Class-based Shared Resource Allocation for Cell-Edge Users in OFDMA Networks

Chetna Singhal, Satish Kumar, Swades De, Nitin Panwar, Ravindra Tonde, Pradipta De

Abstract

In this paper we present a new resource allocation scheme for cell-edge active users to achieve improved performance in terms of a higher system capacity and better quality-of-service (QoS) guarantee of the users, where we utilize the 2-dimensional resource allocation flexibility of orthogonal frequency division multiple access (OFDMA) networks. Here, the mobile stations (MSs) at the cell-edge can maintain parallel connections with more than one base station (BS) when it is in their coverage area. A MS, before handoff to a new BS, seeks to utilize additional resources from the other BSs if the BS through which its current session is registered is not able to satisfy its requirements. The handoff procedure is termed as *split handoff*. The BSs participate in split handoff operation while guaranteeing that they are able to maintain QoS of the existing connections associated with them. In this study, first, we present the proposed shared resource allocation architecture and protocol functionalities in split handoff, and give a theoretical proof of concept of system capacity gain associated with the shared resource allocation approach. Then we provide a differentiated QoS provisioning approach that accounts for the MS speed, its channel quality, as well as the loads at different BSs. Via extensive simulations in Qualnet, the benefits of the proposed class-based split handoff approach is demonstrated. The results also indicate traffic load balancing property of the proposed scheme in heavy traffic conditions.

Index Terms

Split handoff, shared resource allocation, effective capacity, differentiated QoS

C. Singhal is with the Bharti School of Telecom, Indian Institute of Technology (IIT) Delhi, New Delhi, India. S. Kumar was with the Bharti School of Telecom, IIT Delhi, New Delhi. Currently he is with Qualcomm India Pvt. Ltd., Hyderabad, India. S. De is with the Department of Electrical Engineering, IIT Delhi, New Delhi, India. N. Panwar was with the Bharti School of Telecom, IIT Delhi, New Delhi. Currently he is with Cisco Systems, Bangalore, India. R. Tonde was with the Department of Electrical Engineering, IIT Delhi, New Delhi. Currently he is with Samsung India Software Center, Noida, India. P. De is with IBM-India Research Lab, New Delhi, India.

1 INTRODUCTION

Steadily increasing data rate support along with the inherent advantages of wireless access networks, such as easy scalability and low cost of deployment and maintenance, have led to the emergence of broadband wireless access (BWA) as a popular alternative to the wire-line access infrastructure. The data rate landmarks in 4th Generation (4G) wireless broadband access networks, like LTE-A (Long Term Evolution-Advanced) and WiMAX-Mobile (World wide Interoperability for Microwave Access-Mobile), are set around 1 Gbps in downlink and 300 Mbps in uplink as per the IMT-Advanced (International Mobile Telecommunications -Advanced) specifications [1]. To achieve and maintain these very high rates in a wireless environment, mobile devices/stations (MSs) are required to change the base station (BS), if there exists one within the reach of the MS, with, for example, a better link quality. This procedure is called *handoff*. Handoff is performed on the basis of some metric threshold, which can be chosen as per the communication system requirements, application constraints of an individual MS, and speed.

In today's evolving wireless networks, the issues related to handoff are not treated as an isolated physical layer problem. The protocol end points that are located in the BS are needed to be moved from the source BS to the target BS. This relocation can be done in two different ways: (a) protocol status transfer from the source BS to the target BS; (b) protocol reinitialization after the handoff. In LTE networks, the relocation is done using a hybrid approach, where the downlink protocol status is transfered from the source BS to target BS [2], [3]. A packet forwarding approach from the old BS to the new BS has also been considered for the in-flight packets, whose impact on the user connection has been studied in [3]. For uplink traffic, random access procedure is performed at the target BS. The results in [3] showed that the extra load caused by the forwarding is not significant. Also, the impact of forwarding on the end user connections can be reduced by using a scheduling architecture at the transport layer that is able to differentiate among different service classes while allocating the bandwidth.

Maintaining guaranteed quality-of-service (QoS) support at the cell-edge and dimensioning a mobile wireless network are the two major challenges, especially when the user rate demands are high. Therefore, efficient scheduling of BS resources is essential to ensure fairness to the users and high network performance. Many scheduling techniques have been proposed in the literature which take care of user-level fairness and network capacity [4], [5], [6], [7]. The problem aspect

of capacity and fairness guarantee however changes when a MS cannot hold on to the existing connection and decides to perform a handoff, or when a macro-diversity approach is adopted to mitigate the disconnection and loss of packets in downlink and uplink directions.

1.1 Related works

In view of high data rate support, several advanced handoff approaches have been proposed in the recent literature. The handoff process is of two major types: hard handoff and soft handoff. Different variants of hard handoff and soft handoff are semi-soft handoff and fractional handoff in OFDM based systems [8], [9], fast base station switching (FBSS) [10], and macro-diversity handoff (MDHO) [10]. Before performing handoff, an appropriate BS candidate must be chosen and then the handoff procedure should be continued based on the current technology and the specific application constraints of the MS. The exact procedures vary depending on the used technology, and usually within the technology several alternatives are available as well.

A routing based seamless handoff was proposed in [11], where layer-2 and layer-3 handoff progress simultaneously to minimize the handoff delay and packets loss. A packet access router was introduced, which connects to both serving BS and target BS. After the MS sets up a new connection with the target BS, the router starts sending duplicate packets to a target BS. Upon notification of the last packet sent from the serving BS, the target BS starts transmitting the onward packets to the MS. In [12], a fast handoff scheme was proposed for real-time applications. Here, the target BS receives QoS parameters and the channel identity (CID) from the serving BS through the backbone network. The target BS uses these old CIDs for downlink transmission until new CIDs are assigned at the target BS. The arriving packets during the handoff interruption time are buffered at the serving BS and sent to the target BS over the backbone. In a handoff scheme for downlink traffic in CDMA (code division multiple access) systems, called spatial multiplexed soft handoff [13], each participating base station sends only a subset of the main data stream, and the MS reassembles them and restores the main data stream. The motivation was to improve the processing gain of the system. The limitation of this approach in CDMA systems is due to the receive power disparity from the participating BSs.

As suggested in [14], depending on the fading conditions, the cell coverage overlap area where a MS can have access to more than one BS, can extend up to 47% of the total area of the two adjacent cells. Thus, in the handoff zone, the signals from two or more BSs can be exploited even

if the MS is traveling at a high speed, because the coverage overlap region is sufficiently large. To exploit this genuine coverage overlap, simulation-based studies were independently carried out in [15] and [16] for downlink cell-edge users. Their results suggested that, the rate supported to a cell-edge user with resources from multiple BSs is not always better than the normal transmission from a single BS. To this end, we argue that, while evaluating such a possibility, the cell loads and the individual user QoS should be considered while exploiting the signal from more than one BS, because the user QoS will define the need for multi-cell transmission, and the load on the BSs will decide the feasibility and limits of such transmissions.

Interference management in OFDMA networks with dynamic fractional frequency reuse (FFR) and interference aware power control was approached in [17], where the handoff scheme employs disjoint links and beacon. This approach however cannot reduce the delay in handoff. Also, FFR limits the overall achievable system capacity. FFR, interference coordination, and interference cancellation schemes are also discussed in [18], [19] with a system capacity viewpoint. As the new services with diversified QoS requirements are evolving, we need to address individual user QoS along with the other pre-existing constraints.

To enhance user QoS and network capacity performance, besides simple data rate guarantee traffic classification is also important. Explicit differentiated QoS support [20] was proposed for Internet applications to address traffic class-specific resource allocation, where the traffic was categorized into the three specified DiffServ model classes, namely, expedited forwarding (EF), assured forwarding (AF), and best effort (BE). There have been several recent studies on explicit QoS support to the BWA users. A set of works, e.g., the studies in [21], [22], [23], [24], [25], addressed QoS support and service differentiation to the wireless users, which focused on single BS access scenario. In [23], physical signal strength and mobility information were exploited for a better support to MDHO and FBSS. A comparative protocol-level study on vertical handoff across different wireless technologies was presented and differentiated QoS support was discussed in [24]. These studies provide a less detailed account of service differentiation approach. Also, they lack on addressing the scheduling issues when resources are available from multiple BSs.

On class based dynamic resource allocation there have been some recent studies in Ethernet passive optical networks. The dynamic bandwidth allocation strategy in [26] assigns a fixed bandwidth to the EF (e.g., voice) traffic, irrespective of the immediate requirements. The leftover bandwidth is allocated to the AF (e.g., video) traffic first, and then to the BE (data) traffic. The

strategy proposed in [27] limits the allocation to EF and AF traffic to their respective service level agreements (SLAs), while assigning the remaining bandwidth to BE traffic. To achieve fairness among all classes, the bandwidth allocation in [28] is done in three stages: first allocate resource proportional to the queue length of all classes, then prune the allocated resource if it exceeds the respective SLA high value or SLA low value, and finally allocate the excess bandwidth proportionally to all queues. To guarantee strict priority and optimal resource usage, a further modification was suggested in [29] which also addressed traffic burstiness dependent optimal prediction of future traffic. We note that, while the studies in optical access domain provide some basis for differentiated resource allocation, they do not deal with channel rate degradation at the cell-edge and the possibility of shared resource usage from multiple cells.

1.2 Motivation and contribution

Intuitively, resource sharing among the BSs can be beneficial to the cell-edge users to mitigate packet loss, delay, as well as radio link failure during handoff. However, the prior studies did not simultaneously account for the user level parameters, such as QoS and speed, and the network level parameters, such as load distribution among the BSs. To increase the capacity while improving the QoS performance of the active cell-edge users, in this paper we present a new handoff scheme, which we call *split handoff*, in an OFDMA (orthogonal frequency division multiple access) cellular network. The proposed strategy is applicable to the cellular systems with universal frequency reuse (UFR) plan (i.e., with frequency reuse factor $= 1$) [30], as prevalent in LTE and WiMAX systems, as well as with partial frequency reuse plan (with frequency reuse factor < 1). The two dimensional flexibility (in frequency and time domains) of resource allocation of OFDMA systems is utilized to allocate resource to cell-edge users from more than one BS. This allocation from multiple BSs is called *sharing of resources*.

Note that, in contrast with the conventional soft handoff schemes, such as MDHO in WiMAX or the soft handoff variants in CDMA standards, the proposed split handoff scheme does resource sharing with different principles. (a) Instead of *replicating the packets* via the participating BSs to the mobile user, in split handoff the packet stream is *appropriately divided* among the participating BSs to the mobile user. (b) The packet stream splitting principle in split handoff takes care of the link qualities, which allows the participating BSs to choose/adapt the respective modulation and coding schemes independently. (c) The split handoff further considers the BS

loads while dividing the traffic stream across the participating BSs, leading to a more effective traffic load balancing. It may also be noted that, MDHO is not generally considered as an efficient alternative, as it is known to have a high bandwidth resource overhead [31, Ch. 15] due to packet replication and diversity combining from multiple BSs for the MSs in the overlapping coverage region, resulting in a poorer network capacity.

The proposed traffic splitting implementation approach as well as its objectives are different from those in [13]. Unlike in a CDMA based system, traffic splitting via more than one BS to one MS in an OFDMA based cellular access networks, having sub-carrier based frequency and slot based time resource allocation, is a distinctive challenge, especially due to inter-cell interference. Also, in contrast with the goal of increased signal distance in [13], we aim at network capacity gain and user QoS improvement, where load sharing among the BSs is a function of individual cell load and QoS requirement of the individual users.

One of the main contributions of this work is the system architecture for shared resource usage along with the functional details to enable split handoff design. The challenge is to construct a DL-map (downlink sub-carrier frequency and time slot allocation map) at the controller, which manages access between BSs and MSs. Typically, a DL-map contains information associated with a single BS. However, in split handoff design, a MS can communicate with multiple BSs simultaneously. Therefore, the DL-map construction in split handoff must take this into account.

We also present an analytical model to capture capacity gain using the proposed split handoff scheme. To maximize the performance at the cell-edge, the effective capacity concept in a single cell scenario [32] is extended to the proposed shared resource allocation policy. The analytically predicted capacity gain performance is verified via rigorous simulations.

Finally, we provide a framework for shared resource allocation to differentiated service classes (voice, video, and data), which is implemented in Qualnet simulator using WiMAX and the other necessary toolboxes. Performance of the proposed scheme is compared with the standard non-shared resource allocation policy (*hard handoff, with a priori association and authentication with the target BS, i.e., with FBSS option*) as well as with MDHO. We show that, the proposed resource sharing is quite beneficial in terms of capacity gain at the cell edge, especially for the users with stringent QoS requirements. We also show the impact of network load and user speed on the users' QoS performance. It may be highlighted that, the gain in the proposed scheme is achieved without requiring extra network resources, such as power and bandwidth.

1.3 Paper organization

The rest of the paper is presented as follows. In the next section, the proposed system model and split handoff activities are described. Section 3 deals with analysis of user QoS related capacity gain using the proposed split handoff technique. Section 4 presents a practically implementable framework for shared resource allocation to class-based users. Section 5 provides the numerical and simulation results. The paper is concluded in Section 6.

2 System Model

We consider a BWA system built on OFDMA based physical layer technologies, as in LTE or WiMAX standards. The adjacent BSs with overlapping coverage areas can operate on the same frequency band (i.e., with frequency reuse factor $= 1$), as in LTE and WiMAX systems, or they can have different frequency bands. The resources can be shared by a mobile user from multiple BSs based on the user requirements, network resource availability, and coverage conditions. This resource sharing eventually results in handoff when the condition of coverage and network resource availability for sharing are not met, or when the user does not need the sharing of resources. The resulting handoffs can be called *in-handoff* or *out-handoff*. In-handoff occurs when a user starts sharing from multiple BSs but later comes back to its parent BS. Similarly, the out-handoff takes place when the user starts sharing and later it goes to the other BS before the call ends. Before either of these handoffs, since the total resource to a user is 'split' between the neighboring BSs (based on certain metrics, such as load of the respective BSs, the user's QoS requirement, etc.), we call these handoffs combinedly as *split handoff*.

2.1 The proposed system architecture

Keeping in mind the flexibility of the proposed scheme we present a new transport layer queuing system model, where a two-level queuing is employed, as depicted in Fig. 1, for all active users to reduce the impact of user movement (in/out-handoff) on the user connection. Here, the node architecture and its interaction with other network entities is presented. Queuing is applied at the data link control (DLC) layer as well as the transport control (TCP) layer, to distribute the traffic to the BSs which are participating in data transmission of the users in the shared region.

In this paper, scheduling discussions are restricted to the downlink traffic only, although the scheduling principle applies to uplink traffic as well with some modifications in the control

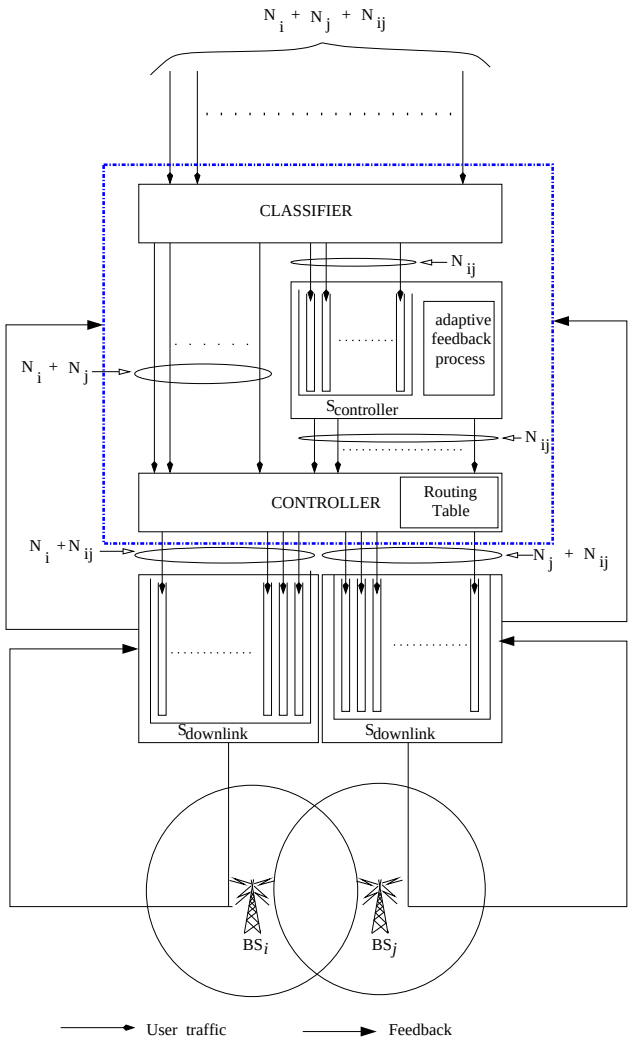

Fig. 1. System architecture, with different downlink queue structure for the shared and non-shared users.

message exchange. Also, in Fig. 1 as well as in the following discussions, we have considered the examples of resource sharing between two BSs, which is extensible to three or more BSs.

Two BSs, BS_i and BS_j are shown in the figure with an overlapping coverage. There are N_i users which are served only by BS_i, N_j users served only by BS_j, and N_{ij} users served by both BS_i and BS_j in shared mode. BS_i maintains queues for $N_i + N_{ij}$ users and serves them using its $S_{downlink}$ scheduler. Likewise, BS_j maintain queues for $N_j + N_{ij}$ users and serves them using $S_{downlink}$. Here we assume that one user has only one class of service at a time. If a single user maintains multiple parallel connections with different QoS requirements, then the scheduling can be easily handled by additional queues, called 'priority queues,' at a BS. A controller directs

the flows from the classifier according to the routing table maintained therein. Controller and classifier are the two logical entities which can be physically co-located. Based on the feedback from the BSs, the classifier is used to distinguish the incoming/outgoing flows if they are of a shared user - served by two BSs, or a non-shared user - served by only one BS. The controller also maintains the queues for all users which can be served by both BSs. Flow scheduling at the controller is according to the rule provided by $S_{controller}$. The parameters considered for splitting of traffic are fed back to the controller using feedback links.

Some of the advantages of the proposed architecture are: (i) centralized routing information maintenance for the subscribers to create multiple parallel connections when necessary; (ii) avoidance of packet duplication, by distributing packets for a cell-edge user across the BSs, thereby minimizing resource wastage; (iii) rule based splitting of traffic by using scheduler $S_{controller}$; (iv) possibility of resource allocation based on traffic classification.

2.2 System functionalities

The controller in the proposed split handoff is connected to the BSs via high-speed wireline or wireless links. Beyond signal transmission-reception over the radio links, the BSs have a very little role to play. Functionality-wise, a controller will perform some extended tasks beyond a conventional BSC (base station controller) or a RNC (radio network controller). The specific activities of a controller in split handoff are: (i) construction of universal DL-map and broadcasting to all BSs, and (ii) scheduling and traffic load balancing by accounting the CINR (carrier-to-interference-and-noise ratio) at the MS from the connected BS and the neighboring BSs, available resources of the neighboring BSs, and QoS requirements of subscribers. The participating BSs are assumed synchronized through the controller.

We have used the following terms and assumptions in this paper. *Primary BS* (PBS) is the BS with which a MS exchanges the management messages as well as data. *Secondary BS* (SBS) is a BS with which the MS exchanges only data. Following the WiMAX standard notations for channel usage, the downlink interval usage code (DIUC) used by a MS with the PBS is denoted as DIUC1, and the DIUC used by a MS with the SBS is denoted as DIUC2. As indicated in the proposed system architecture (Section 2.1), traffic splitting is done at the transport layer. The controller stores the BS IDs and their associated loads. With respect to a particular MS, it stores the MS ID, its MAC address, PBS ID, DIUC1, priority calculated based on the service flow

QoS parameters, and SBS ID and DIUC2 - in case the MS is in contact with two BSs.

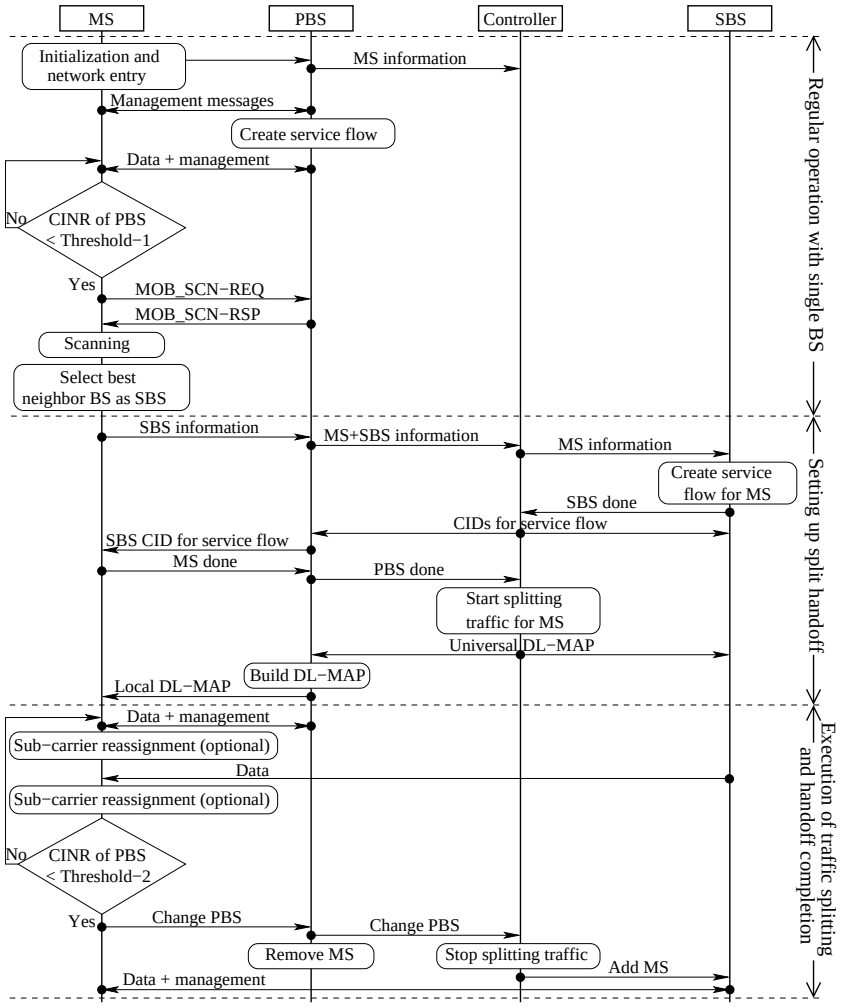

Fig. 2. Timing diagram of a MS session and split handoff process, with only the downlink traffic considered.

The timing diagram of a MS session and a split handoff process is shown in Fig. 2, in which only downlink data traffic is considered. The *initialization procedure* is similar as in a standard service flow set up. The MAC information and DIUC1 of the MS is passed to the controller at the network entry phase. During data and management message exchange with the PBS, the MS sends scanning request (MOB_SCN-REQ) to the PBS when the CINR from the PBS falls below Threshold-1 (which corresponds to a higher-than-the-lowest modulation and coding rate). If the response (MOB_SCN-RSP) from the PBS is positive, the MS starts scanning for a SBS and synchronizes with one, initiating the *split handoff set up* phase. Otherwise it continues its

connection with the PBS only. During scanning, the MS sends (via the PBS) the SBS information (BS ID, DIUC2, traffic load) to the controller. Subsequently, a new CID for data connection with the SBS is created by the controller, which is forwarded to the MS. With the ongoing CID for the PBS and the new CID for the SBS, the *user-level split handoff procedure* starts. The timing information of the bursts (slots) for connecting to the PBS and SBS is notified to the MS by the controller via a universal DL-map. To address user QoS and cell load imbalance, the controller accounts for the QoS priority, buffer status at the MS, and the BS traffic load at the time of burst scheduling. The burst timings for the PBS and SBS are separated within a frame such that the sub-carrier frequency reassignment latency of the MS is sufficiently accommodated. Note that, with frequency reuse factor $= 1$, it may be possible that the MS is connected to the BSs at different time slots over the same assignment of sub-carriers, in which case no reassignment is necessary. On the other hand, in case the adjacent BSs operate at different carrier frequencies, i.e., with frequency reuse factor < 1, split handoff involves sub-carrier reassignment latency. Finally, when the CINR from the PBS falls below Threshold-2 (corresponding to the lowest allowable data rate), a PBS change request is sent to the controller, and at that point the SBS assumes the responsibility of PBS for the MS. This process marks the *end of split handoff*.

3 Analytical Model for Shared Resource Allocation

In this section, we present an analytical framework to develop an intuition of capacity gain via shared resource usage by the cell-edge users. Since hard handoff is the default scheme in LTE/WiMAX standards, we consider this scheme as the baseline. Later, in Section 5 we will demonstrate via numerical and rigorous system simulation studies the capacity gain as a function of user QoS, where the relative performance with respect to soft handoff as well is considered.

3.1 QoS and effective capacity

Theoretically, QoS is defined by the maximum tolerable delay D_{max} for a user (traffic type) beyond which the delay violation probability exceeds a predefined threshold ϵ [32], i.e.,

$$\sup_t \Pr\{D(t) \geq D_{max}\} \leq \epsilon.$$

It was also shown in [32] that, for a dynamic queuing system, where the arrival and service processes are stationary and ergodic, the probability that the delay at time t, $D(t)$ exceeds the

threshold D_{max} can be accurately given as:

$$\sup_{t} \Pr\{D(t) \geq D_{max}\} \approx \Upsilon(\Omega) e^{-\theta(\Omega) D_{max}}, \tag{1}$$

where $\theta(\Omega)$ is a function of constant source rate Ω, D_{max} is a sufficiently large quantity, and $\Upsilon(\Omega)$ is the probability that the delay of a particular packet is non-zero, i.e., $\Upsilon(\Omega) = \Pr\{D(t) > 0\}$ at a randomly chosen time instant t. Here, $\theta(\Omega) > 0$ in (1) is a parameter describing the exponential decay rate of probability of QoS violation. $\theta(\Omega)$ is referred as the QoS exponent. A large value of $\theta(\Omega)$ corresponds to a fast decaying rate, i.e., a *stringent* QoS requirement, and a small value of $\theta(\Omega)$ corresponds to a slow decay rate, i.e., a *loose* QoS requirement.

The effective capacity for a given QoS exponent θ specifies the maximum constant arrival rate that can be supported by the system at the link layer.[1] Effective capacity concept in [32] can be applied to wireless channels with arbitrary physical-layer characteristics. For a discrete-time stationary and ergodic service process with rate $\mu(n)$ and channel service rate $R(m) = \sum_{n=0}^{m} \mu(n)$, the effective capacity is defined as:

$$E_C(\theta) \stackrel{\text{def}}{=} - \lim_{m \to \infty} \frac{1}{\theta m} \ln E\{e^{-\theta R(m)}\}, \tag{2}$$

where m is the block length. For uncorrelated block fading channels, where the service process $\{\mu(n), n = 1, 2, \cdots\}$ is also uncorrelated, the expression (2) is reduced to $E_C(\theta) = -\frac{1}{\theta} \ln E\{e^{-\theta \mu(n)}\}$, given any $n = 1, 2, \cdots$. We further normalize the effective capacity with respect to the frame length T_f and system bandwidth B as:

$$E_C(\theta) = -\frac{1}{\theta T_f B} \ln E\{e^{-\theta \mu}\}, \tag{3}$$

where the product $T_f B$ is the total time-frequency resources available in one frame $S = T_f B$.

3.2 Scheduling of shared users for maximizing capacity

An interesting saturation condition is where the total resource demand is more than the available resources at the BS such that $\sum_{u=1}^{N_i} S^u = S_i$, where S_i is the total resource available per frame in BS_i, and $(0 \leq S^u \leq S_i)$ is the resource allocated to the user u from BS_i in order to maintain

1. Here, dependency of θ on the source rate Ω is not shown for simplicity. In rest of the paper we will follow the same convention unless otherwise mentioned.

its QoS demand. For simplicity without loss of generality we assume, the total resource available at each BS in a cluster with UFR plan are equal, i.e., $S_i = S \ \forall \ i \in N_c$.

From (3), for a user u scheduled from BS_i, the effective capacity can be expressed as:

$$E^u_{C,i}(\theta^u) = -\frac{1}{\theta^u S} \ln E\{e^{-\theta^u \mu^u_i}\}, \tag{4}$$

where θ^u is the QoS exponent of user u and $\mu^u_i = r^u_i S^u$ is the rate provided to user u from BS_i. Here r^u_i is the modulation index of user u from BS_i.

Now, if the same user is scheduled from two BSs, e.g., BS_i and BS_j, then the total effective capacity, which we term as *joint effective capacity* $E^u_{C,joint}(\theta^u)$, is defined as:

$$\begin{aligned} E^u_{C,joint}(\theta^u) &= -\frac{1}{\theta^u S} \ln E\{e^{-\theta^u \mu^u_{i(1)}}\} - \frac{1}{\theta^u S} \ln E\{e^{-\theta^u \mu^u_{j(2)}}\}, \\ &= -\frac{1}{\theta^u S} \ln \left[E\{e^{-\theta^u \mu^u_{i(1)}}\} E\{e^{-\theta^u \mu^u_{j(2)}}\} \right], \end{aligned} \tag{5}$$

subject to the condition that the joint resources from BS_i and BS_j are the same as in (4) and the CINR is above the acceptable threshold γ_{th}. Stated mathematically: $S^u = S^u_{i(1)} + S^u_{j(2)}$ with the conditions $\{\gamma_i, \gamma_j\} > \gamma_{th}$. Here, $\mu^u_{i(k)}$ indicates the kth part of the rate achievable from BS_i for user u. Similarly $S^u_{i(k)}$ indicates the kth part of the resources allocated from BS_i to user u.

Let us denote $p_i = \Pr\{\gamma^u_i \leq \gamma_{th}\}$ and $p_j = \Pr\{\gamma^u_j \leq \gamma_{th}\}$. Then,

$$E\{e^{-\theta^u \mu^u_{i(1)}}\} = e^{-\theta^u \mu^u_{i(1)}}(1 - p_i) + p_i,$$
$$E\{e^{-\theta^u \mu^u_{j(2)}}\} = e^{-\theta^u \mu^u_{j(2)}}(1 - p_j) + p_j.$$

For simplicity of the subsequent expressions, the modulation indices for the user u from the two BSs are assumed equal. That is, $r^u_i = r^u_j \equiv r$. Hence, the expression (5) reduces to:

$$E^u_{C,joint}(\theta^u) = -\frac{1}{\theta^u S} \ln \left[\{e^{-\theta^u r S^u_{i(1)}}(1 - p_i) + p_i\} \cdot \{e^{-\theta^u r S^u_{j(2)}}(1 - p_j) + p_j\} \right], \tag{6}$$
$$\text{s.t.} \quad S^u_{i(1)} + S^u_{j(2)} = S^u \quad \text{and} \quad S^u > S^u_{i(1)}, S^u_{j(2)} > 0.$$

Note that, (4) is the effective capacity when a user is scheduled from only one BS, i.e., from BS_i, whereas (6) is the effective capacity when a user is scheduled from two BSs, i.e., from BS_i and BS_j. To maximize $E^u_{C,joint}$ in (6), first we substitute $S^u_{i(1)}$ with $S^u - S^u_{j(2)}$ and denote:

$$\{e^{-\theta^u r (S^u - S^u_{j(2)})}(1 - p_i) + p_i\} \cdot \{e^{-\theta^u r S^u_{j(2)}}(1 - p_j) + p_j\} = \Lambda(S^u_{j(2)}). \tag{7}$$

By differentiation of (7) with respect to $S^u_{j(2)}$ and equating it to zero, we have,

$$S^u_{j(2)} = \frac{S^u}{2} + \frac{S^u}{2\theta^u r} \ln \left[\frac{(1-p_j)/p_j}{(1-p_i)/p_i} \right], \tag{8a}$$

$$S^u_{i(1)} = \frac{S^u}{2} + \frac{S^u}{2\theta^u r} \ln \left[\frac{(1-p_i)/p_i}{(1-p_j)/p_j} \right]. \tag{8b}$$

By double differentiation of (7), it can easily be proved that the obtained values of $S^u_{i(1)}$ and $S^u_{j(2)}$ maximize the joint effective capacity in (6).

When the MS is in the coverage region of both the BSs, equations (8a) and (8b) give the resource allocation from them, so that the total effective capacity can be increased.

Numerical and simulation results on capacity gain with shared resource usage in a single class traffic environment will be discussed in Section 5.

4 Class-based Shared Resource Allocation

To implement a class based resource allocation policy, we learn from the dynamic bandwidth resource allocation policies in optical access networks [26], [27], [28], [29]. However, the proposed allocation policy in BWA is additionally influenced by the variability of available bandwidth, shared BS resource usage by the cell-edge users, and dynamic cell load conditions.

Although the current standard service differentiation approaches suggest to divide the user traffic into five service classes (e.g., in [10]), for a proof of concept service differentiated shared resource usage, we categorize the user traffic into three classes: P_0 (voice packets), P_1 (video traffic), and P_2 (data traffic) [20]. P_0 traffic is the most delay sensitive, requiring a guaranteed channel bandwidth, P_1 has a higher delay flexibility but requires a minimum bandwidth (rate) guarantee. P_2 traffic has neither delay nor bandwidth guarantee constraints. The proposed priority allocation approach can be easily extended to more number of classes.

We consider time frames of fixed length T_f seconds for downlink resource scheduling. Let, at any instant there be N_i users within the coverage region of BS_i only, N_j users which are within the coverage region of BS_j only, and N_{ij} users in the overlapping coverage region of BS_i and BS_j both. P_0 and P_1 traffic have their respective service level agreements (SLAs) which are the respective upper limits of resource that can be allocated to them. Since P_0 traffic is most delay sensitive, first resources are allocated to this traffic class. Then the resources are allocated to P_1 type of traffic in the first phase. Remaining resource at each base station are calculated as excess resource which is allocated to P_1 traffic in second phase and to the P_2 traffic.

For resource allocation to the mobile users in a class P_c, $c \in \{0, 1, 2\}$, we successively select the user with maximum scheduling function ψ^u [33], where, $\psi^u = \max\left\{\frac{\rho^u}{\tau^u}\right\}$, ρ^u is current bit rate of the user u (based on its channel conditions), τ^u is the user throughput to ensure fairness. To maximize throughput, ρ^u will ensure that the users with best channel conditions is selected while τ^u will ensure that no user experiences starvation. The users with high ρ^u and/or low τ^u will be selected, hence ensuring fairness.

4.1 Predicted resource allocation

Time-frequency resource allocation for the mobile users with different service classes is influenced by the statistical characteristics of the traffic arrival and the channel behavior as well as the usage of the backlog history. To predict the resource required due to the incoming traffic over a frame interval T_f, we adopt a linear predictor [29], [34]:

$$\tilde{S}_{P_c}^{u(\nu)}(n+1) = \sum_{l=0}^{\mathcal{L}_{P_c}-1} \xi_{P_c,l}^{u}(n) S_{P_c}^{u(\nu)}(n-l),$$

where $c \in \{0, 1, 2\}$ and $\mathcal{L}_{P_c}$ is the prediction order - a function of traffic type P_c. $\tilde{S}_{P_c}^{u(\nu)}$ is predicted resource requirement for user u and traffic type P_c due to new arrivals over the interval T_f. $\xi_{P_c,l}^{u}$ is the parameter indicating the impact of the actual resource requirement $S_{P_c}^{u(\nu)}(n-l)$ due to new arrivals in frame $(n-l)$ on the predicted resource requirement $\tilde{S}_{P_c}^{u(\nu)}$ for user u and priority type P_c. $\xi_{P_c,l}^{u}$ is updated by standard least mean square (LMS) algorithm as [34]:

$$\xi_{P_c,l}^{u}(n+1) = \xi_{P_c,l}^{u}(n) + \eta_{P_c}^{u}(n) \frac{\varepsilon_{P_c}^{u}(n)}{S_{P_c}^{u(\nu)}(n)},$$

where $\varepsilon_{P_c}^{u}(n)$ is the prediction error in the nth frame, defined as: $\varepsilon_{P_c}^{u}(n) = S_{P_c}^{u(\nu)}(n) - \tilde{S}_{P_c}^{u(\nu)}(n)$, and $\eta_{P_c}^{u}(n)$ is defined as: $\eta_{P_c}^{u}(n) = \frac{\mathcal{L}_{P_c}}{\sum_{l=0}^{\mathcal{L}_{P_c}-1}\left[S_{P_c}^{u(\nu)}(n-l)\right]^2}$. With the predicted new arrivals, the requested resource $S_{P_c}^{u(r)}(n+1)$ for frame $(n+1)$ and P_c type traffic is

$$S_{P_c}^{u(r)}(n+1) = S_{P_c}^{u(q)}(n) + \tilde{S}_{P_c}^{u(\nu)}(n),$$

where the superscript q indicates the resource required due to the queued traffic.

Note that, it is important to choose an optimum number of taps $\mathcal{L}_{P_c}$ for a given traffic type P_c that would maximize the quality of prediction $\tilde{S}_{P_c}^{u(\nu)}$ using the historical prediction. As studied in [29], while a higher value of $\mathcal{L}_{P_c}$ would increase the prediction accuracy by sharply tracking

the traffic burstiness, it causes a higher lag in predicted traffic - which can also be detrimental to the prediction quality. On the other hand, a very small value of $\mathcal{L}_{P_c}$ may not closely track the burstiness, although it tracks the traffic fast.

4.2 Allocation of resources: P_0 traffic

P_0 traffic has a strict delay constraint. Accordingly, for the users connected to BS_i only, the granted resource in frame $(n+1)$ is: $S^{u(g)}_{P_0,i}(n+1) = \min\{S^{u(r)}_{P_0,i}(n+1), SLA_{P_0}\}$. The allocation to the users in BS_j only is similarly done. For the users in shared region, the resource allocations from BS_i and BS_j are: $S^{u(g)}_{P_0,i(1)}(n+1) = \frac{L_j}{L_i+L_j}\cdot\min\left\{S^{u(r)}_{P_0,i,j}(n+1), SLA_{P_0}\right\}$ and $S^{u(g)}_{P_0,j(2)}(n+1) = \frac{L_i}{L_i+L_j}\cdot\min\left\{S^{u(r)}_{P_0,i,j}(n+1), SLA_{P_0}\right\}$, where L_i and L_j are the respective loads on BS_i and BS_j. This strategy ensures that more resources are allocated from the lightly loaded BS.

4.3 Allocation of resources: P_1 traffic (phase 1)

For the users associated with only BS_i, $(u = 1$ to $N_i)$,

$$S^{u(g)}_{P_1,i}(n+1)|_I = \begin{cases} S^{u(r)}_{P_1,i}(n+1), & \text{if } S^{u(r)}_{P_1,i}(n+1) \le SLA_{P_1}, \\ SLA_{P_1}, & \text{otherwise.} \end{cases}$$

A similar approach is taken for the users associated with only BS_j. For the users in the shared region (for $u = 1$ to N_{ij}), the allocated resource $S^{u(g)}_{P_1,i,j}(n+1)|_I = S^{u(g)}_{P_1,i(1)}(n+1)|_I + S^{u(g)}_{P_1,j(2)}(n+1)|_I$ is shared between BS_i and BS_j based on their loads, as done for P_0 traffic.

After allocation of resources to P_0 traffic and P_1 traffic in the first phase, the remaining resources $S^{(e)}_i$, $S^{(e)}_j$ are calculated:

$$S^{(e)}_i = S_i - \sum_{u=1}^{N_i}\left(S^{u(g)}_{P_0,i} + S^{u(g)}_{P_1,i}|_I\right) - \sum_{u=1}^{N_{ij}}\left(S^{u(g)}_{P_0,i(1)} + S^{u(g)}_{P_1,i(1)}|_I\right),$$

$$S^{(e)}_j = S_j - \sum_{u=1}^{N_j}\left(S^{u(g)}_{P_0,j} + S^{u(g)}_{P_1,j}|_I\right) - \sum_{u=1}^{N_{ij}}\left(S^{u(g)}_{P_0,j(2)} + S^{u(g)}_{P_1,j(2)}|_I\right).$$

4.4 Allocation of resources: P_1 traffic (phase 2) and P_2 traffic

For the users associated with only BS_i, $(u = 1$ to $N_i)$,

$$S^{u(g)}_{P_1,i}(n+1)|_{II} = \begin{cases} S^{u(r)}_{P_1,i}, & \text{if } S^{u(r)}_{P_1,i} \le SLA_{P_1}, \\ SLA_{P_1} + \frac{S^{(e)}_i \cdot S^{u(r)}_{P_1,i}}{\sum_{u=1}^{N_i}\left(S^{u(r)}_{P_1,i}+S^{u(r)}_{P_2,i}\right) + \sum_{u=1}^{N_{ij}}\left(S^{u(r)}_{P_1,i,j}+S^{u(r)}_{P_2,i,j}\right)\frac{L_j}{L_i+L_j}}, & \text{otherwise,} \end{cases}$$

and

$$S_{P_2,i}^{u(g)}(n+1) = \min\left\{ S_{P_2,i}^{u(r)}, \frac{S_i^{(e)} \cdot S_{P_2,i}^{u(r)}}{\sum_{u=1}^{N_i}\left(S_{P_1,i}^{u(r)} + S_{P_2,i}^{u(r)}\right) + \sum_{u=1}^{N_{ij}}\left(S_{P_1,i,j}^{u(r)} + S_{P_2,i,j}^{u(r)}\right)\frac{L_j}{L_i+L_j}} \right\}.$$

Similarly, the resources are allocated to the users in BS_j. For the users in the shared region ($u = 1$ to N_{ij}), the resources are allocated in phase II from the two BSs based on their traffic loads. Thus, the fractional allocation from BS_i, for example, are: $S_{P_1,i(1)}^{u(g)}(n+1)|_{II} = \frac{L_j}{L_i+L_j} \cdot S_{P_1,i,j}^{u(g)}(n+1)|_{II}$, and $S_{P_2,i(1)}^{u(g)}(n+1) = \frac{L_j}{L_i+L_j} \cdot S_{P_2,i,j}^{u(g)}(n+1)$.

5 Numerical and Simulation Results

In light of the practical deployment scenarios, we have considered three different cases: The first scenario deals with two adjacent BSs and a straight line movement trajectory of the MS through them, to justify the gain of the proposed scheme over a typical hard handoff as well as a soft handoff approach (MDHO), verify via simulations the numerical results the system capacity gain with respect to hard handoff, and study the effect of traffic loads in the BSs. The second scenario is with three BSs in sequence, to further generalize the impact of split handoff on different service classes. In the third scenario we have taken a generalized 3-tier (19 cell) network with random movement pattern of the MSs, to study the average network performance.

For verification of numerical results as well as for generalized network performance, the simulations were carried out in Qualnet 5.2 with mobile WiMAX toolbox for emulating broadband wireless access. Urban propagation library was used for channel modeling. Hard handoff was implemented with FBSS option. For implementing shared resource allocation in split handoff and replicated resource allocation in MDHO, the necessary control message modifications in uplink and downlink frames were done in UL-map and DL-map to allow downlink data communication to the MS via more than one BS and control message exchange via the PBS only. In all simulations other than for verification of numerical results, adaptive modulation and coding scheme was enabled so that a MS can select a suitable rate depending on its channel condition.

Following the typical cases of wireless systems with adaptive modulation, the downlink power budget for each user is kept constant. The signals from the center cell as well as from the neighboring cells are considered with path loss and shadow fading. Unless otherwise specified, the default system settings and parameters in the simulations are mentioned in Table 1.

TABLE 1

Default system parameters for simulations

Downlink power per user	20 dBm
Antenna model	Omnidirectional
Channel bandwidth	10 Mbps
Frequency reuse factor	1
Number of subcarriers	2048
Cyclic prefix training length	8 μS
Path loss model	Two-ray
Shadowing mean	0 dB
Shadowing standard deviation	4 dB
Propagation fading model	None
Propagation limit	-111 dBm
BS frame duration	20 ms
BS time-division duplex DL duration	18 ms
Maximum allowed downlink load level	0.7

The class-based traffic parameters are taken as follows. P_0: VoIP traffic - exponentially distributed with average ON time 1.34 s and OFF time 1.67 s; P_1: video traffic - Paris sequence with H.264 variable bit rate encoded at 30 frames per second with 352×288 pixels/frame; P_2 is the best effort data traffic, which is taken as web traffic with mean packet size 1500 Bytes and Poisson distributed packet inter-arrival times with mean 0.133 s. In our simulation studies, we have determined the respective optimum values of prediction order $\mathcal{L}_{P_c}$, for $c \in \{0, 1, 2\}$, as 1, 4, and 2. The ratio of P_0, P_1, and P_2 type users were taken as $10 : 45 : 45$. The performance measures were: packet drop rate for P_0, normalized throughput for P_1, and packet delay for P_2.

5.1 Scenario 1: Two BS case

5.1.1 Enhanced data rate during handoff

A simple simulation scenario with two BSs and one MS is depicted in Fig. 3. The speed of the MS is 5 m/s. For this specific scenario a CBR connection with Packet size 1024 Bytes.

With the two BSs equally loaded, supported data rate with the two handoff schemes at different position of the MS is shown in Fig. 4. In the handoff (coverage overlap) region the throughput achieved with the proposed handoff is much higher than hard handoff as well as MDHO.

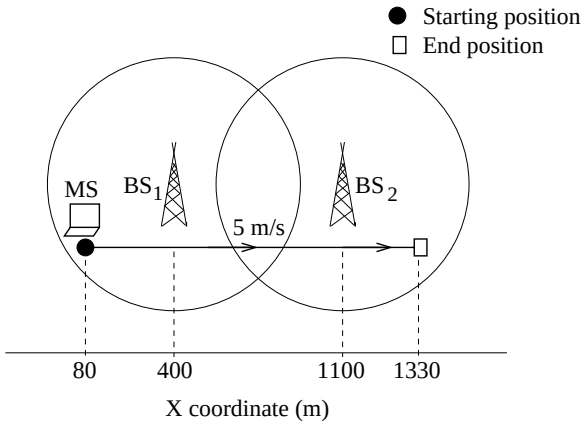

Fig. 3. Two BS scenario with straight line trajectory of MS from one BS to the other.

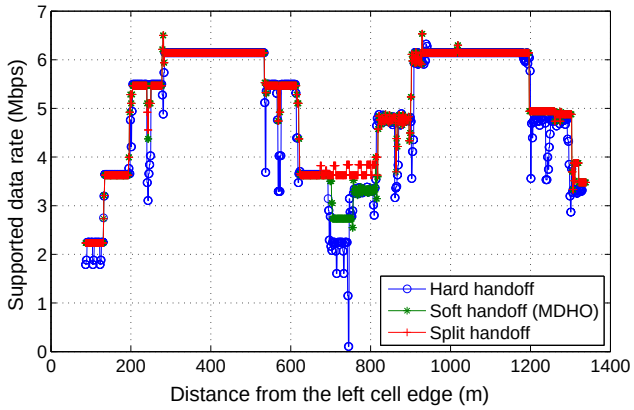

Fig. 4. Comparison of data rate in the three handoff schemes. Mobile speed 5 m/s.

Specifically, our simulation results show that the average throughput maintained by split handoff in the coverage overlap region of the two BSs (spanning about 200 m distance) is 3.69 Mbps in comparison to 2.22 Mbps via hard handoff and 3.07 Mbps via MDHO. Note that, when the MS is at either of the cell boundaries, the performance of all three handoff schemes are equally poor, as there are no sharable network resource for the split handoff in this two-BS scenario.

5.1.2 System capacity gain

The numerical results from the analysis developed in Section 3 on capacity gain of split handoff with respect to hard handoff are compared with the simulation results, where two different service

classes, P_0 and P_1, were individually considered. The capacity gain is computed as:

$$E^{u,\max}_{C,gain}(\theta^u) \stackrel{\text{def}}{=} \frac{E^{u,\max}_{C,joint}(\theta^u) - E^u_C(\theta^u)}{E^u_C(\theta^u)} \cdot 100\%, \tag{9}$$

Here, in each direction of movement of the MS, the number of BSs which can share the resources are limited to two. Beyond the required parameters in Table 1, the parameter values considered for the numerical results are as follows: Path loss factor $= 3$; modulation scheme is 4-QAM (correspondingly, bits per symbol is $r = 2$); distance between the two adjacent BSs is 700 m.

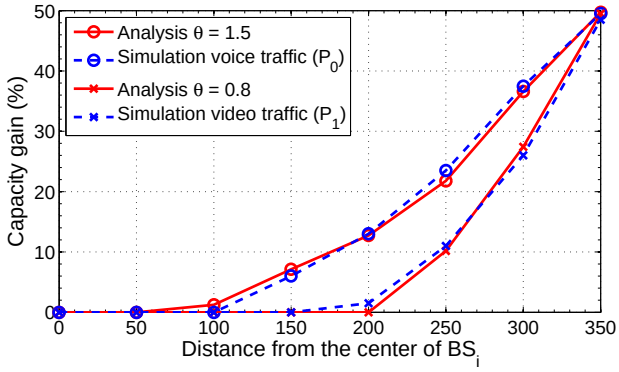

Fig. 5. Comparison of effective capacity gain of split handoff with respect to hard handoff, as defined in (9), ref. experiment scenario in Fig. 8. $\gamma_{th} = 3$ dB, $\rho = 0.9$, and MS speed 5 m/s.

Fig. 5 shows the numerical results for two different values of QoS exponent θ on capacity gain, which are plotted against distance of the MS from the center of one BS to the other. As the distance between the two BSs is 700 m, the results are presented for 0 up to 350 m. The MS is considered traveling at a constant speed of 5 m/s. It can be noted that, the gain is higher for the traffic with a stricter QoS constraint (i.e., with a higher value of θ). The plots also show that, as compared to a user with a loose QoS requirement, the one with a stringent QoS starts benefiting earlier in its trajectory to the adjacent BS. For example, at 200 m distance from the center of the left-side BS (Fig. 3), the analytically predicted capacity gain of P_0 traffic with split handoff is about 26%, whereas that of the P_1 class is still 0. Intuitively, the capacity gain reaches the maximum value in both traffic classes when the MS is equidistant from the two BSs.

Fig. 5 also presents the comparison of analysis and simulation results for the two different classes of traffic, which correspond to two different values of θ. Here, P_0 (VoIP) traffic corresponds to the numerically generated plot with $\theta = 1.5$, and P_1 (video) traffic corresponds

to the numerical result with $\theta = 0.8$. The numerical and simulation results are reasonably well matched. A little difference can be attributed to the fact that, the constant arrival rate assumption, which is a baseline for the effective capacity definition, does not strictly hold in simulations with practical parameters of the two traffic classes.

5.1.3 *Traffic load balancing*

Fig. 6 shows the impact of BS loading on sharing of resources from the two BSs. The mobile speed considered here is 20 m/s. The two curves show the data packets handled by the two sharing BSs. It can be noticed in Fig. 6(a), when the BS_1 is carrying more load (with utilization factor $\rho_1 = 0.67$) as compared to BS_2 ($\rho_2 = 0.33$), the sharing window is positioned closer to the BS_1 center. This is because, for load balancing the BS_2 starts sharing the traffic from the handoff region quite early. Fig. 6(b) shows the same impact on sharing window position when

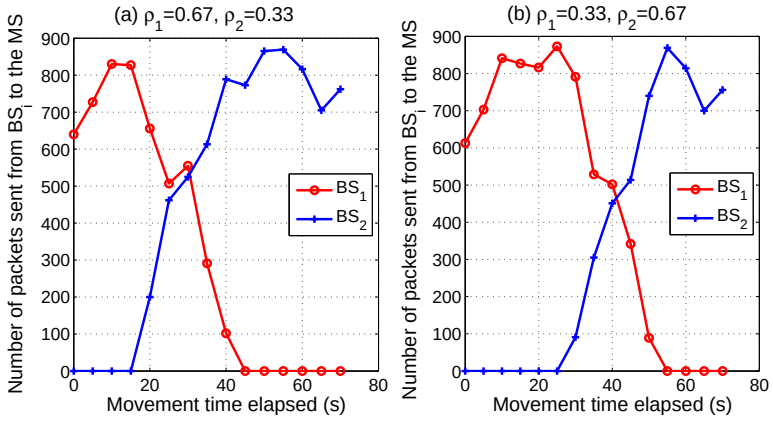

Fig. 6. Distribution of number of data packets handled with movement time elapsed (ref. mobility scenario in Fig. 3) and position of sharing window at different cell loads. Mobile speed 20 m/s.

the load on BS_1 is $\rho_1 = 0.33$ and on BS_2 it is $\rho_2 = 0.67$.

Fig. 7 presents the impact of the proposed scheme compared to the hard handoff and MDHO on packet drop rate. When the MS is close to either of the BSs, i.e., at the positions $X = 400$ m and $X = 1100$ m (cf. Fig. 3), the data packet drop rate is lower. But as the MS goes away from either of the BSs, the packet drop rate increases. In split handoff scheme, the drop rate is significantly less than the hard handoff in the coverage overlap region because of its resource sharing capability. The replicated resource usage in MDHO helps stem the packet drop rate

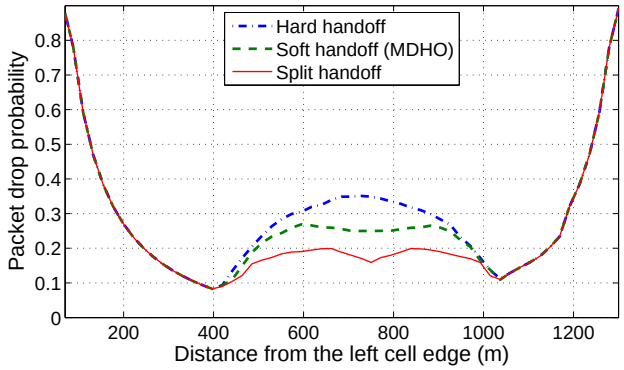

Fig. 7. Comparison of packet drop probability in the three handoff schemes. Mobile speed 20 m/s.

compared to that in hard handoff, but its performance is poorer than split handoff because, by its principle of macro-diversity for all cell-edge users the available resource from the neighboring BSs for the cell-edge users is reduced. In particular, the three curves in Fig. 7 indicate that, the highest packet drop probability during handover is reduced to 20% in split handoff from 35% in hard handoff and 27% in MDHO. In other parts of the trajectory, where there is no sharing possibility, both curves are overlapping, which is intuitive as the unique features of the split handoff scheme is not exploited there.

The key observations on split handoff from this study can be summed up as follows: (a) More data rate can be achieved while sharing the resources from more than one BS. (b) The sharing window shifts towards the BS which is more loaded, demonstrating cooperation and load sharing among the neighboring BSs. (c) The packet drop rate is also less in case of shared resource allocation strategy, as the proposed strategy exploits the additional resource from the neighboring BSs in a more efficient way than in MDHO.

5.2 Scenario 2: Three BS case

A more generalized scenario was considered next with 3 BSs and the MS trajectory from the center of the first BS to the center of the last BS, as depicted in Fig. 8. The simulation results are presented against movement time, starting at the MS position in BS_i and stopping in BS_k. There are two handoff points in the full trajectory. The network carries all three classes of traffic: P_0, P_1, and P_2.

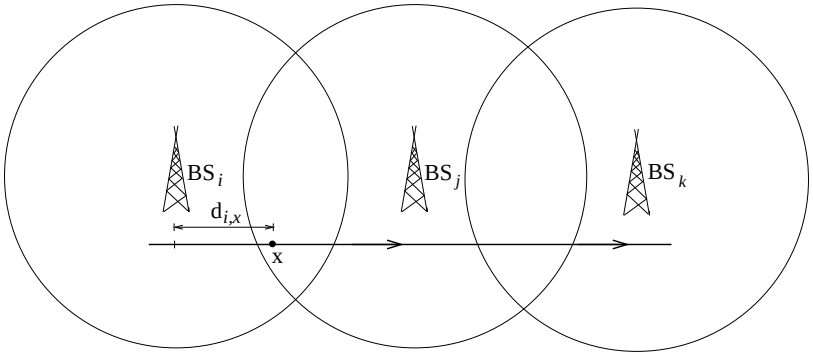

Fig. 8. 3 BS scenario with straight line MS movement. Center-to-center distance $=$ 700m.

P_0 packet drop rate and P_2 packet delay are captured in the three handoff strategies, as shown in Fig. 9. As the MS moves, the number of packets dropped in Fig. 9(a) are nearly the same for all the handoff schemes, except inside the coverage overlap region where the handoff takes place. During hard handoff all stored packets are dropped, and so drop rate during that time is very high. In MDHO, due to *replicated resource usage* from the two BSs, the performance improves improves to some extent. In split handoff, on the other hand, as more than one downlink path is available *through sharing*, the packets can be successfully delivered by taking the diversity advantage. The average delay performance in Fig. 9(b) can be similarly explained. Unlike in

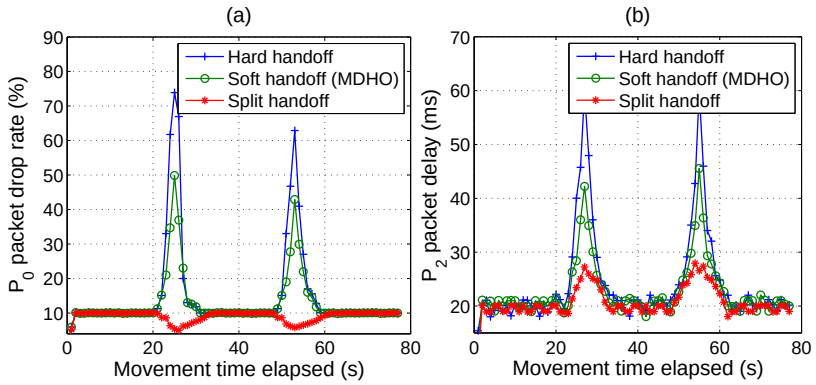

Fig. 9. Comparison of class based performance: (a) packet drop rate for P_0 traffic; (b) average packet delay for P_2 traffic. Mobile speed 20 m/s.

split handoff, in hard handoff as well as in MDHO the packets experiencing bandwidth resource

limitations are dropped during the handoff. They have to be either retransmitted, or tunneled from the first BS to the second, which introduces delay. The added fluctuation in delay performance can be attributed to the best effort nature of P_2 traffic, which is served after guaranteeing the P_0 and P_1 SLAs.

5.3 Scenario 3: Three tier case (a cluster of 19 BSs)

To study the effect of the proposed handoff strategy at a broad level in a typical cellular scenario, we have taken a 3-tier cellular structure with 19 BSs in a cluster, and the user traffic is broadly categorized into three classes: P_0 (VoIP), P_1 (packet video), and P_2 (data).

By the principle of split handoff the resource sharing for a particular MS can be from two or more BSs such that the gain in capacity is maximized. However, the number of connection threads required with the increased number of shared BSs as well as the increased number of MSs and the consequent increase in runtime in our machine (Dell Optiplex 990) are noted to be very high. Therefore, we restrict our simulation scenario and resource sharing between only two neighboring BSs (PBS and SBS). Each BS has up to 38 MSs that are uniformly distributed over the coverage area of a BS. Mobile users are considered moving with a random mobility.

Fig. 10(a) shows the comparison of average packet drop rate of P_0 traffic versus speed during the handoff phase. As shown earlier in Fig. 9, the packet drop rate during handoff for a constant speed are higher in hard handoff and MDHO as compared to the proposed scheme. It can be

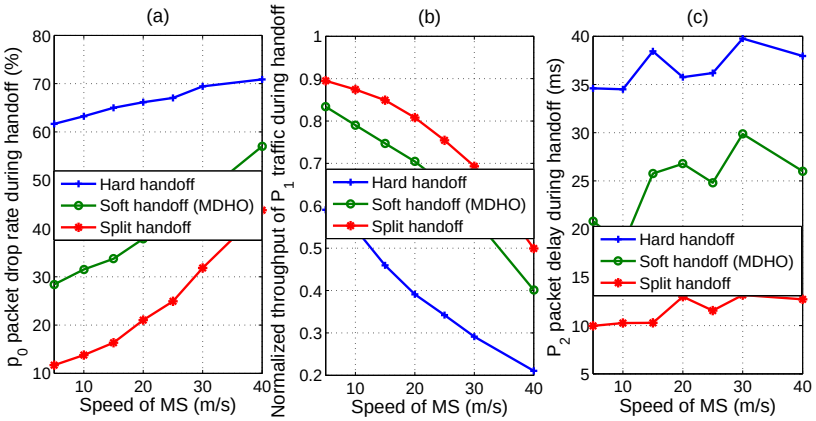

Fig. 10. Class based performance in a multi-cell handoff scenario at different speed of mobile users. (a) P_0 packet drop rate; (b) P_1 packet throughput; (c) P_2 packet delay. Average number of MS per BS is 38.

observed from Fig. 10(a) that, as the speed of the users increases the packet drop rate in the proposed scheme increases at a higher rate. With the increased speed, the performance gain of proposed scheme tends to level off. The reason is that, with a higher mobility, the effective region for resource sharing is decreased, resulting in a reduced performance benefit.

Fig. 10(b) shows the throughput comparison of P_1 (video) traffic versus speed of the mobile users. The throughput has been normalized with respect to the highest achievable throughput. The first observation is that, the throughput decreases in all handoff schemes as the mobile speed increases. This is because, the packet drop rate increases with speed. Secondly, as the speed increases the throughput gain in the proposed scheme with respect to the hard handoff and MDHO diminishes. The reason is same as in Fig. 10(a); as the speed increases the sharing window size shrinks and hence the proposed scheme does not perform as good as in lower speeds. The delay performance of P_2 traffic is also plotted in Fig. 10(c). Although the plots look a little random, the variations in magnitude are rather minuscule. Yet, there is a clear trend of increased delay performance at higher speeds in all three handoff schemes, which has the same intuitive reasoning as in Figs. 10(a) and 10(b). The randomness in the delay plots could be because of the least-priority handling of the P_2 traffic, which is served only after serving P_0 class and P_1 class. The bursty natures of P_0 and P_1 traffic entail that the resource available for P_2 can vary widely from frame to frame, leading to a poor convergence to its average performance. Also, the increase in velocity has very little impact on the delay variation, resulting in a bloated representation of the minor variation in delay.

Fig. 11 shows variation of the same performance parameters as in Fig. 10 for P_0, P_1, and P_2 traffic with traffic load per BS. In Fig. 11(a) the packet drop rate for P_0 traffic is drawn. The drop rate in the proposed scheme is less compared to those in hard handoff and MDHO. But it is noticeable that, as the average load per BS increases, the packet drop rates in hard handoff and MDHO increase at higher rate as compared to the proposed scheme. Thus, the proposed scheme is more beneficial in high load conditions for high QoS applications. Fig. 11(b) shows P_1 throughput performance of the proposed scheme in comparison with hard handoff and MDHO. The throughput of the proposed scheme is quite higher. Moreover, compared to the other two handoff schemes, the throughput has less impact of the number of users per BS on the proposed scheme. As the number of users increases, the throughput tend to decrease. But, in split handoff more opportunities are created to share resource for active handoff users, and as a result the rate

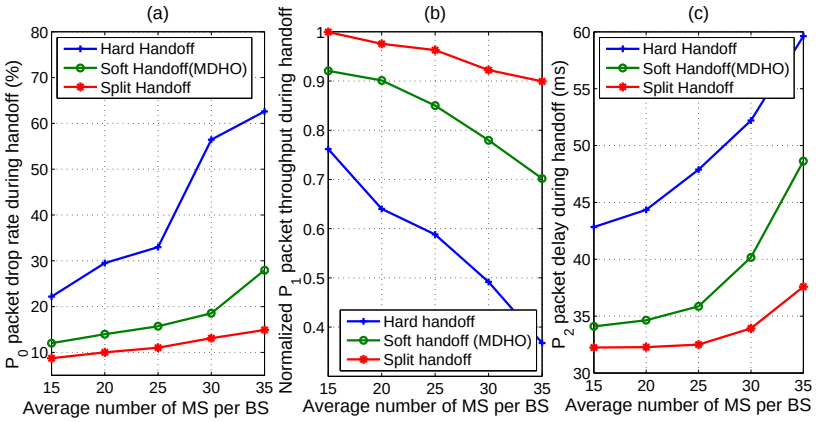

Fig. 11. Class based performance in a multi-cell handoff scenario at different traffic loads. (a) P_0 packet drop rate; (b) P_1 packet throughput; (c) P_2 packet delay. Speed of mobile users is 25 m/s.

of decay in throughput is less. Average delay for P_2 traffic (during handoff) versus traffic load per BS is shown in Fig. 11(c). Again, the delay of the proposed scheme is less as we have seen in the two BS case. The increment in delay in the proposed scheme is less, which is attributed to the same reason of resource sharing. It may be noted here that, in contrast with the impact of speed variation on P_2 delay, the impact of cell load is much significant, and as a result, the trends of delay variation versus cell load is smooth.

We also present single BS statistics in a two-cell environment in terms of percentage of shared resources and position of sharing window with the proposed scheme. In general, the number of users per BS basically represents its load. We recall that, static users are kept constant since they do not take part in handoff. So, as the number of mobile users increases, the load on BS also increases. Fig. 12 shows the effect of load in a BS on the sharing window position and resources shared from the more loaded BS. As the load increases on a BS, it extracts the benefit of sharing by shifting more of its load to the neighboring BS. With respect to the impact on load sharing window, the sharing window position means the location of mid point of sharing zone between two BSs. The center of sharing window shifts towards the BS, which is heavily loaded to distribute its traffic to the nearby BS in form of shared users. So, in this way the split handoff inherently takes care of the load balancing property of a cellular network.

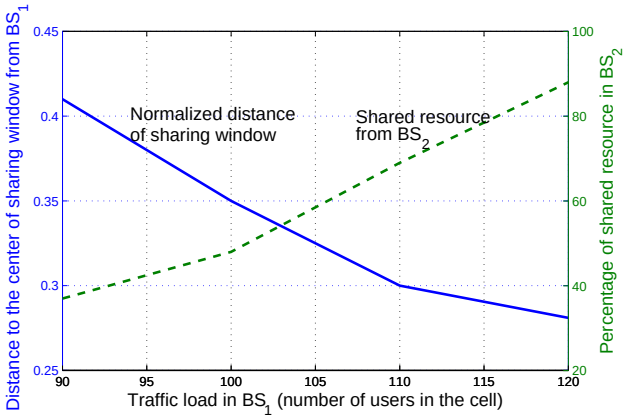

Fig. 12. Load sharing performance versus load in a cell, with the other BS having a fixed load of 100 users. Speed of mobile users is 30 m/s.

6 Conclusion

We have presented a QoS aware handoff strategy in OFDMA broadband cellular networks, called split handoff, that allows sharing of channel resource from more than one BS when the user is in their joint coverage area. We have proposed a system architecture that supports the proposed scheme, outlined the system interactions involved, and provided an analytical framework to quantify the capacity gain by shared resource usage. The capacity enhancement analysis has been validated by network simulation results using Qualnet. Further, we have provided a heuristic class-based shared resource allocation policy for the cell-edge users that aims at maximizing the QoS support to different service classes while maximizing the resource utilization. We have conducted rigorous network simulations in Qualnet simulator, where mobile WiMAX has been considered as an example broadband wireless access network standard. Our results demonstrated that the shared resource allocation in the proposed split handoff scheme significantly improves the network performance with respect to the hard handoff as well as the macro-diversity handoff in terms of system capacity, QoS guarantee, as well as traffic load balancing, without incurring additional network operation cost, such as power consumption and bandwidth usage. The proposed strategy and system model can be customized to scenario-specific requirements.

ACKNOWLEDGMENT

This work has been supported by the Department of Science and Technology (DST) under the grant no. SR/S3/EECE/0122/2010. The authors are thankful to the anonymous reviewers for the constructive criticisms, insightful comments, and valuable suggestions, which have significantly improved the quality of presentation of the paper.

REFERENCES

[1] E. Dahlman, S. Parkvall, and J. Skold, *3G evolution: HSPA and LTE for mobile broadband*, 2nd ed. Academic Press, 2008.

[2] "3GPP Long-Term Evolution (LTE)." [Online]. Available: http://www.3gpp.org/Highlights/LTE/LTE.htm

[3] L. Bajzik, P. Horvath, L. Korossy, and C. Vulkan, "Impact of intra-LTE handover with forwarding on the user connections," in *Proc. IEEE Mobile and Wireless Commun. Summit*, Budapest, Hungary, July 2007.

[4] B. G. Lee, D. Park, and H. Seo, *Wireless Communications Resource Management*, 1st ed. Wiley-IEEE Press, Dec. 2008.

[5] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical QoS guarantees in mobile wireless networks," in *Proc. ACM Intl. Conf. Quality of Service in Heterogeneous Wired/Wireless Networks*, Waterloo, Ontario, Canada, Aug. 2006.

[6] T. Ali-Yahiya, A.-L. Beylot, and G. Pujolle, "An adaptive cross-layer design for multiservice scheduling in OFDMA based mobile WiMAX systems," *Elsevier Computer Commun.*, vol. 32, pp. 531–539, Feb. 2009.

[7] Z. Kong, Y.-K. Kwok, and J. Wang, "A low-complexity QoS-aware proportional fair multicarrier scheduling algorithm for OFDM systems," *IEEE Trans. Vehicular Tech.*, vol. 58, no. 5, pp. 2225–2235, June 2009.

[8] H. Lee, H. Son, and S. Lee, "Semisoft handover gain analysis over OFDM-based broadband systems," *IEEE Trans. Veh. Technol.*, vol. 58, no. 3, pp. 1443–1453, Mar. 2009.

[9] J. Chang, Y. Li, S. Feng, H. Wang, C. Sun, and P. Zhang, "A fractional soft handover scheme for 3GPP LTE-advanced system," in *Proc. IEEE Intl. conf. Commun.*, Dresden, Germany, June 2009.

[10] "Standard for local and metropolitan area networks Part 16: Air interface for fixed and mobile broadband wireless access systems, IEEE Std. 802.16e," 2005.

[11] P. Li, X. Yi, and Y. Pan, "A seamless handover mechanism for IEEE 802.16e systems," in *Proc. Intl. Conf. Commun. Technol.*, Guilin, China, Nov. 2006.

[12] W. Jiao, P. Jiang, and Y. Ma, "Fast handover scheme for real-time applications in mobile WiMAX," *Proc. IEEE ICC*, June 2007.

[13] S. W. Kim, "Spatial-multiplexed soft handoff," in *Proc. IEEE Conf. Wireless Comm. and Networking*, Kowloon, China, Mar. 2007.

[14] L. G. de R. Guedes and M. D. Yacoub, "Overlapping cell area in different fading conditions," in *Proc. IEEE Vehicul. Tech. Conf.*, Chicago, IL, USA, July 1995.

[15] M. R. R. Kumar, S. Bhashyam, and D. Jalihal, "Throughput improvement for cell-edge users using selective cooperation in cellular networks," in *Proc. IEEE/IFIP Conf. Wireless and Optical Commun. Networks*, Surabaya, Indonesia, May 2008.

[16] L. Xu, K. Yamamoto, H. Murata, and S. Yoshida, "Adaptive base station cooperation and subchannel reallocation at cell edge in cellular networks with fractional frequency reuse," in *Proc. IEEE Conf. Pers. Indoor and Mobile Radio Commun.*, Tokyo, Japan, Sept. 2009.

[17] M. M. Wang, T. Ji, J. Borran, and T. Richardson, "Interference management and handoff techniques in ultra mobile broadband communication systems," in *Proc. IEEE Intl. Symp. Spread Spectrum Techniques and Applications*, Bologna, Italy, Aug. 2008, pp. 166–172.

[18] M. C. Necker, "Interference coordination in cellular OFDMA networks," *IEEE Network Mag.*, vol. 22, no. 6, pp. 12–19, Nov.-Dec. 2008.

[19] G. Boudreau, J. Panicker, N. Guo, R. Chang, N. Wang, and S. Vrzic, "Interference coordination and cancellation for 4G networks," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 74–81, Apr. 2009.

[20] S. Shenker, C. Partridge, and R. Guerin, "Specification of guaranteed quality of service," 1997. [Online]. Available: http://www.ietf.org/rfc/rfc2212.txt

[21] J. Chen, W. Jiao, and Q. Guo, "Providing integrated QoS control for IEEE 802.16 broadband wireless access systems," in *Proc. IEEE Vehicul. Tech. Conf. - Fall*, Dallas, TX, USA, Sept. 2005.

[22] M. Ma and B. C. Ng, "Supporting differentiated services in wireless access networks," in *Proc. Intl. Conf. Commun. Systems*, Singapore, Oct. 2006.

[23] H.-J. Yao and G.-S. Kuo, "An integrated QoS-aware mobility architecture for seamless handover in IEEE 802.16e mobile BWA networks," in *Proc. IEEE MOLCOM*, Washington, DC, USA, Oct. 2006.

[24] D. Wright, "Maintaining QoS during handover among multiple wireless access technologies," in *Proc. Management of Mobile Business*, Toronto, Ontario, Canada, July 2007.

[25] H. Zhou and Z. Zhang, "Differentiated statistical QoS guarantees for real-time CBR services in broadband wireless access networks," in *Proc. Intl. Conf. Wireless Communications, Networking, and Mobile Computing*, Chengdu, China, Sept. 2010.

[26] S.-I. Choi and J.-D. Huh, "Dynamic bandwidth allocation algorithm for multimedia services over Ethernet PONs," *ETRI J.*, vol. 24, no. 6, pp. 465–468, Dec. 2002.

[27] Y. Luo and N. Ansari, "Bandwidth allocation for multi-service access on EPONs," *IEEE Optical Commun. Mag.*, vol. 43, no. 2, Feb. 2005.

[28] D. Nowak, J. Murphy, and P. Perry, "Bandwidth allocation in DiffServ enabled ethernet passive optical networks," *IET Commun. J.*, vol. 3, no. 3, pp. 391–401, Mar. 2009.

[29] S. De, V. Singh, H. M. Gupta, N. Saxena, and A. Roy, "A new predictive dynamic priority scheduling in Ethernet passive optical networks," *Elsevier J. Optical Switching and Networking,*, vol. 7, no. 4, pp. 215–223, 2010.

[30] S. Sezginer and H. Sari, "Full frequency reuse in OFDMA-based wireless networks with sectored cells," in *Proc. IEEE WCNC*, Budapest, Hungary, Apr. 2009.

[31] M. D. Katz and F. H. P. Fitzek (Eds.), *WiMAX Evolution: Emerging Technologies and Applications*. John Wiley, 2009.

[32] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

[33] M. Kazmi and N. Wiberg, "Scheduling algorithm for HS-DSCH in a WCDMA mixed traffic scenario," in *Proc. IEEE Personal Indoor, Mobile Radio Commun.*, Beijing, China, Sept. 2003.

[34] S. Haykin, *Adaptive Filter Theory*. Prentice Hall, 4th Ed., 2001.

BIOGRAPHIES

Chetna Singhal completed her M.Tech in Computer Technology from the Electrical Eng. Department, IIT Delhi, in 2010. From June 2010 to July 2011 she worked in IBM Software Lab, Gurgaon, as a Software Engineer. She is currently pursuing Ph.D. at the Bharti School of Telecom, IIT Delhi. Her research interests include handoff studies and cross layer optimization in wireless networks, multimedia multicast, and broadcast technology.

Satish Kumar completed his M.S. from the Bharti School of Telecom, IIT Delhi, in 2011. He is currently associated with Qualcomm India Pvt. Ltd. in the field of wireless technology. His research interests include cooperative wireless communications, scheduling techniques in WCDMA and OFDMA networks, and LTE/WCDMA small cell optimization.

Swades De received his Ph.D. in Electrical Eng. from the State University of New York at Buffalo, in 2004. He is currently an Associate Professor in the Department of Electrical Eng. at IIT Delhi. His research interests include performance study, resource efficiency in multihop wireless and high-speed networks, broadband wireless access, and communication and systems issues in optical networks.

Nitin Panwar completed his M.Tech from the Bharti School of Telecom, IIT Delhi, in 2011. He is currently associated with Cisco Systems India Pvt. Ltd. in the field of networking. His research interests include networking, handoff schemes in wireless networks, scheduling techniques in OFDMA based networks.

Ravindra Tonde is currently working with Samsung India Software Center, Noida, as a Senior Software Engineer. He completed his M.Tech in Computer Technology from the Electrical Eng. Department, IIT Delhi, in 2010, and B.E. in Electronics and Telecommunications from the Pune University in 2008. His research interests are in wireless communication technologies, cloud and convergence technologies.

Pradipta De is a Research Staff Member at IBM Research, India, at New Delhi. He is currently a member of the Telecom and Mobile Research group, where he is involved in projects related to Mobile Enabled Financial Services and Mobile Cloud Computing. He has also worked on projects related to various aspects of Data Center Management and Service Delivery. He holds a Ph.D. from Stony Brook University.