Performance Modeling and Analysis of Wireless Local Area Networks with Bursty Traffic

Submitted by Noushin Najjari to the University of Exeter
as a thesis for the degree of
Doctor of Philosophy in Computer Science
In January 2017

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

To My Family

Abstract

The explosive increase in the use of mobile digital devices has posed great challenges in the design and implementation of Wireless Local Area Networks (WLANs). Ever-increasing demands for high-speed and ubiquitous digital communication have made WLANs an essential feature of everyday life. With audio and video forming the highest percentage of traffic generated by multimedia applications, a huge demand is placed for high speed WLANs that provide high Quality-of-Service (QoS) and can satisfy end user's needs at a relatively low cost. Providing video and audio contents to end users at a satisfactory level with various channel quality and current battery capacities requires thorough studies on the properties of such traffic. In this regard, Medium Access Control (MAC) protocol of the 802.11 standard plays a vital role in the management and coordination of shared channel access and data transmission. Therefore, this research focuses on developing new efficient analytical models that evaluate the performance of WLANs and the MAC protocol in the presence of bursty, correlated and heterogeneous multimedia traffic using Batch Markovian Arrival Process (BMAP). BMAP can model the correlation between different packet size distributions and traffic rates while accurately modelling aggregated traffic which often possesses negative statistical properties.

The research starts with developing an accurate traffic generator using BMAP to capture the existing correlations in multimedia traffics. For validation, the developed traffic generator is used as an arrival process to a queueing model and is analyzed based on average queue length and mean waiting time. The

performance of BMAP/M/1 queue is studied under various number of states and maximum batch sizes of BMAP. The results clearly indicate that any increase in the number of states of the underlying Markov Chain of BMAP or maximum batch size, lead to higher burstiness and correlation of the arrival process, prompting the speed of the queue towards saturation.

The developed traffic generator is then used to model traffic sources in IEEE 802.11 WLANs, measuring important QoS metrics of throughput, end-to-end delay, frame loss probability and energy consumption. Performance comparisons are conducted on WLANs under the influence of multimedia traffics modelled as BMAP, Markov Modulated Poisson Process and Poisson Process. The results clearly indicate that bursty traffics generated by BMAP demote network performance faster than other traffic sources under moderate to high loads.

The model is also used to study WLANs with unsaturated, heterogeneous and bursty traffic sources. The effects of traffic load and network size on the performance of WLANs are investigated to demonstrate the importance of burstiness and heterogeneity of traffic on accurate evaluation of MAC protocol in wireless multimedia networks.

The results of the thesis highlight the importance of taking into account the true characteristics of multimedia traffics for accurate evaluation of the MAC protocol in the design and analysis of wireless multimedia networks and technologies.

Acknowledgement

I would like to extend my thanks to the many people without whom this work would not have been possible.

First and foremost I would like to say a special thank you to my enthusiastic supervisor, Professor Geyong Min for his constant help, support and encouragement. My PhD experience has been amazing and I thank professor Min not only for his tremendous academic support, but also for giving me so many wonderful opportunities.

I am also hugely appreciative to my second supervisor Dr. Jia Hu, for his immense support and constant guidance.

I am forever grateful and indebted to my parents Dr. Hossein Najjari and Mrs. Fatemeh Karami for their unconditional love and priceless support without which I would not be where I am today. I would also like to thank my husband Dr. Shahin Assadinia for his support and patience during these years. Many thanks to my sister Mehrnoush, and brother Sina for being a source of strength. My family are the most important people in my world and I dedicate this thesis to them.

Finally I would like to thank my fellow research students in the lab for creating memorable times, in particular Haozhe Wang and Wang Miao.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACK	Acknowledgment
AIFS	Arbitrary Inter Frame Space
A-MPDU	Aggregate MAC Protocol Data Unit
A-MSDU	Aggregate MAC Service Data Unit
AP	Access Point
BA	Basic Access
BMAP	Batch Markovian Arrival Process
BSS	Basic Service Set
CA	Collision Avoidance
CDMA	Code-Division Multiple Access
CRC	Cyclic Redundancy Check
CSMA/CA	Carrier Sense Multiple Access with Collision Avoidance
CSMA/CD	Carrier Sense Multiple Access with Collision Detection
CTMC	Continuous-Time Markov chain
CTS	Clear-to-Send
CW	Contention Window
DCF	Distributed Coordination Function
DES	Discrete Event Simulation
DIFS	Distributed Inter-Frame Space
DSSS	Direct Sequence Spread Spectrum
EDCA	Enhanced Distributed Channel Access
ESS	Extended Service Set (ESS).
FCFS	First Come First Serve
FHSS	Frequency Hopping Spread Spectrum
FIFO	First In First Out
IBSS	Independent Basic Service Set
IDC	Index of Dispersion for Counts
IEEE	Institute of Electrical and Electronics Engineers
IID	Independent and Identically Distributed
IP	Internet Protocol
IPP	Interrupted Poisson Process
ISM	Industrial Scientific and Medical
LAN	Local Area Network

LRD	Long-Range-Dependent
MAC	Medium Access Control
MAP	Markovian Arrival Process
MIMO	Multi-Input Multi-Output
MMPP	Markov Modulated Poisson Process
NAV	Network Allocation Vector
OFDM	Orthogonal Frequency-Division Multiplexing
PCF	Point Coordination Function
PH	Phase-type
PHY	Physical
QoE	Quality-of-Experience
QoS	Quality-of-Service
RTS	Request-to-Send
SIFS	Short Inter-Frame Space
SRD	Short-Range Dependent
SPP	Switched Poisson Process
TCP	Transmission Control Protocol
TXOP	Transmission Opportunity
VBR	Variable-Bit-Rate
VMPP	Versatile Markovian Point Process
VoIP	Voice over Internet Protocol
Wi-Fi	Wireless Fidelity
WLAN	Wireless Local Area Network

Publications

Some elements of this work have been published or submitted for publication in conference proceedings:

- Noushin Najjari, Geyong Min, and Jia Hu, “Performance Analysis of WLANs under Bursty and Correlated Video Traffic”. 11th International Conference on Frontier of Computer Science and Technology (FCST-2017), Exeter, UK, July 2017.
- Noushin Najjari, Geyong Min, and Jia Hu, “Performance Analysis of WLANs with Heterogeneous and Bursty Multimedia Traffic”, Submitted to 2017 IEEE GLOBECOM.

Other publications in progress:

- Noushin Najjari, Geyong Min, and Jia Hu, “QoE-Based Admission Control in Multimedia WLANs under Bursty Traffic”, to be submitted to IEEE Transactions on Communications.

Chapter 1:

Introduction

Past decade has witnessed rapid development of wireless communication and technologies. Explosive growth in the number of wireless devices such as smartphones, PCs, personal digital assistants and home entertainment systems, along with the rapid formation of advanced multimedia applications, such as Voice-Over-IP (VoIP), Video Conferencing, IPTV, Telemedicine and Internet Gaming have resulted in revolutionary advance and deployment of the wireless technology.

Noted for being the most desired networking technology of choice, IEEE 802.11 based Wireless Local Area Networks (WLANs), also known as Wi-Fi (Wireless Fidelity) [1], have experienced impressive commercial success owing to their low cost and easy deployment. WLANs have connected an immensely expanding range of user-centric WiFi-equipped mobile devices over the last decade. With consumer services and applications that are persistently in need of a ubiquitous network access, WLANs are the preferred means of internet access for users and product developers worldwide. However, the constant increase in demand for high bandwidth and Quality-of-Service (QoS) of high definition multimedia applications in WLANs has made them extremely dense, posing great challenges on their design and deployment. Several technology forecasts predict that by 2020, the global mobile data traffic is expected to increase nearly eightfold. With 75% of the mobile data traffic being video, it is expected for the mobile data traffic to reach an average of 30.6 Exabyte per month, 53 percent higher than it was in 2015 [2]. A

great portion of this increasing high volume traffic is generated and carried through WLANs.

The great strain of constant and continuous high volume generation of multimedia traffic and in particular video will be a stern test for the 802.11-based WLANs. With the emergence of video streaming websites such as Netflix, Hulu, YouTube, and etc., video applications are recognized as significant drivers of current network traffic. As well as video, high volumes of traffic is generated as a result of online and instantaneous data synchronization and backups through mobile devices alongside the use of VoIP applications such as Skype and FaceTime.

Since its first release in 1997, the IEEE 802.11 standard has gone through various stages of development. Nonetheless, the primary aim of this standard, simple and best effort local area communication has always been a priority. However, almost all previous amendments to the standard were aimed at increasing the peak physical data rate through the exploitation of new modulation and coding schemes and recently through the use of Multiple-Input-Multiple-Output (MIMO) antenna mechanisms.

The IEEE 802.11 standard only deals with the two lowest layers of the Open System Interconnection (OSI) reference model: the Medium Access Control layer (MAC) (a sub-layer of the Data Link layer), and the Physical layer (PHY), as shown in Figure 1.1. The MAC layer offers two types of contention free channel access service: 1) a service provided by the Distributed Coordination Function (DCF), and 2) a polling-based service provided by the Point Coordination Function (PCF). These services are available on top of the Physical layer. DCF is originally the

mandatory service utilized by the MAC sub-layer, whereas PCF is provided as an optional service with a lower throughput, and as a result is rarely implemented in commercially available WLANs. The ratified DCF in IEEE 802.11 standard is based on the Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) protocol [3, 4].

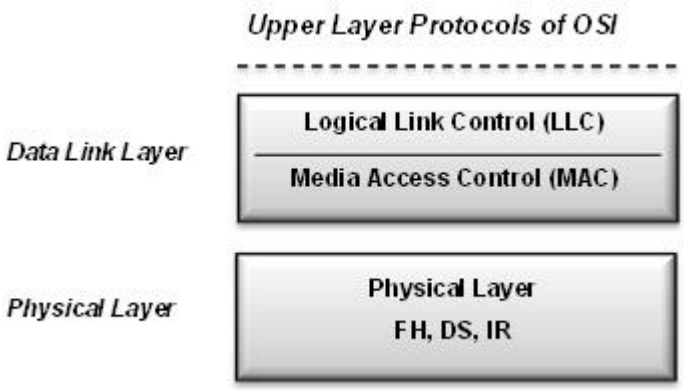


Figure 1.1: The IEEE 802.11 standard focuses on the bottom two levels of the OSI model: PHY and MAC

The legacy IEEE 802.11 a/b/g standards are able to support bandwidth intensive applications such as interactive games and audio or video streaming. However, they cannot guarantee QoS when traffic load increases. In order to overcome these problems and meet the QoS requirements of multimedia applications, the IEEE force group has introduced IEEE 802.11n, 802.11ac and 802.11ad standards which provide higher data rates [1, 5, 6]. With the implementation of Multiple Input Multiple Output (MIMO) antennas, frame aggregation, Block Acknowledgment (BA) and Orthogonal Frequency Division Multiplexing (OFDM), the new IEEE 802.11 standards address the need for increased network capacity, lower power consumption, longer range and ease of use. Due to promises of higher throughput and more performance reliability in the 5GHz of unlicensed band, the WLAN

market is now gradually evolving from IEEE 802.11n to 802.11ac [1] with 802.11ad [7] following. However the goal of IEEE 802.11 standards is to be backward compatible, in particular at the MAC or Data Link layer of all the standard series. For this reason each of the IEEE 802.11 standards would mainly differ in their PHY layer characteristics and the medium access mechanisms are left mostly unchanged.

Majority of existing analytical works on performance modelling of the DCF scheme in MAC layer are built upon the model originally proposed by Bianchi [8]. This model adopts a bi-dimensional discrete-time Markov chain to derive the saturation throughout of DCF. For analytical tractability and simplicity, most existing models have been developed under the assumption of unrealistic working scenarios where the traffic is saturated or the traffic model follows a Poisson process. However, realistic network conditions are non-saturated as very few networks operate in a situation where all nodes have frames to send at all times.

Many high quality measurement studies [7, 9-12] have shown that realistic traffic generated by multimedia applications in wireless and mobile networks exhibits self-similar nature and bursty outlook over a wide range of timescales. Since Self-similarity and burstiness can degrade network performance through requirement of large buffers, causing long delays and high volume of packet loss, it is of most importance that it is taken into consideration in the study and development of highly efficient WLANs. Conventional teletraffic models such as Poisson model were initially successful and analytically simple for modelling the non-bursty traffic behaviour; however these models are no longer applicable and adequate for capturing traffic burstiness of compressed voice and video in modern

communication networks, where batch arrivals, event correlations and burstiness are of paramount importance [13-19].

To this end, this thesis studies and develops novel analytical models for capturing the characteristics of IEEE 802.11 standard MAC sublayer implemented in WLANs under bursty and self-similar traffic using the Batch Markovian Arrival Process (BMAP) [20]. BMAP can accurately model aggregated traffic which often possesses negative statistical properties (correlated arrivals, burstiness and self-similarity [21]). These properties are not only imitated on the level of packet arrivals, but BMAP can also capture the correlation between arriving packet size and the current packet arrival intensity.

1.1. Motivations and Challenges

Performance modelling and analysis has become a necessity in the design and development of computer and communication networks for the purpose of providing the best QoS possible to end users. To this end, accurate analytical models that can capture the real characteristics of network traffics play a pivotal role in maximizing the efficiency of future network designs.

With ever-increasing demands for multimedia applications and large increase in communication traffic over wireless local area networks in recent years, modelling and analysing performance metrics of IEEE 802.11 DCF protocol has become an important factor in design, development and optimization of WLANs. Developing analytical models using processes that are incapable of capturing the true characteristics of modern network traffic can lead to unexpected and incorrect

results. To obtain an in-depth understanding of the performance characteristics of WLANs, significant research efforts have been devoted to developing analytical models for the DCF protocol, however, there is a gap in literature to cover the analysis of this protocol under realistic network traffic characteristics of burstiness, correlation and self-similarity all combined in one comprehensive model. Areas related to this topic that required further research and attention are pointed out in the following.

1. Today's WLANs are integrating and transmitting a diverse range of traffic generated by multimedia applications which significantly differ with each other in their packet arrival rates and packet sizes, including video, voice and text [2, 22]. Although initially successful and analytically simple for modeling the non-bursty traffic behavior, the Poisson model has proven inadequate for capturing traffic burstiness of compressed voice and video in modern communication networks, where batch arrivals, event correlations and burstiness are important factors. The fractal behavior of multimedia traffic should be modeled using statistically self-similar processes [23], which have significantly different theoretical properties from those of the conventional Poisson process. Therefore, in order to accurately evaluate and to obtain a better understanding of the performance characteristics of modern WLANs, it is critical to consider the burstiness, correlation and self-similar characteristics of the traffic transmitting through these networks.

2. Most studies on WLANs have been carried out under unrealistic network conditions of saturated stations and small or infinite buffer sizes for MAC. In real world scenarios, WLANs operate under unsaturated network conditions

with limited but adequate buffer sizes [24-26]. Even though the IEEE 802.11 MAC DCF protocol has been extensively studied, it is imperative to study and investigate its performance behavior under unsaturated bursty and correlated traffic loads in order to develop a comprehensive and accurate analytical model. The model should be able to calculate the important performance measures required for real-time network traffic such as end-to-end delay, frame loss probability and network throughput.

3. In reality, WLANs are non-homogeneous and convey heterogeneous traffics from sources to destinations. For simplicity and tractability, most developed models study the properties of WLANs and the MAC protocol in scenarios where all the stations of the model are homogeneous and generate similar type of network traffic. Therefore, there is a void in literature for an accurate analytical model for non-homogeneous bursty and correlated flows in random access WLANs. An accurate model is required to evaluate and obtain a better understanding of the performance and heterogeneous characteristics of non-homogenous multimedia WLANs under bursty and correlated traffic.
4. With many simulation platforms available for design and analysis of WLANs, none can implement a traffic generator that considers self-similarity, burstiness and correlation characteristics of actual traffics flowing through these networks.

1.2. Research Aims and Contributions

This research is focused on the analysis and enhancement of the IEEE 802.11 MAC protocols in multimedia WLANs under bursty and correlated traffic and real working environments. The main contribution is the design and development of new analytical models for evaluating the impact of the bursty traffic arrival on the MAC scheme under unsaturated traffic loads and finite buffer capacity using the Batch Markovian Arrival Process (BMAP). To this end, the following studies have been carried out:

1. A reliable traffic generator is developed for generating bursty, correlated and self-similar multimedia traffic using BMAP and generalized m-state Markov Modulated Poisson Process (MMPP). The accuracy of the developed traffic generators are validated through modelling and analysis of $BMAP/M/1/N$ and $MMPP/M/1/N$ queues using different input settings and scenarios.

2. Reliable, cost-effective and efficient tools are developed for performance evaluation of 802.11 DCF protocol in WLANs under bursty, correlated and self-similar multimedia traffic using the developed BMAP and m-state MMPP traffic generators.

3. The effect of bursty traffic on important performance metrics of throughput, mean waiting time in queue, mean queuing delay, packet loss probability and energy consumption of packets transmitted in unsaturated WLANs under different settings of traffic intensity, buffer size, and network size are thoroughly studied. An in-depth performance analysis is then carried out to compare the QoS metrics between the gained results from the study of

WLANs under bursty traffic with models using Poisson Process or the 2-state MMPP. The model validations are subject to the traffic parameters obtained from the accurate measurements of the real-world multimedia video resources.

4. The MAC scheme of 802.11 standard is evaluated in the presence of heterogeneous traffic, through developing a versatile analytical model that captures the traffic heterogeneity and models the features of bursty and correlated traffic. The performance results highlight the importance of taking into account the heterogeneous traffic for the accurate evaluation of the MAC scheme in wireless multimedia networks.

5. The accuracy of the developed models is corroborated through extensive comparisons between the analytical model results and those obtained from NS-2 [27] simulation experiments.

1.3. Outline of the Thesis

The rest of this thesis is organized as follows:

Chapter 2: Background and Literature Review

This chapter introduces the background knowledge in regards to IEEE 802.11 standards, the DCF protocol ratified in the MAC layer, ad-hoc WLANs and traffic models, and also methods available in literature for modelling network traffics using real data traces. Finally the chapter covers a detailed literature review on the modelling and analysis of DCF protocols.

Chapter 3: Analysis of the BMAP/M/1 Queue

In Chapter 3, a detailed description on the analytical model for the BMAP/M/1 queue is presented. The formulas for the calculation of the first and second moments of mean waiting time, and mean queue length are presented along with the simulation and analytical results gained from modelling the queue.

Chapter 4: Performance Modelling of Wireless Local Area Networks under Bursty Traffic

Chapter 4 proposes a new analytical model for the MAC DCF scheme under unsaturated conditions with bursty traffic using BMAP. The chapter presents detailed analysis of the developed model and compares the gained analytical and simulation results with scenarios in which simpler processes are used to model network traffic.

Chapter 5: Performance Analysis of Wireless Local Area Networks with Heterogeneous Stations under Bursty Traffic

In this chapter an analytical model for the DCF protocol in WLANs is developed with heterogeneous stations. The analytical model is used to investigate the effects of the DCF protocol on the QoS of WLANs when heterogeneous sources of traffic are present. The chapter presents detailed analysis of the developed model and compares the gained analytical and simulation results with WLANs composed of homogeneous traffic sources.

Chapter 6: Conclusion

The thesis is concluded in this chapter by drawing together all the elements from the preceding chapters and discussing potential avenues of future investigation.

1.4. Summary

This chapter has briefly introduced the main challenges and research objectives of the thesis. Relevant background materials are now introduced in the following chapter before presenting the developed models.

Chapter 2:

Background and Literature Review

2.1. Introduction

Wireless communication technology has received much attention during the last decade. The IEEE 802.11 standard for Wireless Local Area networks [3], also commonly known as Wi-Fi (Wireless Fidelity), is a well-developed technology that has been around for almost 19 years. Today's IEEE 802.11 standards have come a long way since the release of the first version in 1997. The first version was introduced as an alternative or extension to the existing wired LANs which were based on the Ethernet technology and the IEEE 802.3 standard. Since its emergence, the IEEE 802.11 standard has endured continuous amendments and modifications in order to accommodate new functionalities and technologies and also to fulfil the evolving needs of the ever expanding digital world.

Low cost installation, easy deployment, accessibility and flexibility of IEEE 802.11 WLANs has resulted in their widespread use everywhere (e.g., homes, public, enterprise environments and etc.). At the same time, most mobile and wearable devices (e.g., smart watches and smart glasses) are equipped with Wi-Fi interface and technology, and it is expected for Wi-Fi to be increasingly installed on various emerging consumer electronics and embedded systems. The number of people using internet applications as well as the devices connected to the Internet and networks are growing immensely every day. Clearly this results in continuous

increase of the traffic flowing through networks and in particular WLANs. This in turn has and will continue to increase the demand for mobile-rich multimedia content, resulting in the constant need for WLANs to evolve and supply new and effective solutions for tackling the uprising challenges.

2.2. IEEE 802.11 Wireless Local Area Networks

Wireless LANs must meet the requirements of any LAN such as high capacity, ability to cover short distances, full connectivity among attached stations and the ability to broadcast [28]. The main communication medium for WLANs is radio waves, so they transmit data between devices without the need of physical connections. As a result, WLANs can either extend or replace wired LANs to provide the connectivity between a backbone network and the in-building or on-campus users.

IEEE 802.11 is the dominant wireless digital data transmission standard aimed at providing connection within WLANs between portable devices and a fixed network infrastructure. It supports three basic topologies: 1) the Independent Basic Service Set (IBSS), 2) the Basic Service Set (BSS), and 3) the Extended Service Set (ESS). All three configurations are totally supported by the MAC layer specifications.

The standards also define two types of operational modes in WLANs: ad-hoc/IBSS (peer-to-peer) mode and infrastructure mode. In the ad-hoc mode, the wireless nodes are self-configuring and communicate directly or indirectly with each other through the wireless communication medium without any pre-existing or

centralized infrastructure. The IBSS WLANs include a number of wireless stations or nodes that communicate directly with each other on ad-hoc, peer-to-peer basis. This normally results in building a full mesh or partial-mesh topology. Logically, an ad-hoc configuration is analogous to peer-to-peer office network in which no single node is required to function as a server. A very important feature of wireless ad-hoc networks is the multi-hop packet transmission, which is aimed to overcome transmission range limitation of radio transceivers and guides the packets via multiple hops to reach a distant destination node. It is important to note that it is very common to design and build a wireless ad-hoc network through the extension of single-hop based IEEE 802.11 Wi-Fi networks, as these types of networks have become the most sound and feasible solution for WLANs (Figure 2.1).

The Infrastructure mode uses fixed interconnected access points to provide connectivity to mobile and portable wireless devices. They are composed of at least one Access Point (AP) connected to the wired network infrastructure and a set of wireless nodes/ stations, where wireless nodes communicate with each other via the use of AP [3]. This type of configuration is called Basic Service Set (BSS). In 802.11 WLANs, communication goes through the Access Point which acts as a base station. For example if node A wants to communicate with node B, the data from node A flows to the AP and then from the AP to node B.

Extended Service Set (ESS) is a set of two or more BSSs forming a single subnetwork. Configurations of ESS consist of multiple BSS cells that can be linked by either wired or wireless backbones.

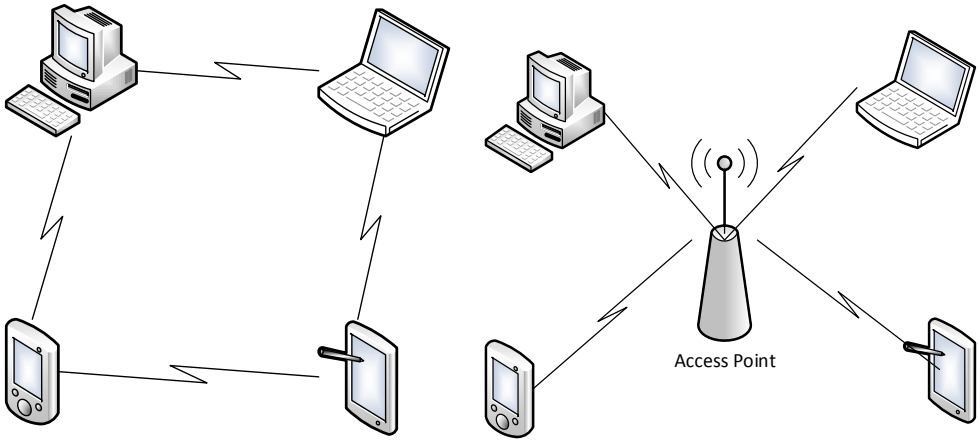


Figure 2.1: (a) Independent and (b) Infrastructure BSS architecture

In general, IEEE 802.11 WLAN standards focus on two main layers: the Medium Access Control (MAC) layer and the Physical (PHY) Layer. The layers allow functional separation of the standards and more importantly they allow for single data protocol to be used with several different radio frequency transmission techniques.

The PHY layer defines the different radio frequency transmission techniques, such as Frequency Hopping Spread Spectrum (FHSS), the Direct Sequence Spread Spectrum (DSSS), etc. The MAC layer provides various services to manage authentication, de-authentication, privacy and most importantly data transfer. The MAC layer will be discussed in more details in section 2.2.3 as it is the main focus of this research.

The following section provides a brief overview of the evolution of the IEEE 802.11 standard over years and shows where the standard lies currently.

2.2.1. Evolution of IEEE 802.11 Standards

The base standard for WLAN communications was released by IEEE in 1997 as IEEE Std. 802.11-1997 [3]. The main scope of the standard was to develop a set of specifications for connecting fixed, portable and mobile stations wirelessly within an area. It defined the specifications of the PHY and MAC layers for a wireless LAN. Initially the standard was designed to be used within the unlicensed spectrum bands of Industrial Scientific and Medical (ISM). This means that the IEEE 802.11 standard works in 2.4 GHz and 5GHz frequency bands, which are globally available [3]. The first version of the standard was capable of providing data rates of 1 to 2 Mbps. In September 1999, the 802.11b “High Rate” amendment was ratified into the 802.11 family which added two higher speeds of 5.5 and 11 Mbps. The basic architecture features and services of IEEE 802.11b, like most up to date versions, were defined by the original 802.11 standard, with changes made only to the PHY layer. The changes resulted in higher data rates and more robust network connectivity. To increase the support of the MAC-layer QoS in WLANs, IEEE 802.11e Enhanced Distributed Channel Access (EDCA) protocol was proposed in 2005. The standard proposes three QoS differentiation schemes in terms of Arbitrary Inter-Frame Space (AIFS), Contention Window (CW), and Transmission Opportunity (TXOP) which are of critical importance for delay-sensitive applications, such as Voice and streaming multimedia [29].

Up until 2009, the 802.11 family included six over-the-air modulation techniques, all involving the same protocol, where the most popular ones were defined by a, b and g amendments to the original standard and are also known as 802.11 legacies. In 2009 another modulation technique which incorporates Multiple-Input Multiple-

Output (MIMO) was introduced with the establishment of the IEEE 802.11n. In the years following, the shortcomings of the first WLAN products resulted in many amendments to the standard and up until today have resulted in the evolution of the IEEE 802.11 specification. During this time throughput enhancements have been the main priority of the IEEE 802.11 development [30]. One of the key solutions to a higher throughput in WLANs has been the adoption of new physical layer techniques. The earliest techniques used the Orthogonal Frequency-Division Multiplexing (OFDM) which increased the maximum data rate up to 54 Mbps. The most recent versions of IEE 802.11 standard benefit from the adoption of Multiple-Input Multiple-Output (MIMO) antenna technologies [31], with IEEE 802.11n, also known as high throughput network, being the first standard to benefit from this advancement in 2009. The maximum data rate defined for IEEE 802.11n standard is 600 Mbps. As part of the amendments to 802.11n several MAC Layer enhancements were presented in [31], such as frame aggregation. One of the advances in the development of 802.11n was to increase the throughput by reducing the MAC layer overhead. For this purpose two approaches to frame aggregations are presented in IEEE 802.11n: 1) the first approach is to Aggregate MAC Service Data Unit (A-MSDU) and, 2) the second approach is to Aggregate MAC Protocol Data Unit (A-MPDU). A-MSDU approach allows different MSDUs destined for the same destination to be sent together in one single MPDU of maximum 7935 bytes. As a result they would share a common MAC header and Cyclic Redundancy Check (CRC) fields. This makes the MAC less dependent to transmission errors as when errors occur the whole packet needs to be retransmitted. However, on the other hand, the A-MPDU aggregates several MPDU sub-frames into a single PHY packet with maximum size of 65536 bytes.

The aggregation happens after the MAC headers are added to each frame; therefore it will allow each sub-MPDU to have its own CRC field.

However the widespread dissemination of mobile devices equipped with diverse networking and multimedia capabilities and extensive use of advanced multimedia applications is intensifying the growth of mobile video traffic, which was already encompassing more than half of the global mobile data traffic by the end of 2013 [32]. Therefore it is obvious that WLANs require specific functions and enhancements to be able to manage and cope with various multimedia applications which would include real-time inter-active audio and video, or streaming live and stored audio and video.

Protocol	Release Date	Operating Frequency (GHz.)	Data Rate (Max)
Legacy	1997	2.4-2.5	2 Mbit/s
802.11a	1999	5	54 Mbit/s
802.11b	1999	2.4-2.5	11 Mbit/s
802.11g	2003	2.4-2.5	54 Mbit/s
802.11e	2005	2.4-5	11 Mbit/s
802.11n	2009	2.4/5	54-600 Mbit/s
802.11ac	2013	5	6.77 Gbit/s
802.11ad	2012	60	7 Gbit/s

Table 2.1: A summary of the IEEE 802.11 standard family

Therefore, to enhance throughput even further two new important amendments have been introduced to the IEEE 802.11n standard which are IEEE 802.11ad and IEEE 802.11ac. IEEE 802.11ac enables throughput of up to 6.77 Gbps and supports multi-user access techniques. IEEE 802.11ad enables throughput of up to

7 Gbps with the possibility of transmitting in the 60 GHz band that provides the opportunity for much wider band channels. Table 2.1 summarizes the important versions of the 802.11 standard family.

2.2.2. Ad-Hoc Networks

Due to their easy deployment and fast configuration, ad-hoc networks have gained increasing popularity in recent years. These networks can usually be set-up in environments where establishing a planned network is impossible or difficult or even not economically feasible, such as training grounds, conference sites, disaster areas, etc. Mainly ad-hoc networks are composed of wireless and potentially mobile stations which do not require any infrastructure or a centralized AP for administration. The administration and utilization of these networks are managed and performed in a distributed peer-to-peer manner. In cases where the stations are mobile, they are free to move around randomly and organize themselves and the connections arbitrarily. As a result, ad-hoc networks usually do not have fixed topologies and their network topology may change unpredictably and rapidly. Therefore it is important that ad-hoc networks are quick in adapting to varying number of stations. Due to limited transmission ranges, situations could occur where the wireless ad-hoc network is not fully connected and as a result the data travels through multiple intermediate stations in a multi-hop mode before reaching the destination [33]. From this point of view ad-hoc networks are considered different to peer-to-peer communication where similar devices directly communicate with each other and all data transmissions take place over single hop connections.

For wireless ad-hoc networks to become more supportive of multimedia applications that have specific QoS requirements, they should provide correct traffic differentiation and support for heterogeneous services. Multimedia traffic is usually delay-sensitive, and video or audio data have a threshold on end-to-end delays. Therefore end-to-end QoS assurance is only possible if every station in the network provides the means for offering QoS guarantees.

One of the most important features of wireless ad-hoc networks is the Medium Access Control (MAC) protocol. The MAC protocol has direct effect on how efficient and reliable data is transmitted in wireless ad-hoc networks as it should address the channel contention and collision problems among stations while effectively utilizing the communication channel. To develop a QoS aware MAC protocol, much attention should be paid to providing a good balance between protocol complexity, signalling overhead, QoS reservation methods, efficient use of resources, energy consumption and most importantly available traffic classes.

2.2.3. Medium Access Control (MAC)

MAC Protocol supplies the functionality required to provide reliable delivery mechanisms for user data over noisy, unreliable wireless media, therefore it plays a pivotal role in wireless networks through data transmission coordination [4]. When two or more stations within a WLAN transmit simultaneously over the wireless channel, collisions can occur. Therefore it is the job of the MAC layer protocol to determine when and how the stations access the shared wireless channel in order to avoid the occurrence of collisions as much as possible.

IEEE 802.11 MAC layer covers three main functional areas of reliable data delivery, medium access control, and security.

802.11 MAC is based on two types of algorithms: 1) distributed access protocols and 2) centralized access protocols. In distributed access protocols the decision to transmit is distributed over all the nodes within the network using a carrier-sense mechanism. In the centralized access protocol there is a centralized decision maker which regulates the transmissions. For ad-hoc networks a distributed access protocol makes sense and can be attractive in wireless LAN configurations that consist of bursty traffic [28].

The next section covers the mechanism of the Distributed Coordination Function ratified in the IEEE 802.11 MAC protocol which plays a major part in transmission coordination of data within ad-hoc WLANs.

2.2.4. Distributed Coordination Function (DCF)

The lower sub layer of the MAC is the Distributed Coordination Function (DCF) [34]. DCF is the basic and most prominent method in the MAC protocol to access the shared medium. It is a random access scheme based on Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) mechanism and uses a contention algorithm to provide access to the channel.

DCF employs two techniques for packet transmission: Basic Access (BA) mechanism and Request-To-Send/Clear-To-Send (RTS/CTS) mechanism. Basic Access mechanism is the default mechanism which uses binary exponential backoff rules for the management of hosts and retransmission of collided packets.

BA is a two-way handshaking technique and is characterized by the immediate transmission of a positive acknowledgement (ACK) by the destination station, upon successful reception of a packet transmitted by the sender station.

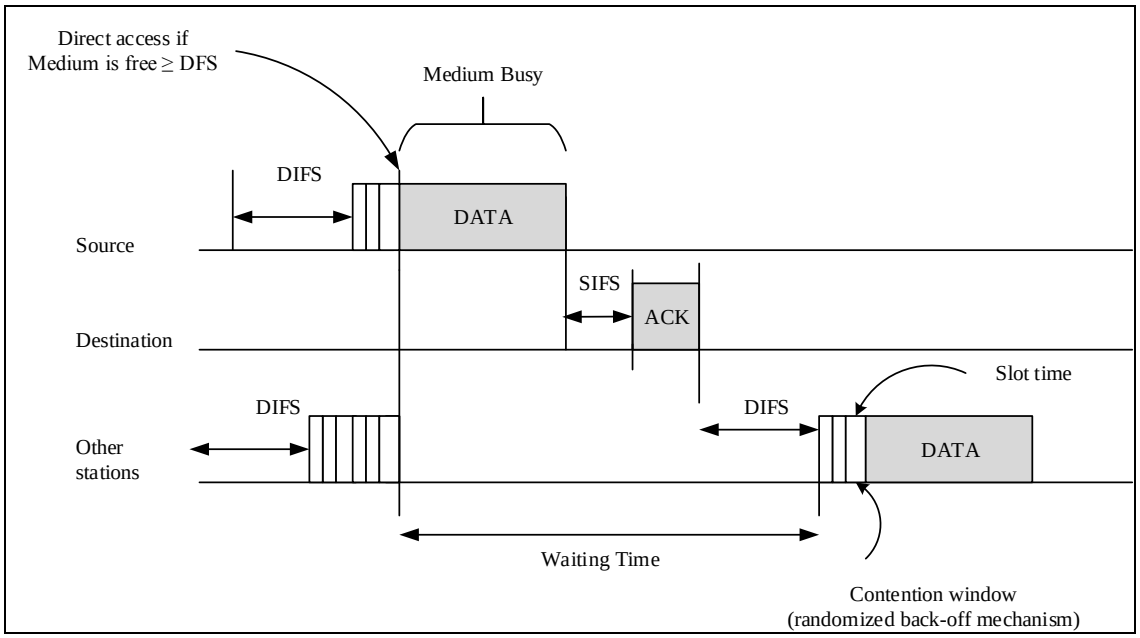


Figure 2.2: The Basic Access mechanism of DCF

A station with a new packet to transmit monitors the channel for a period of time. If the channel is idle for duration of Distributed Inter-Frame Space (DIFS), as shown in Figure 2.2, then the station transmits, otherwise if the channel is sensed busy (either immediately or during the DIFS duration) then the station continues to monitor the channel until it is idle for DIFS duration. When the situation is right, the station generates a random backoff interval before transmitting. This is the Collision Avoidance (CA) feature of the protocol. The backoff procedure also helps to avoid channel capture by allowing the stations to wait a random backoff time between two consecutive new packet transmissions, even if the medium is sensed idle in the DIFS time. The time immediately following the idle DIFS is slotted and

stations are allowed to transmit at the beginning of each time slot, which is dependent on the physical layer.

At each packet transmission the backoff time is uniformly chosen from the range of $[0, W_i - 1]$, where W specifies the current Contention Window (CW) size and i indicates the backoff stage. The value of W depends on the number of transmission failures for a packet. At the beginning of transmission and in the first attempt the value of W is set equal to the minimum contention window size. After each unsuccessful transmission the value of the contention window is doubled up to a maximum value of $W_m = 2^m W$, where m represents the backoff stage. When reaching the maximum, size of the contention window remains at the value of W_m until the transmission is successful or the retransmission attempts reach a retry limit. As long as the channel is sensed idle, the backoff counter is decreased by one for each time slot. When a transmission is detected on the channel, the backoff counter is "frozen" and then reactivated when the channel is sensed idle again for a period of DIFS. The station transmits when the channel counter reaches zero [34]. At the same time, other stations which are in the hearing distance of the transmission of the frame, update their Network Allocation Vector (NAV) to the expected period of time in which the wireless channel will be busy. This is known as Virtual Carrier Sensing mechanism. The stations start their backoff procedure either when the virtual carrier sensing or the physical carrier sensing indicate that the channel is busy [34].

Successful packet reception is signalled by the destination station via sending an ACK frame immediately after a Short Inter-Frame Space (SIFS) from complete reception. The SIFS duration plus the propagation delay is shorter than the DIFS

duration, therefore no other station is able to detect the channel idle for a DIFS until the end of the ACK. If the source does not receive the ACK frame within a specified ACK-Timeout period, or detects the transmission of a different packet on the channel, it reschedules for a retransmission of the packet based on the backoff procedure. The retransmission process continues until the retransmission limit is reached and by that time the packet is dropped.

DCF does not include a collision detection function (i.e., CSMA/CD) because collision detection is not practical on a wireless network. Another consideration in wireless networks is the problem of hidden terminals [35], which is the situation when a station is unable to detect a potential competitor for the channel because they are not within the transmission range of each other. The RTS/CTS shown in Figure 2.3 is a four-way handshaking technique implemented by DCF to help overcome the problem of hidden terminals. A station wanting to transmit a packet listens to the channel until it is sensed idle for a DIFS, then it follows the exponential backoff rule and then transmits a short frame called Request-To-Send (RTS). When the receiving station detects the RTS frame it responds after a SIFS with a Clear-To-Send (CTS) frame. The transmitting station then sends its packet if the CTS frame is received. The RTS and CTS frames carry information about the length of the packet to be transmitted. The listening stations can read this information and update their Network Allocation Vector (NAV). When detecting one frame among the RTS and CTS, a hidden station can delay further transmissions and thus avoid collision [8, 28, 36].

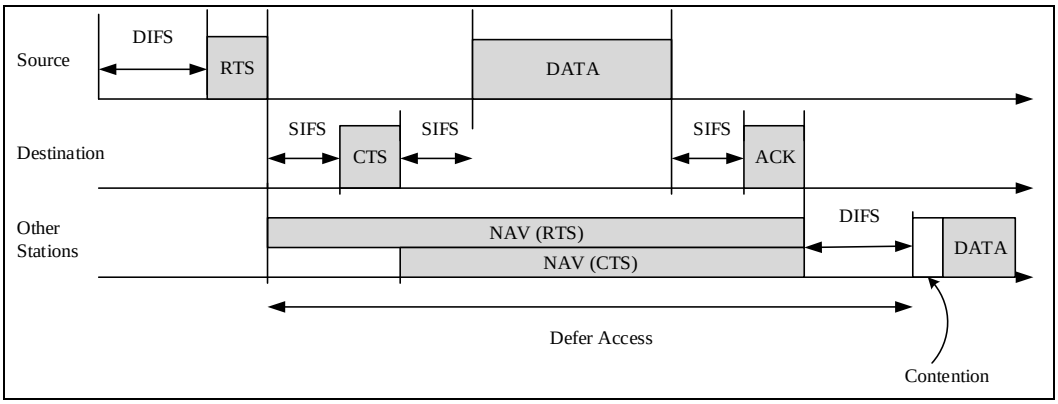


Figure 2.3: The RTS/CTS mechanism of DCF

2.3. Network Planning and Traffic Models

Performance evaluation plays a significant role in the field of computer and telecommunication networks, as optimal utilization of resources along with the end user satisfaction of QoS and QoE (Quality-of-Experience) is of paramount importance. One of the main goals of teletraffic engineering is to develop accurate models using queueing theory and stochastic processes in order to investigate, examine and predict the performance of communication systems with sufficient accuracy for the purpose of providing the necessary QoS demands of applications while controlling the cost [37].

To carry out an accurate network performance study, several important steps are needed where the first and foremost is the analysis and modelling of network traffic characteristics. Modelling and analysis of computer network traffic has been an area of extensive research for a very long time [15, 37-41]. A challenging issue in traffic characterization is obtaining accurate characteristics description of the complex traffic flow in order to form the system inputs. The most convenient

method in this regard is the use of mathematical, stochastic modelling based approach [42]. In modelling features of data streams, having a broad, versatile classes of point processes that can model qualitative features of arrival processes is extremely helpful as they are versatile, analytically and algorithmic tractable and are best candidates for simulation studies.

A method for extending the capabilities of the advanced arrival process is the generalization of the existing ones. The best candidate for generalization is the Poisson Process with exponential interarrival times. However, with the increased complexity of network traffics, the widely used simple descriptors such as the mean rate and peak-to-mean-ratio may not be sufficient enough for accurate representation of the burstiness, correlation and self-similarity characteristics of the network traffic. Several other descriptors have been proposed in literature that capture the correlation of network traffic and help to develop a close representation, such as the Peakedness Coefficient, the Auto-covariance, the Index of Dispersion over an observation time, and the Hurst parameter [43, 44]. An important step would be the good use of these descriptors in a traffic model, such as Markovian models.

The following subsections introduce the above mentioned descriptors and present the popular processes that can be used for traffic modelling such as Poisson Process and Markovian Arrival processes.

2.3.1. Network Traffic Modeling Descriptors

Modelling the traffic generated by network resources and applications is a highly complex task, as the users and protocols influence the behaviour of the applications [45]. However, network traffic simulation and modelling is proven to be very useful for both researchers and industry practitioners. In [40], the authors conducted a traffic analysis on Local Area Networks (LAN) and proposed for the first time that the Internet and LAN network traffic have self-similar characteristics. The result of many studies on network traffics analysis proved the fact that self-similarity, correlation and burstiness properties exists in the traffics generated within Ethernet, Wide Area and ad-hoc networks [40, 46]. Also the traffic generated through the World Wide Web, IP networks and the well-known 802.11 wireless LAN networks show the same characteristics [9, 10, 47-49].

Traffic that is considered bursty on many or all time scales statistically is described as self-similar. Self-similarity is a property characterized by factuality, meaning that self-similar phenomena displays similar structural patterns and the same statistical properties across a wide range of time scales. In the case of stochastic objects like time series, self-similarity is closely related to the so-called “bursty” behaviour (extended periods above the mean) and Long Range Dependence (LRD) at a wide range of time scales. In other words, burstiness in network traffic is the tendency of packets to form dense and sparse regions over a time period [50-52]. Burstiness property has negative effects on queues as the queueing system fed by bursty traffic has to operate at lower utilizations to be able to offer QoS at an acceptable level. In general burstiness is a result of two very important phenomena [52]: heavy-tailed distributions of interarrival times and high correlation between distant

events. The correlation between time distant events is described by the correlation function. The correlation function is also used to classify random processes into Long-Range-Dependant (LRD) or Short-Range-Dependant (SRD) processes. In wireless networks burstiness comes as a consequence of the effects of users, protocols and applications [53, 54].

Substantial body of literature has been devoted to Self-similarity [7, 23, 55-58] as it can considerably deteriorate the user-perceived QoS and degrade network performance by large delays, packet dropping and by requiring large buffers. The intensity and degree of self-similarity is measured by means of the Hurst parameter which is a dimensionless factor. The higher the Hurst parameter of a certain process, the more self-similar the process becomes. Estimations of the Hurst parameter are useful to understand the correlation structure and the evolution of a process, and to thus attain the aforementioned goals which the study of self-similarity is based on.

Assume $X = (X_t: t = 0,1,2, \dots)$ to be a covariance stationary stochastic process with the correlation function of $r(k)$. For each value of $m = 1,2,3, \dots$. let $X^{(m)} = (X_k^{(m)}: k = 1,2, \dots)$ define the new covariance stationary time series (with corresponding correlation function $r(m)$) which can be obtained from averaging the original series of X over non-overlapping and consecutive blocks of size m [50, 52].

- The process X is known as second order self-similar when the correlation function of the aggregated process $X^{(m)}$ is identical to the correlation function of the original process X in the limit of large K:

$$lim_{K\to\infty} \frac{r^{(m)}(K)}{r(K)} = 1, \quad for\ all\ m \tag{2.1}$$

- LRD traffic has a slowly decaying correlation function that makes the traffic bursty [51]. Therefore process X is said to be LRD if its correlation function $r(K)$ decays so slowly that its sum diverges:

$$\lim_{n\to\infty} \sum_{k=1}^{n} r(K) = \infty \tag{2.2}$$

In order for a long-range dependence to occur, the correlation function should drop off with a power-law:

$$r(K) \sim \frac{1}{K^{\alpha-1}}, \quad 1 < \alpha < 2 \tag{2.3}$$

The LRD process whose correlation function decays as a power-law with an exponent α, is a second order self-similar with Hurst parameter[50]:

$$H = (3 - \alpha)/2 \tag{2.4}$$

The Hurst parameter is defined to be limited to the range of $(0 < H < 1,\ H \in \mathbb{R})$. When the Hurst parameter lies in the range of $0 < H < 0.5$, the process is said to be Short-Range Dependent (SRD). A value of $0.5 < H < 1$ indicates self-similarity with positive near neighbour correlation. The more H is closer to 1, the more self-similar the process and therefore it is known as LRD. Graphically speaking, the lower the H is, the noisier or more volatile the process is, while the higher the H is, the smoother it is. When the Hurst value is equal to 0.5, the process is said to be absent of trends and memory less.

As mentioned above, Self-similarity is one way of defining the burstiness of network traffic. Other descriptors that can also be used to describe the burstiness property of network traffic are Squared Coefficient of Variation and the Index of Dispersion for Counts (IDC). The simplest definition for burstiness is the ratio of

peak rate to mean rate, however this does not reveal much information about the source because it does not capture how long the peak rate, which has the greatest effect on the model, can be sustained.

The squared coefficient of variation $c^2(X) = var(X)/E^2(X)$ is a normalized version of the variance of $X, var(X)$, normalized by dividing by the squared mean.

A related but more sophisticated measure is the index of dispersion of counts or IDC, $I_{idc}(t) = var(N(t))/E(N(t))$; which is the variance of the number of arrivals up to time t normalized by the mean number of arrivals. The number of arrivals up to time t, $N(t)$, is the number of interarrival times fitting in the interval $(0, t)$. Thus, the IDC captures similar but more timescale-dependent information than the coefficient of variation.

2.3.2. Poisson Process

The Poisson process is a stochastic counting process and is characterized as a renewal process with exponentially distributed interarrival times. It is one of the most important random processes in probability theory, which has been widely used to model the traffic behaviour and inputs in many communication networks and systems [26, 59-61].

In a Poisson process the events occur continuously and independently of one another and follow an Exponential distribution [62] which is the key to modelling traffics using the Poisson process. The cumulative distribution function of the Exponential distribution is [62]:

$$P(X < x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2.5)$$

In the above equation, x is a stochastic variable, λ is the mean arrival rate of a Poisson process, and $1/\lambda$ is the mean time between two arrivals.

The most important factor of Poisson process is the memory less property, which considerably facilitates the analysis of queuing systems. The number of arrivals found in the Poisson process during a finite interval depends only on the length of the interval and not on its starting point. As a result, the Poisson process is not able to model the bursty behaviour of multimedia traffic with time-varying arrivals. Some researchers [63] state that the Poisson Process cannot accurately model the network traffic as some aspects of network traffic show the properties of asymptotically second-order self-similar processes. Furthermore, both the processes of the interarrival times and of the packet sizes of aggregated traffic appear to be long-range dependent in addition to the packet and byte count.

2.3.3. Point Process

Point processes [64] are generalizations of the Poisson process, which are a class of discrete parameter stochastic processes in which each random variable is interpreted as the point in time when the n-th arrival occurs.

In the special case where the interarrival times are independent and identically distributed (i.i.d), the processes are known as Renewal process [65, 66]. One class of processes that is more general than the class of Renewal processes is the Markov Renewal processes. Markov Renewal processes are bivariate processes

where, as well as the time until the next renewal, the state at the time of renewal is also kept track of [44, 65]. Special cases of Markov Renewal process in which the time until the next arrival is exponentially distributed give rise to several useful and well known arrival processes, such as: Markovian Arrival Process (MAP) and Markov Modulated Poisson Process (MMPP). The generalization of MAP that includes batch arrivals is known as Batch Markovian Arrivals (BMAP) [65].

2.3.3.1. Markovian Arrival Process (MAP)

A popular tool for modelling arrival processes of stochastic systems is Markov Arrival Process (MAP). MAP which was first introduced by Neuts in 1979 [67], is most popular modelling process used in many areas such as queueing systems, reliability systems, telecommunication networks, inventory and supply chain systems, and risk and insurance systems. Versatility in modelling stochastic systems along with the Markovian property that leads to Markovian structures and flexibility in the resulting Markov chains are the most important properties of MAP that have led to its popularity [68-70].

Like the Poisson Process, MAP is also considered to be a popular counting process where the counting is modelled by the transition of the underlying continuous-time Markov Chain. Instead of switching between different arrival rates that depend on the states of the underlying Continuous-Time Markov Chain (CTMC), in MAP, an arrival is triggered by specific transitions between states. Through the utilization of this idea in a systematic manner, Neuts introduced Markov Arrival Process as generalizations of Poisson process, compound Poisson process, and Markov Modulated Poisson Process [67], which was however named

as versatile Markovian Point Process at first. Later in 1990, Lucantoni et. al. renamed this process to Non-renewal Arrival Process [65]. In 1991, Lucantoni [71] also introduced a simple matrix representation for Markovian Arrival processes, which made it easy to interpret parameters of Markovian Arrival Processes and to use Markovian Arrival Processes in stochastic modelling.

Mathematical Definition of MAP:

As mentioned earlier, MAP is a counting process whose arrival rate is governed by a continuous-time Markov chain (CTMC). If the underlying CTMC has m distinct phases called states and currently is in state $,1 \leq i \leq m$, it leaves this state with rate λ_i. This transition ends in state $j, 1 \leq j \leq m\ (i = j\ may\ hold)$, with probability p_{ij}, and triggers an arrival. Or with probability $p'_{ij}, i \neq j,$ this transition could end in state $j, 1 \leq j \leq m$, without triggering any arrival. It is important to note that all outgoing transition probabilities from one state sum up to one:

$$\sum_{j=1}^{m} p_{ij} + \sum_{\substack{j=1 \\ j \neq i}}^{m} p'_{ij} = 1, \quad 1 \leq i \leq m \tag{2.6}$$

The infinitesimal generator matrix Q_{MAP} of the underlying CTMC is given by:

$$Q_{MAP} = D_0 + D_1 \tag{2.7}$$

Let D_0 denote the infinitesimal generator of the underlying CTMC in the case of no arrivals, and assume D_1 to be the rate matrix in the cases of an arrival which leads to a possible state transition of the underlying CTMC. The corresponding D_0 and D_1 are defined as:

$$D_0 = \begin{pmatrix} C_{11} & \cdots & C_{1m} \\ \vdots & \ddots & \vdots \\ C_{m1} & \cdots & C_{mm} \end{pmatrix} \qquad D_1 = \begin{pmatrix} D_{11} & \cdots & D_{1m} \\ \vdots & \ddots & \vdots \\ D_{m1} & \cdots & D_{mm} \end{pmatrix} \tag{2.8}$$

where the elements of D_0 and D_1 matrices are defined as follows:

$$D_0 = [C_{ij}], \quad 1 \le i \le m,\ 1 \le j \le m, \tag{2.9}$$

$$D_1 = [D_{ij}], \quad 1 \le i \le m, 1 \le j \le m, \tag{2.10}$$

where:

$$D_{ij} = \lambda_{ij} p_{ij},\ 1 \le i \le m,\ 1 \le j \le m, \tag{2.11}$$

$$C_{ij} = \lambda_{ij} p'_{ij},\ 1 \le i \le m,\ 1 \le j \le m, i \ne j, \tag{2.12}$$

$$C_{ii} = -\lambda_i, \quad 0 \le i \le m, \tag{2.13}$$

D_1 is a non-negative matrix, with elements that give transition rates of the transitions that trigger an arrival event. D_0 has negative diagonal elements and non-negative off-diagonal elements representing the rates of the hidden transitions. It should be noted that $Q \ne D_0$, which implies that D_0 is invertible and that arrival process does not terminate.

The steady state probability vector of the underlying CTMC is calculated using the following equations:

$$\pi Q_{MAP} = 0 \qquad \text{and} \qquad \pi 1 = 1 \tag{2.14}$$

where 0 is a row vector of zeros, 1 is a column vector of ones, and $\pi = (\pi_1, \pi_2, \dots, \pi_m)$.

The mean steady-state arrival rate λ_{tot} generated by the MAP and squared coefficient of variation are calculated as [72]:

$$\lambda_{tot} = \pi D_1 1 \tag{2.15}$$

$$c_{var}^2 = 2\lambda_{tot}\pi(-D_0)^{-1}e - 1 \tag{2.16}$$

One of the most important measures in the family of MAPs is the lag k $(k > 0)$ correlation which is related to the time dependency of MAPs. It captures the correlation between interarrival times and can be calculated as:

$$\rho_k = \frac{\lambda_{tot}\pi[(-D_0)^{-1}D_1]^k(-D_0)^{-1}e-1}{2\lambda_{tot}(-D_0)^{-1}e-1} \tag{2.17}$$

The superposition of independent Markovian Arrival Processes yields a MAP again [71].

Setting $D_0 = [-\lambda], D_1 = [\lambda]$ and $\pi = (1)$ and following Eq. (2.15), it can be seen that the MAP is equivalent to a Poisson process with rate $\lambda_{tot} = (1)D_1 1 = \lambda$ [44, 65, 73].

2.3.3.2. Markov Modulated Poisson Process (MMPP)

The Markov Modulated Poisson Process (MMPP) is a non-stationary Point Process model, which has been extensively used for modelling time-varying arrival rates and important correlations between interarrival times [38, 74-78]. MMPP was firstly introduced by Naor and Yechiali [79], and later by Neuts [80].

MMPP is a generalization of many processes such as the Poisson process, Interrupted Poisson Process [81], and autoregressive process. In other words, MMPP could be assumed as a Poisson process whose arrival rate is determined by the states of an m-state irreducible CTMC. MMPP has been widely used to model several types of communication traffics, including packetized voice and

image sources, video and data traffic, and those resulting from their integration [74, 76, 82-84].

In a mathematical sense, MMPP is also known to be a subclass of MAP where the arrival matrix (D_1), is a diagonal matrix containing the arrival rate of the states of the underlying CTMC, Figure 2.5. However, the essential difference between MAP and MMPP is whether or not the phase changes just after an arrival.

Mathematical Definition of MMPP:

An MMPP parameterized by an m-state CTMC with infinitesimal generator Q, and m Poisson arrival rates $\lambda_1, \lambda_2, \lambda_3, .., \lambda_m$, can be described as a Poisson process whose arrival rate is given by $\lambda^*[J(t)]$, where $J(t), t \geq 0$, is an m-state irreducible Markov process. When the Markov chain is in state i, arrivals occur according to a Poisson process rate of λ_i. Therefore the arrivals of an MMPP occur according to a Poisson process of arrival rate $\lambda_i, 1 \leq i \leq m$, defined by the current state i of an underlying irreducible Continuous-Time Markov Chain (CTMC) with m states. The counting process of an MMPP is given by the bivariate process $[(J(t), N(t)): t \in T]$, where $N(t)$ is the number of arrivals within a certain time interval of t, $t \in T$, and $0 \leq J(t) \leq m$, is an m-state irreducible Markov process (the underlying CTMC).

$$Q = \begin{bmatrix} -\sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & -\sigma_{mm} \end{bmatrix} \qquad \lambda = \begin{bmatrix} -\lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -\lambda_m \end{bmatrix} \tag{2.18}$$

$$\sigma_i = \sum_{j=1, j \neq i}^{m} \sigma_{ij} \tag{2.19}$$

$$\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_m) \tag{2.20}$$

The element σ_{ij} is the transition rate from state i to state j of the MMPP and σ_{ji} is the rate out of state j to i.

The MMPP model is commonly used in telecommunication traffic modelling and has several attractive properties, such as being able to capture correlations between interarrival times while still remaining analytically tractable.

Special cases of MMPP are the Switched Poisson Process (SPP), which is a two-state MMPP, and the Interrupted Poisson Process (IPP), where in SPP one arrival rate is zero [39, 44, 84-87]. The IPP, also known as the On-Off process [84, 87], is defined as a 2-state MMPP with one arrival rate and is considered an important process for characterizing the bursty properties of network traffic, where packets only arrive during the ON period (state), and the traffic source becomes idle when the model is in OFF period (state), which means no data is generated during this time. The durations of the ON and OFF periods are exponentially distributed with means $1/\sigma_{on}$ and $1/\sigma_{off}$ respectively. The infinitesimal generator matrix and the rate matrix of IPP are as follows:

$$Q = \begin{bmatrix} -\sigma_{on} & \sigma_{on} \\ \sigma_{off} & -\sigma_{off} \end{bmatrix} \qquad \lambda = \begin{bmatrix} \lambda_{on} & 0 \\ 0 & 0 \end{bmatrix} \tag{2.21}$$

The On-Off traffic source is thus a stream of deterministically distributed correlated bursts and silent periods. The mean traffic arrival rate λ_{tot} is given by:

$$\lambda_{tot} = \frac{\lambda_{on}\sigma_{off}}{\sigma_{off}+\sigma_{on}} \tag{2.22}$$

Such features of MMPP have made it very attractive for modelling bursty traffic. And in particular this process is famous for modelling Voice traffic. A two-state

MMPP can be used to model the superposition of multiple On-Off voice sources [88], as superposition and splitting of MMPPs result in a new MMPP.

In many studies, the Markov Modulated Poisson Process has been used for modelling of aggregated On-Off voice sources [38, 88] and video sources [89]. In voice resources, during the talk spurts (i.e., during the On states), traffic is generated at a fixed rate of R (frames/second). And during the silences (i.e., during the Off states) no frames arrive. The On and Off durations are exponentially distributed with $1/\alpha$ and $1/\beta$ ($1/\alpha$, means sojourn time in the On state and $1/\beta$, means sojourn time during the Off state).

To model the correlation characteristics of k On-Off voice sources, a superposition of two-state MMPPs can be used, whose parameters can be determined using the Index of Dispersion for Counts (IDC) matching technique:

$$\lambda_1 = R\frac{\sum_{i=0}^{k'} i\pi_i}{\sum_{j=0}^{k'} \pi_j} \quad \text{and} \quad \lambda_2 = R\frac{\sum_{i=k'+1}^{k} i\pi_i}{\sum_{j=k'+1}^{k} \pi_j} \tag{2.23}$$

$$\sigma_1 = \frac{2(\lambda_2-\lambda_{tot})(\lambda_{tot}-\lambda_1)^2}{(\lambda_2-\lambda_1)\lambda_{tot}(IDC(\infty)-1)} \quad \text{and} \quad \sigma_2 = \frac{2(\lambda_2-\lambda_{tot})^2(\lambda_{tot}-\lambda_1)}{(\lambda_2-\lambda_1)\lambda_{tot}(IDC(\infty)-1)} \tag{2.24}$$

where

$$IDC(\infty) = \frac{1-(1-\alpha/R)^2}{(\alpha/R+\beta/R)^2} \tag{2.25}$$

$$\pi_j = \frac{k!}{j!(k-j)!} q^j (1-q)^{k-j} \tag{2.26}$$

where $q = \frac{\beta}{\alpha+\beta}, k' = [qk], \lambda_{MMPP} = kRq$

As well as voice, self-similar traffic can also be modelled by MMPP through the superposition of L two-state MMPPs. Self-similar traffic is famous for modelling

video in communication networks. The superposition of L two-state MMPPs results in a new MMPP with 2^L sates [55, 82]. Assume each two-state MMPP, named as $MMPP_j$ with subscript j denoting the j-th two-state MMPP ($1 \leq j \leq L$), has the infinitesimal generator Q_j, and the rate matrix Λ_j defined as:

$$Q = \begin{bmatrix} -\delta_{1j} & \delta_{1j} \\ \delta_{2j} & -\delta_{2j} \end{bmatrix} \qquad \Lambda = \begin{bmatrix} \lambda_{1j} & 0 \\ 0 & \lambda_{2j} \end{bmatrix} \tag{2.27}$$

The parameter matrices of the new MMPP, e.g. the infinitesimal generator Q and the arrival rate Λ, are thus calculated as follows (the symbol $\oplus$ denotes the Kronecker sum [90]):

$$Q = Q_1 \oplus Q_2 \oplus \ldots \oplus Q_L \quad \text{and} \quad \Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \ldots \oplus \Lambda_L \tag{2.28}$$

The resulting multi-state MMPP can then be used to characterize self-similar traffic. It should be noted that MMPP in the form of MAP is defined as [20, 71]: $D_0 = Q_{MMPP} - \lambda$ and $D_1 = \lambda$.

2.3.3.3. Batch Markovian Arrival Process (BMAP)

One of the many ways to analyse communication systems is through the application of queueing models, where the arrival process plays a fundamental role. In this regard, Batch Markovian arrival process (BMAP) is defined to be an attractive model for describing backbone packet traffic of communication systems and Internet Protocol [41]. BMAP came to light as Poisson Process was deemed to no longer be accurate enough in capturing specific characteristics of network traffic [38, 39].

The Batch Markovian Arrival process [20, 71] is a subclass of Stochastic Point process that generalizes the standard Poisson Process (and other Point Processes) by allowing “batches” of arrivals, dependent interarrival times, non-exponential interarrival time distributions and correlated batch sizes. It is equivalent to the Versatile Markovian Point Process (VMPP) which was introduced by Neuts [67] and is often referred to as the N-process [91]. The Versatile Markovian Point Process (VMPP) has a more complex notation than the BMAP defined by Lucantoni in [71]. The primary objective of VMPP was to extend the standard Poisson process to account for more complex customer arrival processes in queuing models. Ramaswami [91] incorporated the VMPP, which he called the N-Process in honour of Neuts, as an arrival process to a single-server queue with generally-distributed service times. In [92] the authors show that stationary MAPs or BMAPs are capable of approximating stationary (batch) Point Processes, suggesting the versatility and range of applications of such processes.

A distinguishing feature of the BMAP is the underlying Markovian structure; and is known to represent various arrival patterns including the stationary Poisson Process, Phase type process (PH), the correlated arrivals such as Markov Modulated Poisson Process (MMPP), Interrupted Poisson Process (IPP), Markovian Arrival Process (MAP), etc. [68, 93-96]. BMAP has been extensively studied and applied to various real world systems such as communication or teletraffic systems, production systems, modelling packets with different byte lengths in the Internet, reliability or insurance where batch dependent arrivals are commonly observed, such as customers arriving in batches to a queue,

simultaneous claims in an insurance company and failures occurring at the same time within an electronic device [97-101].

Since BMAP includes arrival of batches with sizes greater than 1 it is possible to say that BMAP is a generalization of MAP. Batches add to the modelling power and flexibility of MAPs, a fact that has been exploited in [102] to model IP traffic.

Mathematical Definition of BMAP:

The Batch Markovian Arrival process can be constructed by generalizing the Poisson Process with batch arrivals to allow for non-exponential times between the arrivals of batches while preserving the underlying Markov structure.

Assume a Poisson process that has a rate value of λ, where the arrival probability of a batch size of j is $p_j, j \geq 1$; in this case $N(t)$ would resemble the number of arrivals in $(0, t]$. As a result, the process $\{N(t)\}$ is considered to be a Markov process on the state space $\{i: i \geq 0\}$ with an infinitesimal generator of the form [71]:

$$Q = \begin{bmatrix} d_o & d_1 & d_2 & \dots \\ & d_o & d_1 & \dots \\ & & d_o & \dots \\ & & \vdots & \end{bmatrix} \tag{2.29}$$

where $d_0 = -\lambda$ and $d_1 = \lambda p_j$ for all $j \geq 1$. In state i, after an exponential sojourn with mean $1/\lambda$, with probability p_j the process jumps to state $i + j$. This transition corresponds to an arrival where j corresponds to the batch size of the arrival.

Batch Markovian Arrival Process is achieved through generalization of the above batch Poisson process. The achieved BMAP allows for non-exponential times

between the arrivals of batches, while still preserving the underlying Markov structure.

To present BMAP, a 2-dimensional Markov process of $\{N(t), J(t)\}$ is used on the state space of $\{(i,j): i \geq 0, 1 \leq j \leq m\}$, where the infinitesimal generator Q has the following structure:

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \cdots & \\ & D_0 & D_1 & D_2 & \ddots & \vdots \\ & & D_0 & D_1 & \cdots & \vdots \\ & & & & \cdots & \end{bmatrix} \tag{2.30}$$

In the matrix presented above, $D_k, k \geq 0$, are $m \times m$ matrices and D_0 has negative diagonal elements and non-negative off-diagonal elements. The row sums in D_0 are less than or equal to zero and the matrix is assumed to be non-singular. In other words D_0 is a stable matrix, and it is considered to govern the transitions of the phase process that do not generate any arrivals. On the other hand, D_k is considered to govern the rate of batch arrivals of size k (with appropriate phase change), $D_k, k \geq 1$.

It is important to note that the strictly negative diagonal elements of D_0 mean that the BMAP process can remain in any phase without any arrival for any finite time interval with positive probability.

Matrix D, $D \neq D_0$, is an irreducible infinitesimal generator which is defined as:

$$D = \sum_{k=0}^{\infty} D_k \tag{2.31}$$

Should $N(t)$ represent a counting variable, and $J(t)$ be an auxiliary state variable, then the above Markov process would define a batch arrival process in which the transitions from a state (i,j) to $(i+k, l), k \geq 1, 1 \leq j, l \leq m$, corresponds to batch

arrivals of size k, therefore batch sizes are completely dependent on the states i and j.

A key quantity for analysing $BMAP$ is the matrix generating function defined as [20, 71]:

$$D(z) = \sum_{k=0}^{\infty} D_k z^k \tag{2.32}$$

Assume π to be the stationary probability vector of the Markov process with generator D, then π satisfies the following conditions:

$$\pi D = 0, \;\; \pi e = 1 \tag{2.33}$$

where e is a column vector of 1's.

As a result, the fundamental arrival rate for the BMAP arrival process, which gives the expected number of arrivals per unit of time, is calculated as:

$$\lambda_{tot} = \pi \frac{dD(z)}{dz}\Big|_{z=1} e$$

$$\lambda_{tot} = \pi \sum_{k=1}^{\infty} k D_k e = \pi d, \qquad \lambda' = {\lambda_{tot}}^{-1} \tag{2.34}$$

$$d = \sum k D_k e \tag{2.35}$$

where λ' is the mean arrival rate.

Assume the underlying Markov process with generator D is in a random state of $i, 1 \leq i \leq m$, and the sojourn time in this state is exponentially distributed with parameter λ_i. At the end of the sojourn time, a transition to another state (or possibly the same state) occurs. This transition may or may not correspond to an arrival epoch. With probability of $p_i\,(j,k)$, $j \geq 1, 1 \leq k \leq m$, there will be a transition

to state k with a batch arrival of size j. Or with probability $p_i\ (0,k)$,$1 \le k \le m, k \ne i$, there will be a transition to state k without any arrival. Therefore we have the following for $1 \le i \le m$:

$$\sum_{\substack{k=1 \\ k \ne i}}^{m} p_i(0,k) + \sum_{j=1}^{\infty} \sum_{k=1}^{m} p_i(j,k) = 1 \tag{2.36}$$

Based on the transition probabilities, it is convenient to represent the evolution of the system in terms of a sequence of matrices $\{D_k, k \ge 0\}$ as:

$$(D_0\ )_{ii} = -\lambda_i \quad 1 \le i \le m$$

$$(D_0\ )_{ik} = \lambda_i p_i\ (0,k) \quad 1 \le i,k \le m, k \ne i$$

$$(D_j\ )_{ik} = \lambda_i p_i\ (j,k) \quad j \ge 1, 1 \le i,k \le m \tag{2.37}$$

D_0 governs transitions in the phase process that do not generate arrivals and D_k governs the transitions that generate batches of size k (with the appropriate phase change).

Therefore the sojourn time in state i is thus exponentially distributed with parameter λ_i, where:

$$\lambda_i = -(D_0)_{ii} \tag{2.38}$$

With $N(t)$ representing the number of arrivals in $(0,t]$ and $J(t)$ representing the auxiliary phase at time t, $P_{ij}(n,t)$ denotes the probability that $Pr\{N(t) = n, J(t) = j | N(0) = 0, J(0) = i\}$ and is the (i,j)-th element of the $m \times m$ matrix known as $P(n,t)$. The $P(n,t)$ matrix represents the probability of n arrivals in $(0,t]$ with a phase transition from i to j. The matrix generating function of the transition probability matrix is defined as:

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t)z^n, \quad for\ |z| \le 1, t \ge 0 \tag{2.39}$$

If a routine argument conditioning is applied to the first transition, the following result can be achieved:

$$P^*(z,t) = e^{D(z)t}, \quad for\ |z| \le 1, t \ge 0 \tag{2.40}$$

where $e^{D(z)t}$ is an exponential matrix. If Eq. (2.40) is differentiated with respect to z while the value of z is set to $z = 1$, the outcome would be the vector:

$$\lambda te + (I - e^{Dt})(e\pi - D)^{-1}\eta \tag{2.41}$$

The ith element of the vector would then give the expected number of arrivals in $(0, t]$ given that the phase at time $t = 0$ is i.

The intensity of group arrivals, λ_g, (arrival rate of batches or groups, excluding batch sizes) can be computed as:

$$\lambda_g = \pi De \tag{2.42}$$

For BMAP with single arrivals, the fundamental arrival rate would be equal to the intensity of group arrivals, $\lambda_{tot} = \lambda_g$. The variance ν of intervals between group arrivals (or variance between interarrival times) is equal to:

$$\nu = 2{\lambda_g}^{-1}\pi(-D_0)^{-1}e - {\lambda_g}^{-2} \tag{2.43}$$

While the correlation coefficient c_{cor} at lag k and the squared variation coefficient c_{var}^2 of intervals between successive group arrivals are given by:

$$\mathrm{c_{cor}}(k) = ({\lambda_g}^{-1}\pi(-D_0)[(D - D_0)(-D_0)^{-1}]^k e - {\lambda_g}^{-2})/\nu \tag{2.44}$$

$$\mathrm{c_{var}^2} = \nu{\lambda_g}^2 \tag{2.45}$$

Probability that an arriving batch is of size L can be given as:

$$P(B = L) = \pi(-D_0)^{-1} D_L e\,, \quad for\ L = 1, \dots, k \tag{2.46}$$

From which the moments of B are obtained as

$$E[B^n] = \pi(-D_0)^{-1} D_n^* e \tag{2.47}$$

where $D_n^* = \sum_{l=1}^{k} l^n D_l$.

2.4. Parameter Estimation

As stated in [41] accurate modelling of network traffic requires matching closely not only the packet arrival process but also the packet size distribution. Matching the arrival process has received considerable attention in literature, whereas packet size distribution is often ignored, mainly because simple and analytically tractable processes such as Poisson are unable to model both arrival and packet size distributions. In this regard, a natural problem that exists with the application of MMPP and BMAP for modelling network traffic is the estimation of their parameters from existing data traces. Most available traffic data traces consists of counts of events during a fixed length interval, such as bytes per second or packets per second, whilst BMAP and MMPP require more than just these important values for accurate fitting and modelling of network traffics. Furthermore, due to computational burdens, estimating parameters for MMPPs and BMAPs with more than a few states is too complex.

Many methods are available in literature for parameter fitting of MMPP processes to real network traces. Most of the methods stem from two very important

techniques of moment-based matching [83, 103] and likelihood-based [104, 105]. Dempster et.al., [106] introduced the Expectation-Maximization (EM) algorithm for computing Maximum Likelihood Estimates (MLE) from incomplete data. Based on this work, Deng and Mark [107] introduced the first approach for adapting the EM algorithm for MMPP, which uses the probability density function on interarrival time and data traces to derive the required parameters. Later, Ryden [105] surveyed the methods based on moment matching and maximum likelihood and proved that MLE methods are strongly consistent. In [108] he tailored the EM algorithm for the MMPP and developed an implementation which was then improved in [109].

Batch Markovian Arrival Process has enough flexibility to describe a wide variety of data and rate fluctuations in many applications. As with MMPP, a challenging issue in the study and application of BMAP is the accurate estimation of its parameters based on real-world systems. Any method used to estimate the parameters has to be as accurate as possible and should keep the number of states small enough for a tractable model. Due to incomplete data, standard statistical techniques such as moment matching cannot be used for BMAP. Since measured trace data does not contain all statistical properties required for the unique specification of a corresponding BMAP, a very common and important method for its parameter estimation would be the Maximum-Likelihood Estimation (MLE) [102]. In this regard, the Expectation-Maximization (EM) algorithm [106] is one of the methods of computing MLEs effectively. It is a statistical framework that computes the MLEs under incomplete data and is particularly useful for stochastic models with many parameters, hence BMAP. The fundamental idea in Maximum-Likelihood Estimation is to find the parameters maximizing the likelihood that the observed

data occurs. The correlation of data is not directly used in these methods of estimating parameters for BMAP, in fact the correlation values enter the model indirectly through the developed transition matrix [110].

Limited number of works exists in literature that focus on developing the required BMAP parameters from existing data traces [41, 102, 111]. In this research, the technique developed in [102] is used for parameter estimation of BMAP from actual data traces, which is based on the EM algorithm. A huge challenge in accurately estimating parameters of BMAP is keeping the number of states in the Markov chain small enough to make the performance models tractable. In general BMAPs are highly-parameterised models while in practice only interarrival times and sizes of batch arrivals or packets are commonly observed, therefore it can be viewed that data observed is actually being generated from a hidden Markov process [112].

2.5. Related Work

Significant amount of research has been carried out on the analysis and modelling of 802.11 protocols. In this section, prior literature relevant to the analytical modelling of DCF and the model-based admission control approaches are surveyed in order to present the existing gaps and demonstrate how the current research covers some of the important issues.

Performance modelling of 802.11 DCF has been widely studied through various approaches in literature; however majority of existing analytical works are built upon the model originally proposed by Bianchi in [8]. This model adopts a bi-dimensional discrete-time Markov chain to describe the exponential backoff

mechanism of the DCF scheme in order to derive the saturation throughput of WLANs under ideal channel conditions, absence of noise and hidden stations.

For analytical tractability and simplicity, Bianchi along with many other researchers have developed their models based on the assumption of unrealistic network scenarios in which the buffers of the stations are assumed to be saturated, or unlimited retries are considered [113-118].

In [24] Malone et al. extend Bianchi's model to the case of unsaturated traffic, taking into consideration the post-backoff behaviour with the assumption of ideal channel conditions and unlimited retries. Daneshgaran et al. in [119] propose another extended model which considers non-ideal channels and unsaturated traffic with the assumption of independence between transmission errors and packet collisions. As a result, they use equivalent probability of failed transmission to replace the collision probability in Bianchi's model and simplify the post-backoff stage with a single idle state and unlimited retries. In [120] the same authors extend Bianchi's bi-dimensional discrete time Markov-chain model by introducing a new state to the Markov chain and through complicated algebraic manipulation on Bianchi's throughput formula they approximately express the throughput as a linear function of the average packet arrival rate. In [121] the authors analyse the non-saturated IEEE 802.11 DCF networks through describing the behaviour of each station using a Markov renewal process, however they consider unlimited sized buffers.

Nevertheless, these simplified assumptions exclude any need for considering queuing dynamics or traffic models for performance analysis while currently WLANs are integrating and transmitting diverse range of traffic generated by

multimedia applications which significantly differ with each other in packet arrival rates and patterns, including video, voice and text [2, 22]. Multimedia applications are more sensitive to packet delay and jitter than traditional data applications, therefore it is unrealistic to assume unlimited MAC retransmissions and buffer sizes, which unnecessarily increase packet delays and jitter. These specific traffic characteristics and strict QoS requirements of multimedia applications in terms of bandwidth, delay, jitter and packet loss tolerance, have posed great challenges in wireless and resource constrained networks [122] and steered many researches towards developing models using realistic networking environments with unsaturated scenarios [24, 25, 75-77, 119, 121, 123-130], or using stochastic processes such as Poisson Process to model the traffic of WLANs [14, 121, 123, 125].

In this regard, the Poisson process and M/M/1 queues played a significant role in the development of many analytical models for the study of DCF protocol. For example in [123], Zhai, Kwon, and Fang demonstrate the usefulness of the Exponential distribution in approximating the MAC service time through the use of M/M/1/K queue in order to model the stations of the WLAN under Poisson traffic. Also in [125] Medepalli and Tobagi present a unified model where the transmission queue of each station is modelled as an M/M/1 queuing system. In [25, 130] the authors use the conventional Poisson Process within an M/G/1 queuing model to analyse the performance of WLANs under unsaturated network conditions. In [121], the authors describe the behaviour of each station of a 802.11-based WLAN as a Markov renewal process using M/G/1 queues.

The main interest of these studies in the use of Poisson Process for modelling the generated traffic within WLANs is the simplicity and tractability of this stochastic process. Finding a stochastic process that can accurately capture the specific characteristics of network traffic while remain analytically tractable is not easy.

Providing the best effort service for the earlier data applications was completely sufficient and could be easily modelled using the conventional Poisson Process which is analytically simple and tractable. However today, popular multimedia applications such as HTTP video streaming (e.g. YouTube), Cloud computing [131] and interactive games feature high-frame transmission rates, enhanced frame density, high levels of burst-like packet loss, latency, jitter and stringent delay constrains to provide a smooth viewing experience. This requires new modelling techniques in order to provide the best possible QoS. With video streaming already consuming the main portion of bandwidth within WLANs [132] and Variable-Bit-Rate (VBR) video traffic exhibiting noticeable burstiness over a wide range of time scales [7, 15, 16, 133], Poisson Process would no longer be adequate for capturing the complex characteristics of traffic generated by todays multimedia applications [7, 13-19, 110, 113, 133].

Besides burstiness, multimedia traffic and in particular video, require models that can capture the self-similarity and correlations between packet size distribution and packet arrival rates. Even though many studies have focused on the analysis of WLANs under multimedia traffic [134-139], to the best of our knowledge little research is available that concentrates on modelling the DCF protocol taking into account important characteristics of burstiness, self-similarity and correlation between the arrival rates and packet sizes. For example in [134] the authors

evaluate the performance of a video streaming applications over ad-hoc networks by varying video quality and network size, but assume all generated video packets have a fixed size; therefore they ignore any possible existing correlation between arrival rates and packet sizes. Even though many studies focus on using the Markov Modulated Poisson Process [74, 75, 140, 141] rather than Poisson process, MMPP [85] is only capable of capturing the burstiness and correlation that might exist between the arrival rates of packets in multimedia traffics and lacks the ability to capture the inherent self-similarity characteristics that exists in for example video traffic or the correlation between packet size distributions. This makes MMPP to be mainly suitable for accurate modelling of voice traffic [76, 142, 143]. Even though it is possible to model self-similarity through super-positioning of multiple MMPPs, as presented in [76], but still MMPP lacks the ability to capture the correlation between arrival rates and packet sizes in highly bursty multimedia traffics. In [76], the authors use the superposition of a number of two-state MMPPs to model the self-similarity properties of video traffic in wireless multimedia networks. Abdrabou and W. Zhuang in [142], use MMPP as a novel way to characterize the service (not the arrival) process of the IEEE 802.11 DCF shared channel to derive its effective capacity. In [143] the authors develop an analytical model to analyse the performance of the power saving protocols of the 802.11 family using MMPP to model general long-range dependent network traffics.

Modelling of packetized video, voice and data traffic has been an interesting topic of research over the years, and has always required the use of stochastic arrival processes. However, the stationary Poisson process, MMPP and others have proven to be inadequate for capturing the characteristics of the generated traffic as

the interarrival times and packet sizes in data streams are strongly correlated, and data is transmitted and received in batches of various sizes. Some studies take into account batch arrivals without correlation [144, 145], and some consider correlation and ignore batch arrivals [146-148].

In this regard BMAP has proven to be able to model dependent and non-exponential inter-arrival time distributions between batches and correlated batch sizes of arriving packets [102]. With no work available in literature that evaluates the performance of 802.11-based WLANs under multimedia traffic through the use of BMAP, this research will be the first in itself to model all three properties of burstiness, correlation and self-similarity of traffic in WLANs, not only in the arrival level but also in packet sizes of the generated traffics.

2.6. Summary

Self-similarity properties that exist in packet interarrival times, form one aspect of the correlation that exists in network traffic. The same properties exists and have been observed for the packet lengths of network traffic [149], meaning that the packet lengths can also be correlated with the interarrival times [41]. It is important to note that all these correlations significantly affect the accuracy of any performance model developed for the analysis of WLANs.

Majority of models designed for WLANs do not consider the variable packet lengths of network traffic which may lead to large errors, hence correct packet length distribution would indeed increase the accuracy of the models [84, 150]. The analytically tractable BMAP model with its batch arrivals is considered to be the

most flexible model from the Markov family that can be used for increased accuracy modelling of network traffic. BMAP can be used to model real time-varying correlated flows as it can capture the joint characterization of variable arrival rates and packet lengths of real traffic. It can be customized so that different packet lengths are represented by batch sizes of arrivals.

When packet lengths are considered to be independent of arrival rates and no batch arrivals are considered, BMAP is simplified and reduced into MAP or MMPP.

The main research of this thesis is based on the use of BMAP as the arrival process of developed models on performance analysis of Wireless Local Area Networks. In summary of the subjects discussed in this chapter, the reason for choosing BMAP as the arrival process is due to the following reasons:

1. The BMAP can not only model batch arrivals, correlation of interarrival times, and variance of interarrival times, but it can also model other subtle characteristics of the arrival process such as the correlation between local intensity of arrivals of batches with the size of the arriving batch [151].
2. Due to available parameter fitting techniques developed for BMAP [102, 111], these great capabilities of BMAP can be used for the analysis and modelling of real multimedia traffic.
3. And finally BMAP entails a unique advantage of combining great complexity and modelling capabilities of Markovian processes with the analytical tractability [152].

Chapter 3:

Modelling and Analysis of the BMAP/M/1 Queuing Systems

3.1. Introduction

Queueing systems under various types of arrivals and service processes have been extensively investigated by researchers due to their applicability in various networking situations, production and manufacturing systems, as covered in the literature. In this regard, traditional queueing analysis using the conventional Poisson Process is proven to be not powerful enough to capture the correlated nature of arrival processes.

To capture batch arrivals with variable rates, Lucantoni [65, 71] introduced the Batch Markovian Arrival Process (BMAP) which is a formal representation of the Versatile Markovian Point Process introduced by Neuts [67]. BMAP generalizes many familiar input processes such as: Markovian Arrival Process (MAP), Markov Modulated Poisson Process (MMPP), Phase-type renewal process, Interrupted Poisson Process (IPP) and Poisson Process. Since its introduction, many published papers have investigated the traffic modelling capability of BMAP as a wide range of real life traffic can be approximated using BMAP for modelling of input processes [41, 70, 102]. A real life example in which the use of BMAP seems to be the natural choice when modelling the arrival process, is in production and

manufacturing systems, where jobs (or customers) arrive in batches from various sources to a common processing centre and therefore the arrival process encompasses complicated characteristics that can no longer be assumed to follow the conventional Poisson process. Overall, the main idea of BMAP is to significantly generalize the Poisson processes and still keep the tractability property for modelling purposes. Also it is important to note that BMAP is a convenient tool for modelling both renewal and non-renewal arrival processes, and can be defined for both discrete and continuous times.

Furthermore, the analysis of queuing systems using BMAP to model the arrival process has received considerable attention in the stochastic modelling community [21, 70, 97-102]. With many useful particular cases, BMAP often leads to algorithmically tractable models. In [20] Lucantoni provided a nice summary of a number of important results for the BMAP/G/1 system.

One of the most significant features of BMAP is the underlying Markovian structure which fits ideally in the context of matrix-analytical solutions for stochastic models. The vast majority of the analysed BMAP/G/1 queuing models are based on the standard matrix analytic-method pioneered by Neuts [153, 154] and further extended by many other researchers (e.g., Lucantoni [20, 71]).

The key ingredient to the matrix analytical is the solution of G, a matrix functional equation. G is related to the busy period of the queue and indirectly, to the behaviour of the arrival process during successive idle periods of the queue. The relationship between the matrix analytical model and the traditional approach is

that the roots in the traditional analysis are eigenvalues of the matrix G in the matrix analytical model.

The remaining of this chapter concentrates on the important mathematical definitions of BMAP/M/1 queue with the sole purpose of introducing methods for calculating main performance measures such as the average queue length, average waiting time in queue of when the queue has limited buffer size. A simulator is developed to simulate m-state BMAP/M/1 queue with batch sizes of maximum K, for the purpose of validation of the developed analytical models of the queue.

To make sure the developed models and simulator for the analysis of m-state BMAP/M/1 queue with maximum batch size of K are correct and valid, a general analytical model and simulator are developed for m-state MMPP/M/1 queue. Due to complexity and for ease of tractability most studies in the literature concentrate on model and simulation of two state BMAP and MMPP queues. To fulfil the existing gap, the developed model and simulator for BMAP and MMPP queues are extended to be able to model and simulate BMAP and MMPP queues of any number of states, and for BMAP they can be adjusted to any maximum batch size.

The models and simulators developed in this section provide the basis of later studies on performance evaluation of Wireless Local Area Networks under bursty, correlated and self-similar traffics, as well as forming a great comparison environment for performance evaluation of BMAP/M/1 and MMPP/M/1 queues.

For the stability of the developed queueing systems in this section, the traffic intensity of the queues (ρ) which is calculated as $\rho = \lambda_{avg} * E(s)$ (with $E(s)$ being

the first moment of the service time distribution of the queue), is assumed to be less than one at all times: $\rho < 1$.

3.2. Study of the Busy Period

BMAP is a natural generalization of the (batch) Poisson arrival process, where the most important parameter towards modeling the performance factors of the BMAP/M/1 queue, is the calculation of the matrix G.

First it is assumed that the service time distribution of the BMAP/M/1 queue is composed of an Exponential distribution $H(x)$ defined with Cumulative Density Function (CDF) of:

$$H(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{3.1}$$

The Laplace transform of this service time distribution, $\widetilde{H}(s)$, is:

$$\widetilde{H}(s) = \frac{\lambda}{\lambda + S} \tag{3.2}$$

With finite mean of $E(s) = \mu_1'$, second and third moments of μ_2' and μ_3' respectively. The traffic intensity of the queue is calculated as:

$$\rho = \lambda_{tot} * E(s) < 1 \tag{3.3}$$

In the context of BMAP/M/1 queue, $G(z, s)$ is the two-dimensional transform of the number of jobs served during, and in the duration of the busy period. It is shown in [71] and [155] that $G(z, s)$ is the solution to:

$$G(z, s) = z \int_0^{\infty} e^{-sx} e^{D[G(z,s)]x} dH(x)$$

$$\equiv zh(sI - D[G(z,s)]) \tag{3.4}$$

It is possible to define the matrices $G(s) \equiv G(1,s)$ and $G \equiv G(0)$ and should also note that the matrix G should satisfy the following equation:

$$G = \int_0^\infty e^{D[G]x} dH(x) \tag{3.5}$$

Matrix G is stochastic when $\rho \leq 1$. For $\rho < 1$, the invariant probability vector g, of the positive stochastic matrix G, satisfies:

$$gG = g\ , ge = 1 \tag{3.6}$$

The vector g is known as the stationary probability vector of the infinitesimal generator $D[G]$. Matrix $D[G]$ has a nice probabilistic interpretation which was originally pointed out in [64]. Since G is calculated as being strictly positive, the off-diagonal entries of $D[G]$ are non-negative. So when the queue is stable, G is stochastic and as a result $D[G]e = 0$, which means that $D[G]$ is the infinitesimal generator of a finite-state, irreducible Markov Process. On the other hand, in the unstable case (i.e. when $\rho > 1$), G is strictly substochastic so that $D[G]$ is a stable matrix. Eq. (3.6) implies that g is the stationary vector of the matrix $D[G]$ and therefore the j-th component of this vector is the stationary probability that the arrival process is in state j given that the server is idle.

Once g is computed, moments of the queue length and waiting time distributions can be immediately computed.

An efficient algorithm is proposed by Lucantoni in [71] for the calculation of G using the matrix analytical methods, where the basic idea is to use the concept of

uniformization. As a result, if Q is the inifinitesimal generator of a continuous Markov process, then based on Eq. (3.5):

$$e^{Qt} = \sum_{n=0}^{\infty} e^{-\theta t} \frac{(\theta t)^n}{n!} L^n \tag{3.7}$$

where $\theta = max_i\{(-D_0)_{ii}\}$ and $L = I + \theta^{-1}Q$, which is a stochastic matrix. Based on these formulas, the new method leads to:

$$G = \sum_{n=0}^{\infty} \gamma_n (I + \theta^{-1} D[G])^n \tag{3.8}$$

where:

$$\gamma_n = \int_0^{\infty} e^{-\theta x} \frac{(\theta x)^n}{n!} d\tilde{H}(x), \quad for\ n \geq 0 \tag{3.9}$$

Thus G can be computed by successively iterating the following recursion:

$$H_{n+1,k} = \left[I + \theta^{-1} D[G]\right] H_{n,k} n = 0,1,2, \ldots, \tag{3.10}$$

$$G_{k+1} = \sum_{n=0}^{\infty} \gamma_n H_{n,k} \tag{3.11}$$

where I is an identity matrix and $H_{0,k} = I$.

The most important factor in this algorithm is the starting value for G_0. If the algorithm starts with $G_0 = 0$, the successive values of G_k will monotonically increase to the unique solution; however the convergence can be slow especially for high values of ρ. Lucantoni [20] suggests that starting with a stochastic matrix leads to extremely fast convergence which appears to be independent of ρ, and so recommends the iteration should start with $G_0 = e\pi$.

After calculating the G matrix using Eqs. (3.10) and (3.11), the value of g is calculated from Eq. (3.6).

3.3. Moments of the Queue Length at Departures

The equations required to calculate the moments of the queue length for the BMAP/M/1 queue at departures require the use of moment matrices of $A^{(i)}(1),\ i = 0, 1, 2, 3$. These moment matrices are defined for the $m \times m$ matrix of mass function $[\tilde{A}_n(x)]_{ij}$ defined as: the probability of given a departure at time 0, which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during that service there were n arrivals. The moment matrices are computed simultaneously as a concentrated matrix:

$$\left[A, A^{(1)}, A^{(2)}, A^{(3)}\right] = \sum_{n=0}^{\infty} \gamma_n L_n \tag{3.12}$$

The value of γ_n is defined in Eq. (3.9). L_0 is an $m \times 4m$ matrix of $[I, 0, 0, 0]$, and is calculated as:

$$L_{k+1} = L_k(I + \theta^{-1}S),\ \ k \geq 0 \tag{3.13}$$

where

$$S = \begin{bmatrix} D & D^{(1)} & D^{(2)} & D^{(3)} \\ 0 & D & 2D^{(1)} & 3D^{(2)} \\ 0 & 0 & D & 3D^{(1)} \\ 0 & 0 & 0 & D \end{bmatrix} \tag{3.14}$$

In order to calculate the moments of the queue length at arbitrary times, the moments of the queue length at departures should be calculated first. To this aim the important value required for the calculation of the queue length is the vector x_0. To define how x_0 is calculated we should first consider the definition for the stationary vector of the Markov chain embedded at departures from the queue. This stationary vector is defined as a joint probability of density of the stationary queue length and the phase of the arrival process, which can be defined as:

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & \dots \\ A_0 & A_1 & A_2 & \dots \\ 0 & A_0 & A_1 & \dots \\ 0 & 0 & A_0 & \dots \\ \dots & & \dots & \end{bmatrix} \tag{3.15}$$

The stationary probability vector $x = (x_0, x_1, \dots)$, where $x_i, i \geq 0$, and x_i are m-vectors, is defined from the P matrix. The x_0 vector defines the stationary probability that a departure leaves the system empty, e.g. x_{0j} resembles the stationary probability that a departure leaves the system empty with the arrival process in state j. By a classical property of Markov chains, the quantity of $(x_{0j})^{-1}$ is the mean recurrence time of the state $(0, j)$ in the Markov chain P.

Using the arguments classical in the theory of Markov renewal process [71], the value of x_0 can be expressed in terms of the invariant probability vector $\boldsymbol{\mathcal{K}}$ of k, which satisfies $\boldsymbol{\mathcal{K}}K = \boldsymbol{\mathcal{K}}$, $\boldsymbol{\mathcal{K}}e = 1$, and the vector $\boldsymbol{\mathcal{K}}^* = K^{(1)}(1)e$, of the row-sum means of $K(z)$. $K(z)$ is defined as the transition matrix generating function of the m-state Markov renewal process generated from observing the chain P only at its

visits to the level $\mathbf{0}$, and recording the indices of the states visited as well as the number transitions in P between consecutive visits to $\mathbf{0}$.

So the matrix $K(z)$ is defined as:

$$K(z) = K(z,0) = z\sum_{v=0}^{\infty} B_v G^v(z) \tag{3.16}$$

$$K = K(1) = K(1,0) = \sum_{v=0}^{\infty} B_v G^v \tag{3.17}$$

Therefore K would be:

$$K = -D_0^{-1}[D[G] - D_0] = I - D_0^{-1}D[G] \tag{3.18}$$

The matrix $D[G]$ is considered to be the infinitesimal generator of a Markov process during the busy period and can be obtained by excising the busy period. Having defined all the necessary values and calculations finally x_0 can be calculated as:

$$x_0 = \frac{\boldsymbol{\mathcal{K}}}{\boldsymbol{\mathcal{K}}\boldsymbol{\mathcal{K}}^*} \tag{3.19}$$

where $\boldsymbol{\mathcal{K}}^*$ is obtained from:

$$\boldsymbol{\mathcal{K}}^* = -D_0^{-1}[DD - [G] + dg][I - A + (e - \boldsymbol{\beta})g]^{-1}e \tag{3.20}$$

The vector $\boldsymbol{\beta}$ whose j-th component is the conditional number of arrivals during a service which starts the arrival process in phase j is explicitly defined as:

$$\boldsymbol{\beta} = \left(\mu'/\lambda'\right)e + (A - I)(e\pi + D)^{-1}d \tag{3.21}$$

For all queues of type M/G/1, the traffic intensity, ρ, can be calculated by $\rho = \pi\boldsymbol{\beta}$ which is also calculate as $\rho = {\mu'}/{\lambda'}$, as expected.

Now we can define the calculations of the moments of queue length at departures and use them to calculate the moments of queue length at arbitrary times for the BMAP/M/1 queue.

The factorial moment vectors of the queue length at departures are given by the quantities $X^{(n)}(1)$, where:

$$X(1) = \pi + -x_0 D_0^{-1} DA(I - A + e\pi)^{-1} \quad (3.22)$$

The final expressions for calculating the first and second moments of the queue length at departures are presented as:

$$X^{(1)}e = \frac{1}{2(1-\rho)}\{XA^{(2)}e + U^{(2)}e + 2\{U^{(1)} - X[I - A^{(1)}]\}(I - A + e\pi)^{-1}\beta\} \quad (3.23)$$

And

$$X^{(2)}e = \frac{1}{3(1-\rho)}\{3X^{(1)}A^{(2)}e + XA^{(3)}e + U^{(3)}e + 3\{U^{(2)} + XA^{(2)} - 2X^{(1)}[I - A^{(1)}]\}(I - A + e\pi)^{-1}\beta\} \quad (3.24)$$

In the above equations, $U(z)$ is defined as $U(z) = -x_0 D_0^{-1} D(z)A(z)$. The derivatives of $X(1) = X$ are written as $X^{(i)} = X^{(i)}(1)$. Also the derivatives of $U^{(i)} = U^{(i)}(1)$ and $A^{(i)} = A^{(i)}(1)$.

3.4. Moments of the Queue Length at Arbitrary Time

The expressions for moments of the queue length at an arbitrary time can be obtained by differentiating the following equation [71]:

$$Y^{(1)}e = X^{(1)}e - \frac{1}{2}\lambda'\pi D^{(2)}e + [\lambda'\pi D^{(1)} - X](e\pi + D)^{-1}D^{(1)}e \tag{3.25}$$

The above equation shows the first moment of the queue length at arbitrary time t. The derivatives are written as $Y^{(i)} = Y^{(i)}(1)$ and $D^{(i)} = D^{(i)}(1)$, for $i \geq 0$. The second moment of the queue length at arbitrary time t is presented as:

$$Y^{(2)}e = X^{(2)}e - \lambda'Y^{(1)}D^{(2)}e - \frac{1}{3}\lambda'\pi D^{(3)}e - 2[X^{(1)} - \lambda'Y^{(1)}D^{(1)} - \lambda'\pi D^{(2)}](e\pi + D)^{-1}D^{(1)}e \tag{3.26}$$

3.5. Moments of the Virtual Waiting Time Distribution

To calculate the moments of the actual waiting time it is required to first calculate the moments of the virtual waiting time or workload distribution. The queueing delay, also known as the virtual waiting time or workload, is the length of time a job awaits in a buffer before transmission.

The virtual waiting time distribution is the joint probability that at an arbitrary time the arrival process is in phase j and that a virtual customer arriving at that time waits at most a time x before entering service. The virtual waiting time distribution is shown as [91]:

$$W_v(s) = s(1-\rho)g\left[sI + D\left(\tilde{H}(s)\right)\right]^{-1}, W_v(0) = \pi \tag{3.27}$$

From this we have the following:

$$w_v(s) = s(1-\rho)g\left[sI + D\left(\widetilde{H}(s)\right)\right]^{-1} e \tag{3.28}$$

where e is a column vector of 1s. Eq. (3.28) illustrates the distribution of the virtual waiting time of the $BMAP/M/1$ queue.

The first two moments of the virtual waiting time distribution are calculated using the following formulas:

$$sw(s) + w(s)D\left(\widetilde{H}(s)\right) = sy_0 \tag{3.29}$$

where $y_0 = (1-\rho)g$. To make the equation easier, the following values are defined:

$$V(s) = D\left(\widetilde{H}(s)\right) \tag{3.30}$$

$$w^{(i)} = w^{(i)}(0), \ \ for \ \ i \geq 1 \tag{3.31}$$

$$V^i = V^{(i)}(0), \ for \ \ i \geq 1 \tag{3.32}$$

Then by subsequently differentiating $V(s)$ we have:

$$V^{(1)} = -\mu_1' D^{(1)} \tag{3.33}$$

$$V^{(2)} = (\mu_1')^2 D^{(2)} + \mu_2' D^{(1)} \tag{3.34}$$

$$V^{(3)} = -(\mu_1')^3 D^{(3)} - 3\mu_1'\mu_2' D^{(2)} - \mu_3' D^{(1)} \tag{3.35}$$

where μ_i' is the i-th moment of the service time distribution $H(s)$. $D^{(n)}$ is calculated using the following formula:

$$D^n = \frac{d^n}{dz^n} D(Z)_{\ z=1} \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} D_k, n \geq 0 \tag{3.36}$$

Note that $\pi D^{(1)} e = \lambda_{tot}$, and the following definition holds $v^i = V^{(i)} e$. Now if we successively differentiate Eq. (3.29) then we will have the moments of the virtual waiting time distribution:

$$w^{(1)} = \left(\frac{1}{2(1-\rho)}\right)\left[2\rho + 1\left(y_0 - \pi V^{(1)}\right)(e\pi + D)^{-1} v_1 + \pi v_2\right] \tag{3.37}$$

$$w'(0) = (w^1 e)\pi - \pi + \left(y_0 - \pi V^{(1)}\right)(e\pi + D)^{-1} \tag{3.38}$$

$$w^{(2)} = \left(\frac{1}{3(1-\rho)}\right)\left[3\left(2w'(0) + 2w'(0)V^{(1)} + \pi V^{(2)}\right)(e\pi + D)^{-1} v_1 - 3w'(0)v_2 - \pi v_3\right] \tag{3.39}$$

3.6. Moments of the Actual Waiting Time

The probability that the incoming customer belongs to a batch of size k is [156]:

$$P_{ba}^{(k)} = \frac{\pi k D_k e}{\lambda_{tot}} \tag{3.40}$$

The probability that an arbitrary customer will occupy a particular position, say n-th $(1 \leq n \leq k)$ in a batch of size k is $1/k$.

The first moment of the actual waiting time of the first customer in the batch of size k is:

$$E[W_k] = \frac{-w^1 D_k e}{\pi D_k e} \tag{3.41}$$

This makes it possible to calculate the first moment of the actual waiting time of an arbitrary customer:

$$E[W] = \sum_{k=1}^{\infty} P_{ba}^{(k)} \left[E[W_k] + \mu_1' \left(\frac{k-1}{2} \right) \right] \tag{3.42}$$

And the second moment of the actual waiting time for an arbitrary customer is calculated as follows:

$$E[W_k^2] = \frac{w_2 D_k e}{\pi D_k e} \tag{3.43}$$

$$w_2 = E[W_k^2]\pi - \pi V^{(2)} + 2w_1 \left(I + V^{(1)} \right) (e\pi + D)^{-1} \tag{3.44}$$

$$E[W^2] = \sum_{k=1}^{\infty} P_{ba}^{(k)} \left[E\left[{W_k}^2 \right] + \mu_2' \left(\frac{k-1}{2} \right) + (k-1) E[W_k] \mu_1' + \frac{(k-1)!}{(k-3)!} \frac{{\mu_1'}^2}{3} \right] \tag{3.45}$$

3.7. Special Cases of the Simplified BMAP/G/1 Queuing System

Based on the arrival processes and the number of state the underlying Markov chain can have, special cases of BMAP can be obtained. Some of the most useful and famous examples are listed below.

3.7.1. The MAP/G/1 Queueing Systems

If all arrival batches of BMAP are of size one, then the BMAP would be Markovian Arrival Process (MAP), Figures 3.1 and 3.2. Therefore it is obvious that $D_k = 0$, for $k > 2$ [65]. From this class there are many well-known arrival processes:

- *Poisson process:*

The conventional Poisson process is a special case of the BMAP process with rate λ, where $D_0 = -\lambda$, $D_1 = \lambda$ and $D_k = 0$, for all $k \geq 2$.

- *PH-renewal Process:*

  The Phase type (PH) renewal process first introduced by Neuts [157] with representation (α, T), is considered to be a $BMAP$ with $D_0 = T$, $D_1 = -Te\alpha$ and $D_k = 0$, for all $k \geq 2$.

  This class of processes also contains the Erlang, E_k, and hyperexponential, H_k, arrival processes as well as finite mixtures of these.

- *Markov-Modulated Poisson Process (MMPP):*

  MMPP is a doubly stochastic Poisson process whose arrival rate is governed by $\hat{\lambda}[J(t)] \geq 0$, where $J(t)$, $t \geq 0$, is and m-state irreducible Markov process. As a result, the arrival rate only takes m values of $\lambda_1, \dots, \lambda_m$, depending on the state of the Markov process. If $\Lambda = diag(\lambda_1, \dots, \lambda_m)$, and the underlying Markov process has the infinitesimal generator of R, then the MMPP is a special case of BMAP where $D_0 = R - \Lambda$, $D_1 = \Lambda$ and $D_k = 0$, for $k \geq 2$, Figures 3.3 and 3.4.

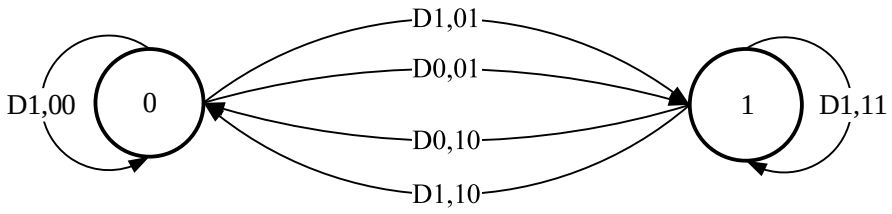


Figure 3.1: Two-state Continuous Time Markov Chain underlying a MAP.

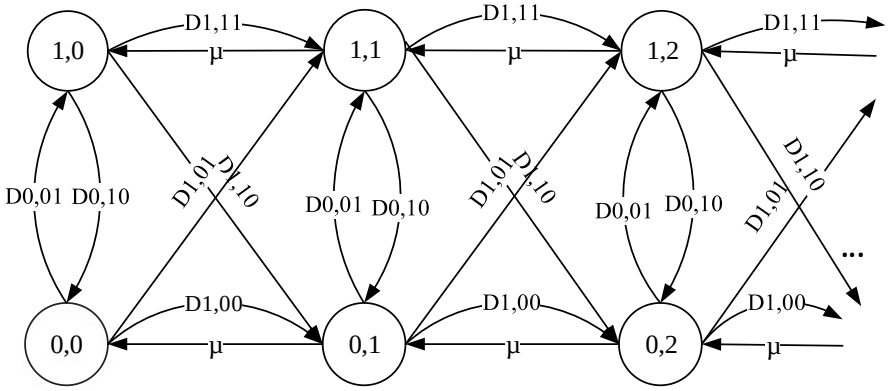
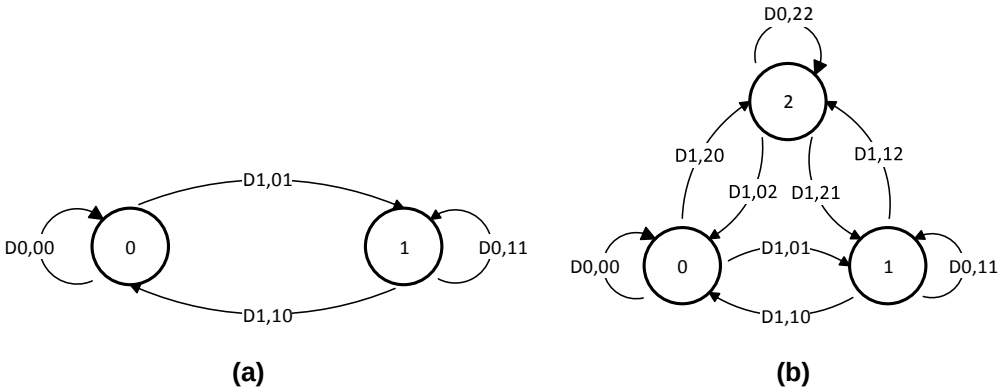


Figure 3.2: State transition diagram of a MAP/M/1 queue with two-state MAP.



(a) **(b)**

Figure 3.3: Two-state (a) and three-state (b) Continuous Time Markov Chain underlying a MMPP.

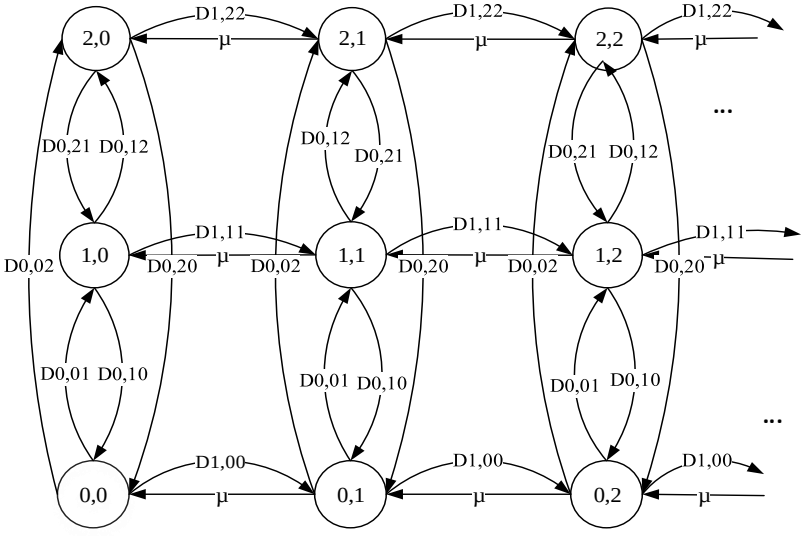


Figure 3.4: State transition diagram of a MMPP/M/1 queue with a three-state MMPP.

3.7.2. Superposition of BMAP's

The class of BMAP's is closed under superposition. The superposition of n independent BMAP's with representations $\{D_k(i)\}, 1 \leq i \leq n$, is also a BMAP with:

$$D_k = D_k(1) \oplus \ldots \oplus D_k(n), k \geq 1 \tag{3.46}$$

The $\oplus$ sign denotes the matrix Kronecker sum [90].

3.8. Analysis of MMPP/M/1 Queuing Systems

As mentioned before BMAP is a generalization of MAP and, MMPP is a special case of MAP where the transition from one state to another does not ignite any arrival. To model the MMPP/M/1 queue in the same format as BMAP/M/1, the following should hold [46]:

$$D_0 = R - \Lambda \text{ and } D_1 = \Lambda \tag{3.47}$$

And $D_k = 0$, for $k \geq 2$.

The 2-state MMPP has received a lot of attention as a simple tractable process which can predict queuing delays very accurately [38]. The equations used for BMAP are reduced to very simple forms for the 2-state MMPP [85]:

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_1 \\ \sigma_2 & -\sigma_2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \quad G = \begin{bmatrix} 1 - G_1 & G_1 \\ G_2 & 1 - G_2 \end{bmatrix} \tag{3.48}$$

where Q presents the infinitesimal generator, Λ represents the arrival rate matrix of each of the states of the underlying markov chain and G shows the structure of the famous G matrix in a 2-state MMPP/M/1 queue.

At the beginning, the value of G_1 should be set to 0. Then recursively the following formulas should be calculated until G_1 and G_2 become stable, the value for G_1 and G_2 are calculated using the following formulas:

$$G_2 = \frac{G_1\sigma_1}{\sigma_1+G_1(\lambda_1-\lambda_2)} \tag{3.49}$$

$$G_1 = 1 - G_2 - H(\sigma_1 + \sigma_2 + \lambda_1 G_1 + \lambda_2 G_2) \tag{3.50}$$

H is the Laplace Transform of the service distribution of the queue under study. From the results of the above equations the value of g for a two state MMPP can be calculated directly as:

$$g = (g_1, g_2) = \frac{1}{G_1+G_2}(G_1, G_2) \tag{3.51}$$

After implementing the analytical model for the specific 2-state MMPP, the model for the general m-state MMPP [85] based on the methods presented in this section was developed for later comparison with BMAP queue. The model for m-sate MMPP is very close to BMAP in method of implementation.

3.9. Simulation of BMAP/M/1 Queuing Systems

An important field of study in the area of computing and communication is development of simulation studies, which form an integral part in the performance

evaluation of computer and communications systems. Simulation is an important method of assisting teletraffic engineers in the design and development of communication networks that enforce Quality of Service (QoS) objectives and improve the cost and performance of such systems [158]. In this regard, traffic modelling which deals with the issue of characterizing the randomness and nature of traffic generated by end users and applications in networks, is the main priority as understanding the nature of traffic in communication systems and selecting the appropriate processes to model it, is crucial in the success of the whole simulation study. An extremely important aspect to focus on when dealing with communication traffic is capturing the correlation characteristics of the traffic traces [37, 46]. This is only possible through developing an accurate analytical model that is capable of capturing every detail and characteristic of real world traffic.

For this purpose as part of this research new traffic generators are developed in C++ programming language that accommodate the properties of Batch Markovian Arrival Process as an input for the simulation of BMAP/M/1/N queue, and m-state MMPP process as an input for the simulation of MMPP/M/1/N queue. The traffic generator developed for m-state MMPP is a generalization of the simulator for a two state $MMPP$ which is very well studied in literature.

The traffic generator developed for BMAP is able to simulate traffics generated based on the underlying Continuous Time Markov Chain of m-state BMAPs with batch size of maximum K in accordance to the behavior of the processes. To start the simulation, the initial state $i_0 \, \epsilon \, S = \{1,2, \dots, M\}$ is randomly selected according to the initial state probability distribution vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_M)$ of the underlying Markov chain of the M-state BMAP with batch size of maximum K. The transitions

between the states of the Markov Chain are done randomly based on the state transition rate probabilities of the BMAP. For example, at the end of an exponentially distributed sojourn time in state i, with mean arrival rate of $\lambda'_i = {}^{-1}/_{D_0(i,i)}$, M possible state transitions can occur, as explained in previous sections. The phase transition matrix during an inter-batch arrival time of jobs accepted in the system is developed as:

$$\overline{D} = (-D_0)^{-1} D_k, \; k \geq 1 \tag{3.52}$$

Having developed the generator of the arrival traffic, the simulations of the queues are executed using Discrete Event Simulation (DES). The simulation scenario for each queuing model is executed for large number of jobs (e.g. 100000) so that the final calculated performance results are generated when the queue has reached a steady state. At this stage all queuing models under study have unlimited buffer sizes and for the purpose of comparison, the outputs of two important performance measures are calculated for each model which are: mean queue length at an arbitrary point in time (in number of jobs in the queue), and mean waiting time in queue for an arbitrary customer (in seconds).

3.9.1. Model Validation and Numerical Results

This section presents validation results of the developed analytical model of the BMAP/M/1 and MMPP/M/1 queues via the developed DES simulators under various settings and scenarios.

Presented analytical results provide the basis of creating stable numerical procedures for calculating the desired performance measures in future study of Wireless Local Area Networks. The validation of the numerical results via simulation demonstrates the accuracy and feasibility of the developed BMAP and MMPP traffic generators, which are then used to study the performance of real world networks under bursty and correlated traffics.

The following subsections contain the detailed information of the parameterized settings used for the analytical models and simulations and the comparison of the output results.

- **Scenario 1: Validation of the Developed Simulators**

To test and validate the simulators and developed analytical models for accuracy and reliability for use in future studies, for the first step a 2-state MMPP/M/1 queue is studied in the format of a simplified m-state BMAP/M/1 with batch size of maximum 1. To validate the developed model and simulator, two methods are used: I) the first method develops the model and simulator of the MMPP/M/1 queue using the simplified analysis of a 2-state MMPP/M/1 queue presented in section 3.9, and II) the second method develops the analytical model and simulator for a 2-state BMAP/M/1 with batch size of maximum 1, which also resembles a 2-state MMPP/M/1 queue. Since the first method has been thoroughly used and tested in previous literature [76, 83, 84, 146], it can act as a safe comparison platform for the validity of the models and DES simulators developed for BMAP of any number of states and batch sizes.

The simulation and analytical model of both methods are executed for different values of overall arrival rates so that the two performance measures of mean queue length and mean waiting time in the queue can be compared for different traffic intensities. The traffic intensity varies from a very small value close to 0.001 up to maximum 1, and for stability of the models during all studies the value of traffic intensity is always kept to less than 1.

For the flexibility and scalability of the simulation experiments, the parameters of the MMPP/M/1 queue are accurately fitted to real-world multimedia applications. The parameters of the queue are obtained from the high-quality measurement of the multimedia application of the G.711 codec voice sources as stated and used in [76], which are as follows:

$$Q = \begin{bmatrix} -0.2 & 0.2 \\ 0.8 & -0.8 \end{bmatrix} \qquad \Lambda = \begin{bmatrix} 6\lambda_2 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

The mean service rate is set to $\mu = 10.0$:

Similar parameters but in the form usable by BMAP/M/1 queue are used for the 2-state MMPP/M/1 queue modelled and simulated as special case of BMAP:

$$D_0 = Q - \Lambda = \begin{bmatrix} -0.2 - 6\lambda_2 & 0.2 \\ 0.8 & -0.8 - \lambda_2 \end{bmatrix} \qquad D_1 = \Lambda = \begin{bmatrix} 6\lambda_2 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

The results from executing the models and the simulations are depicted in Figures 3.5 and 3.6. The figures show a great match between the analytical models and simulations of 2-state MMPP/M/1 queue using the two different introduced methods of implementations. It is clear from the figures that as the traffic intensity of the queue increase as a result of increasing the overall load, the values of the

mean queue length and mean waiting time also gradually increase. This process continues until when the traffic intensity gets closer to 1, at which stage the queue starts to become saturated meaning that the arrival rate of the customers or packets coming in to the queue becomes higher than the service rate and as a result the queueing system starts to over load. This situation in queuing systems is called saturation. Saturation is when any more increase in the load of the queue will have no effect on the performance measures of the queue, and as a result the values of the performance measures stays almost the same from that point onwards.

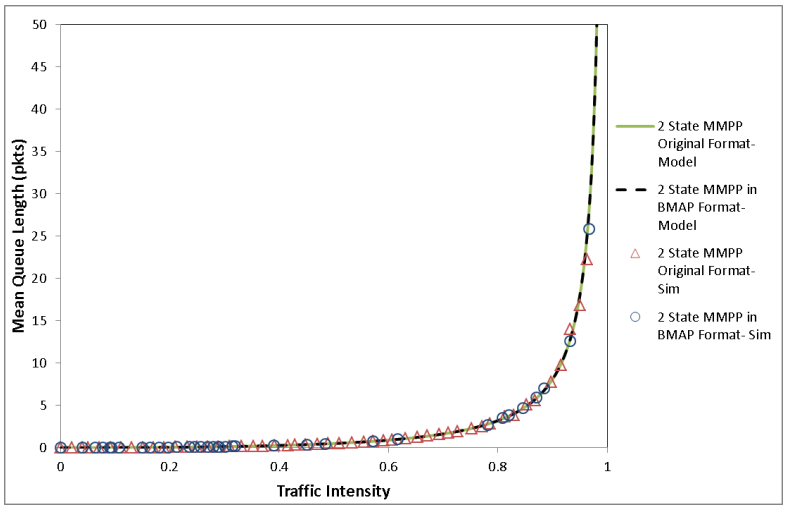


Figure 3.5: Mean queue length in a 2-State MMPP/M/1 queue.

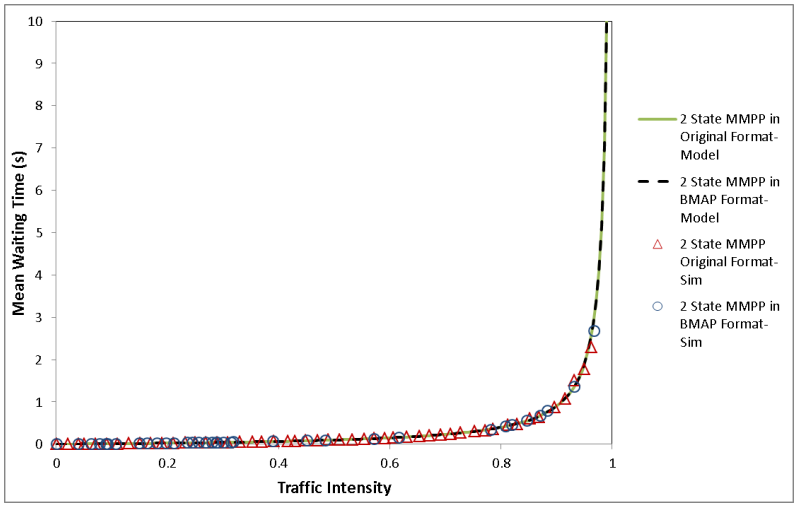


Figure 3.6: Mean waiting time in a 2-State MMPP/M/1 queue.

The great match between the models and simulations of the two methods is a solid proof that the developed simulator and analytical model for m-state BMAP/M/1 queue with any maximum batch size is accurate and reliable to use in future studies.

Since the BMAP/M/1 model and simulator have proven to be accurate, they can be used to test the model and simulator developed for m-state MMPP/M/1 queue. For this purpose, the performance measures of a 5-state MMPP/M/1 queue model and simulation are compared with a 5-state BMAP/M/1 queue with batch size of maximum 1, which also resembles a MMPP/M/1 queue.

The parameterization used for these models and simulations are based on real world multimedia applications adopted using the Expectation-Maximization (EM) method introduced in [102], as discussed previously in section 2.4. The data trace used is obtained from the high quality measurement of the video stream for the film "Tears of Steel", encoded in H.265/HEVC codec [159]. To vary the traffic load of

the queue, the arrival rates of the 5 states of the underlying Markov Chain are varied according to the overall load of the traffic placed on the queue.

The infinitesimal generator matrix of the 5-state MMPP, Q, and the arrival rates of each state for the original analytical model and simulator of MMPP/M/1 queue are set as follows:

$$Q = \begin{bmatrix} -0.3 & 0.08 & 0.06 & 0.10 & 0.06 \\ 0.19 & -0.97 & 0.29 & 0.26 & 0.24 \\ 0.42 & 0.40 & -1.62 & 0.43 & 0.38 \\ 0.64 & 0.60 & 0.51 & -2.30 & 0.56 \\ 0.76 & 0.87 & 0.75 & 0.69 & -3.08 \end{bmatrix}$$

$$\Lambda = \begin{bmatrix} 0.57\lambda_{tot} & 0 & 0 & 0 & 0 \\ 0 & 0.19\lambda_{tot} & 0 & 0 & 0 \\ 0 & 0 & 0.1\lambda_{tot} & 0 & 0 \\ 0 & 0 & 0 & 0.08\lambda_{tot} & 0 \\ 0 & 0 & 0 & 0 & 0.05\lambda_{tot} \end{bmatrix}$$

Similar parameters are used for BMAP/M/1 queue with batch size of maximum one which resembles the 5-state MMPP/M/1 queue. The parameters are used in forms usable by the BMAP model and simulator and they are as follows for mean service rate of $\mu = 10.0$:

$$D_0 = Q - \Lambda = \begin{bmatrix} -0.3 - 0.57\lambda_{tot} & 0.08 & 0.06 & 0.10 & 0.06 \\ 0.19 & -0.97 - 0.19\lambda_{tot} & 0.29 & 0.26 & 0.24 \\ 0.42 & 0.40 & -1.62 - 0.1\lambda_{tot} & 0.43 & 0.38 \\ 0.64 & 0.60 & 0.51 & -2.30 - 0.08\lambda_{tot} & 0.56 \\ 0.76 & 0.87 & 0.75 & 0.69 & -3.08 - 0.05\lambda_{tot} \end{bmatrix}$$

$$D_1 = \Lambda = \begin{bmatrix} 0.57\lambda_{tot} & 0 & 0 & 0 & 0 \\ 0 & 0.19\lambda_{tot} & 0 & 0 & 0 \\ 0 & 0 & 0.1\lambda_{tot} & 0 & 0 \\ 0 & 0 & 0 & 0.08\lambda_{tot} & 0 \\ 0 & 0 & 0 & 0 & 0.05\lambda_{tot} \end{bmatrix}$$

Figures 3.7 and 3.8, respectively depict the mean queue length and mean waiting time for the queues under study.

Once again the results show great accuracy in the results developed using the analytical model and the simulator specifically developed for m-state MMPP/M/1 queue and the model and simulator developed for the study of m-state BMAP/M/1 queue with any batch size.

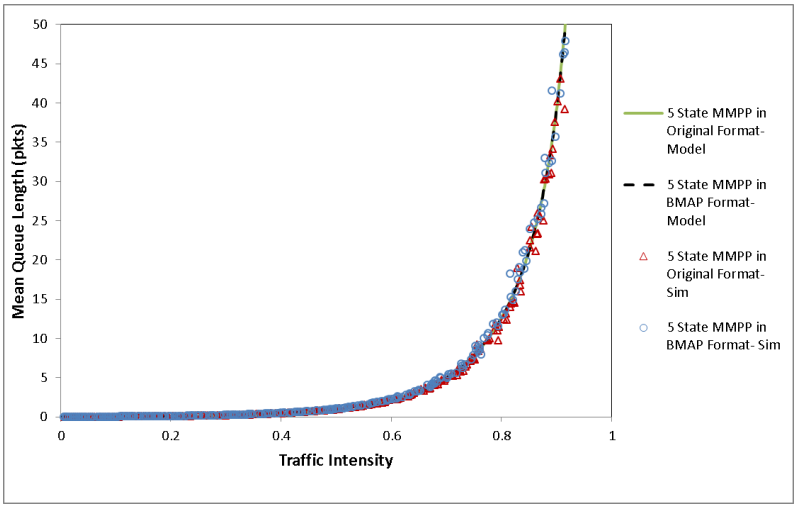


Figure 3.7: Mean queue length in a 5-State MMPP/M/1 queue.

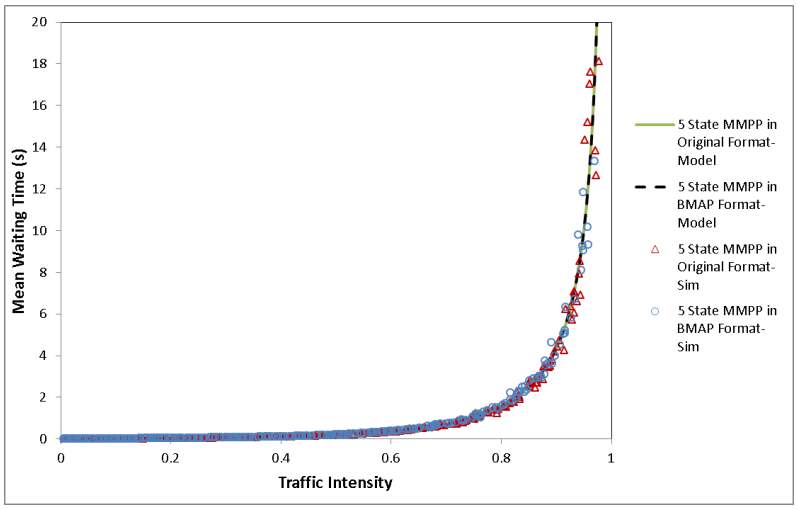


Figure 3.8: Mean waiting time in a 5-State MMPP/M/1 queue.

- **Scenario 2: Effect of Coefficient of Correlation on BMAP/M/1 Queues**

For the second scenario, four different settings of 3-state BMAPs with maximum batch sizes of three are considered and compared to each other for the purpose of studying the effect of burstiness and correlation on the mean queue length and mean waiting time performance measures of queues whose arrival process are modelled using BMAP. For this reason the infinitesimal generator Q, defined for all four BMAP settings are kept the same and only the proportion of the arrival rates at different states of the underlying Markov Chain are varied in accordance to the total load λ_{tot} placed on the queue during each run. Also, the D_0 matrix is defined the same for all four BMAPs, however, as explained, the diagonal elements of the D_0 matrix which represent the arrival rates of the states of the underlying Markov Chain, will vary in each of the BMAPs. This will result in different correlation intensities between the interarrival times in each case under study, which will help to analyse the effect of correlation and burstiness on the performance of the queues. Also, as a result of varying the total queue load during each run, it will be possible to compare the performance measures using the value of traffic intensity (ρ) which varies from a small value close to zero to maximum of one.

The parameterization used for the BMAPs is based on the values calculated and used in [98], therefore the infinitesimal generator Q and the matrix D_0, are defined as:

$$Q = \begin{bmatrix} -0.247533 & 0.156836 & 0.090698 \\ 0.123767 & -0.247533 & 0.123767 \\ 0.090698 & 0.103926 & -0.194622 \end{bmatrix} \quad D_0 = \begin{bmatrix} -\lambda_1 & 0.090698 & 0.090698 \\ 0.090698 & -\lambda_2 & 0.090698 \\ 0.090698 & 0.090698 & -\lambda_3 \end{bmatrix}$$

The BMAPs under study are named as $BMAP(1)$, $BMAP(2)$, $BMAP(3)$ and $BMAP(4)$.

For $BMAP(1)$ the relationship between the arrival rates of the three states and the fundamental arrival rate are defined as: $\lambda_1 = 2.0 \times \lambda_{tot}$, $\lambda_2 = 0.9 \times \lambda_1$, $\lambda_3 = 1.0 \times \lambda_1$, and the calculated average correlation coefficient for $BMAP(1)$ at $lag(1)$ is $\mathrm{c_{cor}} = 0.002$.

The relationship between the arrival rates of the three states of $BMAP(2)$ with the fundamental arrival rate is defined as: $\lambda_1 = 3.0 \times \lambda_{tot}$, $\lambda_2 = 1.0 \times \lambda_1$, $\lambda_3 = 5.0 \times \lambda_1$, the calculated average correlation coefficient for $BMAP(2)$ $lag(1)$ is $\mathrm{c_{cor}} = 0.26$.

For $BMAP(3)$ the relationship between the arrival rates of the three states and the fundamental arrival rate is defined as: $\lambda_1 = 1.0 \times \lambda_{tot}$, $\lambda_2 = 1.0 \times \lambda_1$, $\lambda_3 = 10.0 \times \lambda_1$, and the calculated average correlation coefficient for $BMAP(3)$ $lag(1)$ is: $\mathrm{c_{cor}} = 0.48$.

And finally for $BMAP(4)$, the arrival rates are set as $\lambda_1 = 5.0 \times \lambda_{tot}$, $\lambda_2 = 1.0 \times \lambda_1$, $\lambda_3 = 0.5 \times \lambda_1$, with calculated average correlation coefficient at $lag(1)$ being $\mathrm{c_{cor}} = 0.076$.

For each of the BMAPs, the rest of the matrices of $D_k, 1 \leq k \leq 3$, are calculated using the method introduced in [98] where having D_0 and Q would be used to first calculate the sum of the remaining matrices defined as D. Then using the formula introduced below, the rest of the D_k, $k = \overline{1,3}$ are calculated as follows, where $q = 0.8$:

$$D = Q - D_0 \tag{3.53}$$

$$D_k = Dq^{k-1}(1-q)/(1-q^3) \tag{3.54}$$

The mean service rate of the four examples of BMAP/M/1 queue is set to $\mu = 10.0$. The fundamental arrival rate is increased during the execution of the analytical model and simulation for the purpose of comparing the average waiting time and increase of queue lengths in regards to the increase of the load intensity of the queue.

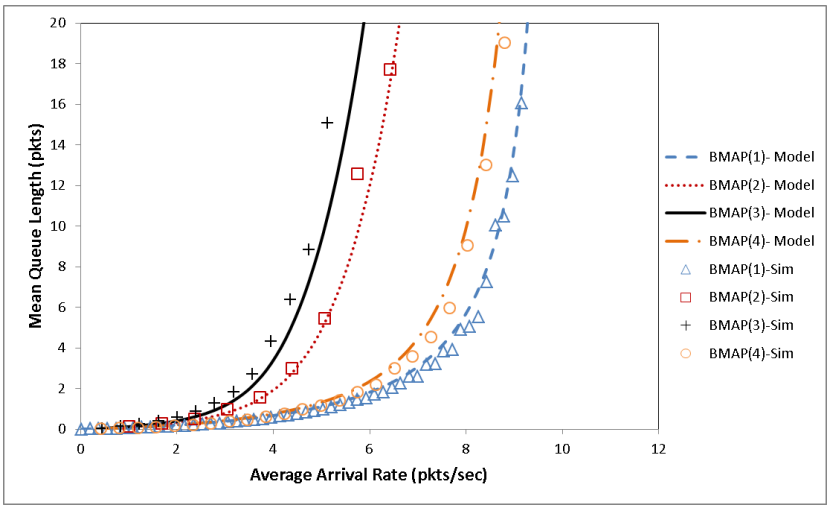


Figure 3.9: Mean queue length in 3-State BMAP/M/1 queue with different average correlation coefficient values.

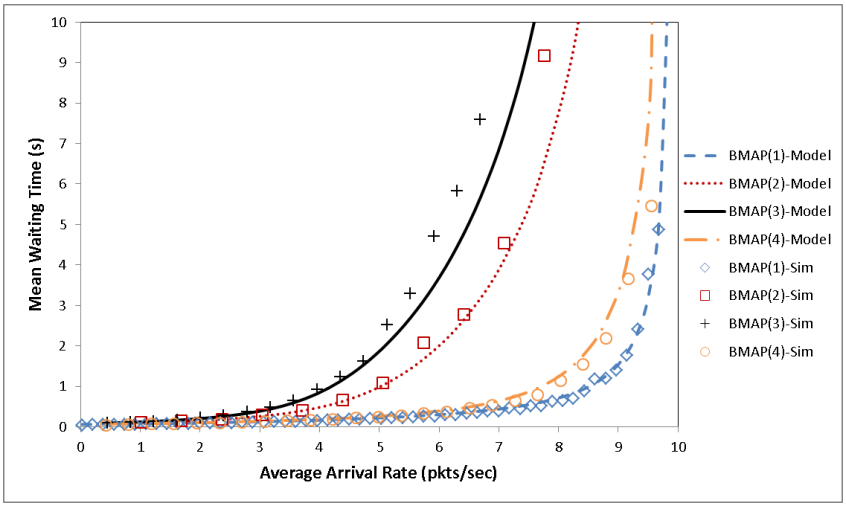


Figure 3.10: Mean Waiting Time in Queue for 3-State BMAP/M/1 queue with different average correlation coefficient values.

Figures 3.9 and 3.10 depict the results of analytical modelling and simulation of the mean queue length and mean waiting time of the BMAP/M/1 queues under various traffic intensities with different auto-correlation coefficients. The results explicitly confirm the fact that the increase in the auto-correlation coefficient of interarrival times of the arrival process increases the burstiness of the incoming traffic which in turn increases the mean waiting time and queue length of the models. In this regard, the mean queue length and mean waiting time of $BMAP(3)$, which on average has the highest auto-correlation coefficient of interarrival times (0.48) at $lag(1)$, reach the saturation level at a point much sooner than the other BMAPs. The mean queue length of this model reaches saturation point when the traffic intensity of the queue is around 4.5 in comparison to $BMAP(1)$ which has the lowest auto-correlation coefficient of interarrival times (0.002) at $lag(1)$. These results show that the auto-correlation coefficient of interarrival times has a great effect on the performance results and dominates the queueing performance.

The increase of the auto-correlation coefficient of interarrival times as the traffic intensity increases on the mean queue length and mean waiting time of the queues are shown in Figures 3.11 and 3.12, respectively. However, for clarity of the figures $BMAP(1)$ results are omitted due to the very small value of the correlation coefficient of interarrival times.

Figure 3.13 shows the relationship between the increase in the traffic intensity and the increase in the correlation coefficient of interarrival times of the queues under study.

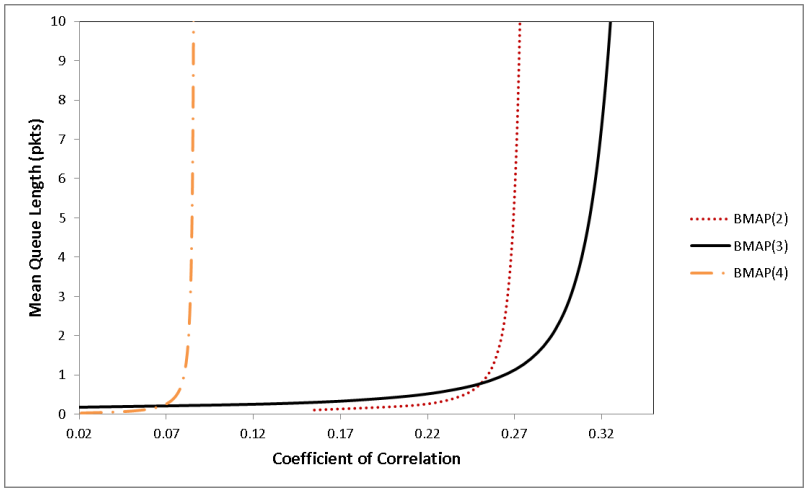


Figure 3.11: Comparison of mean queue length in 3-state BMAP/M/1 queue with correlation coefficient of interarrival times.

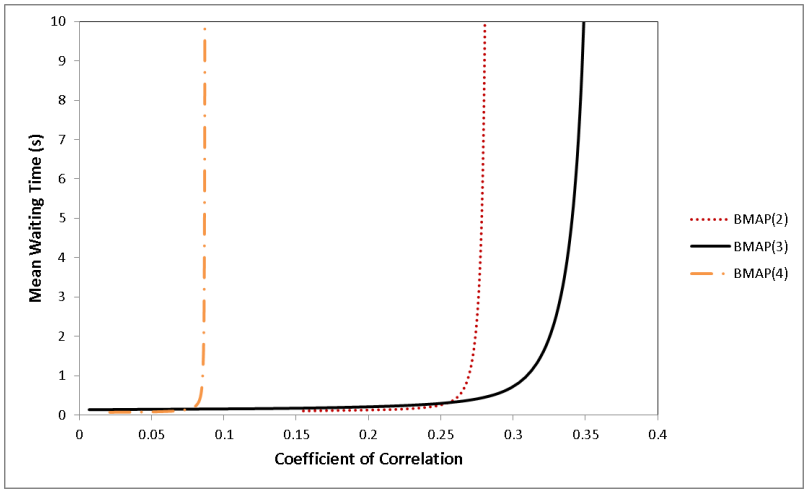


Figure 3.12: Comparison of mean waiting time in 3-State BMAP/M/1 queue with correlation coefficient of interarrival times.

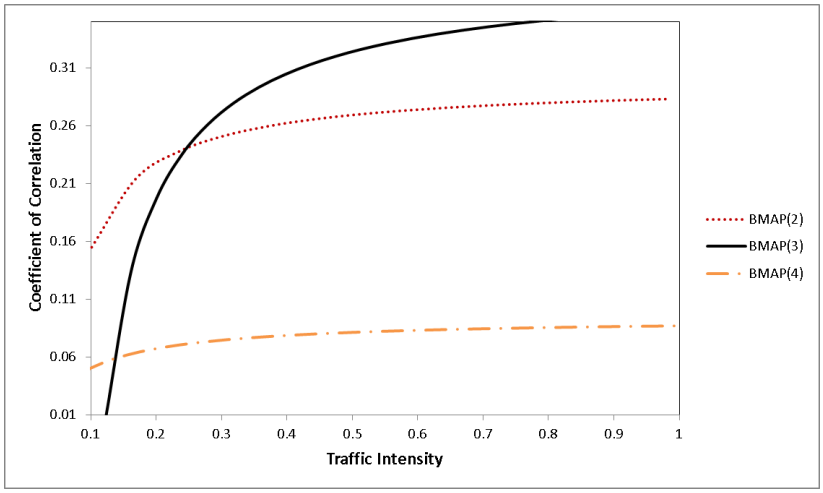


Figure 3.13: Comparison of correlation coefficient of interarrival times in 3-state BMAP/M/1 queues with traffic intensity.

- **Scenario 3: Effect of the Number of States of the Underlying Markov Chain on BMAP/M/1 Queues**

For this scenario, four different BMAP/M/1 queues are considered where the maximum batch sizes are fixed while the number of states of the underlying Markov Chain varies. This provides the basis for the analysis and study of the effect of number of states composing the underlying Markov Chain on the performance of queues whose input is modelled using the BMAP process.

In this section, a set of real data traces are used for developing BMAPs parameters. Again the data traces are obtain from the high quality measurement of the video stream for the film “Tears of Steel”, encoded in H.265/HEVC codec [159]. Utilizing the Expectation Maximization technique as in [102], a set of parameters are estimated such that the resulting model closely matches the statistical properties of the original trace. To be able to compare the BMAPs with different number of states with each other, a single source of data trace should be used. However, just like previous scenarios the overall load on the system and the mean service time are kept the same for all BMAPs at each point of time for comparison purposes. Having estimated the infinitesimal generator and the D_0 matrix using the aforementioned EM algorithm, the rest of the the $D_k, k = \overline{1,3}$ matrices are developed using the same method as in previous scenario via: $D = Q - D_0$ and $D_k = Dq^{k-1}(1-q)/(1-q^3)$, where $q = 0.8$.

The BMAPs under study are named in the form of $BMAP(i,j)$, where i represents the number of states and j represents the maximum batch size. Maximum batch sizes for BMAPs under study are kept to 3 during this scenario. The models and

simulations are studied under varying traffic intensities which is the result of varying arrival rates. The infinitesimal generator and matrix D_0 of each of the BMAPs are as follows:

- BMAP(2,3):

$$Q = \begin{bmatrix} -0.086 & 0.086 \\ 0.549 & -0.549 \end{bmatrix} \qquad D_0 = \begin{bmatrix} -0.86 * \lambda_{tot} & 0.022 \\ 0.174 & 0.14 * \lambda_{tot} \end{bmatrix}$$

- BMAP(3,3):

$$Q = \begin{bmatrix} -0.158 & 0.096 & 0.063 \\ 0.4 & -0.76 & 0.366 \\ 0.66 & 0.765 & -1.43 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} -0.75 * \lambda_{tot} & 0.021 & 0.012 \\ 0.11 & -0.17 * \lambda_{tot} & 0.054 \\ 0.099 & 0.19 & -0.077 * \lambda_{tot} \end{bmatrix}$$

- BMAP(4,3):

$$Q = \begin{bmatrix} -0.27 & 0.09 & 0.08 & 0.08 \\ 0.28 & -0.81 & 0.24 & 0.29 \\ 0.38 & 0.40 & -1.40 & 0.61 \\ 0.81 & 0.54 & 0.72 & -2.1 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} -0.63 * \lambda_{tot} & 0.02 & 0.02 & 0.03 \\ 0.059 & -0.18 * \lambda_{tot} & 0.07 & 0.05 \\ 0.095 & 0.14 & -0.11 * \lambda_{tot} & 0.17 \\ 0.252 & 0.13 & 0.12 & -0.08 * \lambda_{tot} \end{bmatrix}$$

- BMAP(6,3):

$$Q = \begin{bmatrix} -0.39 & 0.06 & 0.08 & 0.07 & 0.09 & 0.09 \\ 0.25 & -1.1 & 0.21 & 0.23 & 0.16 & 0.25 \\ 0.39 & 0.33 & -1.8 & 0.37 & 0.36 & 0.36 \\ 0.37 & 0.45 & 0.44 & -2.38 & 0.46 & 0.66 \\ 0.71 & 0.57 & 0.55 & 0.68 & -3.1 & 0.59 \\ 0.77 & 0.71 & 0.61 & 0.62 & 0.70 & 3.41 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} -0.53 * \lambda_{tot} & 0.01 & 0.02 & 0.02 & 0.02 & 0.02 \\ 0.06 & -0.16 * \lambda_{tot} & 0.05 & 0.08 & 0.06 & 0.07 \\ 0.07 & 0.05 & -0.1 * \lambda_{tot} & 0.09 & 0.11 & 0.06 \\ 0.07 & 0.06 & 0.15 & -0.08 * \lambda_{tot} & 0.13 & 0.14 \\ 0.17 & 0.11 & 0.21 & 0.16 & -0.06 * \lambda_{tot} & 0.22 \\ 0.21 & 0.24 & 0.16 & 0.10 & 0.19 & -0.06 * \lambda_{tot} \end{bmatrix}$$

Figures 3.14 and 3.15 show the results for mean queue length and mean waiting time of the four BMAPs with varying number of states. The results clearly indicate a direct relationship between the increase in the number of the states of the underlying Markov Chain of the BMAP process and the increase in the saturation speed of the queues. When the number of the states is equal to 6, the increasing speed of the mean queue length is faster than when the number of states is less, e.g. 3. The same stands for the mean waiting time; smaller Markov Chain results in the queue reaching the saturation state in a lower speed.

From the outcome of the study it can be concluded that increase in the number of the states of the underlying Markov Chain increases the burstiness of the traffic generated by the BMAP process, and therefore affects the performance of the queue through increasing the speed towards reaching a saturation point.

- **Scenario 4: Effect of Batch Size on BMAP/M/1 Queues**

For this part of studying queues with BMAP as input process, the effect of increase in the variety of batch sizes on the performance of the queue is considered. For this reason, the number of states of the underlying Markov Chain is kept the same for all models and throughout all points of time during the simulations for each run.

However during different runs, the possible maximum batch size is increased. The BMAPs are modelled with a 3 state Markov Chain and are shown as $BMAP(3, j)$, where j resembles the maximum batch size acceptable for each BMAP. The maximum batch sizes considered are $2, 5$ and 10.

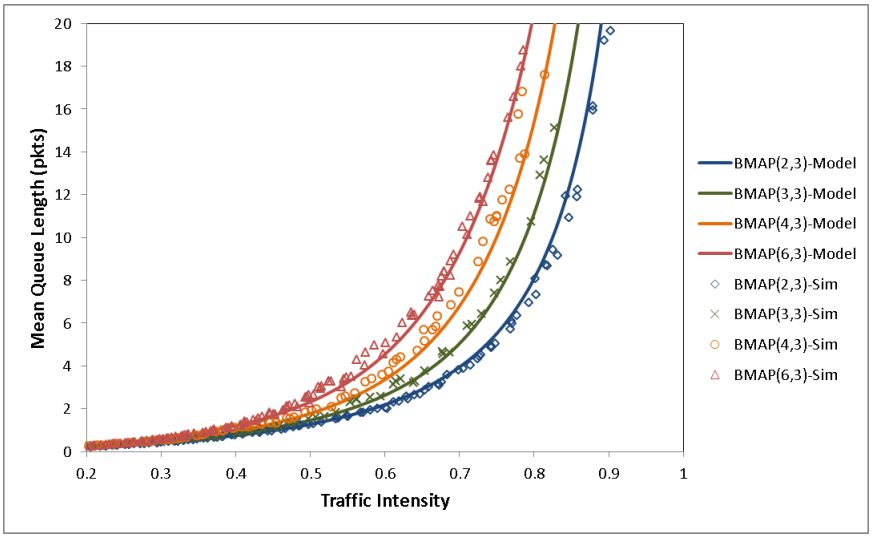


Figure 3.14: Mean queue length for BMAP/M/1 queue with variable number of states in Markov Chain with maximum batch size of 3.

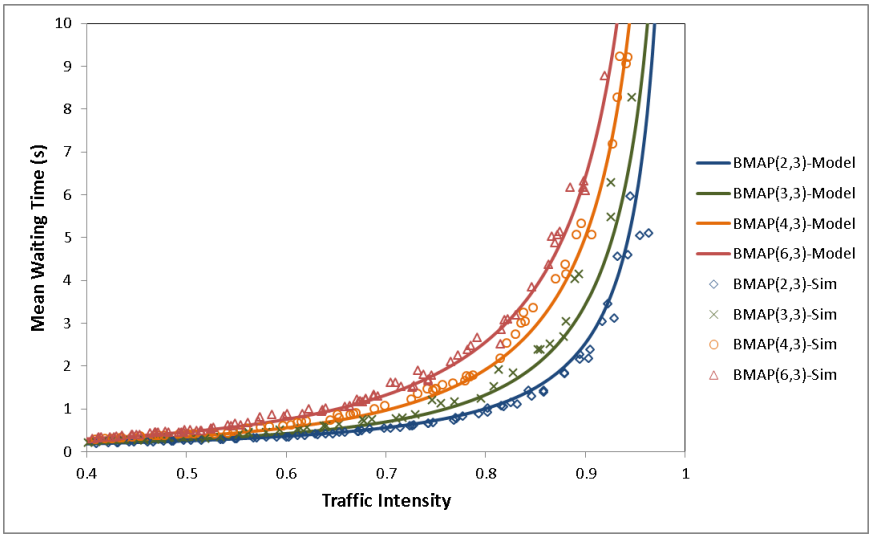


Figure 3.15: Mean waiting time in queue for BMAP/M/1 queues with variable number of states in Markov Chain with maximum batch size of 3.

The parameterization used for the BMAPs is based on the values calculated and used within [98], therefore the infinitesimal generator Q and the matrix D_0, are defined as:

$$Q = \begin{bmatrix} -0.247533 & 0.156836 & 0.090698 \\ 0.123767 & -0.247533 & 0.123767 \\ 0.090698 & 0.103926 & -0.194622 \end{bmatrix}$$

$$D_0 = \begin{bmatrix} -0.3 * \lambda_{tot} & 0.090698 & 0.090698 \\ 0.090698 & -0.34 * \lambda_{tot} & 0.090698 \\ 0.090698 & 0.090698 & -0.36 * \lambda_3 \end{bmatrix}$$

As in previous scenarios, the rest of the matrices of each BMAP, $D_k, 1 \leq k \leq K_{max}$, are calculated using the method introduced in [98] where having D_0 and Q would help to calculate the sum of the remaining matrices defined as D, and then using the introduced formulas in Eq. (3.53) and Eq. (3.54) the rest of the $D_k, k = \overline{1, K_{max}}$ are calculated. However, Eq. (3.54) should be updated each time according to the maximum possible batch size as: $D_k = Dq^{k-1}(1-q)/(1-q^{K_{max}})$, with $q = 0.8$. The mean service rate of all BMAP/M/1 queues is set to $\mu = 10.0$.

Figures 3.16 and 3.17 show the effect of increased maximum batch size on the performance of queues modelled using a 3-state BMAP as an arrival process. From the results it is clear that increase in the possible maximum batch size of the BMAP process increases the speed by which the queue reaches the saturation point. Three different BMAP processes are considered for the same queue with different maximum batch sizes of 2, 5 and 10. The mean queue length and mean waiting times in the BMAP/M/1 queue with maximum batch size of 10 are overall higher than the other two queues. Moreover the performance measures of this queue reaches saturation point a lot quicker than the queues with maximum batch

sizes of 2 and 5. The results for BMAP/M/1 queue with maximum batch size of 5 also show higher performance measures compared to the queue with maximum batch size of 2. However it is important to note that the increase in maximum batch size does not significantly affect the correlation of the interarrival times of the incoming batches to the queue. Instead, the increase in maximum possible batch size for BMAP increases the burstiness of the traffic generated and so the queue reaches the saturation point a lot sooner than when the maximum possible batch size is smaller.

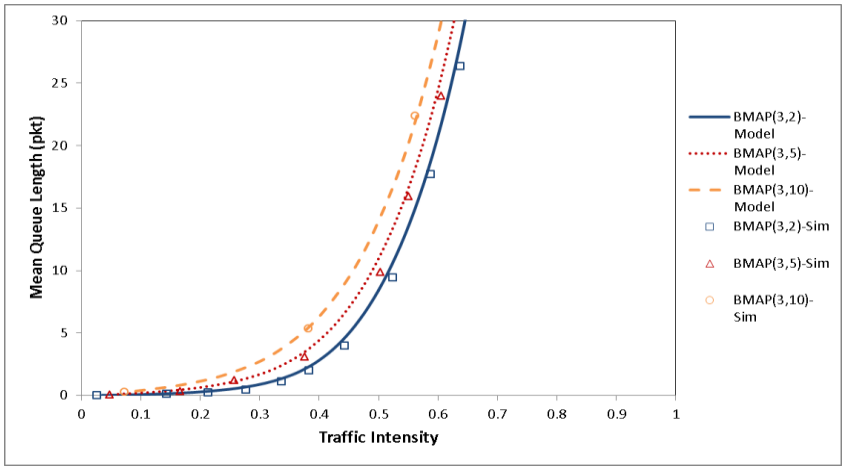


Figure 3.16: Mean queue length for 3-State BMAP/M/1 queues with varying maximum batch size.

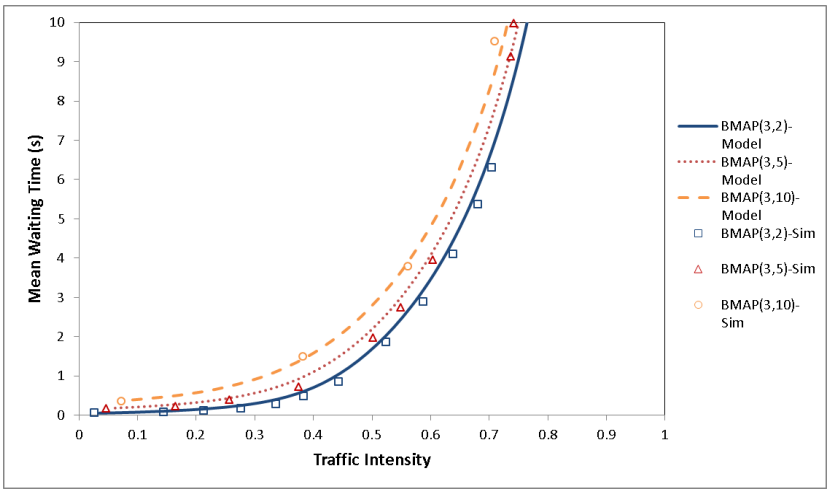


Figure 3.17: Mean waiting time in queue for 3-State BMAP/M/1 queues with varying maximum batch size.

3.10. Summary

This chapter concentrated on the study of MMPP/M/1 queue and mainly BMAP/M/1 queue. The properties of these queues were studied with the aim of developing models and simulators that would form the basis of future studies. It was explained that the G matrix plays and important role in the analysis of BMAP/M/1 queue and calculations of its important performance measures, as it captures the characteristics of the queue during the busy period. A general numerical procedure was presented based on the work of Lucantoni in [71], for developing an analytical solution for the performance measures of the BMAP/M/1 queue. Below is a summary of the main steps of this procedure:

1. Compute matrix A through summation of the $\{A_n\}$ matrices which are the starting point of a long series of numerical computations, which should be computed to high level of accuracy.

2. Compute the G matrix through generation of a sequence of non-negative matrices which increase monotonically to the unique solution required for G. Once G is calculated up to the desired accuracy, the stationary probability vector of g can be generated.

3. Calculate the value of the desired vector β and vector μ. This is the stage at which a powerful accuracy check can be carried out through verification of equation: $g\mu = (1-\rho)^{-1}$, using the computed estimates of g and μ.

4. Calculate matrix $D[G]$ using the famous Horner's method. With the computation of the two important values of $\mathcal{K}$ and K, vector K^* can be gained which can then be used to calculate the x_0 vector.

5. At this stage, all required parameters for calculating the queue length at departures are computed, therefore the first two moments of the queue length distribution at departures ($X^{(1)}(1)e$ and $X^{(2)}(1)e$) can be calculated explicitly in terms of x_0.

6. To calculate the moments of the queue length at arbitrary times, the value of y_0 should be calculated which is then used for the calculation of $Y^{(1)}(1)e$ and $Y^{(2)}(1)e$, the first and second moments of the queue length at arbitrary times. Also the moments of the virtual waiting time distribution can be calculated using y_0, from which the moments of the actual waiting times can then be calculated.

With completion of the analytical study of the BMAP/M/1 queue, the rest of the chapter focused on the study of the queue under different settings. The aim was to 1) validate the developed models and simulators for m-state BMAP/M/1 queue with various maximum batch sizes and m-state MMPP/M/1 queues, 2) to understand the effect of different settings of maximum number of states or batch sizes of the BMAP process on the performance measures of the BMAP/M/1 queue, 3) to study the effect of different batch sizes and number of states on the correlation coefficient of interarrival times and burstiness of the generated traffic using the BMAP process.

From the study of the models and simulations it was noted that even though the increase in the traffic intensity does increase the performance measures of mean waiting time and mean queue length until the queue becomes saturated, different settings have important effects on the speed towards saturation. For example when the BMAP process has higher number of states in the underlying Markov Chain or it has the possibility of having bigger batch sizes, the increase in the mean waiting time and mean queue length towards saturation is faster. It is safe to say that any increase in the maximum batch size or number of states in the underlying Markov Chain increases the burstiness of the traffic and results in higher values of mean waiting time and mean queue length.

At times when the number of states of the underlying Markov Chain and the maximum possible batch size are the same, the increase in the correlation coefficient of interarrival times of the BMAP process increases the burstiness of the generated traffic. In particular the gained results illustrate the dependence of the waiting time of packets and queue length on burstiness, correlation and variation of packet arrival rates.

The results clearly prove that BMAP is a versatile process that can greatly capture the burstiness and correlation of interarrival times and packet size distributions of the multimedia traffic under various settings. Finally, the results also illustrate the fact that BMAP/G/1 queues can be greatly used in modelling wireless communication networks where the network parameters can constantly change, e.g., due to unreliability of some elements of the network, imperfect channels, mobility of the stations, etc.

Chapter 4:

Modelling of Wireless Local Area Networks under Bursty Traffic

4.1. Introduction

Burstiness and self-similarity are two most important characteristics of multimedia traffic generated by various applications in computer networks (compressed video, voice, file transfer and etc.). The two properties play a critical role in determining efficient network design as there is a strong relationship between burstiness and correlation within multimedia traffic, since strong positive correlations are major causes of burstiness.

The weaknesses of the existing analytical models developed for the IEEE 802.11 MAC DCF scheme reported in the current literature are threefold:

1- While existing traffics within networks have the above described properties, many studies on the performance analysis of WLANs still use processes that do not embrace correlation or burstiness characteristics of network traffic, such as the Poisson Process or MMPP.

2- Most existing analytical models reported in the current literature on 802.11 MAC DCF, primarily focus on the analysis of system throughput and access delay, and do not consider other important QoS performance metrics, such as end-to-end delay and energy consumption.

3- Also, the analytical models developed for the MAC DCF scheme are mainly based on unrealistic network assumptions such as saturated stations and infinite transmission buffers.

As a result, this chapter presents the fundamental methodology and components to develop an analytical model for the analysis of WLANs under practical and realistic working conditions. The transmission queue at each station is modelled as a BMAP/G/1/N queuing system where the arrival traffic follows a Batch Markovian Arrival Process with the average arrival rate of λ_{tot} (frames/second), to support multimedia applications including real-time and non-real-time video and voice. The BMAP process enables the modelling of the batch arrivals, variance of interarrival times, correlation of interarrival times, and many other subtle characteristics of the arrival process (e.g. the correlation between the local intensity of arrivals of batches with the size of an arriving batch), which are property of utmost importance in the queueing performance of multimedia traffic. Also BMAP has a unique advantage of combining great complexity and modelling capabilities with the analytical tractability which makes it a great solution for modelling multimedia traffic in WLANs.

The service time of the queueing system is defined as the time interval from the instant that a frame starts contending for the channel to the instant that either the data is acknowledged following successful transmission or data is terminated due to transmission failure. The service time is calculated through modelling the backoff procedure of the frame transmission scheme under unsaturated conditions. The final developed model can be used in real-time traffic control schemes in network elements in order to predict congestion and QoS.

The rest of this chapter is organized as follows: section 4.2 presents the analytical model for the MAC DCF mechanism with the assumption of un-saturated network conditions for all sources. Section 4.3 presents the analytical model for calculation of waiting time, queue length, loss probability and energy consumption of nodes in a WLAN modelled using BMAP/M/1/N queue. In section 4.4 the developed model is validated through NS-2 simulation experiments and performance evaluations are conducted for the comparison of the results from the analytical model with the results gained from simulation. And finally section 4.5 summarizes and concludes this chapter.

4.2. Analytical Model of the IEEE 802.11 MAC DCF Scheme

As a result of massive changes in consumption patterns of digital devices and high demands for multimedia services due to rapid arising and widespread usage of advanced hardware and software technologies, large amounts of traffic are constantly generated and transferred on ubiquitous IEEE 802.11-based WLANs and in particular ad-hoc WLANs which have become imperative in the context of wireless networks.

In this regard, IEEE 802.11 is the dominant standard implemented in majority of digital devices using ad-hoc technology. The standard first introduced in 1997 [3] presents the set of media access and physical layer specifications required for implementing WLANs. The main channel access mechanism provided by IEEE 802.11 is the Distributed Coordination Function (DCF), which allows sharing of the wireless medium through the use of Carrier Sense Multiple Access with Collision

Avoidance (CSMA/CA). In section 2.2.3 of Chapter 2, the mechanism of the DCF function and CSMA/CA in IEEE 802.11 WLANs is explained in details.

Bianchi [8] developed a bi-dimensional Markov chain to model the backoff procedure of the IEEE 802.11 in single hop WLANs, deriving the saturation transmission probability, with the assumption that all stations are always ready for transmission and their transmission queues are assumed to always be non-empty. Duffy et al. in [124], extended Bianchi's model [8] for non-saturated conditions and Wu et al in [160] extended the model to accommodate the case of retry limit. In this section, the analytical models of [124, 160] are extended to develop an analytical model for WLAN under unsaturated network conditions with limited retry and limited buffer size in order to develop a condition closer to realistic networks for applying the multimedia traffic. Therefore, the analytical model is based on the assumption of ideal channel conditions (i.e. no hidden terminals) and fixed number of identical stations (n) in an unsaturated scenario.

An 802.11 WLAN can be considered as a discrete time system which contains multiple generic time slots. In this report, the term time slot is used to denote the time interval between the starts of two consecutive decrements of backoff counter, while the term physical time slot represents a fixed time interval (unit time) specified in the IEEE 802.11 standard [3], which is dependent on the physical layer and accounts for the propagation delay. A generic slot may contain an empty slot, a collision, or a successful transmission.

A station transmits only when its transmission queue is non-empty, therefore the transmission probability τ is calculated by weighting the saturation transmission probability with the probability of the non-empty transmission queue:

$$\tau = (1 - P_0)\tau' \tag{4.1}$$

where P_0 is the probability that the transmission queue of the station is empty, the value of which will be calculated in section 4.3. Eq. (4.1) is calculated with the assumption of no post backoff as stated in [24]. Whenever there is a new packet arrival, the station starts a backoff procedure. τ' is the saturation transmission probability or the stationary probability that the station transmits a packet in a generic (i.e. randomly chosen) slot time, and is calculated as [160]:

$$\tau' = \frac{2(1-2p)(1-p)}{(1-2p)(1-p^{m+1})+(1-p)W\left(1-(2p)^{m'+1}\right)+W2^{m'}p^{m'+1}(1-2p)(1-p)^{m-m'}} \tag{4.2}$$

where W is the minimum contention window size, m is the maximum backoff stage (i.e. retry limit) and m'denotes the maximum number of times that W can be doubled. p is the conditional collision probability and is equal to the probability that at least one of the remaining stations transmits in a given time slot:

$$p = 1 - (1 - \tau)^{n-1} \tag{4.3}$$

The mean service time is the summation result of the average channel access delay, $E[A]$, and average transmission delay, $E[T]$. The channel access delay is the time interval from the moment the frame reaches the head of the queue and contends for the channel until the time it gains access and is ready for transmission. The transmission delay is the time interval of the frame being successfully transmitted. Thus, the mean service time can be shown as follows:

$$E[S] = E[A] + E[T] \tag{4.4}$$

Assuming that the frame is successfully transmitted after experiencing $j, (j \geq 0)$, collisions, its channel access delay would equal to the delay caused by j unsuccessful transmission and the $(j + 1)$ backoff stages. Therefore the channel access delay is calculated as:

$$E[A] = T_c\varphi + \sigma'\delta \tag{4.5}$$

where T_c is the collision time, φ is the average number of collisions before a successful transmission from the station, σ' is the average length of a time slot, and δ represents the average number of time slots the station defers during the backoff stages.

$$\varphi = \sum_{j=0}^{m} \frac{jp^j(1-p)}{(1-p^{m+1})} \tag{4.6}$$

$$\delta = \sum_{j=0}^{m} \sum_{h=0}^{j} \frac{W_h - 1}{2} \frac{p^j(1-p)}{(1-p^{m+1})} \tag{4.7}$$

where p^j is the probability that the frame experiences $j, (0 \leq j \leq m)$, collisions and $\frac{W_h - 1}{2}$ denotes the mean of the backoff counters generated in the h-th $(0 \leq h \leq j)$ backoff stage.

P_{tr} represents the probability that at least one of the remaining stations transmits in a given time slot while the current station is in backoff procedure. When a station transmits, the value of P_{tr} would be equal to the value of p which is given in Eq. (4.3).

P_s is the probability that a transmission occurring on the channel is successful given by the probability that exactly one station transmits on the channel, conditioned on the fact that other stations are in a backoff procedure:

$$P_s = n\tau(1-\tau)^{n-1} \tag{4.8}$$

The average size of a time slot shown by σ', is calculated differently at different stages. When the station is in the backoff stage, the size of the time slot is obtained by considering the fact that the channel is idle with probability:

$$P_{idle} = (1 - P_{tr}). \tag{4.9}$$

When the transmission is successful, it equals to P_s. And finally when there is a collision the size of time slot would equal to $(P_{tr} - P_s)$. Therefore the value of σ' is calculated as follows:

$$\sigma' = (1 - P_{tr})\sigma + P_s T_s + (P_{tr} - P_s)T_c \tag{4.10}$$

where σ is the duration of an empty physical time slot as mentioned in [36].

T_s is the average time the channel is sensed busy by each station because of successful transmission, and T_c is the average time the channel is sensed busy by each station during a collision [8]. The values of T_s, T_c and σ should all be expressed with the same unit.

$$T_s = T_H + T_L + SIFS + 2\Delta + ACK + DIFS \tag{4.11}$$

$$T_c = T_H + T_L + DIFS + \Delta \tag{4.12}$$

where T_H is the average time required to transmit the frame header ($H = PHY_{hdr} + MAC_{hdr}$). Δ is the propagation delay and T_L is the average time required to transmit the longest frame payload.

As mentioned earlier $E[T]$ is the average transmission delay which can be expressed as:

$$E[T] = DIFS + T_H + T_L + SIFS + ACK + 2\Delta \quad (4.13)$$

4.3. BMAP/M/1/N Queueing Analysis of Stations

This section concentrates on the modelling and analysis of transmission queue of the stations within the WLAN. As mentioned earlier, in this model the stations within the WLAN are modelled as BMAP/M/1/N queuing systems [71], where N represents the limited buffer size of each station. The idea is the same as any queueing model, when a frame reaches the head of the transmission queue, the server becomes busy, and as soon as a frame is acknowledged by the destination following a successful transmission, the server becomes free. The service time is dependent on the size of the frame transmitted and is modelled by an exponential distribution function with mean $E[S]$. Thus, the service rate, μ, can be calculated as:

$$\mu = 1/E[s] \quad (4.14)$$

BMAP is characterized by an underlying Continuous Time Markov Chain (CTMC). The number of states in the CTMC represents the number of states considered for BMAP. Figure 4.1 presents a 3-state CTMC of the model with maximum batch size of 3, and Figure 4.2 shows the state transition diagram of the BMAP/M/1/N queue

assuming that the maximum buffer size of each station is N $(N = 50)$ with the maximum batch size of 3:

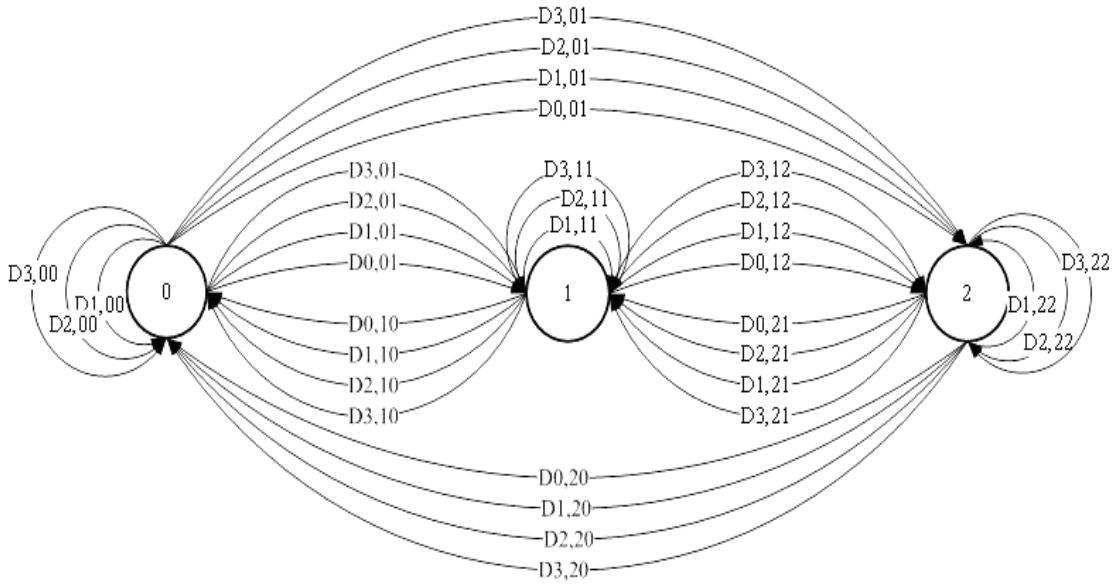


Figure 4.1: Three-state CTMC underlying BMAP with batch size three

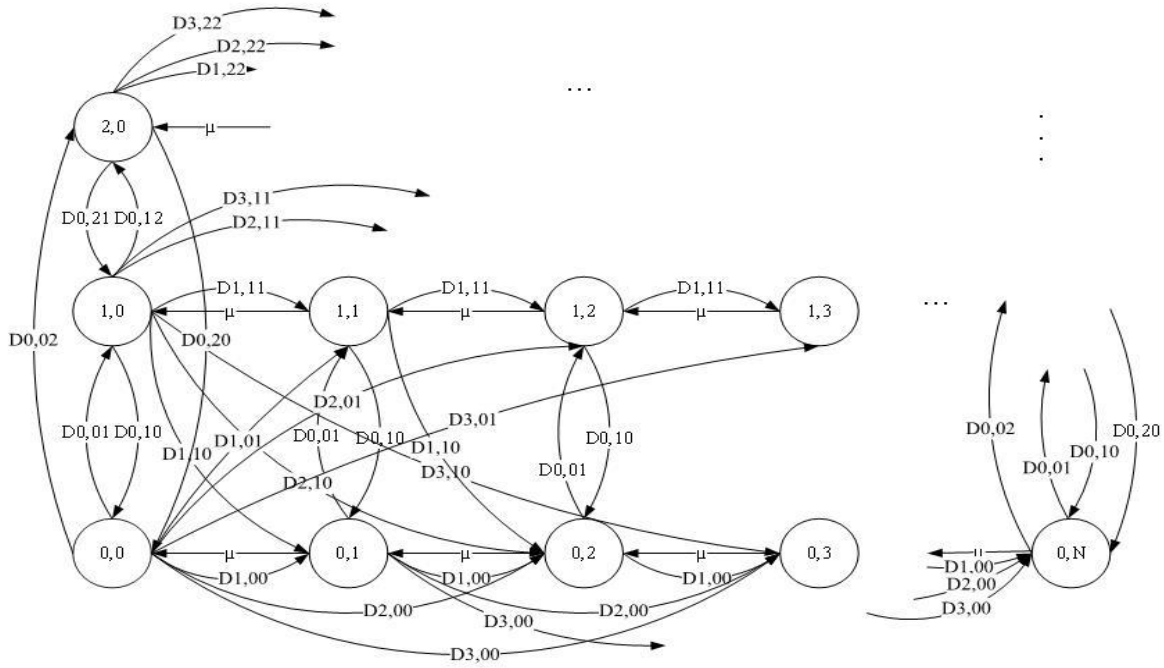


Figure 4.2: State transition diagram of the BMAP/M/1/N (N=50) queue

For clarity and simplicity, the following definitions are presented assuming that a three state BMAP with batch size of maximum three models the nodes of the stations within the network under study. The definitions are definitely extendible to a more general BMAP.

State (η, s), where $(s = 0, \dots, N)$ and $(\eta = 0,1,2)$, represent the case that there are s frames in the queuing system and the three-state BMAP characterizing the traffic of the station is at state η. The transition rate diagram of the BMAP/M/1/N queue is complex. It is characterized based on the states and the batch size of the arrival process. If the batch size is one, meaning that only one frame arrives at the station at one time, then, the transition rate from state (η, s) to $(\eta, s + 1)$ in the CTMC would be determined by the respective value from the matrix of $D_1(\eta, \eta)$, considering that the arrival rate of the generated frame stays the same, meaning that the CTMC is still at the same state. For batch sizes greater than 1, the transition rate from state (η, s) to $(\eta, s + k)$, would be determined by $D_k(\eta, \eta)$ for $(k > 1)$, again with the arrival rate staying the same as before. However if there is a transition from state η to any other state, e.g. $\eta + 1$, with arrivals of batch size of k, then the transition rate would be determined by $D_k(\eta, \eta + 1)$.

The transition rate out of state (η, s) to $(\eta, s - 1)$, would equal to the service rate μ. Whereas the transition rate from state (η, s) to (η', s), where $\eta' = (0,1,2)$ and $\eta' \neq \eta$, would have to be determined by the $D_0(\eta, \eta')$ matrix considering there is no arrival at the time of transition, or by $D_k(\eta, \eta')$ for $(k > 0)$ if there is any arrival during the transition time.

Following the above analysis, the transition rate matrix G, which would be of size $(N+1)\times(N+1)$ (with N being the size of the buffer for each station), of the underlying CTMC of Figure 4.2 can be obtained:

$$G=\begin{bmatrix} -(D_0(0,1)+D_0(0,2)+D_1(0,0)+\cdots) & D_1(0,0) & D_2(0,0) & D_3(0,0) & 0 & \dots \\ \mu & -(\mu+D_0(0,1)+\cdots) & D_1(0,0) & D_2(0,0) & D_3(0,0) & \dots \\ 0 & \mu & -(\mu+D_0(0,1)+\cdots) & D_1(0,0) & D_2(0,0) & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \\ D_0(2,0) & D_1(2,0) & D_2(2,0) & D_3(2,0) & \cdots & \\ \vdots & \ddots & \ddots & \dots & & \end{bmatrix} \tag{4.15}$$

To calculate the steady state probability vector $P=\left(P_{s,\eta}, s=0,1,\dots,N,\ \eta=0,1,2\right)=(P_0,P_1,\dots,P_N)$ of the Markov chain, the transition rate matrix G and the following equations can be used:

$$PG=0 \quad and \quad Pe=1 \tag{4.16}$$

where $P_s=\left(P_{s,0},\ P_{s,1},\ P_{s,2}\right),\ 0\le s\le N$.

Using the method indicated in Eq. (4.17) [85], the above equations can be solved in order to calculate the steady-state vector,

$$P=u(I-\mathcal{R}+eu)^{-1} \tag{4.17}$$

where $\mathcal{R}=I+(\frac{G}{min\{G(j,j)\}})$ denotes the minimum diagonal element of the transition rate matrix G. In Eq. (4.17), u is an arbitrary row vector of $\mathcal{R}$, and e is a column vector of 1's. As a result of the above method, the probability P_0 that the transmission queue of a station is empty is obtained.

4.4. Performance Measures

This section concentrates on the analysis and calculation of unsaturated throughput, end-to-end delay, loss probability and energy consumption of the IEEE 802.11 MAC protocol under bursty traffic. The analytical model is based on the assumption of ideal channel conditions (i.e. no hidden terminals) and fixed number of unsaturated stations. In other words each station transmits only when its transition queue is non-empty.

Basically the arrival process of each station is modelled using the Batch Markovian Arrival Process; the service time distribution is given by the distribution function of $H(s)$ which is assumed to be Exponential, and the service discipline (or the queue discipline of each station) is modelled as First Come First Serve (FCFS). What is important is that the system capacity is finite and is equal to N. This means that the total number of frames in the system of each station must not exceed *N*. Frames arriving when the queue is full are lost and never return. The service times are assumed to be mutually independent and that they do not depend on the arrival process.

π_s denotes the steady-state probability of s frames being in the queuing system of a station and is given by:

$$\pi_s = P_s e, \quad for \ 0 \leq s \leq N \tag{4.18}$$

The probability that the transmission queue of the station is empty, P_0, can be calculated by π_0. Assume π_s' represents the steady-state probability of s frames

being in the queuing system of a station when a frame arrives. In the case of BMAP traffic, π'_s can be written as [161]:

$$\pi'_s = \frac{P_s \Lambda e}{\sum_{s=0}^{N} P_s \Lambda e}, \quad for \;\; 0 \leq s \leq N \tag{4.19}$$

where Λ is a diagonal matrix containing the arrival rates of each state of the underlying CTMC of BMAP. The values of the Λ matrix can be gained from the absolute values of the diagonal elements of the D_0 matrix of BMAP.

The loss probability, the probability that an arriving frame finds the buffer of the station full, is shown by P_b, and can be gained from π'_N.

$$P_b = \pi'_N = \frac{P_N \Lambda e}{\sum_{s=0}^{N} P_s \Lambda e}, \quad for \;\; 0 \leq s \leq N \tag{4.20}$$

With $E[P]$ being the frame payload size and λ_{tot} the fundamental traffic arrival rate of the station, then the throughput, TH, of the station can be computed by:

$$TH = \lambda_{tot} E[P](1 - p_b) \tag{4.21}$$

The end-to-end delay is the time interval from the instant that a frame enters the transition queue of the source station, to the instant that the frame is acknowledged after successful transmission by the destination station. Using the Little's Law, the average end-to-end delay of a frame being transmitted from a source station to a destination station, $E[D]$, can be calculated as:

$$E[D] = \frac{E[N]}{\lambda_{tot}(1-p_b)} \tag{4.22}$$

$E[N]$ is the average number of frames in the queuing system of the station, which can be computed as:

$$E[N] = \sum_{s=1}^{N} s\pi_s \quad (4.23)$$

$\lambda_{tot}(1 - p_b)$ is the effective arrival rate of the transmission queue of the station. When the finite buffer becomes full, any arriving frames are discarded.

4.5. Calculating the Energy Consumption

With the increased maturity of 802.11 technologies over the years and ubiquity of WLANs in providing continuous Internet and network access to many mobile devices, many concentrations have focused on reducing the energy consumption of wireless networks. Since batteries limit the capability of most portable devices, energy conservation has become an increasing concern in the design and implementation of network protocols and technologies. As a result, many researches have devoted their studies into modelling the energy consumption of WLANs with the aim of gaining insights on the power consumption behaviour of real world wireless devices as well as ways to increase the energy efficiency and to prolong the battery life of wireless stations [162-164]. Major proportion of energy consumption in a WLAN interface relates to the contention based MAC protocol, which uses CSMA/CA mechanism [165]. High amount of energy is consumed not only during the active states, but also during the idle states of the protocol.

The energy consumption (in Joules) of a WLAN interface is determined by the power (in Watts) consumed by the WLAN interface during transmission, receiving and idle states (doze), as well as the duration (in hours) the WLAN interface operates in these states. So in general, a WLAN interface is defined by three states of idleness (dozing), transmitting and receiving [31]. The solutions proposed

to reduce the energy consumption of wireless devices mainly evolve around two main issues: 1) reducing the power consumption of the WLAN interface or 2) minimizing the time WLAN interface operates in each of the states, e.g. transmitting or receiving.

As a result, in this section the energy consumption of per successful frame within the developed model is considered and analysed as it could play an important role in design and implementation of WLANs.

The average energy consumed by the interface of each station in a WLAN for successful transmission of a frame is calculated as:

$$E = E_{su} + E_{col} + E_{bf} + E_{em} \tag{4.24}$$

In the above equation, E is composed of four components defined by E_{su}, E_{col}, E_{bf} and E_{em}. E_{su}, is the energy consumed to successfully transmit a frame from source to destination. E_{col} is the energy lost or wasted on collision of the frame while contending for the channel before successful transmission. E_{bf} is the amount of energy consumed during the backoff stages. The energy consumed by the stations when the queue is empty and the station has no pending frame for transmission is shown as E_{em}, which is also known as the doze or idle period.

Even during the backoff stages, the stations consume energy, this is denoted here as e_{ov}, or the energy of overhearing, as the station must continue to monitor the state of the channel and incoming data. When in the transmitting state, the station consumes energy which is denoted as e_{tx}. This is composed of the energy consumed to transmit the data payload as well as MAC header frames. The energy

consumption of WLANs at the transmitting state is higher than that of the receiving state. This is due to the fact that when transmitting a frame, the station must amplify the signal so that the sending frame has enough power to reach its destination successfully.

The energy consumed during the receiving phase is denoted as e_{rx} and the energy consumed during idle periods, when there is no transmitting or receiving of data, is shown as e_{id}. The rest of this section focuses on calculating the components of the energy consumption for each successful transmission of the stations in a WLAN.

The first step is to find and calculate the energy consumed as a result of a successful frame transmission:

$$E_{su} = e_{tx}(T_L + T_H) + e_{rx}T_{ACK} + e_{id}T_{SIFS} \tag{4.25}$$

It is clear that the energy consumed for successful transmission entails the amount of energy used to transmit the header and payload, energy used to receive the acknowledgement and the energy consumed for the idle period of time that the station waits in between of transmission and receiving the acknowledgement, SIFS.

The next step would be to calculate the energy consumed when the station incurs collision. On average each station entails $\varphi(1 - P_d)$ collisions, where φ represents the average number of collisions before a successful transmission from the station, and P_d is the probability that the frame is dropped due to transmission failures and can be calculated as:

$$P_d = {P_{tr}}^{m+1} \tag{4.26}$$

where P_{tr} is the probability that at least one of the remaining stations transmits in a given time slot while the current station is in backoff procedure, which as stated before would be the same as the value of P calculated using Eq. (4.3). Therefore the consumed energy during collisions is expressed as:

$$E_{col} = \left(e_{tx}(T_L + T_H) + e_{id}(T_{DIFS} + T_{SIFS} + T_{ACK})\right) * (\varphi(1 - P_d)) \quad (4.27)$$

The next value to calculate is the energy consumed during the backoff process. If the channel is sensed idle, the backoff counters start to decrement by one physical time slot σ per time. However if there is a collision or a successful transmission from the other stations of the WLAN and the channel is sensed busy, the backoff counter is halted. Based on the equation defined for the calculation of average length of time slot in Eq. (4.10), the energy consumed during backoff stages can be calculated as:

$$E_{bf} = (e_{id}\sigma + e_{ov}(P_s T_s + (P_{tr} - P_s)T_c))(\delta(1 - P_d)) \quad (4.28)$$

The value of the average number of time slot that each station defers during the backoff stages, δ, is already presented in Eq. (4.7). P_s is the probability that a transmission occurring on the channel is successful and can be calculated based on Eq. (4.8).

E_{em} is the last part of Eq. (4.24) that should be calculated. E_{em} represents the amount of energy consumed during the idle state of the station when there is no frame for transmission and the queue is empty. For this purpose, the average time that the transmission queue is empty, T_{em}, should be calculated. This can be done as:

$$T_{em} = E_s(\frac{P_0}{1-P_0}) \quad (4.29)$$

where E_s, the mean service time of a buffer transmitted from a station, is already presented in Eq. (4.4). P_0, the probability that the transmission queue of the station is empty, is calculated in section 4.3. Based on the value defined for T_{em} and regardless of whether the channel is sensed busy or idle, the value of E_{em} can be presented as:

$$E_{em} = T_{em}(e_{id}(1 - P_{tr}) + e_{ov}P_{tr}) \quad (4.30)$$

4.6. Model Validation and Performance Evaluation

In this section, the developed model for performance evaluation of the IEEE 802.11 standard under bursty traffic is validated through extensive simulations using the NS2 [27] simulation environment under various scenarios and conditions.

The developed model is simulated within a Basic Service Set (BSS) of WLANs where n static stations are distributed within a rectangular $100m \times 100m$ grid, and each station generates and transmits traffic to its paired stations. All stations are considered to be within the transmission range of each other and are paired randomly.

The simulations are executed for duration of 600 seconds of NS2 simulation time, which is sufficiently long to gain a stable simulation and reliable performance results. The simulation results are collected after a 10 second warm up period. The remaining simulation settings of the WLAN and stations are summarized in Table. 4.1:

CW_{min}	32	Retry limit (m)	7
CW_{max}	1024	Basic Data Rate	1 Mbps
Slot time	20 μs	Channel Data Rate	11 Mbps
DIFS	50 μs	Propagation delay	2 μs
SIFS	10 μs	ACK Frame Payload	112 bits
MAC hdr	224 bits	PHY Header	192 $bits$

Table 4.1: System parameters for performance analysis of IEEE 802.11 standard under bursty traffic

To validate the developed model, a new traffic generator is developed in NS2 to accommodate the Batch Markovian Arrival Process using C++ and TCL programing languages. The practical usefulness of the analytical results using BMAP depends on how good the model is parameterised based on the original used traffic trace. For flexibility and scalability of the simulation experiments and the analytical model, real-world multimedia applications are adopted using the method presented in [102] for the calculation of the parameters. In each scenario, the details of the used data trace and developed parameters are explained in more details.

The generated traffic in NS2 is then injected into the MAC buffer of the simulated stations where the wireless ad-hoc network will then transmit the traffic through a single-hop route.

- **Scenario 1: Comparison Between 3-state BMAP, 3-state MMPP and Poisson**

In this scenario, the WLAN is composed of n $(n = 10)$ identical stations which are equipped with the $802.11b$ physical layer. The WLAN is assumed to be under ideal wireless channel conditions. The buffer size of all stations is configured to be maximum 50 frames and the size of all data frame payload is set to be 500 Bytes.

To be able to show the accuracy of BMAP in modelling network traffic, four scenarios are considered: I) traffic generated by the stations modelled as a 3-state BMAP with maximum batch size of 3, II) traffic generated by the stations modelled as a 3-state BMAP with maximum batch size of 5, III) traffic generated by the stations modelled using a 3-state MMPP, and IV) traffic generated by the stations modelled by a Poisson Process.

At this stage, as regards to the practical size limits and for ease of modelling and simulation, the BMAP of the queueing model at each station of the WLAN is considered to be composed of three states. The stations are assumed to be able to generate batches of maximum three and five frames at one time for different scenarios; however the accuracy of the model will slightly increase as the CTMC is extended, and a three state CTMC at this stage will generate a good result.

The data trace used is obtained from the high quality measurement of the video stream for the film “Tears of Steel”, encoded in H.265/HEVC codec [159]. Using the data trace and the EM algorithm [102], parameters for a 3-state BMAP with maximum batch sizes of three and five, and the 3-state MMPP are estimated.

To vary the traffic load of the stations, the arrival rate of each state of the underlying Markov Chain of the BMAP process defined in matrix D_0 are varied according to the overall load of the traffic placed on the network. The infinitesimal generator matrix of the 3-state BMAP with batch size of maximum 3, Q, and the D_0 matrix are as follows:

$$D_0 = \begin{pmatrix} -\lambda_1 & 0.01 & 0.03 \\ 0.05 & -\lambda_2 & 0.08 \\ 0.17 & 0.20 & -\lambda_3 \end{pmatrix} \quad Q = \begin{pmatrix} -0.17 & 0.07 & 0.10 \\ 0.34 & -0.7 & 0.36 \\ 0.61 & 0.7 & -1.33 \end{pmatrix}$$

where $\lambda_2 = 0.4 * \lambda_1$ and $\lambda_3 = 0.1 * \lambda_1$ and $\lambda_1 = 0.5 * \lambda_{tot}$, with λ_{tot} being the overall traffic load of the station. For 3-state BMAP with maximum batch size of five the following results were estimated using the same data trace:

$$D_0 = \begin{pmatrix} -\lambda_1 & 0.03 & 0.03 \\ 0.1 & -\lambda_2 & 0.1 \\ 0.2 & 0.12 & -\lambda_3 \end{pmatrix} \qquad Q = \begin{pmatrix} -0.25 & 0.14 & 0.11 \\ 0.49 & -1.1 & 0.57 \\ 0.78 & 0.96 & -1.93 \end{pmatrix}$$

For the purpose of comparison, the developed model is executed using BMAP, MMPP and Poisson processes. To generate comparable results, the overall load of the stations is kept the same in all the models. The stationary probability vector of all three models using BMAP and MMPP is $\pi = [0.722 \quad 0.175 \quad 0.102]$.

With the infinitesimal generator Q and D_0 matrices of the BMAPs defined as above, the rest of the matrices, D_1 to D_k $(k = 3 \; or \; 5)$, are developed using the method presented in [98]. For three-state MMPP the following was estimated for the infinitesimal generator Q using the EM algorithm:

$$Q = \begin{pmatrix} -0.079 & 0.039 & 0.04 \\ 0.180 & -0.416 & 0.24 \\ 0.33 & 0.33 & -0.66 \end{pmatrix}$$

The MMPP with infinitesimal generator Q and arrival rate matrix $\Lambda = diag(\lambda_{1,\dots,}\lambda_m)$ is a $BMAP$ process [71] with $D_0 = Q - \Lambda$, and $D_1 = \Lambda$, and $D_k = 0\ (k > 1)$. Therefore with the infinitesimal generator estimated above, and the arrival rates set as $\lambda_2 = 0.4 * \lambda_1$, $\lambda_3 = 0.1 * \lambda_1$ and $\lambda_1 = 0.5 * \lambda_{tot}$, the result of the models will be comparable. The corresponding Poisson process in the format of BMAP is a process with rate λ, where $D_0 = -\lambda$ and $D_1 = \lambda$.

Figures 4.3 to 4.6 depict the throughput, end-to-end delay, frame loss probability and energy consumption of the WLAN under study with different traffic models. The figures reveal a good match between the analytical and the simulation results, therefore proving the accuracy and reliability of the developed model. There are minor discrepancies between the model and the simulation results at some stages which are mainly due to the approximations taken into consideration to make the model tractable. An instance of the approximations is the assumption made that the collision probability is the same at all times regardless of the backoff stage of the DCF protocol. However the extensive comparisons made between the analytical results and those gained from simulation prove that the model has an acceptable accuracy. It is evident from the results that bursty traffic models have significant impacts on the network performance. At lower rates, the throughput of the WLAN with different models is similar, but as soon as the load reaches around 0.2 Mbps differences start to appear.

For the Poisson and 3-state MMPP the results remain close up to the load of 0.3 Mbps, with continuous increase in the load, the network reaches saturation point for 3-state MMPP while in the Poisson model the throughput continues to increase

until reaching a steady value at 0.34 Mbps. In both the Poisson model and the 3-state MMPP, the WLAN reaches saturation point when the load reaches 0.4 Mbps, whereas for the 3-state BMAP with batch sizes of 3 and 5, the network reaches saturation point earlier, when the load is around 0.2 Mbps. Overall the maximum throughput of the network during saturation is higher for Poisson, 0.34 Mbps, than the other three traffic models. After Poisson, stands the three-state MMPP with 0.29 Mbps, then 3-state BMAP with batch size of maximum three with 0.17 Mbps and finally is the 3-state BMAP with maximum batch size of five with 0.12 Mbps.

The same analysis stands for the frame loss probability and the end-to-end delay. The network has the highest loss when the traffic is modelled using the three-state BMAP with batch size of maximum five, even higher than three-state BMAP with batch size of maximum three. The three-state MMPP has a lower loss probability than the two BMAPs but still has a higher loss rate compared to the Poisson model. The results are logical as BMAP with maximum batch size of five generates more burstiness, as seen in chapter 3, whereas the Poisson process or the MMPP process generates one packet at each instance. The end-to-end delay comparison shows that the delay in the WLANs with highly bursty traffic reaches a high level at even lower loads. When the traffic is modelled using Poisson or MMPP, the network has a period of very low end-to-end delay before a sharp increase which then stays nearly the same during the saturation period. The reason is that with highly bursty traffic models, even at low rates, there are more packets competing for the shared media and therefore the delay for access to the wireless channel increases.

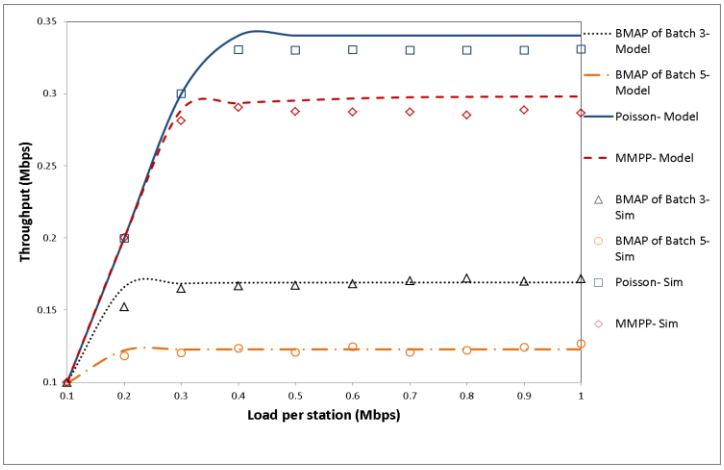


Figure 4.3: Throughput of WLANs with traffic generated by Poisson, MMPP and BMAP with maximum batch size of 3 and 5.

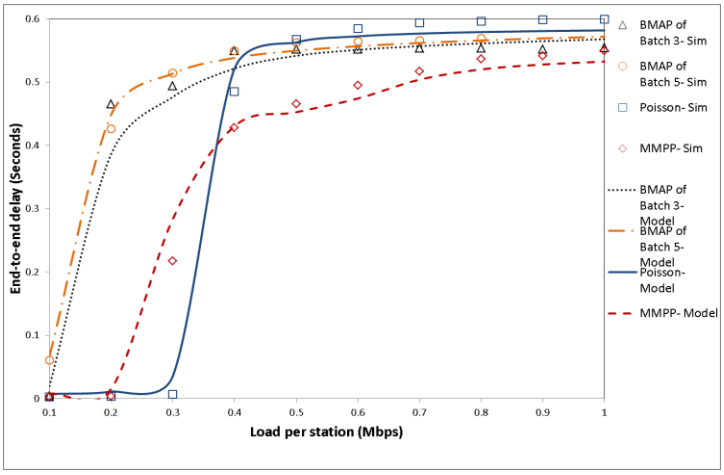


Figure 4.4: End-to-end delay of WLANs with traffic generated by Poisson, MMPP and BMAP with maximum batch size of 3 and 5.

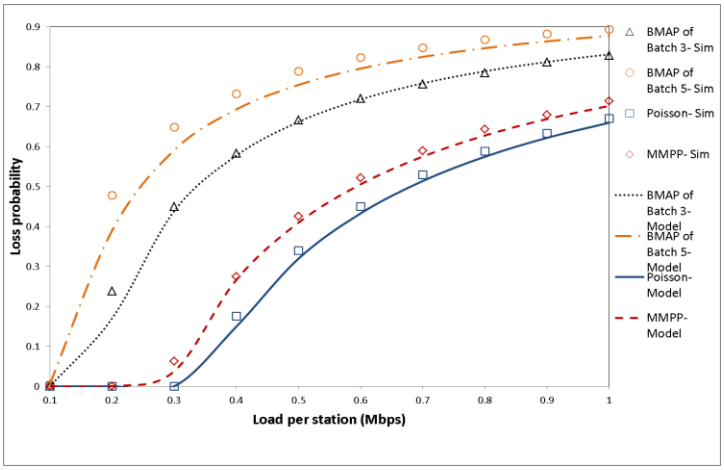


Figure 4.5: Loss probability of WLANs with traffic generated by Poisson, MMPP and BMAP with maximum batch size of 3 and 5.

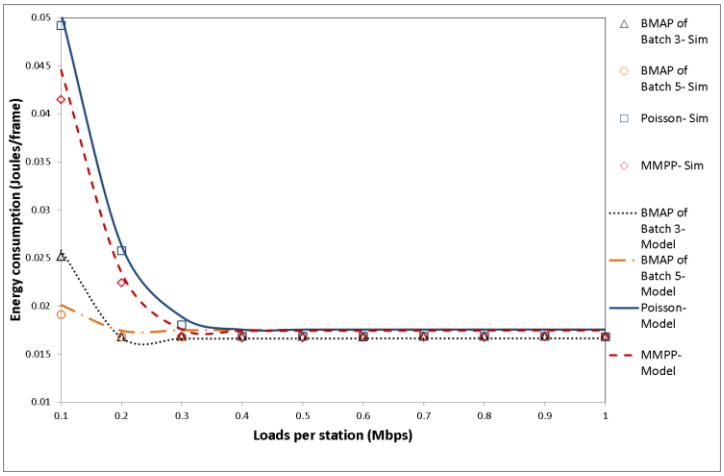


Figure 4.6: Energy consumptions of successful transmissions in WLANs with stations generating traffic using Poisson, 3-state MMPP and 3-state BMAP with maximum batch size of 3 and 5.

The power required for transmitting, receiving, overhearing and being idle are 1.65, 1.4, 1.4, and 1.15 W, respectively. Figure 4.6 illustrates the energy consumption of the WLAN per successful transmission, it is worth mentioning that overall the energy consumption decreases as the traffic load increases, and this is because much energy is wasted in the idle period when the loads are low. Using the same reasoning it is possible to see that the Poisson process has the highest consumption of energy at the start and all through when the network load is low. But from when the load of the stations reaches 0.4 Mbps, the energy consumption of all the models drops down to the same level as they all reach their saturation point, during which little or no idle time is wasted in the DCF scheme, hence the low energy consumption.

Moreover, in general the results show that the throughput and end-to-end delay first increase and as the traffic loads grow they start to stabilize whereas the frame dropping probability keeps rising with the loads.

Overall it can be seen from the results that the QoS measures of stations under bursty traffic are similar to those under Poisson traffic when the network operates at low traffic loads (less than 0.2 Mbps). This is because the collision probability at lower rates is very small hence the low frame loss probability and queueing delays. However from the figures it is very obvious that as the load increases, the network performance and the QoS measures start to drift apart as the load on the network increases rapidly under bursty traffic. This is an obvious proof to the fact that the traffic generated by multimedia applications is highly bursty, correlated and self-similar, and the conventional Poisson model is no longer adequate for modelling current network traffic.

- **Scenario 2: Effect of Buffer Size**

In this scenario, the WLAN is composed of n $(n = 10)$ identical stations equipped with the $802.11b$ physical layer, and ideal wireless channel conditions are assumed. The traffic generated by each station of the WLAN is modelled by a 3-state BMAP with maximum batch size of three. To investigate the impact of the buffer size on the performance of the WLAN under bursty and correlated traffic, the model is studied under three different buffer sizes of 5, 10 and 50 frames while the size of all data frame payload is set to be 500 Bytes.

Depending on the applications and how sensitive they might be to delay, e.g. video applications, or loss, e.g. voice applications, the size of the buffer set on the stations can become an important issue. Enlarging the buffer size can result in lower blocking probability but in turn could increase the end-to-end delay and vice

versa. Therefore in order to reach a reasonable trade-off between these parameters it is important that both performance measures are calculated.

Similar to Scenario1, for parameter estimation of the BMAPs the video stream for the film “Tears of Steel”, encoded in H.265/HEVC codec is used [159], from which the parameters for a 3-state BMAP with maximum batch size of three are estimated using the EM algorithm [102].

To vary the traffic load of the stations, the arrival rate of each state of the underlying Markov Chain of the BMAP process defined in matrix D_0 are varied according to the overall load of the traffic placed on the network. The infinitesimal generator matrix of the 3-state BMAP with batch size of maximum 3, Q, and the D_0 matrix are as follows:

$$D_0 = \begin{pmatrix} -\lambda_1 & 0.01 & 0.03 \\ 0.05 & -\lambda_2 & 0.08 \\ 0.17 & 0.20 & -\lambda_3 \end{pmatrix} \qquad Q = \begin{pmatrix} -0.17 & 0.06 & 0.10 \\ 0.34 & -0.70 & 0.36 \\ 0.61 & 0.71 & -1.33 \end{pmatrix}$$

where $\lambda_2 = 0.2 * \lambda_1$ and $\lambda_3 = 0.1 * \lambda_1$ and $\lambda_1 = 0.7 * \lambda_{tot}$, with λ_{tot} being the overall traffic load of the station. The stationary probability vector of the BMAP is $\pi = [0.722 \quad 0.175 \quad 0.102]$.

Figures 4.7 to 4.10 illustrate the results gained from the analytical models and simulation of the WLAN with different buffer sizes of 5, 10 and 50 (where N resembles the buffer size). The results show that under similar conditions and traffic loads, the WLAN in which the stations have larger buffer sizes achieves a higher throughput and faces lower loss probability while causing an increased delay. The increase in end-to-end delay of the WLAN with buffer size of 50 is

incredibly higher than the other two WLANs with buffer sizes of 5 and 10, whereas the difference in the loss probability and throughput is now as high. This is because the end-to-end delay is calculated for frames that wait in the queue; therefore under similar traffic burstiness and traffic intensity conditions, it takes longer for the frames waiting in the queue of the WLAN with buffer size of 50 to be transmitted and successfully received at the destination station. Moreover, when the buffer size increases from 5 to 10 frames, the improvement of throughput and loss probability is not significant, however between having buffer size of 5 and buffer size of 50 the differences are considerable. As throughput and loss probability are the most important performance metrics of delay-insensitive applications, it is desirable to set a large buffer size for these applications. However, for the delay-sensitive applications such as voice and video, a large buffer results to the high delay that may be intolerable for these inelastic applications. Thus, a small buffer is preferable for delay-sensitive applications.

The energy consumption results show differences for when the load on the stations reaches a higher level. At the beginning as the load on the stations is low the energy consumption of all three WLANs with buffer sizes of 5, 10 and 50 is very close to each other. However as soon as the load reaches around 0.3 Mbps, and the networks start to become saturated the energy consumption of the WLAN with smaller buffer size increases. This is because the stations of that network have to deal with high rates of frame loss and as a result the energy consumption increases. It can be understood from the results that during the saturation period, the throughput, loss probability and energy consumption of all three WLANs reaches a similar steady point.

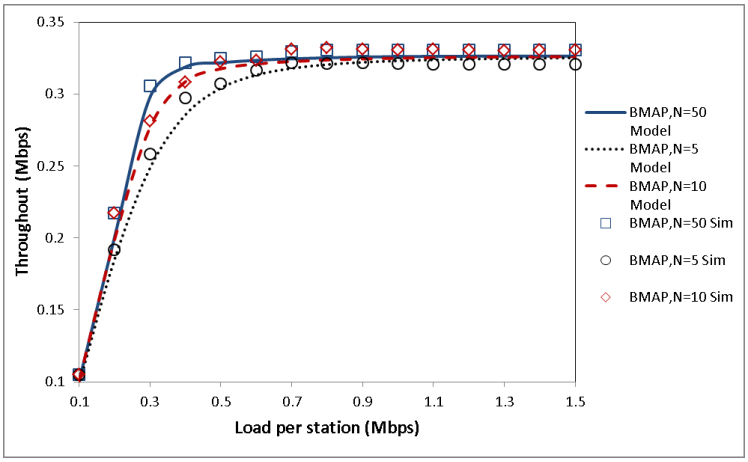


Figure 4.7: Throughput of WLANs with varying buffer sizes and stations generating traffic using 3-state BMAP with maximum batch size of 3.

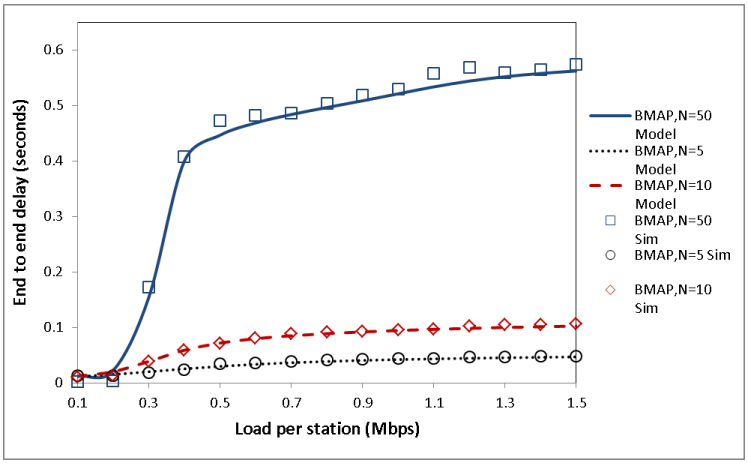


Figure 4.8: End-to-end delay of WLANs with varying buffer sizes and stations generating traffic using 3-state BMAP with maximum batch size of 3.

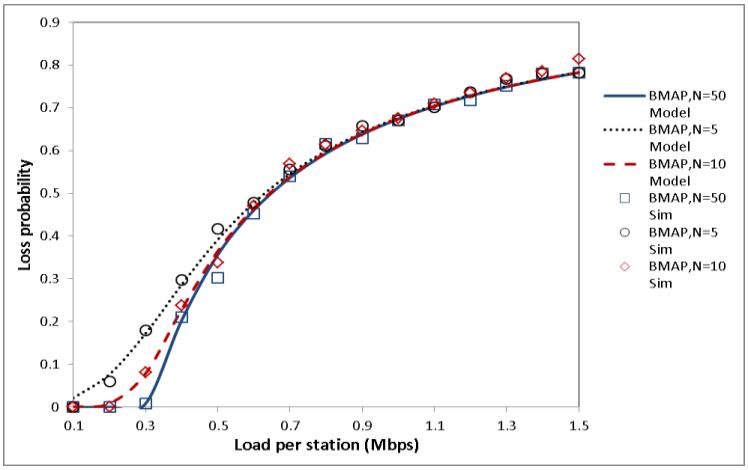


Figure 4.9: Loss probability of WLANs with varying buffer sizes and stations generating traffic using 3-state BMAP with maximum batch size of 3.

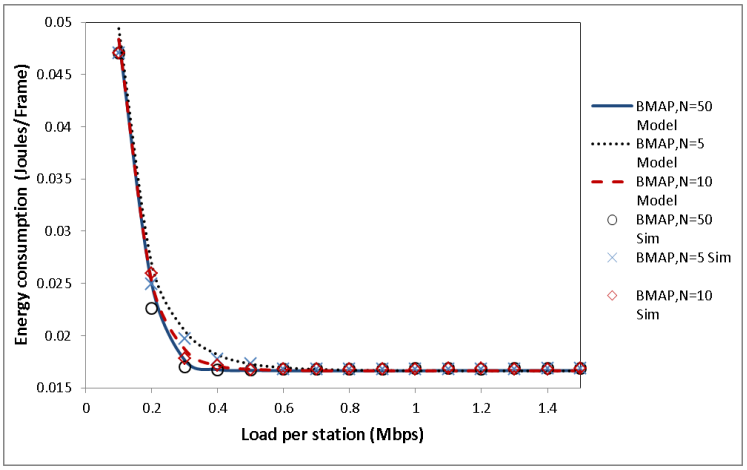


Figure 4.10: Energy consumption of successful transmissions in WLANs with varying buffer sizes and stations generating traffic using 3-state BMAP with maximum batch size of 3.

➢ Scenario 3: Effect of Traffic Burstiness on WLANs

In this scenario the effect of maximum batch size of BMAP process used to model the network traffic is studied on WLANs. The variability in the maximum batch size of BMAP results in different burstiness in the generated traffic of the stations within the WLANs. For each WLAN, $n\ (n = 10)$ identical stations are defined which are each equipped with the $802.11b$ physical layer, and ideal wireless channel conditions are assumed. The BMAP used for generating the traffic is composed of a 3-state underlying Markov Chain. All data frame payloads are set to be 500 Bytes.

Similar to previous scenarios the parameters of the BMAP are estimated from the "Tears of Steel" data trace encoded in H.265/HEVC using the EM algorithm defined in [102].

The BMAPs are modelled with a 3 state Markov Chain and are shown as $BMAP(j)$, where j resembles the maximum batch size acceptable for each BMAP. The maximum batch sizes considered are $3, 5$ and 10.

To vary the traffic load of the stations, the arrival rate of each state of the underlying Markov Chain of the BMAP process defined in matrix D_0 are varied according to the overall load of the traffic placed on the network. The infinitesimal generator matrix of the 3-state BMAP, Q, and the D_0 matrix are as follows:

$$D_0 = \begin{bmatrix} -\lambda_1 & 0.01 & 0.03 \\ 0.05 & -\lambda_2 & 0.08 \\ 0.17 & 0.20 & -\lambda_3 \end{bmatrix} \qquad Q = \begin{bmatrix} -0.17 & 0.06 & 0.10 \\ 0.34 & -0.70 & 0.36 \\ 0.61 & 0.71 & -1.33 \end{bmatrix}$$

where $\lambda_2 = 0.2 * \lambda_1$ and $\lambda_3 = 0.1 * \lambda_1$ and $\lambda_1 = 0.7 * \lambda_{tot}$, with λ_{tot} being the overall traffic load of the station. The stationary probability vector of the BMAP is $\pi = [0.722 \quad 0.175 \quad 0.102]$.

As in Scenario 4 of section 3.10.1, the rest of the matrices of each BMAP, $D_k, 1 \leq k \leq K_{max}$, are calculated by substituting the value of K_{max} with the maximum batch size for each WLAN in Eq. (3.54).

Figures 4.11 to 4.14 depict the gained results for each of the performance measures of throughout, end-to-end delay, loss probability and energy consumption of the WLAN in successful transmissions. It is clear that any increase in burstiness of the traffics generated by the stations of the network highly affect the performance results. The increase in burstiness as a result of generating batches with maximum size of 10 frames significantly reduces the throughput of the network, and increases the frame loss probability and end-to-end delay.

However this increase in maximum possible batch size and burstiness results in less energy consumption in successful transmission when the network is under lower traffic loads. The reason being is that during this time the network is less idle and hence less energy is consumed. So depending on the purpose of the designed network there needs to be a balance between the energy consumption of the stations and the loss and end-to-end delay of the frames.

These results have important points in regards to encoding video traffics transmitted within WLANs. It can be concluded that the higher the variation of the frame sizes used for encoding the video traces, the lower the throughput of the networks is going to be. Also the increase in the end-to-end delay and loss probability show how the increase in frame size variation can degrade the performance of delay sensitive or loss sensitive applications. This is an important outcome that should be incorporated in design and deployment of network applications.

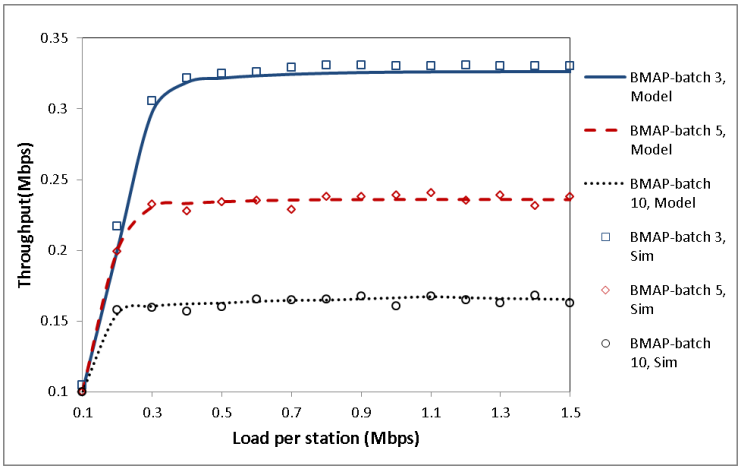


Figure 4.11: Effects of burstiness on throughput, 3-state BMAP/M/1/N with variable maximum batch sizes.

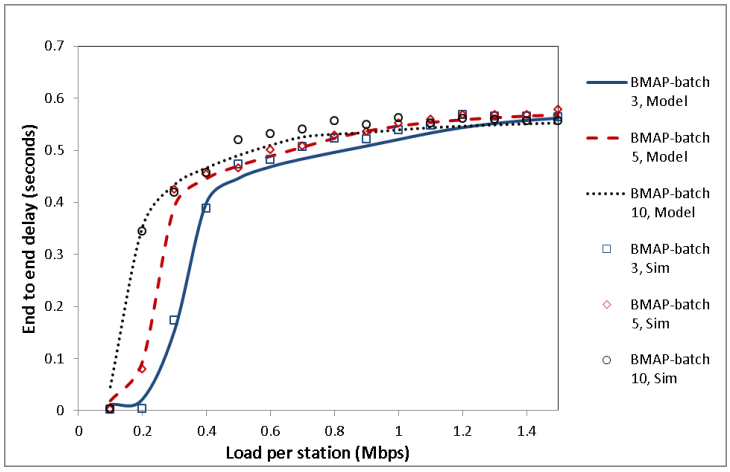


Figure 4.12: Effects of burstiness on end to end delay, 3-state BMAP/M/1/N with variable maximum batch sizes.

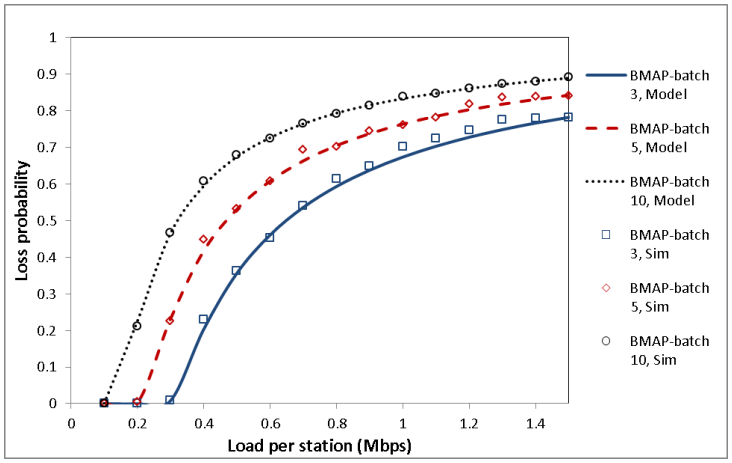


Figure 4.13: Effects of burstiness on loss probability, 3-state BMAP/M/1/N with variable maximum batch sizes.

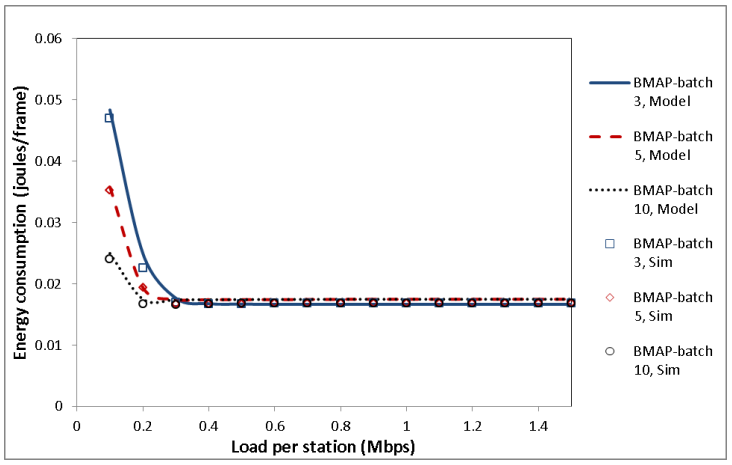


Figure 4.14: Effects of burstiness on energy consumption of successful transmission, 3-state BMAP/M/1/N with variable maximum batch sizes.

- **Scenario 4: Effect of Network Size**

In this scenario the aim is to execute models and simulations that exhibit the explicit relationship between the performance metrics of WLANs and the size of network. For this purpose, four different settings for BMAP are used to model the traffic generation of the stations within four WLANs which are similar with each other in all terms other than the BMAP process. The difference between the BMAPs used is in the maximum possible batch size of the BMAP which are set to $3, 4, 5$ and 10. To identify the WLANs from each other, they are resembled by the maximum batch size of the respective BMAP used as $BMAP(K_{max})$, where K_{max} resembles the maximum batch size. The number of the states of the underlying Markov Chain used for all four BMAPs is kept to three to create comparable situations.

The stations within each WLAN are identical and are each equipped with the $802.11b$ physical layer. The traffic rate, buffer size and data frame payloads at each station are set to $\lambda_{tot} = 0.35$ Mbps, 50 frames and 500 Bytes, respectively. The WLANs are modelled in ideal wireless channel conditions. The number stations in each WLAN is increased from 5 to 50 over the region of $100m \times 100m$.

The parameters of the BMAP in this scenario again are estimated from the "Tears of Steel" data trace encoded in H.265/HEVC using the EM algorithm defined in [102]. The matrix D_0 and the infinitesimal generator matrix, Q, for each BMAP is defined as follows:

$$D_0 = \begin{pmatrix} -\lambda_1 & 0.01 & 0.03 \\ 0.05 & -\lambda_2 & 0.08 \\ 0.17 & 0.20 & -\lambda_3 \end{pmatrix} \qquad Q = \begin{pmatrix} -0.17 & 0.06 & 0.10 \\ 0.34 & -0.70 & 0.36 \\ 0.61 & 0.71 & -1.33 \end{pmatrix}$$

where $\lambda_2 = 0.2 * \lambda_1$ and $\lambda_3 = 0.1 * \lambda_1$ and $\lambda_1 = 0.7 * \lambda_{tot}$, with λ_{tot} being the overall traffic load of the stations defined as 0.35 Mbps. The stationary probability vector of the BMAP is $\pi = [0.722 \quad 0.175 \quad 0.102]$.

As in Scenario 4 of section 3.10.1, the rest of the matrices of each $BMAP(K_{max})$, $D_k, 1 \leq k \leq K_{max}$, are calculated by substituting the value of K_{max} with the maximum batch size (3, 4, 5 or 10) for each WLAN in Eq. (3.54).

Figures 4.15 to 4.17 plot the performance metrics of the WLANs, modelled using BMAPs with different maximum batch sizes, as a function of the network size. From the figures it can be established that when the number of stations is small, the burstiness of the traffic has little impact on the performance of the WLAN since the network is under light loads, except for $BMAP(10)$. Maximum batch size of 10 increases high burstiness in the traffic generated by the stations, and as a result the performance measures are greatly affected. The increase in traffic burstiness reduces the throughput, increases the end-to-end delay and loss probability. This shows that with less bursty traffic, the network can sustain more stations.

Overall in all four settings, when the number of stations reaches 10, the network becomes saturated and the values of the performance measures take abrupt change. For WLANs modelled using $BMAP(3)$, $BMAP(4)$ and $BMAP(5)$ after saturation the values of the end-to-end delays become extremely close to each other.

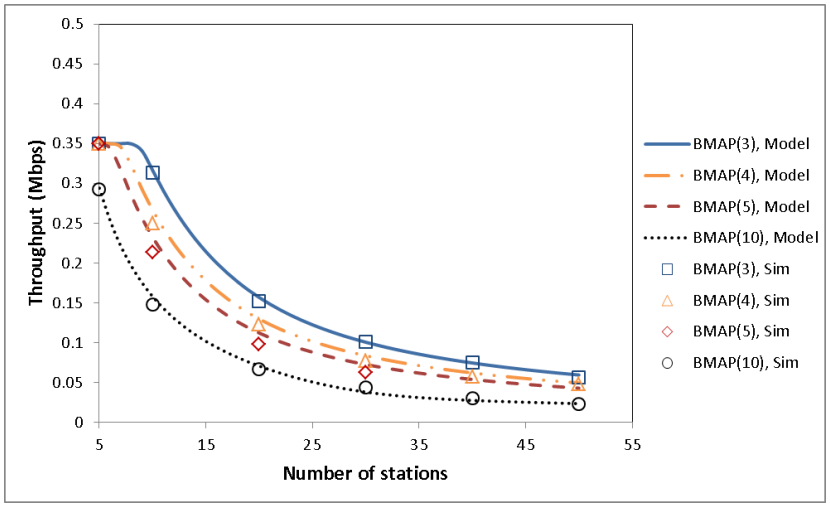


Figure 4.15: Effects of network size and burstiness on throughput, 3-state BMAP/M/1/N with variable maximum batch sizes.

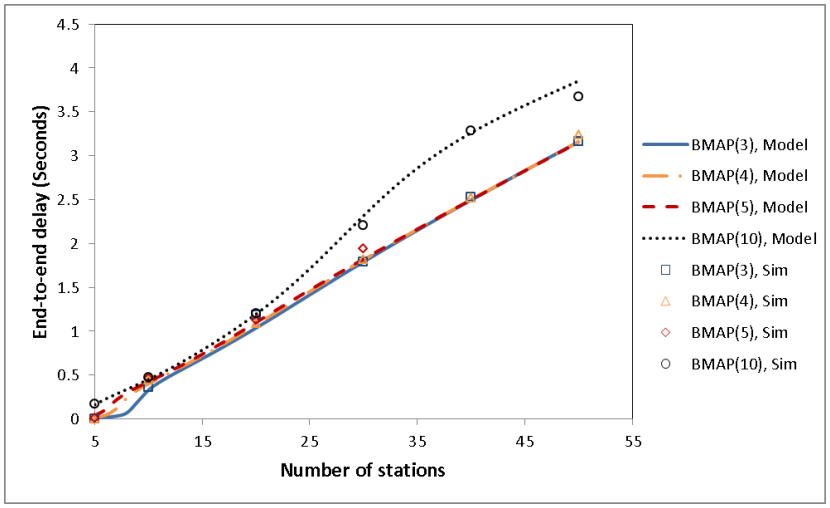


Figure 4.16: Effects of network size and burstiness on end-to-end delay, 3-state BMAP/M/1/N with variable maximum batch sizes.

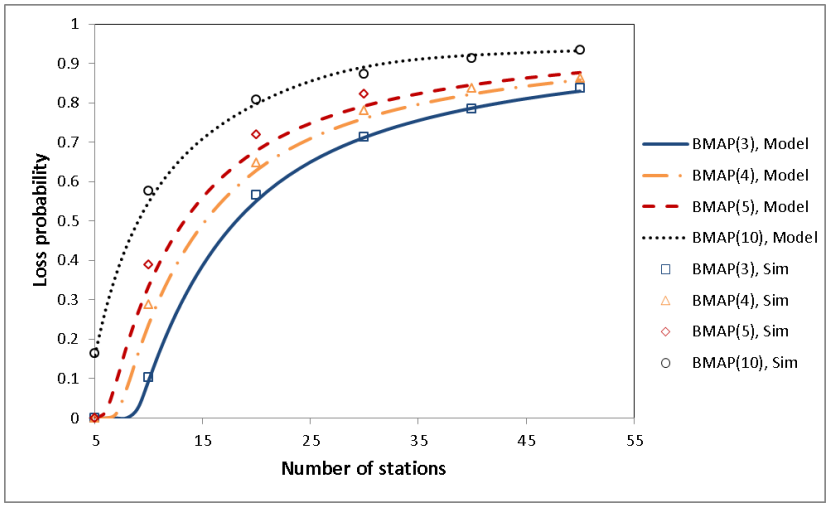


Figure 4.17: Effects of network size and burstiness on loss probability, 3-state BMAP/M/1/N with variable maximum batch sizes.

4.7. Summary

In this chapter a novel analytical model was presented for 802.11 WLANs under bursty traffic. The accuracy of the developed model was verified through comparing the analytical results with extensive NS2 simulation experiments. The expressions of the important QoS performance metrics including throughput, end-to-end delay, frame loss probability, and energy consumption were calculated. A thorough investigation into the impact of the traffic load, traffic burstiness, buffer size, and number of stations on the QoS performance of the 802.11 WLAN was conducted.

The performance results have shown that the bursty traffic can substantially affect the QoS performance of WLANS by reducing the throughput and energy consumption of successful transmission while increasing the end-to-end delay and the loss probability. When the network is under light load, the QoS performance is less affected except for the energy consumption. However, as the network loads

become moderate and heavy, the stations with burstier traffics perceive the degrading QoS than those under less bursty traffic.

Chapter 5:

Modelling and Analysis of Heterogeneous Multimedia WLANs under Bursty Traffic

5.1. Introduction

Heterogeneous wireless networks widely adopt layered video encoding for multimedia video applications. As well as resulting in high Variable Bit Rate (VBR) in video traffics, they often create correlation between the base layer and the enhancement layer(s) of video data [166, 167]. These correlations lead to extreme burstiness and self-similarity in video traffics transmitted over WLANs which emphasizes the importance and necessity of accurate performance models for the analysis of bursty and correlated traffics in wireless networks [168].

Many studies have been carried out on developing models that can accurately capture the properties of video traffics and in particular their burstiness characteristics [76, 141, 169-172]. In [172] the authors thoroughly study the properties of video traffics over networks encoded via multiple encoding schemes and show that the LRD (Long Range Dependence or self-similarity) characteristics of video traffics [7, 173, 174] exposed in previous studies [175] on itself is not sufficient for the modelling and analysis of these types of traffic in modern communication systems. They show with precise results that as well as LRD, video traffics exhibit high bit rate variability also known as burstiness, which fractal

analysis and even the Poisson Process falls short in fully describing. In [171] the authors develop versatile traffic models that focus on the MAC and the Physical (PHY) layers suitable for use in multiple simulation scenarios. This work is an extension to the author's earlier [170] work which adopts the superposition of several IPPs (Interrupted Poisson Process) to produce typical internet traffic over a wireless network. The resulting IPP-based On-Off model is easy to use and scalable enough to provide accurate results and so it was proposed for IEEE 802.16 networking standard [176]. The developed On-Off models have extensively been used since then to model traffic generated by voice applications in various wireless networks [141, 143, 177, 178]. In [76] the authors use superposition of multiple 2-state MMPPs to model the On-Off characteristics of voice. As well as correlation between the variable arrival rates, burstiness in video and other multimedia traffics is also the result of correlation between packet size distributions and packet arrival rates [41]. MMPP can only capture the correlation between the arrival rates and lacks the ability to capture the inherent correlation that exists between packet size distributions and arrival rates. Whereas BMAP [20, 65] has the capability to model dependent and non-exponential inter-arrival time distributions between batches and correlated batch sizes of arriving packets.

However with the availability of the aforementioned analytical results as well as other studies on the analysis of WLANs under multimedia traffic [26, 135, 179]; to the best of our knowledge, there is no work available in literature that considers the use of continuous-time BMAP for traffic generation in the analysis and performance evaluation of heterogeneous WLANs under multimedia traffic.

Since Bianchi's work, many further models have been proposed in literature that present a more refined and accurate model for the DCF [113, 114, 116, 118, 128, 160]. Among the models proposed, most concentration is on non-saturated [25, 119, 121, 123, 124, 180, 181] network conditions and homogeneous traffic sources (i.e. sources with the same distribution and arrival rates). Only limited research is available that considers heterogeneity of non-saturated traffic sources within the network [24, 60, 182-186]. For example in [187] the authors present throughput analysis of heterogeneous IEEE 802.11 DCF WLAN where nodes are grouped based on saturated and un-saturated traffic generation conditions, their rates and backoff window size. However WLANs today are utilized in a wide variety of environments where miscellaneous devices generate tremendous volume of multimedia traffic with distinct characteristics and respective Quality-of-Service (QoS) guarantees [188].

To model varying arrival rates and packet size distributions, this chapter presents a new and practical analytical model for the analysis of the MAC layer in 802.11 WLANs under bursty heterogeneous multimedia traffics using BMAP which calculates the throughput, end-to-end delay and loss probability performance measures. As burstiness, correlation and self-similarity can degrade network performance through long delays, severe packet dropping and large buffer requirements, it is important that they are taken into consideration in the study and development of highly efficient WLANs.

The rest of this chapter is organised as follows: the analytical model for heterogeneous Wireless Local Area Networks subject to heterogeneous

multimedia traffics based on bursty traffic generators is presented in Section 5.2. Section 5.3 validates the accuracy of the model using thorough simulations is NS2, as well as presenting performance evaluations. And finally Section 5.4 concludes and summarizes this chapter.

5.2. Analytical Model

This section presents the analytical model for evaluating the performance of unsaturated WLANs with heterogeneous stations generating bursty and correlated multimedia traffic using the Batch Markovian Arrival Process and background data traffics using the Poisson Process. The transmission queues of the stations generating voice or video traffic are modelled as $BMAP/G/1/N$ queuing system, and the ones generating the background data are modelled as $M/G/1/N$ queueing system.

The analytical model is developed based on the assumption that the stations are divided into G groups of heterogeneous IEEE 802.11 stations, where each group is identified with the label g_i $(1 \leq i \leq G)$. The nodes in the same group are assumed to have identical traffic settings.

While the traffic models differ between the groups, the overall rate of the stations are kept the same for comparison purposes. All stations are unsaturated with a limited sized buffer of N. The quantities of interest calculated by the model are throughput, end-to-end delay and frame loss probability of individual group of stations.

5.2.1. Analysis of the Service Time

Bianchi [8] developed a bi-dimensional Markov chain to model the backoff procedure of the IEEE 802.11 in single hop WLANs, deriving the saturation transmission probability, with the assumption that all stations are always ready for transmission and their transmission queues are assumed to be non-empty at all points in time. D. Malone et al. in [24], extend Bianchi's model for non-saturated and heterogeneous conditions, and Wu et al in [160] extend the model to accommodate the case of retry limit. In this section, the models of [24, 160] are extended to develop an analytical model for WLANs under unsaturated network conditions with heterogeneous stations, retry limit and limited buffer size in order to develop conditions closer to realistic networks. We assume the WLAN under study is composed of fixed number of heterogeneous stations (n_i). The stations are grouped into i $(1 \leq i \leq G)$ groups. Each group (g_i) represents stations with similar type of generating traffics. The stations only transmit when their transmission queues are non-empty; therefore the per-station transmission probability τ_i is calculated by weighting the saturation transmission probability with the probability of the non-empty transmission queue:

$$\tau_i = (1 - P_{i0})\tau_i' \tag{5.1}$$

where P_{i0} is the probability that the transmission queue of the station is empty. Eq. (5.1) is calculated with the assumption of no post backoff as stated in [24]. τ_i' is the saturation transmission probability given as [160]:

$$\tau_i' = \frac{2(1-2P_i)(1-P_i)}{(1-2P_i)\left(1-{P_i}^{m+1}\right)+(1-P_i)W\left(1-(2P_i)^{m'+1}\right)+W2^{m'}{P_i}^{m'+1}(1-2P_i)(1-P_i)^{m-m'}} \tag{5.2}$$

In Eq. (5.2), W is the minimum contention window size, m is the maximum backoff stage (i.e. retry limit) and m'denotes the maximum number of times that W can be doubled. P_i is the conditional collision probability and is equal to the probability that at least one of the remaining stations transmits in a given time slot:

$$P_i = 1 - (1 - \tau_i)^{n_i - 1} \prod_{r \neq i} (1 - \tau_r)^{n_r} \quad for \ \ r = 1, \dots, G \tag{5.3}$$

Eq. (5.1) and Eq. (5.3) are two non-linear equations that can be solved numerically for different values of τ_i and P_i.

The mean service time is the summation of the average channel access delay, $E[A_i]$, and average transmission delay, T_s, calculated as

$$E[S_i] = E[A_i] + T_s \tag{5.4}$$

$E[A_i]$, is the time interval from the instant the frame reaches the head of its transmission queue and starts contending for the channel until the time it wins the contention and is ready for transmission. T_s, is the time interval that the frame is successfully transmitted, with i denoting the transmission from a station in group i.

Assuming that the frame is successfully transmitted after experiencing j $(j \geq 0)$ collisions, its channel access delay would equal to the delay caused by j unsuccessful transmissions and the $(j + 1)$ back off stages, and can be calculated as the following for each of the groups:

$$E[A_i] = T_c \varphi_i + \sigma_i' \delta_i \tag{5.5}$$

where σ_i' is the average length of a time slot and is individually calculated at different stages. T_c is the collision time, φ_i is the average number of collisions

before a successful transmission from a station, and δ_i represents the average number of time slots the station defers during backoff stages:

$$\varphi_i = \sum_{j=0}^{m} \frac{j\, P_j^i\,(1-P_i)}{\left(1-P_i^{m+1}\right)} \qquad \delta_i = \sum_{j=0}^{m} \sum_{v=0}^{j} \frac{W_v-1}{2} \frac{P_j^i\,(1-P_i)}{\left(1-P_i^{m+1}\right)} \tag{5.6}$$

where P_j^i is the probability that the frame experiences j $(0 \le j \le m)$ collisions, and $(\frac{W_v-1}{2})$ denotes the mean of the backoff counters generated in the v-th $(0 \le v \le j)$ backoff stage. Assume that P_{tri} represents the probability that at least one of the remaining stations transmits in a given time slot when a station in group i is in the backoff procedure. When a station in group i transmits, the value of P_{tri} would be equal to the value of P_i which is given in Eq. (5.3). P_{si}^r is the probability that a station in group r $(1 \le r \le G)$ successfully transmits on the channel, when the station in group i is in the backoff procedure:

$$P_{si}^r = \tau_r(1-\tau_r)^{n_r-1}(1-\tau_i)^{n_i-1}\prod_{g\neq r,i}\left(1-\tau_g\right)^{n_g} \qquad 1 \le r \le G,\ 1 \le g \le G \tag{5.7}$$

When the station is in backoff, the size of the time slot is obtained by considering the fact that the channel is idle with probability:

$$P_{idli} = (1 - P_{tri}) \tag{5.8}$$

And the transmission is successful with probability $\sum_{r=1}^{G} P_{si}^r$. A collision occurs with probability $(P_{tri} - \sum_{r=1}^{G} P_{si}^r)$. Therefore the value of σ_i' is calculated as follows, where σ is the duration of an empty physical time slot as mentioned in [3]:

$$\sigma_i' = (1 - P_{tri})\sigma + \sum_{r=1}^{G} P_{si}^r\, T_s + (P_{tri} - \sum_{r=1}^{G} P_{si}^r)T_c \tag{5.9}$$

In the above formula, T_s is the average time for successful transmission, and T_c is the average time the channel is sensed busy by each station during a collision. The calculations for T_s and T_c are as follows [8]:

$$T_s = T_H + T_L + SIFS + 2\Delta + ACK + DIFS$$
$$T_c = T_H + T_L + DIFS + \Delta \quad (5.10)$$

where T_H is the average time required to transmit the packet header, Δ is the propagation delay and T_L is the average time required to transmit the longest packet payload.

5.2.2. Queueing Analysis of Stations

In order to develop a model for WLAN that has heterogeneous stations generating different types of traffic as well as a bursty multimedia traffic, the stations in some groups are modelled as a $BMAP/M/1/N$ queuing system [71], and in other remaining groups the stations are modelled as $M/G/1/N$ queueing systems. The idea is the same as in any queueing model; when a frame reaches the head of the transmission queue, the server becomes busy and as soon as a frame is acknowledged by the destination following a successful transmission, the server becomes free. The service time is dependent on the size of the frame transmitted and in this paper is modelled by an exponential distribution function with mean $E[S_i]$ where $1 \leq i \leq G$. Thus, the service rate, μ_i, for each of the stations within a group can be calculated as: $\mu_i = 1/E[S_i]$.

In the following the two main traffic models used in the paper are presented along with the definition of the queueing systems.

1) BMAP Queueing Analysis of Stations

For the group of stations generating traffic consistent with high definition video traffic, a BMAP composed of a three-state Markov Chain with maximum possible batch size of three is used. The maximum batch size allows us to model three different packet sizes for every video traffic generated synthetically. As regards to the practical size limits and for ease of modelling and simulation, the maximum batch size and number of the states of the underlying Markov Chain are kept to three in this model. Batches of maximum 3 frames at this stage will generate a good result to contemplate on. Figure 5.1 represents the 3-state CTMC of the model with maximum batch size of 3, and Figure 5.2 illustrates the state transition diagram of the $\mathrm{BMAP/M/1/N}$ queue assuming where the maximum buffer size of each station is N $(N = 50)$, with the maximum batch size 3.

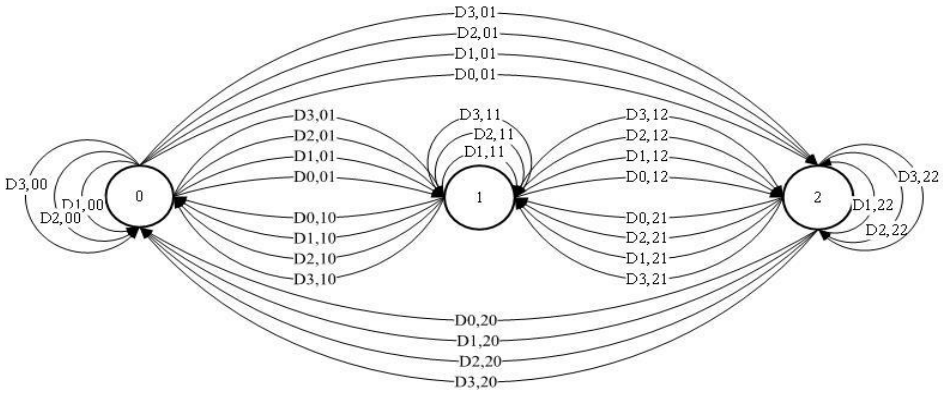


Figure 5.1: Three-state CTMC underlying BMAP with batch size three

State (η, s), where $(s = 0, \dots, N)$ and $(\eta = 0,1,2)$, represent the case that there are s frames in the queuing system and the BMAP characterizing the traffic of the station is at state η. The BMAP is characterized based on the states and the batch size of the arrival process. If the batch size is one, meaning that only one frame arrives at the station at one time, then, the transmission rate from state (η, s) to $(\eta, s+1)$ in the CTMC would be determined by the respective value from the matrix of $D_1(\eta, \eta)$, considering that the arrival rate of the generated frame stays the same (the CTMC is still at the same state). For batch sizes greater than 1, the transmission rate from state (η, s) to $(\eta, s+k)$, would be determined by $D_k(\eta, \eta)$ for $(k > 1)$, again with the arrival rate staying the same as before. However if there is a transition from state η to any other state, e.g. $\eta + 1$, with arrivals of batch size of k, then the transmission rate would be determined by $D_k(\eta, \eta + 1)$. The transition rate out of state η, e.g. from (η, s) to $(\eta, s-1)$, would equal to the service rate μ_i. Whereas the transition rate from state (η, s) to (η', s), where $\eta' = (0,1,2)$ and $\eta' \neq \eta$, would have to be determined by the $D_0\ (\eta, \eta')$ matrix considering there is no arrival at the time of transition, or by $D_k\ (\eta, \eta')$ for $(k > 0)$ if there is any arrival during the transition time.

Following the above analysis, the transition rate matrix, G, of the underlying CTMC of Figure 5.2 can be obtained. The same method can be used for obtaining the transition rate matrix of the superposition of two state BMAPs with maximum batch size of one to model the voice traffics.

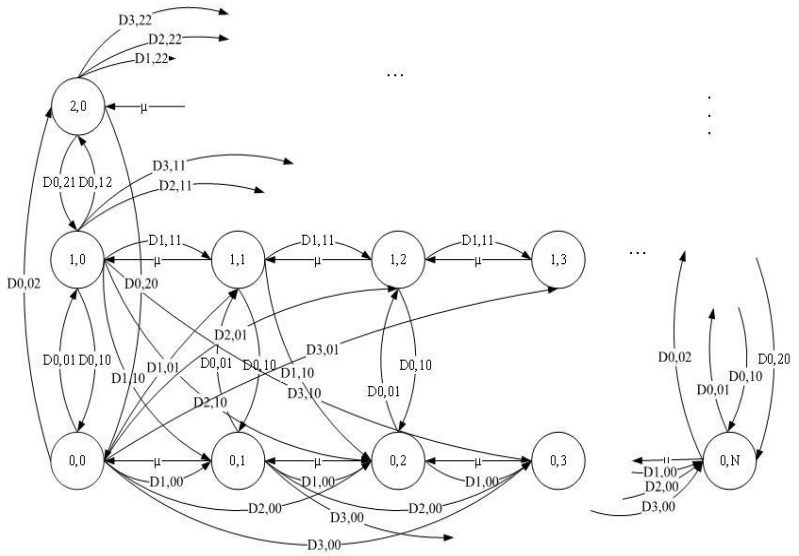


Figure 5.2: State transition diagram of the BMAP/M/1/N (N=50) queue

To calculate the steady state probability vector $P = \left(P_{is,\eta}, s = 0,1,\ldots,N,\ \eta = 0,1,2\right) = (P_0, P_1, \ldots, P_N)$ of the Markov Chain, the transition rate matrix G and the following equations can be used:

$$PG = 0 \quad and \quad Pe = 1 \tag{5.11}$$

where $P_s = \left(P_{is,0},\ P_{is,1},\ P_{is,2}\right),\ 0 \leq s \leq N$. By solving the above equation, P_{i0}, the probability that the transmission queue of a station in group i is empty can be obtained.

The same analysis stands for the stations generating On-Off voice traffic using superposition of M two-state BMAPs with a maximum batch size of one. Figure 5.3 illustrates the underlying CTMC of a two state BMAP with maximum batch size of used for modelling the On-Off voice traffics.

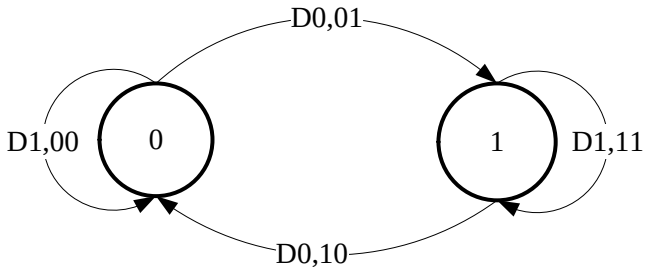


Figure 5.3: Two-state CTMC underlying BMAP with batch size one

2) Poisson Queueing Analysis of Stations

The transmission queue of the stations with non-bursty Poisson traffic (i.e. background data) is modelled as a $M/G/1/N$ queuing system, where N represents the buffer size.

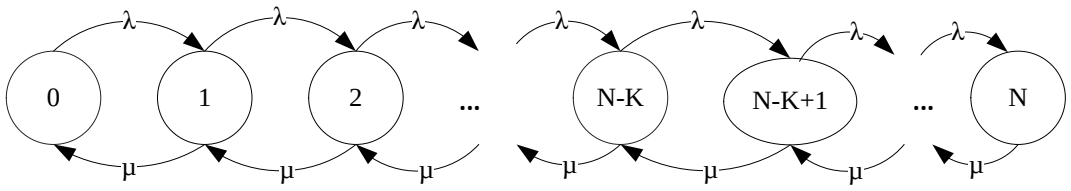


Figure 5.4: State transition diagram of M/G/1/N queue.

Figure 5.4 shows the state-transition-rate diagram of the queuing system for the stations modelled using the M/G/1/N queueing system. The states on the diagram denote the numbers of frames in the system. Transition rate from state s to state $(s + 1)$, $(0 \leq s \leq N - 1)$, is equal to the arrival rate λ of the Poisson process. μ represents the transition rate out of state s to state $(s - K), (K \leq s \leq N)$. This means that the transmission of the frame from a station completes with a mean serving time of μ. The change from state s to 0, $(1 \leq s \leq K - 1)$, indicates that all s frames in the system are transmitted and the transition queue of the station is

empty, hence the unsaturated condition. Again, as before the transition rate matrix, G, of the Markov chain can be obtained from the state transition diagram shown in Figure 5.4. As in BMAP/M/1/N, the steady-state probability vector, $P = (P_{is}, s = 0, 1, \ldots, N)$ of the Markov chain should satisfy Eq. (5.11), from which the probability that the transmission queue of a station in group i is empty can be obtained, P_{i0}.

5.2.3.Performance Measures

Assume π_{si} denotes the steady-state probability of s frames being in the queuing system of a station in group i, which can be calculated as:

$$\pi_{si} = P_s e \quad for \;\; 0 \leq s \leq N \tag{5.12}$$

The probability that the transmission queue of the station in group i is empty, P_{i0}, can be calculated by π_{0i}. Assume π'_{si} represents the steady-state probability of s frames being in the queuing system of a station in group i, when a frame arrives. In the case of BMAP traffic, π'_{si} can be written as:

$$\pi'_{si} = \frac{P_s \Lambda e}{\sum_{s=0}^{N} P_s \Lambda e}, \quad for \;\; 0 \leq s \leq N \tag{5.13}$$

where Λ, for stations modelled using BMAP is a diagonal matrix containing the arrival rates of each state of the underlying CTMC. The values of the Λ matrix for BMAP can be gained from the absolute values of the diagonal elements of the D_0 matrix. For the stations where the traffic is modelled using the Poisson Process, π'_{si} would be equivalent to π_{si}.

The loss probability, which is shown by P_{bi}, can be gained from π'_{Ni}. With $E[P]$ being the frame payload size and λ_i being the fundamental traffic arrival rate of the station in group i, the throughput, TH_i, of the stations in group i can be computed by:

$$TH_i = \lambda_i E[P](1 - P_{bi}) \tag{5.14}$$

Using the Little's Law, the average end-to-end delay of a frame being transmitted from a source to a destination, $E[D_i]$, can be calculated as:

$$E[D_i] = \frac{E[N_i]}{\lambda_i(1-P_{bi})} \tag{5.15}$$

The end-to-end delay is the time interval from the instant that a frame enters the transmission queue of the source station, to the instant that the frame is acknowledged after successful transmission by the destination station. $E[N_i]$ is the average number of frames in the queuing system of the station in group i, which can be computed as:

$$E[N_i] = \sum_{s=1}^{N} s\pi_{si} \tag{5.16}$$

5.3. Model Validation and Performance Evaluation

The accuracy of the developed analytical model has been validated through extensive simulations using the NS2 simulation environment. The newly developed traffic generators are used in NS2 to accommodate the Batch Markovian Arrival Processes for generation of bursty video and voice traffics.

For the flexibility and scalability of the simulation experiments and the analytical model, real-world multimedia applications are adopted using the Expectation-Maximization algorithm presented in [102] to calculate the parameters. The data trace used to generate the required parameters of the video traffic is obtained from the high quality measurement of the video stream for the film “Tears of Steel”, encoded in H.265/HEVC codec [159]. Using the data trace and the EM algorithm [102], parameters for a three-state BMAP with maximum batch size of three are estimated. The estimated infinitesimal generator matrix of the three-state BMAP with batch size of maximum three, Q, and matrix D_0 are as follows:

$$D_0 = \begin{pmatrix} -\lambda_1 & 0.01 & 0.03 \\ 0.05 & -\lambda_2 & 0.08 \\ 0.17 & 0.20 & -\lambda_3 \end{pmatrix} \qquad Q = \begin{pmatrix} -0.17 & 0.07 & 0.10 \\ 0.34 & -0.7 & 0.36 \\ 0.61 & 0.7 & -1.33 \end{pmatrix} \tag{5.17}$$

where $\lambda_1 = 0.7 * \lambda_i$ and $\lambda_2 = 0.2 * \lambda_i$ and $\lambda_3 = 0.1 * \lambda_i$, with λ_i being the overall traffic load of the station in group i.

For modelling the voice traffic, high quality measurements of the G.711 codec [76, 189, 190] is used to obtain the parameters. G.711 has a 64kbps bit rate with packetization intervals of 60ms [189]. This results in the rate of 16.67 packetization/s with a payload size of $64{,}000/(16.67 \times 8) = 480$ Bytes. When modelling G.711 voice sources as an On-Off traffic, the On and Off will be exponentially distributed with mean of $352\,ms$ and $650\,ms$, respectively [19]. The parameters for modelling the superposition of multiple On-Off voice sources using a two-state BMAP can be obtained using Eqs. (2.23) to (2.26).

Simulation experiments are executed in two scenarios. The first scenario, Scenario 1, concentrates on the effect of network size and heterogeneous traffic sources on

performance measure of WLANs, and in a way is a validation of the developed analytical model. The second scenario, Scenario 2, studies the effect of traffic load and heterogeneous traffic sources. In Scenario 2, the overall traffic load of the stations varies over the course of simulation. To transfer the effect of traffic load change, in the models using BMAP, the arrival rates of the D_0 matrix are varied in accordance to the overall load change of each station. Also, based on the overall load of the stations within the model, the number of voice sources super-positioned to generate the On-Off voice traffic dynamically changes so that the overall load of all the stations are kept the same.

5.3.1.Simulation Scenario

The developed model is simulated within a Basic Service Set (BSS) of WLANs where n $(n = 12)$ static stations are distributed within a rectangular $150m \times 150m$ grid, and are classified into three groups with identical number of stations, 4 stations in each group. Each of the groups generates and models a unique network traffic of background data, voice or video. Video traffics generated by a 3-state BMAP with maximum batch size of three, voice generated by superposition of M two-state BMAP with maximum batch size of one, and data traffic generated by the Poisson Process are injected into the MAC buffer of the stations in groups $1,2$ and 3, respectively. Each station generates and transmits traffic to its paired stations. The accuracy of the analytical model is validated under various working conditions which show consistent performance results.

In the simulation scenario, the stations are equipped with the 802.11b physical layer and ideal wireless channel conditions are assumed. The buffer size of all stations is configured to be maximum $N = 50$ frames and the size of all data frame payload is set to be 480 Bytes to match the traffic settings of G.711 voice data traces. The remaining simulation settings of the stations and the WLAN are summarized in Table 5.1. Each simulation is executed for 600 seconds of NS2 simulation time, which is sufficiently long to gain a stable simulation and reliable performance results.

CW_{min}	32	Retry limit (m)	7
CW_{max}	1024	Basic Data Rate	1 Mbps
Slot time	20 μs	Channel Data Rate	11 Mbps
DIFS	50 μs	Propagation delay	2 μs
SIFS	10 μs	ACK Frame Payload	112 bits
MAC hdr	224 bits	PHY Header	192 bits

Table 5.1: Parameters used in the performance analysis

5.3.2.Performance Evaluation

- **Scenario 1: Effect of Network Size on Heterogeneous WLANs**

In this scenario the model and simulation are executed on varying network size, while the load of the stations is kept at $350Kbps$ at all times. In each run the number of stations defined for the WLAN are increased which in turn increases the number of stations forming each of the traffic groups. Figures 5.5 to 5.7 illustrate the results gained for throughout, end-to-end delay and loss probability of each of the traffic groups within the WLAN.

Figure 5.5 shows that the throughput of stations with bursty video traffic is slightly higher than the other two groups as the network size is very small and each group has only two stations. This is logical in the sense that when the network is sparse, all generated traffic will be transmitted to destinations successfully. With BMAP generating highly bursty traffic with higher frame sizes, the throughput of the nodes generating video traffic is higher than the other groups. However, as soon as the network size starts to grow beyond 6 nodes and the network becomes busy, the throughput of the nodes generating video traffic starts to deteriorate and decreases to a lower value compared to the background data and voice. This condition continues to stay the same until the network size reaches around 26 stations. At this stage increase in the number of stations results in the saturation of the network, and the network reaches its capacity when the number of stations reaches 26. From this point onwards, the throughput of all three groups diverts towards a fixed rate of about $0.1\ Mbps$.

The end-to-end delay of the stations within the video traffic group, depicted in Figure 5.6, is slightly higher than the data and voice groups until the network size reaches around 12 stations. From network size of 12 and onwards, the end-to-end delay of the stations in all three groups stays very close and increases at a steady pace. There is a slight difference between the simulations and analytical which could be the result of network saturation. However in general the match between the results gained from simulation and analytical model is acceptable for future analysis.

The loss of the group modelling the video traffic is higher than the other two groups for nearly all network sizes, shown in Figure 5.7. This is logical as BMAP generates

batches of data frames at exponentially distributed times and as a results, the loss probability of stations having their traffics modelled via BMAP would be higher than the other stations. However as the network size grows beyond 30 stations, the network becomes saturated and, as a result, most generated frames are dropped. Due to this fact, the loss probabilities of all three groups of traffics reach the same rate of about 80%.

Overall the results clearly indicate the fact that burstiness and heterogeneity of network traffic has a higher impact when the network size is smalle or moderate. As the network size grows and the number of stations increases, the network reaches a saturation point where all traffic types result in similar network performance measures.

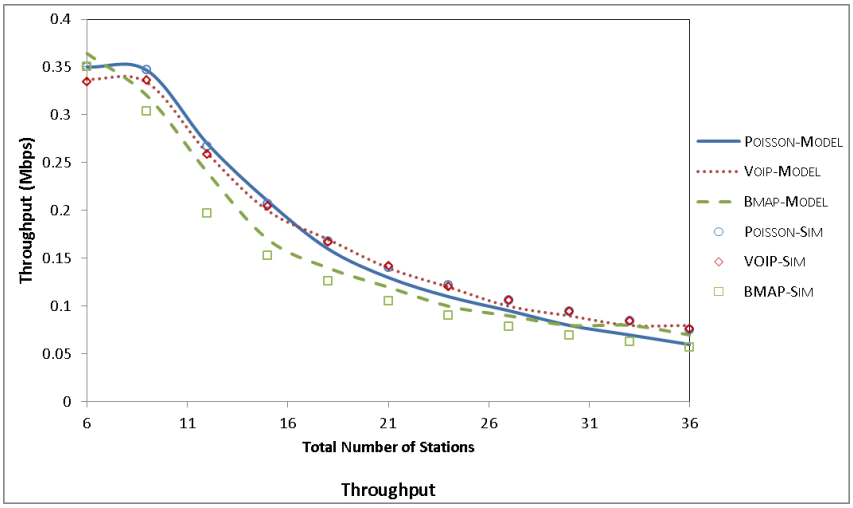


Figure 5.5: Comparison of the Throughput between analytical results and simulation of WLANs with heterogeneous stations under different network size.

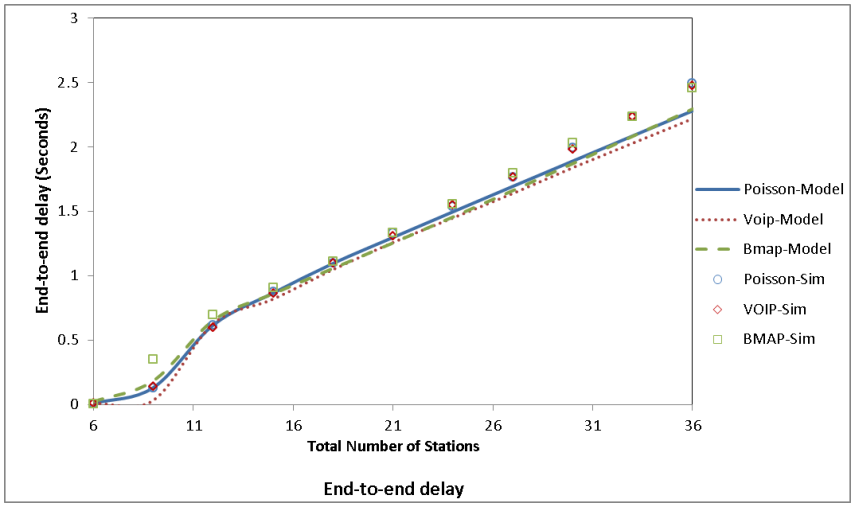


Figure 5.6: Comparison of End-to-end delay between analytical results and simulation of WLANs with heterogeneous stations under different network size.

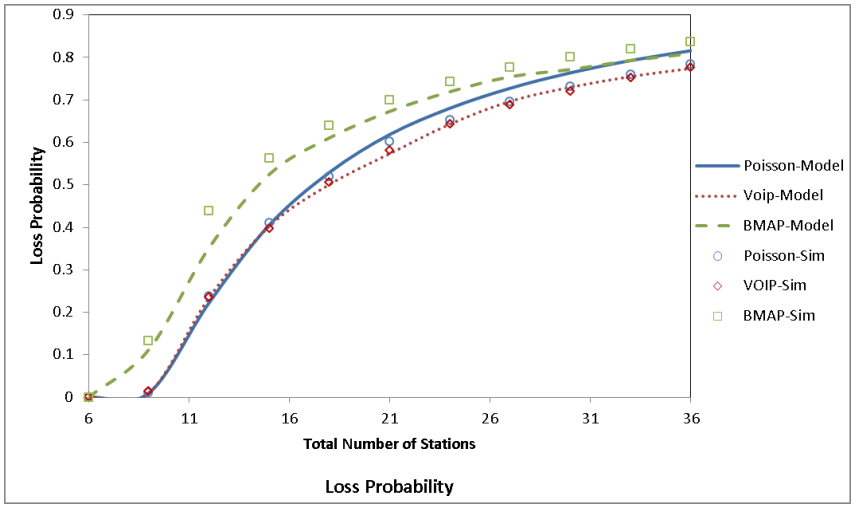


Figure 5.7: Comparison of the Frame Loss Probability between analytical results and simulation of WLANs with heterogeneous stations under different network size.

- **Scenario2: Comparison Between Heterogeneous and Homogeneous Sources**

To investigate the impact of heterogeneous bursty traffic on the performance of WLANs, in this scenario a comparison is carried out between WLANs having

heterogeneous traffic sources and homogeneous traffic sources. The comparison is executed on throughput, end-to-end delay and frame loss probability.

In WLANs with homogeneous traffic sources, all stations generate traffics using either the Poisson Process, superposition of On-Off voice sources, or bursty video traffic with the same BMAP settings of that in the heterogeneous model.

The performance results illustrated in Figures 5.8 to 5.10, reveal the areas where bursty heterogeneous traffics significantly impact the performance of WLANs. Overall, when the network operates at low traffic loads (less than $0.2\ Mbps$) the QoS performance measures of all stations under heterogeneous traffic is close to those operating under homogeneous traffics. This is because of the low collision probability during lower traffic loads, where loss probability and delay are almost zero resulting in the throughput to be very close to the traffic load. However, as the traffic load increases the differences between the QoS performance measures of stations within different traffic groups starts to increase in comparison to stations in homogeneous networks.

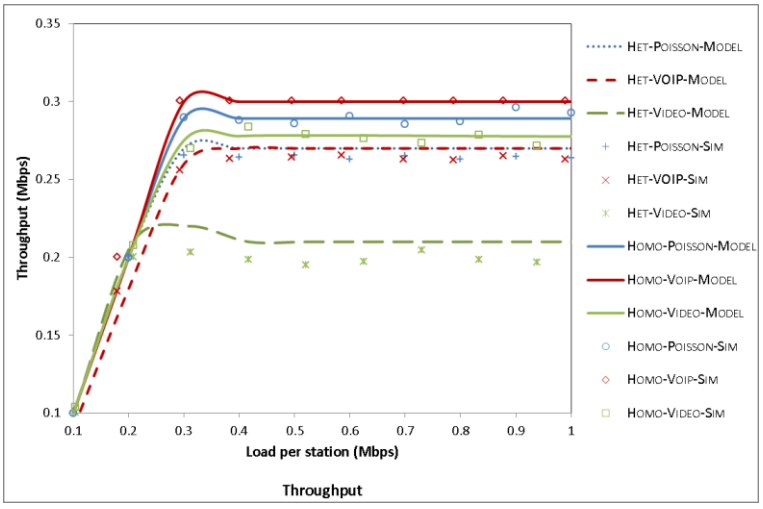


Figure 5.8: Comparison of the Throughput between analytical results and simulation of WLANs with heterogeneous and homogeneous stations.

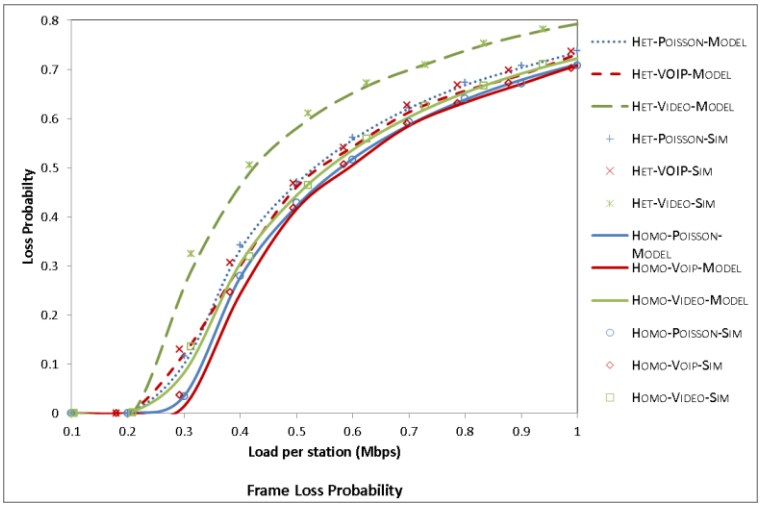


Figure 5.9: Comparison of the Frame Loss Probability between analytical results and simulation of WLANs with heterogeneous and homogeneous stations.

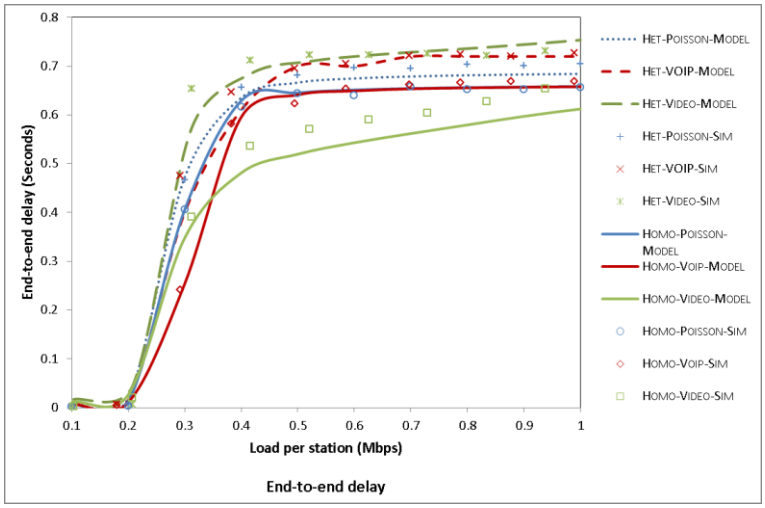


Figure 5.10: Comparison of the End-to-end delay between analytical results and simulation of WLANs with heterogeneous and homogeneous stations.

It is evident from Figure 5.8 that the overall throughput of the stations in all groups of traffic type degrades in a WLAN with heterogeneous traffic sources as soon as the traffic load surpasses $0.2\,Mbps$ and the collisions increase. With BMAP generating bursty traffics with variable frame sizes, it sufferes the most degradation in throughput than the other traffic sources. This effect is also visible in the frame

loss probability shown in Figure 5.9. From the figure it is clear that the stations modelled using BMAP, when in a heterogeneous setting, lose more data due to their higher rate and frame size variation which generates a more bursty traffic compared to On-Off sources or the background traffic.

From Figure 5.10 it is evident that overall the end-to-end delay of the frames transmitted through the WLAN increases when the network is composed of heterogeneous stations compared to homogeneous networks. Again, the end-to-end delay of the frames generated via the BMAP sources has a much higher increase due to the bursty nature of the traffic generated.

5.4. Summary

This chapter has proposed a comprehensive analytical model for 802.11-based WLANs in the presence of unsaturated heterogeneous stations with bursty BMAP and non-bursty Poisson traffics. To obtain the queueing dynamics, the transmission queue at each station generating multimedia traffic has been modelled as a BMAP/M/1/N queueing system, while the stations generating background data have been modelled as M/G/1/N queueing systems. Important QoS performance measures in terms of throughput, end-to-end delay and frame loss probability have been calculated. To validate the accuracy of the proposed analytical model, extensive simulations have been carried out using NS2 simulation experiments. The parameters used for the traffic models have been obtained from accurate measurements of real-world multimedia applications including the G.711 codec voice sources and encoded H.265 video streams.

The analysis of the results clearly demonstrates the importance of adopting heterogeneous traffic sources for accurate performance evaluation of the MAC layer of 802.11 WLANs in the presence of multimedia applications. The developed model could be the basis of deeper analysis on future wireless networks where multiple traffic classes with varying QoS requirements exist in highly dense WLANs.

Chapter 6:

Conclusions and Future Work

Over the past decade, we have witnessed a surge in the development of wireless communication and technologies. Explosive growth in the number of wireless devices such as smartphones, PCs, personal digital assistants and home entertainment systems, along with the rapid formation of advanced multimedia applications have resulted in a revolutionary deployment of the wireless technologies. With all the enhancements taken place on the 802.11 standard, the Medium Access Control mechanism, responsible for wireless medium access control and data transmission has been left almost untouched. Therefore, with multimedia and in particular video traffic exhibiting high burstiness and correlation properties over a wide range of timescales, this thesis focuses on deeper and more accurate performance evaluation methodologies required to capture and analyze QoS performance of the MAC protocol in WLANs which integrate heterogeneous sources of multimedia traffics.

This chapter provides a summary of the works reported in this thesis as well as introducing some future research directions that can stem from the developed analytical models presented in the research.

6.1. Conclusions

This thesis has presented new analytical tools for performance analysis and enhancement of wireless MAC protocols under bursty and correlated multimedia traffic. Throughout the thesis, the accuracy of the developed models have been validated through extensive simulation experiments developed using the NS2 simulator. The proposed analytical models have been used to investigate important QoS performance measures of 802.11 MAC protocols under bursty and correlated traffics. The major achievements in this research are summarized as follows:

- To develop a reliable foundation that formed the basis of later analysis and studies, new traffic generators were developed and thoroughly tested in Chapter 3 for multi-state Batch Markovian Arrival Process with any maximum batch size and multi-state MMPP. The traffic generators were designed and implemented in C++ programming language. To test and validate the developed traffic generators, they were put into test in BMAP/M/1 and MMPP/M/1 queues. The BMAP/M/1 queue was tested under various conditions to find the effects of different settings on the performance of the queue, such as:
    - Effect of different values of coefficient of correlation
    - Effect of the number of states of the underlying Markov chain
    - And effect of different batch sizes

  Since previously developed models and simulators of 2-state MMPP were extensively used in literature, the developed m-state MMPP/M/1 queue was

used as a reliable evaluation tool. This was due to the fact that MMPP is a subclass of Batch Markovian Arrival Process.

- The newly developed traffic generators of Chapter 3 were put in to use to develop a new analytical model for performance evaluation of the MAC protocol in the IEEE 802.11 standard under unsaturated traffic loads and finite buffer capacity. The QoS performance metrics including throughput, end-to-end delay, frame loss probability and energy consumption of the analytical model were derived in Chapter 4. To validate the results of the analytical model, the BMAP and m-state MMPP where implemented in to NS2 network simulator using TCL and C++ programming. Using the simulator and the analytical model, a thorough investigation into the impact of multiple settings on the QoS performance of WLANs were executed:

    - A comprehensive comparison was carried out between the effect of traffics generated using a 3-sate BMAP, 3-state MMPP and Poisson Process on the performance of WLANs.
    - Effect of buffer size was investigated with stations of WLAN generating traffic using a 3-state BMAP with maximum batch size of three.
    - Effect of traffic burstiness was investigated on the WLANs through different maximum batch sizes of BMAP: 3, 5 and 10.
    - And finally the effect of network size was investigated on WLANs with stations generating bursty traffic using a three-state BMAP with maximum batch sizes of three, four, five and ten.

- In Chapter 5 a new analytical model was developed for the MAC DCF protocol in unsaturated WLANs with heterogeneous traffic sources using bursty BMAP and non-bursty Poisson process to model multimedia applications. Again the results gained from the analytical model for throughput, end-to-end delay and frame loss probability were extensively validated using simulation experiments in NS2 subject to the traffic parameters obtained from the accurate measurements of the real-world multimedia voice and video sources. For this model:
    - The impacts of traffic load was investigated when background data was modelled by non-bursty Poisson Process, Voice traffic was modelled using a superposition of two-state BMAPs with maximum batch sizes of one, and the video traffic was modeled using a three-state BMAP with maximum batch size of three.
    - The effect of network size on the performance of the 802.11 MAC was studied when the WLAN was composed of heterogeneous and bursty traffic sources.

The performance results have highlighted the importance of taking into account the heterogeneous traffic for the accurate evaluation and design of the MAC protocol in wireless multimedia networks.

6.2. Future Work

The thesis mainly investigates the QoS performance of MAC protocols in WLANs with multimedia applications. Although the work has emphasized on the main

research objectives, the following future works can be suggested to extend this research into accommodating emerging wireless networks and more general working scenarios.

- **Device-to-Device Communications over Wi-Fi Direct**

Advancements in cellular communication systems have resulted in the emergence of Device-to-Device (D2D) technology for the purpose of significantly improving the performance of cellular systems. D2D enables devices of proximity to directly communicate with each other, therefore mitigating the system overhead while increasing the spectrum utilization through bypassing cellular Base Stations (BSs) or Access Points (APs) [191, 192]. D2D technology facilitates mobile users with instant information sharing (e.g. pictures and videos) even in areas without cellular coverage or APs. With the emergence of the so called mobile ad-hoc clouds as a result of D2D technology [193], many open challenges still exist in the efficiency of the design and deployment of D2D connections over wireless technology. To this end using the appropriate traffic processes to model the data exchanged between the digital devices could result in increased accuracy and efficiency of future designs.

- **Traffic offloading and resource sharing**

Increase in mobile data traffic has placed immense pressure on capacity improvement of heterogeneous networks especially cellular networks. To alleviate this pressure, new techniques of traffic offloading and resource sharing have been

proposed in literature such as massive multiple-input multiple-output (MIMO), heterogeneous Dense WiFis, direct device-to-device communications, etc. In spite of these cutting-edge techniques, the limited licensed spectrum is still the principal bottleneck for capacity improvement. To alleviate the existing problems, Wi-Fi offloading is envisioned as a promising solution to utilize the various benefits of Wi-Fi and cellular networks together via the migration of traffic from cellular to Wi-Fi networks. Even though traffic offloading uses the unlicensed bands for delivering cellular data traffic [194-197], due to the existence of the DCF protocol in the MAC layer, guaranteeing the QoS of cellular traffic is a challenging issue. Moreover, the type and volume of the offloaded traffic plays a pivotal role in the efficiency of this process; therefore an accurate process should be designed that avoids over saturation and excessive packet collisions within the Wi-Fi network. For this purpose, accurate analytical models are required to model the miscellaneous types of traffics generated and offloaded onto Wi-Fi in order to increase the efficiency of the designs.

- **Software-Defined Networking:**

Software Defined Networking (SDN) is an emerging architecture with promising properties for the next-generation Internet [198-200]. SDN decouples the control plane, responsible for making network forwarding decisions, from the data plane, responsible for data forwarding. This decoupling enables more centralized control where coordinated decisions directly guide the network to desired operating conditions. The unprecedented network programmability provided by SDN provides the grounds for handling the explosive growth of data generated by smart mobile

devices and the pervasiveness of content-rich multimedia applications. However to develop a realistic tractable analytical model that takes into account the real characteristics of traffics generated by multimedia applications, new processes such as BMAP should be taken into account to model the traffic transmitting through these networks.

References

[1] "IEEE Standard for Information technology-- Telecommunications and information exchange between systemsLocal and metropolitan area networks-- Specific requirements--Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications--Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz," *IEEE Std 802.11ac-2013 (Amendment to IEEE Std 802.11-2012, as amended by IEEE Std 802.11ae-2012, IEEE Std 802.11aa-2012, and IEEE Std 802.11ad-2012),* pp. 1-425, 2013.

[2] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html. (2016, may). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Paper*.

[3] "IEEE Standard for Information Technology- Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," *IEEE Std 802.11-1997,* pp. i-445, 1997.

[4] L. Kleinrock and F. Tobagi, "Packet Switching in Radio Channels: Part I - Carrier Sense Multiple-Access Modes and Their Throughput-Delay Characteristics," *IEEE Transactions on Communications,* vol. 23, pp. 1400-1416, 1975.

[5] "IEEE Draft Standard for Information technology--Telecommunications and information exchange between systems Local and metropolitan area networks--Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment: Pre-Association Discovery," *IEEE P802.11aq/D3.0 October 2015 (Amendment to IEEE Std 802.11REVmc, as amended by IEEE Std 802.11ah-2016 and IEEE Std 802.11ai-2016),* pp. 1-46, 2016.

[6] "ISO/IEC/IEEE International Standard for Information technology--Telecommunications and information exchange between systems--Local and metropolitan area networks--Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band (adoption of IEEE Std 802.11ad-2012)," *ISO/IEC/IEEE 8802-11:2012/Amd.3:2014(E),* pp. 1-634, 2014.

[7] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications,* vol. 43, pp. 1566-1579, 1995.

[8] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications,* vol. 18, pp. 535-547, 2000.

[9] L. Qilian, "Ad hoc wireless network traffic-self-similarity and forecasting," *IEEE Communications Letters,* vol. 6, pp. 297-299, 2002.

[10] O. Tickoo and B. Sikdar, "On the impact of IEEE 802.11 MAC on traffic characteristics," *IEEE Journal on Selected Areas in Communications,* vol. 21, pp. 189-203, 2003.

[11] R. A. Joarder, S. Parveen, H. Sarwar, S. K. Sanyal, and S. Rafique, "Analysis of real-time multimedia traffic in the context of self-similarity," in *Electrical and Computer Engineering, 2008. ICECE 2008. International Conference on*, 2008, pp. 618-623.

[12] H. T. William, P. T. Desmond, E. Z. Rodger, F. M. Nicholas, and W. M. Jon, "On the SelfSimilar Nature of Ethernet Traffic (Extended Version)," in *The Best of the Best:Fifty Years of Communications and Networking Research*, ed: Wiley-IEEE Press, 2007, pp. 517-531.

[13] K. Park and W. Willinger, "Self-Similar Network Traffic: An Overview," in *Self-Similar Network Traffic and Performance Evaluation*, ed: John Wiley & Sons, Inc., 2002, pp. 1-38.

[14] A. Abdrabou and W. Zhuang, "Service Time Approximation in IEEE 802.11 Single-Hop Ad Hoc Networks," *IEEE Transactions on Wireless Communications,* vol. 7, pp. 305-313, 2008.

[15] G. V. d. Auwera, P. T. David, and M. Reisslein, "Traffic and Quality Characterization of Single-Layer Video Streams Encoded with the H.264/MPEG-4 Advanced Video Coding Standard and Scalable Video Coding Extension," *IEEE Transactions on Broadcasting,* vol. 54, pp. 698-718, 2008.

[16] G. V. d. Auwera, P. T. David, and M. Reisslein, "Traffic characteristics of H.264/AVC variable bit rate video," *IEEE Communications Magazine,* vol. 46, pp. 164-174, 2008.

[17] F. H. Li, Y. Xiao, and J. Zhang, "Variable bit rate VOiP in IEEE 802.11e wireless LANs," *IEEE Wireless Communications,* vol. 15, pp. 56-62, 2008.

[18] G. Min, J. Hu, and M. E. Woodward, "A Dynamic IEEE 802.11e TXOP Scheme in WLANs under Self-Similar Traffic: Performance Enhancement and Analysis," in *2008 IEEE International Conference on Communications*, 2008, pp. 2632-2636.

[19] J. W. So, "Performance Analysis of VoIP Services in the IEEE 802.16e OFDMA System With Inband Signaling," *IEEE Transactions on Vehicular Technology,* vol. 57, pp. 1876-1886, 2008.

[20] D. M. Lucantoni, "The BMAP/G/1 QUEUE: A Tutorial," presented at the Performance Evaluation of Computer and Communication Systems, Joint Tutorial Papers of Performance '93 and Sigmetrics '93, 1993.

[21] A. Chydzinski and R. Winiarczyk, "On the blocking probability in batch Markovian arrival queues," *Microprocess. Microsyst.,* vol. 32, pp. 45-52, 2008.

[22] E. Charfi, L. Chaari, and L. Kamoun, "PHY/MAC Enhancements and QoS Mechanisms for Very High Throughput WLANs: A Survey," *IEEE Communications Surveys & Tutorials,* vol. 15, pp. 1714-1735, 2013.

[23] K. Park and W. Willinger, *Self-similar network traffic and performance evaluation*: Wiley Online Library, 2000.

[24] D. Malone, K. Duffy, and D. Leith, "Modeling the 802.11 Distributed Coordination Function in Nonsaturated Heterogeneous Conditions," *IEEE/ACM Transactions on Networking,* vol. 15, pp. 159-172, 2007.

[25] O. Tickoo and B. Sikdar, "Modeling Queueing and Channel Access Delay in Unsaturated IEEE 802.11 Random Access MAC Based Wireless Networks," *IEEE/ACM Transactions on Networking,* vol. 16, pp. 878-891, 2008.

[26] Y. Gao, C. W. Tan, Y. Huang, Z. Zeng, and P. R. Kumar, "Characterization and Optimization of Delay Guarantees for Real-Time Multimedia Traffic Flows in IEEE 802.11 WLANs," *IEEE Transactions on Mobile Computing,* vol. 15, pp. 1090-1104, 2016.

[27] *The Network Simulator–NS-2*.

[28] W. Stallings, *Wireless communications & networks*: Pearson Education India, 2009.

[29] "IEEE Standard for Information technology--Local and metropolitan area networks--Specific requirements--Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications - Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements," *IEEE Std 802.11e-2005 (Amendment to IEEE Std 802.11, 1999 Edition (Reaff 2003),* pp. 1-212, 2005.

[30] B. Bellalta, A. Vinel, P. Chatzimisios, R. Bruno, and C. Wang, "Research advances and standardization activities in WLANs," *Computer Communications,* vol. 39, pp. 1-2, 2/15/ 2014.

[31] "IEEE Standard for Information technology-- Local and metropolitan area networks-- Specific requirements-- Part 11: Wireless LAN Medium Access Control (MAC)and Physical Layer (PHY) Specifications Amendment 5: Enhancements for Higher Throughput," *IEEE Std 802.11n-2009 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009),* pp. 1-565, 2009.

[32] C. T. Report, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Up-date,2013-2018," www.cisco.com2014.

[33] M. Natkaniec, K. Kosek-Szott, S. Szott, J. Gozdecki, A. Głowacz, and S. Sargento, "Supporting QoS in Integrated Ad-Hoc Networks," *Wireless Personal Communications,* vol. 56, pp. 183-206, 2011.

[34] IEEE, "Part 11:Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer in the 5 GHz Band, IEEE Std. 802.11a/d7.0, 1999.," ed, 1999.

[35] F. Tobagi and L. Kleinrock, "Packet Switching in Radio Channels: Part II - The Hidden Terminal Problem in Carrier Sense Multiple-Access and the Busy-Tone Solution," *IEEE Transactions on Communications,* vol. 23, pp. 1417-1433, 1975.

[36] "IEEE Standard for Information Technology - Telecommunications and information exchange between systems - Local and Metropolitan networks - Specific requirements - Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications: Higher Speed Physical Layer (PHY) Extension in the 2.4 GHz band," *IEEE Std 802.11b-1999,* pp. 1-96, 2000.

[37] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communications Magazine,* vol. 32, pp. 70-81, 1994.

[38] H. Heffes and D. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer

performance," *IEEE Journal on selected areas in communications,* vol. 4, pp. 856-868, 1986.

[39] M. F. Neuts, "Modelling Data Traffic Streams," in *Teletraffic and Datatraffic in a Period of Change: ITC-13: Proceedings of the Thirteenth International Teletraffic Congress, Copenhagen, Denmark, June 19-26, 1991*, 1991, p. 1.

[40] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," *SIGCOMM Comput. Commun. Rev.,* vol. 23, pp. 183-193, 1993.

[41] P. Salvador, A. Pacheco, and R. Valadas, "Modeling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs," *Computer Networks,* vol. 44, pp. 335-352, 2004.

[42] R. Puigjaner, N. N. Savino, and B. Serra, *Computer Performance Evaluation: Modelling Techniques and Tools*: Springer, 2003.

[43] J. L. Peterson, "Petri net theory and the modeling of systems," 1981.

[44] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi, *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*: John Wiley & Sons, 2006.

[45] P. Nicopolitidis, F. Zarai, and M. S. Obaidat, "Modeling and Simulation of Computer Networks and Systems," ed: Morgan Kaufmann, 2015.

[46] A. Adas, "Traffic models in broadband networks," *IEEE Communications Magazine,* vol. 35, pp. 82-89, 1997.

[47] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: evidence and possible causes," *IEEE/ACM Transactions on networking,* vol. 5, pp. 835-846, 1997.

[48] C. Oliveira, K. Jaime Bae, and T. Suda, "Long-range dependence in IEEE 802.11b wireless LAN traffic: an empirical study," in *Computer*

Communications, 2003. CCW 2003. Proceedings. 2003 IEEE 18th Annual Workshop on, 2003, pp. 17-23.

[49] M. U. Ilyas and H. Radha, "Long Range Dependence of IEEE 802.15.4 Wireless Channels," in *2008 IEEE International Conference on Communications*, 2008, pp. 4261-4265.

[50] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on networking,* vol. 2, pp. 1-15, 1994.

[51] T. Karagiannis, M. Molle, and M. Faloutsos, "Long-range dependence ten years of Internet traffic modeling," *IEEE Internet Computing,* vol. 8, pp. 57-64, 2004.

[52] P. Abry, R. Baraniuk, P. Flandrin, R. Riedi, and D. Veitch, "Multiscale nature of network traffic," *IEEE Signal Processing Magazine,* vol. 19, pp. 28-46, 2002.

[53] X. Sun, H. Cui, R. Liu, J. Chen, and Y. Liu, "Multistep ahead prediction for real-time VBR video traffic using deterministic echo state network," in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, 2012, pp. 928-931.

[54] O. C. Kwon, Y. Go, and H. Song, "An Energy-Efficient Multimedia Streaming Transport Protocol Over Heterogeneous Wireless Networks," *IEEE Transactions on Vehicular Technology,* vol. 65, pp. 6518-6531, 2016.

[55] G. Min and M. Ould-Khaoua, "A performance model for wormhole-switched interconnection networks under self-similar traffic," *IEEE Transactions on Computers,* vol. 53, pp. 601-613, 2004.

[56] X. Jin and G. Min, "Performance Modelling of Hybrid PQ-GPS Systems under Long-Range Dependent Network Traffic," *IEEE Communications Letters,* vol. 11, pp. 446-448, 2007.

[57] X. Jin and G. Min, "Modelling and analysis of priority queueing systems with multi-class self-similar network traffic: a novel and efficient queue-decomposition approach," *IEEE Transactions on Communications,* vol. 57, pp. 1444-1452, 2009.

[58] M. Abu-Tair, G. Min, Q. Ni, and H. Liu, "An adaptive medium access control scheme for mobile ad hoc networks under self-similar traffic," *The Journal of Supercomputing,* vol. 53, pp. 212-230, 2010.

[59] H. Daehyoung and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology,* vol. 35, pp. 77-92, 1986.

[60] S. H. Nguyen, H. L. Vu, and L. L. H. Andrew, "Performance Analysis of IEEE 802.11 WLANs With Saturated and Unsaturated Sources," *IEEE Transactions on Vehicular Technology,* vol. 61, pp. 333-345, 2012.

[61] L. Dong, Y. Shu, H. Chen, and M. Ma, "Packet delay analysis on IEEE 802.11 DCF under finite load traffic in multi-hop ad hoc networks," *Science in China Series F: Information Sciences,* vol. 51, pp. 408-416, 2008.

[62] A. L. Kleinrock, *Theory, Volume 1, Queueing Systems* vol. 1: Wiley-Interscience 1975.

[63] W. Willinger, V. Paxson, and M. S. Taqqu, "Self-similarity and heavy tails: structural modeling of network traffic," in *A practical guide to heavy tails*, J. A. Robert, E. F. Raisa, and S. T. Murad, Eds., ed: Birkhauser Boston Inc., 1998, pp. 27-53.

[64] W. A. Gardner, *Introduction to Random Processes: with applications to signals and systems*: Macmillan Publishing Company, 1986.

[65] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts, "A single-server queue with server vacations and a class of non-renewal arrival processes," *Advances in Applied Probability,* pp. 676-705, 1990.

[66] K. S. Trivedi, *Probability & statistics with reliability, queuing and computer science applications*: John Wiley & Sons, 2008.

[67] M. F. Neuts, "A versatile Markovian point process," *Journal of Applied Probability,* pp. 764-779, 1979.

[68] Q.-L. Li and Y. Q. Zhao, "A MAP/G/1 queue with negative customers," *Queueing Systems,* vol. 47, pp. 5-43, 2004.

[69] S. Tang and B. L. Mark, "Analysis of opportunistic spectrum sharing with markovian arrivals and phase-type service," *IEEE Transactions on Wireless Communications,* vol. 8, pp. 3142-3150, 2009.

[70] S. Yang, W. Song, and Z. Zhong, "Packet-Level Performance Analysis for Video Traffic over Two-Hop Mobile Hotspots," *IEEE Wireless Communications Letters,* vol. 1, pp. 137-140, 2012.

[71] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process," *Communications in Statistics. Stochastic Models,* vol. 7, pp. 1-46, 1991.

[72] M. F. Ncuts, "ALGORITHMIC PROBABILITY: A Collection of Problems," *Journal of Applied Mathematics and Stochastic Analysis,* vol. 9, pp. 229-230, 1996.

[73] H. Okamura, T. Dohi, and K. S. Trivedi, "Markovian Arrival Process Parameter Estimation With Group Data," *IEEE/ACM Transactions on Networking,* vol. 17, pp. 1326-1339, 2009.

[74] Z. Ye, R. E.-. Azouzi, and T. Jimenez, "Analysis and modelling Quality of Experience of video streaming under time-varying bandwidth," in *2016 9th IFIP Wireless and Mobile Networking Conference (WMNC)*, 2016, pp. 145-152.

[75] G. Min, J. Hu, W. Jia, and M. E. Woodward, "Performance Analysis of the TXOP Scheme in IEEE 802.11e WLANs with Bursty Error Channels," in

2009 IEEE Wireless Communications and Networking Conference, 2009, pp. 1-6.

[76] G. Min, J. Hu, and M. E. Woodward, "Performance Modelling and Analysis of the TXOP Scheme in Wireless Multimedia Networks with Heterogeneous Stations," *IEEE Transactions on Wireless Communications,* vol. 10, pp. 4130-4139, 2011.

[77] J. Hu, G. Min, and M. E. Woodward, "Analysis and Comparison of Burst Transmission Schemes in Unsaturated 802.11e WLANs," in *IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference*, 2007, pp. 5133-5137.

[78] L. Liu, X. Jin, and G. Min, "Modelling an Integrated Scheduling Scheme under Bursty MMPP Traffic," in *Advanced Information Networking and Applications Workshops, 2009. WAINA '09. International Conference on*, 2009, pp. 212-217.

[79] U. Yechiali and P. Naor, "Queuing problems with heterogeneous arrivals and service," *Operations Research,* vol. 19, pp. 722-734, 1971.

[80] M. F. Neuts, "A queue subject to extraneous phase changes," *Advances in Applied Probability,* pp. 78-119, 1971.

[81] I. Ichiro, "Superposition of interrupted Poisson processes and its application to packetized voice multiplexers," *Teletraffic Science IT C-12,* pp. 1399-1405, 1989.

[82] A. T. Andersen and B. F. Nielsen, "A Markovian approach for modeling packet traffic with long-range dependence," *IEEE journal on Selected Areas in Communications,* vol. 16, pp. 719-732, 1998.

[83] S. Shah-Heydari and T. Le-Ngoc, "MMPP modeling of aggregated ATM traffic," in *Electrical and Computer Engineering, 1998. IEEE Canadian Conference on*, 1998, pp. 129-132.

[84] L. Muscariello, M. Meillia, M. Meo, M. A. Marsan, and R. L. Cigno, "An MMPP-based hierarchical model of Internet traffic," in *Communications, 2004 IEEE International Conference on*, 2004, pp. 2143-2147 Vol.4.

[85] W. Fischer and K. Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook," *Performance Evaluation,* vol. 18, pp. 149-171, 1993/09/01 1993.

[86] M. Van Hoorn and L. Seelen, "The SPP/G/1 queue: a single server queue with a switched Poisson process as input process," *Operations-Research-Spektrum,* vol. 5, pp. 207-218, 1983.

[87] I. W. Habib and T. N. Saadawi, "Access control of bursty voice traffic in ATM networks," *Computer Networks and ISDN Systems,* vol. 27, pp. 1411-1427, 1995/09/01 1995.

[88] J.-W. So, "Performance analysis of VoIP services in the IEEE 802.16 e OFDMA system with inband signaling," *IEEE Transactions on Vehicular Technology,* vol. 57, pp. 1876-1886, 2008.

[89] M. Schwartz, *Broadband integrated networks* vol. 19: Prentice Hall PTR New Jersey, 1996.

[90] A. Graham, "Kronecker Products and Matrix Calculus: With Applications," *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, 1982, 130,* 1982.

[91] V. Ramaswami, "The N/G/1 queue and its detailed analysis," *Advances in Applied Probability,* pp. 222-261, 1980.

[92] S. Asmussen and G. Koole, "Marked point processes as limits of Markovian arrival streams," *Journal of Applied Probability,* pp. 365-372, 1993.

[93] V. Subramanian and R. Srikant, "Tail probabilities of low-priority waiting times and queue lengths in MAP/GI/1 queues," *Queueing systems,* vol. 34, pp. 215-236, 2000.

[94] S. Shioda, "Departure process of the MAP/SM/1 queue," *Queueing systems,* vol. 44, pp. 31-50, 2003.

[95] I. J.-B. F. Adan and V. G. Kulkarni, "Single-server queue with Markov-dependent inter-arrival and service times," *Queueing Systems,* vol. 45, pp. 113-134, 2003.

[96] H. W. Lee, S. H. Cheon, E. Y. Lee, and K. C. Chae, "Workload and waiting time analyses of MAP/G/1 queue under D-policy," *Queueing Systems,* vol. 48, pp. 421-443, 2004.

[97] A. Dudin and V. Klimenok, "A Retrial BMAP/PH/N queueing system with Markov modulated retrials," in *Future Internet Communications (BCFIC), 2012 2nd Baltic Congress on*, 2012, pp. 246-251.

[98] A. N. Dudin, A. A. Shaban, and V. I. Klimenok, "Analysis of a queue in the BMAP/G/1/N system," *International Journal of Simulation,* vol. 6, pp. 13-23, 2005.

[99] Q.-L. Li, Y. Ying, and Y. Q. Zhao, "A BMAP/G/1 retrial queue with a server subject to breakdowns and repairs," *Annals of Operations Research,* vol. 141, pp. 233-270, 2006.

[100] A. N. Dudin, V. M. Vishnevsky, and J. V. Sinjugina, "Analysis of the BMAP/G/1 queue with gated service and adaptive vacations duration," *Telecommunication Systems,* vol. 61, pp. 403-415, 2016.

[101] J. Rodríguez, R. E. Lillo, and P. Ramírez-Cobo, "Dependence patterns for modeling simultaneous events," *Reliability Engineering & System Safety,* vol. 154, pp. 19-30, 10// 2016.

[102] A. Klemm, C. Lindemann, and M. Lohmann, "Modeling IP traffic using the batch Markovian arrival process," *Performance Evaluation,* vol. 54, pp. 149-173, 2003.

[103] R. Gusella, "Characterizing the variability of arrival processes with indexes of dispersion," *IEEE Journal on Selected Areas in Communications,* vol. 9, pp. 203-211, 1991.

[104] K. S. Meier-Hellstern, "A fitting algorithm for Markov-modulated Poisson processes having two arrival rates," *European Journal of Operational Research,* vol. 29, pp. 370-377, 1987.

[105] T. Rydén, "Parameter estimation for Markov modulated Poisson processes," *Stochastic Models,* vol. 10, pp. 795-829, 1994.

[106] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological),* pp. 1-38, 1977.

[107] L. Deng and J. W. Mark, "Parameter estimation for Markov modulated Poisson processes via the EM algorithm with time discretization," *Telecommunication Systems,* vol. 1, pp. 321-338, 1993.

[108] T. Rydén, "An EM algorithm for estimation in Markov-modulated Poisson processes," *Computational Statistics & Data Analysis,* vol. 21, pp. 431-447, 1996.

[109] W. J. Roberts, Y. Ephraim, and E. Dieguez, "On Rydén's EM algorithm for estimating MMPPs," *IEEE Signal Processing Letters,* vol. 13, p. 373, 2006.

[110] D. P. Heyman and D. Lucantoni, "Modeling multiple IP traffic streams with rate limits," *IEEE/ACM transactions on networking,* vol. 11, pp. 948-958, 2003.

[111] L. Breuer, "An EM Algorithm for Batch Markovian Arrival Processes and its Comparison to a Simpler Estimation Procedure," *Annals of Operations Research,* vol. 112, pp. 123-138, 2002.

[112] Y. Ephraim and N. Merhav, "Hidden markov processes," *IEEE Transactions on information theory,* vol. 48, pp. 1518-1569, 2002.

[113] E. Ziouva and T. Antonakopoulos, "CSMA/CA performance under high traffic conditions: throughput and delay analysis," *Computer Communications,* vol. 25, pp. 313-321, 2/15/ 2002.

[114] F. Chuan Heng and J. W. Tantra, "Comments on IEEE 802.11 saturation throughput analysis with freezing of backoff counters," *IEEE Communications Letters,* vol. 9, pp. 130-132, 2005.

[115] X. Yang, "Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs," *IEEE Transactions on Wireless Communications,* vol. 4, pp. 1506-1515, 2005.

[116] T. Sakurai and H. L. Vu, "MAC Access Delay of IEEE 802.11 DCF," *IEEE Transactions on Wireless Communications,* vol. 6, pp. 1702-1710, 2007.

[117] J. S. Vardakas, M. K. Sidiropoulos, and M. D. Logothetis, "Performance behaviour of IEEE 802.11 distributed coordination function," *IET circuits, devices & systems,* vol. 2, pp. 50-59, 2008.

[118] I. Tinnirello, G. Bianchi, and Y. Xiao, "Refinements on IEEE 802.11 Distributed Coordination Function Modeling Approaches," *IEEE Transactions on Vehicular Technology,* vol. 59, pp. 1055-1067, 2010.

[119] F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin, "Unsaturated Throughput Analysis of IEEE 802.11 in Presence of Non Ideal Transmission Channel and Capture Effects," *IEEE Transactions on Wireless Communications,* vol. 7, pp. 1276-1286, 2008.

[120] F. Daneshgaran, M. Laddomada, F. Mesiti, and M. Mondin, "On the Linear Behaviour of the Throughput of IEEE 802.11 DCF in Non-Saturated Conditions," *IEEE Communications Letters,* vol. 11, pp. 856-858, 2007.

[121] X. Zhang, "A New Method for Analyzing Nonsaturated IEEE 802.11 DCF Networks," *IEEE Wireless Communications Letters,* vol. 2, pp. 243-246, 2013.

[122] M. Natkaniec, K. Kosek-Szott, S. Szott, and G. Bianchi, "A Survey of Medium Access Mechanisms for Providing QoS in Ad-Hoc Networks," *IEEE Communications Surveys & Tutorials,* vol. 15, pp. 592-620, 2013.

[123] H. Zhai, Y. Kwon, and Y. Fang, "Performance analysis of IEEE 802.11 MAC protocols in wireless LANs: Research Articles," *Wirel. Commun. Mob. Comput.,* vol. 4, pp. 917-931, 2004.

[124] K. Duffy, D. Malone, and D. J. Leith, "Modeling the 802.11 distributed coordination function in non-saturated conditions," *IEEE Communications Letters,* vol. 9, pp. 715-717, 2005.

[125] K. Medepalli and F. A. Tobagi, "Towards Performance Modeling of IEEE 802.11 Based Wireless Networks: A Unified Framework and Its Applications," in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, 2006, pp. 1-12.

[126] J. Hu, G. Min, and M. E. Woodward, "Modeling of IEEE 802.11e Contention Free Bursting Scheme with Heterogeneous Stations," in *2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, 2007, pp. 88-94.

[127] J. Hu, G. Min, M. E. Woodward, and W. Jia, "A Comprehensive Analytical Model for IEEE 802.11e QoS Differentiation Schemes under Unsaturated Traffic Loads," in *2008 IEEE International Conference on Communications*, 2008, pp. 241-245.

[128] E. Felemban and E. Ekici, "Single Hop IEEE 802.11 DCF Analysis Revisited: Accurate Modeling of Channel Access Delay and Throughput for Saturated and Unsaturated Traffic Cases," *IEEE Transactions on Wireless Communications,* vol. 10, pp. 3256-3266, 2011.

[129] P. K. Wong, D. Yin, and T. T. Lee, "Analysis of Non-Persistent CSMA Protocols with Exponential Backoff Scheduling," *IEEE Transactions on Communications,* vol. 59, pp. 2206-2214, 2011.

[130] L. Feng, J. Li, and X. Lin, "A New Delay Analysis for IEEE 802.11 PCF," *IEEE Transactions on Vehicular Technology,* vol. 62, pp. 4064-4069, 2013.

[131] J. Wu, C. Yuen, N. M. Cheung, J. Chen, and C. W. Chen, "Enabling Adaptive High-Frame-Rate Video Streaming in Mobile Cloud Gaming Applications," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 25, pp. 1988-2001, 2015.

[132] C. Company. (2015, Feburary). *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014–2019*.

[133] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H.263 video traces for network performance evaluation," *IEEE Network,* vol. 15, pp. 40-54, 2001.

[134] R. AlTurki and R. Mehmood, "Multimedia Ad Hoc Networks: Performance Analysis," in *Computer Modeling and Simulation, 2008. EMS '08. Second UKSIM European Symposium on*, 2008, pp. 561-566.

[135] T. Hayajneh and G. Al-Mashaqbeh, "Multimedia traffic over WLANs: QoS support and performance evaluation," in *Information and Communication Systems (ICICS), 2014 5th International Conference on*, 2014, pp. 1-6.

[136] S. Perez, H. Facchini, A. Dantiacq, G. Cangemi, and J. Campos, "An evaluation of QoS for intensive video traffic over 802.11e WLANs," in *Electronics, Communications and Computers (CONIELECOMP), 2015 International Conference on*, 2015, pp. 8-15.

[137] D. Li and J. Pan, "Performance evaluation of video streaming over multi-hop wireless local area networks," *IEEE Transactions on Wireless Communications,* vol. 9, pp. 338-347, 2010.

[138] E. Setton, Y. Taesang, Z. Xiaoqing, A. Goldsmith, and B. Girod, "Cross-layer design of ad hoc networks for real-time video streaming," *IEEE Wireless Communications,* vol. 12, pp. 59-65, 2005.

[139] H. p. Shiang and M. V. D. Schaar, "Multi-user video streaming over multi-hop wireless networks: A distributed, cross-layer approach based on priority

queuing," *IEEE Journal on Selected Areas in Communications,* vol. 25, pp. 770-785, 2007.

[140] Y. Wu, G. Min, K. Li, and A. Y. Al-Dubai, "A Performance Model for Integrated Wireless Mesh Networks and WLANs with Heterogeneous Stations," in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, 2009, pp. 1-6.

[141] Y. Wu, G. Min, and L. T. Yang, "Performance Analysis of Hybrid Wireless Networks Under Bursty and Correlated Traffic," *IEEE Transactions on Vehicular Technology,* vol. 62, pp. 449-454, 2013.

[142] A. Abdrabou and W. Zhuang, "Stochastic delay guarantees and statistical call admission control for IEEE 802.11 single-hop ad hoc networks," *IEEE Transactions on Wireless Communications,* vol. 7, pp. 3972-3981, 2008.

[143] Y. He, R. Yuan, and W. Gong, "Modeling Power Saving Protocols for Multicast Services in 802.11 Wireless LANs," *IEEE Transactions on Mobile Computing,* vol. 9, pp. 657-671, 2010.

[144] F. N. Gouweleeuw, "The loss probability in finite-buffer queues with batch arrivals and complete rejection," *Probability in the Engineering and Informational Sciences,* vol. 8, pp. 221-227, 1994.

[145] M. Bratiychuk and A. Chydzinski, "On the loss process in a batch arrival queue," *Applied Mathematical Modelling,* vol. 33, pp. 3565-3577, 2009.

[146] A. Chydzinski, R. Wojcicki, and G. Hryn, "On the Number of Losses in an MMPP Queue," in *International Conference on Next Generation Wired/Wireless Networking*, 2007, pp. 38-48.

[147] R. Nagarajan, J. F. Kurose, and D. Towsley, "Approximation techniques for computing packet loss in finite-buffered voice multiplexers," *IEEE Journal on Selected Areas in Communications,* vol. 9, pp. 368-377, 1991.

[148] A. Baiocchi and N. Blefari-Melazzi, "Steady-state analysis of the MMPP/G/1/K queue," *IEEE Transactions on Communications,* vol. 41, pp. 531-534, 1993.

[149] B. Emmert, A. Binzenhöfer, D. Schlosser, and M. Weiß, "Source traffic characterization for thin client based office applications," in *Meeting of the European Network of Universities and Companies in Information and Communication Engineering*, 2007, pp. 86-94.

[150] A. Baiocchi, N. B. Melazzi, M. Listanti, A. Roveri, and R. Winkler, "Loss performance analysis of an ATM multiplexer loaded with high-speed on-off sources," *IEEE Journal on Selected Areas in Communications,* vol. 9, pp. 388-393, 1991.

[151] A. Chydzinski and B. Adamczyk, "Transient and stationary losses in a finite-buffer queue with batch arrivals," *Mathematical Problems in Engineering,* vol. 2012, 2012.

[152] S. Chakravarthy, "The batch Markovian arrival process: A review and future work," *Advances in probability theory and stochastic processes,* vol. 1, pp. 21-49, 2001.

[153] M. F. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach*: Courier Corporation, 1981.

[154] M. Neuts, "Structured stochastic matrices of M/G/1 type and their applications," *Marcel Decker Inc., New York,* 1989.

[155] V. Ramaswami, "From the matrix-geometric to the matrix-exponential," *Queueing Systems,* vol. 6, pp. 229-260, 1990.

[156] S. Söhnlein, "Traffic-based decomposition of multi-class networks," MSc, Department of Computer Science, University Erlangen, 2005.

[157] M. F. Neuts, *Probability distributions of phase type*: Purdue University. Department of Statistics, 1974.

[158] P. E. Wirth, "The role of teletraffic modeling in the new communications paradigms," *IEEE Communications Magazine,* vol. 35, pp. 86-92, 1997.

[159] *http://trace.eas.asu.edu/videotraces2/4k/tos_1920_265_B0/20/*.

[160] W. Haitao, P. Yong, L. Keping, C. Shiduan, and M. Jian, "Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement," in *INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, 2002, pp. 599-607 vol.2.

[161] K. S. Meier-Hellstern, "The analysis of a queue arising in overflow models," *IEEE Transactions on Communications,* vol. 37, pp. 367-372, 1989.

[162] A. Rice and S. Hay, "Measuring mobile phone energy consumption for 802.11 wireless networking," *Pervasive and Mobile Computing,* vol. 6, pp. 593-606, 2010.

[163] G. Y. Li, Z. Xu, C. Xiong, C. Yang, S. Zhang, Y. Chen*, et al.*, "Energy-efficient wireless communications: tutorial, survey, and open issues," *IEEE Wireless Communications,* vol. 18, pp. 28-35, 2011.

[164] S.-L. Tsao and C.-H. Huang, "A survey of energy efficient MAC protocols for IEEE 802.11 WLAN," *Computer Communications,* vol. 34, pp. 54-67, 2011.

[165] J.-P. Ebert, S. Aier, G. Kofahl, A. Becker, B. Burns, and A. Wolisz, "Measurement and Simulation of the Energy Consumption of a WLAN Interface," 2002.

[166] D. Fiems, V. Inghelbrecht, B. Steyaert, and H. Bruneel, "Markovian characterisation of H. 264/SVC scalable video," in *International Conference on Analytical and Stochastic Modeling Techniques and Applications*, 2008, pp. 1-15.

[167] J. Benita and R. Jayaparvathy, "Comparative performance analysis of subcarrier assignment for real-time video traffic," *IET Networks,* vol. 4, pp. 304-313, 2015.

[168] N. Changuel, B. Sayadi, and M. Kieffer, "Control of distributed servers for quality-fair delivery of multiple video streams," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 269-278.

[169] D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, "Methods for performance evaluation of VBR video traffic models," *IEEE/ACM Transactions on Networking,* vol. 2, pp. 176-180, 1994.

[170] C. Baugh, "4IPP traffic model for IEEE 802.16. 3," IEEE802.16.3c-00/51, 2000.

[171] C. Baugh, J. Huang, R. Schwartz, and D. Trinkwon, "Traffic model for 802.16 tg3 mac/phy simulations," IEEE 802.16 Broadband Wireless Access Working Group2001.

[172] I. Reljin, A. Samčović, and B. Reljin, "H. 264/AVC video compressed traces: multifractal and fractal analysis," *EURASIP Journal on Advances in Signal Processing,* vol. 2006, pp. 1-13, 2006.

[173] M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," in *ACM SIGCOMM computer communication review*, 1994, pp. 269-280.

[174] G. He and J. C. Hou, "On sampling self-similar Internet traffic," *Computer Networks,* vol. 50, pp. 2919-2936, 2006.

[175] R. Riedi and J. L. Véhel, "Multifractal properties of TCP traffic: a numerical study," INRIA, 1997.

[176] I. L. M. S. Committee, "IEEE Standard for local and metropolitan area networks Part 16: Air interface for fixed and mobile broadband wireless access systems amendment 2: Physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1," *IEEE Std 802.16-2004/Cor 1-2005,* 2006.

[177] I. Papapanagiotou, J. S. Vardakas, G. S. Paschos, M. D. Logothetis, and S. A. Kotsopoulos, "Performance evaluation of IEEE 802.11 e based on ON-

OFF traffic model," in *Proceedings of the 3rd international conference on Mobile multimedia communications*, 2007, p. 17.

[178] L. Muscariello, M. Mellia, M. Meo, M. A. Marsan, and R. L. Cigno, "Markov models of internet traffic and a new hierarchical MMPP model," *Computer Communications,* vol. 28, pp. 1835-1851, 2005.

[179] H. Zhou, B. Li, Q. Qu, and Z. Yan, "An analytical model for Quality of Experience of HTTP video streaming over wireless ad hoc networks," in *Signal Processing, Communication and Computing (ICSPCC), 2013 IEEE International Conference on*, 2013, pp. 1-5.

[180] C. H. Foh, M. Zukerman, and J. W. Tantra, "A Markovian Framework for Performance Evaluation of IEEE 802.11," *IEEE Transactions on Wireless Communications,* vol. 6, pp. 1276-1265, 2007.

[181] Q. Zhao, D. H. Tsang, and T. Sakurai, "A simple and approximate model for nonsaturated IEEE 802.11 DCF," *IEEE Transactions on Mobile Computing,* vol. 8, pp. 1539-1553, 2009.

[182] H. M. K. Alazemi, A. Margolis, J. Choi, R. Vijaykumar, and S. Roy, "Stochastic modelling and analysis of 802.11 DCF with heterogeneous non-saturated nodes," *Computer Communications,* vol. 30, pp. 3652-3661, 12/10/ 2007.

[183] K. Kosek-Szott, "A comprehensive analysis of IEEE 802.11 DCF heterogeneous traffic sources," *Ad Hoc Networks,* vol. 16, pp. 165-181, 2014.

[184] M. F. Tuysuz, M. Ucan, and D. Ayneli, "Energy-efficient medium access control over IEEE 802.11 wireless heterogeneous networks," in *2015 IEEE/CIC International Conference on Communications in China (ICCC)*, 2015, pp. 1-6.

[185] Y.-B. Chen, G.-Y. Lin, and H.-Y. Wei, "A Dynamic Estimation of the Unsaturated Buffer in the IEEE 802.11 DCF Network: a Particle Filter Framework Approach."

[186] X. Ling, L. X. Cai, J. W. Mark, and X. Shen, "Performance analysis of IEEE 802.11 DCF with heterogeneous traffic," in *2007 4th IEEE Consumer Communications and Networking Conference*, 2007, pp. 49-53.

[187] Y. Gao, X. Sun, and L. Dai, "Throughput Optimization of Heterogeneous IEEE 802.11 DCF Networks," *IEEE Transactions on Wireless Communications,* vol. 12, pp. 398-411, 2013.

[188] D. Moltchanov, "On-line state detection in time-varying traffic patterns," in *International Conference on Next Generation Wired/Wireless Networking*, 2007, pp. 49-60.

[189] L. X. Cai, X. S. Shen, J. W. Mark, L. Cai, and Y. Xiao, "Voice Capacity Analysis of WLAN With Unbalanced Traffic," *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY,* vol. 55, 2006.

[190] D. Gao, J. Cai, C. H. Foh, C.-T. Lau, and K. N. Ngan, "Improving WLAN VoIP capacity through service differentiation," *IEEE Transactions on Vehicular Technology,* vol. 57, pp. 465-474, 2008.

[191] H. Sun, M. Sheng, X. Wang, Y. Zhang, J. Liu, and K. Wang, "Resource allocation for maximizing the device-to-device communications underlaying LTE-Advanced networks," in *2013 IEEE/CIC International Conference on Communications in China - Workshops (CIC/ICCC)*, 2013, pp. 60-64.

[192] W. Shen, B. Yin, X. Cao, L. X. Cai, and Y. Cheng, "Secure device-to-device communications over WiFi direct," *IEEE Network,* vol. 30, pp. 4-9, 2016.

[193] J. Pedersen, A. G. i. Amat, I. Andriyanova, and F. Brännström, "Distributed Storage in Mobile Wireless Networks With Device-to-Device Communication," *IEEE Transactions on Communications,* vol. 64, pp. 4862-4878, 2016.

[194] Q. Chen, G. Yu, H. Shan, A. Maaref, G. Y. Li, and A. Huang, "Cellular Meets WiFi: Traffic Offloading or Resource Sharing?," *IEEE Transactions on Wireless Communications,* vol. 15, pp. 3354-3367, 2016.

[195] B. H. Jung, N. O. Song, and D. K. Sung, "A Network-Assisted User-Centric WiFi-Offloading Model for Maximizing Per-User Throughput in a Heterogeneous Network," *IEEE Transactions on Vehicular Technology,* vol. 63, pp. 1940-1945, 2014.

[196] Y. He, M. Chen, B. Ge, and M. Guizani, "On WiFi Offloading in Heterogeneous Networks: Various Incentives and Trade-Off Strategies," *IEEE Communications Surveys & Tutorials,* vol. 18, pp. 2345-2385, 2016.

[197] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in Heterogeneous Networks: Modeling, Analysis, and Design Insights," *IEEE Transactions on Wireless Communications,* vol. 12, pp. 2484-2497, 2013.

[198] M. Haiyan, Y. Jinyao, P. Georgopoulos, and B. Plattner, "Towards SDN based queuing delay estimation," *China Communications,* vol. 13, pp. 27-36, 2016.

[199] I. T. Haque and N. Abu-Ghazaleh, "Wireless Software Defined Networking: A Survey and Taxonomy," *IEEE Communications Surveys & Tutorials,* vol. 18, pp. 2713-2737, 2016.

[200] K. Wang, H. Li, F. R. Yu, and W. Wei, "Virtual Resource Allocation in Software-Defined Information-Centric Cellular Networks With Device-to-Device Communications and Imperfect CSI," *IEEE Transactions on Vehicular Technology,* vol. 65, pp. 10011-10021, 2016.