On the Trade-Off between Quality of Experience and Energy Efficiency in a Heterogeneous Cellular Network

by

Abbas Farrokhi

Submitted to the Graduate School of Engineering
and Natural Sciences in partial fulfillment of the
requirements for the degree of Master of Science

Sabancı University

December, 2015

On the Trade-Off between Quality of Experience and Energy Efficiency in a Heterogeneous Cellular Network

by Abbas Farrokhi

APPROVED BY:

Prof. Dr. Özgür Erçetin
(Thesis Supervisor) .....

Prof. Dr. Özgür Barış Akan .....

Assoc. Prof. Dr. Özgür Gürbüz .....

DATE OF APPROVAL: 30/12/2015

To my beloved family...

On the Trade-Off between Quality of Experience and Energy Efficiency in a Heterogeneous Cellular Network

Abbas Farrokhi

Electronics Engineering, Master's Thesis, 2015

Thesis Supervisor: Prof. Dr. Ozgur Ercetin

1 Abstract

Keywords: ***Energy efficiency, heterogeneous cellular network, sleep/awake strategy, buffer starvation, start-up delay, quality of experience***

Two important issues in the current mobile cellular networks are: Firstly, the traffic on the internet has shifted from the file downloads to the video and audio streaming, secondly, the energy efficiency of cellular networks is a major concern. Particularly, the ever-increasing number of users with the exponential growth of high-data-rate traffic demand creates new challenges for wireless access providers. On the one hand, service providers want to satisfy the growing mobile data traffic demands but on the other hand, they try to reduce the operational costs and carbon emissions by decreasing the energy consumption.

In this work, we explicitly quantify the tradeoff between quality of experience (QoE) and energy efficiency in a heterogeneous cellular network. We investigate an optimal resource

on-off switching framework that minimizes the energy consumption of a heterogeneous cellular network while satisfying a desired level of quality of user experience. Considering an ON/OFF bursty arrival process, we introduce recursive equations to obtain the buffer starvation probability, as a QoE metric, of a mobile device (MD) for streaming services. The MD is in the coverage area of a femtocell base station (FBS) which is implemented at the cell edge of a macrocell base station (MBS). The buffer starvation event occurs whenever the mobile device's buffer gets empty, and after each such event, the media player of the MD restarts the service after a certain amount of packets are prefetched (start-up or initial buffering delay). Our results have the potential to reduce carbon emissions of cellular networks by reducing the energy consumption throughout the network, while guaranteeing a target starvation probability to mobile users.

Acknowledgements

Foremost, I would like to express my deepest gratitude to my supervisor, Prof. Özgür Erçetin, for the continuous support of my MSc study, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I have been amazingly fortunate to have a supervisor who gave me the freedom to explore on my own, and at the same time the guidance to recover when my steps faltered. Prof. Özgür Erçetin taught me how to question thoughts and express ideas. He has helped me through extremely difficult times to overcome many crisis situations in completion of this work. I should confirm that the research included in this work could not have been successfully performed without the support of Prof. Özgür Erçetin. The most important attitude that I learned from Prof. Özgür Erçetin is that, it does not matter how challenging the problems are, I should never give up.

In addition, I would like to express my gratitude to my MSc oral examination committee members, Dr. Özgür Gürbüz and Prof. Özgür Barış Akan, for taking their time serving on defense exam committee. Meanwhile, I am thankful for the time and effort they put into reading my thesis and greatly appreciate their valuable comments.

I would also like to thank ARGELA technologies for supporting this research during my master studies.

Last but not least, I would like to thank my family for supporting me spiritually throughout my life. They were always supporting me and encouraging me with their best wishes.

Contents

List of Figures

List of Tables

2 Introduction

2.1 Motivation

During the past decade, due to extensive popularization of smart mobile terminals, there has been an increasing demand for wireless communication services, which have extended beyond telephony services to include audio and video streaming [1]. Accordingly, the conventional homogeneous cellular networks based on the careful deployment of MBSs have faced a serious challenge to meet this overwhelming demand of network capacity, as they suffer poor signal quality for indoor and cell edge users. In order to address this issue and provide better coverage, heterogeneous cellular networks have been introduced in the LTE-Advanced standardization [2–4]. A heterogeneous network, as shown in Fig. 1, uses a mixture of macrocells and small cells such as microcells, picocells, and femtocells.

Even though implementing of small cell networks is seen to be a promising way of catering to the ever increasing traffic demands, the dense and random deployment of small cells and their uncoordinated operation raise important questions about the implication of energy efficiency in such multi-tier networks [5–8]. In fact, the huge development of information and communication technology (ICT) industry has become one of the leading sources of energy consumption and is expected to grow exponentially in the future [9]. The rapid increase in energy cost and CO2 emissions has made the network operators realize the importance of designing their networks in an energy efficient manner [10–12]. The energy consumption of the cellular networks mostly comes from the BSs [13]. It has been estimated that the energy consumption of BSs is about 60% to 80% of that of the whole cellular network [11], [13–16].

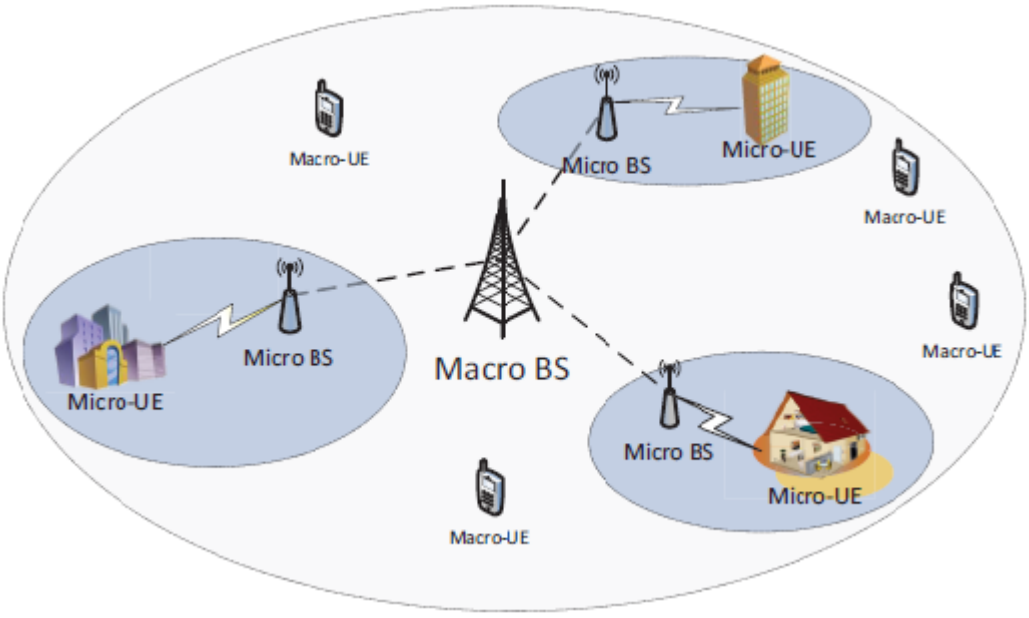

Figure 1: Typical Heterogeneous Cellular Network Architecture

There are many ways to reduce the energy consumption of BSs in Cellular networks: from topological management (e.g., the deployment of micro BSs and/or relays [17–19]) to hardware design (e.g., more energy efficient power amplifiers [20] and natural resource for cooling [21]). Moreover, since the energy consumption of a BS mainly comes from the cooling, controller, baseband signal processor and other circuits (these energy consumptions are known as the fixed power consumption of a BS), rather than the transmit power which consumes only 3.1% [22], turning BSs into sleep mode whenever possible is another promising strategy to reduce the energy consumption [13], [23–36]. Because of the high fluctuations in traffic demand over space and time in cellular networks [12], [37], some BSs could be switched off when the traffic load in their coverage area is low, and the users in sleeping cells can be served by neighboring active BSs [36]. Nevertheless, applying sleep/active strategy and turning some BSs into sleep mode may deteriorate the Quality of Service (QoS). Therefore, in order to make a tradeoff between QoS and energy efficiency of cellular networks, researchers have been investigating different active/sleep schedules while guaranteeing acceptable QoS such as delay [16], coverage performance [38], blocking probability [39], and spectral efficiency [40].

Motivated by the ever-growing demand for video streaming services in the past decade, the QoE that we consider in this work is the starvation probability of a MD buffer. The probability of buffer starvation is an important performance measure for video streaming services, as the quality of the video perceived by mobile devices is strongly related to the starvation event. The probability of starvation (also known as jitter probability) for video streaming services has been investigated in [41–45]. The event of starvation happens when the buffer gets empty, and after each such event, the media player of the MD resumes the service when there is a certain amount of packets accumulated in the buffer (prefetching). According to studies in [46–48], the user perceived video quality (or QoE equivalently) is deteriorated by two major parameters, the large start-up delay and the frequent starvation. Therefore, the media streaming service is under the influence of two factors which are the prefetching process and the starvation event. In fact, as the prefetching process gets shorter the starvation event occurs with a higher probability, and a longer prefetching process results in a larger start-up (initial buffering) delay.

2.2 Contributions

In this work, we introduce recursive equations, based on the approach in [49], to obtain the starvation probability of a buffer for an ON/OFF bursty arrivals and in a time-slotted queuing system. Unlike [41] where the authors obtain the buffer starvation probability of a user that could be served only through a single source, we evaluate the starvation probability while the MD is within the coverage area of two BSs namely MBS and FBS. Accordingly, the MD may be served by either of these BSs depending on which one is in active mode. On the other hand, analyzing the aggregated active/sleep period length distribution analytically has been an unsolved challenging problem in the literature. In [50], the authors use Monte Carlo simulations to investigate the characteristics of the OFF-period length distribution in an aggregated ON/OFF process. In our work, using a three state Markov chain and applying the first step analysis, we investigate analytically the aggregated active/sleep period length distribution for the first time in the literature.

2.3 Organization of Work

This work is organized as follows: Chapter 2 is allocated to introduction. In chapter 3, we go over literature. In 3.1, we explain what the green communication is and why we need it. Video streaming and buffer starvation probability are discussed in 3.2, 3.3, respectively. In chapter 4, we describe the system model under consideration. We introduce the BSs' sleep/awake strategy in 4.2, and then the energy consumption model is given in 4.3. In chapter 5, we present the calculation of buffer starvation probability with an ON/OFF bursty arrival. In order to derive the buffer starvation probability, we first introduce a recursive equation in 5.1. The relative recursive equation includes a term which denotes the probability of number of packets leaving the MDs buffer during an inter-arrival period of time. To obtain this term, we first derive the probability generating function of inter-arrival time in 5.2, and then in 5.3, we obtain the distribution of number of packets that leave the MD's buffer during this period. We formulate our optimization problem in 5.4. In chapter 6, we validate our analysis via simulation results. Lastly, our conclusions and future work directions are given in chapter 7.

3 Literature Review

3.1 Green Wireless Communications

During the past decade, the demand for wireless communication services has been increasing exponentially, as the smart mobile terminals are getting more and more popular. This issue has led to a wide deployment of wireless access networks to meet the overwhelming demand for network capacity, and accordingly, the power consumption of the cellular networks has significantly increased [51]. The high energy consumption of cellular networks has arisen environmental and financial concerns for both network operators and users, and QoE considerations for the end users.

Regarding the environmental concerns, the research in [52] reports that ICT contributes 2 percent of the total CO_2 emissions worldwide, and this amount is expected to increase to 4 percent by 2020. In addition, the study in [53] indicates that the high energy consumption of BSs and MDs leads to high heat dispersion and electronic pollution. From a financial perspective, energy costs constitute a significant part of the operating expenses of network operators [38, 54]. In particular, according to [55], energy costs of network operators range from 18 percent to 32 percent of their operational expenditure. From a mobile user QoE aspect, in [56], it is claimed that more than 60 percent of users complain about their mobile devices' battery life time. Moreover, the gap between the energy demand and the battery capacity offered to mobile users is significantly increasing [57]. Because of the above mentioned problems, there has been a crucial necessity for energy efficient solutions in cellular networks. The conducted research works in this direction are referred to *green* solution, where the term *green* indicates that the proposed solutions are friendly to the environment.

In the following, we go over the set of research work carried out in the direction of green communications.

Studies on the energy footprint of mobile networks indicate that energy consumption of the cellular networks mostly comes from the BSs. In particular, the research in [16] indicates that BSs contributes 60% to 80% of the total energy consumed by a cellular network. Reducing the transmitted power of BSs [58] is not sufficient enough to reduce the energy consumption of cellular networks, as a major part of the energy comes from load-independent components, such as cooling, controller, baseband signal processor, and other circuits. Different approaches aiming at minimizing the power consumption in BSs can be divided into following three types. First, increasing the number of cells in order to reduce the cell size, which brings the BSs and mobile terminals close to each other, i.e., it reduces the distance between BSs and MTs, and accordingly the BSs need on average a lower transmission power for establishing a reliable communication. Secondly, introducing femtocells and indoor distributed antenna systems which use MIMO technology. This approach has been proposed to maintain high spectral efficiency through reducing the co-channel interference resulted from frequency reuse between the femtocells. Thirdly, according to [54] BSs in a mobile network are typically underutilized. Motivated by high fluctuations of traffic load which is shown in Fig. 2, in order to reduce the energy consumption of cellular networks, some BSs could be switched off when the traffic load in their coverage area is low. However, turning some BSs into sleep mode may deteriorate the QoS or QoE required by the end users. Therefore, an important set of works on green communications has been dedicated to the tradeoff between energy efficiency and QoS/QoE.

In [16], the authors formulate a total cost minimization problem that provide a tradeoff between flow-level performance and energy consumption. Under the assumption of *time-scale separation*, the authors decouple user association and dynamic BS on/off operation and purely focus on each of these problems. In [28], the optimal design of proportion of off-mode MBSs, active MBSs' transmission power power, and small cell BSs density has been investigated to minimize BS energy consumption while satisfying MDs power con-

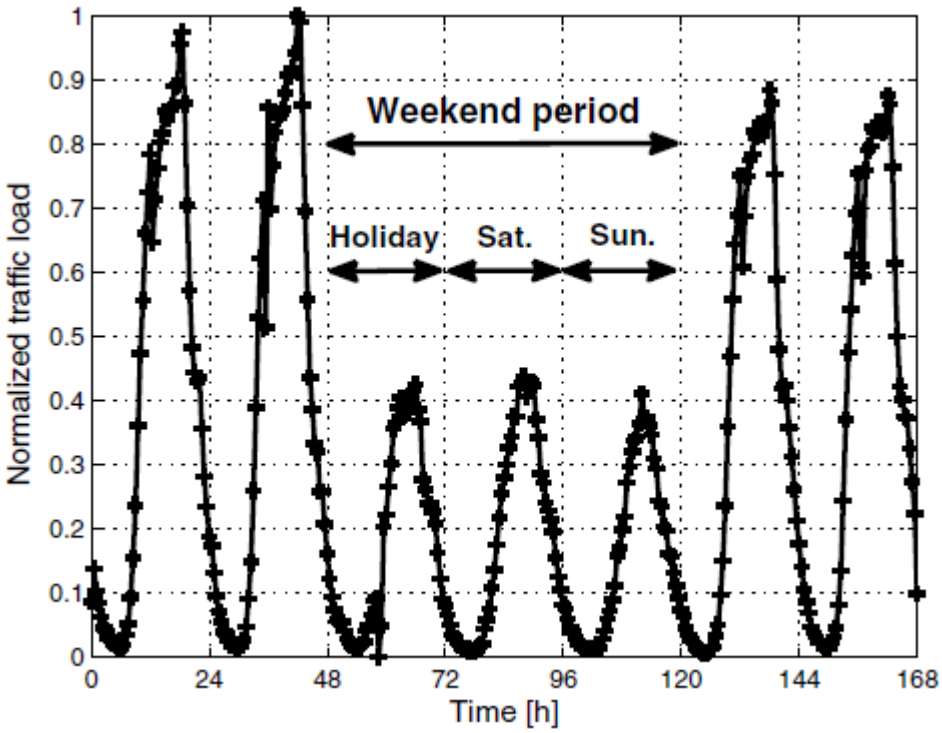

Figure 2: Normalized real traffic load (voice call information) during one week that are recorded by an anonymous cellular operator

sumption constraint. In [38], based on two different sleeping mechanisms, random sleeping and strategic sleeping, the authors model different optimization problems to minimize the power consumption of BSs subject to a coverage probability constraint for both homogeneous and heterogeneous cellular networks. In [39], a dynamic programming (DP) problem has been formulated to minimize the energy consumption while satisfying a target blocking probability as the QoS metric.

In [59], in an OFDMA cellular network, assuming that the base stations and mobile devices are located according to homogeneous Poisson point process with different rates, they study the impact of switching off base stations on the total expected power consumption, on the coverage, and on the amount of radiation to the humans body, under the following two conditions, considering interference when the traffic is heavey and neglecting radio interference for a light traffic. In [33], the authors introduce a 2-level scheme that adjusts cell sizes between two fixed values to optimize the energy consumed in a mobile network. They show that there are factors such as base station technology, data rates, and traffic demands that

significantly affect the choice of optimal cell size to minimize the energy consumption in a mobile network. In [60], the authors study energy consumption of different cellular network architectures. Specifically, they compare the transmit power consumption between a single large cell with multiple antennas, multiple small cell with a single antenna at each cell, and a large cell with a distributed antenna. They claim that macrocells with distributed antennas have the best energy efficiency compare to other two architectures under perfect channel state information (CSI).

The research work in [61] formalizes the problem of jointly optimizing the base stations transmit power and user association to minimize overall power consumption while guaranteeing a target blocking probability, and then an iterative algorithm, which takes the inter-cell interference into account, has been proposed to solve the problem. The work in [29] aims in minimizing the energy consumption of the overall heterogeneous cellular network while satisfying the Quality of Service (QoS) required by the mobile terminals. Using Markov Decision Processes, the authors study optimal sleep/awake schemes for the femtocell base stations , which are deployed within macrocell BS, for the following three different cases, the information on traffic load and user localization in the cell is complete, partial or delayed. In [62], the authors propose a theoretical framework for BS energy saving that considers dynamic BS operation and user association jointly. They formulate a total cost minimization problem to analyze the tradeoff between flow-level performance and energy consumption in a cellular network. By decomposing the general problem into two subproblems, energy-efficient user association and energy-efficient BS operation, they investigate these problems separately. For the user association problem, they propose an optimal energy-efficient user association policy and also introduce a distributed implementation. For the dynamic BS operation problem, they propose greedy-on and greedy-off algorithms.

The research in [63] investigates a component-level deceleration technique in BS operation, called speed-scaling, which can conserve dynamic power effectively during periods of time when traffic is low, while guaranteeing full coverage at all times. Their main goal is to evaluate an equilibrium which results from the interaction between speed-scaling and load bal-

ancing for green cellular networks, and then introduce a distributed iterative optimal speed control and user association policy. [64] study how small cell network deployment can meet the growth in data traffic demand, while reducing both the cost and energy consumed. The authors first obtain the closed form expressions for capacity , energy consumption and total cost, and then present a tradeoff between Capacity, Energy and Cost. In [40], using stochastic geometry and markov chain theory, the authors compute the analytical expressions of average energy consumption and spectrum efficiency. Afterward, they analyze the tradeoff between spectral efficiency and energy minimization of a heterogeneous cellular network with a sleep control in which small cell BSs are adaptively turned into active or sleep mode based on traffic variation.

The work in [65] characterize the trade-off between initial delay and the usage cost for guaranteeing a target interruption probability. In the first part, the author study the problem of efficient streaming in technology-heterogeneous settings, where the mobile device is able to receive a stream from different servers. To alleviate the duplicate packet reception problem which is a major challenge in multi-server systems, the authors propose a random linear network coding (RLNC) across packets within each block of the media file. In the second part, considering an unreliable wireless channel, and assuming that each server delivers packets according to a Poisson process with a known rate, they aim at developing an algorithm that switches between the free and the costly servers to satisfy the desired QoE, which are the probability of interruption throughout media playback and the initial buffering delay, at the minimum cost. The authors use a Markov Decision Process, with a probabilistic constraint to formulate the optimal control problem, and using the Hamilton-Jacobi-Bellman (HJB) equation, a characterization of the optimal policy has been presented.

In a heterogeneous cellular network, small cell BSs can be powered by distributed electricity generators that utilize green energy provided by renewable resources such as wind and solar. However, due to user traffic and energy supplies' dynamic behavior, managing the green energy powered small cell BSs to maximize the utilization of green energy is an challenging issue. In networks powered by renewable sources,The fundamental design issue is how to

utilize the gained energy to provision traffic demands of users in the network. Intelligent Cell brEathing (ICE) is an approach which is proposed in [66] to minimize the maximal energy depleting rates (EDRs) of Low-power base stations (LBSs), and hence maximizing the utilization of the green energy. Due to the limited energy storage, the energy consumption of LBSs might be larger than their energy storage, and thus these BSs are not able to serve all the users with green energy. Accordingly, the relative users will be served by the high power BS (HBS) which consume the main-grid energy. The authors in [66] demonstrate how ICE enables more users to be served with green energy by balancing the energy consumptions among LBSs, and therefore reduces the on-grid energy consumption.

Considering multicell cooperation technique, the authors in [67] aim at improving the energy efficiency of cellular networks through the following three approach. The first approach is traffic-intensity-aware multicell cooperation, which adapts the network layout according to users traffic demands to reduce the number of active base stations by switching off BSs with a light traffic load. The second approach is energy-aware multicell cooperation, which switches mobile users from on-grid base stations to green BSs powered by renewable sources such as wind or solar, and accordingly, reducing the on-grid energy consumption. In the third approach, the authors study coordinated multipoint transmissions for improving the energy efficiency of cellular networks. In [68], by define the network energy efficiency as the ratio of the network total throughput to the network total power consumption, the authors investigate the optimal transmission power of micro BSs through formulating an optimization problem to maximize the network energy efficiency while ensuring a target traffic coverage ratio. In [69], considering both greedy and round-robin scheduling schemes at the active BSs, for a typical sleeping cell user, the authors study two user association schemes as Maximum best-case Mean channel Access Probability (MMAP) and Maximum Received Signal Power (MRSP)-based user association. After the association is performed, the authors evaluate the spectral and energy efficiencies by obtaining the exact access probability for a mobile user in a sleeping BS and the statistics of its received signal and interference powers.

Cognitive radio and cooperative relaying are two other promising technologies that enable

green communications in mobile networks. The main goal of cognitive radio is collecting information on the spectrum usage and trying to access the unused frequency bands, in order to compensate for the spectrum under-utilization. In order to understand how using spectrum more efficiently can reduce power consumption, we refer to Shannones capacity formula [70], where we find the trade-off between the bandwidth and power. In [71], it is claimed network operator can save up to 50 percent of power through dynamically managing their spectrum or sharing of spectrum which allow channel bandwidths to be increased. On the other hand, owing to some undesirable properties of wireless communication channels, such as shadowing effects, different types of fading and large path losses, serving the users which are far from the BS requires a high transmission power from the relative BS to establish a reliable communication. This high transmission power not only generates high levels of interface at nearby users and BSs, but also leads to a high power consumption in the cellular network. Cooperative communication is able to improve the MIMO systems in terms of coverage enlarging and capacity enhancement [72]. Cooperative relaying technique divides a direct link between BS and mobile terminals into several shorter links [73]. This technique alleviates the wireless channel impairments such as path loss, and accordingly, BS and relays need a lower transmission power for establishing a reliable communication. Enabling green communication through cooperative techniques can be obtained by two different approaches. The first approach aims at provisioning service to more users with less power consumption, and is done through installing fixed relays within the coverage area of networks. And the second approach is exploiting the mobile terminals to behave as relays.

In [74], the authors introduce an energy-efficient video streaming system for mobile terminals. The video streaming system simultaneously benefits two communication channels, cellular links which carry media content from the media Cloud to mobile terminals and WiFi links enabling cooperation among mobile terminals. The latter one is due to the fact that utilization of short-range communication links not only result in higher data rate and shorter delays, two crucial performance metrics for multimedia streaming, but also according to [75], it improves the energy efficiency on mobile terminals. The research in [76] studies the mobile devices' optimal sleep policy which minimizes both the energy consumption and the system

response delay simultaneously, while sustaining the desired balance between the two. The authors consider the following three different sleep period length distribution for the mobile terminals, Exponential distribution, Hyper-exponential distribution which is for the case when there is a prior distribution on the parameter of the exponential distribution, and the parameter itself is unknown, and finally General distribution.

3.2 Video Streaming

Due to widespread development of ICT industry, ubiquity of internet access, and increasing usage of mobile devices, there is an ever-growing demand for video streaming services. For example, web video constitutes up to more than 37 percent of total traffic during peak hours in USA [77]. In 2014, YouTube and Netflix hog up to 49 percent of the fixed access Internet traffic, in North America, and YouTube solely constitutes 20 percent of the mobile internet traffic [78]. In addition, according to the predictions of Cisco Visual Networking Index, 79 percent of all internet traffic will be based on internet protocol (IP) video traffic [79] in 2018. In contrast to exponential growth of internet traffic, bandwidth provision usually falls behind. In this context, maintaining a satisfactory QoE of streaming service has been a crucial challenge for network operators. In the literature, various QoS metrics, such as statistical delay bound [80], are proposed to measure the quality of video perceived by mobile devices.

3.3 Buffer Starvation Probability

The probability of buffer starvation (also known as jitter probability), as an important performance metric, has various applications in different fields, such as video streaming services. Accordingly, there has been a great effort to investigate the starvation probability, as a QoE metric, to improve to the video quality experienced by mobile devices. In [41], to obtain the distribution of the number of starvations for a single file, two approaches has been introduced. The first approach, which is suitable for independent and identically distributed (i.i.d.) arrival process, is based on Ballot theorem. According to the Ballot theorem, for a file of size N packets of which x_1 packets is initially buffered before the service starts, the buffer starvation probability is given as follows.

$$P_s = \sum_{k=x_1}^{N-1} \frac{x_1}{2k-x_1} \binom{2k-x_1}{k-x_1} p_a^{k-x_1} (1-p_a)^k, \tag{1}$$

$$p_a = \frac{\lambda}{\lambda+\mu}, \tag{2}$$

where p_a denotes the probability of packet arrival in a system in which the packet arrival and the packet departure are modeled according to Poisson processes with rates λ and μ, respectively. Their second approach is based on a recursive equation which can be used to obtain the starvation probabilities for more complicated arrival process. In [42], the authors aim to to investigate the amount of initial buffering needed for meeting a target interruption probability during a video file streaming. Modeling the receivers buffer as a queue with Poisson arrivals and deterministic departures, i.e. M/D/1, they provide upper and lower bounds on the minimum initial buffering required to keep the playback interruption probability below a target level. Additionally, they show that for arrival rates slightly larger than the playback rate, the minimum initial delay for a given jitter probability remains bounded as the file size grows, and the interruption probability is given by:

$$P_{jitter} = e^{-I(R)T_1}, \tag{3}$$

where *I(R)* denotes the reliability function and is the largest root of $\gamma(r) = r + R(e^{-r} - 1)$, T_1 denotes the initial buffering delay. In [43], the video streaming problem has been investigated over both constant and variable bit-rate (VBR) channels. To prevent the interruption event, they specify the minimum required buffer for a given video stream and a deterministic VBR channel. In [44], the authors aim to optimize the streaming of VBR encoded video over a random VBR channel through, by investigating the fundamental tradeoffs between the initial buffering delay, the end user buffer size, and the jitter-free playback probability. In [45], considering the receiver buffer as a $G/G/1/\infty$ and $G/G/1/N$ queue,respectively, with arbitrary packet arrival and packet departure processes, they compute the interruption probability using the diffusion approximation, and try to optimize the user perceived video quality in terms of the initial buffering delay, seamless video playback and packet loss rate. In [81], the authors study the effect of feedback based rate controls on video streaming service. Modeling the receiver buffer as a finite-capacity single-server queue with a state-dependent arrival process, the buffer-overflow and buffer starvation probabilities are analyzed using a discrete-time Markov chain.

4 Problem Definition

4.1 System Model

We consider a heterogeneous cellular network consisting of two base stations where a FBS is implemented within the coverage area of a MBS. Our main goal is to optimize the energy consumption of this heterogeneous cellular network while satisfying user QoE, which in this work is guaranteeing a target buffer starvation probability for streaming services. To this end, we consider a single media file with finite size N. The media content is pre-stored in the media server (e.g., video on demand (VoD) service). After a request by the MD, the server (either MBS or FBS) segments the file into packets and transfers them to the MD. In order to correctly model the packet arrivals to the media player of the MD, we should consider several important points. Firstly, we consider an ON/OFF bursty traffic model where the sources (BSs) may stay for relatively long durations in ON and OFF modes, and the packet arrival occurs only when a BS is in ON mode. Secondly, we divide the time into small slots with duration h, and denote by ρ_m, ρ_f the probability of packet arrival from MBS and FBS to the media player of MD in a time slot, respectively. Assuming that in a continuous-time scenario the packet arrival from MBS and FBS is modeled according to poisson processes with rates λ_m and λ_f, respectively, ρ_m and ρ_f can be defined as

$$\rho_m = (\lambda_m h) e^{-\lambda_m h}, \tag{4}$$

$$\rho_f = (\lambda_f h) e^{-\lambda_f h}. \tag{5}$$

Thirdly, we denote by ψ_m the probability that MBS is active, given that the *system* is in ON mode, and by ψ_f the probability that FBS is active, given that the *system* is in ON mode. Note that the *system* is active whenever either MBS or FBS is in ON mode. To obtain the probabilities ψ_m, ψ_f, we use the Markov chain of our system model as shown in Fig. 1. The state space of this Markov chain is $\{(s_m(k), s_f(k)) \ : s_i(k) \in (\text{ON}, \text{OFF}) \ ; \ i = m,f\}$, where $s_m(k)$, $s_f(k)$ denote the state of MBS and FBS at time k, and t_{ij} denotes the transition probability from state i to state j. In Fig. 1, the states 0, 1, 2 denote the state space (OFF,OFF), (OFF,ON), and (ON,OFF), respectively. In this model, since we consider the MD that is under the coverage area of both FBS and MBS, only one base station is needed to be in active mode for serving this MD. Moreover, keeping in mind that we more prefer the FBS rather than MBS to server the end user, there will be one transition from the state (ON,OFF), which denotes the MBS to be active and FBS to be in sleep mode, to state (ON,ON) and there would not be any transition back to (ON,OFF), instead there would be one transition from state (ON,ON) to (OFF,ON), which denotes the MBS to be in sleep and FBS be in active mode, and again there would not be any transition back to state (ON,ON). Accordingly, omitting the state (ON,ON) in our system model, will not effect the general problem of minimizing the cellular network energy consumption. However, the consequence of this assumption is that the MD cannot be served while both MBS and FBS are in active mode.

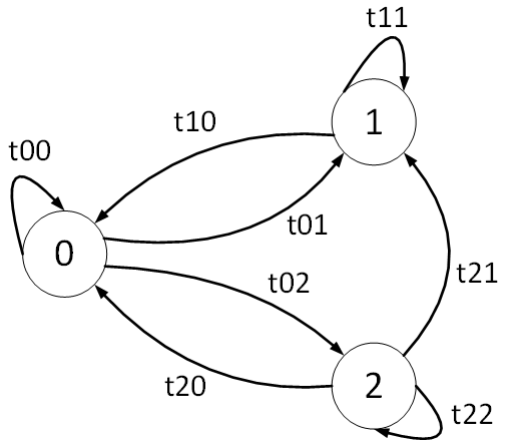

Figure 3: Markov Chain

We denote by π_1, π_2 the steady state probabilities of states 1 and 2, i.e., the proportion of time that each FBS and MBS is in active mode at the steady state. Therefore, we obtain ψ_m and ψ_f as follows.

$$\psi_m = \frac{\pi_2}{\pi_2 + \pi_1}, \tag{6}$$

$$\psi_f = \frac{\pi_1}{\pi_2 + \pi_1}. \tag{7}$$

Using equations (4)-(7), we obtain the probability of packet arrival to the media player of a MD during a time slot h as follows.

$$\varsigma = \psi_m \rho_m + \psi_f \rho_f. \tag{8}$$

The probability ς denotes the probability of packet arrival from either MBS or FBS to the MD's buffer during a time-slot h. We model the arrival process as a bernoulli process with this success probability ς. In addition, we assume that at the buffer of a MD packet departure follows an exponential distribution with rate μ. Using this assumption we obtain the probability of packet departure, denoted by ω, in a time-slot h as follows.

$$\omega = 1 - e^{-\mu h}. \tag{9}$$

Considering ω as the probability that a packet completes its service during a small time slot h, we model the service process at the media player of the MD as a bernoulli process with probability ω.

4.2 Base Stations Active/Sleep Schedules

The active/sleep period durations of BSs are modeled as four independent and identically distributed (i.i.d.) random variables. More specifically, we model the active period durations of MBS and FBS according to an exponential distribution with rates α_m and α_f, respectively. The sleep period durations of the BSs are modeled as exponential distributions with rates β_m, β_f for macrocell and femtocell, respectively. Recall that we are considering those users that are under the coverage area of femtocell base station, so the users could be served by FBS or MBS depending on which one is active. However, we assume that the arrival rate from a FBS is more than that a MBS provides to the MDs.

4.3 Energy Consumption Model

The expected energy consumption of this cellular network is given by:

$$E_{total} = E_f + \gamma E_m, \tag{10}$$

where E_m, E_f denote the expected energy consumptions of MBS and FBS, respectively. In this model, we assume that the energy consumption of the MBS is γ times more than what a FBS consumes per unit time. Note that E_f and E_m are proportional to the average amount of time that each FBS and MBS spends in active mode. Accordingly, to obtain E_f and E_m, we compute the average amount of time the system spends in states 1 and 2 of the Markov chain shown in Fig. 1, respectively, in steady state. The energy consumed by a FBS per unit time is considered as 5 Microjoule (5μJ). In addition, based on [12], the fixed power (load-independent) consumption of MBS and FBS are set to 118.7 and 4.8 W. Accordingly, the real value of γ can be approximately considered as 25.

Table 1: List of Parameters

h	Time is divided into small slots and h denotes the duration of each time-slot
N	File Size, i.e., number of packets of the considered file.
t_{ij}	The transition probability from state i to state j
ψ_m, ψ_f	The probability that MBS/FBS is active, given that the *system* is in ON mode
λ_m, λ_f	Poisson Packet arrival rate from MBS/FBS
π_1, π_2	The proportion of time that FBS/MBS is in active mode at the steady state
ρ_m, ρ_f	Probability of packet arrival from MBS and FBS to the media player of MD in a time slot
ς	Probability of packet arrival from either MBS or FBS to the MD's buffer during a time-slot
ω	Probability of packet departure during a time-slot
α_m, α_f	MBS/FBS active period exponential durations rates
β_m, β_f	MBS/FBS sleep period exponential durations rates
E_m, E_f	The expected energy consumptions of MBS/FBS
p	Probability of packet arrival during an active period
τ	Packet inter-arrival time to MD's buffer
ν	Number of packets that leave the MD's buffer during the inter-arrival time τ
$P_i(m)$	Probability of starvation for a file of n packets, given that there are i packets in the buffer of the MD upon arrival of the first packet of this file

5 Buffer Starvation Analysis

5.1 Probability of Starvation for an ON/OFF Bursty Arrival

In this section, we define a recursive approach to obtain the buffer starvation probability of a MD that is in the coverage area of the FBS, and FBS is implemented within the coverage of the MBS, i.e. the mobile device could be served by both BSs. Note that the buffer starvation analysis developed in this work can be used both for video and audio streaming services in a video on demand basis, where the finite media file is pre-stored in the media server. We denote by $P_i(n)$ the probability of starvation for a file of n packets, given that there are i packets in the buffer of the MD upon arrival of the first packet of this file. In our system, we aim to obtain the starvation probability for streaming a file of size N packets (with a typical packet size of 1460 bytes) while x packets of this file are prefetched before the service begins. Therefore, the starvation probability in our system model corresponds to $P_i(n)$ with $i = x-1$ and $n = N-x+1$. To compute $P_i(n)$, we introduce recursive equations. To this end, we define a quantity $Q_i^{ON}(k)$, $0 \leq i \leq N-1$, $0 \leq k \leq i$, which is the probability that k packets out of i leave the MD's buffer upon an arrival at the ON state, i.e., there is no packet arrival when the system is in OFF mode (both BSs are switched off). To apply the recursive equations, we start from the case $n = 1$.

$$P_i(1) = 0, \ \ \forall i \geq 1. \tag{11}$$

When the file size is 1 and the only packet observes a non-empty queue, the probability of starvation is zero. If i is zero, i.e. upon arrival we find the buffer empty, the starvation occurs

for sure, thus yielding

$$P_0(n) = 1, \;\; n = 1, ..., N. \tag{12}$$

For $n \geq 2$, we have the following recursive equation:

$$P_i(n) = \sum_{k=0}^{i+1} Q_{i+1}^{ON}(k) P_{i+1-k}(n-1), \;\; 0 \leq i \leq N-1. \tag{13}$$

According to (13), when the first packet of the file arrives and finds i packets in the system, the starvation does not happen. However, the starvation might happen in the service of remaining $n-1$ packets. Upon the arrival of the next packet, k packets out of $i+1$ leave the system with probability $Q_{i+1}^{ON}(k)$. Since the total number of packets is N, the starvation probability must satisfy $P_i(n) = 0$ for $i+n > N$. In order to obtain $P_i(n)$ using (13), we should first obtain the term $Q_i^{ON}(k)$.

5.1.1 Number of Packets Leaving the Mobile Device's Buffer During an Inter-Arrival Period

As it is mentioned earlier, the term $Q_i^{ON}(k)$ denotes the probability that k packets out of i leave the buffer of the MD during an inter-arrival period. First, we denote the random variable (r.v.) of inter-arrival period by τ, and let $T(z) = \mathsf{E}[z^{\tau}]$ be its probability generating function. Secondly, we denote by ν the r.v. of the number of packets that leave the MD's buffer during an inter-arrival period, and let $N(z) = \mathsf{E}[z^{\nu}]$ be its probability generating function. Using the probability generating function $T(z)$, we obtain the probability generating function of the number of bernoulli departures, with a success probability as defined in (9), during the inter-arrival period τ, i.e. we obtain $N(z)$ from $T(z)$. Finally, by evaluating the inverse transform of $N(z)$, we obtain the probability mass function (pmf) of r.v. ν, from which we obtain the term $Q_i^{ON}(k)$.

Considering that an arbitrary packet has been generated by the system, we denote the time

period from the instant at which this packet is generated until the point when the system goes to sleep mode, i.e., both MBS and FBS goes to sleep mode, by active period number 1, and the following sleep period by sleep period number 1. Then, we number the subsequent active (sleep) periods by the numbers $2,3,\ldots$. We define the event ϕ_m as the event in which the next packet arrives during active period number m, (m=$1,2,\ldots$). The probability of ϕ_m is given as follows.

$$\mathsf{Pr}(\phi_m) = q^{m-1}p, \;\; m \geq 1, \tag{14}$$

where p denotes the probability of packet arrival in an active period, and $q = 1-p$. In other words, the probability p denotes the event in which the time duration from the beginning of an active period until the next packet arrival is less than or equal to the duration of that active period. We let r.v. R denote the time duration from the beginning of an active period until the next packet arrival in that active period, and r.v. Y denote the time duration of an active period. In order to obtain p, we shoud first derive the distributions of random variables Y and R.

5.1.2 System's Aggregated Active Period Length Distribution

According to our system model which is shown in Fig. 1, and using the first step analysis we obtain the probability mass function (pmf) of r.v. Y as follows.

$$T_1(1) = t_{10}, \tag{15}$$

$$T_1(k) = t_{11}T_1(k-1), \tag{16}$$

$$T_2(1) = t_{20}, \tag{17}$$

$$T_2(k) = t_{22}T_2(k-1) + t_{21}T_1(k-1), \tag{18}$$

where $T_i(.)$ denotes the number of steps that it takes to get to state zero, given that we are initially at state $i(i=1,2)$, and t_{ij} denotes the transition probability from state i to state j. To obtain $T_1(k)$ in a closed formula, we rewrite (16) as follows:

$$\begin{aligned} T_1(k) &= t_{11}T_1(k-1) = t_{11}^2T_1(k-2) = ... \\ &= t_{11}^{k-1}T_1(1) = t_{11}^{k-1}t_{10}, \quad k = 1,2,3,.... \end{aligned} \tag{19}$$

To obtain $T_2(k)$ in a closed formula, we rewrite (18) as follows.

$$\begin{aligned} T_2(k) &= t_{22}T_2(k-1) + t_{21}T_1(k-1) \\ &= t_{22}[t_{22}T_2(k-2) + t_{21}T_1(k-2)] + t_{21}T_1(k-1) \\ &= t_{22}^2T_2(k-2) + t_{22}t_{21}T_1(k-2) + t_{21}T_1(k-1) = ... \\ &= t_{22}^{k-1}T_2(1) + t_{22}^{k-2}t_{21}T_1(1) + t_{22}^{k-3}t_{21}T_1(2) + ... + t_{21}T_1(k-1). \end{aligned} \tag{20}$$

Inserting (17) and (19) in (20) results in:

$$\begin{aligned} T_2(k) &= t_{22}^{k-1}t_{20} + t_{22}^{k-2}t_{21}t_{10} + t_{22}^{k-3}t_{21}t_{11}t_{10} + ... + t_{21}t_{11}^{k-2}t_{10} \\ &= t_{22}^{k-1}t_{20} + t_{21}t_{10}[t_{22}^{k-2} + t_{22}^{k-3}t_{11} + t_{22}^{k-4}t_{11}^2 + ... + t_{11}^{k-2}] \\ &= t_{22}^{k-1}t_{20} + t_{21}t_{10}t_{22}^{k-2}\sum_{r=2}^{k}\left(\frac{t_{11}}{t_{22}}\right)^{r-2} \\ &= t_{22}^{k-1}t_{20} + t_{21}t_{10}\left(\frac{t_{22}^{k-1} - t_{11}^{k-1}}{t_{22} - t_{11}}\right), \quad k = 1,2,3,.... \end{aligned} \tag{21}$$

Using (19) and (21), we obtain the aggregated active period length distribution as follows.

$$F_Y(y) = t_{11}^{y-1} t_{10} \psi_f + \left(t_{22}^{y-1} t_{20} + \frac{t_{21} t_{10} (t_{22}^{y-1} - t_{11}^{y-1})}{t_{22} - t_{11}} \right) \psi_m, \tag{22}$$

where t_{ij} denotes the transition probability from state i to state j.

5.1.3 Distribution of Inter-Arrival time *R*

Considering that the packet arrival to the MD's buffer is modeled as a Bernoulli process with a success probability defined in (8), we obtain the pmf of r.v. R as follows.

$$F_R(r) = \varsigma(1-\varsigma)^{r-1}, \;\; r = 1,2,3,.... \tag{23}$$

5.1.4 Packet Arrival Probability During an Active Period

By the use of $F_Y(y)$ and $F_R(r)$ we obtain the probability of packet arrival during an active period as follows.

$$p = \mathsf{Pr}(R \leq Y) = \sum_{y=1}^{\infty} F_Y(y) \sum_{r=1}^{y} F_R(r) = \sum_{y=1}^{\infty} F_Y(y) \sum_{r=1}^{y} \varsigma(1-\varsigma)^{r-1}$$

$$= \sum_{y=1}^{\infty} F_Y(y)\varsigma[1 + (1-\varsigma) + (1-\varsigma)^2 + ... + (1-\varsigma)^{y-1}]$$

$$= \sum_{y=1}^{\infty} F_Y(y)\varsigma\frac{1-(1-\varsigma)^y}{\varsigma} = \sum_{y=1}^{\infty} F_Y(y) - \sum_{y=1}^{\infty} F_Y(y)(1-\varsigma)^y$$

$$= \sum_{y=1}^{\infty} (t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)t_{11}^{y-1} + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m t_{22}^{y-1} \tag{24}$$

$$- \sum_{y=1}^{\infty} [(t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)t_{11}^{y-1}(1-\varsigma)^y + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m t_{22}^{y-1}(1-\varsigma)^y]$$

$$= (t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)\frac{1}{1-t_{11}} + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m\frac{1}{1-t_{22}}$$

$$- (t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)\frac{1-\varsigma}{1-(1-\varsigma)t_{11}} - (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m\frac{1-\varsigma}{1-(1-\varsigma)t_{22}}$$

$$= d_1 + d_2$$

where the values of d_1 and d_2 are given as follows.

$$d_1 = (\frac{1}{1-t_{11}} - \frac{1-\varsigma}{1-(1-\varsigma)t_{11}})(t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m), \tag{25}$$

$$d_2 = (\frac{1}{1-t_{22}} - \frac{1-\varsigma}{1-(1-\varsigma)t_{22}})(t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m. \tag{26}$$

5.2 Probability Generating Function of Inter-Arrival Period

In Fig. 2, we illustrate the inter-arrival time τ in terms of three subsections, i.e., C_k, S_k, D_k. Note that C_k denotes the time duration of active period number k given that this period ends before the arrival of the next packet. S_k denotes the time duration of sleep period number k. D_k denotes the time duration from the beginning of active period number k until the arrival of the next packet given that this packet has arrived in this active period. Note that C_k, S_k, D_k, $k \geq 1$, are i.i.d. random variables. In Fig. 2, the sleep events point to the time instances at which both BSs are in sleep mode, and thus the system is in sleep mode. The activation events, after a sleep event, point to the time instances at which either FBS or MBS wakes up, i.e. the system is in active mode.

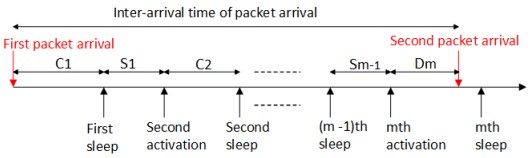

Figure 4: Illustration of random variables τ, C_k, S_k, and D_k, $k \geq 1$, given that the next packet arrival occurs in active period number m.

Accordingly, using (14) we define the probability generating function of inter-arrival time τ as follows.

$$\begin{aligned} T(z) = \mathsf{E}[z^{\tau}] &= \sum_{m=1}^{\infty} \mathsf{Pr}(\phi_m)\mathsf{E}[z^{\tau} \mid \phi_m], \\ &= \sum_{m=1}^{\infty} pq^{m-1}\mathsf{E}[z^{\sum_{k=1}^{m-1}(C_k+S_k)+D_m}], \\ &= W(z)\sum_{m=1}^{\infty} pq^{m-1}(U(z)V(z))^{m-1}, \\ &= W(z)\frac{p}{1-qU(z)V(z)}, \end{aligned} \tag{27}$$

where $U(z)$, $V(z)$, $W(z)$ denote the probability generating functions of random variables C_k,

S_k, and D_k, $k \geq 1$, respectively, are derived in the following three subsections.

5.2.1 Probability Generating Function of Random Variable *C*

In order to obtain the probability generating function $U(z)$, we first need to obtain the pmf of r.v. C_k. To this end, we first derive the probability $\mathsf{Pr}(C_k > m)$, and then we obtain the cdf of random variable C_k as $Z_{C_k}(m) = 1 - \mathsf{Pr}(C_k > m)$. Finally, from the obtained cdf, we derive the pmf of r.v. C_k as $F_{C_k}(m) = Z_{C_k}(m) - Z_{C_k}(m-1)$.

$$\begin{aligned}
\mathsf{Pr}(C_k > m) = \mathsf{Pr}(Y_k > m \mid N_k) &= \frac{\mathsf{Pr}(Y_k > m,\, N_k)}{\mathsf{Pr}(N_k)} \\
&= \frac{\mathsf{Pr}(Y_k > m,\, N_{k-1},\, Y_k < R_k)}{\mathsf{Pr}(N_{k-1},\, Y_k < R_k)} \\
&= \frac{\mathsf{Pr}(m < Y_k < R_k)\mathsf{Pr}(N_{k-1})}{\mathsf{Pr}(Y_k < R_k)\mathsf{Pr}(N_{k-1})} \\
&= \frac{\sum_{r=m+2}^{\infty} F_{R_k}(r) \sum_{y=m+1}^{r\text{-}1} F_{Y_k}(y)}{\sum_{r=2}^{\infty} F_{R_k}(r) \sum_{y=1}^{r\text{-}1} F_{Y_k}(y)}.
\end{aligned} \tag{28}$$

In (28), if we set $m = 0$, the numerator and denominator will be the same. Hence, we first obtain the numerator and then set $m = 0$ in the obtained result to get the answer for the denominator.

$$\sum_{r=m+2}^{\infty} F_{R_k}(r) \sum_{y=m+1}^{r\text{-}1} F_{Y_k}(y)$$

$$= \sum_{r=m+2}^{\infty} F_{R_k}(r) \sum_{y=m+1}^{r\text{-}1} [((t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)t_{11}^{y-1} + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m t_{22}^{y-1})]$$

$$= \sum_{r=m+2}^{\infty} F_{R_k}(r)[\frac{1}{1-t_{11}}(t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)(t_{11}^{m} - t_{11}^{r-1}) + \frac{1}{1-t_{22}}(t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m(t_{22}^{m} - t_{22}^{r-1})]$$

$$= \frac{\varsigma}{1-t_{11}}(t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m) \sum_{r=m+2}^{\infty} (t_{11}^{m} - t_{11}^{r-1}) + \frac{\varsigma}{1-t_{22}}(t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m \sum_{r=m+2}^{\infty} (t_{22}^{m} - t_{22}^{r-1})$$

$$= c_1 t_{11}^{m}(1-\varsigma)^m + c_2 t_{22}^{m}(1-\varsigma)^m, \tag{29}$$

where c_1 and c_2 are given as follows.

$$c_1 = \frac{\varsigma(1-\varsigma)}{1-t_{11}}(\frac{1}{1-(1-\varsigma)} - \frac{t_{11}}{1-(1-\varsigma)t_{11}})(t_{10}\psi_f - \frac{t_{21}t_{10}\psi_m}{t_{22}-t_{11}}), \tag{30}$$

$$c_2 = \frac{\varsigma(1-\varsigma)}{1-t_{22}}(\frac{1}{1-(1-\varsigma)} - \frac{t_{22}}{1-(1-\varsigma)t_{22}})(t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m. \tag{31}$$

By setting $m = 0$ in (29) we obtain the denominator of (28). Hence, we rewrite the (28) as follows.

$$\Pr(C_k > m) = \frac{\sum_{r=m+2}^{\infty} F_{R_k}(r) \sum_{y=m+1}^{r\text{-}1} F_{Y_k}(y)}{\sum_{r=2}^{\infty} F_{R_k}(r) \sum_{y=1}^{r\text{-}1} F_{Y_k}(y)}$$

$$= \frac{c_1 t_{11}^m (1-\varsigma)^m + c_2 t_{22}^m (1-\varsigma)^m}{c_1 + c_2} \tag{32}$$

From (32), we obtain the pmf of random variable C_k as follows.

$$F_{C_k}(m) = Z_{C_k}(m) - Z_{C_k}(m-1) = (1 - \Pr(C_k > m)) - (1 - \Pr(C_k > m\text{-}1))$$

$$= \frac{c_1}{c_1 + c_2}(1 - t_{11}(1-\varsigma)) t_{11}^{m-1}(1-\varsigma)^{m-1} + \frac{c_2}{c_1 + c_2}(1 - t_{22}(1-\varsigma)) t_{22}^{m-1}(1-\varsigma)^{m-1} \tag{33}$$

Using (33) we obtain the probability generating function $U(z)$ as follows.

$$U(z) = \mathsf{E}[z^C] = \sum_{r=1}^{\infty} z^r F_{C_k}(r)$$

$$= \frac{c_1}{c_1 + c_2}(1 - t_{11}(1-\varsigma)) \sum_{r=1}^{\infty} z^r t_{11}^{r-1}(1-\varsigma)^{r-1} + \frac{c_2}{c_1 + c_2}(1 - t_{22}(1-\varsigma)) \sum_{r=1}^{\infty} z^r t_{22}^{r-1}(1-\varsigma)^{r-1}$$

$$= \frac{c_1(1 - t_{11}(1-\varsigma))z}{(c_1 + c_2)(1 - t_{11}(1-\varsigma)z)} + \frac{c_2(1 - t_{22}(1-\varsigma))z}{(c_1 + c_2)(1 - t_{22}(1-\varsigma)z)}. \tag{34}$$

5.2.2 Probability Generating Function of Random Variable D

To obtain the probability generating function $W(z)$, we first need to obtain the pmf of r.v. D_k. To this end, we first derive the probability $\mathsf{Pr}(D_k > m)$, and then we obtain the cdf of random variable D_k as $Z_{D_k}(m) = 1 - \mathsf{Pr}(D_k > m)$. Finally, from the obtained CDF, we will end up to the pmf of r.v. D_k as $F_{D_k}(m) = Z_{D_k}(m) - Z_{D_k}(m-1)$.

$$\mathsf{Pr}(D_k > m) = \mathsf{Pr}(R_k > m \mid \phi_k) = \mathsf{Pr}(R_k > m \mid R_k \leq Y_k, N_{k-1})$$

$$= \frac{\mathsf{Pr}(R_k > m,\ R_k \leq Y_k,\ N_{k-1})}{\mathsf{Pr}(R_k \leq Y_k,\ N_{k-1})} = \frac{\mathsf{Pr}(m < R_k \leq Y_k)\mathsf{Pr}(N_{k-1})}{\mathsf{Pr}(R_k \leq Y_k)\mathsf{Pr}(N_{k-1})} \tag{35}$$

$$= \frac{\mathsf{Pr}(m < R_k \leq Y_k)}{\mathsf{Pr}(R_k \leq Y_k)} = \frac{\sum_{y=m+1}^{\infty} F_{Y_k}(y) \sum_{r=m+1}^{y} F_{R_k}(r)}{\sum_{y=1}^{\infty} F_{Y_k}(y) \sum_{r=1}^{y} F_{R_k}(r)},$$

where the term N_k denotes the event in which the next packet arrival does not happen in the first k active period. Note that the denominator in (35) is equal to $p$given in (24). Substituting (22) and (23) in (35) results in:

$$\Pr(D_k > m) = \frac{\sum_{y=m+1}^{\infty} F_{Y_k}(y) \sum_{r=m+1}^{y} \varsigma(1-\varsigma)^{r-1}}{p}$$

$$= \frac{\sum_{y=m+1}^{\infty} F_{Y_k}(y)(1-\varsigma)^m - \sum_{y=m+1}^{\infty} F_{Y_k}(y)(1-\varsigma)^y}{p}$$

$$= [\sum_{y=m+1}^{\infty} \left((t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)(1-\varsigma)^m t_{11}^{y-1} + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m (1-\varsigma)^m t_{22}^{y-1} \right)$$

$$- \sum_{y=m+1}^{\infty} \left((t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}(1-\varsigma)^y \psi_m) t_{11}^{y-1} + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m (1-\varsigma)^y t_{22}^{y-1} \right) (1-\varsigma)^y \,]/p$$

$$= [(t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)\frac{t_{11}^m (1-\varsigma)^m}{1-t_{11}} + (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m \frac{t_{22}^m (1-\varsigma)^m}{1-t_{22}}$$

$$- (t_{10}\psi_f - \frac{t_{21}t_{10}}{t_{22}-t_{11}}\psi_m)\frac{t_{11}^m (1-\varsigma)^{m+1}}{1-(1-\varsigma)t_{11}} - (t_{20} + \frac{t_{21}t_{10}}{t_{22}-t_{11}})\psi_m \frac{t_{22}^m (1-\varsigma)^{m+1}}{1-(1-\varsigma)t_{22}} \,]/p$$

$$= \frac{d_1 t_{11}^m (1-\varsigma)^m + d_2 t_{22}^m (1-\varsigma)^m}{d_1 + d_2}. \tag{36}$$

The values of d_1, d_2 are given in (25) and (26), respectively. From (36), we obtain the pmf of random variable D_k as follows:

$$F_{D_k}(m) = Z_{D_k}(m) - Z_{D_k}(m-1) = (1 - \mathsf{Pr}(D_k > m)) - (1 - \mathsf{Pr}(D_k > m\text{-}1)) = \mathsf{Pr}(D_k > m\text{-}1) - \mathsf{Pr}(D_k > m)$$

$$= \frac{d_1 t_{11}^{m-1}(1-\varsigma)^{m-1} + d_2 t_{22}^{m-1}(1-\varsigma)^{m-1}}{d_1+d_2} - \frac{d_1 t_{11}^{m}(1-\varsigma)^{m} + d_2 t_{22}^{m}(1-\varsigma)^{m}}{d_1+d_2}$$

$$= \frac{d_1(1-t_{11}(1-\varsigma))}{d_1+d_2} t_{11}^{m-1}(1-\varsigma)^{m-1} + \frac{d_2(1-t_{22}(1-\varsigma))}{d_1+d_2} t_{22}^{m-1}(1-\varsigma)^{m-1}. \tag{37}$$

Using (37) we obtain the probability generating function $W(z)$ as follows:

$$W(z) = \mathsf{E}[z^D] = \sum_{r=1}^{\infty} z^r F_{D_k}(r)$$

$$= \frac{d_1(1-t_{11}(1-\varsigma))}{d_1+d_2} \sum_{r=1}^{\infty} z^r t_{11}^{r-1}(1-\varsigma)^{r-1} + \frac{d_2(1-t_{22}(1-\varsigma))}{d_1+d_2} \sum_{r=1}^{\infty} z^r t_{22}^{r-1}(1-\varsigma)^{r-1}$$

$$= \frac{d_1(1-t_{11}(1-\varsigma))z}{(d_1+d_2)(1-t_{11}(1-\varsigma)z)} + \frac{d_2(1-t_{22}(1-\varsigma))z}{(d_1+d_2)(1-t_{22}(1-\varsigma)z)}. \tag{38}$$

5.2.3 Probability Generating Function of Random Variable S

Using the Markov chain shown in Fig. 1, we obtain the pmf of sleep period as follows.

$$F_{S_k}(m) = t_{00}^{m-1}(t_{01} + t_{02}). \tag{39}$$

From this pmf we obtain the probability generating function of r.v. S_k as follows.

$$
\begin{aligned}
V(z) = \mathsf{E}[z^S] &= \sum_{r=1}^{\infty} z^r F_{S_k}(r) \\
&= (t_{01} + t_{02}) \sum_{r=1}^{\infty} z^r t_{00}^{r-1} = \frac{(t_{01} + t_{02})z}{1 - t_{00}z}
\end{aligned}
\tag{40}
$$

Inserting the obtained generating functions, (34), (38) and (40) in (27) gives us the probability generating function of inter-arrival time τ. In the following subsection, we obtain the probability generating function of random variable ν and then, by comparing it to the probability generating function *V(z)*, we derive the pmf of random variable ν.

5.3 Distribution of Number of Packets Leave the MD's Queue During an Inter-Arrival Time

Recalling that r.v. ν denotes the number of packets that leave the buffer of the MD during an inter-arrival period τ and its probability generating function is denoted by $N(z)$. We express the generating function of r.v. ν as follows.

$$
N(z) = \mathsf{E}[z^{\nu}] = \sum_{t=1}^{\infty} \mathsf{E}[z^{\nu} | \tau = t] F_{\tau}(t) = \sum_{t=1}^{\infty} F_{\tau}(t) \sum_{k=0}^{t} z^k \mathsf{Pr}(\nu = k | \tau = t),
$$

$$
= \sum_{t=0}^{\infty} F_{\tau}(t) \sum_{k=0}^{t} z^k \binom{t}{k} \omega^k (1-\omega)^{t-k} = \sum_{t=1}^{\infty} F_{\tau}(t)(1-\omega)^t \sum_{k=0}^{t} \binom{t}{k} (\frac{\omega z}{1-\omega})^k, \tag{41}
$$

$$
= \sum_{t=0}^{\infty} F_{\tau}(t)(1-\omega)^t (1 + \frac{\omega z}{1-\omega})^t = \sum_{t=1}^{\infty} F_{\tau}(t)(1 + \omega(z-1))^t,
$$

where $F_{\tau}(t)$ denotes the pmf of inter-arrival period τ, and ω denotes the probability of service completion as defined in (9). On the other hand, the probability generating function of r.v. τ is equal to

$$T(z) = \mathsf{E}[z^{\tau}] = \sum_{t=1}^{\infty} z^t F_{\tau}(t). \tag{42}$$

By comparing the equations (41) and (42), we conclude the following expression which results in the probability generating function of r.v. ν

$$\begin{aligned} N(z) &= T(1+\omega(z-1)), \\ &= \frac{W(1+\omega(z-1))p}{1-qU(1+\omega(z-1))V(1+\omega(z-1))}. \end{aligned} \tag{43}$$

Let $F_{\nu}(t)$ denote the pmf of r.v. ν. By evaluating the inverse transform of $N(z)$, we obtain $F_{\nu}(t)$. Meanwhile, recall that $Q_i^{ON}(k)$ denotes the probability that k packets out of i leave the MD's buffer during the inter-arrival period τ. According to [49], the term $Q_i^{ON}(k)$ is obtained as follows.

$$Q_i^{ON}(k) = F_{\nu}(k), \;\; 0 \leq k \leq i-1, \tag{44}$$

$$Q_i^{ON}(i) = \sum_{n=i}^{\infty} F_{\nu}(n). \tag{45}$$

Therefore, inserting (44) and (45) in the recursive equation (14) gives us the probability of starvation for streaming a file with size N, given that there are x packets (start-up delay) accumulated in the buffer before the service begins.

5.4 Formulating an Optimization Problem

In this section, we use the results from the previous sections to investigate the energy efficiency related optimization problem subject to a QoE constraint in terms of starvation probability. We formulate an optimization problem that minimizes the energy consumption of

heterogeneous cellular network while guaranteeing a target buffer starvation probability for a MD as follows.

$$\begin{aligned} \underset{\beta_m,\beta_f}{\text{Minimize}} \quad & E_{total} = E_f + \gamma E_m \\ \text{s.t.} \quad & P_i(n) \leq \varepsilon, \\ & i = x-1, \end{aligned} \tag{46}$$

where E_m, E_f denote the expected value of the energy consumptions of MBS and FBS, respectively, and x denotes the start-up delay. Using the Markov chain shown in Fig. 1, we obtain the values of E_m, E_f as the proportion of time that MBS and FBS are in active mode in the steady state. We also assume that a femtocell's energy consumption is $1/\gamma$ of that a MBS consumes per unit time. Note that β_m, β_f denote the rates at which MBS and FBS go to active mode, respectively. Therefore, E_m, E_f increase with the increase in rates β_m, β_f. On the other hand, the starvation probability, which is given in (14), is a decreasing function of β_m and β_f. In order to solve the above problem, we first find the values of β_m and β_f that satisfy the buffer starvation constraint with equality, and then, we solve the minimization problem considering the β_m^*'s and β_f^*'s.

6 Numerical Results

In this section, we first investigate the energy minimization problem, and then compare the buffer starvation probability of a MD in a heterogeneous and a homogeneous cellular network. In the following, we set γ to 10 if not explicitly mentioned.

6.1 Energy Consumption Optimization Subject to a QoE Constraint

Fig. 5 illustrates the minimum amount of energy consumed for streaming a file with size N packets while guaranteeing a target starvation probability ε which is set to 0.15. The file size in this experiment ranges between 100 and 300 in terms of packets, and the start-up delay x is set to 50 packets. The energy consumed by a FBS is considered as 5 Microjoule (5μJ) per unit time. For the heterogeneous network, we let the rates α_m, α_f be 0.1 and 0.15, respectively. The rate β_m varies between 0.01 and 0.11, and the rate β_f varies between 0.05 and 0.15. We set λ_m, λ_f, μ, and time-slot h to 1.5, 1.7, 1, and 10^{-5}, respectively. In the case of homogeneous cellular network with a single MBS, in order to satisfy the QoE constraint (i.e., the buffer starvation probability to be less than or equal to ε), the rate β_m should vary between 0.11 and 0.21. The expected energy consumption of the network increases with the increase in the rates at which the BSs go to active mode. Nevertheless, our system model, in which the MD is covered by two BSs, significantly reduces the expected energy consumption of cellular network while guaranteeing a target starvation probability in comparison to the case where the MD is covered only by a MBS. In order to quantify the energy savings in MBS by using a FBS, in Fig. 5, we denote by green line the MBS's expected energy consumption in a heterogeneous mobile network. Accordingly, the difference between green and red lines, which is the energy that can be saved in MBS by implementing a FBS, demonstrates the

efficacy of our system in term of reducing the MBS's energy consumption for streaming a file with various sizes.

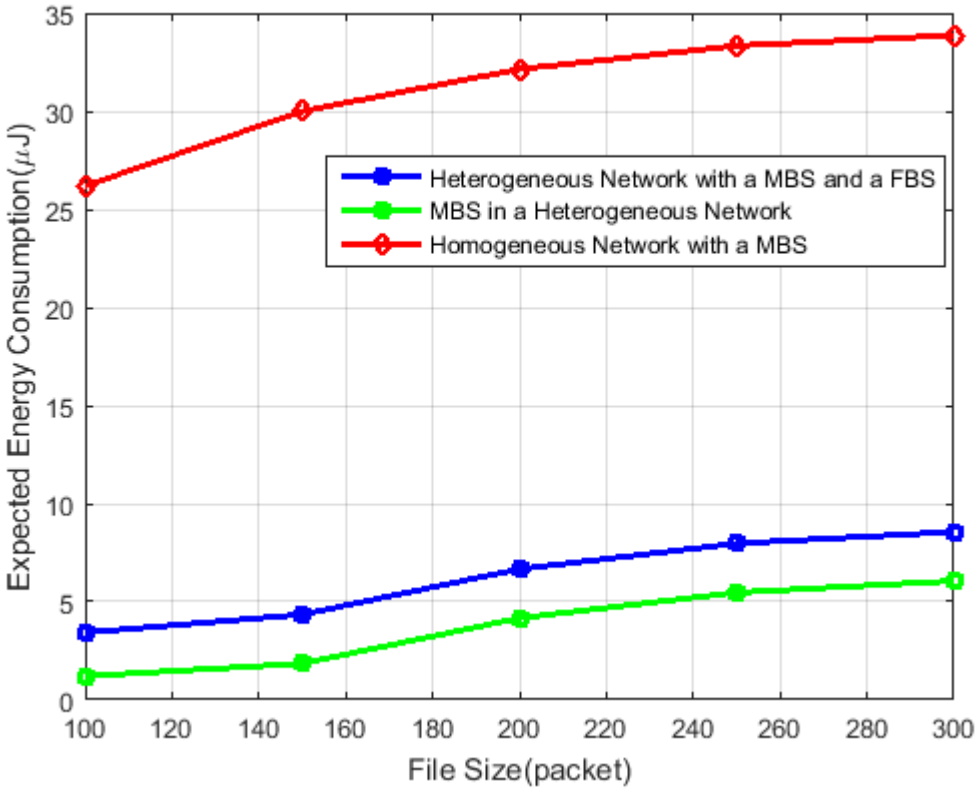

Figure 5: Expected energy consumption of cellular networks with initial-buffering delay x= 50, target starvation probability $\varepsilon = 0.15$, and $\gamma = 10$

6.2 Cellular Network Expected Energy Consumption for Different values of γ

In the following two experiments, we depict the impact of γ on the energy consumption of mobile networks. The fixed power (load-independent) consumption of MBS and FBS are typically set to 118.7 and 4.8 W, respectively [12]. Accordingly, the real value of γ, which denotes the ratio of energy consumption between a MBS and a FBS, can be approximately considered as 25. Fig. 6 and 7, illustrate the minimum amount of energy consumed for streaming a file with size N packets while ensuring a target starvation probability ε and for $\gamma = 25$, $\gamma = 40$, respectively. The file size in these experiments ranges between 100 and 300 packets, and the start-up delay x is set to 50 packets. The rest of parameters are the same as in part 6.1. According to these figures, as we increase the value of γ, the efficacy of our system in terms of reducing the expected energy consumption gets much better compared to the case of homogeneous mobile network. Furthermore, with a bigger γ, we can save more energy in MBS by using a FBS as the gap between red and green lines gets bigger with the increase in γ. The reason behind this is the fact that in a heterogeneous mobile network, the MD is able to get service from the FBS, and accordingly, we can keep the MBS mostly in sleep mode. However, in the homogeneous network, the MBS has to be more often in active to be able to satisfy the user required QoE. Therefore, this issue leads to a high energy consumption especially when the γ gets larger.

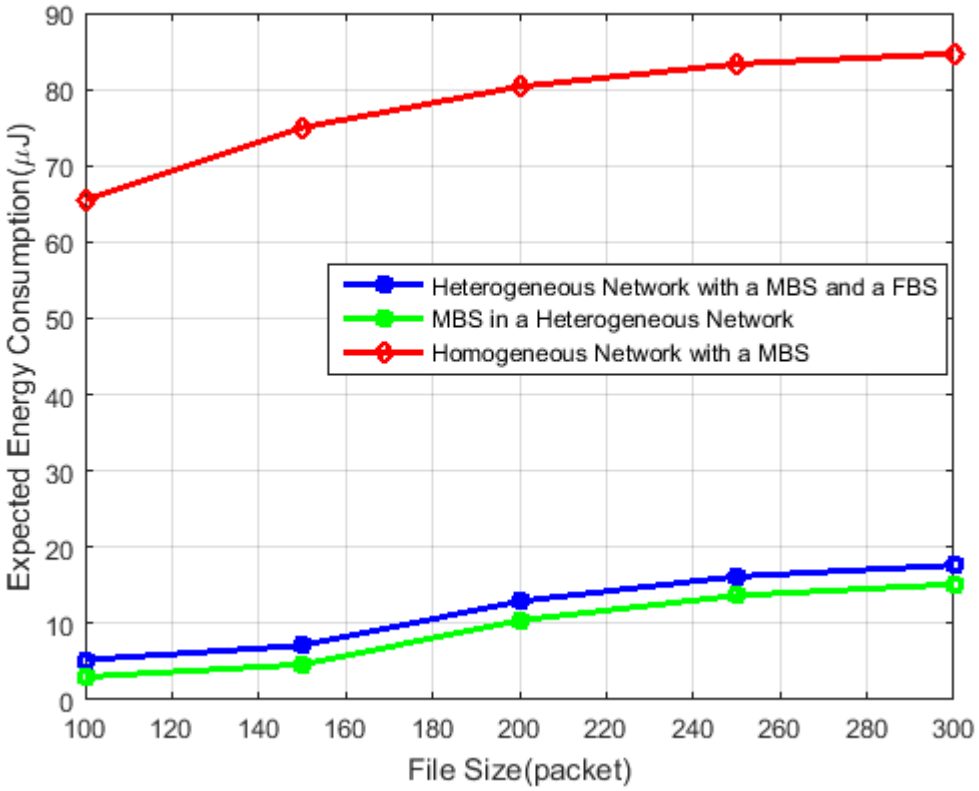

Figure 6: Expected energy consumption of cellular networks with initial-buffering delay x= 50, target starvation probability $\varepsilon = 0.15$, and $\gamma = 25$

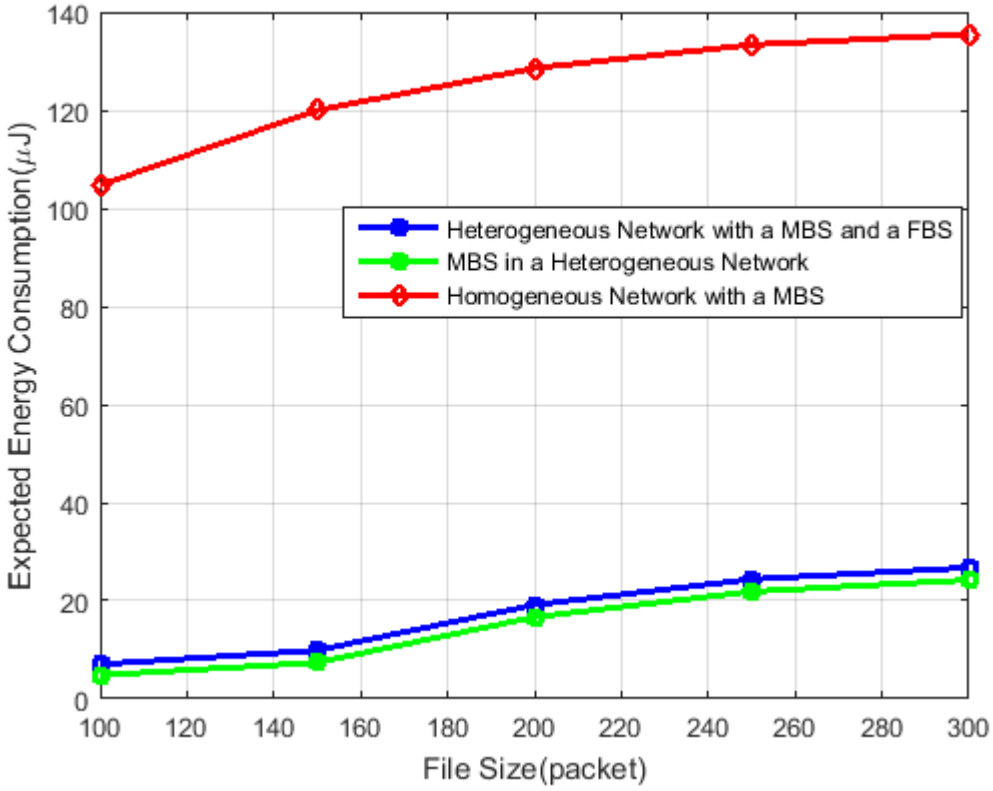

Figure 7: Expected energy consumption of cellular networks with initial-buffering delay x= 50, target starvation probability $\varepsilon = 0.15$, and $\gamma = 40$

6.3 Buffer Starvation Probability with respect to File Size

In Fig. 8, we plot the buffer starvation probability with initial buffering delay $x = 30$. The file size increases from 100 to 600 packets. In order to make a fair comparison in two systems, for the heterogeneous mobile network we set the active/sleep mode rates as $\alpha_m = \beta_m = 0.1$, $\alpha_f = \beta_f = 0.15$, however, in the in the homogeneous network, letting the rate α_m, the rate at which MBS goes to sleep mode, be 0.1, we raise the rate at which MBS goes to active mode, β_m, to 0.163 such that the expected energy consumption in two systems would be the same. The probability of buffer starvation increases with the increase in file size, however, the probability of having a buffer starvation while streaming a file (with the same size) in our system with two BSs is much less than the case where the MD could be served only through a single MBS. Moreover, as the file size increases, the starvation probability in a system with one MBS increases with a much faster rate compared to our system. The reason is that in a heterogeneous cellular network, the MD could be served by either MBS or FBS, and since the arrival rate from a FBS is usually more than that of a MBS, a starvation event occurs less frequently.

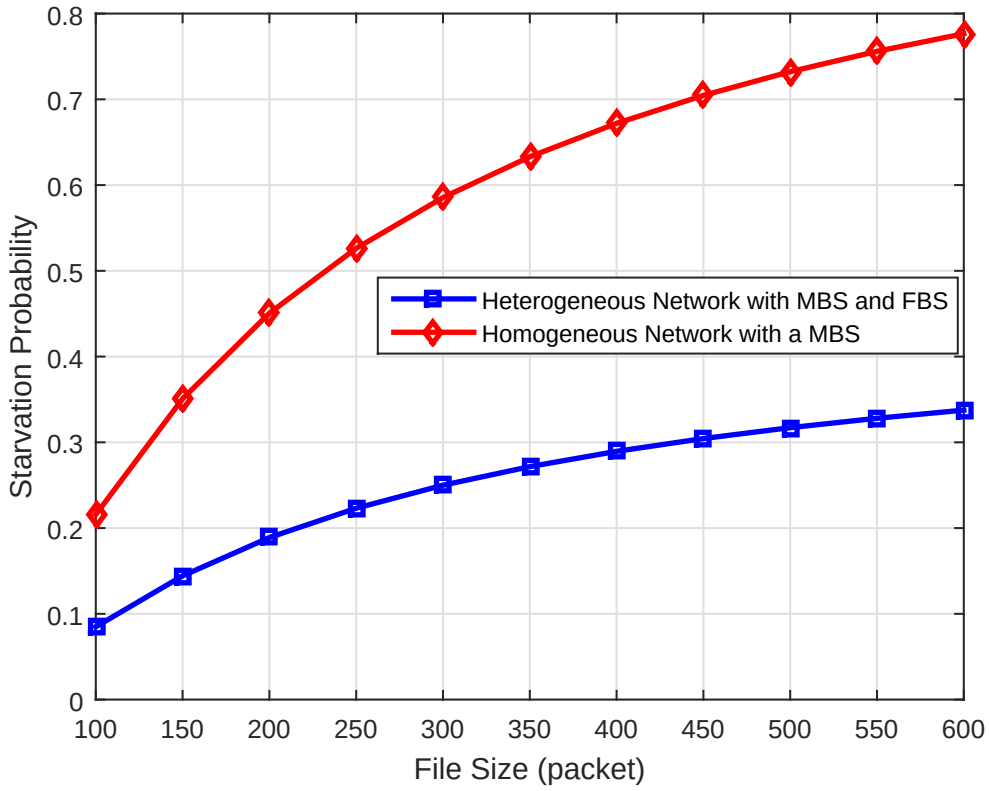

Figure 8: Buffer starvation probability in two systems with the same energy consumption and with the initial-buffering delay $x = 30$

6.4 Buffer Starvation Probability with respect to Start-Up Delay

Fig. 9 depicts the impact of start-up delay on the starvation probability. In this set of experiments, $N = 600$ and the start-up delay varies between 30 and 100 packets. In order to make a fair comparison in two systems, for the heterogeneous mobile network we set the active/sleep rates as $\alpha_m = \beta_m = 0.1$, $\alpha_f = \beta_f = 0.15$, however, in the in the homogeneous network, letting the rate α_m, the rate at which MBS goes to sleep mode, be 0.1, we raise the rate at which MBS goes to active mode, β_m, to 0.163 such that the expected energy consumption in two systems would be the same. λ_m, λ_f, and h are set to 1.5, 1.7, 10^{-5}, respectively. First, for the same file size and the same start-up delay, the starvation probability of a MD in our system model is much less than that of a MD in a system with a single MBS. Second, a slight increase in start-up delay can greatly improve the starvation probability in our system compared to the case where the MD is in a homogeneous cellular network.

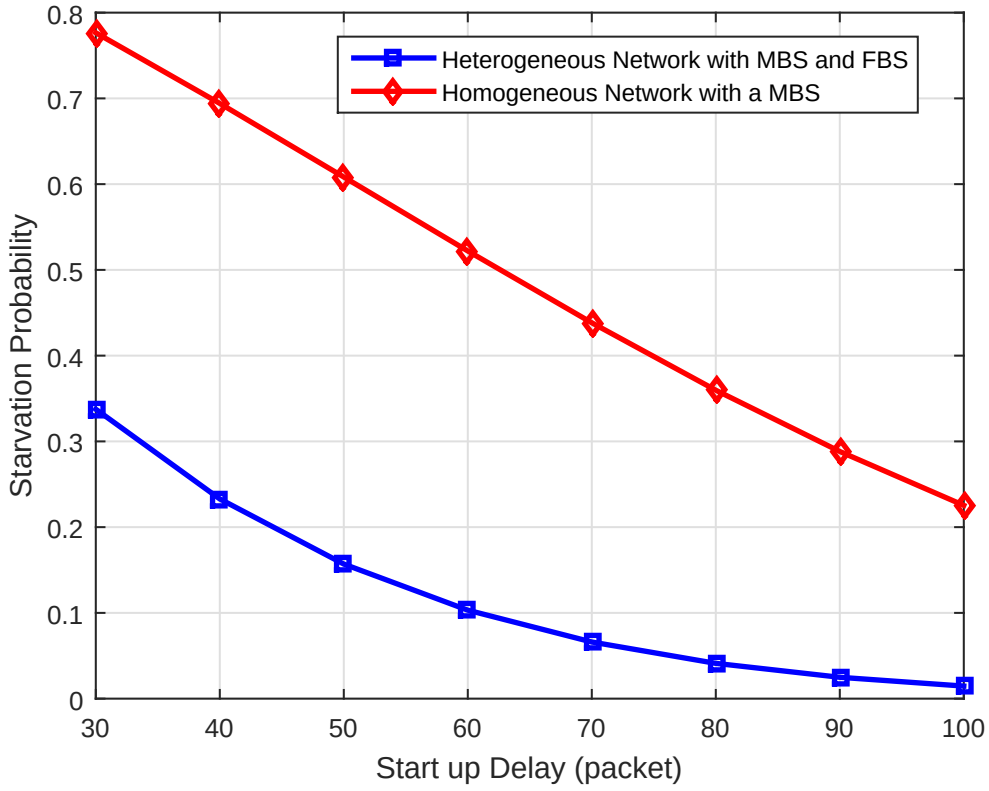

Figure 9: Buffer starvation probability in two systems with the same expected energy consumption and for streaming a file of size $N = 600$.

In the following simulation experiment, we investigate the effects of FBS arrival rate on the probability of starvation for different start up delay threshold. According to the Fig. 10, as we expected, reducing the FBS packet arrival rate λ_f from 1.7 to 1.5, while keeping all the other parameters the same as in part 6.3, increases the probability of starvation for streaming a video file of size N=600.

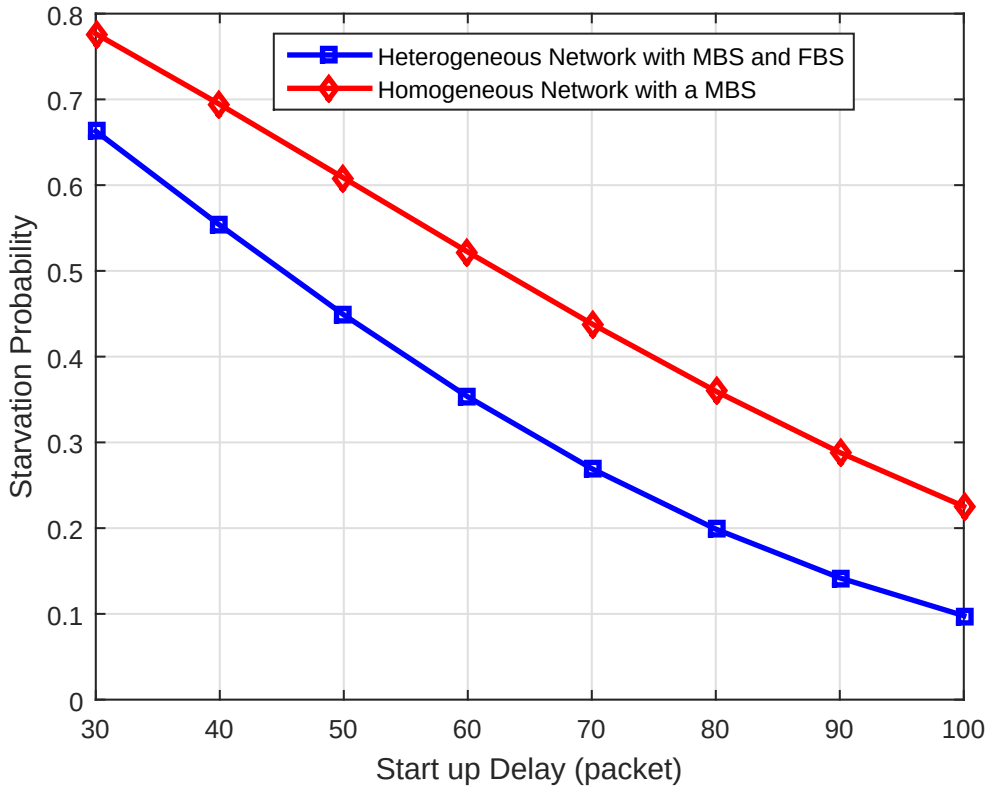

Figure 10: The effect of FBS packet arrival rate λ_f on starvation probability for streaming a file of size N= 600

6.5 Buffer Starvation Probability with respect to FBS packet arrival rate

In Fig. 11, for streaming a file of size N=600 packets with initial buffering delay x=30 packets, we characterize how FBS Poisson packet arrival rate λ_f affects the starvation probability. We let λ_m, and h be 1.5, 10^{-5}, respectively. Further, we increase the FBS Poisson packet arrival rate λ_f from 1 to 2. The active/sleep mode duration rates are as $\alpha_m = \beta_m = 0.1$, $\alpha_f = \beta_f = 0.15$. Clearly, the buffer starvation probability decreases with increase in the packet arrival rate λ_f to mobile device, as the buffer of the mobile device receive more packets per unit time with increase in the arrival rate.

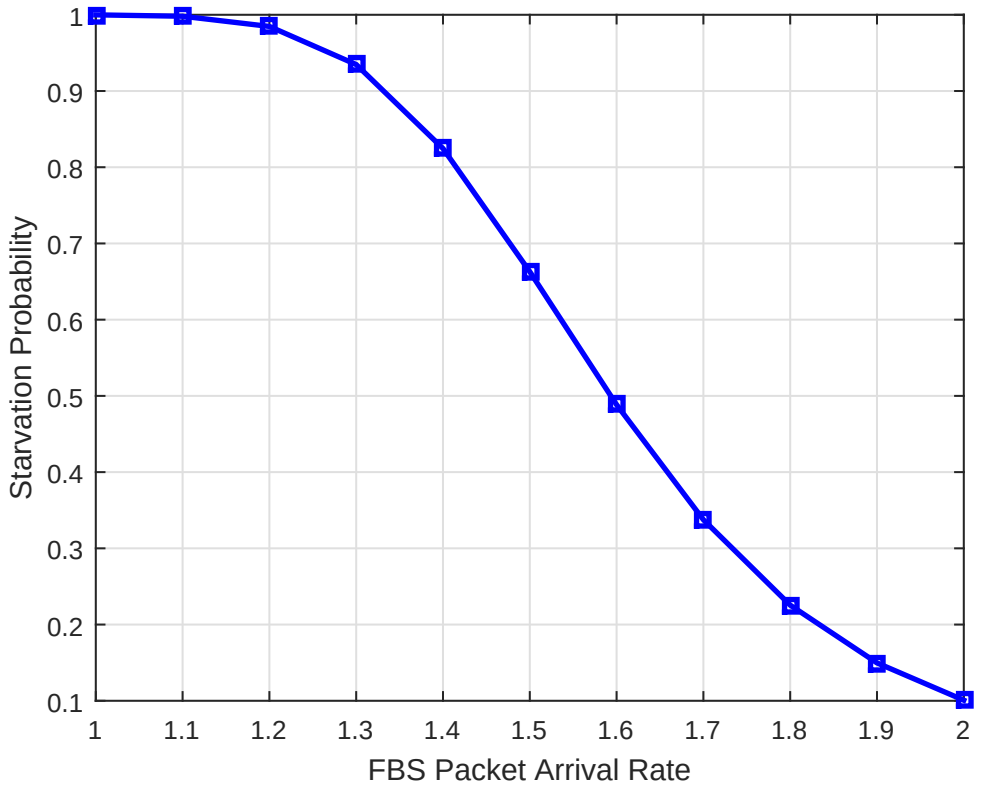

Figure 11: Buffer starvation probability for streaming a file of size N=600 packets and initial buffering delay x=30 packets.

6.6 Buffer Starvation Probability with respect to Energy Consumptions

In this experiment, we illustrate how the starvation probability is related to the energy consumption of BSs in a heterogeneous cellular network. We set the file size N, and start-up delay x to 300 and 60, respectively. The rate of going to active mode for MBS increases from 0.01 to 0.11, and this rate for FBS ranges between 0.05 and 0.15. The values of h, λ_m, λ_f are the same as in the first part of this section. It is clear that the starvation probability increases with the decrease in MBS's and FBS's energy consumption.

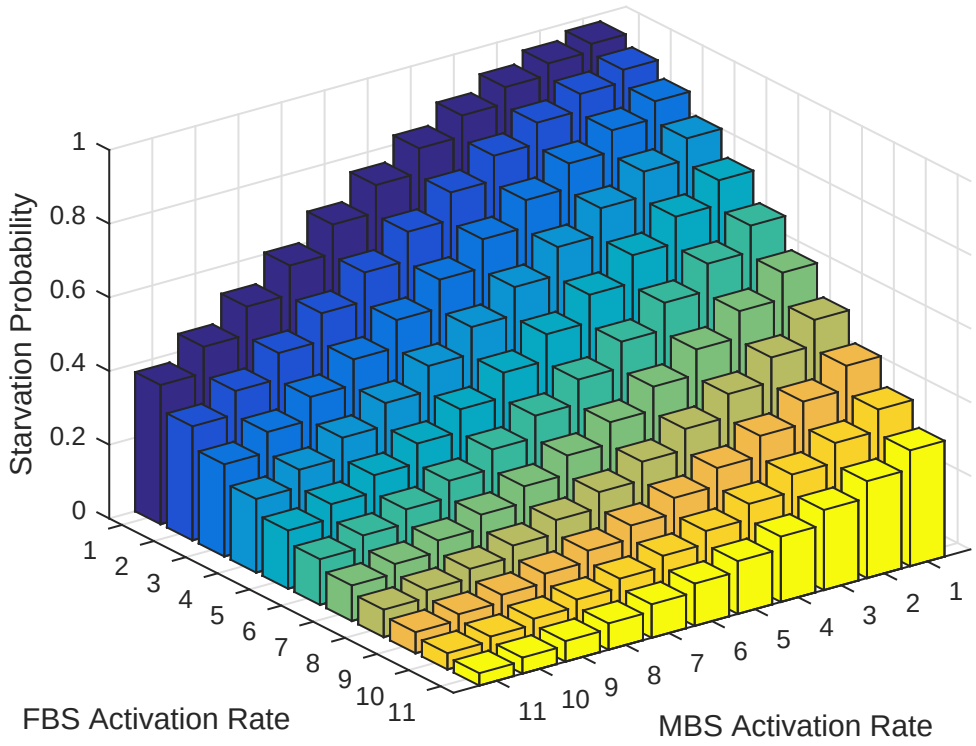

Figure 12: Buffer starvation probability with N= 300 and x= 60.

6.7 Buffer Starvation Probability with Respect to initial waiting time and Video File Size

Fig. 13 illustrates how the probability of starvation at the buffer of a MD is dependent to the both initial buffering delay and the length of the video going to be streamed. We let α_m, β_m be 0.1, and α_f, β_f be 0.15. λ_m, λ_f, and h are set to 1.5, 1.7, 10^{-5}. We increase the file size N from 150 to 650 packets, and increase the initial buffering delay from 30 to 80 packets. Clearly, as the file size increases, for a given start-up threshold, the probability of starvation increases as well. On the other hand, the starvation probability decreases, for a given file size, with increase the initial buffering delay.

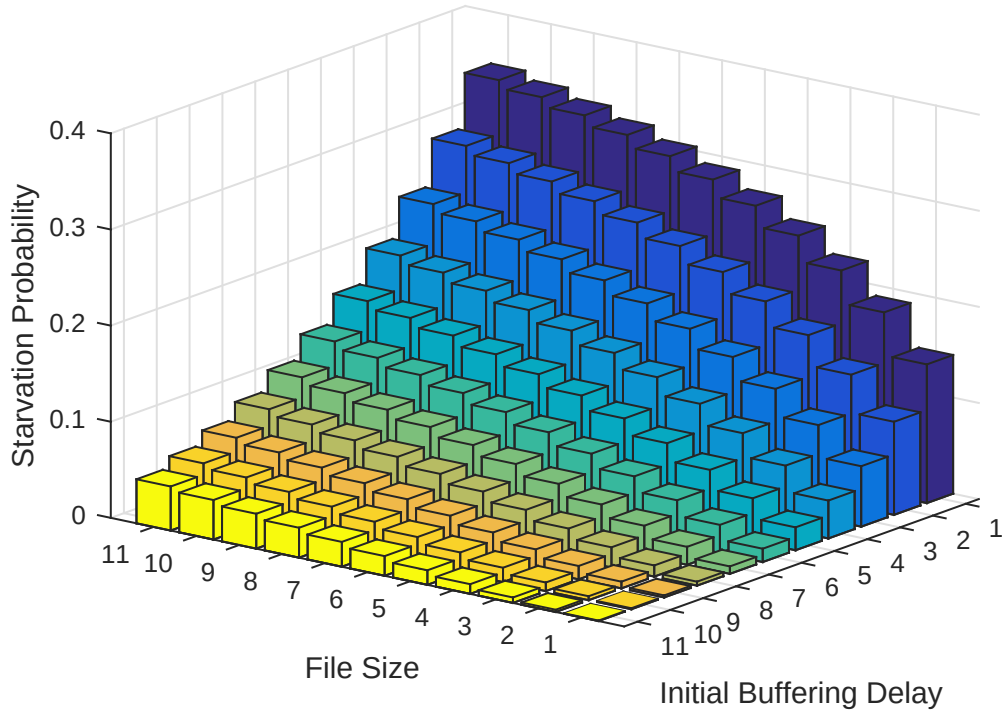

Figure 13: The changes in buffer starvation probability with file size N and initial waiting time x

6.8 Comparison of Buffer Starvation Probability in two different system models each with a single BS

In all the above simulations, we compare our system model composed of two BSs with another system model which contains a single BS and is actually taken from the work in [41].In order to validate the results of our numerical experiments, in Fig. 14, we plot the buffer starvation probability with respect to video file size for our system model and the system described in [41] while considering only a single BS in both system models. The goal is to check if the results of these two system models with one BS match each other or not. In this experiment, fhe file size increases from 100 to 600. The initial buffering delay x, the Poisson arrival rate λ_f, and the time-slot h are set to 30, 1.5, 10^{-5}, respectively. We let $\alpha_f = \beta_f = 0.1$, $\alpha_m = \beta_m = 0$. By setting β_m to 0, we confirm that only one BS in our system model is able to go to active mode, i.e., our system is operating with a single BS. As it is shown in Fig. 9, the figures from two different system models completely match one another, which help us to verify our numerical simulations.

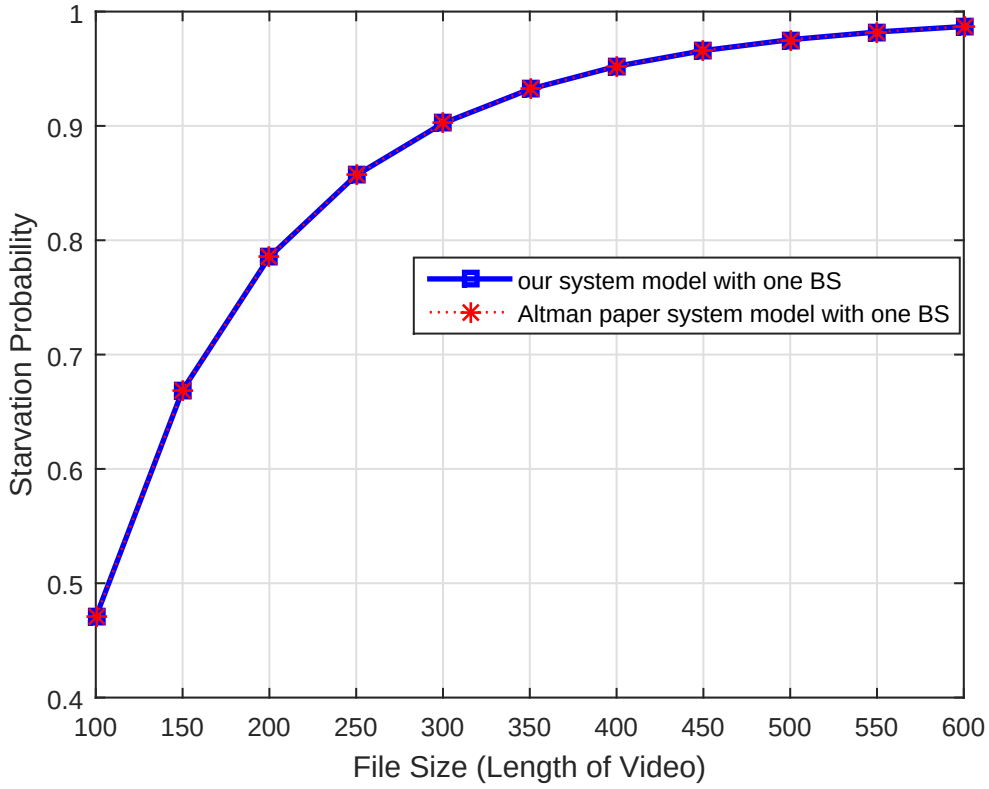

Figure 14: Buffer starvation probability with initial-buffering delay x= 30 in two different system models while both systems contain a single BS

7 Conclusions and Future Works

7.1 Conclusion

In this work, we have investigated the tradeoff between quality of experience and energy efficiency in a heterogeneous cellular network with two BSs, MBS and FBS. Proposing a simple system model, where MBS and FBS goes into sleep and active modes randomly throughout the system operation, we demonstrate the efficacy of heterogeneous cellular networks in minimizing the energy consumption while improving the QoE of MDs. The QoE that we consider in this work is the buffer starvation probability of a MD streaming a media file with finite size. For an on/off bursty traffic, we derived the buffer starvation probability of a MD in a system with multiple servers, where the MD could be served by a MBS or FBS depending on which one is in active mode, for the first time. In addition, by the use of a three state Markov chain and applying the first step analysis we investigated the aggregated active/sleep period length distribution analytically. The simulation results reveal that the proposed system model provides significant energy savings compared to a homogeneous cellular network. Finally, the proposed framework in this work can be used both for the energy efficient design and operation of different types of base stations in a heterogeneous networks and for improving the mobile devices' quality of experiences. In order to implement the proposed sleep/awake strategy in the real network, we should note that switching off BSs deteriorates the end user's desired QoE. Accordingly, we first investigate the impact of switching off BSs on the starvation probability for streaming a finite media file, and then based on the results, we obtain MBS/FBS optimal activation rate, β_m/β_f, and the optimal rate by which MBS/FBS should go to sleep mode, α_m/α_f, in a way that a threshold buffer starvation probability constraint is guaranteed.

7.2 Future Works

We believe that our model can be used as a useful starting point for future studies on interruption analysis in video streaming for mobile devices in a system with multiple servers, and specially in studying the aggregated active/sleep mode duration distribution. Interesting future direction to extend this work include developing analytical approaches towards analyzing the buffer starvation probability of mobile devices in a heterogeneous network with more than one femtocell base station.

References

[1] M. Ismail and W. Zhuang, *Cooperative networking in a heterogeneous wireless medium*. Springer, 2013.

[2] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 10–21, June 2011.

[3] D. Lopez-Perez, I. Guvenc, G. de la Roche, M. Kountouris, T. Quek, and J. Zhang, "Enhanced intercell interference coordination challenges in heterogeneous networks," *Wireless Communications, IEEE*, vol. 18, no. 3, pp. 22–30, June 2011.

[4] A. Ghosh, N. Mangalvedhe, R. Ratasuk, B. Mondal, M. Cudak, E. Visotsky, T. Thomas, J. Andrews, P. Xia, H. Jo, H. Dhillon, and T. Novlan, "Heterogeneous cellular networks: From theory to practice," *Communications Magazine, IEEE*, vol. 50, no. 6, pp. 54–64, June 2012.

[5] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: a survey," *Communications Magazine, IEEE*, vol. 46, no. 9, pp. 59–67, 2008.

[6] I. Ashraf, F. Boccardi, and L. Ho, "Sleep mode techniques for small cell deployments," *Communications Magazine, IEEE*, vol. 49, no. 8, pp. 72–79, 2011.

[7] J. Hoydis, M. Kobayashi, and M. Debbah, "Green small-cell networks," *Vehicular Technology Magazine, IEEE*, vol. 6, no. 1, pp. 37–43, 2011.

[8] T. Q. Quek, W. C. Cheung, and M. Kountouris, "Energy efficiency analysis of two-tier heterogeneous networks," in *Wireless Conference 2011-Sustainable Wireless Technologies (European Wireless), 11th European*. VDE, 2011, pp. 1–5.

[9] M. A. Marsan, L. Chiaraviglio, D. Ciullo, and M. Meo, "Optimal energy savings in cellular access networks," in *Communications Workshops, 2009. ICC Workshops 2009. IEEE International Conference on*. IEEE, 2009, pp. 1–5.

[10] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczók, "The global footprint of mobile communications: The ecological and economic perspective," *Communications Magazine, IEEE*, vol. 49, no. 8, pp. 55–62, 2011.

[11] Y. Chen, S. Zhang, S. Xu, and G. Y. Li, "Fundamental trade-offs on green wireless networks," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 30–37, 2011.

[12] G. Auer, V. Giannini, C. Desset, I. Godor, P. Skillermark, M. Olsson, M. A. Imran, D. Sabella, M. J. Gonzalez, O. Blume *et al.*, "How much energy is needed to run a wireless network?" *Wireless Communications, IEEE*, vol. 18, no. 5, pp. 40–49, 2011.

[13] T. Chen, Y. Yang, H. Zhang, H. Kim, and K. Horneman, "Network energy saving technologies for green wireless access networks," *Wireless Communications, IEEE*, vol. 18, no. 5, pp. 30–38, 2011.

[14] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Toward energy-efficient operation of base stations in cellular wireless networks," *Green Communications: Theoretical Fundamentals, Algorithms and Applications*, p. 435, 2012.

[15] C. Han, T. Harrold, S. Armour, I. Krikidis, S. Videv, P. M. Grant, H. Haas, J. S. Thompson, I. Ku, C.-X. Wang *et al.*, "Green radio: radio techniques to enable energy-efficient wireless networks," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 46–54, 2011.

[16] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 8, pp. 1525–1536, September 2011.

[17] A. J. Fehske, F. Richter, and G. P. Fettweis, "Energy efficiency improvements through micro sites in cellular mobile radio networks," in *GLOBECOM Workshops, 2009 IEEE*. IEEE, 2009, pp. 1–5.

[18] P. Rost and G. Fettweis, "11 green communications in cellular networks with fixed relay nodes," *Cooperative Cellular Wireless Networks*, p. 300, 2011.

[19] K. Son, E. Oh, and B. Krishnamachari, "Energy-aware hierarchical cell configuration: from deployment to operation," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*. IEEE, 2011, pp. 289–294.

[20] D. Brubaker, "Optimizing performance and efficiency of pas in wireless base stations: Digital pre-distortion reduces signal distortion at high power levels," *White Paper, Texas Instruments*, 2008.

[21] A. Ericsson, "Sustainable energy use in mobile communications," *white paper, EAB-07: 02l80l Uen Rev C*, 2007.

[22] H. Karl *et al.*, "An overview of energy-efficiency techniques for mobile communication systems," *Report of AG Mobikom WG7*, 2003.

[23] S.-E. Elayoubi, L. Saker, and T. Chahed, "Optimal control for base station sleep mode in energy efficient radio access networks," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 106–110.

[24] B. Rengarajan, G. Rizzo, and M. Ajmone Marsan, "Bounds on qos-constrained energy savings in cellular access networks with sleep mode," 2011.

[25] G. Jie, Z. Sheng, and N. Zhisheng, "A dynamic programming approach for base station sleeping in cellular networks," *IEICE Transactions on Communications*, vol. 95, no. 2, pp. 551–562, 2012.

[26] S. Mclaughlin, P. M. Grant, J. S. Thompson, H. Haas, D. Laurenson, C. Khirallah, Y. Hou, R. Wang *et al.*, "Techniques for improving cellular radio base station energy efficiency," *Wireless Communications, IEEE*, vol. 18, no. 5, pp. 10–17, 2011.

[27] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 5, pp. 2126–2136, 2013.

[28] J. Peng, P. Hong, and K. Xue, "Stochastic analysis of optimal base station energy saving in cellular networks with sleep mode," *Communications Letters, IEEE*, vol. 18, no. 4, pp. 612–615, 2014.

[29] L. Saker, S.-E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *Selected Areas in Communications, IEEE Journal on*, vol. 30, no. 3, pp. 664–672, 2012.

[30] M. Wildemeersch, T. Q. Quek, C. H. Slump, and A. Rabbachin, "Cognitive small cell networks: Energy efficiency and trade-offs," *Communications, IEEE Transactions on*, vol. 61, no. 9, pp. 4016–4029, 2013.

[31] X. Guo, S. Zhou, Z. Niu, and P. R. Kumar, "Optimal wake-up mechanism for single base station with sleep mode," in *Teletraffic Congress (ITC), 2013 25th International*. IEEE, 2013, pp. 1–8.

[32] K. Samdanis, D. Kutscher, and M. Brunner, "Self-organized energy efficient cellular networks," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2010 IEEE 21st International Symposium on*. IEEE, 2010, pp. 1665–1670.

[33] S. Bhaumik, G. Narlikar, S. Chattopadhyay, and S. Kanugovi, "Breathe to stay cool: adjusting cell sizes to reduce energy consumption," in *Proceedings of the first ACM SIGCOMM workshop on Green networking*. ACM, 2010, pp. 41–46.

[34] L. Chiaraviglio, D. Ciullo, M. Meo, M. A. Marsan, and I. Torino, "Energy-aware umts access networks," *Proc. of IEEE W-GREEN*, pp. 1–8, 2008.

[35] E. Oh and B. Krishnamachari, "Energy savings through dynamic base station switching in cellular wireless access networks," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*. IEEE, 2010, pp. 1–5.

[36] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell zooming for cost-efficient green cellular networks," *Communications Magazine, IEEE*, vol. 48, no. 11, pp. 74–79, 2010.

[37] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary users in cellular networks: A large-scale measurement study," in *New frontiers in dynamic spectrum access networks, 2008. DySPAN 2008. 3rd IEEE symposium on*. IEEE, 2008, pp. 1–11.

[38] Y. S. Soh, T. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 5, pp. 840–850, May 2013.

[39] G. Jie, Z. Sheng, and N. Zhisheng, "A dynamic programming approach for base station sleeping in cellular networks," *IEICE Transactions on Communications*, vol. 95, no. 2, pp. 551–562, 2012.

[40] J. Peng, H. Tang, P. Hong, and K. Xue, "Stochastic geometry analysis of energy efficiency in heterogeneous network with sleep control," *Wireless Communications Letters, IEEE*, vol. 2, no. 6, pp. 615–618, 2013.

[41] Y. Xu, E. Altman, R. El-Azouzi, M. Haddad, S. Elayoubi, and T. Jimenez, "Analysis of buffer starvation with application to objective qoe optimization of streaming services," *Multimedia, IEEE Transactions on*, vol. 16, no. 3, pp. 813–827, 2014.

[42] A. ParandehGheibi, M. Médard, A. Ozdaglar, and S. Shakkottai, "Avoiding interruptionsa qoe reliability function for streaming media applications," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 5, pp. 1064–1074, 2011.

[43] T. Stockhammer, H. Jenkăc, and G. Kuhn, "Streaming video over variable bit-rate wireless channels," *Multimedia, IEEE Transactions on*, vol. 6, no. 2, pp. 268–277, 2004.

[44] G. Liang and G. Liang, "Effect of delay and buffering on jitter-free streaming over random vbr channels," *Multimedia, IEEE Transactions on*, vol. 10, no. 6, pp. 1128–1141, 2008.

[45] T. H. Luan, L. X. Cai, and X. Shen, "Impact of network dynamics on user's video quality: Analytical framework and qos provision," *Multimedia, IEEE Transactions on*, vol. 12, no. 1, pp. 64–78, 2010.

[46] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4, pp. 362–373, 2011.

[47] F. Dobrian, A. Awan, D. Joseph, A. Ganjam, J. Zhan, V. Sekar, I. Stoica, and H. Zhang, "Understanding the impact of video quality on user engagement," *Communications of the ACM*, vol. 56, no. 3, pp. 91–99, 2013.

[48] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs," *Networking, IEEE/ACM Transactions on*, vol. 21, no. 6, pp. 2001–2014, 2013.

[49] I. Cidon, A. Khamisy, and M. Sidi, "Analysis of packet loss processes in high-speed networks," *Information Theory, IEEE Transactions on*, vol. 39, no. 1, pp. 98–108, 1993.

[50] M. Wellens, J. Riihijarvi, and P. Mahonen, "Modelling primary system activity in dynamic spectrum access networks by aggregated on/off-processes," in *Sensor, Mesh and Ad Hoc Communications and Networks Workshops, 2009. SECON Workshops '09. 6th Annual IEEE Communications Society Conference on*, June 2009, pp. 1–6.

[51] G. Fettweis and E. Zimmermann, "Ict energy consumption-trends and challenges," in *Proceedings of the 11th International Symposium on Wireless Personal Multimedia Communications*, vol. 2, no. 4, 2008, p. 6.

[52] M. Ismail and W. Zhuang, "Network cooperation for energy saving in green radio communications," *Wireless Communications, IEEE*, vol. 18, no. 5, pp. 76–81, 2011.

[53] G. Miao, "Energy-efficient uplink multi-user mimo," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 5, pp. 2302–2313, 2013.

[54] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 56–61, 2011.

[55] I. Humar, X. Ge, L. Xiang, M. Jo, M. Chen, and J. Zhang, "Rethinking energy efficiency models of cellular networks with embodied energy," *Network, IEEE*, vol. 25, no. 2, pp. 40–49, 2011.

[56] X. Ma, M. Sheng, Y. Zhang, and X. Wang, "Concurrent transmission for energy efficiency in heterogeneous wireless networks," *IEEE Trans. Wireless Commun*.

[57] G. Miao, N. Himayat, Y. Li, and A. Swami, "Cross-layer optimization for energy-efficient wireless communications: a survey," *Wireless Communications and Mobile Computing*, vol. 9, no. 4, pp. 529–542, 2009.

[58] J.-M. Kelif, M. Coupechoux, and F. Marache, "Limiting power transmission of green cellular networks: Impact on coverage and capacity," in *Communications (ICC), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1–6.

[59] E. Altman, M. K. Hanawal, R. ElAzouzi, and S. Shamai, "Tradeoffs in green cellular networks," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 3, pp. 67–71, 2011.

[60] E. Kurniawan and A. Goldsmith, "Optimizing cellular network architectures to minimize energy consumption," in *Communications (ICC), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4771–4775.

[61] B. Rengarajan, G. Rizzo, M. Ajmone Marsan, and B. Furletti, "Qos-aware greening of interference-limited cellular networks," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a*. IEEE, 2013, pp. 1–9.

[62] K. Son, H. Kim, Y. Yi, and B. Krishnamachari, "Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 8, pp. 1525–1536, 2011.

[63] K. Son and B. Krishnamachari, "Speedbalance: Speed-scaling-aware optimal load balancing for green cellular networks," in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2816–2820.

[64] W. Guo and T. O. Farrell, "Capacity-energy-cost tradeoff in small cell networks," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*. IEEE, 2012, pp. 1–5.

[65] A. ParandehGheibi, A. Ozdaglar, and M. Medard, "Qoe-aware media streaming in technology and cost heterogeneous networks," *arXiv preprint arXiv:1203.3258*, 2012.

[66] T. Han and N. Ansari, "Ice: Intelligent cell breathing to optimize the utilization of green energy," *Communications Letters, IEEE*, vol. 16, no. 6, pp. 866–869, 2012.

[67] ——, "On greening cellular networks via multicell cooperation," *Wireless Communications, IEEE*, vol. 20, no. 1, pp. 82–89, 2013.

[68] X. Li, X. Zhang, and W. Wang, "An energy-efficient cell planning strategy for heterogeneous network based on realistic traffic data," in *Computing, Management and Telecommunications (ComManTel), 2014 International Conference on*. IEEE, 2014, pp. 122–127.

[69] H. Tabassum, U. Siddique, E. Hossain, M. Hossain *et al.*, "Cellular downlink performance with base station sleeping, user association, and scheduling," *arXiv preprint arXiv:1401.7088*, 2014.

[70] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[71] O. Holland, V. Friderikos *et al.*, "Green spectrum management for mobile operators," in *GLOBECOM Workshops (GC Wkshps), 2010 IEEE*. IEEE, 2010, pp. 1458–1463.

[72] R. Pabst, B. H. Walke, D. C. Schultz, P. Herhold, H. Yanikomeroglu, S. Mukherjee, H. Viswanathan, M. Lott, W. Zirwas, M. Dohler *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *Communications Magazine, IEEE*, vol. 42, no. 9, pp. 80–89, 2004.

[73] X. J. Li, B.-C. Seet, and P. H. J. Chong, "Multihop cellular networks: Technology and economics," *Computer Networks*, vol. 52, no. 9, pp. 1825–1837, 2008.

[74] M. R. Zakerinasab and M. Wang, "A cloud-assisted energy-efficient video streaming system for smartphones," in *Quality of Service (IWQoS), 2013 IEEE/ACM 21st International Symposium on*. IEEE, 2013, pp. 1–10.

[75] P. E. M. D. Katz *et al.*, "Power consumption and spectrum usage paradigms in cooperative wireless networks," in *Cooperation in Wireless Networks: Principles and Applications*. Springer, 2006, pp. 365–386.

[76] A. P. Azad, S. Alouf, E. Altman, V. Borkar, and G. S. Paschos, "Optimal control of sleep periods for wireless terminals," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 8, pp. 1605–1617, 2011.

[77] [online], "Available: http://techcrunch.com/2010/11/19/web-video-37-percent-internet-traffic/."

[78] Sandvine, "The global internet phenomena report:," *2h 2014*, Jun 2014.

[79] Cisco, "Cisco visual networking : Forecast and methodology," *2013-2018*, Jun 2014.

[80] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *Wireless Communications, IEEE Transactions on*, vol. 2, no. 4, pp. 630–643, 2003.

[81] M. Mitsumura, H. Masuyama, S. Kasahara, and Y. Takahashi, "Buffer-overflow and starvation probabilities for video streaming services with application-layer rate-control mechanism," in *Proceedings of the 6th International Conference on Queueing Theory and Network Applications*. ACM, 2011, pp. 134–138.