asynchronous output. The m -cascade of cell can be implemented to achieve $n \times m$ -bits Register; and
- The MUXF5, MUXF6, MUXF7 multiplexers are required accordingly for the selection which is very important to achieve a variable output channel.

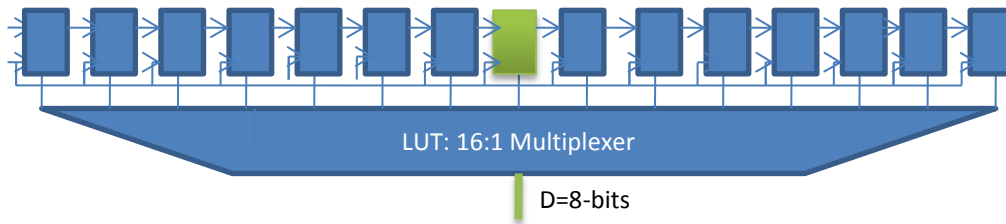


Figure 7.4: An SRL16 Register

We did not direct our project to implement such a register since some of its characteristics rely on the network parameters:

- The buffer sizing depends on the Round Trip Time (RTT) and the number of RTP packet window by exploiting its transport multiplexing and/or its compressed header scheme.
- The transmission link capacity utilization relies on the fluctuations in the queue as analyzed in our approach, which in turn depends on the buffer size.
- The bit rate estimates that are achieved through measurements in compliance with the Link layer of the access network with the aim to model the LAN traffic.

The operability of such a queue in the above conditioned hardware and software will be done in future studies.

7-4) PROJECT RESULTS

Our main results are listed as follows:

- The diffusion approximation achieves the exponential server that gives rise to the Geometric probability distribution. We have seen by inference that the Diffusion can be approximated accurately into to a Geometrical distribution with decrement factor $e^{rate < 0}$. This has been developed in the application for univariate and multivariate Diffusions whereby we define the probability distribution of the process to be the following:

If $\hat{\rho} = e^{rate < 0}$, then we shall write: $P(n) = (1 - \rho)\hat{\rho}^{n-1}$, $rate$ is the diffusion index taking into account the drift and the volatility in the queue.

- The diffusion approximation, like the fluid method accommodates variable bit rates and outputs a linear distribution (waiting time or queue length), thus filtering out the positive correlations that cause high transfer delay, which is the cause of packet loss. This result confirmed by the constant transition epochs in the queue, comes from the fact that, unlike the exponential server with a fixed rate which is defined *a priori*, the Diffusion server achieves

instead the exponential server in the aggregate traffic which is a function of the fluctuations in the queue, therefore outputting the required rate such that the random variable $(X(t), Q(t))$ of state and queue content is a Renewal process.

- Furthermore, the waiting time and queue length satisfy the linear growth in the domain of the Diffusion process. As a consequence, one observes that unlike the Fluid method that presents a linearity that increases to infinity, the Diffusion approximation through the MLE determines the bounds of the process.

Also, since the Diffusion queue is the Fluid method that takes into account the various time scales and the related changes, it gives accurately better values compared to other queue solutions focusing on the randomness in the queue but with a specificity that the heavier the traffic, the best it fits with the Diffusion approximation in terms of delay and queue length.

- The delay in the queuing system therefore becomes significant only for traffic intensity closer 1 erlang (100% transmission link capacity utilisation) without exhibiting a self-similar traffic or heavy tail traffic. It turns out that the Diffusion approximation simulates queues with probability $\Pr[Q > 0] \approx 0$ or better for $\Pr[X > x] \approx 0$ of buffer content excess, as long as the traffic intensity is less than 1 erlang: this means that as long as the utilization of the transmission link capacity is less than 100% (100% is not advised to achieve in network planning), the queue delay has trend to be neglected. This result has been achieved in both statistical multiplexers proposed herein. It comes from the fact that the drift (fluctuations of the queue mean) and the volatility (variation of the standard deviation) have set through the MLE a margin in the queue length design which is $\sqrt{n}(X_{mle} - \bar{X})$ as the central limit theorem implies. We can therefore optimize the resources involved in the communications.

The buffer sizing is given by the window packet size (WS) which takes into account the RTP application multiplexing on top of the transport level of the protocol suites and the Round Trip Time (RTT).

If the transmission link capacity is C packets/sec, the buffer size is given by $C \times RTT$ packets. Or in the other way, if the buffer size N is defined, the processing rate of the Multipoint Processor (MP) is given by $N \times WS / RTT = C$.

In our case, the RTP can use its multiplexing functionality to present N window sizes of voice packets. The Diffusion process and the central limit theorem as we applied in this project yield: $W \rightarrow N(N \times WS, N \times Var)$ where W is the sum of window sizes. This gives as standard deviation $Sd(W) = \sqrt{N \times Var} = \sqrt{N}(X_{mle} - \bar{X}) = \sqrt{N}(\alpha WS)$.

The buffer size is therefore given by $W = C \times RTT \pm \alpha C \times RTT / \sqrt{N}$ from where the level of confidence can be set accordingly. But the result obtained here suggests small size buffer is suitable for the queue.

In optimizing the bandwidth, the above results suggest that the queue length is a function of the transmission link capacity. Therefore the optimization of the transmission link depends on the queue length distribution and on one parameter that is the packet loss ratio, which as part of our studies is defined as:

$$PLR = \int_C^{\infty} \left(\frac{r - C}{\mu_{\bar{R}}} \right) f(r) dr = \frac{E[R(t) - C]^+}{E[R(t)]},$$

which is constant and where $f(r)$ is the Gaussian distribution with mean $\mu_{\bar{R}}$ and variance $\sigma_{\bar{R}}^2$.

For this project, we saw that the small size buffer is required to satisfy the traffic flow whether it is synchronous or asynchronous with the Diffusion Approximation. This ratio is maintained in a small range defined by the drift of the Diffusion process in compliance with the ITU-T in terms of packet loss.



7-5) CRITICAL ANALYSIS

We have reached as the results show that the sequence $((X(t), Q(t)))$ is a renewal process over $N + 1$ output levels. These results are conform to the fluid model in terms of outputting variable bit rate and join the assumption of a renewal process of the MMPP analysis conducted by Heffes and Lucantoni [160] in terms of moments. It is noteworthy to mention how much these two queue solutions have been proven effective in handling television variable bit rate signals, voice and data packets in the statistical multiplexer. However, we cannot conclude yet the media convergence.

In fact, we have used for simplicity reasons, the reducibility property of the Diffusion process to achieve that Markov Renewal process from a correlated non-renewal process, leaning only on the characteristics of voice packets we presented earlier in this thesis (Chapter 3).

In real world, where control information matters, while the voice signal follows the UDP transport mechanism, its control information in contrast is carried by the TCP transport mechanism. Therefore, the Fokker-Plank equation (6.22) cannot be reduced as we did in (6.26) since $i \neq j$. The irreducibility property of the Diffusion process, which is a more general approach,

is needed to determine the coefficients of the log-expansion expression (6.27) in view to model our one server FIFO queue. This irreducibility solution can also be used, although out of our scope, to capture the variability of audio and video in IPTV.

Also, in this simulation, we worked with standard bit rates assuming zero bit errors, which is not the case in the real world since the bit errors that may occur in the LAN may affect the bit rates or the latter may need a recovery mechanism that could add another delay. The bit rate estimate needs to be determined through measurements and processed by fitting operations according to the sampling time of a processing cycle of the MP. These bit-estimates can be carried on through our next studies.

7-6) CONCLUSIONS

The results obtained in this chapter, whether Level 1 or Level 2 multiplexer show that we can accommodate variable bit rate sources without experiencing an extra delay for heavy load traffic when using the Diffusion approximation process. Previous studies have shown that this delay is the result of cumulative correlations due to the traffic's burstiness. Herein, we achieved two queues, MG/G/1 and MMRP/G/1, with a linear waiting time showing besides the fact that the delay is neglected (probability equal to zero for level 1 and near zero for level 2), but also that the variability is filtered out.

This result is confirmed by the simulation where the transition epochs in the Diffusion queue exhibit a Markov renewal process to achieve multiple queue content levels.

The multiple content level of the queue shows its ability of outputting variable bit rates. These multiple levels can be dynamically accommodated as shown by the SRL register, where the look up table sets asynchronously the pointer to the required flip-flop to achieve n -bits/s transfer rate.

This ultimate point leaves us with saliva in the mouth, leading to the future works, which will be described in the next chapter.

CHAPTER 8

CONCLUSION AND FUTURE WORK

8-1) INTRODUCTION

In this project we have set as a goal the optimization of resource allocation in an H.323 network. A two level multiplexing scheme is then proposed as a compromise in terms of queue delay or queue size and bandwidth management. But the bandwidth has never been a problem at all regarding service rates available today in terms of transmission link, when it comes to transporting real time services over the Internet. The problem lies in the delays observed in the queue which are the major cause of packet loss, therefore the low level of quality offered by those services. By investigating the different queue models of different related works, we aimed to show different problems that these models' applications are experiencing in view to justify our proposal: the use of the Diffusion approximation as a compromise for future statistical multiplexers. This choice is backed by the answer of the research question we set at the beginning of this thesis.



8-2) ANSWERING THE RESEARCH QUESTION

The queue waiting time or the queue length has been smoothed through the properties of the Gaussian distribution, bringing therefore an answer to our research question:

- The Renewal process shows that the correlation in the queue has been filtered out, justifying the choice of the Diffusion process as a statistics analytical tool.
- Large queue sizes result when the traffic from multiple sources starts arriving in bursty state leading to a sustained transmission at the peak rate; that causes with time a build-up in the queue size. This is not the case when using the Diffusion process. The confirmation comes when comparing the Diffusion and the Exponential queues which allows us to conclude that the Diffusion process approximated into a Renewal process leads to less delay in the queue.
- A small delay system means a small buffer size. The delay system is determined from the RTT (Round Trip Time) and the transmission link capacity in which we add the margin defined by the Diffusion approximation to yield a sojourn time in the system that is a function of the type of sources involved. Thus, different types of sources with

different bit rates are accommodated within a queue served by a single server with a service time that depends on $(X(t), Q(t))$. The ability to adapt frame sizes in response to traffic variation shows the improved performance in terms of delay system (buffer) and bandwidth. This results in a better utilization of buffer and bandwidth achieving the optimization of resource allocation.

- The RTT plays a key role in terms of resource allocations. The less it is, the better the resources are optimized. Therefore, the H.323 Multipoint Control Units must be implemented at the edges bridging the network where the routing, the CoS, and the E-model requirements are also implemented, indicating the required Endpoints configuration.

8-3) OUR CONTRIBUTION

Smaller buffer size are needed: we have seen that for an 8 KHz sampling rate signal, 3 Mbytes buffer size for a 250 ms buffering is needed when using Diffusion approximation while 5 Mbytes is required for Poisson queue. We bore in the mind the New Generation Network as mentioned in Chapter 1, whereby the intelligence is moved to the transport part of the network, outlining the key role the access network will have to play in such a case. Our intention was not to reduce the buffer size regarding the current trend of memory in the market with the price decreasing every year, but instead to propose at the access network layer, a statistical multiplexing scheme capable of processing more signals within the same time interval without experiencing any extra delay. This quality of service improvement can result to providing voice service to millions of users, increasing therefore revenues necessary for the return of investment.

Transcoding is avoided: transcoding is required to manage resources, adding another delay for the on-going processing. Adapting the output level to the multiplexer input and the process phase could allow avoiding the transcoding.

Service degradation minimized: in the ITU-T Rec. H.323, we saw that dynamic changes of bandwidth are requested when congestion occurs, causing service degradation. With the variable bit rate output reached in this thesis, the bandwidth changes request during conversations would be reduced, therefore minimizing service degradation.

8-4) FUTURE WORK

We have just started in this scope a wide open field in terms of work. It is noteworthy to recall that not that much has been done in the Diffusion process as solution for a queue model. But

the analysis done herein with the introduction of the Maximum Likelihood Estimators (MLE) is proof that this solution has to be carried out in the real world.

We determine the condition of this implementation. The implementation of such a multiplexer as we said, need a Shift Register combined with the Look-Up-Table 4-bits address called SRL 16 and a MUXFx device as selector. This will be among the future work since this entails some parameters as RTT and the bit rate of LAN Ethernet traffic.

The bit rates estimates of the Ethernet traffic are obtained through the fitting procedures. The procedure allows avoiding bit-error correction that may add another delay in the bit rates recovery.

We left aside some properties of the MLE, particularly in the multivariate environment. The MLE is sensitive to outliers; parsing data in different groups may result in a multimodal statistical structure. This could give an opportunity to define a new form of prioritization of the queue for different classes of traffic to better control the real time services in the network, rather than relying as we did on the relative QoS called CoS.

Also, the analysis carried out in this thesis is based on reducible time homogeneous Diffusion processes. This has left wide open the non-reducible Diffusion processes as well as time non-homogeneous Diffusion processes.

We have in terms of resource allocation focused on the buffer where the transfer delay matters and the impact of the buffer sizing on allocation of the bandwidth. We mention as a performance parameter, the packet loss ratio. The evaluation of this parameter has to be taken into account in network management by providing a monitoring tool for this particular development.

REFERENCES

- [1] ITU-T Recommendation E.439: "Test call measurement to assess N-ISDN 64Kbps circuit," March 2000.
- [2] ITU-T Recommendation E.416: "Network management principles and functions for B-ISDN traffic," March 2000.
- [3] ITU-T Recommendation E.351: "Routing of multimedia connections across TDM-ATM and IP based networks," Sept 2006.
- [4] F. Gonzalez, "VoIP in Public Networks: Issues, Challenges and Approaches," *Infrastructure and Application Summit ITU-T. Africa*, Johannesburg, 2001.
- [5] D. Wischik and N. Mckeown, "Buffer Sizes for Core Routers", *SIGCOMM Computer Communication Review*, Vol. 35, No. 2, pp. 75-78, July 2005.
- [6] ITU-T Recommendation G.709, "Dense Wavelength Division Multiplexing: Interface for OTN," 2001.
- [7] ITU-T Recommendation G.704, "Synchronous frame structure used at 1544, 6312, 2048 Kbps," Oct. 1998.
- [8] ITU-T Recommendation G.783, "Characteristics of Synchronous Digital Hierarchy Equipment Functional blocks," 2006.
- [9] VoIP Network monitoring Report, Nextone U.K., May 2010.
- [10] IEEE 802.3u 100Base-T Fast Ethernet at 100 Mbps, 1995.
- [11] ITU-T Recommendation H.323, "Visual Telephone Systems and Terminal Equipments for LANs which provide none guaranteed QoS," 1996.
- [12] IEEE 802.3i, "10Base-T Ethernet at 10 Mbps," 1990.
- [13] IEEE 802.3j, "10Base-F Ethernet at 10 Mbps, 100Base-TX/T4/FX Ethernet at 100 Mbps," 1993.
- [14] IEEE 803.3z, MAC Frame with Gigabit Carrier Extension, 1998.
- [15] G. Eichler, "Implementing Integrated and Differentiated Services for the Internet with ATM Networks: A Pratical Approach". *IEEE Comm. Mag.*, Vol. 38, pp. 132-142, 2000.
- [16] IEEE 802.1P, "Traffic class expediting and dynamic multicast filtering," 1998.

- [17] IEEE 802.Q: "Standard for LAN and WAN MAC bridges and Virtual Bridges," 1998.
- [18] IEEE 802.1D, "Standard for LAN and WAN Media Access Control (MAC) Bridges," 1998.
- [19] IEEE 802.11, "Implementing wireless LAN Computer Communication in the 2.4 GHz frequency band," Sept. 1999.
- [20] IEEE 802.16, "Wireless LAN and recommended practices to support developments and deployments of broadband wireless MAN in 10-66 GHz," 2001.
- [21] RFC 1349, "Priority queue for Upstream traffic based on ToS field," July 1992.
- [22] RFC 791, "Internet protocol specification of how traffic travels over Internet," Sept. 1981.
- [23] RFC 1812, "Requirements for IP version 4 routers," June 1995.
- [24] RFC 2474, "Definition of the Differentiated service field in the IPV4," Dec. 1998.
- [25] RFC 2205, "Resource Reservation Protocol, V1, Functional Specifications," Sept. 1997.
- [26] RFC 2225 "Classical IP and ARP over ATM, V1, Functional Specifications," Sept. 1997.
- [27] D. Awduche, "MPLS and Traffic Engineering in IP Networks". *IEEE Comm. Mag.*, Vol. 37, pp. 42-47, 1999.
- [28] ITU-T Rec. G.113, The E-model, "Transmission Impairments due to speech processing," Nov. 2007.
- [29] ITU Rec G.107 The E-model, "A computational use in transmission planning," Feb 12, 2003.
- [30] ITU G.109, "Definition of Categories of Speech Transmission Quality," Sept. 1999.
- [31] ITU Rec G.113, "Transmission Impairments," *World Telecommunication Standardization Conference*, Helsinki, March, 1993.
- [32] RFC 1180, "Steps in Forwarding IP Datagram from Source to Destination Hosts," Jan. 1991.
- [33] ITU-T Rec. G.711, "Pulse Code Modulation of Voice Frequencies," Nov. 1988.
- [34] ITU-T Rec. G.722, "7 KHz Audio-Coding within 64 kbps," Nov. 1988.
- [35] ITU-T Rec. G.723, "Dual Rate Speech for Multimedia Communication Transmitting at 5.3 and 6.3 kbps," 1996.
- [36] ITU-T Rec. G.728, "Coding of Speech at 16 kbps using Low-delay Code," Sept. 1992.

- [37] ITU-T Rec. G.729, "Coding of Speech at 8 Kbps using Conjugate Structure." Erratum, 2006.
- [38] ITU-T Rec. H.225.0, "Call Signalling Protocols and Media Stream Packetization for Packet-based Networks," Dec. 2009.
- [39] ITU-T H.200/AV-120 Series Recommendations, "Service Requirements for Visual Telephone Services," 1993.
- [40] ITU-T Rec. Q-931, "ISDN Connection Control Signalling Protocol, Part of DSS1," May 1998.
- [41] ITU-T Rec. H.245, "Control protocol for multimedia communication," May 2006.
- [42] CCITT X.208 on "Abstract Syntax Notation Specification Rules and Structures for Representing, Encoding, Transmitting, and Decoding Data in Telecommunication and Computer Networking," 1988.
- [43] M. Dom, "Overview of R.A.S protocol," *Telecommunication Protocols*, Feb. 2007.
- [44] RFC 3550, "A Transport Protocol for Real Time Application (RTP/RTCP)," July 2003.
- [45] RFC 1889, "A RTP Transport Protocol for Real Time Application," Jan. 1996.
- [46] ITU-T Rec. H.310, "Broadband Audiovisual Communication System and Terminals," 1998.
- [47] ITU-T Rec. H.321, "Adaptation of H.320 Visual Telephone Terminals to B-ISDN environment," Feb. 1998.
- [48] ITU-T Rec. H.322, "Visual Telephone Systems and Terminals Equipments for LAN which provide a Guaranteed QoS," March 1996.
- [49] ITU-T Rec. H.324, "Terminals for Low Bit rate Multimedia Communication," June 2001.
- [50] ITU-T Rec.V.70, "Procedures for simultaneous transmission of data and digital encoded voice signal over GSTN or over 2-wire based point-to-point telephone type circuit," July 1996.
- [51] ITU-T Rec. H.231, "Multipoint Control Units for Audiovisual Systems using Digital Channels up to 1920 kbps," July 1997.
- [52] ITU-T Rec. T.120, "Data Protocol for Multimedia Conferencing," July 1996.
- [53] M. Lewis, "A/ μ law: A Companding Implementation," Application Notes.
- [54] MPEG Audio, "MPEG Audio Compression Basics," Dec. 1999.

- [55] ITU-T Rec. H.221, "Frame Structure for a 64 to 1920 Kbps Channel in Audiovisual Teleservices," March 2009.
- [56] ITU-T Rec. H.242, "System for Establishing Communication between Audiovisual Terminal and Digital Channels up to 2 Mbps," March 2009.
- [57] ITU-T Rec. H.243, "Procedures for Establishing Communication between three or more Audiovisual Terminals using Digital Channels up to 1920 kbps," Oct. 2005.
- [58] ITU-T Rec. G.732, "Characteristics of Primary PCM Equipments," Nov. 1988.
- [59] W. Wu, H. Bulut, A. Uyar, and G. Fox, "Adaptating H.323 Terminals in Service Oriented Collaboration System," *Internet Computing IEEE*, Vol. 9, pp. 43-50, 2005.
- [60] K. Chandra, "The Statistical Multiplexing", *Encyclopedia of Telecommunications*, John Wiley and Sons, Inc, 2003.
- [61] J. W. Roberts and J. T. Virtamo, "The Superposition of Periodic Cell Arrival in an ATM Multiplexer," *IEEE Trans. Comm.*, Vol. 39, pp. 298-303, 1991.
- [62] V. S. Frost, and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Commun. Mag.* Vol. 32, No. 4, pp. 70-81, 1994.
- [63] R. Gusela, "Characterizing the Variability of Arrival Processes with Indexes of Dispersion," *IEEE J. Select. Areas in Commun.*, Vol. 9, No. 2, pp. 203-211, 1991.
- [64] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Network.*, Vol. 2, No. 1, pp. 1-15, 1994.
- [65] P. Jeulenkovic, and A. Lazar, "Multiplexing On-Off Sources with Super exponential On periods: Part 1," *In Proc. INFOCOM 97*, Kobe, Japan, pp. 187-195, 1997.
- [66] J. Medhi, *Stochastic models in Queuing Theory*, New York: Academic, 1991.
- [67] J. Hunter, *Mathematical Techniques for Applied Probabilities in Discrete Time Models: Basic Theory*, Vol. 1, New York,:Academic, 1983
- [68] J. LeBoudec, "An Efficient Solution Method for Markov Models of ATM Links with loss priorities," *IEEE J. Select. Areas Commun.*, Vol. 9, pp. 408-417, 1991.
- [69] C. Blondia, "A Discrete Batch Markov Arrival Process as B-ISDN Traffic Model," *J. Open. Res. Stat. Comput. Sci.*, Vol. 32, No. 3, pp. 3-32, Belgium, 1992.

- [70] M. F. Neuts, "A versatile Markov point process," *Journal Applied Probability*, Vol. 16, pp. 764-779, Dec 1979.
- [71] M. F. Neuts and J. Li, "The Bernoulli Splitting of a Markovian Arrival Process," www.maths.adelaide.edu.au/jli/papers/aplit.pdf, 2002.
- [72] A. T. Anderson, A. Jensen, and B. F. Nielsen, "Modelling and Performance Study of Packet Traffic with Self-similar Characterization over Several Time Scales with MAP," *In Proc. 12th Nordic Teletraffic seminar*, Espoo, Finland, pp. 269-283, 1995.
- [73] O. Hashida, Y. Takayashi, and S. Shimogawa, "Switched Batch Bernoulli Process and The Discrete Time SBBP/G/1 Queue with Application to Statistical Performance," *IEEE J. Select. Areas Commun.*, Vol. 9, pp. 394-401, 1991.
- [74] D. Lucantoni, "New Results on The Single Server Queue with a BMAP," *Stochastic models*, Vol. 7, pp. 1-46, 1991.
- [75] C. Blondia, "A Discrete-time batch Markovian arrival process as B-ISDN Traffic Model," *Belgian J. Oper. Res. Stat. Comput. Sci.* Vol. 32, No. 3, pp. 3-23, 1992.
- [76] A. Allen, "Probability, Statistics and Queuing Theory," New York, Academic, 1978.
- [77] A. Papoulis, *Probability Random Variables and Stochastic Processes*, 3rd Edition, New York: McGraw Hill, 1991.
- [78] B. Q. Li and C.-L. Hwang, "Queue Response to Input Correlation Function: Discrete Spectral Analysis," *IEEE/ACM Trans. Networking*, Vol. 1, No. 5, pp. 522-533, 1993.
- [79] B. Q. Li and C.-L. Hwang, "Queue Response to Input Correlation Function: Continuous Spectral Analysis," *IEEE/ACM Trans. Networking*, Vol. 1, No. 6, pp. 678-692, 1993.
- [80] W. Gardner, "Introduction to Random Processes", 2nd ed., New York: McGraw-Hill, 1989.
- [81] A. V. Oppenheim, and R. Schaffer, *Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1989.
- [82] M. F. Neuts, "Models Based on The Markov Arrival Process," *IEICE Transaction on Communications*, E75B, pp. 1255-1265, 1992.
- [83] M. F. Neuts, "A versatile Markovian Point Process," *J. Appl Prob.* Vol. 14, pp. 764-779, 1979.
- [84] D. Cox and H. Miller, *The Theory of Statistic Processes*. London: Chapman and Hall, 1978.

- [85] D. Cox and P. Lewis, "The Statistical Analysis of Series of Events," *In Methuen's Monographs on Applied Probability and Statistics*, London: Methuen Press, 1996.
- [86] A. Adas, "Traffic models in Broadbands networks," *IEEE Commun. Mag.*, Vol. 35, pp. 82-89, July 1994.
- [87] A. Rueda and W. Kinsner, "A Survey of Traffic Characterization Technique in Telecommunication Networks," *In Proc. IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 96)* Calgary, Alta Canada, pp. 830-833, 1996.
- [88] A. Burger, "Desirable Properties of Traffic Descriptors for ATM Connections in a Broadband ISDN," *In Proc. Discussion Circles, 14th Int. Teletraffic Congress (ITC 14)*, Antibes, France: Elsevier, pp. 233-242, 1994.
- [89] K. Sriram and W. Whitt, "Characterizing the Superposition of Arrival Process in Packet Multiplexers for Voice and Data", *IEEE J. Select Areas Commun.*, Vol. 4, No. 6, pp. 833-846, 1986.
- [90] M. Livny, B. Melamed, and K. Tsolis, "The Impact of autocorrelation on queuing systems," *Manage. Sci.*, Vol. 39, No. 3, pp. 322-339, 1993.
- [91] A. K. Erlang, "The theory of probabilities and telephone conversation," *Nyt Tidsskrift Mathematic B*, vol. 20, pp. 33, 1909.
- [92] F. Pollaczek, "An analytic method for the treatment of queuing problems," *Proceedings of the Symposium on Congestion Theory*, University of North Carolina, Chapel Hill, pp. 1-42, 1964.
- [93] N. T. J. Bailey, "A continuous time treatment of a single queue using generating functions." *Journal of Royal Statistical Society*, Ser. B., Vol. 16, pp. 288-291, 1956.
- [94] L. Takacs, *Introduction to the theory of queues*, New York: Oxford University Press, 1962.
- [95] D. G. Kendall, "Some problems in the theory of queues," *Journal of the Royal Statistical Society*, Sev. B, Vol. 13, pp. 151-185, 1951.
- [96] Q. Ren, and H. Kobayashi, "Diffusion Approximation Analysis of an ATM Statistical Multiplexer with Multiple Types of traffic. Part I: Equilibrium State Solutions," Princeton Univerity, Princeton, NJ 08544, USA, 1993.
- [97] Q. Ren, and H. Kobayashi, "Diffusion Approximation Modeling for Markov Modulated Bursty Traffic and Its Applications to Bandwidth Allocation in ATM Networks," *IEEE J. Select. Areas Comm.*, Vol. 16, No. 5, 1998.

- [98] M. Reiser and H. Kobayashi, "Accuracy of the Diffusion Approximation for some Queuing Systems," *IBM J. Res. Develop.*, March 1974.
- [99] D. Stiliadis and A. Varma, "Latency Rate Servers: A General Model for Analysis of Traffic Scheduling Algorithm," *IEEE/ACM Transactions on Networking*, Vol. 5, No. 5, pp. 611-614, 1998.
- [100] J. Nagle, "On Packet Switches with Finite Storage," *IEEE Trans. Comm.*, Vol. 35, No. 4, pp. 435-438, 1987.
- [101] J. Nagle, "On Packet Switches with Finite Storage," *RFC 970, IETF*, December 1985.
- [102] M. Wurtzler, "Analysis and Simulation of WRED Queues," *MS Project Defense*, Department of Electrical Engineering and Computer Sciences, University of Kansas, Dec. 2003.
- [103] V. B. Iversen, *Teletraffic Engineering and Network Planning*, DTU Course 34340.
- [104] E. Senata, *Non-Negative Matrices and Markov Chains*, 2nd Edition, Berlin: Springer-Verlag, 1981.
- [105] S. Fuhrman and J. LeBoudec, "Burst and Cell Level Models of ATM Buffers," *In Proceedings of the 13th International Teletraffic Congress (ITC13)*, pp. 975-980, 1991.
- [106] D. R. Miller, "Computation of steady-state for M/M/1 priority," *Operation Research*, Vol. 29, pp. 945-958, 1981.
- [107] M. Neuts, *Matrix-Geometric Solutions in Stochastic Model*, Baltimore, MD: Johns Hopkins Univ., 1981.
- [108] M. Neuts, *Structure Stochastic Matrices of M/G/1 Type and their Applications*, New York: Marcel Dekker, 1989.
- [109] J. Abate and W. Whitt, "The Fourier Series Method for Inverting Transforms of Probability Distributions," *Queuing Systems*, Vol. 10, pp. 5-88, 1995.
- [110] A. El Walid and D. Mitra, "Statistical Multiplexing with Loss Priorities in Rate-based Congestion Control of High Speed Networks," *IEEE Trans. Commun.*, Vol. 42, No. 11, pp. 2989-3002, 1994.
- [111] H. Michel, and K. Laevens, "Teletraffic Engineering in a Broad Band Era," *Proceedings of the IEEE*, Vol. 85, No 12, pp. 2024-2025, December 1997.
- [112] F. Yegenoglu and B. Jabbari, "Characterization and Modelling of Aggregate Traffic for Finite Buffer Statistical Multiplexer," *Comput. Networks ISDN Syst.*, Vol. 26, pp. 1169-1185, 1994.

- [113] A. E. Eckberg, Jr, "Generalized peakedness of teletraffic processes", *10th Int. Teletraffic Congress*, paper 4.4B-3, Montreal, 1983.
- [114] A. E. Eckberg, Jr, "Approximation for bursty (and smoothed) Arrival Queuing delays based on generalized peakedness.", *11th Int. Teletraffic Congress*, paper 3.1A-4-1, Kyoto, Japan, 1985.
- [115] K. Chandra, "Statistical Multiplexing," *Wiley Encyclopedia of Telecommunications*, John Wiley and Sons, 2003.
- [116] A. A. Fredericks, "Congestion in blocking system: A simple approximation technique," *Bell Syst. Tech. J.*, Vol. 59, pp. 805-827, 1980.
- [117] K. Sriram P. K. Varshiney and J. G. Shanthikumar, "Discrete-Time Analysis of Integrated Voice-Data Multiplexers with and without Speech Activity Detector," *IEEE J. Select. Areas Commun.*, Vol. SAC-1, Special Issue on Packet Switched Voice and Data Communication, Dec. 1983.
- [118] D. R. Cox, "The analysis of non-markovian stochastic processes by the inclusion of supplementary variables," *Proceedings of Cambridge Philosophy Society*, Vol. 51, pp. 441-443, 1955.
- [119] W. Whitt, "Approximating a Point Process by a Renewal Process: Two Basic Methods." *Oper. Res.*, Vol. 30, No. 1, pp. 125-147, January-February 1982.
- [120] B. Reid, *Elements of the theory of Markov Birth-death Process and their application*, Dover Publication, Inc Mineola, New York, A. T., 1988.
- [121] P. Nain, "Basic Elements of Queuing Theory. Application to the Modelling of Computer Systems," *Lecture Notes*, INRA, 06902 Sophia Antipolis, France.
- [122] F. B. Nielsen, "Queuing systems: Modelling, Analysis and Simulations," *Research Report 259: Dpt. Inf., Univ. Oslo*, 1988.
- [123] I. Adam, O. Boxma, and D. Perry, "The G/M/1 Queue revisited," *Mathematical methods of Operation Research*, Vol. 62, pp. 437-452, 2005.
- [124] D. G. Kendall, "Stochastic Processes Occurring in the Theory of Queue and their Analysis by the Method of Embedded Markov Chain," *Ann. Math. Stat.* Vol. 24, pp. 338-354, 1953.
- [125] L. Takacs, *Introduction to the Theory of Queue*, New York: Oxford University Press, 1962.
- [126] K. Bose, *The G/M/1, G/G/1, G/G/m, M/G/m/m queues*, Samjay, 2002.
- [127] L. Kleinrock, *Queuing Systems, Volume 1: Theory*, Wiley, 1975.

- [128] J. C. S. Lui, "G/G/1 Queue Systems," *Computer systems Performance Evaluation, lecture notes*, Department of Computer Sciences, The Chinese University of Hong Kong.
- [129] D. Sloughter, "Liouville's Theorem," Furma University, Mathematics Vol.39, 2004.
- [130] H. Kobayashi, "Bounds for Waiting Time in Queuing Systems," Computer Science Department, *IBM Thomas J. Watson Research Centre*, Yorktown Heights, New York 10598.
- [131] S. L. Brumell, "Some Inequalities for Parallel Server Queues," *Operation Research*, Vol. 19, pp. 402-413, 1971.
- [132] M. M. Marjanovic, and Z. Kadelburg, "A proof of Chebyshev's inequality," *The teaching of mathematic.*" Vol. 10, No. 2, pp. 107-108, 2007.
- [133] F. Feller, *An Introduction in probability theory and its applications, Vol. 1*, John Wiley and Sons, New York, 1957.
- [134] J. F. C. Kingman, "On Queue in Heavy Traffic," *Journal of the Royal Statistical Society, Ser. B*, Vol. 24, pp. 383-392, 1962.
- [135] J. F. C. Kingman, "Some Inequalities for the Queue GI/G/1," *Biometriks*, Vol. 49, pp.315-324, 1962.
- [136] K. T. Marshall, "Some Inequalities in the Queuing," *Operation Research*, Vol. 16, pp. 651-665, 1968.
- [137] R. Cruz, "A Calculus for Network delay, part I: Network Elements in Isolation," *IEEE Trans. Infos. Theory*, Vol. 37, pp. 1034-1056, 1995.
- [138] I. Adam, and J. Resing, *Queuing Theory*, Department of Mathematics and Computer Sciences, Eindhoven University of Technology: The Netherlands, 2002.
- [139] J. D. C. Little, "A proof for queuing formula $L=\lambda W$," *Operational Research*, Vol. 9, pp. 383-387, 1961.
- [140] H. Zhan, and D. Shi, "Explicit solution for M/M/1 Preemptive queue," *International Journal of Information and Management Sciences*, 2010.
- [141] P. Gevros, "Congestion control mechanism and the best effort service model," *IEEE Netw.*, Vol. 15, No. 3, pp. 16-26, 2001.
- [142] K. Chandrayana, B. Sakdar, and S. Kalyanaraman, "Scalable configuration of RED queue parameters," *IEEE Workshop on High Performance Switching and Routing*, pp. 185-189, 2001.

- [143] J. M. Pitts, X. Wang, Q. Yang, and J. A. Schormans, "Excess-rate queuing theory for M/M/1/RED with application to VoIP," *Electronics letters*, Vol. 42, No. 20, Sept. 2006.
- [144] Y. Berra, "Steady-State M/M/1-Type Loops," Section 2.1, chapter 1: M/M/1 Queue, Lecture Notes.
- [145] W. Whitt, "Approximating the Arrival Process in a G/M/1/ ∞ System," *In Management Sciences*, Vol. 27, No. 6, pp. 621-628, June 1981.
- [146] R. B. Cooper, *Introduction to Queuing Theory*. McMillan, New York, 1972
- [147] A. Kuczura, "The Interrupted Poisson Process as an Overflow Process," *Bell System Tech. J.*, Vol. 52, pp. 437-448, 1973.
- [148] B. Wallstrom, "Congestion Studies in Telephone Systems with Overflow Facilities," *Ericson Technics*, Vol. 3, pp. 189-351, 1966.
- [149] R. I. Wilkinson, "Theories of Toll Traffic Engineering in the U.S.A," *Bell System Tech. J.*, Vol. 35, pp. 421-514, 1956.
- [150] D. L. Iglehart, "Functional limit theorem for the queue GI/G/1 in light traffic," *Advanced Applied Probability*, Vol. 3, pp. 269-281, 1971.
- [151] P. J. Kuehn, "Approximate Analysis of General Queuing Networks by Decomposition," *IEEE Trans. Comm.*, Vol. COM-27, pp. 113-126, 1978.
- [152] P. M. Morse, *Queues, Inventories and Maintenance*, Wiley, New York, 1958.
- [153] B. D. Choi, B. C. Shin, K. B. Choi, D. H. Han, and J. S. Jang, "Priority queue with two-state Markov modulated arrival," *Comm. IEEE Proceedings*, Vol. 145, pp. 152-158, 1996.
- [154] J. Abate, and W. Whitt, "Calculating Time Dependence Performance measures for the M/M/1 Queue," *IEEE Communication theory*, 1999.
- [155] P. Romano, B. Ciciani, A. Santoro, and F. Quaglia, "Fast computation of Hyper-exponential Approximation of Response Time Distribution of MMPP/M/1 Queues," 41st *IEEE annual symposium*, 2008.
- [156] P. T. Brady, "A statistical Analysis of On-Off Patterns in 16 Conversations," *Bell Syst. Tech. J.*, Vol. 47, No. 1, pp. 73-91, Jan. 1968.
- [157] B. Sinclair, "The M/G/1 Queue," *Connexions Module: m10819*, Version 2:4, June 2005.
- [158] R. B. Cooper, *Introduction to Queuing Theory*, 3rd Edition, CEEPress Books, 1990.

- [159] G. B. Erasmus, "Confidence Limits for Expected Waiting times of two Queuing Models," *Orion*, Vol. 20, No. 1, 2004.
- [160] 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*, Vol. SAC-4, No. 6, 1986.
- [161] S. Shah-Heydari, and T. Le-Ngoc, "MMPP Modelling of Aggregated ATM Traffic," Concordia University, Montreal.
- [162] M. Schwartz, *Telecommunication Networks*, Addison-Wesley, 1986.
- [163] M. L. Luhanga, *Fluid Approximation of an Integrated Packet Voice and Data Multiplexer*, Ph. D Thesis, Columbia University, New York, 1984.
- [164] R. D. Cox, and H. D. Miller, *The Theory of Stochastic Processes*, John Wiley and Sons, Inc., New York, 1965.
- [165] G. F. Newell, *Applications of Queuing Theory*, Chapman and Hall, Ltd., London, 1971.
- [166] D. P. Gaver, and G. S. Shelder, "Approximate Models for Processors Utilization in Multi programmed Computers Systems," *SIAM J. Computing*, Vol. 2, No. 3, pp. 183-192, 1973.
- [167] S. Robert and J.-Y. Leboudec, "On Markov Modulated Chain Exhibiting Self-similar over Finite Time-scale," *In Proc. Performance, 96, Performance Evaluation*, vols. 27/28, pp. 159-173, 1996.
- [168] H. Kobayashi, "Applications of the Diffusion Approximation to Queuing Networks: Part I and II," *IBM Research Reports RC 3943, and RC 4054*, New York 1972
- [169] Y. Nonaka, and S. Nogami, "Evaluation of the Diffusion Approximation for G/G/1 Queuing Model," *IEICE Tech. Rep.*, IN2009-93, Vol. 109, No. 327, pp. 35-40, Dec. 2009.
- [170] J. W. Roberts, "Performance Evaluation and Design of Multiservice Networks," *COST 224*, pp. 108-110, 1991.
- [171] J. M. Fernandez, "The Relationship between a MMPP and a MMRP," *Network and System Modeling Group*, 1992.
- [172] B. Patrick, *Probability and Measure*, 3rd Edition, John Wiley and Sons, 1995.
- [173] B. Do Chuong, "The Multivariate Gaussian distribution," www.cs229.stanford.edu/section/gaussians.pdf, October 2008.

- [174] F. Feller, *Introduction to Probability Theory and its Applications, Vol. 2*, John Wiley & Sons, New York, 1971.
- [175] F. Hans, "A History of the Central Limit Theorem: From Classical to Modern Probability Theory," *Sources and Studies in the History of Mathematics and Physical Sciences*, New York, Springer, 2011.
- [176] K. Tsoukatos and A. Makowski, "Heavy Traffic Analysis for a Multiplexer Driven by M/G/ ∞ Input Process," in *Proc. ITC15*, Washington DC, pp. 497-506, 1997.
- [177] P. Carlson, and M. Fiedler, "Multifractal Products of Stochastic Processes: Fluid Flow Analysis," In *Proc. NTS-15*, Lund, 2000.
- [178] J. M. Harrison, and A. Zeevi, *Dynamic Scheduling of a Multi Class Queue in the Halfin-Whitt Heavy Traffic*. Revised July, 2002, March 2003.
- [179] D. R. Cox, and H. D. Miller, *Theory of Stochastic Process*, Methuen, 1965.
- [180] Y. Sahalia, and R. Kimmer "Maximum Likelihood Estimation of Stochastic Volatility Model," *Financial Economic*, No. 83, pp. 413-452, 2007.
- [181] Y. Sahalia, "Maximum Likelihood Estimation of Discrete Sampled Diffusion: A closed form Approximation Approach," *Econometrica*, No. 70, pp. 223-261, 2002.
- [182] S. Cyganowski, P. Kloeden, and J. Ombach, *Elementary Probability to Stochastic Differential Equation with MAPPLE*, Springer, Berlin, 2001.
- [183] A. Friedman, *Partial Differential equation of Parabolic Type*, Prentice-Hall, Englewood Cliffs.
- [184] Y. A. Sahalia, "Closed Form Likelihood Expansions for Multi-variate Diffusion," *The Anals of Statistics*, Vol. 36, No. 2, pp. 906-937, 2008.
- [185] A. T. Anderson, A. Jensen, and B. F. Nielsen, "Modelling and Performance Study of Packet-Traffic with Self-Similar Characteristics over several Time-scales with Markovian arrival Process," In *Proc. 12th Nordic Teletraffic Seminar*, pp. 269-283, Finland, 1995.
- [186] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "Statistical Analysis and Stochastic Modelling of Self-Similar Data Traffic," In *Proc. ITC14*, pp. 319-328, France, 1994.
- [187] A. Eramilli, J. J. Gordon, and W. Willinger, "Application of Fractals in Engineering for Realistic Traffic Processes," In *Proc. ITC14*, pp. 35-44, France, 1994.

[188] R. W. Wolf, "Poisson Arrival times Averages," *Operation Research*, Vol. 30, pp. 223-231, 1982.

[189] J. Ried, and A. R. Ward, "An Approximate Analysis of a buffered CSMA/CD," *IEEE Trans. on Comm.*, Vol. 36, pp. 932-941, 1988.



APPENDIX A

STATISTICAL MULTIPLEXING

A-1: Self-Similar and Long-range dependence process

Assume X_k to be a stationary process with mean $E[X_k] = \bar{X}$ and autocorrelation function [185]

$$r_i = \frac{C(i)}{\text{Var}[X_k]}.$$

Note that this is not a standard definition.

Consider the processes $\{X_k^{(m)}\}$, $m = 0, 1, 2, \dots$ which are constructed out of X_k [184] as

$$X_k^{(m)} = \sum_{n=0}^{m-1} \frac{X_{kn+n}}{m},$$

that is the m -order moment over the non overlapping blocks of size m .

The processes $\{X_k^{(m)}\}$ are also stationary with mean \bar{X} and autocorrelation function r_i^m .

The process X_k is called asymptotically second order self-similar if for $m, i \rightarrow \infty$.

All-order self-similar is present when, for all m , the process $mX_k^{(m)}$ is statistically identical to the process $m^H X_k$, showing that both processes have same multivariate distributions as proved by Anderson [72].

Self-Similar is defined in a multitude of equivalent ways [186]:

- Slowly decaying variances: $\text{Var}[X_k^{(m)}] \approx m^{-\beta}$ as $m \rightarrow \infty$, where $0 < \beta < 1$.
- Slowly decaying autocorrelation function $r(i) \approx i^{-\beta}$ as $i \rightarrow \infty$.
- The power spectral density $S(f)$ that behaves like that of $1/f$ noise around the origin: $S(f) \approx f^{-(1-\beta)}$ as $f \rightarrow 0$.

Self-Similarity also implies long-range dependence [186], that means $\sum_{-\infty}^{+\infty} r(i) = \infty$, with both converging to Hurst parameter $H = 1 - \beta/2$ [187].

A process characterized by $\sum_{-\infty}^{+\infty} r(i) < \infty$ is said to be short-range dependent. It differs from a long-range dependent process in the sense that:

- The variance $Var[X_k^{(m)}] \approx m^{-1}$;
- The autocorrelation $r(i) \approx \gamma^{-1}$ for large i decays exponentially fast;
- The power spectral density is finite (almost constant) around the origin;
- The process analysis can be approximated as a second-order pure noise for large m ($r^{(m)}(i) \rightarrow 0, i \neq 0, m \rightarrow \infty$).

A-2: Matrix geometric solution

Miller approach [106] which exploits Neuts [107] matrix-geometric solution is also a good approach of queue model. But, let us outline some results of the development to show its geometric aspect:

- The matrix-geometric method

The system under consideration is a Markov process, the state space of which consists of the boundary states $(0, j)$, where $j = 0, 1, 2, \dots, n$ and a semi-infinite range of states (i, j) where $i = 1, 2, \dots, \infty$ and $j = 0, 1, 2, \dots, m$. The states are ordered lexicographically in the following manner: $(0, 0), (0, 1), \dots, (0, n); (1, 0), (1, 1), \dots, (1, m); (2, 0), (2, 1), \dots, (2, m); \dots$

The set of boundary states $\{(0, 0), (0, 1), \dots, (0, m)\}$ is called *level 0* and the set of states $\{(i, 0), (i, 1), \dots, (i, m)\}, i \geq 1$ is called *level i*. The state space is partitioned according to these levels and by grouping the entries according to the number of customers we get the generator matrix [111] in the form:

$$Q = \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & \dots \\ B_{10} & B_{11} & A_0 & 0 & \dots \\ B_{20} & A_2 & A_1 & A_0 & \dots \\ B_{30} & A_3 & A_2 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

where the matrices characteristics: B_{00} a $(n+1) \times (n+1)$ dimensions; B_{01} $(n+1) \times (m+1)$ dimensions B_{i0} , $i \geq 1$, a $(m+1) \times (n+1)$ dimension and B_{11} , A_i , $i \geq 0$ are squared of $m+1$ dimensions.

Define the generator $A = \sum_{i=0}^{\infty} A_i$, which is a generator describing the Markov process Q in the

Orthogonal direction only. Assume an irreducible Markov process with a one transition class matrix generator A . For the stability, the Markov process Q is ergodic if and only if:

$$\pi A_0 e = \pi \sum_{i=2}^{\infty} (i-1) A_i e,$$

where $e = (1,1,\dots,1)'$ and $\pi = (\pi_0, \pi_1, \dots, \pi_m)$ is the vector probability at equilibrium of the Markov process with generator A . This implies: $\pi A = 0$ and $\pi e = 1$.

In the matrix form, these equilibrium probabilities can be written as $p_i = \{(p(i,0), p(i,1), \dots, p(i,m))\}$. Provided the geometric form obtained in the scalar development, we can write: $p_i = p_1 R^{i-1}$, $i = 1, 2, \dots$ where the matrix R is the minimal non-negative solution of the equation $\sum_{i=0}^{\infty} R^i A_i = 0$.

This equation can be written by substitution of the equilibrium probabilities as follows:

$$\sum_{k=0}^{\infty} p_{i+(k-1)} A_k = 0 \text{ gives } p_{i-1} \sum p_k A_k = 0 \text{ with } p_{i-1} > 0.$$

The spectral radius of the matrix R is less than 1 so that $I - R$ is invertible.

The boundary equation for p_0 and p_1 are the same as in the M/M/1 type model above, the only difference is the replacement of the matrix quadratic equation for R . With the normalization of the probability, we can write:

$$R = -(A_0 + \sum_{k=2}^{\infty} R^k A_k) A_1^{-1}.$$

The resolution of this equation necessitates an approximation of the infinite sum to a constant value, for instance K , and then by successive substitutions in the equation

$$R_{l+1} = -(A_0 + \sum_{k=2}^K R_l^k A_k) A_1^{-1}, \text{ compute the approximation.}$$

It is important to note the existence of another solution method of queuing model called "spectral expansion method". It is computational difficult when the system becomes large, therefore we will not analyse it here. Instead, we will try by example of a simple queue system to see how it is computed.

Basically, with the same respect of Markov process, the equilibrium probability is given by:

$$p_i = \sum_{k=0}^m c_k y_k x^{i-1}, \quad i = 1, 2, \dots,$$

where x_0, x_1, \dots, x_m are the roots of the unit circle of

$$\det \left(\sum_{k=0}^{\infty} x^k A_k \right) = 0.$$

The vector $y_k, k = 0, 1, 2, \dots, m$ is a non-zero solution of the equation:

$$y \sum_{k=0}^{\infty} x^k A_k = 0.$$

As in the M/M/1 vector solution, the mean number of jobs in the system is:

$$E[N] = \sum_{i=1}^{\infty} i p_i e = \sum_{i=1}^{\infty} i p_0 R^i e = p_0 R (I - R)^{-2} e,$$

and applying Little's theorem [139] we can determine the waiting time:

$$W = E[N] \lambda^{-1}.$$

A-3: M/M/1 queue with priority

In their development, Zhan and Shi [140] choose to prioritize traffic class of type 1 over type 2 and takes into account two priority rules: preemptive-resume priority and non-preemptive priority.

– Preemptive-resume priority queue

The preemptive resume priority hands absolute priority to a particular class of traffic. As chosen, if type 1 has absolute priority over type 2 it means that when type 1 arrives while type 2 is in service, the server interrupts the latter to process type 1 service and resumes the type 2 traffic at its interrupted point once there is no longer type 1 traffic.

Let the random variables N_i and S_i be the number and sojourn time of type i packets in the system.

According to the rule, type 2 does not exist for type 1, therefore; type 1 packets follow the results obtained for M/M/1 standard:

$$E[S_1] = \frac{1/\mu}{1-\rho_1}, \quad \text{and} \quad E[N_1] = \frac{\rho_1}{1-\rho_1}.$$

Since the service time is exponentially distributed with a constant mean, the total number of packets in the queue is independent of their service discipline. Thus, this number is the same as if they were served in order:

$$E[N] = E[N_1] + E[N_2] = \frac{\rho}{1-\rho} = \frac{\rho_1 + \rho_2}{1-(\rho_1 + \rho_2)},$$

from where one can derive type 2 number of packets by:

$$E[N_2] = \frac{\rho_1 + \rho_2}{1-\rho_1-\rho_2} - \frac{\rho_1}{1-\rho_1} = \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)},$$

and by applying Little's theorem, one get the mean sojourn time:

$$E[S_2] = \frac{E[N_2]}{\lambda_2} = \frac{1/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

– Non-preemptive priority

In the non-preemptive priority, the nearly “absolute priority” is handed to a particular class of traffic. In this case for instance, type 1 has the nearly absolute priority. Nearly absolute priority means that the service of type 2 is not interrupted at the time type 1 packets arrive. Type 1 has to wait until type 2 service is completed. Provided the exponential distribution of the service time, type 1 packets experience as mean sojourn time:

$$E[S_1] = E[N_1] \frac{1}{\mu} + \frac{1}{\mu} + \rho_2 \frac{1}{\mu},$$

where the last term expresses the time type 1 will wait until type 2 service completes. According to PASTA property [188], the fraction of time type 2 packets are served, ρ_2 , is the probability that type 1 finds a type 2 packets in service. According to the relation $E[N_1] = \lambda_1 E[S_1]$, we obtain:

$$E[S_1] = \frac{(1+\rho_2)/\mu}{1-\rho_1}, \quad \text{and} \quad E[N_1] = \frac{(1+\rho_2)\rho_1}{1-\rho_1}.$$

As stated earlier, the number of packets in the system is independent of the order they are served, then:

$$E[N] = E[N_1] + E[N_2] = \frac{\rho_1 + \rho_2}{1 - (\rho_1 + \rho_2)} .$$

Therefore the mean number and mean sojourn time of type 2 packets in the system is given by:

$$E[N_2] = \frac{(1 - \rho_1(1 - \rho_1 - \rho_2))/\rho_2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} \text{ and } E[S_2] = \frac{(1 - \rho_1(1 - \rho_1 - \rho_2))/\mu}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} .$$

A-4: M/M/1/N queue model

The M/M/1/N queue is the same as M/M/1 queue with the difference that $n = N$. This type of queue is also referred to as a blocking model because of the finite value of n .

The steady state equations for the mode are defined as:

$$-\rho P_0 + P_1 = 0, \quad n = 0,$$

$$\rho P_{n-1} - (\rho + 1)P_n + P_{n+1} = 0, \quad n = 1, 2, 3, \dots, N-1,$$

$$\rho P_{n-1} - P_N = 0, \quad n = N.$$

The density function of the queue length is now given by:

$$P_n = \frac{1 - \rho}{1 - \rho^{N+1}} \rho^n \text{ with } P_0 = \frac{1 - \rho}{1 - \rho^{N+1}} .$$

The expected number of entities in the system is reduced to:

$$E[N_s] = \frac{\rho[1 - (N+1)\rho^N + N\rho^{N+1}]}{(1 - \rho)(1 - \rho^{N+1})} .$$

Since the queue length is limited, we might experience losses which are avoided by introducing the effective arrival rate [144] denoted λ_{eff} :

$$\lambda_{eff} = \lambda(1 - P_N) .$$

The expected number of packets in the queue is therefore:

$$E[Q_N] = E[N_s] - \frac{\lambda_{eff}}{\mu} = \frac{\rho^2[1 - N\rho^{N-1} + (N-1)\rho^N]}{(1 - \rho)(1 - \rho^{N+1})} .$$

Finally, the expected waiting time per packet in the queue is derived as previously:

$$E[W_N] = \frac{E[Q_N]}{\lambda_{eff}} = \frac{\lambda[(\mu^N - \lambda^N) - N\lambda^{N-1}(\mu - \lambda)]}{\mu(\mu - \lambda)(\mu^N - \lambda^N)}.$$

A-5: MMPP/G/1 queue

It is assumed that the source generates traffic to a single-server queue with general service time distribution $F_S(\cdot)$ with moments $m_i, i \geq 1$ [160]. Let the process $\{\tau_n, n \geq 0\}$ be the successive departure epochs. Let X_n be the number of packets in the system and j_n the phase of the Markov process at the departure epoch. The process $\{X_n, j_n, \tau_{n+1} - \tau_n : n \geq 0\}$ is a Semi-Markov process on the state space $N \times M$.

Define the conditional probability $P_{ij}(n, t)$ of n arrivals under the Markov process starting in state i to end in state j during $[0, t]$ time interval.

This can be interpreted in the system at the departure instants:

$$A(x) = \int_0^x P(n, t) dF_S(t), \quad n \geq 0, \quad t \geq 0.$$

Define also the probability that the first packet reaches an empty system at time $t \leq x$, one can write $U(x) = \int_0^x e^{-(R-\Lambda)t} \Lambda dt$. Then as in the previous matrix determination, after having defined $B(x) = U(x) * A(x)$. This gives rise to the transition probability matrix:

$$Q(x) = \begin{pmatrix} B_0(x) & B_1(x) & B_2(x) & \cdots \\ A_0(x) & A_1(x) & A_2(x) & \cdots \\ 0 & A_0(x) & A_1(x) & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

The solution required the system to be stationary in terms of the number of packets and the phase of the MMPP considered at an embedded arbitrary departure time. In particular, this stationary system which is such that a departure leaves the system empty with the MMPP being in state j can be represented as $x(0, j)$. This represents a recursive system with returns to $(0, j)$. Since the probability here is just the complementary of the expected number of transition in the chain with transition matrix $Q = \lim_{x \rightarrow \infty} Q(x)$ between the returns to $(0, j)$, it turns out from this structure that the system may move up many levels and down only one

level at a time. Thus, the mean number of jumps from $i+1$ to i , which are independent of i , is the base in obtaining the mean recurrence time to level 0. Let study this structure.

Define the probability matrix $G(n) = \{G_{jk}(n)\}$ of the embedded Markov Chain starting from $(i+1, j)$, reaches the level i for the first time in state (i, k) (level i) in exactly n jumps. This first passage gives rise to the moment generating function given by the discrete Laplace-Stieltjes Transform:

$$G(z) = \sum_{n=0}^{\infty} z^n G(n)$$

which must satisfy the following condition: according to the structure of Q and the fact that the stability of the system is obtained if and only if the matrix $G(1) = G$ is stochastic: starting in $(i+1)$, i will be surely reached (probability equal one). This is the case if the intensity of traffic is such that $\rho = \pi\lambda m_1 < 1$, where $\lambda = \Lambda e$ and m_1 is the mean service time. Thus:

$$G(z) = z \sum_{n=0}^{\infty} A_n G z^n, \text{ where } A_n = \lim_{x \rightarrow \infty} A_n(x).$$

If g is the equilibrium value of the vector G , this implies $gG = g$ and $ge = 1$. Also, let the vector T whose elements $\{t_j\}$ are the number of jumps from $(i+1, j)$ to level i :

$$T = \left(\frac{dG}{dz} \Big|_{z=1} \right) e = (I - G + eg)[I - A + eg - Kg]^{-1} e,$$

where $A = \sum_{n=0}^{\infty} A_n = \int_0^{\infty} e^{Rt} dF_S(t)$ and $K = \sum_{n=0}^{\infty} nA_n e = \rho e + (R + e\pi)^{-1}(A - I)\lambda$.

Define the matrix $L(n) = \{L_{ij}(n)\}$ indexed by its entries L considered as the probability starting in level $(0, i)$, the first return to level 0 is in state $(0, j)$ after exactly n jumps of the embedded Markov Chain. The matrix generating function of the discrete Laplace-Stieltjes $L(z) = \sum_0^{\infty} z^n L(n)$ is given by:

$$L(z) = z \sum_0^{\infty} B_n G(z)^n = UG(z) \text{ where } B_n = \lim_{x \rightarrow \infty} B_n(x) \text{ and } U = (\Lambda - R)^{-1} \Lambda.$$

As previously, the stability is obtained if and only if the matrix $L = L(1)$ is stochastic: that is the return to level 0 is surely reached (probability 1). Define then a stationary vector l of the matrix UG such that $lUG = l$ and $le = 1$.

Let $O = \{o_j\}$ be the vector of the number of jumps of the process just before the first return to any state in level 0, starting in state $(0, j)$, $O = \{o_j\}$ is given by:

$$O = \frac{dL(z)}{dz} \Big|_{z=1}, \quad e = UT.$$

Now, by observing the MMPP only at returns to level 0 with related modulator matrix Q , and keeping track of the number of jumps of such process, which are discrete time jumps, the successive state $(0, j)$ visited, and the time interval between returns to level 0 is a Markov renewal process with transition probability function $L(n)$.

From the Markov renewal theory [174], the mean recurrence time of state $(0, j)$ in such a process is given by $E[\tau_{n+1} - \tau_n] = lOo_j^{-1}$.

But this mean recurrence time the expected number of jumps in the Markov Chain between successive returns to $(0, j)$ and therefore the vector $x_0 = x(0, j) = l(UT)^{-1}$, giving us the returns to level 0.

Considering the equilibrium probability $y(0, x)_{ij}$ at state $(0, j)$ at an arbitrary point of time.

Let $R_{kl}(x)$ be the expected number of visits to state (k, l) within interval $[0, x]$ of the Markov renewal process $Q(x)$ given that it started in state (i, j) at time $t=0$. One can write by conditioning the departure before t as:

$$y(0, l, t) = \Pr[X(t) = 0, J(t) = l \mid X(0) = i, J(0) = j] = \sum_{k=1}^m \int_0^t dR_{0k}^{ij} [e^{(R-\Lambda)(t-u)}]_{kl}.$$

By applying the renewal theorem [174]:

$$y(0, l) = \lim_{t \rightarrow \infty} y(0, j, t)_{ij} = \sum_{k=1}^m \frac{1}{m(0, k)} \left[\int_0^\infty e^{(R-\Lambda)t} dt \right]_{kl},$$

where $m(0, k)$ is the mean recurrence time of $(0, k)$ in the process $Q(x)$.

By respecting the continuous time condition between returns to level 0, one can determine the expression of $m(0, k)$, then in turn obtain the expression of the vector $y_0 = (\pi\lambda)x_0(\Lambda - R)^{-1}$, which is related to the traffic intensity by $y_0e = 1 - \rho$.

Define the virtual waiting time $V(t)$ as the time that elapses from the time a packet arrives in the system till he enters the service. The conditioned probability on the last departure before time t is given by [160]:

$$\begin{aligned} & \Pr[0 < V(t) < x, J(t) = j \mid X(0) = i, J(0) = l] \\ &= \sum_{k=1}^m \sum_{v_1=1}^{\infty} dR_{v_1 k}^{il}(\tau) \sum_{v_2=0}^{\infty} P_{kj}(v_2, t - \tau) \cdot \int_{w=0}^x dF_S(t + w - \tau) F_S^{(v_1+v_2-1)}(x - w) \\ &+ \sum_{k=1}^m \int_{\tau=0}^t dR_{0k}^{il}(\tau) \int_{u=0}^{t-\tau} \sum_{p=1}^m [e^{(R-\Lambda)u} \Lambda du]_{kp} \cdot \sum_{v_2=0}^{\infty} P_{pj}(v_2, t - \tau - u) \cdot \int_{w=0}^x dF_S(t + w - \tau - u) F_S^{v_2}(x - w), \end{aligned}$$

where F_S^n is the n -fold convolution of the service time distribution with itself and the conditional probability obtained as follows:

The last departure at time τ occurred before time t and left $v_1 \geq 1$ packets in the system. Within that time interval, there were $v_2 \geq 0$ arrivals so that at time t there are $v_1 + v_2$ packets in the system. The packet that entered the service at τ got served at time $t + w$; then for a packet that arrived at time t and entered the service before time $t + x$, $x > w$, there must be $v_1 + v_2 - 1$ departures in the interval time $x - w$.

In the second term, the last departure leaves the system empty, but there is a need to keep track of the MMPP phase during the idle period.

Consider the vector $W(x)$ whose components $W_j(x)$ are given by:

$$W_j(x) = \lim_{t \rightarrow \infty} \Pr[V(t) \leq x, J(t) = j \mid X(0) = i, J(0) = l].$$

Denote by $W(s)$ the Laplace-Stieltjes transform of $W(x)$, by applying the renewal theorem we get the virtual waiting time of the MMPP/G/1 given by [160]:

$$W(s) = \begin{cases} sy_0 [sI + R - \Lambda F_S(s)]^{-1}, & s > 0, \\ \pi, & s = 0. \end{cases}$$

A-6: Fluid method solution: determination of the initial vector

Luhanga [163] proposed the solution of the differential equation (5.) and it is given by:

$$p_k(x_1) = \pi_k + \sum_{n=0}^{N-(n_1+1)} K_n \{\phi_n\}_k e^{S_n x_1}, \quad k = 0, 1, 2, \dots, N,$$

where S_n is the n -th negative eigenvalue of the matrix MD^{-1} , ϕ_n is the right eigenvector of the matrix MD^{-1} and $\{\phi_n\}_k$ is the k -th of the eigenvector, K_n are the normalized constraints which are to be determined, n_1 is the number of voice packets corresponding to C_1 , and π_k is the probability of having k sources in talk spurts in the queue.

Proceeding in the same manner for the forward Kolmogorov's equation with the only difference set for the rate of changes of the buffer. The assumption made here are the following:

A proportion of the transmission capacity reserved for the exclusive use of data is $C_2 = \gamma C$. Data packets may also use the transmission capacity reserved for voice packets if there is no voice traffic. The data packets traffic drifts in the queue r_k^d are:

$$r_k^d = \begin{cases} \lambda_2 - \max(C_2, C - R_k), & k \neq 0, n_1 = 0, \\ \lambda_2 - C, & k = 0, n_1 = 0. \end{cases}$$

The conditional probability that $x_1 = 0$ given k voice packets in talk spurts are obtained from the solution of the differential equation as follows:

$$\Pr[x_1 = 0, N = k] = 1 + \frac{1}{\pi_k} \sum_{n=0}^{N-(n_1+1)} K_n \{\phi_n\}_k = \tau_k.$$

So r_k^d becomes $r_k^d = [\lambda_2 - \max\{C_2, C - R_k\}] \cdot \Pr[x_1 = 0, N = k] + [\lambda_2 - C_2] \cdot (1 - \Pr[x_1 = 0, N = k])$.

The elements for the diagonal rate matrix for the data packets in the matrix differential equation are obtained from r_k^d .

The stability of the system in such a case, according to Lindley's theorem, invokes [128]:

$$\sum_{k=0}^N \pi_k r_k^d < 0.$$

The number of negative eigenvalues n_e of the matrix MD^{-1} for data packets traffic is equal to the number of values k satisfying $r_k^d > 0$.

Let I_{-i} be a set of the number of active voice sources k , such that $k = 0, 1, \dots, I_0 - 1$.

Thus, the number of negative eigenvalues is given therefore by $n_e = N - I_0' + 1$, where

$$I_0' = \sup \left[-\frac{\lambda_2}{\tau_k} + \left(\frac{1}{\tau_k} - 1 \right) C_2 \right] \text{ and } \lceil \cdot \rceil \text{ is the smallest integer greater than its argument.}$$

The solution of the matrix equation for the data packet traffic is given by:

$$p_k(x_2) = \pi_k + \sum_{n=0}^{n_e-1} K_n^d \{ \phi_n^d \}_k e^{S_n^d x_2},$$

the parameters of this solution are similar to the one of the voice solution with the subscript of d to mean data.

The use of the fluid approximation ensures that the probability of data packet traffic in the system, given that the number of active voice sources is greater than or equal to I_0' , is zero since the number of active voices k for which $k > I_0'$ represents that data packet traffic.

Thus, $\pi_k = - \sum_{n=0}^{n_e-1} K_n^d \{ \phi_n^d \}_k e^{S_n^d x_2}$. This equation represents n_e simultaneous equations to solve for the unknown K_n^d , the normalization constraint constant. If we assumed an infinite waiting room for the queue, the data packet probability at its equilibrium is set to the asymptotical value, leading to the data queue length value:

$$Q_d = \sum_0^{n_e-1} \frac{K_n^d}{S_n^d} (e \cdot \phi_n^d),$$

with $e = [1, 1, \dots, 1]'$ and ϕ_n^d is the column eigenvector.

A-7: Bit estimation of the Ethernet traffic

The most important thing is that the Ethernet frame has a fixed timestamp which is used as reference to evaluate bit error. We are interested in the magnitude of that error which we want to be as small as possible.

The percentage of error in bit rate estimations is given by the ratio timestamp accuracy and time sample as follows:

$$\%(Error) = \frac{T_A}{T_s},$$

where T_A is the timestamp accuracy obtained from measures and the sample time interval T_s determined by the base scale.

This ratio gives us the size or the magnitude of the error. The timestamp accuracy specifies how much a frame can be shifted with respect to the standard timestamp in the link layer frame. A frame can be shifted at most T_A seconds.

In the worst case, this can result in $T_A \times C$ bits to be placed in the incorrect interval. Thus, a rough estimate of the error is given by $\%Error = \frac{T_A C}{T_s C}$.

The table below standardizes the values of error given in percentage based on some typical timestamp accuracies [177]. The error sets the lower limit on the timescales that can be used. For most of measurements software, 1 ms is commonly used.

Table A.1: Bit error

T_s	$T_A=100$ ns	$T_A=10$ μ s	$T_A=1$ ms	$T_A=10$ ms	$T_A=1$ s
1 s	10^{-5}	10^{-3}	10^{-1}	1	10^2
1 ms	10^{-2}	1	10^2	10^3	10^5
1 μ s	10	1000	10^5	10^6	10^8

- Moment estimation

The moment estimation is based on measurements. The goal of the measurements is to estimate the bit rate with as few errors as possible.

- Bit Rate Estimation

The bit rate within an interval i is the number of bits that has arrived during that sample interval i divided by the sample interval duration. One can write:

$$B_i = \frac{b_0 + \sum_{k=1}^N b_k + b_{N+1}}{T_s},$$

where b_0 is the number of bits of previous frame ending within the sample interval, b_{N+1} is the number of the bits of the frame starting within the sampling interval. The sampling interval is determined by the desired base scale t_0 which in turn is obtained by the sample rate as $t_0 = 1/F_s$. Figure A.1 illustrates this.

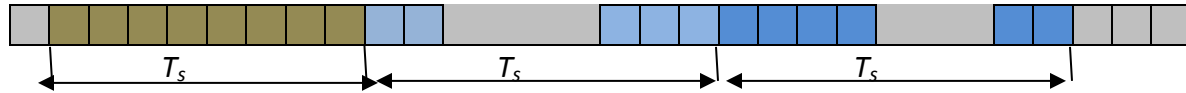


Figure A.1: $B_1 = \frac{8 \times 8bits}{T_s}$ $B_2 = \frac{5 \times 8bits}{T_s}$ $B_3 = \frac{6 \times 8bits}{T_s}$

The goal as previously stated is to reduce the bit error that could take time to recover over other techniques hence introducing another delay. The bit estimation is done as follows:

- Moment Estimation

Based on bit rate estimations, we can evaluate the moments at different timescales. To achieve this, we need first to set the base timescale which determine the sampling time interval as well as a scaling factor.

Note the base time scale is given by $t_0 = T_s = \frac{1}{F_s}$.

For simplicity, since the scaling factor specifies the distance between timescales, it is an integer for our case which respects the requirement of jitter avoidance, a requirement that states: the voice packets in an H.323 network must be delivered at integer factor of the audio frame length, we can write: $t_i = St_{i-1}$ where S is the scaling factor.

Define X_t to be the bit rate sample at the k -th timescale. The i -th moment at that timescale $E[X_t^i]$ is given by [177]:

$$E[X_t^i] = \frac{1}{M_k} \sum_{j=1}^{M_k} (X_{t,j})^i ,$$

where M_k is the number of samples present in X_t bits frame. This is formed from immediate samples of lower timescales:

$$X_{t,j} = \frac{1}{S} \sum_{m=(j-1)S+1}^{jS} X_{k-1,m} ,$$

with X_0 being the base scale computed from measurements.

APPENDIX B

E-MODEL TABLES

Table B.1: Speech transmission quality

R Value Range	Speech Transmission Category	User Impression
$90 \leq R < 100$	<i>Best</i>	<i>Very satisfied</i>
$80 \leq R < 90$	<i>High</i>	<i>Satisfied</i>
$70 \leq R < 80$	<i>Medium</i>	<i>Some users dissatisfied</i>
$60 \leq R < 70$	<i>Low</i>	<i>Many Users dissatisfied</i>
$50 \leq R < 60$	<i>Poor</i>	<i>Almost all users dissatisfied</i>

The transmission rating factor according to E-model defined in ITU-Rec G.107 and G.109 is related to 5 speech transmission qualities as in the table.



Table B.2: Packetization delay

Codec Type	Data Rate kbits/s	Delay ms
G.711 μ	64	1
G.711 A	64	1
G.726-32	32	1
G.729	8	25
G.723 MPMLQ	6.3	67.5
G.723 ACELP	5.3	67.5

Table B.3: Equipment impairment factor vs low bit rate signals

Codec Type	Operating Rate kbits/s	I_e Value
PCM G.726, G.727	64	0
ADPCM G.726, G.727	40	2
ADPCM G.721, G.726, G.727	32	7
ADPCM G.726, G.727	24	25
ADPCM G.726, G.727	16	50
LD-CELP G.728	16	7
LD-CELP G.728	12.8	20
CS-ACELP G.729	8	10
CS-ACELP G.729-A +VAD	8	11
RPE-LTP GSM 06.10, full rate	13	20
VSELP GSM 06.20, Half rate	5.6	23
ACELP GSM 06.60 Enhanced full rate	12.2	5
ACELP G.723.1	5.3	19
MP-MLQ G.723.1	6.3	15

Table B.3 shows the values for some codecs of interest against " I_e " from ITU Rec G.113 [13]. As the table shows, there are voice codecs that can provide significant saving in bandwidth for voice transportation at the expense of quality degradation " I_e ".

APPENDIX C

CODE FUNCTIONS

C-1 Test of Gaussian Arrival

```

    % Test of Gaussian Arrival
    %
    % Default parameters source 1
t1=1e-03; % packetization time interval
r1=64e+03; % Arrival rate
n1=1000; % number of sources 1
a1=650; % sources 1 active
bet1=1/a1*t1; % mean active time
al1=100e-03; % mean silence time
    % Default parameters source 2
t2=16e-03;
r2=32e+03;
n2=1000;
a2=650;
bet2=1/a2*t2;
al2=al1;

    % Statistics source 1- mean and variance
mn1=al1*n1/(al1+bet1);
var1=al1*bet1*n1/(al1+bet1)^2;
    % Statistics source 2
mn2=al2*n2/(al2+bet2);
var2=al2*bet2*n2/(al2+bet2)^2;

    % The Maximum likelihood estimator

```

```

%
Ye=(n1+n2)/2;
% Aggregate process
%
mr=r1*mn1+r2*mn2;
avar=r1*var1+r2*var2;
y=100:200:2000;
Nq=(1/avar^2)*(1/sqrt(2*pi))*exp((-1/2)*((y-Ye).^2/avar^2));
plot(y,Nq), grid

```

C-2 Test Diffusion Process

```

%Test Probability distribution Diffusion Approximation
%
% Default parameters
% Packetization delay sources
t1=1e-03;
t2=16e-03;
% Arrival rates
r1=64e+03;
r2=32e+03;
% Number of sources
n1=1000;
n2=1000;
% Number of Active sources
a2=360;
a1=480;

```

% Silence period fixed supposed exponentially distributed

alpha1=100e-03;

alpha2=alpha1;

c=100e+06; % Transmission Capacity link

% Active source period

beta2=a2*t2;

beta1=a1*t1;

% Mean Diffusion Process

y1=alpha1*n1/(alpha1+beta1);

y2=alpha2*n2/(alpha2+beta2);

% Aggregate arrival statistics: mean and variance

mr=r1*y1+r2*y2;

vary1=alpha1*beta1*n1/(alpha1+beta1)^2; % Diffusion variance

vary2=alpha2*beta2*n2/(alpha2+beta2)^2;

varag=r1^2*vary1+r2^2*vary2;

alpha=2*n1*alpha1*beta1/(alpha1+beta1)+2*n2*alpha2*beta2/(alpha2+beta2);

% The probability is significant only if arrival rate >= service rate

c=128e+03; % capacity available

C-3 Test Queue fluctuations

% test Queue fluctuations

beta=mr-c;

% Process decrements

ri=exp(-2*beta/alpha);

% fluctuations in the queue

ri=0.1:0.1:0.5;

```

ro=0.9;
n=1:1:10;
% probability distribution
Pn=ro.*(1-ri).^ri.^(n-1);
plot(n,Pn), grid

```

C-4 Test waiting Time On-Off Diffusion Process

% Test Waiting Time On-Off Diffusion Queue

%

% Default Parameters

```

n1=1500; % number of samples sources type 1
n2=1000;
a1=22; % average number of active source per sample
a2=15;
r1=64e+03; % source rate type 1
r2=32e+03;
t1=1e-03; % packetization delay
t2=16e-03;
beta1=a1*t1; % mean active time
beta2=a2*t2;
alpha1=100e-03; % mean silence period assumed exponentially distributed
alpha2=alpha1;
my1=alpha1*n1/(alpha1+beta1); % infinitesimal mean
my2=alpha2*n2/(alpha2+beta2);
mr=my1*r1+my2*r2;
vary1=alpha1*beta1*n1/(alpha1+beta1)^2; % infinitesimal variance

```

```

vary2=alpha2*beta2*n2/(alpha2+beta2)^2;
varag=r1^2*vary1+r2^2*vary2;
alpha=2*n1*alpha1*beta1/(alpha1+beta1)+2*n2*alpha2*beta2/(alpha2+beta2);
c=2e+06;
beta=(r1*my1-c)+(r2*my2-c);
ri=exp(-2*beta/alpha); % decrement factor
ro=0.1:0.2:2;
ri=exp(-2*c.*(ro-1)./alpha); % Decrement factor
wait=ro./c.*(1-ri);
plot(ro,wait), grid

```

C-5 Log-Test Multivariate Diffusion Process

```

% Test Log-Probability Multivariate Diffusion
%
% Default parameters
% coefficient of variation
C1=2;
C2=2;
% Probability transition matrix
P11=0.6;
P12=0.4;
P21=0.5;
P22=0.5;
% Mean sojourn time
alpha1=1/200e-3;
alpha2=1/100e-3;

```

% Infinitesimal mean

$B = [\alpha_1 * P_{11} \ \alpha_2 * P_{12}; \ \alpha_1 * P_{21} \ \alpha_2 * P_{22}]$;

% Number of Samples

$X_1 = 3000$;

$X_1 = 1500$;

$X_2 = 1500$;

$X = (X_1, X_2)$;

$X = [X_1; X_2]$;

% infinitesimal variance

$A_{11} = ((C_1 - 1) * P_{11} + 1) * P_{11} * \alpha_1 * X_1$;

$A_{12} = ((C_1 - 1) * P_{12} + 1) * P_{12} * \alpha_1 * X_1$;

$A_{21} = ((C_2 - 1) * P_{21} + 1) * P_{21} * \alpha_2 * X_2$;

$A_{22} = ((C_2 - 1) * P_{22} + 1) * P_{22} * \alpha_2 * X_2$;

$A = [A_{11} \ A_{12}; \ A_{21} \ A_{22}]$;

% standard deviation

$\sigma = \sqrt{A}$;

% Lamperti Transform

$Y = (1./\sigma) * X$;

$Y_1 = Y(1,1)$;

$Y_2 = Y(2,1)$;

$K = (1./\sigma) * B * \sigma$;

% Mean service per type of source at the equilibrium

$X_{e1} = 1/C * 5$;

$X_{e2} = 1/10 * C$;

$X_e = [X_{e1} \ X_{e2}]'$;

$n = (1./\sigma) * X_e$;




```

n1=n(1,1);
n2=n(2,1);
K11=K(1,1);
K12=K(1,2);
K21=K(2,1);
K22=K(2,2);
X0=[1000; 1000]; % number at t=0
X01=X0(1,1);
X02=X0(2,1);
Y0=(1./sigma)*X0;
Y01=Y0(1,1);
Y02=Y0(2,1);
% Coefficient of the transform determination
Cy_1=(-1/2)*((Y1-Y01)^2+(Y2-Y02)^2);
gama=Y;
gama1=gama(1,1);
gama2=gama(2,1);
Cy0=(-1/2)*(Y1-Y01)*((Y1+Y01-2*gama1)*K11+(Y2+Y02-2*gama2)*K12);
Cy0=(-1/2)*(Y1-Y01)*((Y1+Y01-2*gama1)*K11+(Y2+Y02-2*gama2)*K12)+(-1/2)*(Y2-
Y02)*((Y1+Y01-2*gama1)*K21+(Y2+Y02-2*gama2)*K22);
a1=(1/2)*(K11-((Y01-n1)*K11+(Y02-n2)*K12)^2);
a2=(1/2)*(K22-((Y01-n1)*K21+(Y02-n2)*K22)^2);
a3=(-1/2)*(Y1-Y01)*((Y01-n1)*(K11^2+K21^2)+(Y02-n2)*(K11*K12+K21*K22));
a4=(1/24)*((Y1-Y01)^2)*(-4*K11^2+K12^2-2*K12*K21-3*K21^2);
a5=(-1/2)*(Y2-Y02)*((Y01-n1)*(K11*K12+K21*K22)+(Y02-n2)*(K12^2+K22^2));
a6=(1/24)*((Y2-Y02)^2)*(-4*K22^2+K21^2-2*K12*K21-3*K12^2);
a7=(-1/3)*(Y1-Y01)*(Y2-Y02)*(K11*K12+K21*K22);

```

```

Cy1=a1+a2+a3+a4+a5+a6+a7;
b1=(-1/12)*(2*K11^2+2*K22^2+(K12+K21)^2);
b2=(1/6)*(Y1-Y01)*(K12-K21)*((Y01-n1)*(K11*K12+K21*K22)+(Y02-n2)*(K12^2+K22^2));
b3=(1/12)*(Y1-Y01)^2*(K12-K21)*(K11*K12+K21*K22);
b4=(1/12)*(Y2-Y02)^2*(K21-K12)*(K11*K12+K21*K22);
b5=(1/6)*(Y2-Y02)*(K21-K12)*((Y01-n1)*(K11^2+K21^2)+(Y02-n2)*(K11*K12+K21*K22));
b6=(1/12)*(Y1-Y01)*(Y2-Y02)*(K12-K21)*(K22^2+K12^2-K11^2+K21^2);
Cy2=b1+b2+b3+b4+b5+b6;
delta=0:100:3000; % Sampling interval vector
m=2; % bivariate Diffusion Process
delta=100:100:3000; % Log-Probability of the Diffusion function of sampling interval
Ly=(-m/2)*log(2*pi.*delta)+(Cy_1./delta)+Cy0+(Cy1.*delta)+(Cy2/2).*(delta.^2);
plot(delta, Ly), grid

```



% Test Probability Distribution Multivariate Diffusion Process

%

% Default Parameters

C2=2; % Coefficient of variation

C1=3;

p11=0.6; %Probability Transition

P12=0.4;

P21=0.5;

P22=0.5;

B=[alpha1*P11 alpha2*P12; alpha1*P21 alpha2*P22]; % Diffusion mean

X1=1400; % Number of Samples

X2=1600;

```

X=[X1,X2]';
X0=[X01 X02]';
X01=1000; % Number at time t=0
X02=1200;
X0=[X01,X02]';
C=100e+6; % Link capacity
Xe1=1/5*C; % Service time at the equilibrium
Xe2=1/10*C;
Xe=[Xe1,Xe2]';
A11=((C1-1)*P11+1)*P11*alpha1*X1; % Diffusion variance
A12=((C1-1)*P12+1)*P12*alpha1*X1;
A21=((C2-1)*P21+1)*P21*alpha2*X2;
A22=((C2-1)*P22+1)*P22*alpha2*X2;
A=[A11 A12; A21 A22];
sigma=sqrt(A); % Standard deviation
Y=(1./sigma)*X; % Lamperti transform
Y=inv(sigma)*X;
Y1=Y(1,1);
Y2=Y(2,1);
K=inv(sigma)*B*sigma;
K11=K(1,1);
K12=K(1,2);
K21=K(2,1);
K22=K(2,2);
n=(1./sigma)*Xe; % The mean number at the equilibrium
n1=n(1,1);

```

```

n2=n(2,1);
Y0=(1./sigma)*X0;
Y01=Y0(1,1);
Y02=Y0(2,1);
delta=1e-3; % sampling time
ri=exp(-K*delta); % Decrement factor
ro=0:0.1:1; % traffic intensity
[V E]=eig(exp(-K*delta)); % Eigen Characteristics
E1=E(1,1);
E2=E(2,2);
x=1:10:100; % Queue length
F=0.2e-79*(exp(E1*x)+exp(E2*x)); % Probability Distribution Normalized to 1
plot(x,F), grid

```



C-6 Simulation mm1 queue

Function Algorithm

```

Function [systime, systsize]= simmm1[n,lambda,mu]
% simmm1 simulate M/M/1 queue with arrival rate of
% lambda, a service time rate of
% mu.
%
% [systime, systsize]=simmm1[n,lambda,mu]
%
% Inputs: n – number of Markov Chain Transition
% lambda – arrival rate
% mu – service time rate

```

%

% Outputs: systime – cumulative transition times

% systsize – system size

%

Function code

% Default Exponential parameters

n=26;

lambda=64e+03; % arrival rate

mu=100e+06; % service rate

i=0; % system initially at level 0

tjump(1)=0; % time at first jumping is 0

systsize=i; % system size at level I at initial time

for k=2:n

if i==0

sertemp=0

else

sertemp=mu

end

r0=lambda+mu;

itime=-log(rand/r0) % interarrival times in the system

% exponentially r0 distributed

p=lambda/r0;

if(rand<p);

i=i+1; % birth

else

```

i=i-1; % death
end
systsize(j)=l; % system size at time i
tjump(j)=ltime;
end % for loop
hist(systsize)
systime=[0 cumsum(tjump)];
stairs(systime), grid

```

C-7 M/G/1 Diffusion Queue Simulation

Function Algorithm

```

Function [systime, systsize]= simmg1(N(my1+my1, Vary1+Vary2), y1,y2)
% simmg1 simulates random numbers from normal distribution
% N(mr,var) which reaches values
% y1 and y2
%
% Inputs: y1 sample mean source 1
%         y2 sample mean source 2
Function [N]=simml(N(mr,var),y1,y2)
% simml simulates the maximum Gaussian estimates from the
% normal distribution
%
% [N]=simml(N(mr,var),y1,y2)
% Inputs y1 – sample source 1
%         y2 – sample source 2
%

```

```

% Output N – mean samples size
%
Function [jumpn]=simgeom(N,ri)
% simgeom random numbers from geometric distribution
%  $p(n)=(1-ro)ri^n, n>0,$ 
% [jumpn]=simgeom(N,ri)
% Input: N – mean sample size
%         ri – decrements
%
% Output: jumpn – vector number of transition
%
% outputs: systime – system time
%         Sysysize – system size
%
Code Program
% Default Parameter
n1=30; %  $E_1$  incoming link
n2=24; %  $T_1$  incoming link
a1=22; % active slots source 1
a2=16; % active slots source 2
beta1=a1*t1; % mean active period source 1
beta2=a2*t2; % mean active period source 2
alpha1=100e-03; % silence period
alpha2=alpha1;
y1=alpha1*n1/(alpha1+beta1); % diffusion number of source 1
y2=alpha2*n2/(alpha2+beta2); % diffusion number of source 2

```

```
r1=64e+03; % rate source 1
r2=32e+03; % rate source 2
mr=r1*y1+r2*y2; % aggregated traffic rate
sav=(n1*alpha1)/(alpha1+beta1)+(n2*alpha2)/(alpha2+beta2); % sample variance
sam=mr-c; % sample mean
ri=exp(-2*sam/sav); % decrement factor
ro=mr/c; % traffic intensity
[mn,varn,mnci,varci]=normfit(normrnd(y1,y2)); % Maximum Gaussian estimator
N=mn;
arrt=(1/mr)*rand(1,N); % interarrival times
cumarrt=[0 cumsum(arrt)];
ri=0.5; % Fixed Fluctuation imposed by the MLE
jumpn=floor(log(rand(1,N))/log(1-ri)); % transistion points
n=length(jumpn); % number of transtions
cumjumpn=[0 cumsum(jumpn)]; % system size
c=100e+06; % LAN transmission rate
servt=(1/c)*rand(1,n); % service time intervals
cumservt=[0 cumsum(servt)]; % cumulative system time
for i=1:n
systsize(i)=cumjumpn(i);
systime(i)=cumservt(i);
end
hist(systsize)
stairs(systime)
```