The Impact of Non-Neutrality on the Quality of Service and Energy Efficiency of a Separation Architecture

A. Arotiba[1], A. Fisusi[1], T. Yesufu[1], A. A. Olawole[1]
[1]Department of Electronic and Electrical Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.

ABSTRACT

A separation architecture is a heterogeneous network specifically designed to reduce the energy consumption of cellular networks. While Quality of Service (QoS) and energy efficiency of this architecture have been studied extensively, the impact of non-neutrality practices like paid prioritisation on these critical network metrics has not previously been investigated. In this paper, we address this gap by presenting a study that evaluated the QoS and energy efficiency of a non-neutral, energy-efficient scenario in a separation architecture. In the study, a novel non-neutral resource allocation scheme based on the concept of paid prioritisation, which involves assigning resources to network users in accordance with their classes of subscription fees, was developed. This scheme was then combined with a topology management scheme that reduces energy consumption of the architecture by turning base stations off and back on based on traffic loads, idle waiting periods, and blocked requests. The combination of these schemes created the non-neutral, energy-efficient scenario evaluated. System-level simulations were carried out to compare the performance of two versions of the non-neutral scenario with a conventional net-neutral scenario in terms of QoS and energy efficiency. Simulation results showed that although non-neutrality led to lower average file transfer delay relative to the net-neutral paradigm, it worsened blocking probability and throughput in the system. Furthermore, non-neutrality achieved similar energy efficiency performance as the net-neutral approach at low traffic loads, reaching up to 67.5% Energy Reduction Gain; however, at high traffic loads, it led to significantly poorer energy efficiency.

KEYWORDS

Heterogeneous Networks
Resource Allocation
Paid Prioritisation
Energy Reduction Gain
Delay
Throughput

1. INTRODUCTION

Separation Architectures have been proposed as a candidate solution to the critical need for cellular networks to be operated in an energy-efficient manner (Xu, He, Zhang, Chen, & Xu, 2013). They are Heterogeneous Networks (HetNets) with their control and data planes separated, enabling high-power control macrocell base stations to handle low data rate applications, coverage and control signalling, while low-power data small cell base stations handle mainly high data rate user traffic (Capone, Filippini, Gloss, & Barth, 2012). This design makes it possible for idle small cell base stations to be switched off to reduce energy consumption without compromising coverage or significantly deteriorating Quality of Service (QoS), especially when the network experiences low traffic intensity. Conventionally, resource allocation in these networks has been driven by technical goals, such as good QoS and energy efficiency, rather than economic or other goals that deliberately violate Network Neutrality. Network Neutrality (or Net Neutrality) has been described as a principle that requires the treatment of packets over the internet equally, regardless of the source, destination, type or content (Choi, Jeon, & Kim, 2018; Maillé & Tuffin, 2019). In contrast, Non-Neutrality involves practices like paid prioritisation, blocking or throttling of specific applications and zero-rating (Gharakheili, Vishwanath, & Sivaraman, 2016). These practices allow QoS and energy efficiency to be influenced not only by technical factors, but also by discriminatory economic and operator strategies.

Although several studies have been carried out on QoS and energy efficiency in separation architectures, the impact of non-neutrality on both has not been investigated to the best of our knowledge. The debate about whether to enforce net neutrality or not has been going on for several years all over the world (Maillé & Tuffin, 2019, 2022). It has been suggested that paid prioritisation, a form of non-neutrality, can lead to the creation of fast lanes and slow lanes on the internet (Economides, 2017). While all forms of non-neutrality have policy implications, it is paid prioritisation that most significantly changes the technical logic of a cellular network's real-time resource allocation scheduler. Unlike zero-rating or blocking, paid prioritisation requires the scheduler to dynamically allocate resources based on economic contracts rather than purely on technical efficiency. Furthermore, this sophisticated policy requires a centralised controller with a global view of user subscription tiers and network conditions in multiple small cell base stations, a mechanism absent in conventional, decentralised networks. The control macrocell base station of the separation architecture precisely provides this capability. Therefore, the research gap this paper addresses is the impact of paid prioritisation (a form of non-neutrality) on the QoS and energy efficiency of a separation architecture, the cellular network type that is particularly suited to its practical implementation.

Cellular networks have become a major means of accessing the internet worldwide. In conventional, net-neutral frameworks of cellular networks, the resource allocation scheduler already manages different QoS classes to enable accurate and timely delivery of associated user or application data. It allocates resources — such as time slots and frequency bands — to meet the QoS requirements of each application, sometimes prioritising real-time applications over non-real-time services. However, though the network strives to provide a good experience for all, it may not be able to guarantee that service quality will not degrade due to systemic factors like high network congestion or inter-cell interference.

A non-neutral regime based on paid prioritisation introduces an extra layer of protection of price-based service guarantee over and above the QoS-based guarantee of conventional net-neutral networks. A user paying a premium fee is not just allocated a superior channel at the start of their session; the network is contractually obligated to actively maintain that QoS throughout the user's transmission. This imposes a new, dynamic, and overriding constraint on the resource allocation scheduler. The system must now proactively deploy mechanisms to continuously shield premium users from the effects of congestion and interference to uphold this ongoing guarantee. It is this shift from a QoS-based guarantee that balances the needs of all users against fluctuating system-wide conditions, to one that must also enforce price-driven service guarantees for a select group of users that creates complex trade-offs for overall network QoS performance and energy efficiency. Understanding how the QoS and energy efficiency performance of this peculiar non-neutral resource allocation paradigm is different from the conventional net-neutral case in a separation architecture motivated this study.

The main contributions of this paper are threefold. First, we

Corresponding Author: A. Fisusi (bimbofisusi@oauife.edu.ng)
The editor responsible for coordinating the review of this article and approving its publication was E. Obayiuwana
P-ISSN: 1115-9782 **e-ISSN:**

propose the Paid Prioritisation-based Clustering Capability Rating (PPCCR) scheme, a novel resource allocation scheme that modifies an existing energy-aware scheme to explicitly incorporate the economic constraints of a non-neutral, paid prioritisation policy. Second, and most importantly, this work provides the first quantitative analysis of the three-way trade-off between non-neutrality, QoS, and energy efficiency within a separation architecture. While prior work has explored energy efficiency in these architectures, the impact of superimposing a non-neutral economic policy onto the resource allocation logic has, to our knowledge, not been studied. Finally, our work reveals new and important insights into the impact of paid prioritisation on energy efficiency and QoS of a separation architecture. By comparing our proposed non-neutral scenarios with a net-neutral baseline, we quantify the precise impact of this policy on key performance metrics, demonstrating that while it can improve average file transfer delay, it comes at a significant cost to overall system blocking probability, throughput, and energy efficiency, particularly under high traffic loads.

The rest of the paper is organised as follows. Related work on QoS and energy efficiency in separation architectures is discussed in Section 2. The methodology utilised for implementing the study is described in Section 3. Section 4 details the results and discussions. Practical Considerations and Future Work are discussed in Section 5. Finally, the conclusions of the paper are provided in Section 6.

2. RELATED WORK

In conventional cellular network deployments, macrocell base stations (BSs) covering a large area have often been used alone in the access network. However, such deployments are not capable of meeting the high data rates and high network capacity requirements necessitated by the rapidly increasing demand for data traffic in an energy-efficient manner (Hoydis, Kobayashi, & Debbah, 2011). As a result, the Separation Architecture was proposed by Xu et al. (2013) to address the energy efficiency challenge in cellular networks. This architecture was referred to as Hyper-cellular network by Zhao et al. (2013) and Control-Data Separation Architecture (CDSA) by Mohamed, Onireti, Imran, Imran, & Tafazolli (2015).

Typically, several small cell BSs are deployed within the coverage areas of macrocell BSs in HetNets. In HetNets based on the separation architecture, small cell BSs can be turned off, especially at low traffic loads, to reduce the energy consumed in the network without compromising the coverage of the network. Furthermore, A User Equipment (UE) can connect and communicate with a macrocell BS and a small cell BS simultaneously in this type of architecture. This kind of dual connection has been standardised for 4G in Release 12 of the 3rd Generation Partnership Project (3GPP) and is termed "dual connectivity" (3GPP, 2015). This concept is extended in 5G to allow a UE to connect to a 4G BS and a 5G BS simultaneously (Agiwal, Kwon, Park, & Jin, 2021).

Several studies have been carried out on QoS and energy efficiency in separation architectures. A cell-on-demand approach was proposed by Capone et al. (2012) to activate small cell BSs in sleep mode at macrocell BSs when needed to serve users. The choice of the most suitable small cell BS to allocate a user to was based on the location of the user rather than conventional signal strength metrics. The cell-on-demand approach achieved significantly higher energy efficiency than conventional cellular networks based on macrocell BSs alone. Similarly, a macrocell BS based small cell activation algorithm was also proposed by Ternon, Agyapong, & Dekorsy (2015). However, unlike Capone et al. (2012), that used the location of users, this algorithm determined the most suitable small cell BS to activate by considering the level of Signal to Interference plus Noise Ratio (SINR) and availability of adequate resources to satisfy user data rate requests. Despite user connection delays, the algorithm still achieved up to 45% energy saving and 25% throughput improvement relative to a baseline scheme, which does not support sleep mode operation of small cell BSs. A database-aided small cell activation scheme was developed by Ternon, Agyapong, Hu, & Dekorsy (2014), but unlike Capone et al. (2012) and Ternon, Agyapong, & Dekorsy (2015), in addition to macrocell BSs, small cell activations were done at UEs. Signal to Noise Ratio (SNR) values of small cell BSs at different locations reported by UEs were used to create a database of SNR values. When small cell BSs entered sleep mode, their SINRs were computed from the stored SNR values and used in small cell activation decisions. In high user density scenarios, this database-aided scheme achieved up to 40% energy saving relative to one without sleep modes for small cell BSs.

The energy efficiency and spectral efficiency of a separation architecture with different frequency bands utilised in the small cell and the macrocell layers were evaluated by Mukherjee & Ishii (2013). The values of these metrics for the separation architecture were compared with those of a conventional HetNet with the same frequency band for the small cell and macrocell layers and without data plane and control plane separation. The separation architecture was shown to outperform the conventional HetNet with regard to the energy efficiency and spectral efficiency of the small cell layer. The proportion of active small cell BSs which can be turned off was investigated in a separation architecture by Zhang, Gong, Zhou, & Niu (2015). It was shown by numerical evaluation for two typical daily traffic profiles that 50% of small cell BSs can be turned off on average at low load, and an additional 10% can be switched off if the macrocell tier borrows bandwidth from the small cell tier. Zhisheng, Guo, Zhou, & Kumar (2015) investigated the tradeoff between the total energy consumption of a data small cell BS and its mean overall delay. The data BS was deployed within the coverage of a control macrocell BS. Analytical equations were obtained for sleep/wake-up policies when the M/G/1 vacation queue was used to model the data BS. It was shown that energy consumption varied linearly with delay under varying close-down times before the data BS sleeps, but it varied non-linearly with delay under varying total packet arrivals before the data BS wakes up.

How different assumptions made for different power models of BSs affect energy saving in the small cell layer of a separation architecture was studied by Fisusi, Grace, & Mitchell (2017). Mathematical expressions were derived for energy savings over short timescales and long timescales for single and multiple BS scenarios in terms of the power consumption of different operating states of BSs. It was shown through theoretical analysis and simulation that under the same system settings, the energy savings achieved with different power models are different and are functions of model-specific state changes of BSs that result in significant energy savings. A mathematical framework was developed by Zhu, Wang, & Qian (2019) to investigate the optimal number of small cell BSs that should cooperatively serve users to achieve maximum energy efficiency in a separation architecture. An equation was derived for the lower bound of the average downlink rate, and based on this equation, formulas for energy efficiency and area spectral efficiency were obtained in terms of the number of cooperating small cell BSs in a cluster. It was deduced from simulation results that an optimal cluster size exists that maximises energy efficiency and also meets the minimum area spectral efficiency requirement.

An algorithm was proposed by Baidowi & Chu (2020) to maximise energy efficiency by jointly optimising user associations to small cell BSs and the proportion of small cell BSs to switch off in a separation architecture. The decision to switch off small cell BSs was made based on the traffic load they support. Users associated with small cell BSs to be turned off were handed over to other small cell BSs with the highest Signal to Interference Ratio (SIR). A bio-inspired algorithm was proposed by Sherif

& Haci (2023) for a separation architecture to maximise energy efficiency through the selection of appropriate modes of operation (on, standby, sleep, or off) for small cell BSs at the macrocell BS. Bias function values were used to optimise the BS power consumption in different operation modes. The proposed algorithm determined the optimal bias function values for the BSs. Energy efficiency maximisation achieved with the proposed scheme was better compared to conventional sleep mode management schemes.

A reinforcement learning algorithm with linear function approximation was implemented by Ozturk et al. (2021) at the macrocell BS of a separation architecture to offload traffic and switch small cell BSs on/off. The proposed algorithm determined the optimal switching on/off policy without visiting all states by utilising a compact representation of the action-value function. The proposed algorithm achieved network throughput comparable to a scheme which never switches off small cell BSs and energy saving similar to the optimal Exhaustive Search (ES) algorithm. A graph representation based unsupervised learning algorithm was proposed by Tan et al. (2022) for determining the small cell BSs to switch off/on at the macrocell BS based on the instantaneous power consumption and traffic load of the BSs. The small cell BSs were of different types (pico, femto, micro, Remote Radio Head (RRH)) and power consumption modes. The proposed algorithm learnt the optimal policy for energy saving while maintaining users' QoS appreciably. The scheme achieved energy efficiency gains of 10.41% relative to a baseline scheme without BSs switching off. Even though only 75.76% of the peak performance measured under the ES algorithm was achieved by the proposed scheme, it is more scalable and less computationally complex.

Although QoS and energy efficiency have been extensively studied in the separation architecture, as observed in the previous studies in this area, the impact of non-neutrality on these two critical cellular network metrics has not been studied. Hence, this paper addresses this gap in the literature by presenting a study on the impact of paid prioritisation, the form of non-neutrality that most significantly affects how resource allocation schedulers work, on a separation architecture. Specifically, in the study, a resource allocation scheme, termed Paid Prioritisation-based Clustering Capability Rating (PPCCR) that allows radio resources to be allocated discriminately to users based on subscription fees paid (paid prioritisation), was developed for a separation architecture. The PPCCR scheme was combined with the topology management scheme developed by Fisusi, Grace, & Mitchell (2017), which switches BSs on/off to reduce the energy consumed by the network. The combined resource allocation and topology management schemes created a non-neutral, energy-efficient scenario, and two versions of this scenario - with different percentages of preferentially treated users (premium users) - were evaluated in the study. The QoS and energy efficiency performances of the two non-neutral scenarios were compared with those of a net-neutral scenario through system-level simulations. This was done to establish how the energy efficiency and QoS of a non-neutral policy differ from the net-neutral case, thereby gaining insight into the impact of non-neutrality on energy efficiency and QoS in a separation architecture.

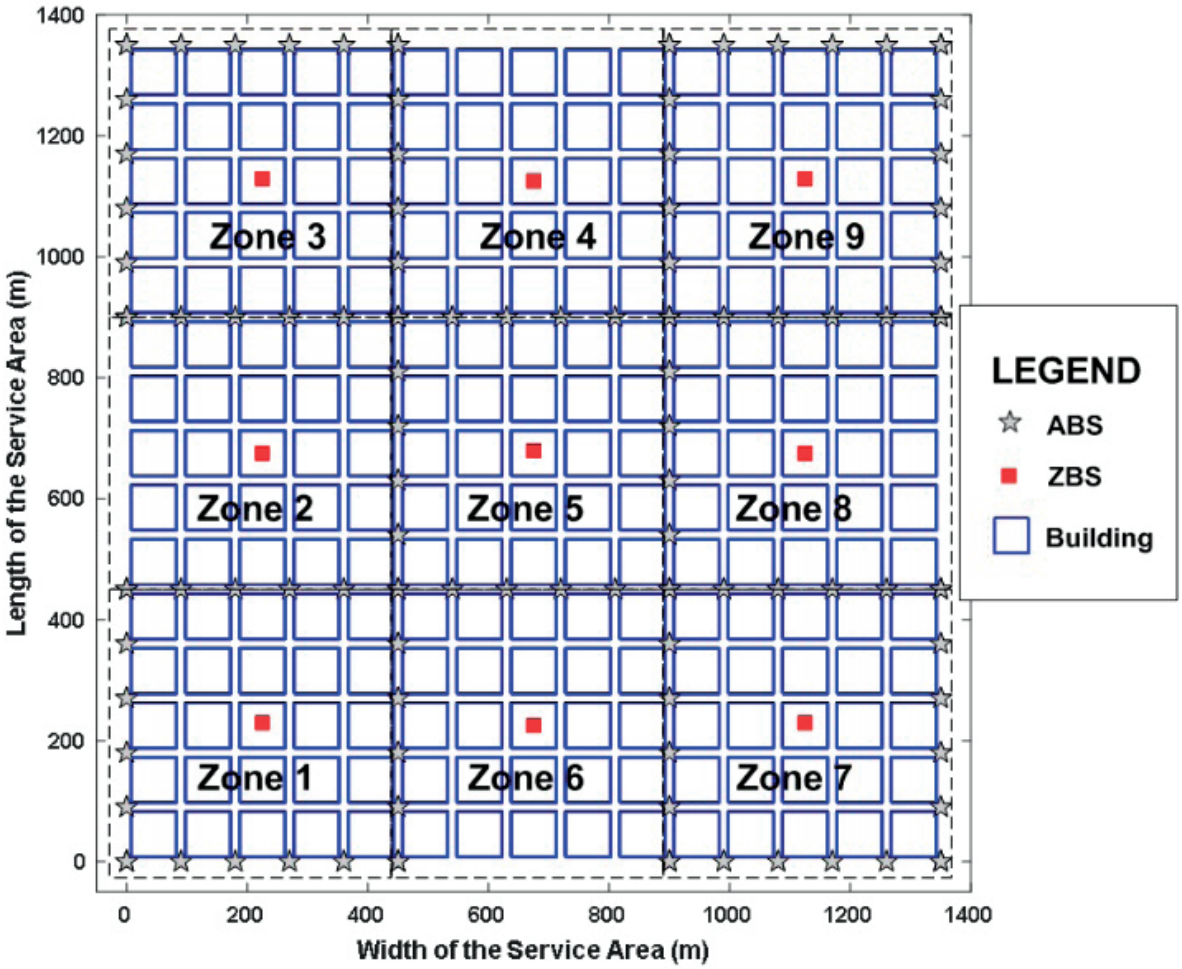

Figure 1 BuNGee Separation Architecture (Fisusi, Grace, & Mitchell, 2017)

3. METHODOLOGY

The study was carried out to investigate the impact of non-neutrality on the QoS and energy efficiency of a separation architecture. A novel non-neutral energy-efficient scheme, Paid Prioritisation-based Clustering Capability Rating (PPCCR) scheme, developed in the study, prioritised premium users over normal

users and clustered all users around as few as possible small cell base stations in its resource allocation decisions to create energy-saving opportunities. These savings were realised by a complementary topology management scheme that switched idle base stations into a low-power sleep mode. The combination of these two schemes created a non-neutral energy-efficient scenario. The effectiveness of the combined approach was evaluated through extensive Monte Carlo simulations, which compared two versions of the non-neutral scenarios against a net-neutral scenario to assess the impacts on both QoS and energy efficiency.

A detailed description of the separation architecture utilised and its implementation is provided subsequently. This is followed by a thorough explanation of the implementation of the proposed PPCCR scheme. The system-level simulations carried out to compare the performance of the non-neutral and net-neutral scenarios are then presented.

3.1 System Model

Uplink data traffic in the separation architecture proposed by Fisusi, Grace, & Mitchell (2017) was considered in the study. This architecture is a modification of the European Union FP7 Beyond Next Generation (BuNGee) mobile broadband network (Roth et al., 2010) into a separation architecture through the introduction of control macrocell BSs into the access network, which originally contained only small cell BSs. This modified architecture, termed BuNGee Separation Architecture hereafter, is shown in Figure 1.

The network contains, as shown in Figure 1, a dense deployment of low-cost, low-power, below-rooftop small cell BSs (termed Access Base Stations (ABSs)) outdoors along the streets in a regular pattern. The service area is partitioned into square zones with control macrocell BSs, termed Zone Base Stations (ZBSs), deployed at the centre of the zones to provide ubiquitous coverage and handle the control functions in their respective zones.

Figure 2 shows, in detail, a single zone, such as Zone 3, to illustrate the ABS antenna directions and the frequency plan for the BuNGee Separation Architecture. Each ABS has two directional antennas facing opposite directions, either North and South or East and West. Horizontal rows of ABSs (such as ABSs 3, 4, 5 and 6) serve the North and South directions; while vertical columns of ABSs (such as ABSs 9, 10, 11, and 12) serve the West and East directions. At the junctions of two streets, two ABSs are deployed; one belonging to the horizontal row and the other to the vertical column.

Four unique frequency bands, F_1, F_2, F_3 and F_4 (one for each direction and each with a bandwidth of 10 MHz), are allocated to the ABS (or small cell) layer. The 10 MHz bandwidth of each band is subdivided into 10 equal subbands or subchannels. Each ABS utilises two of these bands for its two antennas. In order to minimise interference, as shown in Figure 2, the antennas of neighbouring ABSs (such as ABSs 3 and 4 or ABSs 23 and 24) pointing in a similar direction use different frequency bands. Also, the antennas of opposite ABSs (such as ABSs 3 and 18 or ABSs 9 and 24) directed along the same street use different frequency bands. The ZBSs utilise a 10 MHz band different from the ABS bands.

Only one Mobile Station (MS) can utilise a subchannel at a given time. The MSs are uniformly distributed along the streets. The service area represents a high-density urban hotspot like a city centre, where ultra-high traffic intensities are generated at peak hours and dense small cell BS deployments are particularly suited to serve. The assumption of a uniform user distribution was made to ensure the comparative analysis of policies for consistent and unbiased heavy-load conditions, rather than one influenced by a specific intra-hotspot clustering pattern. Each MS antenna is an omnidirectional radiator.

The ABSs handle the data requests of the MSs under the supervision of the ZBSs in their zones. ABSs can relay channel quality and other information to the ZBSs in their zones with low delay through backhaul links. Some ABSs are associated with more than one zone and thus more than one ZBS. Information exchange between such ABSs and all ZBSs associated with their zones is also possible through backhaul links.

Each ZBS makes the decision regarding which ABS is most suitable to handle the data request of an MS in its zone. This is achieved by requiring the channel quality measurement (SINR in the study) related to the MS to be reported by all ABSs in the zone to the ZBS. The ZBS then determines the best ABS to serve

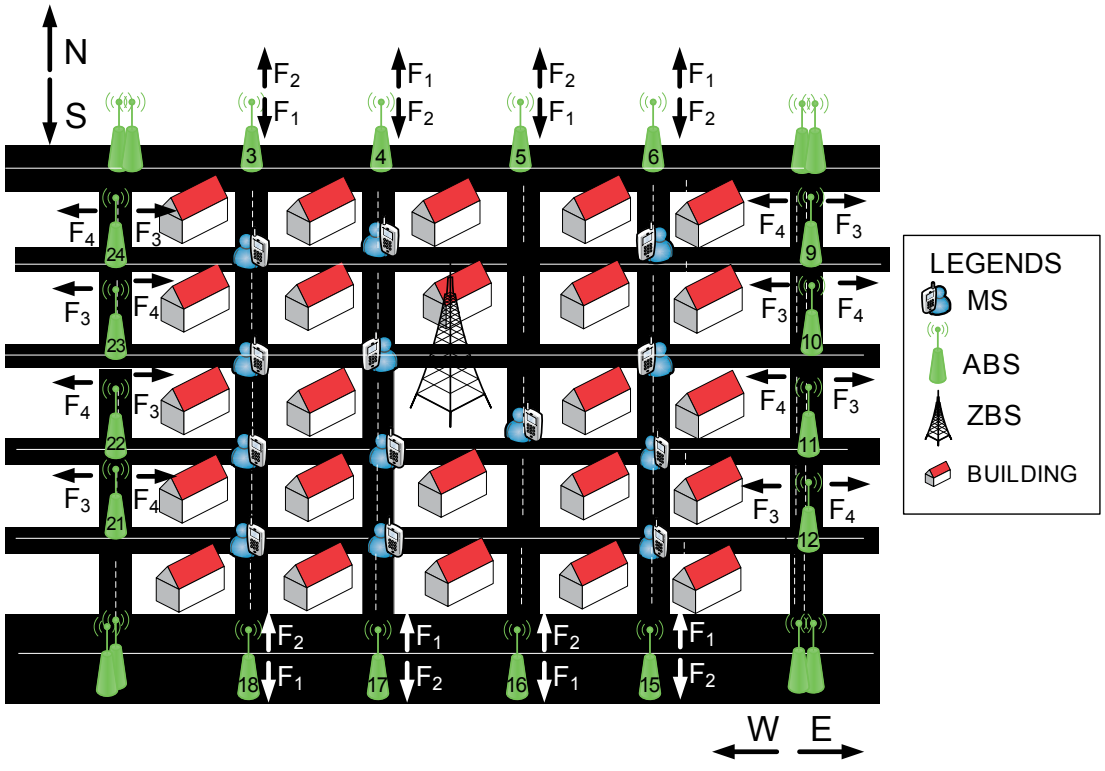

Figure 2 ABS Antenna Directions and The Frequency Plan

the MS request based on the reported channel quality and other parameters required by the resource allocation scheme utilised. Each MS makes one uplink request per time and utilises one subchannel for uplink transmission.

In the study, for the non-neutral scenarios, two kinds of MSs - premium and normal MSs - were distributed over the network. Premium MSs were assumed to have paid a higher subscription fee for a guaranteed uplink data rate of 3 Mbps. This value is the minimum upload speed recommended for live streaming High-Definition (HD) video 1080p at 30 frames per second (fps) on popular platforms like YouTube. Moreover, it has been shown that video bitrates of 2.8 Mbps and 4.0 Mbps are sufficient to achieve "Fair" and "Good" quality playback, respectively, for H.264 coded Full HD 1080p content at 30 fps (Uhrina, Holesova, Bienik, & Sevcik, 2021). Therefore, this 3 Mbps guaranteed data rate ensures that premium users can be provided at least a "Fair" quality live stream at all times, with the potential to achieve "Good" quality streams when they are allocated channels with data rates of 4 Mbps and above during uncongested periods. The upload speed of the premium users was chosen to be high enough for typical and popular uplink-focused mobile applications. In contrast, the normal MSs were assumed to have paid the basic subscription fee with much lower QoS guarantees.

The WINNER II B1 Urban Micro-cell propagation model (Kyosti et al., 2007) was used to estimate the losses (path loss and shadowing) between ABSs and MSs. Uplink data transmission rates *R* were estimated by the Truncated Shannon Bound (TSB) (3GPP, 2009):

$$\mathrm{R}=\begin{cases}0; & \text{for SINR} < \mathrm{SINR_{min}}\\ \alpha \log_2(1+\mathrm{SINR}); & \text{for } \mathrm{SINR_{min}} < \mathrm{SINR} < \mathrm{SINR_{max}}\\ \mathrm{R_{max;}} & \text{for SINR} > \mathrm{SINR_{max}}\end{cases} \quad (1)$$

$\propto$ represents the attenuation factor, $SINR_{min}$ denotes the lower limit of suitable uplink SINR, and $SINR_{max}$ denotes the SINR at which the highest possible link data rate is attained. The TSB parameters for BuNGee (Jiang et al., 2012) are $\propto = 0.65$, $SINR_{min} = 1.8$ dB, $SINR_{max} = 21$ dB, and $R_{max} = 4.54$ bps/Hz. These values were obtained through the link-to-system mapping methodology used in the BuNGee project (Jiang et al., 2012), which calibrated the TSB model to accurately reflect the practical performance of the uplink as determined by detailed link-level simulations.

A new uplink request from an MS was admitted if the SINR of the MS met the threshold for its class and its interference to any existing user transmission would not deteriorate link quality below the agreed level. The new uplink request was blocked if no ABS could satisfy these conditions. The QoS metrics used were blocking probability, average file transfer delay and throughput. The file transfer delay measured the time from when a user's uplink request was sent to the ZBS to when the file was completely and successfully transmitted to the serving ABS. Throughput measured the total number of bits which were successfully transmitted in a period in bits/second. Only the user information-carrying bits were considered. Overhead bits such as packet headers and error detection bits were not considered.

In a separation architecture, the control macrocell BSs (the ZBSs in the study) are always on and never switched off to satisfy the requirement of ubiquitous network coverage in the service area. Energy saving is achieved by putting some data small cell BSs (ABSs in this work) into sleep mode when the traffic intensity permits. ABSs consume significantly lower power in the sleep mode than when they are transmitting data to MSs (transmitting mode) or when they are receiving data from MSs (receiving mode). The study focused on energy saving in the ABS (or small cell) layer of the network under the net-neutral and non-neutral regimes for the case of uplink data traffic. The total energy consumed by the ABSs, E_{ABS} , was calculated using the energy model proposed by Han, Grace, & Mitchell (2012) :

$$\mathrm{E_{ABS}}=\sum_{\mathrm{d}=1}^{\mathrm{n_{abs}}}\Big(\mathrm{t_{sleep,d}p_{sleep,d}}+\mathrm{t_{Rx,d}}\frac{\mathrm{p_{Rx,d}}}{\mu_{\mathrm{RF}}}+\mathrm{t_{Tx,d}}\frac{\mathrm{p_{Tx,d}}}{\mu_{\mathrm{RF}}}+\mathrm{t_{idle,d}P_{idle,d}}+\mathrm{n_{wakeup,d}E_{wakeup}}\Big)\frac{1}{(1-\mu_{\mathrm{c}})} \quad (2)$$

n_{abs} represents the total number of ABSs in the network; P_{sleep}, P_{Rx}, P_{Tx} and P_{idle} represent the power consumed by an ABS in sleep, receiving, transmitting and idle modes, respectively. An ABS is in idle mode when it is on but not transmitting or receiving data; however, it is on standby to serve user requests in this mode. The idle mode power consumption equals the no-load, static power consumption of an ABS, which is due to power expended on non-radio frequency aspects like power supply and cooling (Budzisz et al., 2014). t_{sleep}, t_{Rx}, t_{Tx} and t_{idle} represent the cumulative time an ABS spends in sleep, receiving, transmitting and idle modes, respectively. μ_{RF} represents the power amplifier efficiency, and μ_c represents losses incurred in the battery and power supply. n_{wakeup} represents the total number of times an ABS wakes up from sleep, while E_{wakeup} represents the energy expended during the waking-up process. The time it takes an ABS to switch from one mode to another is assumed to be negligible.

In the study, P_{Tx} was assumed to be zero for all ABSs since the focus was on the uplink direction. Also, P_{idle} was assumed to be equal to P_{Rx} because for low-power small cell BSs, the dependency of power consumption on traffic load is negligible (Auer et al., 2011). Table 1 specifies the values for the parameters in equation (2).

Energy efficiency was measured in terms of Energy Consumption Rating (ECR) and Energy Reduction Gain (ERG) in the study. ECR estimates the energy consumed in transmitting a bit of information and is measured in joules per bit. According to the Green Radio Project (Han et al., 2011), ECR is given by

$$\mathrm{ECR}=\frac{\text{Energy Consumed in the Network}}{\text{Successfully transmitted information bit}} \quad (3)$$

The ERG is often used in the literature to measure energy efficiency (e.g. He et al., 2010; Maleki & Abolhassani, 2014; Turyagyenda, O'Farrell, & Guo, 2012). It estimates the energy efficiency of a test scheme relative to a baseline scheme using ECR as follows:

$$\mathrm{ERG}=\frac{\mathrm{ECR_{baseline}}-\mathrm{ECR_{test}}}{\mathrm{ECR_{baseline}}}\times 100\% \quad (4)$$

Table 1 Parameters for the ABS Energy Consumption (Han, Grace, & Mitchell, 2012)

Parameter	Value
Power in the receiving mode	5W
Power in sleep mode	250mW (assumed 5% of receiving mode)
Efficiency of RF	20%
Efficiency of supply loss	10%
ABS max transmit power	5W
Wakeup Energy	50J

3.2 Proposed Non-Neutral Resource Allocation Scheme

Previous resource allocation schemes developed for the BuNGee architecture were net-neutral in nature. Hence, when these schemes were utilised, there was no discrimination in the allocation of resources based on subscription fees or the identity of users. The non-neutral resource allocation scheme proposed in the study is a modification of a previous net-neutral energy-efficient resource allocation scheme, the Normalized Clustering Capability Rating (NCCR) scheme (Fisusi, Grace, & Mitchell, 2013), developed for the BuNGee architecture. The modifica-

tion of NCCR was done to enable the allocation of resources to different classes of users based on subscription fees paid (paid prioritisation) whilst still achieving energy savings. This modified scheme is referred to as the PPCCR scheme. The scheme considers two classes of MSs or users: normal and premium users. Premium users pay an additional fee to guarantee a higher QoS than normal users. The foundational algorithm for this scheme was developed in Arotiba (2019).

Specifically with PPCCR, when an MS sends a request for an uplink subchannel to a ZBS, the ZBS requests the channel quality (SINR values in the study) of uplink subchannels of the ABSs in its zone. The ZBS shortlists ABS subchannels with uplink SINR values that satisfy the admission threshold for the user class of the MS. Next, the ZBS creates a second list containing only ABS subchannels from the initial list that will not result in interferences high enough to degrade data rates of ongoing MS transmissions below guaranteed levels. Then, among the ABS subchannels in the second list, the ZBS shortlists the subchannel with the highest uplink SINR from each ABS to create a third list. Hence, the third list will contain only one subchannel from each ABS. Finally, the subchannel of the ABS which has the highest value of NCCR in the third list is selected and allocated to the MS.

If an ABS has a higher NCCR value than another ABS, it implies that it has a higher capability to cluster more MSs around itself than the other ABS because of its location and current load. Selection of ABSs with higher NCCR values in resource allocation decisions results in fewer active ABSs and more idle ABSs that can go to sleep, leading to higher energy savings. A detailed explanation of the computation of the NCCR value for an ABS is provided in the energy efficiency study by Fisusi, Grace, & Mitchell (2013), where the NCCR scheme was first proposed.

Resource allocation decisions based on the proposed PPCCR scheme are carried out as follows: Let $\mathrm{L_i}$ denote the current load level of an ABS *i* (defined in terms of the total number of users or MSs currently being served by the ABS), while $\mathrm{L_{max}}$ denotes the maximum load capacity of an ABS. The normalized load $\mathrm{x_i}$ on ABS *i* is given by:

$$\mathrm{x_i} = \frac{\mathrm{L_i}}{\mathrm{L_{max}}} \tag{5}$$

Let $\mathrm{X} = (\mathrm{x_1} \;\; \mathrm{x_2} \;\; \dots \;\; \mathrm{x_m})$ denote the vector of normalized loads of all *m* ABSs in the zone of the ZBS and MS in question, while *S* denote an $m \times t$ matrix of pre-connection subchannel SINR values for ABSs given by:

$$\mathrm{S} = \begin{pmatrix} \mathrm{s_{1,1}} & \mathrm{s_{1,2}} & \dots & \mathrm{s_{1,t}} \\ \mathrm{s_{2,1}} & \mathrm{s_{2,2}} & \dots & \mathrm{s_{2,t}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{s_{m,1}} & \mathrm{s_{m,2}} & \dots & \mathrm{s_{m,t}} \end{pmatrix} \tag{6}$$

where *m* denotes the number of ABSs in the zone and *t* denotes the number of uplink subchannels of an ABS; $s_{i,j}$ is, therefore, the SINR of subchannel *j* of ABS *i* available for uplink transmission *(1≤i≤m; 1≤j≤t)*. Let c_i denote the NCCR value of the *ith* ABS, so $\mathrm{C} = (\mathrm{c_1} \;\; \mathrm{c_2} \;\; \dots \;\; \mathrm{c_m})$ denotes the vector of NCCR values of the ABSs in the zone. Also, let $\mathrm{K} = (\mathrm{k_1} \;\; \mathrm{k_2} \;\; \dots \;\; \mathrm{k_n})$ denote the vector of the user classes of all *n* MSs in the system; hence k_q denotes the user class of an arbitrary *qth* MS.

Furthermore, let SINR_{th} denote the admission SINR threshold for an MS; SINR_n and SINR_p are the admission SINR thresholds for normal and premium users, respectively. SINR_g denotes the SINR for guaranteed data rate; and SINR_{min} is the minimum SINR for user admission and data transmission. The admission SINR thresholds are derived from the TSB equation (equation (1)). This equation determines the theoretical minimum SINR (SINR_{min}) required for user admission. However, in our scheme, both user classes are admitted at SINR values significantly higher than this minimum to ensure service quality. Premium users are assigned an SINR threshold, SINR_p, based on the TSB equation that guarantees their agreed data rates. Normal users are assigned a lower SINR threshold, SINR_n, reflecting lower service tier. The specific values of SINR_{min} , SINR_p, and SINR_n used in the study are stated and justified in the system-level simulations section (section 3.3).

In addition, let V be an $m \times t$ matrix of post-connection subchannel SINR values for ABSs given by:

$$\mathrm{V} = \begin{pmatrix} \mathrm{v_{1,1}} & \mathrm{v_{1,2}} & \dots & \mathrm{v_{1,t}} \\ \mathrm{v_{2,1}} & \mathrm{v_{2,2}} & \dots & \mathrm{v_{2,t}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{v_{m,1}} & \mathrm{v_{m,2}} & \dots & \mathrm{v_{m,t}} \end{pmatrix} \tag{7}$$

where $v_{i,j}$ is the uplink SINR of subchannel *j* of ABS *i* already being used for an MS uplink transmission. Therefore, whereas S is a record of the SINR values of the subchannels of ABSs available for MS connection, V contains information about the SINR values of the subchannels of ABSs already being utilised for MS uplink transmissions.

Also, let *W* be an $m \times t$ matrix of the identity of the MSs utilising the ABS subchannels given by:

$$\mathrm{W} = \begin{pmatrix} \mathrm{w_{1,1}} & \mathrm{w_{1,2}} & \dots & \mathrm{w_{1,t}} \\ \mathrm{w_{2,1}} & \mathrm{w_{2,2}} & \dots & \mathrm{w_{2,t}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{w_{m,1}} & \mathrm{w_{m,2}} & \dots & \mathrm{w_{m,t}} \end{pmatrix} \tag{8}$$

where $w_{i,j}$ is the identity of the MS utilising subchannel *j* of ABS *i* for uplink transmission. Similarly, let *Y* be an $m \times t$ matrix of the frequency bands of the ABS subchannels given by:

$$\mathrm{Y} = \begin{pmatrix} \mathrm{y_{1,1}} & \mathrm{y_{1,2}} & \dots & \mathrm{y_{1,t}} \\ \mathrm{y_{2,1}} & \mathrm{y_{2,2}} & \dots & \mathrm{y_{2,t}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{y_{m,1}} & \mathrm{y_{m,2}} & \dots & \mathrm{y_{m,t}} \end{pmatrix} \tag{9}$$

where $y_{i,j}$ is the frequency band of subchannel *j* of ABS *i*.

The ABS subchannel assigned to an arbitrary *q*th MS is determined by the ZBS as follows:

1. All the elements of ***X, S, C, V*** and ***W*** are initially set to zero (before the first MS arrives in the network).

2. When the *qth* MS arrives and requests for a subchannel, the ZBS requests uplink SINR $s_{i,j}$ in respect of the *qth* MS for all unoccupied subchannel *j* of ABS *i* with $w_{i,j} = 0$ from all active ABSs in the zone of the MS and updates S. At this stage, the uplink SINRs of MS *q* over subchannels that are not being used by other MSs are requested from the ABSs and stored.

3. The ZBS requests for the user class of MS *q*.

4. The ZBS verifies the condition $\mathrm{s_{i,j}} \geq \mathrm{SINR_{th}}$ for all values of *i* and *j* and reset any $\mathrm{s_{i,j}} < \mathrm{SINR_{th}}$ to zero. $\mathrm{SINR_{th}} = \mathrm{SINR_n}$ if MS *q* is a normal user, but if it is a premium user $\mathrm{SINR_{th}} = \mathrm{SINR_p}$. $\mathrm{SINR_n}$ and $\mathrm{SINR_p}$ are the minimum allowed SINR values that newly arriving normal and premium users, respectively, must achieve on an available channel for it to be considered for allocation. This step ensures the admission threshold condition is satisfied for the new incoming MS *q*. At the end of this stage, the ZBS retains only uplink subchannels that meet the admission threshold criterion for the user class of the MS in the subchannel shortlist.

5. For each $\mathrm{s_{i,j}} > 0$, the ZBS recalculates post-connection uplink SINR $\mathrm{v_{f,j}} (\mathrm{f} \neq \mathrm{i}; \quad 1 \leq \mathrm{f} \leq \mathrm{m})$ for all *f's* that satisfy the criteria $\mathrm{v_{f,j}} > 0$, $\mathrm{w_{f,j}} > 0$ and $\mathrm{y_{f,j}} = \mathrm{y_{i,j}}$, taking into consideration the potential interference from MS *q*. For every recalculated $\mathrm{v_{f,j}} < \mathrm{SINR_g}$ for the user class of already transmitting MS $w_{f,j}$, the ZBS resets

the associated $s_{i,j}$ to zero. $SINR_g = SINR_p$ for premium users, but $SINR_g = SINR_{min}$ for normal users. $SINR_g$ is the minimum SINR (guaranteed SINR) that an already-connected user's link must be able to maintain after the new user is added, while $SINR_{min}$ is the minimum value of SINR permitted for admitting users into the network and for uplink data transmission. This step ensures all subchannels that will introduce interferences that will degrade ongoing MS transmissions in the zone below the guaranteed SINR levels are dropped. The ZBS also coordinates with neighbouring ZBSs via a logical interface (like the X_2 in 4G or the X_n in 5G) to prevent a new MS admitted into its zone from degrading existing MS transmissions in neighbouring zones. It is important to note that this interference check is highly efficient as it is bounded by a small, fixed constant because of the frequency plan used in the BuNGee Separation Architecture. This kind of interference management centric frequency plan is a standard feature of cellular networks. Due to the peculiar positioning of ABS antennas to manage interference, uplink transmission on a subchannel can cause interference to a maximum of 5 subchannels in its zone and 48 subchannels in neighbouring zones in the worst-case scenario of the most central zone (Zone 5 in Figure 1) with 8 neighbouring zones. Hence, a maximum of 5 intra-zone and 48 inter-zone interference checks are ever needed for a single subchannel. This is a constant-time operation, O(1), for each of the candidate $m \times t$ subchannels. At the end of this stage, only uplink subchannels in the shortlist from the previous stage that will not degrade the QoS of ongoing MS transmissions below guaranteed levels are retained in the updated subchannel shortlist by the ZBS.

6. For each row i of $S (1 \leq i \leq m)$, the ZBS determines the highest value of SINR $s_{i,h}$ in the row $(1 \leq h \leq t)$ and resets all row elements $s_{i,j}$ for which $j \neq h$ to zero. Hence, S will contain at most one non-zero element per row. This step ensures only the suitable uplink subchannel with the highest SINR from each ABS is shortlisted at this stage of the resource allocation decision.

7. For each row i of S with a non-zero element, the ZBS computes c_i (the NCCR for ABS i) and updates C. Then, the ZBS determines the element of C with the greatest value $c_u (1 \leq u \leq m)$ and resets all the elements of each row i of S for which $i \neq u$ to zero. Thus, S will now contain only one non-zero element $s_{u,h}$, which is the SINR of the suitable subchannel with the highest SINR on the suitable ABS with the highest NCCR. This subchannel is allocated to MS q.

8. ZBS updates X, W and V to account for the newly admitted MS. Before the next MS request in its zone, the ZBS resets all the elements of S *and* C to zero. Step 1, which is the initialisation stage, is not repeated for MS uplink subchannel requests after the first arrival.

The non-neutral PPCCR scheme was developed to allocate resources based on paid prioritisation while operating in an energy-efficient manner. Energy efficiency was achieved mainly by reducing the number of active ABSs. In the study, the PPCCR scheme, which is a resource allocation scheme, was complemented by the topology management scheme proposed by Fisusi, Grace, & Mitchell (2017). The PPCCR scheme's decisions conditioned the network state by determining the real-time traffic load on each ABS. The topology management scheme then monitored this state, and based on its own rules regarding traffic load, idle waiting periods, and blocked requests, it switched ABSs into and out of sleep mode. The combination of the non-neutral resource allocation scheme PPCCR and the topology management scheme of Fisusi, Grace, & Mitchell (2017) resulted in a non-neutral energy-efficient scenario. Two different versions of this non-neutral scenario were implemented and compared with a net-neutral energy-efficient scenario through system-level simulations.

3.3 System-Level Simulations

Monte Carlo simulations were carried out using MATLAB in the BuNGee separation architecture. Each simulation result was obtained after 50,000 iterations. This implies that each data point on the graphs presented in the Results and Discussion section represents a mean value obtained after 50,000 simulation iterations. This large number of iterations ensured that the averages for the plotted points sufficiently converged with minimal variance.

In the service area, 9 ZBSs and 112 ABSs were deployed as shown earlier in Figure 1. A total of 6,000 MSs, consisting of both premium and normal MSs, were distributed uniformly along the streets of the service area. This dense population of users depicts highly populated areas such as city centres. Each MS, whether normal or premium, arrived in the network with a request to upload a file of 2 MB. MS arrival into the network constituted a Poisson arrival process with a mean arrival rate of λ. Four unique frequency bands were allocated to the ABS layer and each band was subdivided into 10 subchannels as earlier stated. The simulation parameters are stated in Table 2.

Table 2 Simulation Parameters

Parameter	Value
Deployment area dimension	1350 m ×1350 m
Street width	15 m
Building block size	75 m × 75 m
ABS antenna height	5 m
MS antenna height	1.5 m
Carrier frequency	3.5 GHz
MS transmit power	23 dBm
ABS maximum gain	17 dBi
MS Antenna gain	0 dBi
Noise Floor	-114 dBm/MHz
Admission SINR (Premium users ($SINR_p$))	14 dB
Admission SINR (Normal users ($SINR_n$))	10 dB
Minimum SINR ($SINR_{min}$)	1.8 dB
Maximum SINR ($SINR_{max}$)	21 dB

Two implementation versions were considered in evaluating the performance of the non-neutral energy-efficient scenario. In the first version, termed non-neutral scenario I, the MS population consisted of 20% premium and 80% normal users. In contrast, the second implementation version, referred to as non-neutral scenario II, consisted of 60% premium and 40% normal users (see Table 3). These two non-neutral energy-efficient scenarios were based on the use of the PPCCR scheme for resource allocation and the topology management scheme of Fisusi, Grace, & Mitchell (2017) for switching ABSs on and off. The non-neutral energy-efficient scenarios were compared with a net-neutral energy-efficient scenario, which was based on the NCCR scheme for resource allocation and the topology management scheme of Fisusi, Grace, & Mitchell (2017) for switching ABSs on and off.

The three scenarios studied are associated with three different premium user populations, selected to model the impact of a non-neutral premium service at distinct stages of market adop-

Table 3 User Distribution for Different Scenarios

Scenario	Ratio of Premium to Normal MSs	Number of Premium MSs	Number of normal MSs
Highest SINR	0:100	0	6000
Net-neutral	0:100	0	6000
Non-neutral I	20:80	1200	4800
Non-neutral II	60:40	3600	2400

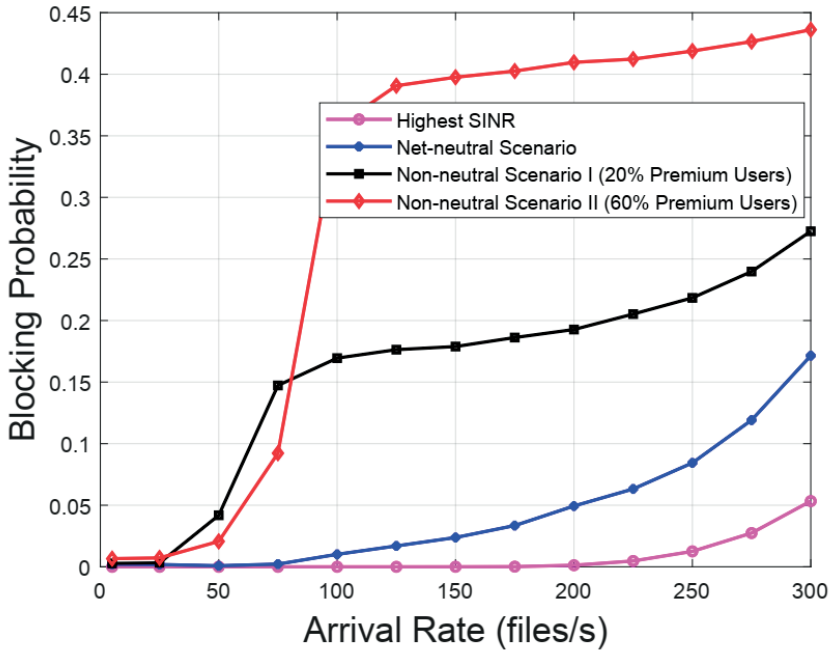

Figure 3 Comparison of Blocking Probability of Different Scenarios

tion: 0%, 20%, and 60% for the net-neutral, non-neutral I, and non-neutral II scenarios, respectively. The 0% case represents the baseline net-neutral scenario before a premium service is introduced. The 20% case models a state of early adoption, where the service has attracted a significant but minority user base. The 60% case models a mature market with mass adoption, where the premium service has become the majority choice. Direct comparison of these scenarios - representing zero, early, and mass adoption - demonstrated the impact of different levels of non-neutrality on the energy efficiency and QoS of a separation architecture.

Different classes of users have different admission SINR thresholds. The admission SINR threshold for premium users ($SINR_p$) was set at 14 dB to match the SINR required to provide the guaranteed 3 Mbps data rate for this user group's HD video live streaming. From the Truncated Shannon Bound (TSB) equation (equation (1)), the SINR equivalent to 3 Mbps is approximately 13.71 dB, which was finally rounded up to 14 dB. In contrast, the admission SINR thresholds for normal users ($SINR_n$) was set at 10 dB. This value is significantly higher than the minimum possible admission SINR ($SINR_{min}$) of 1.8 dB from the Truncated Shannon Bound (TSB) equation. This is common practice in cellular network resource management. Although the normal user SINR threshold is lower than the premium value, it limits the interference new users introduce to the system and provides them with sufficient headroom in a network that is not severely congested against interference from other subsequent users. Moreover, 10 dB corresponds to a data rate of approximately 2.25 Mbps and will adequately support many popular applications, including Standard Definition (SD) video streaming, social media, and web browsing.

The Highest SINR resource allocation scheme with no ABSs ever switched off was used as the baseline scheme for estimating energy saving for non-neutral and net-neutral scenarios. This Highest SINR scheme, so called by Fisusi, Grace, & Mitchell (2017), was utilised in an energy efficiency study on the BuNGee architecture by Han, Grace, & Mitchell (2012). This scheme connects an MS to the ABS which offers the highest uplink SINR. Thus, MSs are typically connected to the ABSs that are closest to them and first choices in terms of uplink SINR. Rather than concentrate or cluster MSs around a few active ABSs, the Highest SINR scheme tends to allocate MSs over a large number of ABSs; hence, it is not energy efficiency focused.

As shown in Table 3, under the net-neutral scenario and the Highest SINR scheme, all MSs were normal users. A blocking probability target of 5% or lower was assumed as the system's desired range of operation. The QoS metrics were blocking probability, average file transfer delay and throughput, while the energy efficiency metrics were ECR and ERG. The ERG of the non-neutral and net-neutral scenarios were obtained with the Highest SINR scheme serving as the baseline scheme.

4. RESULTS AND DISCUSSIONS

Blocking probability performances of the scenarios over a range of file arrival rates from very low to very high are shown in Figure 3. The highest SINR scheme performs best overall with regard to blocking probability. This is as a result of all ABSs being available to serve MSs since all ABSs are always on. Also, MSs always connect to the closest and highest SINR ABSs. This enables the avoidance of interference associated with connection to distant and lower SINR choices. Furthermore, the blocking probabilities of all the scenarios have similar performances at very low traffic

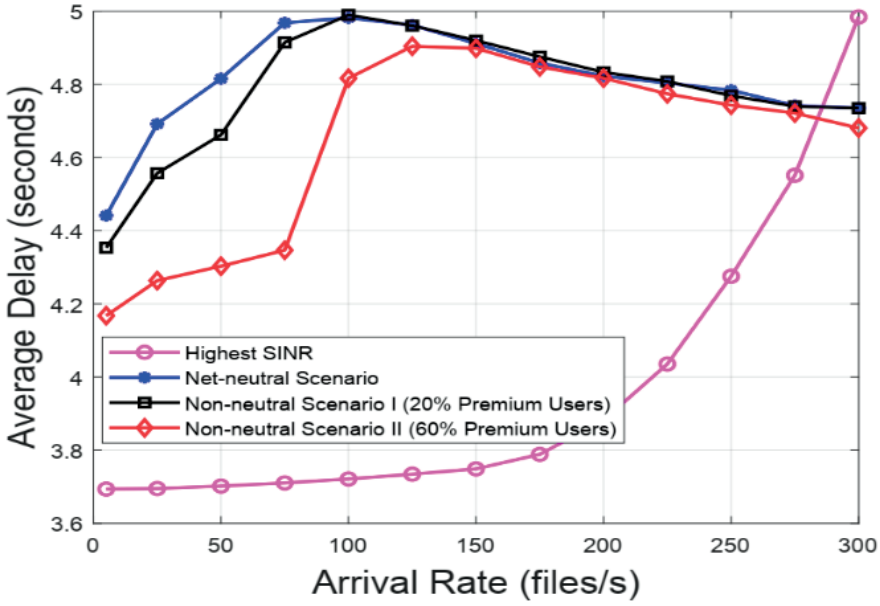

Figure 4 Comparison of Average File Transfer Delay of Different Scenarios

load (below 25 files/s) since only a few MSs would be active simultaneously and interference would be minimal. However, the blocking probability of the non-neutral scenario II is significantly higher than those of other scenarios above the file arrival rate of 75 files/s. This can be attributed to two factors: the premium users are predominant in this scenario and the high SINR requirement of 14 dB (corresponding to 3 Mbps upload speed) for premium users becomes less available to admit this kind of users when ABSs are switched off. The need to maintain the 14 dB SINR for admitted premium users throughout their file transfers, rather than 1.8 dB for normal users, also leads to higher blocking in the non-neutral scenario II than in other scenarios with fewer or no premium users. Furthermore, the non-neutral scenario I with 20% premium users has a significantly poorer blocking probability performance than the net-neutral scenario with no premium users at all. Hence, the higher the percentage of premium users in the network, the poorer the blocking probability beyond low traffic load intensities. Also, the non-neutral scenarios surpass the blocking probability target of 5% at a lower traffic load (above 50 files/s) than the net-neutral scenario (above 200 files/s). Largely, non-neutrality (in the form of paid prioritisation) results in significantly poorer blocking probability performance than the net-neutral paradigm.

As can be observed in Figure 4, the Highest SINR scheme offers the best average file transfer delay performance. This is because MSs can connect to the complete set of ABSs and are usually assigned to the closest and highest SINR ABSs. Faster file transfer (and thus lower delay) is possible with this scheme relative to the other scenarios that permit switching off ABSs and lower SINR ABS choices. For traffic loads below 150 file/s, the average file transfer delay performance of the non-neutral scenario II is much better than that of the non-neutral scenario I, which is, in turn, better than the net-neutral scenario's case. An important reason is that the non-neutral scenario II has a higher percentage of premium MSs (60%), which require a higher SINR guarantee that leads to faster file transfer speed, than the non-neutral scenario I (with 20% premium users) and the net-neutral scenario (with no premium users). Hence, for users admitted into the system in the non-neutral scenario II, the average file transfer delay is lower than in the other two scenarios. For this same reason, the non-neutral scenario I with a slightly higher premium user population has a slightly better delay performance than the net-neutral scenario. For traffic loads above 150 files/s, the performances of the non-neutral scenarios

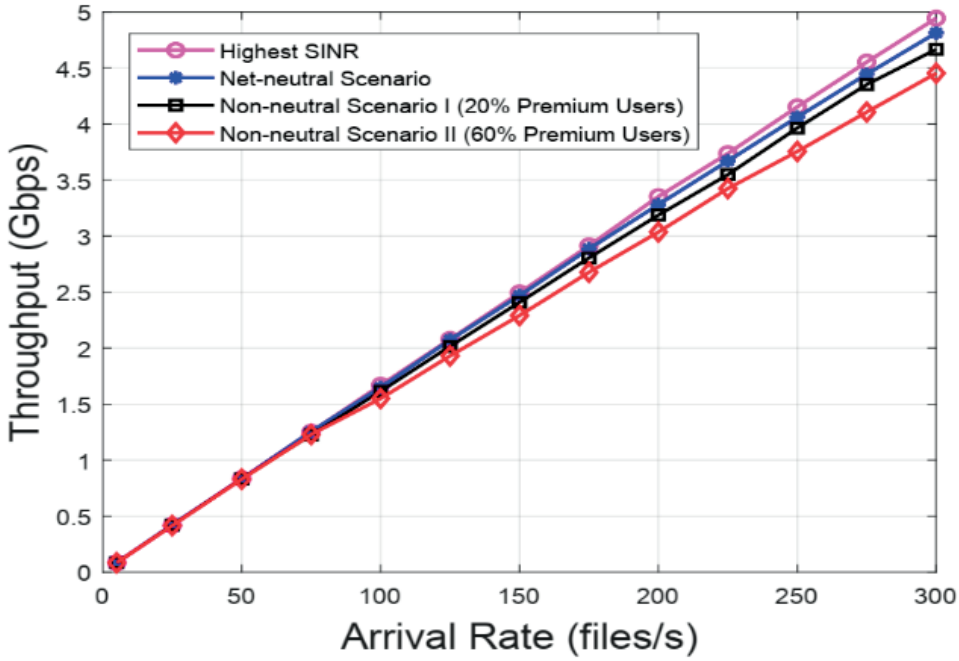

Figure 5 Comparison of Throughput of Different Scenarios

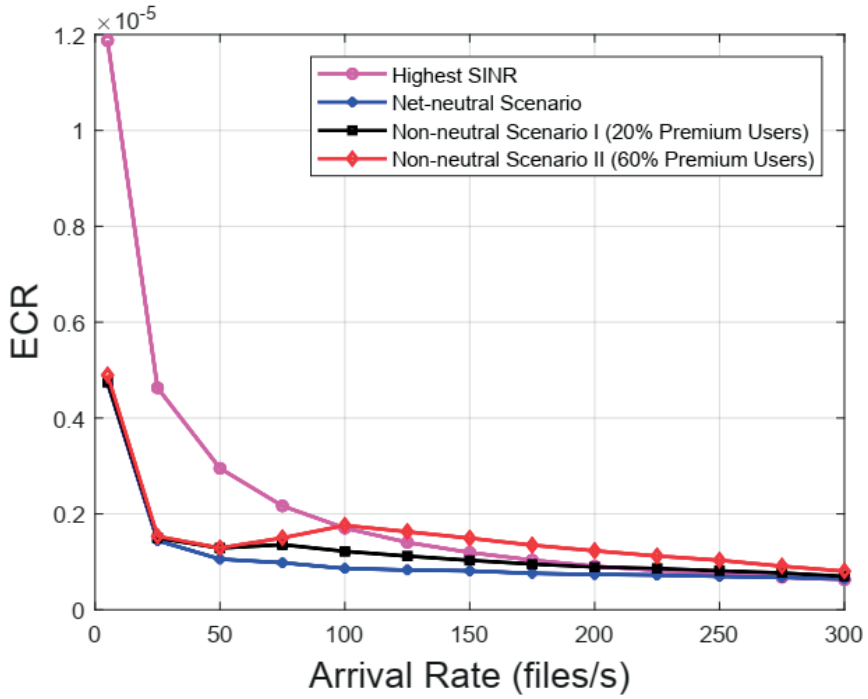

Figure 6 Comparison of Energy Consumption Rating of Different Scenarios

and the net-neutral scenario are quite similar. This can be attributed to the higher interference level in the network at higher traffic load levels, which results in more blocked MS requests and channel request re-attempts, and consequently, similar high average file transfer delays for these three scenarios. In general, a higher population of premium users in the network results in a lower (or better) average file transfer delay if ABSs are permitted to be switched off, especially at low traffic levels. Overall, non-neutrality leads to better average file transfer delay performance than the net-neutral method of allocating resources.

Figure 5 presents the throughput performance for the scenarios. It is important to note that retransmission overhead is considered negligible, as the resource allocation scheme proactively maintains all active links above the minimum SINR for the duration of their transmission by controlling interference from newly admitted users. The penalty for poor channel conditions or high interference is not retransmission overhead, but rather an increase in blocking probability or file transfer delay, as users must wait for suitable resources to become available. Since our throughput metric already excludes protocol overhead and only measures successfully transmitted information bits for admitted users, it serves as a direct and valid measure of goodput. Figure 5 reveals that all the scenarios have very similar throughput performances at low traffic loads (below 100 files/s). However, at higher traffic load levels (above 100 files/s), the Highest SINR scheme has the highest throughput, since it permits more MS to be admitted into the network due to the lower admission SINR than the non-neutral case and the lower blocking probability than the other scenarios as seen in Figure 3. The net-neutral scenario has the next best throughput performance and this can be linked to the lower blocking probability than the non-neutral scenarios (see Figure 3) resulting from QoS guarantee for all users (0 premium users) being just 1.8 dB, thus enabling admission of more users into the network than the non-neutral scenarios (with some premium users). The non-neutral scenario II has the worst throughput performance as it has the highest premium user population and thus more users must meet the higher admission SINR condition of 14 dB and maintain this level of SINR at all times. However, at high traffic load levels with the interference in the network being high, it will be challenging to provide the high QoS guarantees for many premium users concurrently. Thus, a higher blocking rate and lower throughput relative to the non-neutral scenario I and the net-neutral scenario are the case for the non-neutral scenario II. Generally, the higher the premium user population, the lower the throughput, especially at higher traffic loads. In essence, non-neutrality leads to poorer throughput performance relative to the net-neutral approach.

Figure 6 shows the ECR performances for the different scenarios. ECR measures the energy consumed per bit for successfully transmitted data. Therefore, a high value of ECR indicates poor ECR performance and poor energy efficiency. Only user information-carrying bits were considered in the study; protocol overheads and error correction code bits were not considered. More so, users only begin to transmit files when a suitable channel becomes available, which is maintained at the user's guaranteed SINR or above until the file is successfully transmitted. Therefore, packet loss during a file transfer does not occur in this model. Instead, the negative effects often observed at high traffic loads are driven by contention at the access level. Severe congestion and high interference lead to a high volume of failed channel requests. These failed requests must be re-initiated by the user, which leads to higher blocking probability, increased file transfer delay as users wait longer for a channel, lower overall throughput, and consequently a higher ECR. In Figure 6, at low traffic load (below 100 files/s), the Highest SINR scheme has a significantly higher ECR value relative to the other scenarios. This can be attributed to the fact that throughput performances are similar for all scenarios at low traffic load (see Figure 5), and ABSs can be put to sleep to reduce energy consumption in the other scenarios, whereas under the Highest SINR scheme, all the ABSs are always on. However, at higher traffic load (above 100 files/s), the ECR performance of the Highest SINR scheme approaches those of the other scenarios. Moreover, it even outperforms the non-neutral scenario II at higher traffic levels. This is due to an increasingly lower opportunity to save energy by switching ABSs off under the other scenarios as traffic load increases and the lower throughput performances of these scenarios compared to that of the highest SINR scheme.

Furthermore, the ECR performances of the non-neutral scenarios are similar to the net-neutral scenario at low traffic intensity (below 75 files/s). This can be attributed to the similar throughput performances of the scenarios (see Figure 5) and their comparable energy consumption since a few MSs are active simultaneously at low load and interference is minimal.

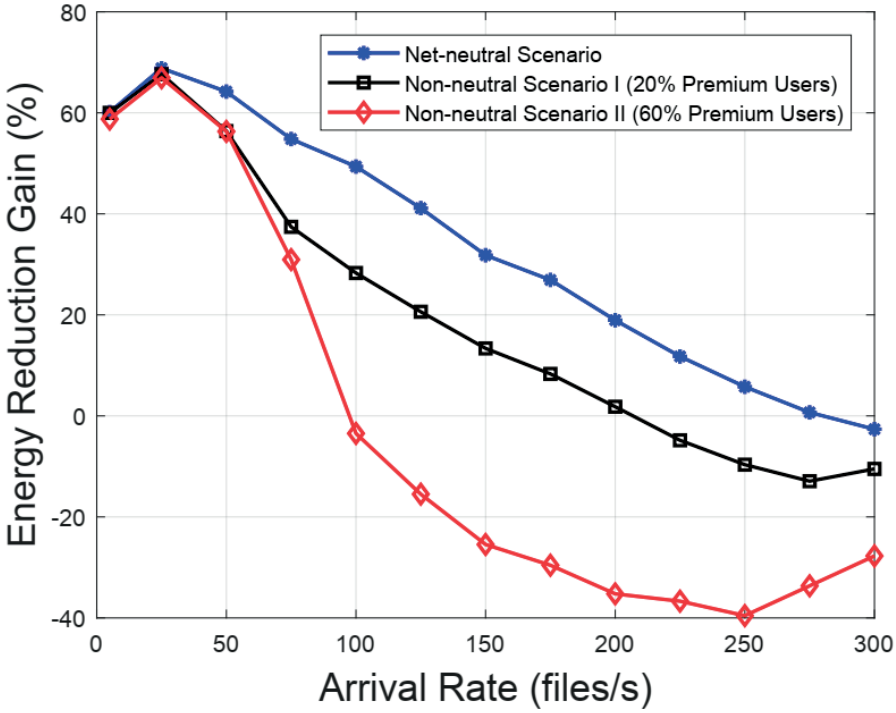

Figure 7 Comparison of Energy Reduction Gain of Different Scenarios

However, at higher traffic levels, the higher admission SINR and QoS guarantee of 14 dB for premium users (compared to 10 dB admission SINR and 1.8 dB QoS guarantee SINR for normal users) necessitates more ABSs to serve MS requests. Hence, the energy consumption of both non-neutral scenarios is higher than that of the net-neutral case and since the throughput of the non-neutral scenarios is lower than that of the net-neutral scenario, the ECR performance of the net-neutral scenario is, therefore, better than those of the non-neutral scenarios. The non-neutral scenario II with the lowest throughput performance has a poorer ECR performance than the non-neutral scenario I due to the higher premium user populations requiring more ABSs to be active and resulting in higher energy consumption.

Figure 7 displays the ERG performances for the non-neutral and net-neutral scenarios relative to the Highest SINR scheme. Just like the ECR performances, the ERG performances of the non-neutral scenarios and the net-neutral scenario are similar at low traffic loads (below 50 files/s). The net-neutral scenario achieves up to 69% ERG, while the non-neutral scenarios I and II result in up to 67.5% and 67% ERG, respectively, at low traffic loads. However, the net-neutral scenario achieves significantly better ERG than the non-neutral scenarios at higher traffic loads. Furthermore, the non-neutral scenario II has the worst ERG performance, just as observed under the ECR performances of the scenarios. The observed trend in the ERG performances of the scenarios is due to the difference in the SINR requirements for admission and QoS guarantees for the premium and normal users, and the corresponding difference in the number of active ABSs needed to handle MS requests, as explained earlier under the ECR performances. Based on the ECR and ERG performances, it can be concluded that non-neutrality leads to poorer energy efficiency than the conventional net-neutral approach.

5. PRACTICAL CONSIDERATIONS AND FUTURE WORK

The study presented in this paper focused on the performance of the non-neutral scenario (based on the PPCCR scheme) relative to the conventional net-neutral scenario using system-level simulations. Nevertheless, the consideration of the real-world deployment constraints and economic viability of the proposed PPCCR algorithm is also important. This will help to establish the possibility of utilising this algorithm in real networks.

In both 4G and 5G, a UE can transmit a Sounding Reference Signal (SRS) to a base station, which the base station will then use for channel quality estimation across a large bandwidth and subsequent uplink resource allocation decision (Chaki, Shikida, & Muraoka, 2024; Shin & Shin, 2016). This UE-assisted decision-making can be exploited in implementing the PPCCR algorithm in practical networks. In our scheme, each ZBS will run the PPCCR algorithm. Specifically, each ZBS will configure a newly arriving MS (the equivalent of UE in the studied architecture) to transmit a single SRS, which is then measured simultaneously by all the active ABSs in its zone. Each active ABS will then generate a multiple channel quality measurement report and send it to the ZBS via a logical interface like the X_2 interface in 4G or X_n interface in 5G.

Interference between neighbouring zones is managed through inter-ZBS communication over the same logical interface. The ZBS will utilise these detailed channel quality reports and neighbouring zone interference information to make uplink resource allocation decisions for the MS uplink request. The ZBS needs only this information from the ABSs and neighbouring ZBSs, other information (such as the load on ABSs and CCRs of ABSs) needed for the resource allocation decision is computed locally at the ZBS. Information exchange related to the channel quality report and interference between base stations can be carried out in real time with the logical interfaces implemented with fibre or millimetre wave. The additional request for ABSs to listen to SRS and report channel quality to ZBS can be implemented through software updates without significant modification to existing interface protocols.

A rigorous analysis of the PPCCR algorithm's complexity reveals it to be highly efficient. The dominant operations, such as information gathering (Step 2), SINR threshold check (step 4), highest SINR candidate selection (Step 6), and resource assignment (Step 7), all scale linearly with a complexity of $O(m \times t)$ for a zone with m ABSs and ABSs with t subchannels. Even the crucial interference check (Step 5) also has a complexity of $O(m \times t)$. Due to the system's frequency plan, the check is a constant-time operation $(O(1))$ for each candidate subchannel, as it only needs to check a small, fixed number of co-channel users – a maximum of 5 intra-zone and 48 inter-zone checks, as explained earlier in section 3.2. The algorithm's overall linear $O(m \times t)$ complexity is computationally trivial for modern 4G/5G

base stations with high-speed processors, confirming its scalability for real-time deployment.

The PPCCR algorithm is channel-aware and inherently exploits multi-user diversity. By evaluating all suitable subchannels for a user and selecting the one with the highest SINR (Step 6), the algorithm matches the user to the resource where their channel conditions are strongest. This process optimises the admitted user's performance and simultaneously avoids allocating that user to a channel that, while poor for them, may be ideal for a different user arriving later. This intelligent assignment leaves a wider array of channel conditions available for subsequent users, improving overall system efficiency. While this approach is effective, a valuable direction for future work is to extend the framework to include more explicit scheduling policies, such as proportional fairness, to further balance system-wide objectives.

Another important future direction for this work is the investigation of the economic aspects of paid prioritisation in cellular networks. This can include investigation of the premium pricing strategy that balances premium and normal users' QoS, analysis of customers' willingness to pay premium rates and return on investment gains, and assessment of the potential trade-offs between return on investment and QoS under different traffic levels or times of the day in a non-neutral energy-efficient network.

Furthermore, a detailed multi-parameter sensitivity analysis will result in a more robust assessment of the proposed scheme. Therefore, the future work will necessarily include the investigation of a wider range of key parameters. This includes varying the premium user population more continuously, exploring different SINR thresholds, and incorporating diverse user classes with various file size distributions. Such an extensive sensitivity analysis will provide deeper insights into the optimal configuration and performance gains/losses of the PPCCR algorithm beyond the preliminary results presented in this paper.

The results of the study presented in this paper reveal complex fairness dynamics that may not be obvious beforehand. An increase in overall system blocking probability with increasing premium user population was observed. This implies that the high SINR requirement for admitting premium users into the network results in a situation where premium users arriving earlier and admitted into the network negatively impact the chances of subsequent premium users and normal users in the non-neutral scenario relative to subsequent users in a net-neutral scenario. This reveals a complex fairness dynamics that is not simply about establishing parity between premium and normal users. The investigation of this complex fairness dynamics using well-established metrics, such as Jain's Fairness Index, is an important and necessary direction for future work.

As the first investigation into the effects of a non-neutral policy on the energy efficiency and QoS of a separation architecture, this study employs a simplified user and traffic model. The binary classification of "premium" and "normal" users, along with the uniform file size assumption, was utilised to establish a clear and foundational understanding of the core performance trade-offs. While this approach is a necessary starting point, real-world networks serve a multitude of service classes (e.g., VoIP, IoT, interactive gaming) with vastly different QoS requirements and traffic characteristics. A critical direction for future work is to build upon this foundational study by extending the PPCCR algorithm and our performance analysis to incorporate these more realistic, multi-tiered user models and diverse traffic profiles.

6. CONCLUSIONS

In this paper, we present a study that created and evaluated non-neutral, energy-efficient scenarios to determine the impact of non-neutrality on the QoS and energy efficiency of a separation architecture. Each non-neutral, energy-efficient scenario was created by integrating two key components. The first is a novel resource allocation scheme developed in the study, the Paid Prioritisation-based Clustering Capability Rating (PPCCR) scheme, which allocates resources based on user subscription fees. The second is a topology management scheme that reduces network energy consumption by switching base stations on and off according to traffic loads, idle waiting periods, and blocked requests. Two different versions of the non-neutral, energy-efficient scenario were implemented: the first, termed non-neutral scenario I, featured 20% premium users and 80% normal users, while the second, non-neutral scenario II, comprised 60% premium users and 40% normal users. The QoS and energy efficiency performance of these non-neutral scenarios was compared with that of a net-neutral scenario through system-level simulations. The simulation results revealed that non-neutrality (in the form of paid prioritisation) resulted in better average file transfer delay than the net-neutral paradigm; however, this was achieved at the cost of poorer blocking probability and lower throughput. Furthermore, while non-neutrality led to up to 67.5% ERG at low traffic load intensities, at higher loads, it led to significantly poorer energy efficiency relative to the net-neutral approach.

REFERENCES

3GPP. (2009). *TR 36.942: Evolved Universal Terrestrial Radio Access (EUTRA); Radio Frequency (RF) system scenarios version 8.2.0 Release 8*. Retrieved from http://www.etsi.org/deliver/etsi_tr/136900_136999/136942/08.02.00_60/tr_136942v080200p.pdf

3GPP. (2015). *Overview of 3GPP Release 12 V0.2.0*. Retrieved from https://www.3gpp.org/ftp/Information/WORK_PLAN/Description_Releases/

Agiwal, M., Kwon, H., Park, S., & Jin, H. (2021). A survey on 4G-5G dual connectivity: Road to 5G implementation. *IEEE Access, 9*, 16193-16210.

Arotiba, A. L. (2019). *Development of a non-neutral energy efficient radio resource management scheme for cellular networks* (Master's thesis). Obafemi Awolowo University, Ile-Ife, Nigeria.

Auer, G., Giannini, V., Godor, I., Skillermark, P., Olsson, M., Imran, M. A., . . . Blume, O. (2011, May). *Cellular energy efficiency evaluation framework*. Paper presented at the IEEE 73rd Vehicular Technology Conference (VTC Spring), Yokohama, Japan.

Baidowi, Z. M. P. A., & Chu, X. (2020). An optimal energy efficiency of a two-tier network in control-data separation architecture. *Journal of Communications, 15*(7), 545-550.

Budzisz, L., Ganji, F., Rizzo, G., Marsan, M. A., Meo, M., Yi, Z., . . . Wolisz, A. (2014). Dynamic resource provisioning for energy efficiency in wireless access networks: A survey and an outlook. *IEEE Communications Surveys & Tutorials, 16*(4), 2259-2285. doi: 10.1109/COMST.2014.2329505

Capone, A., Filippini, I., Gloss, B., & Barth, U. (2012, October). *Rethinking cellular system architecture for breaking current energy efficiency limits*. Paper presented at the Sustainable Internet and ICT for Sustainability (SustainIT), Pisa, Italy.

Chaki, P., Shikida, J., & Muraoka, K. (2024). *Resource allocation for sounding reference signal in 5G massive MIMO under channel aging*. Paper presented at the 2024 IEEE Globecom Workshops (GC Wkshps), Washington, DC, USA.

Choi, J. P., Jeon, D. S., & Kim, B. C. (2018). Net neutrality, network capacity, and innovation at the edges. *The Journal of Industrial Economics, 66*(1), 172-204.

Economides, N. (2017). A case for net neutrality. Retrieved from https://spectrum.ieee.org/a-case-for-net-neutrality

Fisusi, A., Grace, D., & Mitchell, P. (2013, June). *Energy efficient cluster-based resource allocation and topology management for beyond next generation mobile broadband networks*. Paper presented at the IEEE International Conference on

Communications Workshops (ICC), Budapest, Hungary.
Fisusi, A., Grace, D., & Mitchell, P. (2017). Energy saving in a 5G separation architecture under different power model assumptions. *Computer Communications, 105*, 89-104.
Gharakheili, H. H., Vishwanath, A., & Sivaraman, V. (2016). Perspectives on net neutrality and Internet fast-lanes. *ACM SIGCOMM Computer Communication Review, 46*(1), 64-69.
Han, C., Harrold, T., Armour, S., Krikdis, I., Videv, S., Grant, P. M., . . . Hanzo, L. (2011). Green radio: radio techniques to enable energy-efficient wireless networks. *IEEE Communications Magazine, 49*(6), 46-54.
Han, Y., Grace, D., & Mitchell, P. (2012, August). *Energy efficient topology management for beyond next generation mobile broadband systems*. Paper presented at the International Symposium on Wireless Communication Systems (ISWCS), Paris, France.
He, J., Loskot, P., O'Farrell, T., Friderikos, V., Armour, S., & Thompson, J. (2010, August). *Energy efficient architectures and techniques for Green Radio access networks*. Paper presented at the 5th International ICST Conference on Communications and Networking in China (CHINACOM), Beijing, China.
Hoydis, J., Kobayashi, M., & Debbah, M. (2011). Green small-cell networks. *IEEE Vehicular Technology Magazine, 6*(1), 37-43. doi: 10.1109/MVT.2010.939904
Jiang, T., Li, P., Liu, C., Khan, N., Grace, D., Burr, A., & Oestges, C. (2012). *BuNGee Deliverable: D4.1.2 Simulation Tool(s) and Simulation Results*. Retrieved from https://cordis.europa.eu/docs/projects/cnect/7/248267/080/deliverables/001-BuNGeeD412UoYv1022052012.pdf
Kyosti, P., Meinilä, J., Hentilä, L., Zhao, X., Jämsä, T., Schneider, C., . . . Milojević, M. (2007). *IST-4-027756 WINNER II D1.1.2. WINNER II channel models*. Retrieved from http://www.ist-winner.org/WINNER2-Deliverables/D1.1.2v1.1.pdf
Maillé, P., & Tuffin, B. (2019). *Neutral and non-neutral countries in a global internet: What does it imply?* Paper presented at the 16th International Conference on Economics of Grids, Clouds, Systems, and Services (GECON), Leeds, UK.
Maillé, P., & Tuffin, B. (2022). *From net neutrality to ICT neutrality*. Cham, Switzerland: Springer.
Maleki, A. D., & Abolhassani, B. (2014). New scheduling scheme for green communications in long term evolution networks. *IET Communications, 8*(14), 2438-2444. doi: 10.1049/iet-com.2013.0408
Mohamed, A., Onireti, O., Imran, M. A., Imran, A., & Tafazolli, R. (2015). Control-data separation architecture for cellular radio access networks: A survey and outlook. *IEEE Communications Surveys & Tutorials, 18*(1), 446-465.
Mukherjee, S., & Ishii, H. (2013, April). *Energy efficiency in the phantom cell enhanced local area architecture*. Paper presented at the IEEE Wireless Communications and Networking Conference (WCNC), Shanghai, China.
Ozturk, M., Abubakar, A. I., Nadas, J. P. B., Rais, R. N. B., Hussain, S., & Imran, M. A. (2021). Energy optimization in ultra-dense radio access networks via traffic-aware cell switching. *IEEE Transactions on Green Communications and Networking, 5*(2), 832-845.
Roth, Z., Goldhamer, M., Chayat, N., Burr, A., Dohler, M., Bartzoudis, N., . . . Bucaille, I. (2010, June). *Vision and architecture supporting wireless GBit/sec/km2 capacity density deployments*. Paper presented at the Future Network and Mobile Summit, Florence, Italy.
Sherif, A., & Haci, H. (2023). A novel bio-inspired energy optimization for two-tier wireless communication networks: A grasshopper optimization algorithm (GOA)-based approach. *Electronics, 12*(5), 1216.
Shin, E., & Shin, J. (2016, February). *Sounding reference signal measurement in LTE system*. Paper presented at the 18th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, South Korea.
Tan, K., Bremner, D., Le Kernec, J., Sambo, Y., Zhang, L., & Imran, M. A. (2022). Graph neural network-based cell switching for energy optimization in ultra-dense heterogeneous networks. *Scientific Reports, 12*(1), 21581.
Ternon, E., Agyapong, P., Hu, L., & Dekorsy, A. (2014, April). *Database-aided energy savings in next generation dual connectivity heterogeneous networks*. Paper presented at the IEEE Wireless Communications and Networking Conference (WCNC), Istanbul, Turkey.
Ternon, E., Agyapong, P. K., & Dekorsy, A. (2015). Performance evaluation of macro-assisted small cell energy savings schemes. *EURASIP Journal on Wireless Communications and Networking, 2015*(1), 1-23.
Turyagyenda, C., O'Farrell, T., & Guo, W. (2012). Energy efficient coordinated radio resource management: a two player sequential game modelling for the long-term evolution downlink. *IET Communications, 6*(14), 2239-2249.
Uhrina, M., Holesova, A., Bienik, J., & Sevcik, L. (2021). Impact of scene content on high resolution video quality. *Sensors, 21*(8), 2872.
Xu, X., He, G., Zhang, S., Chen, Y., & Xu, S. (2013). On functionality separation for green mobile networks: concept study over LTE. *IEEE Communications Magazine, 51*(5), 82-90.
Zhang, S., Gong, J., Zhou, S., & Niu, Z. (2015). How many small cells can be turned off via vertical offloading under a separation architecture? *IEEE Transactions on Wireless Communications, 14*(10), 5440-5453.
Zhao, T., Yang, P., Pan, H., Deng, R., Zhou, S., & Niu, Z. (2013, April). *Software defined radio implementation of signaling splitting in hyper-cellular network*. Paper presented at the Proceedings of the second workshop on Software radio implementation forum, Shanghai, China.
Zhisheng, Z., Guo, X., Zhou, S., & Kumar, P. R. (2015). Characterizing energy-delay tradeoff in hyper-cellular networks with base station sleeping control. *IEEE Journal on Selected Areas in Communications, 33*(4), 629-642. doi: 10.1109/JSAC.2015.2393494
Zhu, Q., Wang, X., & Qian, Z. (2019). Energy-efficient small cell cooperation in ultra-dense heterogeneous networks. *IEEE Communications Letters, 23*(9), 1648-1651.