Load Dependent Optimal ON-OFF Policies in Cellular Heterogeneous Networks

Deeksha Sinha
Electrical Engineering,
Indian Institute of Technology Bombay.
deekshasinha@iitb.ac.in

Veeraruna Kavitha
Industrial Engineering and Operations Research,
Indian Institute of Technology Bombay.
vkavitha@iitb.ac.in

Abhay Karandikar Electrical Engineering,
Indian Institute of Technology Bombay.
karandi@ee.iitb.ac.in

Abstract

The use of small cells has been proposed to increase system capacity by installation of base stations close to user location. Proximity of the base station with the user equipment also implies lesser power requirement for transmitting the same information. Thus one may expect improvement in energy efficiency. But installing a large number of base stations can also lead to an increase in the total energy consumption of the system. To combat this, mechanisms have been proposed to switch OFF these base stations at times of low load. In this report, we consider the problem of finding the fraction of base stations that can be switched OFF while maintaining quality of service (measured in terms of the average waiting time of users), for given load conditions. We also obtain the optimal switch OFF pattern. We do this in two steps. First, we determine the optimal ON-OFF pattern of base stations and user-base station association policy for a fixed fraction of base stations to be switched OFF. Then, we find the maximum fraction of base stations that can be switched OFF for given load conditions.

I. Introduction

The number of mobile subscriptions in the world is increasing at an exponential rate. From 4.6 billion at the end of 2009 [1], the number has increased to 6.8 billion at the end of 2012 [2], which is almost 96% of the world's population. Newer technologies like small cell networks, are being proposed to take care of these enormous demands. Alongside, more energy resources are being consumed to meet the increased demands. Today, Information and Communication Technology (ICT) sector is responsible for 2% of the world's total carbon emissions [3]. The GeSI SMART 2020 report [4] indicates that improvements in ICT could reduce the projected greenhouse gas emission of 2020 by 16.5%. Thus, we need to focus on new ways of meeting the increasing demands of cellular users and simultaneously making the system more energy efficient.

Energy efficiency of a cellular network can be enhanced at various levels - hardware components, system design and protocol level. A breakdown of the power consumption of a wireless network indicates that base stations consume about 80% of the total energy [5]. Though the energy consumption of the base station varies with the load, its fixed power consumption is quite high [6]. Thus, a base station consumes significant energy once it is ON. Hence, mechanisms have been proposed in which base stations are switched OFF or put in a 'sleep' mode during low load/traffic hours [5]–[12]. Typically there is a conservative deployment of the base stations to cater to the peak load, nevertheless, via these mechanisms the energy consumption can be controlled.

To meet the growing demands, the Third Generation Partnership Project (3GPP) standards have introduced the possibility of a heterogeneous network [13]. In this report, we focus on heterogeneous networks with cells of different sizes together covering an area. In these networks, in addition to the already existing macrocells, we would also have micro, pico or femto cells. As the number of users in any area varies significantly during the day [11], these small cells become redundant at some times. Since now the same area is covered by possibly two types of base stations (e.g., macro and a femto base station), one can use both the base stations during peak loads while one (mostly macro base station) is sufficient for low load periods. Thus, the sleep wake mechanisms become important in the context of small cell/heterogeneous networks.

A. *Previous Works*

In the recent years, a number of sleep-wake up mechanisms have been proposed. Sleep wake up algorithms in a homogeneous network setting i.e. a network consisting of only macro-cells are described in [6], [7], [8], [11]. Networks with both 2G and 3G users are studied in [12].

There are also some proposals for small cell based heterogeneous networks. In [9], the authors study wake up mechanisms with the focus on hardware of a small cell. In [14], the authors propose an offline optimized controller and elaborate on practical issues like activation time and ping-pong effect which are encountered while deploying sleep-wake up mechanisms.

In [5], a Markov Decision Process (MDP) based framework is used to study a network with femto and macro base stations. The optimal sleep-wake up policy based on the traffic and user localization is formulated as a solution to the MDP and can be evaluated using numerical methods. In this paper, the situations with complete, partial and delayed traffic information are considered separately. The above scheme provides a general solution which needs to be numerically computed, while in our work we obtain directly the optimal policy for a commonly encountered scenario. Further, we have derived a class of optimal policies, parametrized by the load/traffic conditions.

In [10], a system having linearly placed base stations with unidirectional antenna has been considered. They derive the optimal ON-OFF base station policy in two contexts: a) for a fixed fraction of base stations to be switched OFF, or b) for each base station to be switched OFF for atleast a given fraction of time. The optimal policy has been obtained using the tools of multimodularity - the counterpart of convex functions over integer sets.

In this work, we considerably build up on the above work. We extend it to a heterogeneous setting containing picocells and a macrocell and obtain again the optimal ON-OFF policy for a given switch-OFF ratio. Further and more importantly we study the finer structural properties of the optimal (bracket) policy and using this study we obtain the optimal fraction to be switched OFF given the load conditions. *In totality, given the load conditions and the QoS constraints, we obtain the optimal operational policies.* Also, we consider the use of bidirectional antennas. Using the structural properties so obtained, *we have a closed form expression for the average waiting time with bracket (optimal) ON-OFF policy.*

Any cellular network aims at providing good quality of service (QoS) to the users. For ensuring good service, for example, it might want to maintain the average waiting time of customers or the call blocking probability below a certain limit. Typically the QoS measures depend both upon the load factor (which depends upon the arrival rate and average work size) of the users as well as the number of available base stations to serve them. As the network also aims to minimize the energy consumption, depending on the traffic, we can operate a variable number of base stations which would meet the QoS requirements and consume only the minimum required amount of energy. In other words, one can switch OFF base stations when the traffic is low and we precisely take up this task: a) we first consider the design of optimal ON-OFF and user-base station association policies for the series of PBSs in such a heterogeneous network, given that a fraction η of them need to be switched OFF; b) and then, we evaluate the optimal

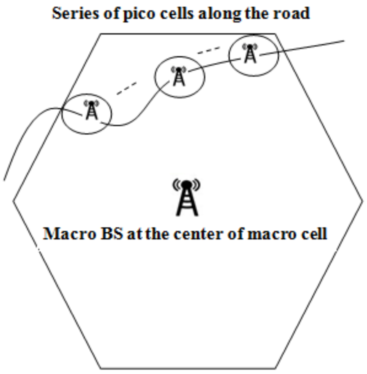

Series of pico cells along the road
Macro BS at the center of macro cell

Fig. 1. Heterogeneous Network

η to meet the QoS requirements. Towards this, we compute an expression for the desired QoS parameter both in terms of the load factor and the number of active servers (base stations), using queuing theoretic analysis.

We have considered an area with a major street, carrying significant traffic. Pico base stations (PBSs) have been installed along this street to meet the demands of the users. We have formulated a joint optimization problem to minimize the average waiting time and derived an optimal ON-OFF policy[1] and a user-base station association strategy. In this case, the optimal ON-OFF pattern is given by a *bracket sequence* which has a very simple form and can be easily computed. Further, we determined a closed form expression for minimum average waiting time, given the fraction of OFF PBSs. Using this, we obtained the optimal switch OFF fraction for any given load condition and QoS requirement.

B. Organization of the report

The remainder of the report is organized as follows. We describe the system model in Section II and then understand the performance of the system in section III. In section IV, we present the optimal control policies i.e. the ON-OFF pattern for PBSs and the user-base station association policy for a given switch OFF ratio. In section V, we find the maximum fraction of PBSs that can be switched OFF while preserving the QoS of the users. We conclude our work in Section VI. (Appendix A contains a brief review of concepts in multimodularity and proofs of theorems and lemmas are presented in Appendix B and Appendix C.)

II. System Description

When a cellular network has to be designed in a particular area, major infrastructural layout of the area is often known. Using this knowledge a better network can be designed. We consider one such scenario, where a busy road passes through a macro cell (Figure 1). Heavy traffic is usually generated on such roads, which can burden the macro base station (MBS). This is a good example, where a heterogeneous network can be deployed: the load of the MBS can be shared by a series of PBSs placed along the road. A similar situation will arise when a metro line passes through a macro cell. Base stations will need to be installed at various intermediate points along the metro line to cater to the demands of the large number of users who are traveling in the metro.

In our system model, PBSs are placed uniformly (at points $0,\ d,\ 2d, \cdots$) along the street/metro line (which lies in the area covered by the macro cell). [2] The street/line can have curvatures, bends etc as in Figure 1. But this can be transformed into a straight line via a homomorphism, as in most of the cases the street is straight locally. Thus, further analysis is done assuming the street to be a straight line.

[1] The terms ON-OFF policy, activation policy and sleep-wake up mechanism have been used interchangeably in this report.

[2] Throughout the below text, 'street' would refer to both metro line and a busy street.

1) Pico Base Stations: The system supports a finite choice of transmission rates (determined by modulation scheme, channel coding, etc). We assume here, that a PBS serves any user in its own cell (covering a distance $d/2$ on either side) with a fixed rate. This rate, θ^0, is determined by the distance d (the farthest user in its cell is located $d/2$ away) and fading parameters of the region. If a PBS is OFF and if one of its immediate neighbors is ON, the neighboring PBS serves the users at a smaller rate θ^1. The distance of the farthest user from this PBS ($3d/2$) along with fading and shadowing determine the choice of θ^1, however we can assume $\theta^1 < \theta^0$. Similarly, we can define θ^i as the rate that the ith PBS from the user's location would provide. Then, we would have

$$\theta^0 > \theta^1 > \cdots > \theta^i > \theta^{i+1} > \cdots \ .$$

As the users in one cell (e.g. $[-d/2, d/2]$) are served with one rate, we refer them as belonging to the midpoint (i.e. 0 in this case). To be more precise: All users arriving anywhere in the interval $[i+(2k-1)d/2,\ i+(2k+1)d/2]$ are modeled to have arrived at point $i+kd$. This representation divides all the users on the line at uniformly spaced points $(0, d, 2d, \cdots Nd)$ for some $N < \infty$. Each of these points represent all the users within $d/2$ distance from it. Thus, we can view the system as *queues of users* at points $0, d, 2d, \cdots$.

We assume that users are arriving independently. Also, the rate of arrivals is uniform throughout the line as can be expected on a highway or a metro line. As discussed above, in our representation, arrivals at one point are actually marked by arrivals in length d. Thus, the distribution of arrivals at the points $(0, d, 2d, \cdots Nd)$ are identical and are independent of each other. Also, the arrivals into each queue follow a Poisson process with rate λ.

2) Macro Base Station: We assume that the macro base station employs power control to ensure that the received rate is same for all users. Thus, we assume that the MBS serves any user on the street with the rate (θ^M), irrespective of its exact location. The analysis would also hold if MBS employs different rates of transmission and if all the possible rates are such that they lie between θ^J and θ^{J+1} for some J. We believe that this analysis would go through even if MBS serves the users from amongst a finite set of transmission rates, depending upon the fading and shadowing as long as the fading and shadowing characteristics remain same throughout the street (which is a reasonable assumption given that the MBS is far away from the street). In this case, the queue of users being served by the MBS would be modelled by an M/G/1 queue.

A. Control Policies

We first aim to determine the configuration which has minimum average waiting time that can be achieved by the users when a fraction η of the base station need to be switched OFF.

There are two control policies relevant to this purpose: a) The PBS activation policy which determines which PBSs are OFF b) Given an activation policy, a user-base station association policy which specifies the base station to which a user should connect to.

User-Base Station Association Policy: Each user can connect to any PBS which is ON or the MBS. If the PBS of its own cell is ON, the user obtains service at best rate (θ^0) from its PBS. When this PBS is OFF, an association decision has to be made.

Among all active BSs, a user gets connected to the one from which the received signal strength is the maximum. With high probability, this is the nearest ON PBS or an MBS if all the 'significant' neighboring PBSs are OFF. Thus we make the following natural choice of parametrized association policies referred to as *J-association policy*: Connect to the MBS if all the J neighbouring PBSs are OFF and if one or more of the J neighbours are ON, then connect to the nearest ON PBS.

To be more precise, we fix a number J and define the association order of the user among the base stations as - PBS at its own position $>$ the two PBSs at distance $d >$ the two PBSs at distance $2d \cdots >$ the two PBSs at distance $Jd >$ macro base station. The user connects to the first active (ON) base station based on this order and all users in a queue get served by the same base station.

Also, a PBS can be serving multiple queues of users simultaneously. *We assume that the resources (like channel bandwidth) of the OFF PBSs are appropriately reallocated among the ON PBSs.* Thus, if a PBS is serving multiple queues, its capacity increases proportionately so that the rates of the users are not affected by the fact that the PBS is also serving other queues. *As the power consumption of the PBS does not vary significantly with load [6], we assume that the power consumption does not change much with this increase in the number of users it it serving.*

III. System Performance

Let the *activation vector* $\mathbf{a} \in \{0,1\}^N$ represent the status of the base stations - $a_i = 1$ if the PBS at position i is OFF and it is 0 if the PBS is ON. *We assume that the PBS at* 0 *is always ON.* With large[3] N, this restriction does not alter the performance. In this report, bold letters like **a** represent an N length sequence while a partial sequence is defined by $\mathbf{a}_j^k := [a_j, a_{j+1}, \cdots, a_k]$.

Performance at n-th point: When the system uses activation vector **a** and J-association policy, then the transfer rate i.e. the rate at which the user at the point nd (user in the nth queue) is offered service, $\theta_n(\mathbf{a}, J)$ (for $n > 0$) is,

$$\theta_n(\mathbf{a}, J) = \begin{cases} \theta^M & \text{if } \ a_k = 1 \ \forall \ k : |n-k| \le J \\ \theta^{s_n} & \text{otherwise} \end{cases}$$

$$\text{where, } \ s_n = \inf_{k:|n-k|\le J} \left\{ |n-k| : a_k = 0 \right\}.$$

For technical purposes , we define $\theta_n(\mathbf{a}, J) = \theta^0$ for $n \le 0$. Note that the system performance can be controlled only via $\{\theta_n(.,.)\}_{n\ge 1}$ (as the system consists of PBSs starting from location 0 and the PBS at 0 is always ON).

The queue at each point among $(0, d, 2d, \cdots Nd)$ is served by a BS (PBS or MBS depending upon $(\mathbf{a}, J)$) independent of other points. Thus, we can model each point as an independent queue with Poisson arrivals at rate λ. We *assume that the amount of information S each user has to transmit is exponentially distributed with mean s.* Given the activation vector **a** and J-association policy, a user who arrives at point nd, is served at rate $\theta_n(\mathbf{a}, J)$. Thus, the user occupies the server of the serving BS for a random time, $S/\theta_n(\mathbf{a}, J)$. Hence, for any given pair of policies $(\mathbf{a}, J)$, the service time is exponentially distributed with mean $s/\theta_n(\mathbf{a}, J)$. Thus, at point nd we have an M/M/1 queue with average waiting time given by ([15]):

$$W_n(\mathbf{a}, J) \ = \ w(\theta_n(\mathbf{a}, J)) \text{ with } w(\theta) := \frac{\lambda}{\frac{\theta}{s}\left(\frac{\theta}{s} - \lambda\right)}. \tag{1}$$

Here $w(\theta)$ represents the average waiting time of a user in an M/M/1 queue with parameters λ and s/θ and hence the average of all users in the system when each of them is being served at the rate θ. The average waiting time of a typical user is obtained by first conditioning on its position of arrival and then taking average over the arrival position. By our assumption, a user is equally likely to arrive at any point nd. Thus, the average waiting time of a typical user equals :

$$\frac{1}{N}\sum_{n=0}^{N} W_n(\mathbf{a}, J).$$

We are considering a long street and *hence assume that N is sufficiently large.* Thus, the average waiting time of a typical user is well approximated by the limit[4]:

$$\overline{W}(\mathbf{a}, J) = \limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} W_n(\mathbf{a}, J).$$

[3] The major streets or metro lines run over kilometers and the pico cells are usually separated by few hundreds of meters and hence N can be large.

[4] As the limit may not exist for every activation vector **a**, we take an upper bound given by the limit superior. We will show that the optimal sequence is periodic, which in turn makes the optimal $\{W_n\}$ periodic and then the limit superior equals the limit.

(Note $W_0(.,.) = w(\theta^0)$ is fixed)

The first aim of this report is to determine the pair $(\mathbf{a}, J)$[5] which minimizes $\overline{W}(\mathbf{a}, J)$ while switching OFF an appropriate fraction of PBSs, i.e.,

$$\min_{\mathbf{a},J} \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} W_n(\mathbf{a}, J)$$
$$\text{subject to } \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} a_n = \eta. \tag{2}$$

Let $\bar{J}$ represent the distance of the farthest PBS from which the transfer rate is better than that provided by the MBS, i.e.

$$\bar{J} := \max_k \left\{ k : \theta^k > \theta^M \right\}.$$

We will show that $\bar{J}$-association policy is the optimal one.

IV. Optimal policies given switch OFF ratio η

A. Optimal On-Off Policy

We begin with the derivation of the optimal activation policy given the J-association policy and the condition that η fraction of the PBSs needs to be switched OFF. Towards this, we first consider the following optimization for any fixed $J \le \bar{J}$:

$$\min_{\mathbf{a}} \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} W_n(\mathbf{a}, J)$$
$$\text{subject to } \liminf_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} a_n \ge \eta, \tag{3}$$

and show that the optimizer satisfies the constraint with equality.

The solution to this problem is obtained, using concepts of multimodularity. A brief overview of multimodular functions has been presented in Appendix A and we obtain the proof of the following theorem in Appendix B.

Theorem 1. *For $J \le \bar{J}$, the solution of (3) is given by a bracket sequence $\mathbf{a}^*$ for some $\beta \in [0, 1)$ where*
$\boldsymbol{a}^* := \{a_n\}_{n\ge1}\,;\ a_n = \lfloor n\eta + \beta \rfloor - \lfloor (n-1)\eta + \beta \rfloor.$ □

In this text, $\lfloor . \rfloor$ and $\lceil . \rceil$ represent the floor and ceil functions respectively.

Every β defines an optimal policy. Without loss of generality, we consider the policy with $\beta = 0$ for all further discussions.

The bracket sequence $\mathbf{a}^*$ in fact, satisfies (3) with an equality (Lemma 5.1 in [16]). Thus, the bracket policy is also optimal if the inequality in (3) is replaced by an equality.

It is easy to see that if η is rational ($\eta = k_1/k_2$), then the bracket sequence is periodic with period k_2. Since rational numbers are dense in $\mathbb{R}$, we indeed assume a rational η in all our discussions below. Thus, when the optimal policy is used the ON-OFF pattern of the PBSs will be periodic.

Further, it should be noted that this optimal bracket policy depends only upon η and is not influenced by other factors, like J, θ^0, θ^M etc. This policy also has a very simple and regular form which permits easy calculation.

[5] With $N \to \infty$, $\mathbf{a}$ now is in $\{0, 1\}^\infty$.

B. Optimal user-base station association policy

As seen in the previous section for any $J \leq \bar{J}$, once the switch OFF ratio η is fixed, the optimal activation policy is independent of J. We now obtain the optimal J with the following theorem (proof can be found in Appendix B).

Theorem 2. *$(\boldsymbol{a}^*, \bar{J})$ is the solution for the optimization problem (2) i.e. $\overline{W}(\boldsymbol{a}, J) \geq \overline{W}(\boldsymbol{a}^*, \bar{J}) \ \ \forall \ J$ and for all activation vectors $\boldsymbol{a}$ in which the fraction of base stations that are switched OFF is η.* □

Thus given η, the fraction of PBSs to be switched OFF, the policies minimizing the average waiting time are $\mathbf{a}^*$ and $\bar{J}$. That is, via the above the theorem we prove our intuitions are correct: for any η it is optimal to connect to the PBSs as long as the signal from them is better than that from MBS (with high probability). In the next section, we find the maximum η *(among rational numbers)* which satisfies the QoS requirement *for any given traffic/load conditions (λ and s). The load conditions are reflected via the term $w(\theta)$ of (1).*

V. Optimal switch OFF ratio η

A network is usually designed to maintain certain desired QoS level throughout the day, irrespective of the time varying load conditions. We assume that the average waiting time needs to be maintained below $\overline{W}_{QoS}$. For any load, there exists a range of switch OFF ratios (η), which meet this QoS requirement. On the other hand, higher η implies higher fraction of PBSs which are switched OFF and thus, lower energy consumption. Hence, we consider the following optimization problem -

$$\sup_{\mathbf{a},J} \ \eta \quad \text{subject to} \tag{4}$$
$$\overline{W}(\mathbf{a}, J) \leq \overline{W}_{\text{QoS}} \quad \text{and} \qquad \liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} a_n \geq \eta.$$

From the previous sections, for a given η, the minimum waiting time is given by $\overline{W}(\mathbf{a}^*, \bar{J})$. Let $\overline{W}^*(\eta) := \overline{W}(\mathbf{a}^*, \bar{J})$ represent this optimal value for a given η. Note here, $\mathbf{a}^*$ depends only upon η. We obtain solution to (4) in two steps: a) we obtain an explicit expression for $\overline{W}^*(\eta)$ in terms of η and show that it is monotone in η; b) we show that the η', that satisfies the equation $\overline{W}^*(\eta') = \overline{W}_{QoS}$, is the required solution.

A. Explicit expression for $\overline{W}^*(\eta)$

We study finer structural properties of the bracket policies which help us obtain the expression for $\overline{W}^*(\eta)$.

$\overline{W}^*(\eta)$ depends on two factors - the rates at which users are being served in different queues and the frequency of each such rate. As we deal with rational η, we take $\eta = k_1/k_2$ for integers k_1 and k_2. We know that the sequence $\mathbf{a}^*$ is periodic with period k_2. Thus for any i,

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} W_n(\mathbf{a}^*, \bar{J}) = \frac{1}{k_2}\sum_{n=i}^{i+k_2-1} W_n(\mathbf{a}^*, \bar{J}). \tag{5}$$

We thus consider a block of k_2 queues. Users in each of these k_2 queues are served with a common rate from among the set $\left\{\theta^0, \theta^1, \cdots \theta^{\bar{J}}, \theta^M\right\}$.

In the activation vector $\mathbf{a}^*$, the fraction of OFF PBSs is exactly equal to η and hence in a block of k_2 PBSs, the number of OFF PBSs will be $\eta \times k_2 = k_1$.

1) Analysis of the bracket policy: Let us analyze the activation vector $\mathbf{a}^*$ which is expressed as $\mathbf{a}^*_n = \{\lfloor \eta n \rfloor - \lfloor \eta(n-1) \rfloor\}_{n \geq 1}$. Consider a block of k_2 consecutive PBSs from $n_0 = mk_2$ to $(m+1)k_2 - 1$ for any integer m. We have,

$$\begin{aligned} a_{n_0} &= \lfloor \eta n_0 \rfloor - \lfloor \eta(n_0 - 1) \rfloor \\ &= \lfloor mk_1 \rfloor - \lfloor mk_1 - \frac{k_1}{k_2} \rfloor \\ &= mk_1 - (mk_1 - 1) \\ &= 1. \end{aligned}$$

To find the next OFF PBS, we need to find the smallest integer $s > 0$ such that $a_{n_0+s} = 1$ i.e. such that

$$\lfloor \eta(n_0 + s) \rfloor - \lfloor \eta(n_0 + s - 1) \rfloor = 1.$$

For any integer $s > 0$, we have $\lfloor \eta(n_0 + s - 1) \rfloor \geq mk_1$. Thus, the required s is the smallest integer s with

$$\lfloor \eta(n_0 + s) \rfloor = mk_1 + 1 \text{ and } \lfloor \eta(n_0 + s - 1) \rfloor = mk_1.$$

But

$$\Big\lfloor \eta(n_0 + s) \Big\rfloor = \Big\lfloor \frac{k_1}{k_2}(mk_2 + s) \Big\rfloor = mk_1 + \Big\lfloor s\frac{k_1}{k_2} \Big\rfloor.$$

So, we need the smallest s such that, $\Big\lfloor s\frac{k_1}{k_2} \Big\rfloor = 1$ and we obtain the following:

Lemma 1. $\Big\lceil p\frac{k_2}{k_1} \Big\rceil$ *is the smallest* s *such that* $\Big\lfloor s\frac{k_1}{k_2} \Big\rfloor = p$. □

(Proof in Appendix C)
Thus, position of the next OFF PBS after n_0, is given by $n_0 + \lceil \frac{k_2}{k_1} \rceil$.

Proceeding in the same manner, the pth next OFF PBS can be found out by solving for the smallest s such that $\lfloor (mk_2 + s)\frac{k_1}{k_2} \rfloor = mk_1 + p$ and $\lfloor (mk_2 + s - 1)\frac{k_1}{k_2} \rfloor = mk_1 + p - 1$. Using Lemma 1 again, we get $s = \lceil p\frac{k_2}{k_1} \rceil$. Thus,

Lemma 2. *For* $n_0 = mk_2$*, where* m *is an integer and* $\eta = \frac{k_1}{k_2}$*,*

$$\begin{aligned} \boldsymbol{a}_{n_0+i} &= 1 \textit{ if } i = \Big\lceil \frac{p}{\eta} \Big\rceil \textit{ for some } p \in \{1, 2 \cdots\} \\ &= 0 \textit{ otherwise.} \end{aligned}$$

□

We call a queue to be of type j if users in that queue are being served with the rate θ^j *by a PBS*. Let the total number of types of queues which are being served by PBSs in the block be $l(\eta)$. The rest, if any left, are served by MBS. It is easy to see that if there exists a queue of type j, then there will exist queues of types i whenever $0 \leq i \leq j$. Thus, $l(\eta) = i$ would mean that the set of possible rates being delivered by the PBSs is $\{\theta^0, \theta^1, \cdots, \theta^{i-1}\}$ and for every rate in this set, there will exist at least one queue being served at that rate. We have the following result -

Lemma 3. $l(\eta) = r + 1$ *for* $h(r-1) < \eta \leq h(r)$ *where* r *is an integer and* $h(r) := \dfrac{2r}{1 + 2r}$. □

This is obtained by proving the following steps:

1) $\{\eta : l(\eta) = r + 1\} \subset \{\eta \leq h(r)\}$

2) $\{\eta \leq h(r)\} \subset \{\eta : l(\eta) \leq r+1\}$

3) $\{h(r-1) < \eta \leq h(r)\} = \{\eta : l(\eta) = r+1\}$.

The detailed proof of each step can be found in Appendix C.

Determining frequency of each type of queue

Having found the different types of possible queues for a given η, we need to determine the number of each type of queue in the k_2 block. We obtain these frequencies and then the final expression for $\bar{W}^*(\eta)$ in the following:

Theorem 3. *When $\eta = k_1/k_2$ and when $h(r-1) < \eta \leq h(r)$ for some r, with $\gamma := \min\{r-1, \bar{J}\}$ we have the following in a block of k_2 consecutive PBSs:*
1. $(1-\eta)$ fraction of the PBSs are ON.
2. $2(1-\eta)$ fraction of PBSs are of type θ^i for each $1 \leq i \leq \gamma$.
3. Remaining are either of type r (if $r-1 < \bar{J}$) or are connected to the MBS.
The expression for the minimum average waiting time is given by (6)

$$\overline{W}^*(\eta) = w(\theta^x) - (1-\eta)\left(\sum_{k=0}^{\min(r-1,\bar{J})} w(\theta^k) b_{r,k} + w(\theta^x)\left(1 + 2\min(r-1,\bar{J})\right)\right) \textit{ with} \tag{6}$$

$$x = \begin{cases} r & \textit{if } r-1 < \bar{J}, \\ M & \textit{if } r-1 \geq \bar{J} \end{cases}, \quad b_{r,k} = \begin{cases} -1 & \textit{if } k = 0, \\ -2 & \textit{if } 1 \leq k \leq \min(r-1,\bar{J}) \end{cases}$$

when $h(r-1) < \eta \leq h(r)$ with $h(r) = \frac{2r}{1+2r}$.

□

Here we present a brief sketch of the proof while the details are in Appendix C. We first show that the minimum distance between any two consecutive ON PBSs is $2r-1$. Let S_i be the set containing i-th ON PBS and its $r-1$ neighbors on each side (which are OFF). Each such set will contain a queue at the ON PBS location being served at the rate θ^0 and two queues being served at $\theta^i, \forall\ 1 \leq i \leq \gamma$. Rest of the PBSs (excluding $\cup_i S_i$), if any, are either connected to the MBS or are of type r. Using this, we get equation (6) as the expression for minimum waiting time when η fraction of the PBSs have to be switched OFF.

B. Solution of (4)

By the above analysis, we derived an expression for the minimum average waiting time possible for a given η.

As seen from (6), $\overline{W}^*(\eta)$ is piecewise continuous and linearly increasing function of η. The continuity at interval boundaries (at $h(r)$ for any r) can be seen by calculating the left and right limits of the waiting time at the boundary (Lemma 5 and Lemma 6 in Appendix C). Thus, $\overline{W}^*(\eta)$ is a continuous non-decreasing function of η.

The average waiting time will be the least, i.e. $w(\theta^0)$, when all the PBSs are ON. It will be the maximum, i.e. $w(\theta^M)$, when all the PBSs are OFF and all the queues are being served by the MBS. Thus, average waiting time always takes values between $w(\theta^0)$ and $w(\theta^M)$. Hence, if $\overline{W}_{\text{QoS}} \geq w(\theta^M)$, then all PBSs can be switched OFF. If $\overline{W}_{\text{QoS}} < w(\theta^0)$, then it is not possible to meet the QoS requirement with the given system parameters.

When $w(\theta^0) \leq \overline{W}_{\text{QoS}} \leq w(\theta^M)$, the fraction η' which satisfies $\overline{W}^*(\eta') = \overline{W}_{\text{QoS}}$, is given by:

$$\eta' = 1 - \frac{w(\theta^x) - \overline{W}_{\text{QoS}}}{w(\theta^x)\left(1 + 2\min(r'-1, \bar{J}\right) + \displaystyle\sum_{k=0}^{\min(r'-1,\bar{J})} w(\theta^k) b_{r',k}}$$

where r' is such that $\overline{W}^*(h(r'-1)) < \overline{W}_{\text{QoS}} \leq \overline{W}^*(h(r'))$.

Theorem 4. *η' is the solution to the optimization problem (4).*

The proof of the above theorem can be found in Appendix C.

Thus, η' is the maximum fraction of PBSs that can be switched OFF for the given load conditions (reflected through η's dependence on $w(\theta)$) while meeting the QoS constraint.

VI. Conclusions

We considered a heterogeneous network containing a macro cell and a series of pico cells placed along a major street and obtained the system performance, the average waiting time. Two control policies were relevant in this context: 1) activation policy indicating the ON-OFF status of pico base stations; 2) user base station association policy for a given activation policy. Using multi-modularity tools, for any given switch OFF ratio we showed that a periodic sequence (called bracket sequence) is an optimal activation policy while 'connect to that ON base station, which maximizes the transfer rate' is shown to be the optimal user base station association policy. This pair of policies jointly minimize the average waiting time. We also showed that the optimal activation policy depends only upon the switch OFF ratio and is independent of transfer rates and other system parameters. Further, we have an explicit expression for this policy which can easily be evaluated. One of the important contributions of this work is that, we obtained the explicit expression for the average waiting time under the optimal policies, i.e., the optimal average waiting time. We did this by obtaining the finer structural properties of the bracket policies. Using this expression, we obtained the optimal switch OFF ratio (among rational numbers) for any given load conditions (specified in terms of arrival rates and average amount of information to be transferred) which meets the QoS requirements.

As future work, we intend to understand the system performance when fading is considered in the signals of the PBSs and the MBS. We would also like to relax the assumption of users arriving uniformly. In this case, it would be more appropriate to consider decentralized policies where every PBS will switch ON-OFF depending on the traffic load and probably the number of waiting users at its own queue.

Acknowledgment

This work is supported by the India-UK Advanced Technology Centre (IU-ATC) of Excellence in Next Generation Networks, systems and services and funded by the Department of Science and Technology (DST), Government of India.

References

[1] International Telecommunication Union. (2009) The World in 2009: ICT Facts and Figures. [Online]. Available: http://www.itu.int/ITU-D/ict/material/Telecom09-flyer.pdf

[2] ITU. (2013) The World in 2013: ICT Facts and Figures. [Online]. Available: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013.pdf

[3] Z. Hasan, H. Boostanimehr, and V. K. Bhargava, "Green Cellular Networks: A Survey, Some Research Issues and Challenges," *IEEE Communications Surveys and Tutorials*, vol. 30, no. 4, 2011.

[4] GeSI. (2008) GeSI SMART 2020 report. [Online]. Available: http://gesi.org/SMARTer2020

[5] L. Saker, S. E. Elayoubi, R. Combes, and T. Chahed, "Optimal Control of Wake Up Mechanisms of Femtocells in Heterogeneous Networks," *IEEE Journal on Selected Areas in Communication*, vol. 30, no. 3, pp. 664–672, 2012.

[6] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Towards Dynamic Energy Efficient Operation of Cellular Network Infrastructure," *IEEE Wireless Communications Magazine*, vol. 49, pp. 56–61, June 2011.

[7] Z. Niu, Y. Wu, J. Gong, and Z. Yang, "Cell Zooming for Cost-Efficient Green Cellular Networks," *IEEE Communications Magazine*, vol. 48, no. 11, pp. 74–79, 2010.

[8] G. Micallef, P. Mogensen, and H. Scheck, "Cell Size Breathing and Possibilities to Introduce Cell Sleep Mode ," *IEEE European Wireless Conference*, pp. 111–115, 2010.

[9] I. Ashra, F. Boccardi, and L. Ho, "Power Savings in Small Cell Deployments via Sleep Mode Techniques," *International Conference on Information Networking*, pp. 307–311, 2010.

[10] S. Ramnath, V. Kavitha, and E. Altman, "Open Loop Optimal Control of Base Station Activation for Green Networks," *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, pp. 161–166, 2011.

[11] P. Chunyi, L. Suk-Bok, L. Songwu, L. Haiyun, and L. Hewu, "Traffic-Driven Power Saving in Operational 3G Cellular Networks," *MobiCom'11 Proc. of the 17th annual international conference on Mobile Computing and Networking*, pp. 121–132, 2011.

[12] L. Saker, S. E. Elayoubi, and H. O. Scheck, "System Selection and Sleep Mode for Energy Saving in Cooperative 2G/3G Networks," *IEEE 70th Vehicular Technology Conference Fall (VTC 2009-Fall)*, pp. 1–5, 2009.

[13] "3GPP TR 36.814, Further Advancements for E-UTRA Physical Layer Aspects," Tech. Rep., Mar 2010. [Online]. Available: http://www.qtc.jp/3GPP/Specs/36814-900.pdf

[14] S. E. Elayoubi, L. Saker, and T. Chahed, "Optimal Control for Base Station Sleep Mode in Energy Efficient Radio access Networks," *IEEE INFOCOMM Proc.*, pp. 106–110, April.

[15] R. W. Wolff, *Stochastic Modelling and the Theory of Queues*. Prentice Hall, 1989.

[16] B. Hajek, "Extremal Splittings of Point Processes," *Mathematics of Operations Research*, vol. 10, no. 4, pp. 543–556, 1985. [Online]. Available: http://pubsonline.informs.org/doi/abs/10.1287/moor.10.4.543

[17] E. Altman, B. Goujal, and A. Hordijk, *Discrete-Event Control of Stochastic Networks Multimodularity and Regularity*. Springer, 2001.

[18] D. Sinha, V. Kavitha, and A. Karandikar, "Load dependent optimal on-off policies in cellular heterogeneous network," IIT Bombay, Tech. Rep., Jan 2014. [Online]. Available: www.ieor.iitb.ac.in/files/faculty/kavitha/OptActivation.pdf

APPENDIX A: REVIEW OF MULTI-MODULARITY

(This section has been reproduced from [10] for a brief summary of concepts in multi-modularity. More details can be found in [17])

Definition 1: A function $f : \{0,1\}^n \to R$ *is Multimodular if*

$$f_n(\mathbf{a}+\mathbf{v}) + f_n(\mathbf{a}+\mathbf{u}) \geq f_n(\mathbf{a}) + f_n(\mathbf{a}+\mathbf{u}+\mathbf{v})$$

for all $\mathbf{a} \in \{0,1\}^n$ *and for all* $\mathbf{u}, \mathbf{v} \in F$ *(the Multimodular base) with* $\mathbf{u} \neq \mathbf{v}$ *and such that* $\mathbf{a}+\mathbf{u},\ \mathbf{a}+\mathbf{v},\ \mathbf{a}+\mathbf{u}+\mathbf{v} \in \{0,1\}^n$.

The Multimodular base F *contains the vectors* $-e_1, s_2, s_3, \cdots, s_n, e_n$*, where*

$$-e_1 = (\text{-1 } 0\ 0\ 0\ 0 \cdots\ 0\ 0), \quad s_2 = (1\ \text{-1 } 0\ 0\ 0 \cdots\ 0\ 0)$$
$$s_3 = (0\ 1\ \text{-1 } 0\ 0 \cdots\ 0\ 0), \qquad s_4 = (0\ 0\ 1\ \text{-1 } 0 \cdots\ 0\ 0)$$
$$\vdots$$
$$s_N = (0\ 0\ 0\ 0\ 0 \cdots\ 1\ \text{-1}) \ \ \textit{and} \ \ e_N = (0\ 0\ 0\ 0\ 0 \cdots\ 0\ 1).$$

Definition 2: The bracket sequence $\boldsymbol{a}^*(\eta,\beta) := \{a_n(\eta,\beta)\}$ *with rate* $\eta \in [0,1)$ *and initial phase* $\beta \in [0,1)$ *is defined as*

$$a_n(\eta,\beta) = \lfloor n\eta + \beta \rfloor - \lfloor (n-1)\eta + \beta \rfloor$$

Theorem 5. *A bracket sequence* $\boldsymbol{a}(\eta,\beta)$ *for any* $\beta \in [0,1)$ *minimizes the cost*

$$\underset{N\to\infty}{lim\ sup} \tfrac{1}{N}\textstyle\sum_{n=1}^{N} f_n(a_1,\cdots,a_n)$$

over all the sequences that satisfy

$$\underset{N\to\infty}{lim\ inf} \tfrac{1}{N}\textstyle\sum_{n=1}^{N} a_n \geq \eta$$

where $\eta \in [0,1)$*, under the following assumptions:*
1) f_n *is multimodular* $\forall\ n$*.*
2) $f_n(a_1,\cdots,a_n) \geq f_{n-1}(a_2,\cdots,a_n)\ \forall\ n > 1$
3) $\forall$ *sequence*$\{a_n\}$ *,* $\exists$ *a sequence*$\{b_n\}$ $\forall\ n, m$ *with* $n > m$*, such that*

$$f_n(b_1,\cdots,b_{n-m},a_1,\cdots,a_m) = f_m(a_1,\cdots,a_m)$$

4) $\forall\ n$*, the functions*$f_n(a_1,\cdots,a_n)$ *are increasing in* $a_i \forall\ i$*.* □

APPENDIX B: OPTIMALITY OF BRACKET SEQUENCE

Theorem 6. *For $J \leq \bar{J}$, the function $f_n(\boldsymbol{a}_1^n) := W_{n-J}(\boldsymbol{a}, J)$ is multimodular for every n.*

Proof. All the sequences in this proof are n length vectors and also J is fixed. Hence, we use the shorthand notation **a** in place of $\mathbf{a}_1^n$ for all the vectors and $\theta_n(\mathbf{a})$ in place of $\theta_n(\mathbf{a}, J)$.
We need to prove

$$f_n(\mathbf{a}+\mathbf{v}) + f_n(\mathbf{a}+\mathbf{u}) \geq f_n(\mathbf{a}) + f_n(\mathbf{a}+\mathbf{u}+\mathbf{v}) \tag{7}$$

$\forall$ $\mathbf{a} \in \{0,1\}^n$ and $\forall$ $\mathbf{u}, \mathbf{v} \in F$ (the multimodular base) with $\mathbf{u} \neq \mathbf{v}$ and such that $\mathbf{a}+\mathbf{u}$, $\mathbf{a}+\mathbf{v}$, $\mathbf{a}+\mathbf{u}+\mathbf{v} \in \{0,1\}^n$.
Let us define $s_1 = -e_1$ and $s_{n+1} = e_n$.
Let $\mathbf{v} = s_j$ and $\mathbf{u} = s_l$ where $j, l \in \{1, n+1\}$ and $l \neq j$. Without loss of generality, we assume that $l > j$. Consider $j \in \{2, n\}$. Since we can only consider such **v** for which **a + v** $\in \{0,1\}^n$, the sequence **a** should have $a_{j-1} = 0$ and $a_j = 1$. Also, we will have

$$(a+v)_{j-1} = 1, (a+v)_j = 0,$$
$$a_i = (a+v)_i \quad \forall\ i \neq j, j-1.$$

Therefore, adding **v** to **a** implies that the PBS located at position $j-1$ which was ON is turned OFF and the PBS at position j is turned ON.

When $\mathbf{v} = s_1 = -e_1$, we would have

$$a_1 = 1,\ (a+v)_1 = 0 \text{ and } a_i = (a+v)_i \quad \forall\ i \neq 1.$$

Note that the PBS at position 0 is always switched ON. Thus, $a_0 = 0 \ \ \forall$ **a**.

Similarly, when $\mathbf{v} = s_{n+1} = e_n$, we would have

$$a_n = 0, (a+v)_n = 1 \text{ and } a_i = (a+v)_i \quad \forall\ i \neq n.$$

Thus, addition of $-e_1$ switches ON the first base station and addition of e_n switches OFF the last base station.
All the above will also hold when **v** is replaced by **u** and j by l.

Note that l cannot be equal to $j+1$ i.e. $l > j+1$ since $u = s_l$ implies $a_{l-1} = 0$ i.e. $a_j = 0$ if $l = j+1$. But, $v = s_j$ implies $a_j = 1$. Hence, we have an inconsistency if $l = j+1$.
Thus, we only need to consider $j \in [1, n]$, $l \in [2, n+1]$ $\forall$ $l > j+1$.

Let the closest active PBS on the left of the $(n-J)$th user and the closest active PBS on the right of $(n-J)$th user be at positions $K_L(\mathbf{a})$ and $K_R(\mathbf{a})$ respectively i.e.

$$K_L(\mathbf{a}) = \max_{0 \leq k \leq n-J} \{a_k = 0\}$$

$$K_R(\mathbf{a}) = \begin{cases} n+1 & \text{if } a_i = 1\ \forall\ n-J \leq i \leq n \\ \min\limits_{k \geq n-J} \{a_k = 0\} & \text{otherwise} \end{cases}$$

$K_R(\mathbf{a})$ is assigned value $n+1$ when none of the J PBSs to the right of the user are ON. This signifies that the user will be connected either to $K_L(\mathbf{a})$ or the MBS.

Let $B(\mathbf{a})$ represent the base station to which the $(n-J)$th user is connected. $B(\mathbf{a})$ is either $K_L(\mathbf{a})$ or $K_R(\mathbf{a})$ or the MBS. *Unless there is a change in the base station to which the user is connected, his average waiting time will not change.*

If $n \leq J$, we will have $n - J \leq 0$. By definition, $\theta_{n-J}(\mathbf{b}) = \theta^0$ for any activation vector **b**. Thus,

$$\theta_{n-J}(\mathbf{a}) = \theta_{n-J}(\mathbf{a}+\mathbf{u}) = \theta_{n-J}(\mathbf{a}+\mathbf{v}) = \theta_{n-J}(\mathbf{a}+\mathbf{u}+\mathbf{v}).$$
$$\therefore f_n(\mathbf{a}+\mathbf{v}) = f_n(\mathbf{a}+\mathbf{u}) = f_n(\mathbf{a}) = f_n(\mathbf{a}+\mathbf{u}+\mathbf{v}) = w(\theta^0)$$

and (7) is satisfied. Now, we focus on $n > J$. If

$$j-1 < l-1 < K_L(\mathbf{a}) \text{ or } K_R(\mathbf{a}) < j-1 < l-1 \text{ or}$$
$$j-1 < K_L(\mathbf{a}) \text{ and } (l-1) > K_R(\mathbf{a}),$$

then even after adding **u** or **v** or **u+v** to **a**, the nearest ON PBS to the $(n-J)$th user on both its sides remain unchanged. Therefore,

$$f_n(\mathbf{a}+\mathbf{v}) = f_n(\mathbf{a}+\mathbf{u}) = f_n(\mathbf{a}) = f_n(\mathbf{a}+\mathbf{u}+\mathbf{v}).$$

Thus, (7) is satisfied. Now, let us divide the rest of the possibilities into three scenarios -

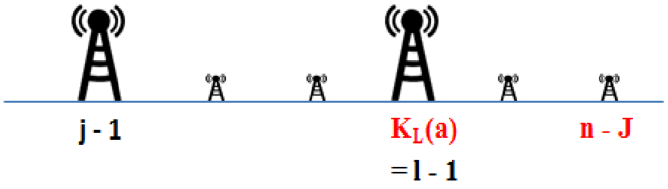
1) $j-1 < K_L(\mathbf{a}),\ l-1 = K_L(\mathbf{a})$ or $l-1 = K_R(\mathbf{a})$

2) $j-1 = K_L(\mathbf{a})$ or $j-1 = K_R(\mathbf{a}),\ l-1 > K_R(\mathbf{a})$

3) $j-1 = K_L(\mathbf{a}),\ l-1 = K_R(\mathbf{a})$

Considering each of them one-by-one,

Case 1: $j-1 < K_L(\mathbf{a}),\ l-1 = K_L(\mathbf{a})$ **or** $l-1 = K_R(\mathbf{a})$



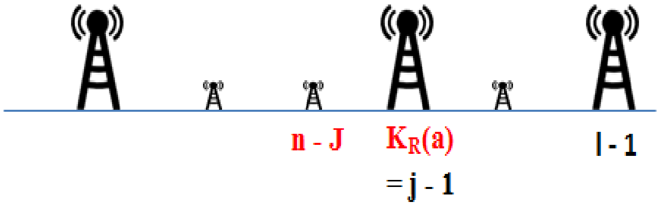
$\mathbf{j - 1 < K_L(a) , l - 1 = K_L(a)}$

As $j-1 < K_L(\mathbf{a})$ switching OFF the PBS at $j-1$ and switching ON the PBS at j or only switching on the PBS at position 1 (when $j=1$) will not affect the $(n-J)$th user. ($j \neq K_L(\mathbf{a})$ as $a_j = 1$ and $a_{K_L(\mathbf{a})} = 0$.) Thus, $B(\mathbf{a}) = B(\mathbf{a}+\mathbf{v})$. Hence,

$$f_n(\mathbf{a}+\mathbf{v}) = f_n(\mathbf{a}) \text{ and similarly, } f_n(\mathbf{a}+\mathbf{u}+\mathbf{v}) = f_n(\mathbf{a}+\mathbf{u}).$$

Therefore, (7) is satisfied for all possible values of **v** and **u** when $j-1 < K_L(\mathbf{a}),\ l-1 = K_L(\mathbf{a})$ or $l-1 = K_R(\mathbf{a})$.

Case 2: $j-1 = K_L(\mathbf{a})$ **or** $j-1 = K_R(\mathbf{a}),\ l-1 > K_R(\mathbf{a})$



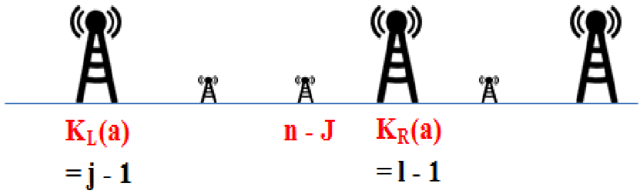
$\mathbf{j - 1 = K_R(a) , l - 1 > K_R(a)}$

Similar to the previous case, we will now have,

$$f_n(\mathbf{a}+\mathbf{u}) = f_n(\mathbf{a}) \text{ and } f_n(\mathbf{a}+\mathbf{v}+\mathbf{u}) = f_n(\mathbf{a}+\mathbf{v}).$$

Therefore, (7) is satisfied for all possible values of **v** and **u** when $j-1 < K_L(\mathbf{a}),\ l-1 = K_L(\mathbf{a})$.

Case 3: $j-1 = K_L(\mathbf{a}),\ l-1 = K_R(\mathbf{a})$



j – 1 = K_L(a) , l - 1 = K_R(a)

(The arguments given below are applicable $\forall\ j, l$)
As $l \leq n+1$, we have $K_R(\mathbf{a}) \leq n$.

$$\therefore K_R(\mathbf{a}) - (n-J) \leq n - (n-J) \tag{8}$$

i.e. the distance between the $(n-J)th$ user and $K_R(\mathbf{a})$ is less than or equal to J which implies $(n-J)$th user is connected to either the PBS at $K_L(\mathbf{a})$ or the PBS at $K_R(\mathbf{a})$. Thus, the user would not be connected to the MBS with this activation vector $\mathbf{a}$. As $K_R(\mathbf{a}) = K_R(\mathbf{a}+\mathbf{v})$, the same arguments hold when the activation vector is $\mathbf{a}+\mathbf{v}$. Thus, even with this activation vector , the user would not be connected to the MBS.

We know that $K_L(\mathbf{a}) \leq n-J$. If $K_L(\mathbf{a}) = n-J$ then by definition, $K_R(\mathbf{a}) = n-J$. Then, $l-1 = n-J = j-1$. This is an inconsistency as $l > j+1$.

Clearly, $K_L(\mathbf{a}+\mathbf{v}) = 1 + K_L(\mathbf{a})$. Thus, when $j-1 = K_L(\mathbf{a}) < n-J$, addition of $\mathbf{v}$ switches ON a PBS closer to the user. Therefore, the transfer rate of the user cannot decrease. Thus,

$$\theta_{n-J}(\mathbf{a}+\mathbf{v}) \geq \theta_{n-J}(\mathbf{a}).$$
$$\Rightarrow f_n(\mathbf{a}+\mathbf{v}) \leq f_n(\mathbf{a}).$$

(Waiting time monotonically decreases with the rate θ.) Using the same arguments we get,

$$f_n(\mathbf{a}+\mathbf{u}+\mathbf{v}) \leq f_n(\mathbf{a}+\mathbf{u}). \tag{9}$$

Now, consider the following sub-cases -

1) $K_R(\mathbf{a}) - (n-J) < (n-J) - K_L(\mathbf{a})$
Here, $B(\mathbf{a}) = K_R(\mathbf{a})$. Also, $K_L(\mathbf{a}+\mathbf{v}) = 1 + K_L(\mathbf{a})$ and $K_R(\mathbf{a}+\mathbf{v}) = K_R(\mathbf{a})$. Thus,

$$K_R(\mathbf{a}+\mathbf{v}) - (n-J) \leq (n-J) - K_L(\mathbf{a}+\mathbf{v}).$$
$$\Rightarrow B(\mathbf{a}+\mathbf{v}) = K_R(\mathbf{a}+\mathbf{v}) = K_R(\mathbf{a}) = B(\mathbf{a}).$$
$$\therefore f_n(\mathbf{a}+\mathbf{v}) = f_n(\mathbf{a}). \tag{10}$$

Adding (9) and (10), we get
$f_n(\mathbf{a}+\mathbf{v}) + f_n(\mathbf{a}+\mathbf{u}) \geq f_n(\mathbf{a}) + f_n(\mathbf{a}+\mathbf{u}+\mathbf{v})$.

2) $K_R(\mathbf{a}) - (n-J) \geq (n-J) - K_L(\mathbf{a})$
Here, $B(\mathbf{a}) = K_L(\mathbf{a})$. Clearly, $K_L(\mathbf{a}) = K_L(\mathbf{a}+\mathbf{u})$ and $K_R(\mathbf{a}+\mathbf{u}) \geq K_R(\mathbf{a})$. So we have,

$$K_R(\mathbf{a+u}) - (n-J) \geq (n-J) - K_L(\mathbf{a+u}).$$
$$\Rightarrow B(\mathbf{a}+\mathbf{u}) = K_L(\mathbf{a}+\mathbf{u}) = K_L(\mathbf{a}) = B(\mathbf{a}).$$
$$\therefore f_n(\mathbf{a}+\mathbf{u}) = f_n(\mathbf{a}). \tag{11}$$

Now, let us consider the situation when the activation vector is $\mathbf{a}+\mathbf{v}$. As $K_R(\mathbf{a}) = K_R(\mathbf{a}+\mathbf{v})$, from (8), we have

$$K_R(\mathbf{a}+\mathbf{v}) - (n-J) \leq n - (n-J).$$

Thus, even when the activation vector is $\mathbf{a}+\mathbf{v}$ the user would not be connected to the MBS but would be connected to either $K_R(\mathbf{a}+\mathbf{v})$ or $K_L(\mathbf{a}+\mathbf{v})$.
Using the hypothesis,

$$K_R(\mathbf{a}+\mathbf{v})-(n-J)\geq(n-J)-K_L(\mathbf{a}+\mathbf{v}).$$

Hence, the user is connected to $K_L(\mathbf{a}+\mathbf{v})$. (In the case of equality in the above, user can be connected either to $K_L(\mathbf{a}+\mathbf{v})$ or $K_R(\mathbf{a}+\mathbf{v})$ but this does not make a difference to $f_n(\mathbf{a}+\mathbf{v})$ as both of them are equidistant from the user.) Therefore,

$$f_n(\mathbf{a}+\mathbf{v}+\mathbf{u})=f_n(\mathbf{a}+\mathbf{v}) \tag{12}$$

Adding (11) and (12), we get
$f_n(\mathbf{a}+\mathbf{v})+f_n(\mathbf{a}+\mathbf{u})=f_n(\mathbf{a})+f_n(\mathbf{a}+\mathbf{u}+\mathbf{v})$.

Thus, for $K_L(\mathbf{a})=j-1$ and $K_R(\mathbf{a})=l-1$, (7) is satisfied. Therefore, for all $\mathbf{u},\mathbf{v}\in F$ and $\mathbf{u}\neq\mathbf{v}$, we have proved that

$$f_n(\mathbf{a}+\mathbf{v})+f_n(\mathbf{a}+\mathbf{u})\geq f_n(\mathbf{a})+f_n(\mathbf{a}+\mathbf{u}+\mathbf{v}).$$

Hence, we conclude that $f_n(\mathbf{a})=W_{n-J}(\mathbf{a})$ is multimodular. □

Proof of Theorem 1. As J remains constant throughout the proof, we will use a shorthand notation of $W_n(\mathbf{a}),\theta_n(\mathbf{a})$ instead of $W_n(\mathbf{a},J),\theta_n(\mathbf{a},J)$ respectively. Define

$$f_n(\mathbf{a}_1^n)=W_{n-J}(\mathbf{a},J).$$

This proof is obtained using Theorem 5. We will verify the validity of its assumptions.
1) $f_n(\mathbf{a}_1^n)=W_{n-J}(\mathbf{a})$ is multimodular from Theorem 6.

2) For assumption 2, we need to prove $\forall\ n>1$ that

$$f_n(a_1,\cdots,a_n)\geq f_{n-1}(a_2,\cdots,a_n).$$

It is sufficient to show

$$W_{n-J}(a_1,\cdots,a_n,\cdots)\geq W_{(n-1)-J}(a_2,\cdots,a_n,\cdots).$$
$$\text{(or) }\theta_{n-J}(a_1,\cdots,a_n,\cdots)\leq\theta_{(n-1)-J}(a_2,\cdots,a_n,\cdots).$$

If $(n-J)\leq 0$, then the assumption is true because

$$\theta_{n-J}(a_1,\cdots,a_n,\cdots)=\theta_{(n-1)-J}(a_2,\cdots,a_n,\cdots)=\theta^0.$$

Now, let us consider $(n-J)>0$. Define activation vector

$$\mathbf{c}=(c_1,c_2,\cdots,c_{n-1}):=(a_2,\cdots,a_n).$$

By our assumption PBS at 0 is ON, or equivalently $c_0=a_0=0$. Recall that θ_{n-J} is the rate at $n-J$ point while $\theta_{(n-1)-J}$ is the rate at its left neighbor. With the changes in activation vectors, we have c_0 where a_1 was originally present. If $a_1=0$, then no user would have been connected to PBS at position 0. On removing a_1, we have c_0 where a_1 was originally present. Thus, there will be no change in user's transfer rate. If $a_1=1$, then by replacing it with c_0, we are switching ON a PBS which was previously OFF. This cannot result in a decrease in the transfer rate of any user. Therefore, $\forall\ n>1$

$$\theta_{n-J}(a_1,\cdots,a_n,\cdots)\leq\theta_{(n-1)-J}(a_2,\cdots,a_n,\cdots).$$

3) We have $a_0=0$. As on adding $\mathbf{b}$, b_{n-m} takes the position of a_0, we take $\mathbf{b}$ such that $b_{n-m}=0$. With such a choice clearly,

$$f_n(b_1,\cdots,b_{n-m},a_1,\cdots,a_m)=f_m(a_1,\cdots,a_m).$$

4) Switching ON a PBS cannot decrease a user's transfer rate i.e. $\theta_{n-J}(a_1, \cdots, a_{i-1}, 0, a_{i+1}, \cdots, a_n, \cdot) \geq \theta_{n-J}(a_1, \cdots, a_{i-1}, 1, a_{i+1}, \cdots, a_n, \cdot) \quad \forall\ i.$
As $f_n(\mathbf{a}_1^n) = W_{n-J}(\mathbf{a})$ is a decreasing function of $\theta_{n-J}(\mathbf{a})$, $W_{n-J}(\mathbf{a})$ will be increasing in $a_i \quad \forall\ i$. Thus, all the assumptions of Theorem 5 hold . Now,

$$\begin{aligned}\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} f_n(a_1, \cdots, a_n) &= \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} W_{n-J}(\mathbf{a}, J)\\ &= \limsup_{N\to\infty}\frac{1}{N}\left(\sum_{n=1-J}^{0} W_0(\mathbf{a}, J) + \sum_{n=1}^{N-J} W_n(\mathbf{a}, J)\right)\\ &= \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N} W_n(\mathbf{a}, J).\end{aligned}$$

(For $n - J \leq 0, W_{n-J}(\mathbf{a}) = w(\theta^0)$.)
Thus, using Theorem 5, the optimization problem (3) has the solution as the bracket policy sequence $\mathbf{a}^*$. □

Proof of Theorem 2. When the J-association policy is being used, the user being served in the nth queue can be associated to 0th, 1st, $\cdots$ Jth nearest PBS or the MBS. (In the following discussion, by 'user', we mean the user being served in the nth queue).
Thus, we have the following cases-
1) User is being served by a PBS:
Let the distance of this PBS from the user be l i.e.

$$l(\mathbf{a}, J) := \min_{0\leq k\leq J} \{a_{n+k} = 0 \text{ or } a_{n-k} = 0\}.$$

Clearly, $l(\mathbf{a}, J) = l(\mathbf{a}, J+1)$. Therefore,

$$\theta_n(\mathbf{a}, J+1) = \theta_n(\mathbf{a}, J).$$

2) User is being served by the the MBS i.e. $\theta_n(\mathbf{a}, J) = \theta^M$:

$$\begin{aligned}\theta_n(\mathbf{a}, J+1) = \theta^M &\mathbf{1}\{a_{n+J+1} = a_{n-J-1} = 1\}\\ &+ \theta^{J+1}\mathbf{1}\{a_{n+J+1}a_{n-J-1} = 0\}\end{aligned}$$

where $\mathbf{1}(.)$ represents the indicator function. Thus,

$$\begin{aligned}&\theta_n(\mathbf{a}, J+1) \geq \theta_n(\mathbf{a}, J) \text{ if } J < \bar{J} \text{ and}\\ &\theta_n(\mathbf{a}, J+1) \leq \theta_n(\mathbf{a}, J) \text{ if } J \geq \bar{J}.\end{aligned}$$

From the two cases above,

$$\begin{aligned}\theta_n(\mathbf{a}, 1) \leq \theta_n(\mathbf{a}, 2) \leq \cdots \theta_n(\mathbf{a}, \bar{J}) \text{ and}\\ \theta_n(\mathbf{a}, \bar{J}) \geq \theta_n(\mathbf{a}, \bar{J}+1) \geq \theta_n(\mathbf{a}, \bar{J}+2) \geq \cdots .\end{aligned}$$

$$\therefore \theta_n(\mathbf{a}, J) \leq \theta_n(\mathbf{a}, \bar{J}) \quad \forall\ \mathbf{a},\ J.$$

But $W_n(\mathbf{a}, J)$ is a monotonically decreasing function of $\theta_n(\mathbf{a}, J)$. Thus,

$$W_n(\mathbf{a}, J) \geq W_n(\mathbf{a}, \bar{J}) \quad \forall\ \mathbf{a},\ J.$$

Averaging over all n's, we get $\overline{W}(\mathbf{a}, J) \geq \overline{W}(\mathbf{a}, \bar{J})\ \forall\ \mathbf{a},\ J.$

From Theorem 1, $\overline{W}(\mathbf{a}, \bar{J}) \geq \overline{W}(\mathbf{a}^*, \bar{J})$ for all $\mathbf{a}$ in which the fraction of base stations switched OFF is equal to η. Thus, we have, $\overline{W}(\mathbf{a}, J) \geq \overline{W}(\mathbf{a}^*, \bar{J}) \quad \forall\ J$ and for all $\mathbf{a}$ satisfying the above condition. □

APPENDIX C: PROOFS RELATED TO WAITING TIME ANALYSIS

Proof of Lemma 1. Let us check if $s = \lceil p\frac{k_2}{k_1} \rceil - 1$ satisfies the equality. We have,

$$\left(p\frac{k_2}{k_1} - 1\right)\frac{k_1}{k_2} \leq \left(\left\lceil p\frac{k_2}{k_1}\right\rceil - 1\right)\frac{k_1}{k_2} < \left(p\frac{k_2}{k_1}\right)\frac{k_1}{k_2}.$$

$$\therefore p - 1 \leq \left(\left\lceil p\frac{k_2}{k_1}\right\rceil - 1\right)\frac{k_1}{k_2} < p.$$

Hence, $\left\lfloor\left(\left\lceil p\frac{k_2}{k_1}\right\rceil - 1\right)\frac{k_1}{k_2}\right\rfloor = p-1$. As $\left\lfloor s\frac{k_1}{k_2}\right\rfloor$ is non-decreasing in s, it is less than p for all $s \leq \left\lceil p\frac{k_2}{k_1}\right\rceil - 1$. Now, consider $s = \left\lceil p\frac{k_2}{k_1}\right\rceil$. We have,

$$p\frac{k_2}{k_1}\frac{k_1}{k_2} \leq \left\lceil p\frac{k_2}{k_1}\right\rceil\frac{k_1}{k_2} < \left(1 + p\frac{k_2}{k_1}\right)\frac{k_1}{k_2}.$$

$$\therefore p \leq \left\lceil p\frac{k_2}{k_1}\right\rceil\frac{k_1}{k_2} < p + \frac{k_1}{k_2}.$$

Hence, $\left\lfloor\left\lceil p\frac{k_2}{k_1}\right\rceil\frac{k_1}{k_2}\right\rfloor = p.$ □

Proof of Lemma 3. We want to find an expression for $l(\eta)$ in terms of η. Towards this, we proceed by proving the following:

1) $\{\eta : l(\eta) = r + 1\} \subset \{\eta \leq h(r)\}$
2) $\{\eta \leq h(r)\} \subset \{\eta : l(\eta) \leq r + 1\}$
3) $\{h(r-1) < \eta \leq h(r)\} = \{\eta : l(\eta) = r + 1\}$

Step 1: $\{\eta : l(\eta) = r + 1\} \subset \{\eta \leq h(r)\}$

We first determine the permissible η for $l(\eta) = i \;\; \forall\, i$. Define

$$d_n(1) = \inf_{j>0}\left\{a^*_{n+j} = 1\right\} \text{ and}$$
$$d_n(k) = \inf_{j>d_n(k-1)}\left\{a^*_{n+j} = 1\right\}.$$

Note that, $d_n(k)$ represents the distance of the kth next OFF PBS from the nth PBS. Clearly, $d_n(k) \geq k$.

Let us consider the various possible values of $l(\eta)$.

1) $l(\eta) = 1$ means all users are being served at the rate θ^0. This is possible only when all the PBSs are ON. This can happen only when $\eta = 0$.

2) $l(\eta) = r + 1$ - This means that the rates of each queue in the k_2 block is one of $\theta^0, \theta^1, \cdots, \theta^r$ and no queue is served at rate θ^{r+1} or lesser. This will happen when the number of consecutive PBSs which are OFF is at most $2r$. This means for any OFF PBS at position $n, d_n(2r) > 2r$. In particular, this is true for $n = n_0$ (recall that $n_0 = mk_2$ and $a_{n_0} = 1$ i.e. the PBS located at n_0 is OFF). Thus, from Lemma 2 $\forall\; p > 2r$

$$\left(n_0 + \left\lceil p\frac{k_2}{k_1}\right\rceil\right) - \left(n_0 + \left\lceil (p-2r)\frac{k_2}{k_1}\right\rceil\right) > 2r.$$

$$\therefore \left\lceil\frac{p}{\eta}\right\rceil - \left\lceil\frac{p-2r}{\eta}\right\rceil > 2r. \tag{13}$$

If possible let $\eta > h(r)$. Then,

$$-(2r+1) < -\frac{2}{\eta}r < -2r.$$

$$\therefore \left\lceil -\frac{2}{\eta}r \right\rceil = -2r.$$

Let us take $p = nk_1$ for some integer n. Then,

$$\left\lceil \frac{p}{\eta} \right\rceil - \left\lceil \frac{p-2r}{\eta} \right\rceil = nk_2 - (nk_2 + \left\lceil -\frac{2r}{\eta} \right\rceil) = 2r.$$

Thus, we have found a value of integer p for which equation (13) is not satisfied. Therefore, $l(\eta)$ cannot be $r+1$ for $\eta > h(r)$. Hence,

$$l(\eta) = r+1 \implies \eta \leq h(r). \tag{14}$$

Step 2: $\{\eta \leq h(r)\} \subset \{\eta : l(\eta) \leq r+1\}$

Now, let us consider the case when $\eta \leq \dfrac{2r}{1+2r}$ and check if it is possible to have $l(\eta) > r+1$. Assume, $l(\eta) \geq r+2$. This implies that there exists two ON PBSs such that there are atleast $2r+1$ consecutive OFF PBSs between them. Thus, for some q,

$$a_q = 0, a_{q+1} = 1, a_{q+2} = 1, \cdots a_{q+2r+1} = 1.$$

From Lemma 2, location of each OFF PBS can be written in the form $n_0 + \left\lceil \dfrac{z}{\eta} \right\rceil$ for some integer z. Thus, there exists a z such that,

$$q+1 = n_0 + \left\lceil \frac{z}{\eta} \right\rceil,\ q+2 = n_0 + \left\lceil \frac{z+1}{\eta} \right\rceil, \cdots,$$

$$q+2r+1 = n_0 + \left\lceil \frac{z+2r}{\eta} \right\rceil.$$

Using the first and last equation, we get

$$\left\lceil \frac{z}{\eta} \right\rceil + 2r = \left\lceil \frac{z+2r}{\eta} \right\rceil.$$

As ceiling of a number is strictly less than one more than the number,

$$\left\lceil \frac{z+2r}{\eta} \right\rceil < \frac{z}{\eta} + 2r + 1.$$

$$\Rightarrow \frac{z+2r}{\eta} < \frac{z}{\eta} + 2r + 1.$$

$$\Rightarrow \frac{2r}{\eta} < 2r + 1.$$

$$\Rightarrow \eta > \frac{2r}{1+2r}.$$

This is a contradiction. Thus, $l(\eta) \leq r+1$ for $\eta < h(r)$.

Step 3: $\{h(r-1) < \eta \leq h(r)\} = \{\eta : l(\eta) = r+1\}$
Now consider $h(r-1) < \eta \leq h(r)$.
We know that $l(\eta) \leq r+1$. We want to find out the exact value of $l(\eta)$.
Assume $l(\eta) \leq r = (r-1)+1$. But from equation (14), $l(\eta) \leq (r-1)+1$ implies $\eta \leq h(r-1)$. This is a contradiction.
Thus, for $h(r-1) < \eta \leq h(r)$, the only possible value of $l(\eta)$ is $r+1$. □

Lemma 4. *If* $\lfloor q\eta \rfloor = \lfloor (q-1)\eta \rfloor = s$ *and* $\lfloor (q+p)\eta \rfloor = \lfloor (q+p-1)\eta \rfloor = s+o$*, then* $o \leq p-1$.

Proof.

$$\begin{aligned}
(q+p-1)\eta &= q\eta + (p-1)\eta. \\
&< q\eta + (p-1). \\
\therefore \lfloor (q+p-1)\eta \rfloor &\leq \lfloor q\eta + p - 1 \rfloor. \\
&= \lfloor q\eta \rfloor + p - 1. \\
\therefore s + o &\leq s + (p-1). \\
o &\leq p-1.
\end{aligned}$$

□

Lemma 5. $\overline{W}^*(\eta)$ *is a non-decreasing function of* η.

Proof. Consider $h(r-1) < \eta \leq h(r)$ for some r. We examine the derivative of η in this interval.
Let us start with the case when $r-1 < \bar{J}$.

$$\begin{aligned}
\frac{d\overline{W}^*(\eta)}{d\eta} &= \sum_{k=0}^{r} w(\theta^k) b_{r,k} + w(\theta^{r-1}(1+2(r-1))) \\
&= -w(\theta^0) - 2w(\theta^1) - 2w(\theta^2) \cdots \\
&2w(\theta^{r-1}) + (2r-1)w(\theta^r) \\
&\geq -w(\theta^0) - 2w(\theta^r) - 2w(\theta^r) \cdots \\
&2w(\theta^r) + (2r-1)w(\theta^r) \\
&= -(2(r-1)+1)w(\theta^r) + (2r-1)w(\theta^r) \\
&= 0.
\end{aligned}$$

Thus, the derivative is non-negative. Hence, $\overline{W}^*(\eta)$ is non-decreasing function of η in the interval $h(r-1) < \eta \leq h(r), \forall\ r$. As $\overline{W}^*(\eta)$ is continuous at the interval boundaries i.e. at $\eta = h(r)$, we conclude that $\overline{W}^*(\eta)$ is non-decreasing function of $\eta, \eta \in [0,1]$. With similar arguments, even when $r-1 \geq \bar{J}$, we can show that $\overline{W}^*(\eta)$ has a non-negative derivative. □

Proof of Theorem 3. Consider two consecutive ON PBSs. Let their positions be q and $q+p$, where p is the distance between the two PBSs. Thus, we have, $a_q = a_{q+p} = 0$.
Let $h(r-1) < \eta \leq h(r)$ for some r. Thus, $l(\eta) = r+1$.
Let $\lfloor q\eta \rfloor = \lfloor (q-1)\eta \rfloor = s$ and $\lfloor (q+p)\eta \rfloor = \lfloor (q+p-1)\eta \rfloor = s+o$. Using Lemma 4, we get $o \leq p-1$.
We have $s \leq q\eta - \eta$ and $(q+p)\eta < s+o+1$.

$$\begin{aligned}
&\Rightarrow \quad (q+p)\eta < q\eta - \eta + o + 1. \\
&\Rightarrow \quad \eta < \frac{o+1}{1+p}. \\
&\Rightarrow \quad \eta < \frac{p}{1+p} \text{ (from Lemma 4).}
\end{aligned}$$

It is easy to see that $\frac{i}{i+1}$ is an increasing function of i. We also know that $h(r-1) = \frac{2(r-1)}{1+2(r-1)} < \eta$. Thus, $p > 2(r-1)$ i.e. $p \geq 2r-1$. Thus, the minimum distance between two consecutive ON PBSs is $2r-1$. This means that there are atleast $2r-2$ consecutive OFF PBSs between any two ON PBSs. Thus, if we define S_i to be the set containing the ith ON PBS and its $r-1$ neighbors on each side, then the sets S_is will be disjoint. Further, each set will contain one queue being served at the rate θ^0 (the queue at the ON PBS) and two queues being served at the rate $\theta^i, 1 \leq i \leq r-1$. Thus, number of θ^i queues for $1 \leq i \leq r-1$ is twice the number of θ^0 queues i.e. twice the number of ON PBSs.
In the activation vector $\mathbf{a}^*$, the fraction of OFF PBSs is exactly equal to η and hence in a block of k_2 PBSs, the number of OFF PBSs will be $\eta \times k_2 = k_1$. Hence, the number of ON PBSs = $k_2 - k_1$. Thus if $r-1 < \bar{J}$, number of θ^i queues will be $2(k_2-k_1)$ for $1 \leq i \leq r-1$. Hence, number of θ^r queues is $k_2 - (1+2(r-1))(k_2-k_1) = k_1(2r-1) - 2k_2(r-1)$.
Therefore,

$$\begin{aligned}\overline{W}^*(\eta) =& \frac{1}{k_2}\left((k_2-k_1)\,w(\theta^0) + 2\,(k_2-k_1)\,w(\theta^1) + \right.\\ & 2\,(k_2-k_1)\,w(\theta^2) + \cdots + 2\,(k_2-k_1)\,w(\theta^{r-1}) + \\ & \left. w(\theta^r)\,(k_1(2r-1) - 2k_2(r-1))\right)\\ =& \,(1-\eta)\,w(\theta^0) + 2\,(1-\eta)\,w(\theta^1) + \cdots + \\ & 2\,(1-\eta)\,w(\theta^{r-1}) + w(\theta^r)\,(\eta(2r-1) - 2(r-1))\,.\end{aligned}$$

Similarly, when $r-1 \geq \bar{J}$, then number of θ^i queues will be $2(k_2-k_1)$ for $1 \leq i \leq \bar{J}$. Rest of the queues i.e. $k_1(2\bar{J}+1) - 2k_2\bar{J}$ queues will be connected to the MBS. Combining these two cases, we get equation (6) as the expression for minimum waiting time when η fraction of the PBSs have to be switched OFF. □

Proof of Theorem 4. If possible, assume there exist $\tilde{\eta}$ and policies $(\mathbf{a}, J)$ such that

$$\tilde{\eta} = \liminf_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} a_n \text{ and satisfying } \overline{W}(\mathbf{a}, J) \leq \overline{W}_{QoS},$$

and such that $\tilde{\eta} \geq \eta'$. Let $\mathbf{a}^*(\tilde{\eta})$ be used to represent the bracket sequence with switch OFF ratio $\tilde{\eta}$. By optimality of policies $(\mathbf{a}^*(\tilde{\eta}),\ \bar{J})$, we have

$$\overline{W}^*(\tilde{\eta}) = \overline{W}(\mathbf{a}^*(\tilde{\eta}), \bar{J}) \leq \overline{W}(\mathbf{a}, J) \leq \overline{W}_{QoS}.$$

On the other hand, by monotonicity of $\overline{W}^*(.)$ (see [18]) we have, $\overline{W}^*(\tilde{\eta}) > \overline{W}^*(\eta') = \overline{W}_{QoS}$. This is a contradiction. Thus, η' is the optimal value in (4) and $\mathbf{a}' := \mathbf{a}^*(\eta')$ and $J = \bar{J}$ are the optimal pair of policies solving the optimization problem (4). □

Lemma 6. *Waiting time is a continuous function of η.*

Proof. We know that if $h(r-1) < \eta \leq h(r)$, then
1)If $r-1 < \bar{J}$,

$$\begin{aligned}\overline{W}^*(\eta) = \quad & (1-\eta)\,w(\theta^0) + 2\,(1-\eta)\,w(\theta^1) + \cdots + \\ & 2\,(1-\eta)\,w(\theta^J) + w(\theta^r)\,(\eta(2r-1) - 2(r-1))\,.\end{aligned} \tag{15}$$

2) If $r-1 \geq \bar{J}$,

$$\begin{aligned}\overline{W}^*(\eta) = \quad & (1-\eta)\,w(\theta^0) + 2\,(1-\eta)\,w(\theta^1) + \cdots + \\ & 2\,(1-\eta)\,w(\theta^{r-1}) + w(\theta^M)\left(\eta(2\bar{J}+1) - 2\bar{J}\right).\end{aligned} \tag{16}$$

We need to check if the waiting time is continuous at the boundaries i.e. for $\eta = h(r)$. For this value of η, we will find the right and left limits $(\overline{W}_r, \overline{W}_l)$. $\overline{W}_l$ will be evaluated based on the expression of $\overline{W}$

in the region $h(r-1) < \eta \leq h(r)$ whereas $\overline{W}_r$ will be evaluated based on the expression of $\overline{W}$ in the region $h(r) < \eta \leq h(r+1)$.
We can have 3 cases depending on relation between r and $\bar{J}$:

1) $r < \bar{J}$
In this case, we will use equation (15) for evaluating both $\overline{W_l}$ and $\overline{W_r}$ (because $r-1 < \bar{J}$ and $r < \bar{J}$).

$$\begin{aligned} \overline{W}_l &= (1-\eta)\, w(\theta^0) + 2\,(1-\eta)\, w(\theta^1) + \cdots + \\ &\quad 2\,(1-\eta)\, w(\theta^{r-1}) + w(\theta^r)\,(\eta(2r-1) - 2(r-1)) \\ &= X + w(\theta^r)\,(\eta(2r-1) - 2(r-1)) \\ &= X + \frac{2}{1+2r} w(\theta^r) \end{aligned}$$

where, $X = (1-\eta)\, w(\theta^0) + 2\,(1-\eta)\, w(\theta^1) + \cdots + 2\,(1-\eta)\, w(\theta^{r-1})$.

$$\begin{aligned} \overline{W}_r &= (1-\eta)\, w(\theta^0) + 2\,(1-\eta)\, w(\theta^1) + \cdots + \\ &\quad 2\,(1-\eta)\, w(\theta^{r-1}) + w(\theta^r)\,(\eta(2r-1) - 2(r-1)) + \\ &\quad w(\theta^{r+1})\,(\eta(2(r+1)-1) - 2(r)) \\ &= X + w(\theta^r)\,(\eta(2r-1) - 2(r-1)) + \\ &\quad w(\theta^{r+1})\,(\eta(2(r+1)-1) - 2(r)) \\ &= X + \frac{2}{1+2r} w(\theta^r) + w(\theta^{r+1}) \times 0 \\ &= \overline{W}_l. \end{aligned}$$

Thus, in this case $\overline{W}^*(\eta)$ is continuous at the boundaries.

2) $r = \bar{J}$
In this case, we will use (15) for evaluating $\overline{W_l}$ and (16) for evaluating $\overline{W_r}$ (because $r-1 < \bar{J}$ and $r \geq \bar{J}$). As derived previously,

$$\overline{W}_l = X + \frac{2}{1+2r} w(\theta^r).$$

$$\begin{aligned} \overline{W}_r &= (1-\eta)\, w(\theta^0) + 2\,(1-\eta)\, w(\theta^1) + \cdots + \\ &\quad 2\,(1-\eta)\, w(\theta^J) + w(\theta^M)\left(\eta(2\bar{J}+1) - 2\bar{J}\right) \end{aligned}$$

Using $r = \bar{J}$,

$$\begin{aligned} &= X + 2\,(1-\eta)\, w(\theta^r) + \\ &\quad w(\theta^M)\left(\eta(2\bar{J}+1) - 2\bar{J}\right) \\ &= X + \frac{2}{1+2r} w(\theta^r) + w(\theta^M) \times 0 \\ &= \overline{W}_l. \end{aligned}$$

Thus, in this case also $\overline{W}^*(\eta)$ is continuous at the boundaries.

3) $r > \bar{J}$
In this case, we will use (16) for evaluating both $\overline{W_l}$ and $\overline{W_r}$ (because $r-1 \geq \bar{J}$ and $r \geq \bar{J}$). From

(16),

$$\overline{W}^*(\eta) = (1-\eta)\, w(\theta^0) + 2\,(1-\eta)\, w(\theta^1) + \cdots + \\ 2\,(1-\eta)\, w(\theta^J) + w(\theta^M)\left(\eta(2\bar{J}+1) - 2\bar{J}\right).$$

Clearly, this expression does not depend on r. Thus value of $\overline{W}^*(\eta)$ will be same using the expressions for $h(r-1) < \eta \leq h(r)$ and $h(r) < \eta \leq h(r+1)$. Hence, it will be continuous at the boundary i.e. when $\eta = h(r)$ for any integer r.

Therefore, in all three cases, we obtain continuity of waiting time at the boundaries. The continuity of waiting times at points other than the boundaries can be easily seen from the two expressions of $\overline{W}^*(\eta)$. Thus, the waiting time is continuous for all η. □