



# User-Cognizant Scalable Video Transmission over Heterogeneous Cellular Networks **39**

Liang Wu and Wenyi Zhang

## Contents

Introduction	1306
Related Work	1308
A User-Cognizant Solution	1310
System Model	1311
Layered Video Model	1311
Network Model	1312
Transmission Schemes	1313
UE Load and Sub-band Occupancy	1315
UE Load	1315
Sub-band Occupancy	1316
SINR Distribution and Data Rate	1318
SINR Distribution	1318
Data Rate	1327
System Performances	1328
LD Transmission	1329
LHDA Transmission	1330
CD Transmission	1332
Simulation and Discussion	1333

---

L. Wu (✉)

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

Huawei Technologies, shanghai, China

e-mail: [tuohai@mail.ustc.edu.cn](mailto:tuohai@mail.ustc.edu.cn)

W. Zhang

Key Laboratory of Wireless-Optical Communications, Chinese Academy of Sciences, Hefei, China

Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

e-mail: [wenyizha@ustc.edu.cn](mailto:wenyizha@ustc.edu.cn)

Conclusion and Future Directions.....	1339
References.....	1340

---

### Abstract

With the increase of mobile video applications in people's daily life as well as industrial manufacture, such as video streaming, surveillance, and so on, video has been the main service in cellular networks. Operators and service providers are struggling to enhance the mobile video service, while user requirements for abundant, high-definition, and low-delay video have nearly drained the transmission capacity of current networks. Moreover, the large population of user equipments (UEs) exhibit differentiated video demands and various network transmission environments. Traditional networking, which is static and base station (BS) concentric, can hardly deal with these challenges. Thus, adaptive video transmission schemes are needed by jointly considering the interplay among user demand, video source characteristics, and networking. This work focuses on user-cognizant scalable video transmission over heterogeneous cellular networks. The video source is encoded using scalable video coding, which enables dynamic adaption of source information to the requirements of UEs and is suitable for cellular networks in which the transmission link quality varies substantially over space and time. Three novel transmission schemes are proposed, layered digital transmission, layered hybrid digital-analog transmission, and cooperative digital transmission. Leveraging tools from stochastic geometry, a comprehensive analysis is conducted focusing on three key performance metrics: outage probability, high-definition probability, and average distortion. The associated spectrum allocation and video transmission are chosen based on the user-cognizant information, such as the requirements for video service, wireless channel status, and the connections with the BSs. The results show that the proposed user-cognizant transmission schemes can provide a scalable video experience for UEs.

---

### Keywords

Coordinated multipoint · Heterogeneous cellular networks · Hybrid digital-analog · Scalable video coding · Stochastic geometry · User cognizant

---

## Introduction

Emerging wireless communication technologies as well as powerful and versatile mobile terminal devices have changed people's daily life, and the data traffic grows explosively, among which a substantial portion is attributed to multimedia such as mobile video. According to the Cisco Visual Networking Index, mobile video is expected to grow at an average growth rate of 62% until 2020, and within the 30.6 exabytes of data per month crossing mobile networks by 2020, 23.0 exabytes will be video related, such as video on demand, real-time streaming video, video

conferencing, and so on. The ever-increasing demand for abundant, high-definition, and low-latency mobile video brings great challenges to the mobile network with time-varying wireless channel. Moreover, with the release of different types of UEs, the requirements on data rate of video transmission vary in a wide range.

Compared to the IP transmission network and cellular core network, the bottleneck of the end-to-end transmission degrading the Quality of Experience (QoE) of video lies in the radio access network due to user traffic congestion and packet loss. Current LTE/LTE-A networks are not inheritably built for QoE-aware video delivery. The application-specific information exists at Packet Data Network Gateway (P-GW), while the wireless channel quality and connection status are restrictively known by eNodeB.

State-of-art design of cellular networks is base station (BS) concentric, which means that the resource allocation and transmission schedule are completed at the BSs and are not on-demand for UEs. The information bits are treated equally and the transmission strategy is not UE specific. As for the video transmission of a typical UE, the UE would require the video content with different video qualities based on its terminal capacity. Meanwhile, the UE can choose different service mechanisms provided by the cellular network based on the connection status.

Taking into account of time-varying wireless channel conditions and congestion, video streams are adapted to reduce the transmission bitrates. Traditionally, rate adaptation of video streams is realized by packet/frame dropping or transcoding with some serious drawbacks, since packet/frame dropping significantly degrades the video quality and transcoding is computationally complex. Advanced source coding techniques provide a new dimension of dynamically provisioning wireless resources for the varying requirements and the varying link conditions of UEs, thus creating the possibility of extracting video scaled in multiple dimensions, e.g., spatial, temporal, and quality. Scalable Video Coding (SVC) is an extension of the H.264/MPEG-4 AVC video compression standard [1] and has been evolved to Scalable High-Efficiency Video Coding (SHVC) [2], in which the bitstream is encoded into multiple layers, namely, a base layer (BL) and at least one enhancement layer (EL). The quality of reconstructed video depends on the number of layers decoded and stays the same until a higher enhancement layer is successfully decoded. The number of layers and their code rates may be determined by the requirement and the link condition of the subscribing UE.

On the other hand, cellular networks are evolving from a homogenous architecture to a composition of heterogeneous networks, comprised of various types of base stations (BSs) [3, 4]. Each type of BSs has its characteristic transmit power and deployment intensity: for example, macro BSs (MBSs) have larger transmit power, aiming at providing global coverage; femto access points (FAPs) are small BSs targeted for home or small business usages. As the distance between a UE and its serving FAP is small, the UE enjoys a high-quality link and achieves power savings. Meanwhile, the reduced transmission range also enhances spatial reuse and alleviates multiuser interference. In addition, different types of BSs can transmit cooperatively the same video content to the UE to enhance the quality of experience. The user-cell association approach for heterogeneous networks should be addressed

to exploit context information as well as channel-related information extracted from UEs. Generally speaking, there are two different service modes, separate mode and cooperative mode. In separate mode, the macro cells and the small cells (e.g., femto cells) transmit different video streams to the UEs in a manner of dual connectivity. In cooperative mode, both the macro cells and small cells transmit the same video content to the UE in a manner of coordinated multipoint transmission. In this paper, we study the problem of scalable transmission over heterogeneous networks and demonstrate the performances of several user-cognizant transmission schemes to exploit the combination of multi-layer video transmission and multitier cellular networks.

## Related Work

The prior works that consider video transmission over wireless networks mainly focus on scalable coding of video source or adaptive networking techniques separately. Furthermore, UE is regarded as a dummy terminal, and thus their differentiated demand and status are neglected. The analysis usually focuses on homogeneous networks, and the common feature of the layered structure of SVC and HCNs is not exploited. In [5], an overview of SVC and its relationship to mobile content delivery are discussed focusing on the challenges due to the time-varying characteristics of wireless channels. In [6], a per-subcarrier transmit antenna selection scheme is employed to support multiple scalable video sequences over a downlink cognitive network, and the outage probability is reduced because of video scalability. In [7], real-time use cases of mobile video streaming are presented, for which a variety of parameters like throughput, packet loss ratio, and delay are compared with H.264 single-layer video under different degrees of scalability. In [8], the proposed scheme employs WiFi: the BL is always transmitted over a reliable network such as cellular, whereas the EL is opportunistically transmitted through WiFi. Technical issues associated with the simultaneous use of multiple networks are discussed. In [9], HCNs with storage-capable small-cell BSs are studied: versions and layers of video have different impacts on the delay-servicing cost tradeoff, depending on the user demand diversity and the network load.

Some works related to QoE-aware or adaptive strategies have also been studied previously. Chen et al. [10] proposed an admission control strategy that was designed to maximize the number of video users satisfying the QoE constraints on their second-order statistics. Although the admission control strategy damages the QoE of the blocked users, the overall percentage of users satisfying the QoE constraints among both admitted and blocked users can be significantly improved. Thakolsri et al. [11] proposed a content-aware scheduling and resource allocation, taking into account the content characteristics of the video streams, and performs video rate adaptation at the BS. Fu et al. [12] proposed a QoE-aware video delivery by considering the hierarchical architecture of LTE/LTE-A network. Different video flows are marked at the core network to transform the video content information into QoE-aware priority classes. A packet dropping strategy addressing the transmission

capacity at the eNodeB is also proposed. But the priority marking process at the core network is unaware of channel status.

Considering the wireless video transmission techniques, there are several approaches to enhance the QoE of video users. The above literature is based on digital transmission, consisting of digital source coding (e.g., quantization and entropy coding), digital modulation (e.g., QPSK, 64QAM), and digital channel coding (e.g., turbo or LDPC). Unfortunately, digital transmission results in cliff effect. The cliff effect refers to the drastic degradation in video quality when the signal strength fades below the decoding threshold (as opposed to a graceful degradation). There exist certain SINR thresholds at which the video quality changes drastically; in between these thresholds, the quality stays approximately constant. The recently revitalized analog transmission has shown promising potential in handling channel variations and user heterogeneities for wireless video communication. The analog scheme consists of analog source coding and analog modulation that directly maps a source signal into a linearly transformed channel signal without channel coding. SoftCast [13] is an analog video broadcast scheme that transmits a linear transform of the video signal without quantization, entropy coding, or channel coding. It is claimed to realize continuous quality scalability. However, information-theoretic studies (such as [14, 15]) show that analog schemes with linear mapping (from source signals to channel signals) are relatively inefficient for video transmission, while hybrid digital-analog transmission is asymptotically optimal under matched channel conditions for optimally chosen power allocations between the analog and digital parts. The hybrid digital-analog scheme combines digital with analog schemes, transmitting digital and analog signals simultaneously using TDMA, FDMA, or superposition transmission. The authors in [16] propose a hybrid digital-analog scheme for broadcasting, showing a substantial performance gain. However, these works did not consider the impacts of HCNs and the spatial distribution of wireless networks, let alone the design of scalable transmission algorithms utilizing the structure of HCNs.

Moreover, coordinated multipoint (CoMP) transmission is intensively studied to enhance the system performance of LTE-A. By coordinating multiple BSs, the interference at the UE can be alleviated, or multiple received signals can be merged. The studies in [17, 18] evaluate the potential system gain of CoMP and discuss the appropriate deployment scenarios. The authors also review the necessary techniques of signal processing, backhaul link design, and supported protocols. There exist two types of cooperative transmissions, namely, coherent and noncoherent joint transmission. Many previous studies considered noncoherent joint transmission because it requires less channel status information. The authors in [19, 20] analyze the performance of noncoherent joint transmission in heterogeneous cellular networks and give the distribution of SINR for a user in a random position and cell edge, respectively. In addition, the impact of channel status information is also studied. Most previous works neglect the spectrum sub-band allocation, but the same sub-band is required when two BSs transmit cooperatively. Bang et al. [21] combines frequency fraction reuse and CoMP, and proposes an allocation to minimize the system power. Zhang et al. [22] and Kosmanos et al. [23] propose a

joint sub-band allocation and power optimization scheme to improve the spectrum efficiency for LTE-A when BS and relay transmit cooperatively.

In order to give a theoretic analysis of the system performance, stochastic geometry has been utilized as an effective tool for modeling and analyzing cellular networks; see, e.g., [24–26] and references therein. Generally, the spatial distribution of BSs is modeled as a spatial point process, such as the homogeneous Poisson point process (PPP) for single-tier networks, for which the coverage probability is derived in [27]. For HCNs, the spatial distribution of heterogeneous BSs is often modeled as multiple independent tiers of PPPs, and several key statistics are analyzed in [20, 28]. A comprehensive treatment of the application of stochastic geometry in wireless communication and content can be found in [29, 30].

## A User-Cognizant Solution

Most existing works on heterogenous networks have focused on admission control, resource allocation, and transmission coordination. The serving BS associated to a particular UE is assigned based on indicators of the wireless link quality at the UEs, such as the received signal strength indicators (RSSIs) or the SINRs. The same priority is allocated to each information bit for different video data flows in the scheduling stage. All these networking designs again verify that the current network is inefficient for video transmission. To enhance the QoE, one promising approach for efficient networking is by making the network better informed of its environment and user requirements.

Thus, considering scalable video transmission over HCNs, we have previously proposed two user-cognizant transmission schemes. In [31], the common layer structures of both video source data and network topology are employed to enhance the video transmission, based on the user's video requirement and association status. In [32], analog transmission of enhancement layer of video stream is proposed in order to make the video reception quality changing continuously with channel quality, thus degrading the staircase effect of digital transmission. Here the cooperative transmission is taken into consideration, where the macro BS and small BS work in a manner of coordinated multipoint transmission.

In all, three user-cognizant transmission schemes are proposed, which are layered digital transmission, layered hybrid digital-analog transmission, and cooperative digital transmission, respectively. The user is cognizant of its video service requirement, mobility, and connection status.

An analytical performance assessment of user-cognizant SVC transmission over two-tier HCNs utilizing tools from stochastic geometry is studied. The contributions of the are:

1. Three user-cognizant transmission schemes are proposed to enhance the QoE of video users exploiting the interplay among user demand, video source characteristic, and networking.

2. An analytical framework is proposed for scalable video transmission exploiting the common feature of a layered structure of SVC and HCNs. A digital and a hybrid digital-analog transmission scheme are proposed and studied. The impact of UE load, i.e., the number of UEs served in a cell, is also considered.
3. The power allocation between the digital BL signal and the analog EL signal is also analyzed to minimize the average distortion. The hybrid digital-analog scheme can further improve the system performance by avoiding the cliff effect and realizing continuous quality scalability when the proportion of frequency resource allocated to the femto tier exceeds a certain threshold.
4. A noncoherent joint transmission cooperative scheme is proposed, and moreover, the impact of sub-band allocation is also studied.

The remaining part of this paper is as follows: section “[System Model](#)” describes the system model, including the transmission schemes and spectrum allocation methods. Section “[UE Load and Sub-band Occupancy](#)” derives the distributions of the number of UEs per cell and sub-band occupancy probabilities. Section “[SINR Distribution and Data Rate](#)” derives the SINR and data rate distributions. Section “[System Performances](#)” evaluates the performance metrics, namely, outage probability, HD probability, and average distortion. Section “[Simulation and Discussion](#)” presents simulation results and related discussions. Section “[Conclusion and Future Directions](#)” concludes this paper.

---

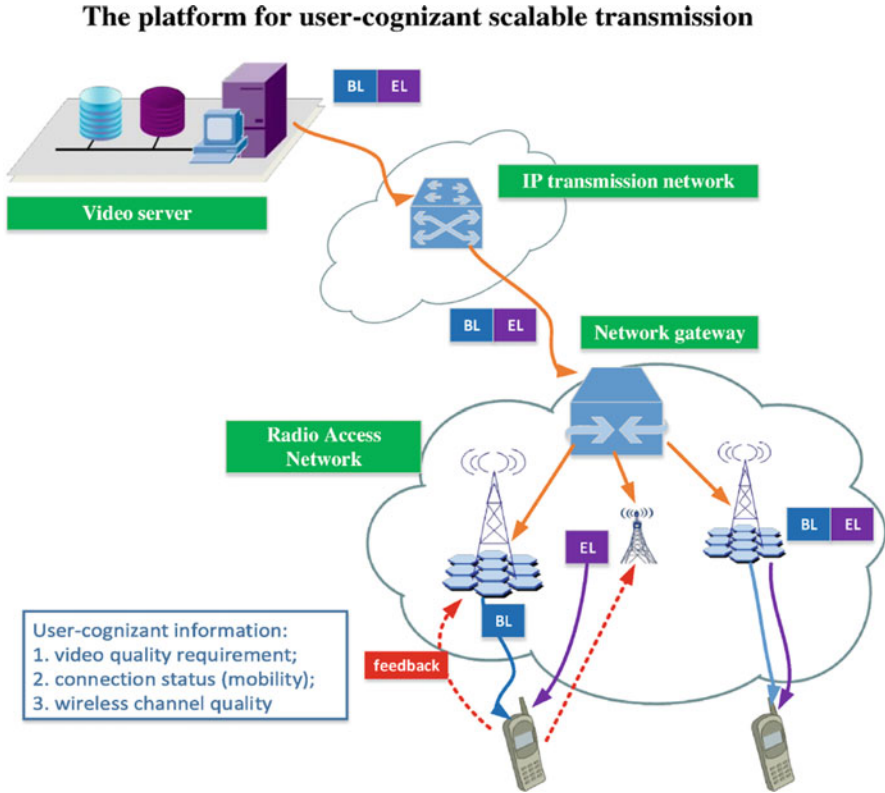
## System Model

The downlink performance of SVC over a two-tier HCN is considered; see Fig. 1. When a video user needs to request for a particular video, it would first collect the information about video quality requirement due to possessing ability of the UE and connection status due to its mobility and the network topology. The user sends the user-cognizant information to the serving BS (or BSs), and the BS or BSs choose an appropriate transmission strategy. The video data stream is traversing through the video server, IP transmission network, cellular core network, radio access network, and finally reaching at the user.

## Layered Video Model

The SVC video content is split into two layers, BL and EL. The BL is always modulated into a digital signal, and the data rate is  $R_B$ , while the EL is modulated into a digital signal or an analog signal in these transmission schemes accordingly. If the EL is modulated into a digital signal, then the data rate is  $R_E$ . Here we focus on the streaming video service; the video can be decoded successfully when the data rate requirements of the BL and the EL are met.

Actually, the proposed analytical framework can be extended to video signals that are encoded to  $J$  layers using a fine granularity, and the BS chooses the first



**Fig. 1** Illustration of the platform for user-cognizant scalable transmission

$J_1$  layers for the BL and the following  $J_2$  layers for the EL based on the channel quality for each UE, where  $J_1 + J_2 \leq J$ .

It should be aware that SVC allows three types of scalable encoding (spatial, temporal, SNR quality) to be combined and create a single layer [1,5]. The proposed layered video model is generic and is not restricted by the specifications of the layered encoding and the optimal selection of scalability combinations. Each layer is generated by some combinations of video scalabilities, and the required data rates are the main parameters from the view of networking.

### Network Model

The two-tier HCN consists of two types of BSs, namely, MBSs and FAPs. These two types of BSs are modeled by two independent tiers of homogeneous PPPs,  $\Phi_{mb}$  and  $\Phi_{fb}$ , whose intensities are  $\lambda_{mb}$  and  $\lambda_{fb}$ , respectively. FAPs aim at providing network access to UEs in their vicinity within a coverage radius  $R_f$ . Suppose that there exist



$N$  sub-bands each of bandwidth  $W$ . The transmit powers of an MBS and an FAP over each sub-band are set as  $P_m$  and  $P_f$ , respectively. The path loss model is  $r^{-\alpha}$ , and here for simplicity, it is assumed that the path loss exponent is the same for MBS and FAP, and the effect of shadowing is ignored. The small-scale fading distribution is exponential with mean unity in squared magnitude, i.e., Rayleigh fading. The fading is assumed to be frequency flat within each sub-band and independent among different sub-bands. The noise variance at each UE is denoted by  $\sigma^2$ .

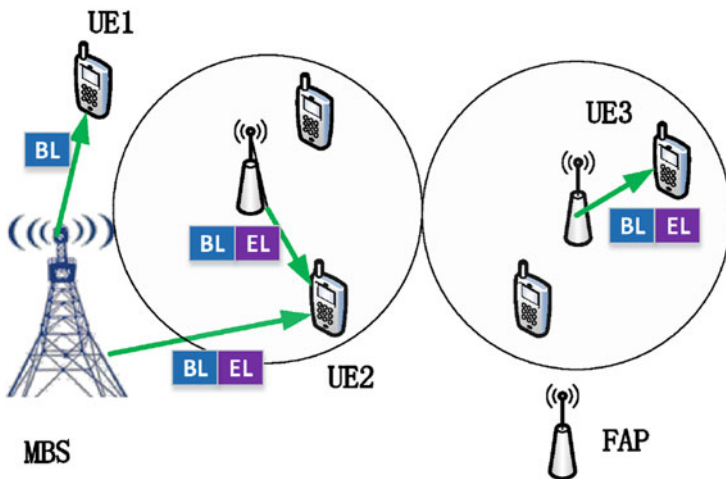
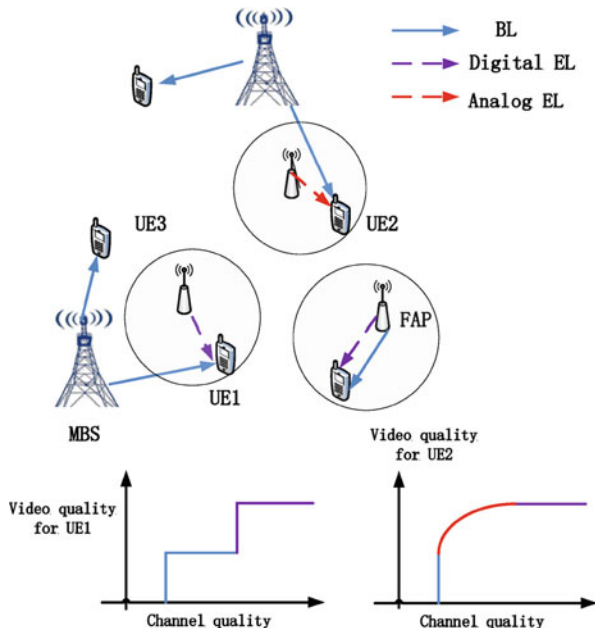
There are two types of UEs, macro UEs and femto UEs. The locations of macro UEs form a homogeneous PPP  $\Phi_{\text{mu}}$  with intensity  $\lambda_{\text{mu}}$ , and each macro UE connects to the nearest MBS. The locations of the femto UEs form a Matern cluster process  $\Phi_{\text{fu}}$  [30] with parent process  $\Phi_{\text{fb}}$  (the FAPs), i.e., the UEs in each cluster form a finite PPP of intensity  $\lambda_{\text{fu}}$  on the disk of radius  $R_f$  centered at each FAP, implying that the mean number of users per cluster is  $\bar{U}_f = \lambda_{\text{fu}}\pi R_f^2$ . Each femto UE connects to the FAP located at the parent point of the corresponding cluster, called the parent FAP. The access mechanism is as follows: a femto UE always connects to its parent FAP when accessing a femto BS and connects to the MBS closest to its parent FAP when accessing a macro BS; a macro UE can only connect to the nearest MBS, even if it is situated within the coverage of an FAP. This corresponds to a closed-access femto network, in which only subscribers are allowed to be served by an FAP.

## Transmission Schemes

Considering the connection status of UEs and their differentiated demand, three transmission schemes are proposed, i.e., layered digital (LD) [31], layered hybrid digital-analog (LHDA) [32], and cooperative digital (CD). For the macro UEs, they only connect to the MBS. Since the MBS aims at providing the coverage service, macro UEs attempt to obtain their BL contents from their serving MBSs and forego the EL. For the femto UEs, since they are covered by the MBS and the FAP, they have two choices: one is that they attempt to obtain their EL contents from their serving FAPs, and they attempt to obtain their BL contents either from their serving MBSs with probability  $p$  or from their serving FAPs with probability  $1 - p$ . The other one is that they receive the video contents which are transmitted cooperatively by the MBS and the FAP in the manner of CoMP.

1. *LD transmission*: See Fig. 2. Both the BL and the EL are modulated into digital signals. For a macro UE, the data stream of encoded BL signals is transmitted from the serving MBS. For a femto UE, the data stream of encoded BL signals for small SINR or jointly encoded signals of both the BL and the EL for large SINR is transmitted from the serving FAP when  $p = 0$ ; the digital BL data stream is transmitted from its serving MBS, while the digital EL data stream is transmitted from its serving FAP when  $p = 1$ ; a mixed transmission is adopted when  $0 < p < 1$ .
2. *LHDA transmission*: See Fig. 2. The BL is modulated into a digital signal, while the EL is modulated into an analog signal. For a macro UE, the data stream of

**Fig. 2** Qualitative illustration of the performances of LD and LHDA transmissions. LD transmission shows a staircase effect, while LHDA transmission shows continuous quality reception with respect to the channel quality



**Fig. 3** Illustration of the CD transmission model

encoded BL signals is transmitted from the serving MBS. For a femto UE, the superposition of the digital BL signal and the analog EL signal is transmitted from the serving FAP when  $p = 0$ ; the digital BL data stream is transmitted from its serving MBS, while the analog EL data stream is transmitted from its serving FAP when  $p = 1$ ; a mixed transmission is adopted when  $0 < p < 1$ .

3. *CD transmission*: See Fig. 3. Both the BL and the EL are modulated into digital signals. For a macro UE, the data stream of encoded BL signals is transmitted from the serving MBS. For a femto UE, if it can claim the same sub-band from both the MBS and the FAP, then the data stream of jointly encoded signals of both the BL and the EL is transmitted from the serving MBS and FAP cooperatively in the manner of noncoherent joint transmission, otherwise, the data stream of jointly encoded signals of both the BL and the EL is transmitted from the serving FAP.

Since the video source is encoded into multiple layers, different layers are transmitted to the UE based on the channel quality, thus providing scalable video quality. Specifically, for those UEs in less favorable conditions, only the BL with relatively low data rate is received in order to ensure basic video experience. When the channel quality improves, the EL is also received for enhanced video experience. Thus, the LD and CD transmissions can provide two-level scalable video for the UEs, and LHDA can provide a continuous quality scalability.

---

## UE Load and Sub-band Occupancy

### UE Load

Since the distribution of femto UEs in an FAP coverage disk is a PPP with intensity  $\lambda_{fu}$ , the number of femto UEs connected to an FAP is a Poisson random variable (r.v.) with mean  $\bar{U}_f$ ,

$$\mathbb{P}\{U_f = i\} = \frac{(\bar{U}_f)^i}{i!} e^{-\bar{U}_f}, \quad i = 0, 1, \dots \quad (1)$$

For LD and LHDA transmissions, an MBS not only serves the macro UEs situated in its Voronoi cell but also the femto UEs that belong to the FAPs in this Voronoi cell and connect to the MBS to receive the BL contents. We denote the number of macro UEs in the Voronoi cell as  $U_{MBS}$  and the total number of femto UEs served by the MBS as  $U_{FAP}$ , which is given by  $U_{FAP} = \sum_{i=1}^{N_c} N_{f,i}$ , where  $N_c$  denotes the number of the FAPs in the Voronoi cell and  $N_{f,i}$  denotes the number of femto UEs which belong to the  $i$ th FAP but connect to the MBS to receive the BL contents. The total number of UEs served by an MBS is thus

$$U_m = U_{MBS} + U_{FAP}. \quad (2)$$

$U_{MBS}$  is conditionally independent of  $U_{FAP}$  given the area of the Voronoi cell. Denote the area of a Voronoi cell by  $S$ ; the probability generating function (pgf) of  $U_m$  conditioned on  $S$ , denoted by  $G_m(z | S)$ , is

$$G_m(z | S) = G_{MBS}(z | S)G_{FAP}(z | S), \quad (3)$$

where  $G_{\text{MBS}}(z | S)$  and  $G_{\text{FAP}}(z | S)$  are the pgfs of  $U_{\text{MBS}}$  and  $U_{\text{FAP}}$  conditioned on  $S$ , respectively.

$U_{\text{MBS}}$  is a Poisson r.v. with mean  $\lambda_{\text{mu}}S$ , and the conditional pgf of  $U_{\text{MBS}}$  is

$$G_{\text{MBS}}(z | S) = e^{\lambda_{\text{mu}}S(z-1)}. \quad (4)$$

Since a femto UE attempts to connect to its serving MBS with probability  $p$  in LD and LHDA transmissions, a thinning occurs, i.e.,  $N_{f,i}$  is a Poisson random variable with mean  $p\bar{U}_f$ . Meanwhile,  $N_c$  is also a Poisson r.v. with mean  $\lambda_{\text{fb}}S$  because of the PPP distribution of the FAP locations.  $U_{\text{FAP}}$  is a compound Poisson r.v. with conditional pgf

$$G_{\text{FAP}}(z | S) = e^{\lambda_{\text{fb}}S(e^{p\bar{U}_f(z-1)}-1)}. \quad (5)$$

There is no known closed form expression of the probability density function (pdf) of the area  $S$  of the typical Poisson Voronoi cell, but the following approximation [33]

$$f_S(x) \approx \frac{(\lambda_{\text{mb}}c)^c}{\Gamma(c)} x^{c-1} e^{-c\lambda_{\text{mb}}x}, \quad (6)$$

where  $c = \frac{7}{2}$  and  $\Gamma(c) = \int_0^\infty t^{c-1} e^{-t} dt$ , has been known to be handy and sufficiently accurate (see, e.g., [34]). Aided by this approximation, with some manipulations, the pgf of  $U_m$  is

$$G_m(z) = c^c \left( c - \frac{\lambda_{\text{mu}}}{\lambda_{\text{mb}}}(z-1) + \frac{\lambda_{\text{fb}}}{\lambda_{\text{mb}}} \left( 1 - e^{p\bar{U}_f(z-1)} \right) \right)^{-c}, \quad (7)$$

and the distribution of  $U_m$  follows as

$$\mathbb{P}\{U_m = i\} = \frac{G_m^{(i)}(0)}{i!}, \quad i = 0, 1, \dots, \quad (8)$$

where  $G_m^{(i)}(0)$  is the  $i$ -th derivative of  $G_m(z)$  evaluated at  $z = 0$ .

For CD transmission, all the femto UEs attempt to connect to the MBS to obtain cooperative gain; thus distribution of  $U_m$  is similar to that in LD and LHDA transmissions with  $p = 1$ .

## Sub-band Occupancy

Since the number of served UEs for each BS is random, the sub-band frequency resource will be underutilized in some BSs and overutilized in some other BSs. As

the UE loads in the MBS and the FAP are different under different transmission schemes, the sub-band occupancy is calculated accordingly.

### Spectrum Allocation for LD and LHDA

Of the  $N$  sub-bands, let  $N_m$  sub-bands be allocated to the macro tier and  $N_f$  sub-bands to the femto tier. Each UE requires one sub-band for each transmission. We consider the following two spectrum allocation methods [35]:

1. *Orthogonal spectrum allocation:* The  $N$  sub-bands are split as  $N = N_m + N_f$ , where the  $N_m$  sub-bands used by all the MBSs of the macro tier are orthogonal to those  $N_f$  sub-bands used by all the FAPs of the femto tier. So there is no inter-tier interference.

It is assumed that the available sub-bands are uniformly and independently allocated to the UEs by the BS. There are  $N_m$  available sub-bands for the MBS, and each sub-band is equally likely to be chosen. If the number of UEs is smaller than that of sub-bands, the MBS randomly chooses  $U_m$  out of the total  $N_m$  sub-bands. Otherwise, all the sub-bands are chosen. The probability that a sub-band is used by an MBS is

$$P_{\text{busy}}^{\text{m},\perp} = \frac{1}{N_m} \sum_{i=0}^{\infty} \min\{i, N_m\} \mathbb{P}\{U_m = i\}, \quad (9)$$

and similarly the probability that a sub-band is used by an FAP is

$$P_{\text{busy}}^{\text{f},\perp} = \frac{1}{N_f} \sum_{i=0}^{\infty} \min\{i, N_f\} \mathbb{P}\{U_f = i\}. \quad (10)$$

2. *Non-orthogonal spectrum allocation:* Compared with the orthogonal case, here the two sets of sub-bands may overlap: each MBS (resp. FAP) independently randomly selects  $N_m$  (resp.  $N_f$ ) sub-bands from the  $N$  sub-bands. The values of both  $N_m$  and  $N_f$  can be chosen from 1 to  $N$  flexibly and need not add to  $N$ . So there is inter-tier interference, while the available spectrum will be abundant as  $N_m$  and  $N_f$  grow large.

For the non-orthogonal case, both the MBS and the FAP choose a sub-band randomly from  $N$  sub-bands, so the probability that a sub-band is used by an MBS is

$$P_{\text{busy}}^{\text{m},\mathcal{L}} = \frac{1}{N} \sum_{i=0}^{\infty} \min\{i, N_m\} \mathbb{P}\{U_m = i\}, \quad (11)$$

and similarly the probability that a sub-band is used by an FAP is

$$P_{\text{busy}}^{\text{f},\mathcal{L}} = \frac{1}{N} \sum_{i=0}^{\infty} \min\{i, N_f\} \mathbb{P}\{U_f = i\}. \quad (12)$$

The spatial point process of BSs that use a given sub-band is an approximately independent thinning of the original point process  $\Phi_{\text{mb}}$  (resp.  $\Phi_{\text{fb}}$ ) by the probability  $P_{\text{busy}}^{\text{m},s}$  (resp.  $P_{\text{busy}}^{\text{f},s}$ ), denoted by  $\tilde{\Phi}_{\text{mb}}$  (resp.  $\tilde{\Phi}_{\text{fb}}$ ) with the intensity  $\tilde{\lambda}_{\text{mb}} = \lambda_{\text{mb}} P_{\text{busy}}^{\text{m},s}$  (resp.  $\tilde{\lambda}_{\text{fb}} = \lambda_{\text{mb}} P_{\text{busy}}^{\text{m},s}$ ) [34], where the superscript  $s \in \{\perp, \not\perp\}$  indicates whether the orthogonal or the non-orthogonal spectrum allocation method is used.

### Spectrum Allocation for CD

Since both the macro UEs and femto UEs connect to the MBS, if the number of UEs connected to MBS  $U_{\text{m}} \leq N$ , each UE is allocated one sub-band, otherwise, if  $U_{\text{m}} \geq N$ , all the UEs share the sub-bands in a round-robin mechanism. Then the FAPs allocate sub-bands to femto UEs in a similar way by comparing  $U_{\text{f}}$  and  $N$ . For a femto UE, if it is chosen by both the MBS and the FAP, the MBS and the FAP allocate the same sub-band to the UE, thus making it working in a cooperative mode. Otherwise it is only served by the FAP and works in a noncooperative mode.

Since both the macro and femto tiers employ the total  $N$  sub-bands, similar to that of the LD case, the probability that a sub-band is used by an MBS is

$$P_{\text{busy}}^{\text{m,CoMP}} = \frac{1}{N} \sum_{i=0}^{\infty} \min\{i, N_{\text{m}}\} \mathbb{P}\{U_{\text{m}} = i\}, \quad (13)$$

and similarly the probability that a sub-band is used by an FAP is

$$P_{\text{busy}}^{\text{f,CoMP}} = \frac{1}{N} \sum_{i=0}^{\infty} \min\{i, N_{\text{f}}\} \mathbb{P}\{U_{\text{f}} = i\}. \quad (14)$$

---

## SINR Distribution and Data Rate

### SINR Distribution

The complementary cumulative distribution function (ccdf) of the SINR is defined as  $\mathcal{P}(\theta) = \mathbb{P}\{\text{SINR} > \theta\}$ , where  $\theta$  is the SINR threshold. The SINR distributions of a UE connected to the MBS and the FAP are derived under three transmission schemes.

### LD Transmission

For analytical tractability, we assume that both the BL and the EL are modulated into digital signals according to a Gaussian codebook.

For the typical UE which is assumed to be located at the origin and connected to its MBS, the received signal denoted by  $Y$  can be written as

$$\begin{aligned}
Y &= P_m^{1/2} \|x_0\|^{-\alpha/2} h_{x_0} X_{x_0} + \sum_{x \in \tilde{\Phi}_{\text{mb}} \setminus \{x_0\}} P_m^{1/2} \|x\|^{-\alpha/2} h_x X_x \\
&+ \kappa \sum_{y \in \tilde{\Phi}_{\text{fb}}} P_f^{1/2} \|y\|^{-\alpha/2} h_y X_y + Z,
\end{aligned} \tag{15}$$

where the first item of right side of the equation denotes the received signal symbol, the second and the third items denote the interference symbols from the macro and the femto tier, respectively, and  $Z$  denotes the Gaussian noise with zero mean and variance  $\sigma^2$ . We use  $x_0$  to denote the location of the serving MBS.  $X_{x_0}$  is the signal symbol, while  $X_x$  is the interference symbol transmitted by the interfering MBS  $x$ .  $X_{x_0}, X_x \sim \mathcal{CN}(0, 1)$ .  $X_y$  is the interference symbol transmitted by the interfering FAP  $y$ , and  $X_y \sim \mathcal{CN}(0, 1)$ . The indicator  $\kappa \in \{0, 1\}$  indicates the orthogonal and non-orthogonal spectrum allocation methods, respectively.

Thus, the received SINR is

$$\gamma_{\text{LD}}^{\text{m}} = \frac{P_m \|x_0\|^{-\alpha} |h_{x_0}|^2}{I_m + \kappa I_f + \sigma^2}, \tag{16}$$

where  $I_m = \sum_{x \in \tilde{\Phi}_{\text{mb}} \setminus \{x_0\}} P_m \|x\|^{-\alpha} |h_x|^2$  is the interference from the macro tier, and  $I_f = \sum_{y \in \tilde{\Phi}_{\text{fb}}} P_f \|y\|^{-\alpha} |h_y|^2$  is the interference from the femto tier.

For the typical femto UE which is assumed to be located at the origin and connected to its FAP, the received signal can be written as

$$\begin{aligned}
Y &= P_f^{1/2} \|y_0\|^{-\alpha/2} h_{y_0} X_{y_0} + \sum_{y \in \tilde{\Phi}_{\text{fb}} \setminus \{y_0\}} P_f^{1/2} \|y\|^{-\alpha/2} h_y X_y \\
&+ \kappa \sum_{x \in \tilde{\Phi}_{\text{mb}}} P_m^{1/2} \|x\|^{-\alpha/2} h_x X_x + Z,
\end{aligned} \tag{17}$$

where  $y_0$  denotes the location of the serving FAP. Note that the FAP transmits the encoded EL signals only or the jointly encoded signals of both the BL and the EL to the typical UE based on user request.  $X_{y_0}$  is the signal symbol transmitted by the serving FAP, and  $X_y$  is the interference symbol transmitted by the interfering FAP  $y$ .

Thus, the received SINR is

$$\gamma_{\text{LD}}^{\text{f}} = \frac{P_f \|y_0\|^{-\alpha} |h_{y_0}|^2}{I_f + \kappa I_m + \sigma^2}, \tag{18}$$

where  $I_f = \sum_{y \in \tilde{\Phi}_{\text{fb}} \setminus \{y_0\}} P_f \|y\|^{-\alpha} |h_y|^2$  denotes the interference from the femto tier and  $I_m = \sum_{x \in \tilde{\Phi}_{\text{mb}}} P_m \|x\|^{-\alpha} |h_x|^2$  denotes the interference from the macro tier.

The following theorem gives the cdf of the SINR for the typical UE,

**Theorem 1.** *For LD transmission, the cdf of the SINR for the typical UE connected to its serving MBS is*

$$\begin{aligned}
\mathcal{P}_{\text{LD}}^{\text{m}}(\theta) &= \mathbb{P}\{\gamma_{\text{LD}}^{\text{m}} > \theta\} \\
&= \int_0^\infty \pi \lambda_{\text{mb}} \exp\left(-\pi v(\lambda_{\text{mb}} + \tilde{\lambda}_{\text{mb}}\rho(\theta, \alpha))\right. \\
&\quad \left.- \frac{\theta v^{1/\delta}\sigma^2}{P_{\text{m}}} - \kappa \left(\frac{P_{\text{f}}\theta}{P_{\text{m}}}\right)^\delta v \tilde{\lambda}_{\text{fb}}\delta\pi^2 \text{csc}(\delta\pi)\right) \text{d}v, \tag{19}
\end{aligned}$$

and the ccdf of the SINR for the typical femto UE connected to its serving FAP is

$$\begin{aligned}
\mathcal{P}_{\text{LD}}^{\text{f}}(\theta) &= \mathbb{P}\{\gamma_{\text{LD}}^{\text{f}} > \theta\} \\
&= \int_0^{R_{\text{f}}^2} \frac{1}{R_{\text{f}}^2} \exp\left(-\frac{\theta v^{1/\delta}\sigma^2}{P_{\text{f}}} - \delta\pi^2 \text{csc}(\delta\pi)\theta^\delta v \left(\tilde{\lambda}_{\text{fb}} + \kappa \left(\frac{P_{\text{m}}}{P_{\text{f}}}\right)^\delta\right) \tilde{\lambda}_{\text{mb}}\right) \text{d}v, \tag{20}
\end{aligned}$$

where  $\delta = 2/\alpha$ ,  $\tilde{\lambda}_{\text{mb}} = \lambda_{\text{mb}} P_{\text{busy}}^{\text{m,s}}$ ,  $\tilde{\lambda}_{\text{fb}} = \lambda_{\text{fb}} P_{\text{busy}}^{\text{f,s}}$ , and  $\rho(\theta, \alpha) = \theta^\delta \int_{\theta^{-\delta}}^\infty \frac{1}{1+x^{1/\delta}} \text{d}x$ . In orthogonal spectrum allocation,  $\kappa = 0$ , while in non-orthogonal spectrum allocation,  $\kappa = 1$ .

*Proof.* Let  $\|x_0\|$  be the distance from the typical UE to its serving MBS, which is the nearest MBS, so the pdf of  $\|x_0\|$  is  $f_{\|x_0\|}(r) = e^{-\lambda_{\text{mb}}\pi r^2} 2\pi \lambda_{\text{mb}} r$ .

The SINR experienced by the typical UE connected to its serving MBS is given by  $\gamma_{\text{LD}}^{\text{m}} = \frac{P_{\text{m}}\|x_0\|^{-\alpha}|h_{x_0}|^2}{I_{\text{m}} + \kappa I_{\text{f}} + \sigma^2}$ , where  $I_{\text{m}} = \sum_{x \in \tilde{\Phi}_{\text{mb}} \setminus \{x_0\}} P_{\text{m}}\|x\|^{-\alpha}|h_{x_0}|^2$  is the interference from the macro tier, and  $I_{\text{f}} = \sum_{y \in \tilde{\Phi}_{\text{fb}}} P_{\text{f}}\|y\|^{-\alpha}|h_y|^2$  is the interference from the femto tier.  $\kappa \in \{0, 1\}$  is the indicator that whether the orthogonal or the non-orthogonal spectrum allocation is used. Due to the independent thinning approximation, the set of interfering MBSs is a PPP  $\tilde{\Phi}_{\text{mb}}$  with intensity  $\tilde{\lambda}_{\text{mb}}$ , and the set of interfering FAPs is a PPP  $\tilde{\Phi}_{\text{fb}}$  with intensity  $\tilde{\lambda}_{\text{fb}}$ .

The ccdf of the SINR experienced by the typical UE connected to its serving MBS

$$\begin{aligned}
\mathcal{P}_{\text{LD}}^{\text{m}}(\theta) &= \mathbb{P}\{\gamma_{\text{LD}}^{\text{m}} > \theta\} \\
&= \int_0^\infty 2\pi \lambda_{\text{mb}} r e^{-\pi \lambda_{\text{mb}} r^2} \mathbb{P}\left\{\frac{P_{\text{m}}|h_{x_0}|^2 r^{-\alpha}}{I_{\text{m}} + \kappa I_{\text{f}} + \sigma^2} > \theta\right\} \text{d}r \\
&\stackrel{(a)}{=} \int_0^\infty 2\pi \lambda_{\text{mb}} r e^{-\pi \lambda_{\text{mb}} r^2 - \frac{\theta r^\alpha \sigma^2}{P_{\text{m}}}} \mathcal{L}_{I_{\text{m}} + \kappa I_{\text{f}}}\left(\frac{\theta r^\alpha}{P_{\text{m}}}\right) \text{d}r. \tag{21}
\end{aligned}$$

where (a) follows from  $|h_{x_0}|^2 \sim \text{Exp}(1)$ .

After excluding the serving BS  $x_0$ ,  $\tilde{\Phi}_{\text{mb}} \setminus \{x_0\}$  is still a PPP, so we apply the pgf of PPP to obtain the Laplace transform of  $I_{\text{m}}$



$$\begin{aligned}\mathcal{L}_{I_m}(s) &= \exp\left(-2\pi\tilde{\lambda}_{mb}\int_r^\infty\left(1-\frac{1}{1+sP_mx^{-\alpha}}\right)xdx\right) \\ &= e^{-\pi\tilde{\lambda}_{mb}r^2\rho\left(\frac{sP_m}{r^\alpha},\alpha\right)}.\end{aligned}\quad (22)$$

Since  $\tilde{\Phi}_{fb}$  is a PPP, the Laplace transform of  $I_f$  is

$$\begin{aligned}\mathcal{L}_{I_f}(s) &= \exp\left(-2\pi\tilde{\lambda}_{fb}\int_0^\infty\left(1-\frac{1}{1+sP_fx^{-\alpha}}\right)xdx\right) \\ &= e^{-\delta\pi^2\csc(\delta\pi)\tilde{\lambda}_{fb}(sP_f)^\delta}.\end{aligned}\quad (23)$$

Substituting (22) and (23) into  $\mathcal{P}_{LD}^m(\theta)$ , we can obtain (19).

Let  $y_0$  be the distance between the typical femto UE and its serving FAP. Since femto UEs are uniformly distributed in the circular coverage area of radius  $R_f$  of each FAP, the pdf of  $y_0$  is given by  $f_{y_0}(r) = \frac{2r}{R_f^2}$ .

The received SINR for the typical femto UE connected to its serving FAP follows as  $\gamma_{LD}^f = \frac{P_f\|y_0\|^{-\alpha}|h_{y_0}|^2}{I_f + \kappa I_m + \sigma^2}$ , where  $I_f = \sum_{y \in \tilde{\Phi}_{fb}} P_f \|y\|^{-\alpha} |h_y|^2$  is the interference from the femto tier, and  $I_m = \sum_{x \in \tilde{\Phi}_{mb}} P_m \|x\|^{-\alpha} |h_{x_0}|^2$  is the interference from the macro tier.

The ccdf of the SINR experienced by the typical femto UE connected to its serving FAP is

$$\begin{aligned}\mathcal{P}_{LD}^f(\theta) &= \mathbb{P}\{\gamma_{LD}^f > \theta\} \\ &= \int_0^{R_f} \frac{2r}{R_f^2} \mathbb{P}\left\{\frac{P_f|h_{y_0}|^2 r^{-\alpha}}{I_f + \kappa I_m + \sigma^2} > \theta\right\} dr \\ &= \int_0^{R_f} \frac{2r}{R_f^2} e^{-\frac{\theta r^\alpha \delta^2}{P_f}} \mathcal{L}_{I_f + \kappa I_m}\left(\frac{\theta r^\alpha}{P_f}\right) dr,\end{aligned}\quad (24)$$

which, after expanding the Laplace transform of  $I_m$ ,  $I_f$ , and further manipulations, leads to (20).

### LHDA Transmission

The BL is modulated to a digital signal, while the EL is modulated to an analog signal. The digital modulation is based on a Gaussian codebook, and the EL signal after analog modulation is also modeled as a Gaussian source with zero mean and unit variance [36, 37]. For analog modulation, it is assumed that the source bandwidth is equal to the channel bandwidth [14, 16].

For the typical UE which is assumed to be located at the origin and connected to its serving MBS, the received signal can be written as

$$\begin{aligned}
Y &= P_m^{1/2} \|x_0\|^{-\alpha/2} h_{x_0} X_{x_0} + \sum_{x \in \tilde{\Phi}_{mb} \setminus \{x_0\}} P_m^{1/2} \|x\|^{-\alpha/2} h_x X_x \\
&+ \kappa \sum_{y \in \tilde{\Phi}_{fb}} P_f^{1/2} \|y\|^{-\alpha/2} h_y X_y + Z,
\end{aligned} \tag{25}$$

which is nearly the same as (15) in LD transmission, the difference lies in that  $X_y$  is the analog EL interference symbol or the superposition of digital BL and analog EL interference symbol transmitted by the interfering FAP  $y$  based on the transmission scheme of  $y$ , and  $X_y \sim \mathbb{CN}(0, 1)$ .

Thus the received SINR is

$$\gamma_{\text{LHDA}}^m = \frac{P_m \|x_0\|^{-\alpha} |h_{x_0}|^2}{I_m + \kappa I_f + \sigma^2}, \tag{26}$$

where  $I_m = \sum_{x \in \tilde{\Phi}_{mb} \setminus \{x_0\}} P_m \|x\|^{-\alpha} |h_x|^2$  is the interference from the macro tier and  $I_f = \sum_{y \in \tilde{\Phi}_{fb}} P_f \|y\|^{-\alpha} |h_y|^2$  is the interference from the femto tier.

For the typical femto UE which is assumed to be located at the origin and connected to its FAP, according to the transmission scheme, it receives only the EL, or it receives the superposition of the digital BL signal and the analog EL signal.

- *Case 1:* The typical femto UE connected to its FAP receives only the EL. The received signal for the typical femto UE is

$$\begin{aligned}
Y &= P_f^{1/2} \|y_0\|^{-\frac{\alpha}{2}} h_{y_0} X_{y_0}^E + \sum_{y \in \tilde{\Phi}_{fb} \setminus \{y_0\}} P_f^{1/2} \|y\|^{-\frac{\alpha}{2}} h_y X_y \\
&+ \kappa \sum_{x \in \tilde{\Phi}_{mb}} P_m^{1/2} \|x\|^{-\frac{\alpha}{2}} h_x X_x + Z,
\end{aligned} \tag{27}$$

where  $X_{y_0}^E$  is the EL signal symbol transmitted by the serving FAP and  $X_y$  is the interference symbol transmitted by the interfering FAP  $y$ .

Thus, the received SINR for the femto UE connected to its FAP to receive the EL is

$$\gamma_{\text{LHDA}}^f = \frac{P_f \|y_0\|^{-\alpha} |h_{y_0}|^2}{I_f + \kappa I_m + \sigma^2}, \tag{28}$$

where  $I_f = \sum_{y \in \tilde{\Phi}_{fb} \setminus \{y_0\}} P_f \|y\|^{-\alpha} |h_y|^2$  is the interference from the femto tier and  $I_m = \sum_{x \in \tilde{\Phi}_{mb}} P_m \|x\|^{-\alpha} |h_x|^2$  is the interference from the macro tier.

- *Case 2:* The typical femto UE connected to its FAP receives the superposition of the digital BL signal and the analog EL signal. The received signal for the typical femto UE is

$$\begin{aligned}
Y &= \|y_0\|^{-\alpha/2} h_{y_0} \left( \sqrt{P_f^B} X_{y_0}^B + \sqrt{P_f^E} X_{y_0}^E \right) + \sum_{y \in \Phi_{fb} \setminus \{y_0\}} P_f^{1/2} \|y\|^{-\alpha/2} h_y X_y \\
&+ \kappa \sum_{x \in \tilde{\Phi}_{mb}} P_m^{1/2} \|x\|^{-\alpha/2} h_x X_x + Z,
\end{aligned} \tag{29}$$

where  $X_{y_0}^B$  is the BL signal symbol transmitted by the serving FAP, and  $X_{y_0}^E$  is the EL signal symbol transmitted by the serving FAP.

Thus, the received SINR for the typical femto UE connected to its FAP to receive the BL, denoted by  $\gamma_{\text{LHDA}}^{f,B}$ , is

$$\gamma_{\text{LHDA}}^{f,B} = \frac{P_f^B \|y_0\|^{-\alpha} |h_{y_0}|^2}{P_f^E \|y_0\|^{-\alpha} |h_{y_0}|^2 + I_f + \kappa I_m + \sigma^2}. \tag{30}$$

Successive interference cancellation (SIC) [38] is adopted to demodulate the EL signal. Conditioned on the successful reception of the BL, the received SINR for the typical femto UE connected to the FAP to receive the EL signal, denoted by  $\gamma_{\text{LHDA}}^{f,E}$ , is

$$\gamma_{\text{LHDA}}^{f,E} = \frac{P_f^E \|y_0\|^{-\alpha} |h_{y_0}|^2}{I_f + \kappa I_m + \sigma^2}. \tag{31}$$

The following theorem gives the ccdf of the SINR for the typical UE:

**Theorem 2.** *For LHDA transmission, the ccdf of the SINR for the typical UE connected to its serving MBS is*

$$\mathcal{P}_{\text{LHDA}}^m(\theta_B) = \mathbb{P}\{\gamma_{\text{LHDA}}^m > \theta_B\} = \mathcal{P}_{\text{LD}}^m(\theta_B); \tag{32}$$

*the ccdf of the SINR for the typical femto UE connected to its serving FAP to receive the EL is given by:*

$$\mathcal{P}_{\text{LHDA}}^f(\theta_E) = \mathbb{P}\{\gamma_{\text{LHDA}}^f > \theta_E\} = \mathcal{P}_{\text{LD}}^f(\theta_E), \tag{33}$$

*and the joint ccdf of the SINR for the typical femto UE connected to its serving FAP to receive the superposition of the digital BL and the analog EL is given by (40).*

*Proof.* Similar to the derivation of  $\mathcal{P}_{\text{LD}}^m(\theta)$ , the ccdf of  $\gamma_{\text{LHDA}}^m$  follows as:

$$\mathcal{P}_{\text{LHDA}}^m(\theta) = \mathbb{P}\{\gamma_{\text{LHDA}}^m > \theta\} = \mathcal{P}_{\text{LD}}^m(\theta). \tag{34}$$

According to the transmission scheme, the FAP transmits the analog EL signal with probability  $p$  or the superposition of the digital BL signal and the analog EL signal with probability  $1 - p$ .

*Case 1:* The received SINR for the typical femto UE connected to the FAP receives the EL follows as  $\gamma_{\text{LHDA}}^f = \frac{P_f \|y_0\|^{-\alpha} |h_{y_0}|^2}{I_f + \kappa I_m + \sigma^2}$ . Similar to the derivation of  $\mathcal{P}_{\text{LD}}^f(\theta)$ , the ccdf of  $\gamma_{\text{LHDA}}^f$  follows as

$$\mathcal{P}_{\text{LHDA}}^f(\theta) = \mathbb{P}\{\gamma_{\text{LHDA}}^f > \theta\} = \mathcal{P}_{\text{LD}}^f(\theta). \quad (35)$$

*Case 2:* The received SINR for the typical femto UE connected to the FAP receives the superposition of the digital BL signal and the analog EL signal is

$$\gamma_{\text{LHDA}}^{f,B} = \frac{P_f^B \|y_0\|^{-\alpha} |h_{y_0}|^2}{P_f^E \|y_0\|^{-\alpha} |h_{y_0}|^2 + I_f + \kappa I_m + \sigma^2}, \quad (36)$$

where  $P_f^E \|y_0\|^{-\alpha} |h_{y_0}|^2$  is the interference of the superposed EL.

$$\begin{aligned} \mathbb{P}\{\gamma_{\text{LHDA}}^{f,B} > \theta\} &= \int_0^\infty \frac{2r}{R_f^2} e^{-\frac{\theta r^\alpha \sigma^2}{P_f^B - \theta P_f^E}} \mathcal{L}_{I_f + \kappa I_m} \left( \frac{\theta r^\alpha}{P_f^B - \theta P_f^E} \right) dr \\ &= \int_0^{R_f^2} \frac{1}{R_f^2} \exp \left( -\frac{\theta v^{1/\delta} \sigma^2}{P_f^B - \theta P_f^E} - \delta \pi^2 \csc(\delta \pi) \theta^\delta v \right. \\ &\quad \left. \times \left( \tilde{\lambda}_{\text{fb}} \left( \frac{P_f}{P_f^B - \theta P_f^E} \right)^\delta + \kappa \tilde{\lambda}_{\text{mb}} \left( \frac{P_m}{P_f^B - \theta P_f^E} \right)^\delta \right) \right) dv. \end{aligned} \quad (37)$$

SIC is adopted to decode the EL signal. After successful reception of the BL, the received SINR for the EL signal is  $\gamma_{\text{LHDA}}^{f,E} = \frac{P_f^E \|y_0\|^{-\alpha} |h_{y_0}|^2}{I_f + \kappa I_m + \sigma^2}$ .

The ccdf of  $\gamma_{\text{LHDA}}^{f,E}$  follows as:

$$\begin{aligned} \mathbb{P}\{\gamma_{\text{LHDA}}^{f,E} > \theta\} &= \int_0^\infty \frac{2r}{R_f^2} e^{-\frac{\theta r^\alpha \sigma^2}{P_f^E}} \mathcal{L}_{I_f + \kappa I_m} \left( \frac{\theta r^\alpha}{P_f^E} \right) dr \\ &= \int_0^{R_f^2} \frac{1}{R_f^2} e^{-\frac{\theta v^{1/\delta} \sigma^2}{P_f^E} - \delta \pi^2 \csc(\delta \pi) \theta^\delta v} \left( \tilde{\lambda}_{\text{fb}} \left( \frac{P_f}{P_f^E} \right)^\delta + \kappa \tilde{\lambda}_{\text{mb}} \left( \frac{P_m}{P_f^E} \right)^\delta \right) dv. \end{aligned} \quad (38)$$

The joint ccdf of  $\gamma_{\text{LHDA}}^{f,B}$  and  $\gamma_{\text{LHDA}}^{f,E}$  is

$$\begin{aligned} \mathcal{P}_{\text{LHDA}}(\theta_B, \theta_E) &= \mathbb{P}\{\gamma_{\text{LHDA}}^{f,B} > \theta_B, \gamma_{\text{LHDA}}^{f,E} > \theta_E\} \\ &= \mathbf{1} \left( \theta_B > \frac{\theta_E P_f^B}{(1 + \theta_E) P_f^E} \right) \int_0^{R_f^2} \end{aligned}$$

$$\begin{aligned}
& \frac{1}{R_f^2} e^{-\frac{\theta_B v^{1/\delta} \sigma^2}{P_f^B - \theta_B P_f^E} - \delta \pi^2 \csc(\delta \pi) \theta_B^\delta v} \left( \frac{\tilde{\lambda}_{fb} P_f^\delta + \kappa \tilde{\lambda}_{mb} P_m^\delta}{(P_f^B - \theta_B P_f^E)^\delta} \right) dv \\
& + \mathbf{1} \left( \theta_B \leq \frac{\theta_E P_f^B}{(1 + \theta_E) P_f^E} \right) \int_0^{R_f^2} \\
& \frac{1}{R_f^2} e^{-\frac{\theta_E v^{1/\delta} \sigma^2}{P_f^E} - \delta \pi^2 \csc(\delta \pi) \theta_E^\delta v} \left( \tilde{\lambda}_{fb} \left( \frac{P_f}{P_f^E} \right)^\delta + \kappa \tilde{\lambda}_{mb} \left( \frac{P_m}{P_f^E} \right)^\delta \right) dv. \quad (39)
\end{aligned}$$

$$\begin{aligned}
\mathcal{P}_{\text{LHDA}}(\theta_B, \theta_E) &= \mathbb{P}\{\gamma_{\text{LHDA}}^{\text{f,B}} > \theta_B, \gamma_{\text{LHDA}}^{\text{f,E}} > \theta_E\} \\
&= \mathbb{P}\left\{ |h_{x0}|^2 > \frac{\theta_B I_{\text{total}}}{(P_B - \theta_B P_E) \|x_0\|^{-\alpha}}, |h_{x0}|^2 > \frac{\theta_E I_{\text{total}}}{P_E \|x\|^{-\alpha}} \right\} \\
&= \mathbb{P}\left\{ |h_{x0}|^2 > \max\left( \frac{\theta_B I_{\text{total}}}{(P_B - \theta_B P_E) \|x_0\|^{-\alpha}}, \frac{\theta_E I_{\text{total}}}{P_E \|x\|^{-\alpha}} \right) \right\} \\
&= \mathbb{P}\{\gamma_{\text{LHDA}}^{\text{f,B}} > \theta_B\} \mathbf{1}\left(\theta_B > \frac{\theta_E P_f^B}{(1 + \theta_E) P_f^E}\right) \\
&\quad + \mathbb{P}\{\gamma_{\text{LHDA}}^{\text{f,E}} > \theta_E\} \mathbf{1}\left(\theta_B \leq \frac{\theta_E P_f^B}{(1 + \theta_E) P_f^E}\right), \quad (40)
\end{aligned}$$

where  $I_{\text{total}} = I_f + \kappa I_m + \sigma^2$ .

### CD Transmission

For the macro UE, the video transmission from the serving MBS is the same as that of LD; thus the ccdf of the SINR denoted by  $\mathcal{P}_{\text{CD}}^m(\theta)$  is equal to  $\mathcal{P}_{\text{LD}}^m(\theta)$  with  $p = 1$  and  $\kappa = 1$ .

For the femto UE, based on the sub-band allocation from the MBS and the FAP, it can work in a cooperative or noncooperative modes.

- In the noncooperative case, since the femto UE can only connect to the FAP, the received signal can be written as

$$\begin{aligned}
Y &= P_f^{1/2} \|y_0\|^{-\frac{\alpha}{2}} h_{y_0} X_{y_0} + \sum_{y \in \tilde{\Phi}_{fb} \setminus \{y_0\}} P_f^{1/2} \|y\|^{-\frac{\alpha}{2}} h_y X_y \\
&\quad + \sum_{x \in \tilde{\Phi}_{mb}} P_m^{1/2} \|x\|^{-\frac{\alpha}{2}} h_x X_x + Z, \quad (41)
\end{aligned}$$

Thus, the received SINR is

$$\gamma_{CD}^{f,\text{non}} = \frac{P_f \|y_0\|^{-\alpha} |h_{y_0}|^2}{I_f + I_m + \sigma^2}. \tag{42}$$

The ccdf of the  $\gamma_{CD}^{f,\text{non}}$ , denoted as  $\mathcal{P}_{CD}^{f,\text{non}}(\theta)$ , is equal to  $\mathcal{P}_{LD}^f(\theta)$  with  $p = 1$  and  $\kappa = 1$ .

- In the cooperative case, since the femto UE is served jointly by the MBS and the FAP, the received signal can be written as

$$Y = P_m^{1/2} \|x_0\|^{-\alpha/2} h_{x_0} X_0 + P_f^{1/2} \|y_0\|^{-\alpha/2} h_{y_0} X_0 + \sum_{x \in \tilde{\Phi}_{mb} \setminus \{x_0\}} P_m^{1/2} \|x\|^{-\alpha/2} h_x X_x + \sum_{y \in \tilde{\Phi}_{fb} \setminus \{y_0\}} P_f^{1/2} \|y\|^{-\alpha/2} h_y X_y + Z, \tag{43}$$

where the first and second items of right side of the equation denote the received signal symbols from the serving MBS and FAP, respectively, and the following two items denote the interference symbols from the macro and femto tiers, respectively

The noncoherent joint transmission is adopted, and the SINR of the received signal is

$$\gamma_{CD}^{f,\text{CoMP}} = \frac{\left| P_m^{1/2} \|x_0\|^{-\alpha/2} h_{x_0} + P_f^{1/2} \|y_0\|^{-\alpha/2} h_{y_0} \right|^2}{I_m + I_f + \sigma^2}, \tag{44}$$

where the interference from the macro tier is  $I_m = \sum_{x \in \tilde{\Phi}_{mb} \setminus \{x_0\}} P_m \|x\|^{-\alpha} h_x$ , and the interference from the femto tier is  $I_f = \sum_{y \in \tilde{\Phi}_{fb} \setminus \{y_0\}} P_f \|y\|^{-\alpha} h_y$ .

**Theorem 3.** For CD transmission, when the femto UE works in a cooperative mode, the ccdf of the SINR is

$$\begin{aligned} &\mathcal{P}_{CD}^{f,\text{CoMP}}(\theta) \\ &= \int_0^\infty \int_0^{R_f} e^{-\frac{\mu\theta\sigma^2}{P_m r_m^{-\alpha} + P_f r_f^{-\alpha}} - \pi \tilde{\lambda}_{mb} r_m^2 \rho \left( \frac{\theta}{P_f r_f^{-\alpha}} \cdot \alpha \right) - \pi \tilde{\lambda}_{fb} r_f^2 \rho \left( \frac{\theta}{P_m r_m^{-\alpha}} \cdot \alpha \right)} - \lambda_{mb} \pi r_m^2 \\ &\quad \times 2\pi \lambda_{mb} r_m \frac{2r_f}{R_f^2} dr_m dr_f. \end{aligned} \tag{45}$$

*Proof.* Let  $\|x_0\|$  be the distance from the typical femto UE to its serving MBS, and the pdf of  $\|x_0\|$  is  $f_{\|x_0\|}(r_m) = e^{-\lambda_{mb}\pi r_m^2} 2\pi \lambda_{mb} r_m$ .

Let  $\|y_0\|$  be the distance from the typical femto UE to its serving FAP, and the pdf of  $r_f$  is  $f_{\|y_0\|}(r_f) = \frac{2r_f}{R_f^2}$ .

The cdf of SINR  $\gamma_{CD}^{f,CoMP}$  is

$$\begin{aligned} \mathcal{P}_f^{\text{CoMP}}(\theta) &= \mathbb{P}\{\gamma_f^{\text{CoMP}} > \theta\} \\ &= \mathbb{P}\left\{\left|P_m^{\frac{1}{2}}r_m^{-\frac{\alpha}{2}}h_m + P_f^{\frac{1}{2}}r_f^{-\frac{\alpha}{2}}h_f\right|^2 > \theta(I_m + I_f + \sigma^2)\right\} \\ &\stackrel{(a)}{=} \mathbb{E}_{I_m, I_f, r_m, r_f}\left\{e^{-\frac{\mu\theta(I_m + I_f + \sigma^2)}{P_m r_m^{-\alpha} + P_f r_f^{-\alpha}}}\right\} \\ &= \mathbb{E}_{r_m, r_f}\left\{e^{-\frac{\mu\theta\sigma^2}{P_m r_m^{-\alpha} + P_f r_f^{-\alpha}}}\mathcal{L}_{I_m}\left(-\frac{\mu\theta}{P_m r_m^{-\alpha} + P_f r_f^{-\alpha}}\right)\right. \\ &\quad \left.\mathcal{L}_{I_f}\left(-\frac{\mu\theta}{P_m r_m^{-\alpha} + P_f r_f^{-\alpha}}\right)\right\}, \end{aligned} \quad (46)$$

where (a) follows that  $h_m$  and  $h_f$  are independent Gaussian variable  $\mathcal{N}(0, 1)$ .

The Laplace transform of  $I_m$  is

$$\begin{aligned} \mathcal{L}_{I_m}(s) &= \exp\left(-2\pi\tilde{\lambda}_{mb}\int_{r_m}^{\infty}(1 - \mathcal{L}_h(sP_m x^{-\alpha}))x dx\right) \\ &= e^{-\pi\lambda'_m r_m^2 \rho\left(\frac{sP_m}{\mu r_m^\alpha}, \alpha\right)}. \end{aligned} \quad (47)$$

The Laplace transform of  $I_f$  is

$$\begin{aligned} \mathcal{L}_{I_f}(s) &= \exp\left(-2\pi\tilde{\lambda}_{fb}\int_{r_f}^{\infty}(1 - \mathcal{L}_h(sP_f x^{-\alpha}))x dx\right) \\ &= e^{-\pi\lambda'_f r_f^2 \rho\left(\frac{sP_f}{\mu r_f^\alpha}, \alpha\right)}. \end{aligned} \quad (48)$$

## Data Rate

The instantaneous data rate that a sub-band channel of bandwidth  $W$  can accommodate is  $R = W \log_2(1 + \text{SINR})$ . For LD transmission, since both the MBS and the FAP transmit digital signals; the channel from the typical UE to its serving MBS can accommodate the data rate  $R_m = W \log_2(1 + \gamma_{LD}^m)$ , and the channel from the typical UE to its serving FAP can accommodate the data rate  $R_f = W \log_2(1 + \gamma_{LD}^f)$ . For LHDA transmission, only the BL is modulated to a digital signal, so the data rate is defined only for the BL, the channel from the typical UE to its serving MBS

can accommodate the data rate  $R_m = W \log_2(1 + \gamma_{\text{LHDA}}^m)$ , and the channel from the typical UE to its serving FAP can accommodate data rate  $R_f = W \log_2(1 + \gamma_{\text{LHDA}}^{f,B})$ .

The actually achieved UE data rates, after taking into consideration the UE load and sub-band occupancy, are given below. Without loss of generality, we take an MBS as an example. When the number of UEs in a macro cell does not exceed the total number of sub-bands (i.e.,  $U_m \leq N_m$ ), each UE can exclusively occupy a sub-band, and its achieved data rate is  $R_m$ ; when  $U_m > N_m$ , the  $U_m$  UEs share the  $N_m$  sub-bands, and the data rate is thus discounted into  $\frac{N_m}{U_m} R_m$ , assuming a round-robin sharing mechanism. So the average achieved data rate of a UE served by an MBS is given by:

$$R_{\text{mu}} = \xi_m R_m, \quad (49)$$

where  $\xi_m$  is the scheduling index denoting the probability that a UE is scheduled by the MBS,

$$\xi_m = \frac{\sum_{i=1}^{N_m} \mathbb{P}\{U_m = i\} + \sum_{i=N_m+1}^{\infty} \mathbb{P}\{U_m = i\} \frac{N_m}{i}}{1 - \mathbb{P}\{U_m = 0\}}. \quad (50)$$

Similarly, the average achieved data rate of a UE served by an FAP is given by

$$R_{\text{fu}} = \xi_f R_f, \quad (51)$$

where  $\xi_f$  is the scheduling index denoting the probability that a UE is scheduled by the FAP,

$$\xi_f = \frac{\sum_{i=1}^{N_f} \mathbb{P}\{U_f = i\} + \sum_{i=N_f+1}^{\infty} \mathbb{P}\{U_f = i\} \frac{N_f}{i}}{1 - \mathbb{P}\{U_f = 0\}}. \quad (52)$$

---

## System Performances

In this section we evaluate several important performance metrics, namely, the outage probability, the HD probability, and the average distortion. The outage probability is the probability that a UE cannot receive the BL, namely, the UE data rate is less than  $R_B$ . The HD probability is the probability that a UE can receive high-definition content, i.e., both the BL and the EL, namely, the UE data rate is greater than  $R_B + R_E$ . The average distortion evaluates the difference between the received video and source video, which is measured using the distortion-rate function. Note that, for LHDA transmission, the HD probability for the femto UE is not defined since the EL is transmitted as an analog signal and the data rate for an analog signal is undefined.



## LD Transmission

For a macro UE, only the BL is transmitted from its serving MBS, so the outage probability, denoted by  $P_{\text{out}}^{\text{LD},m}$ , is

$$\begin{aligned} P_{\text{out}}^{\text{LD},m} &= \mathbb{P}\{R_{\text{mu}} < R_B\} \\ &= \mathbb{P}\left\{\gamma_{\text{LD}}^m < 2^{\frac{R_B/\xi_m}{W}} - 1\right\} \\ &= 1 - \mathcal{P}_{\text{LD}}^m\left(2^{\frac{R_B/\xi_m}{W}} - 1\right). \end{aligned} \quad (53)$$

For a femto UE, it either connects to its serving MBS with probability  $p$  or its serving FAP with probability  $1 - p$  to receive the BL, so the outage probability, denoted by  $P_{\text{out}}^{\text{LD},f}$ , is

$$\begin{aligned} P_{\text{out}}^{\text{LD},f} &= p\mathbb{P}\{R_{\text{mu}} < R_B\} + (1 - p)\mathbb{P}\{R_{\text{fu}} < R_B\} \\ &= p\mathbb{P}\left\{\gamma_{\text{LD}}^m < 2^{\frac{R_B/\xi_m}{W}} - 1\right\} + (1 - p)\mathbb{P}\left\{\gamma_{\text{LD}}^f < 2^{\frac{R_B/\xi_f}{W}} - 1\right\} \end{aligned} \quad (54)$$

$$= p\left(1 - \mathcal{P}_{\text{LD}}^m\left(2^{\frac{R_B/\xi_m}{W}} - 1\right)\right) + (1 - p)\left(1 - \mathcal{P}_{\text{LD}}^f\left(2^{\frac{R_B/\xi_f}{W}} - 1\right)\right). \quad (55)$$

To receive the high-definition video content, a femto UE receives the BL from the MBS and receives the EL from the FAP with probability  $p$ , or it receives both the BL and the EL from the FAP with probability  $1 - p$ . Thus, the HD probability for a femto UE, denoted by  $P_{\text{HD}}^f$ , is

$$\begin{aligned} P_{\text{HD}}^f &= p\mathbb{P}\{R_{\text{mu}} > R_B, R_{\text{fu}} > R_E\} + (1 - p)\mathbb{P}\{R_{\text{fu}} > R_B + R_E\} \\ (a) &= p\mathbb{P}\{R_{\text{mu}} > R_B\}\mathbb{P}\{R_{\text{fu}} > R_E\}(1 - p)\mathbb{P}\{R_{\text{fu}} > R_B + R_E\} \\ &= p\mathbb{P}\left\{\gamma_{\text{LD}}^m > 2^{\frac{R_B/\xi_m}{W}} - 1\right\}\mathbb{P}\left\{\gamma_{\text{LD}}^f > 2^{\frac{R_E/\xi_f}{W}} - 1\right\} \\ &\quad + (1 - p)\mathbb{P}\left\{R_{\text{fu}} > 2^{\frac{(R_B + R_E)/\xi_f}{W}} - 1\right\} \\ &= p\mathcal{P}_{\text{LD}}^m\left(2^{\frac{R_B/\xi_m}{W}} - 1\right)\mathcal{P}_{\text{LD}}^f\left(2^{\frac{R_E/\xi_f}{W}} - 1\right) + (1 - p)\mathcal{P}_{\text{LD}}^f\left(2^{\frac{(R_B + R_E)/\xi_f}{W}} - 1\right), \end{aligned} \quad (56)$$

where (a) follows from the tier independence approximation.

The distortion-rate function  $D(R)$  [14, 39] is used to measure the distortion per source sample when the source rate is  $R$  bits/sample. As the bandwidth of a sub-band is  $W$  and the data rate of the BL (resp. the EL) is  $R_B$  (resp.  $R_E$ ), the source rate is  $\frac{R_B}{W}$  (resp.  $\frac{R_E}{W}$ ). Since the source signal is modeled as a Gaussian signal with zero mean and unit variance, the distortion of the received video signal can be divided into three cases based on the reception. If the BL is not decoded correctly,

the distortion is  $D_0 = 1$ ; if the BL is decoded correctly while the EL is not, then the distortion is  $D_B = 2^{-2\frac{R_B}{W}}$ ; if both the BL and the EL are decoded correctly, the distortion is  $D_{HD} = 2^{-2\frac{R_B+R_E}{W}}$ .

The average distortion for femto UEs, denoted by  $D_{LD}$ , is given by

$$D_{LD} = P_{out}^{LD,f} D_0 + (1 - P_{out}^{LD,f} - P_{HD}^f) D_B + P_{HD}^f D_{HD}. \tag{57}$$

### LHDA Transmission

For a macro UE, only the BL is digitally transmitted from its serving MBS; just the same as that in LD transmission, the outage probability  $P_{out}^{LHDA,m}$  is

$$P_{out}^{LHDA,m} = \mathbb{P}\{R_{mu} < R_B\} = P_{out}^{LD,m}. \tag{58}$$

For a femto UE, since it receives the BL from the MBS with probability  $p$  or receives the BL from the FAP with probability  $1 - p$ , the outage probability, denoted by  $P_{out}^{LHDA,f}$ , is

$$\begin{aligned} P_{out}^{LHDA,f} &= p\mathbb{P}\{R_{mu} < R_B\} + (1 - p)\mathbb{P}\{R_{fu} < R_B\} \\ &= p\mathbb{P}\left\{\gamma_{LHDA}^m < 2^{\frac{R_B/\xi_m}{W}} - 1\right\} + (1 - p)\mathbb{P}\left\{\gamma_{LHDA}^{f,B} < 2^{\frac{R_B/\xi_f}{W}} - 1\right\} \\ &= p\left(1 - \mathcal{P}_{LHDA}^m\left(2^{\frac{R_B/\xi_m}{W}} - 1\right)\right) + (1 - p)\left(1 - \mathcal{P}_{LHDA}\left(2^{\frac{R_B/\xi_f}{W}} - 1, 0\right)\right). \end{aligned} \tag{59}$$

The femto UE has two choices to receive the video content, and the average distortion is calculated accordingly.

1. *Case 1:* The femto UE receives the BL from MBS and receives the EL from FAP. Since the EL signal is analog, an MMSE estimator is employed for the estimation of the EL, and thus we have  $MMSE = \frac{1}{1 + \gamma_{LHDA}^f}$ , where  $\gamma_{LHDA}^f$  is the received SINR. Since there are multiple femto UEs in a FAP, a round-robin mechanism is used to schedule time slots for each femto UE to transmit the EL. If a UE is scheduled, its distortion for the EL is MMSE; otherwise, its distortion is unity. So the distortion is  $e_{LHDA} = \xi_f \cdot \frac{1}{1 + \gamma_{LHDA}^f} + (1 - \xi_f) \cdot 1$ . Since the EL is estimated only if the BL is decoded successfully, the cdf of  $e_{LHDA}$  conditioned on the successful reception of the BL is given by

$$\begin{aligned} \mathbb{P}\{e_{LHDA} < T \mid R_{mu} \geq R_B\}(a) &= \mathbb{P}\{e_{LHDA} < T\} \\ &= \mathbb{P}\left\{\xi_f \frac{1}{1 + \gamma_{LHDA}^f} + (1 - \xi_f)1 < T\right\} \end{aligned}$$

$$\begin{aligned}
&= \mathbb{P} \left\{ \gamma_{\text{LHDA}}^f > \frac{1-T}{T-1+\xi_f} \right\} \\
&= \mathcal{P}_{\text{LHDA}}^f \left( \frac{1-T}{T-1+\xi} \right), \tag{60}
\end{aligned}$$

where (a) follows from the tier independence approximation.

Since for a positive random variable  $X$ ,  $\mathbb{E}\{X\} = \int_{t>0} \mathbb{P}\{X > t\} dt$ , the mean distortion for the EL, denoted by  $D_E$ , is

$$\begin{aligned}
D_E &= \mathbb{E}\{e_{\text{LHDA}} \mid R_{\text{mu}} \geq R_B\} \\
&= 1 - \xi_f + \int_{1-\xi_f}^1 \left( 1 - \mathcal{P}_{\text{LHDA}}^f \left( \frac{1-T}{T-1+\xi_f} \right) \right) dT. \tag{61}
\end{aligned}$$

Since the EL corresponds to the residual between the BL and the source signal, the distortion when both the BL and the EL are received, denoted by  $D_{\text{HD}}$ , is given by  $D_{\text{HD}} = D_B D_E$ .

So the average distortion for the femto UE in *Case 1*, denoted by  $D_{\text{LHDA}}^{(1)}$ , is

$$\begin{aligned}
D_{\text{LHDA}}^{(1)} &= \mathbb{P}\{R_{\text{mu}} < R_B\} D_0 + \mathbb{P}\{R_{\text{mu}} \geq R_B\} D_{\text{HD}} \\
&= 1 - \mathcal{P}_{\text{LHDA}}^m \left( 2^{\frac{R_B}{\xi_m W}} - 1 \right) + \mathcal{P}_{\text{LHDA}}^m \left( 2^{\frac{R_B}{\xi_m W}} - 1 \right) 2^{-2R_B} \\
&\quad \times \left( 1 - \xi_f + \int_{1-\xi_f}^1 \left( 1 - \mathcal{P}_{\text{LHDA}}^f \left( \frac{1-T}{T-1+\xi_f} \right) \right) dT \right). \tag{62}
\end{aligned}$$

2. *Case 2*: The femto UE receives both the BL and the EL from the FAP. Since the EL signal is analog and superposed with the digital BL signal, an MMSE estimator is employed for the estimation of the EL conditioned on the correct reception of the BL; thus we have  $\text{MMSE} = \frac{1}{1+\gamma_{\text{LHDA}}^{\text{f,E}}}$ , where  $\gamma_{\text{LHDA}}^{\text{f,E}}$  is the received SINR after the cancellation of the BL. The distortion for the EL is  $e_{\text{LHDA}} = \xi_f \frac{1}{1+\gamma_{\text{LHDA}}^{\text{f,E}}} + (1-\xi_f)1$ . The cdf of  $e_{\text{LHDA}}$  conditioned on the successful reception of the BL is given by

$$\begin{aligned}
\mathbb{P}\{e_{\text{LHDA}} < T \mid R_{\text{fu}} \geq R_B\} &= \mathbb{P} \left\{ \xi_f \frac{1}{1+\gamma_{\text{hc}}^{\text{f,E}}} + (1-\xi_f)1 < T \mid R_{\text{fu}} \geq R_B \right\} \\
&= \mathbb{P} \left\{ \gamma_{\text{LHDA}}^{\text{f,E}} > \frac{1-T}{T-1+\xi_f} \mid \gamma_{\text{LHDA}}^{\text{f,B}} > 2^{\frac{R_B/\xi_f}{W}} - 1 \right\} \\
&= \frac{\mathcal{P}_{\text{LHDA}} \left( 2^{\frac{R_B/\xi_f}{W}} - 1, \frac{1-T}{T-1+\xi_f} \right)}{\mathcal{P}_{\text{LHDA}} \left( 2^{\frac{R_B/\xi_f}{W}} - 1, 0 \right)}. \tag{63}
\end{aligned}$$

Then, we can obtain the distortion of the EL as

$$\begin{aligned}
 D_E &= \mathbb{E}\{e_{\text{LHDA}} < T \mid R_{\text{fu}} \geq R_B\} \\
 &= 1 - \xi_f + \int_{1-\xi_f}^1 \left( 1 - \frac{\mathcal{P}_{\text{LHDA}}\left(2^{\frac{R_B}{W\xi_f}} - 1, \frac{1-T}{T-1+\xi_f}\right)}{\mathcal{P}_{\text{LHDA}}\left(2^{\frac{R_B}{W\xi_f}} - 1, 0\right)} \right) dT. \quad (64)
 \end{aligned}$$

So the average distortion for the femto UE in *Case 2*, denoted by  $D_{\text{LHDA}}^{(2)}$ , is

$$\begin{aligned}
 D_{\text{LHDA}}^{(2)} &= \mathbb{P}\{R_{\text{fu}} < R_B\}D_0 + \mathbb{P}\{R_{\text{fu}} \geq R_B\}D_{\text{HD}} \\
 &= 1 - \mathcal{P}_{\text{LHDA}}\left(2^{\frac{R_B/\xi_f}{W}} - 1, 0\right) + \mathcal{P}_{\text{LHDA}}\left(2^{\frac{R_B/\xi_f}{W}} - 1, 0\right)2^{-2R_B} \left( 1 - \xi_f \right. \\
 &\quad \left. + \int_{1-\xi_f}^1 \left( 1 - \frac{\mathcal{P}_{\text{LHDA}}\left(2^{\frac{R_B/\xi_f}{W}} - 1, \frac{1-T}{T-1+\xi_f}\right)}{\mathcal{P}_{\text{LHDA}}\left(2^{\frac{R_B/\xi_f}{W}} - 1, 0\right)} \right) dT \right). \quad (65)
 \end{aligned}$$

Since a femto UE follows *Case 1* with probability  $p$  and follows *Case 2* with probability  $1 - p$ , the average distortion for a femto UE, denoted by  $D_{\text{LHDA}}$ , is

$$D_{\text{LHDA}} = pD_{\text{LHDA}}^{(1)} + (1 - p)D_{\text{LHDA}}^{(2)}. \quad (66)$$

## CD Transmission

For a macro UE, only the BL is transmitted from MBS, so the outage probability, denoted by  $P_{\text{out}}^{\text{CD,m}}$ , is

$$\begin{aligned}
 P_{\text{out}}^{\text{CD,m}} &= \mathbb{P}\{R_{\text{mu}} < R_B\} \\
 &= 1 - \mathcal{P}_{\text{CD}}^{\text{m}}\left(2^{\frac{R_B/\xi_m}{W}} - 1\right). \quad (67)
 \end{aligned}$$

For a femto UE, if it is chosen by the MBS and the FAP, it works in a cooperative mode with the probability  $\eta = \xi_m/\xi_f$ . Thus, the outage probability, denoted by  $P_{\text{out}}^{\text{CD,f}}$  is

$$\begin{aligned}
 P_{\text{out}}^{\text{CD,f}} &= \mathbb{P}\{R_{\text{fu}} < R_B\} \\
 &= \eta \left( 1 - \mathcal{P}_{\text{CD}}^{\text{f,CoMP}}\left(2^{\frac{R_B/\xi_f}{W}} - 1\right) \right) + (1 - \eta) \left( 1 - \mathcal{P}_{\text{CD}}^{\text{f,non}}\left(2^{\frac{R_B/\xi_f}{W}} - 1\right) \right). \quad (68)
 \end{aligned}$$

For a femto UE, the HD probability, denoted as  $P_{\text{HD}}^{\text{CD},f}$ , is

$$\begin{aligned} P_{\text{HD}}^{\text{CD},f} &= \mathbb{P}\{R_{\text{fu}} > R_{\text{B}} + R_{\text{E}}\} \\ &= \eta \mathcal{P}_{\text{CD}}^{\text{f,CoMP}} \left( 2^{\frac{(R_{\text{B}}+R_{\text{E}})/\xi_{\text{f}}}{W}} - 1 \right) + (1 - \eta) \mathcal{P}_{\text{CD}}^{\text{f,non}} \left( 2^{\frac{(R_{\text{B}}+R_{\text{E}})/\xi_{\text{f}}}{W}} - 1 \right). \end{aligned} \quad (69)$$

## Simulation and Discussion

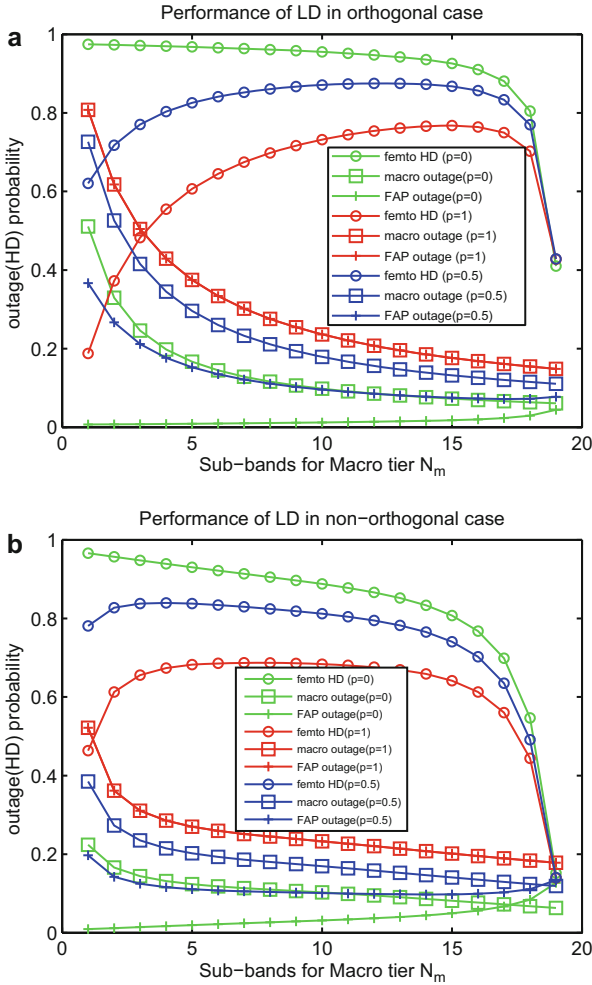
In this section, the outage probabilities, the HD probabilities, and the average distortions are evaluated for the three transmission schemes. Meanwhile, the optimal power allocation for the digital BL and the analog EL for LHDA transmission is assessed. Unless otherwise specified, the system parameters are listed in Table 1.

Figure 4a displays the performance of LD transmission in the orthogonal case. In that case,  $N_{\text{m}}$  sub-bands for the macro tier and  $N_{\text{f}}$  sub-bands for the femto tier that are orthogonal with  $N_{\text{m}} + N_{\text{f}} = N$ . As  $N_{\text{m}}$  increases, more resources are allocated to the macro tier, and the outage probabilities decrease for both macro UEs and femto UEs, except that the femto UE outage probabilities slightly increase for very large values of  $N_{\text{m}}$ . The HD probability of the femto UE with  $p = 0$  decreases with  $N_{\text{m}}$  because the EL transmission via FAPs deteriorates as the resources for the femto tier are reduced. The HD probabilities of the femto UE for  $p = 0.5$  and  $p = 1$  increase for small  $N_{\text{m}}$  and then decrease as  $N_{\text{m}}$  grows large, reflecting the tension between the resources for the BL transmission and the EL transmission.

Figure 4b displays the performance of LD transmission in the non-orthogonal case. For comparison with Fig. 4a, we still let  $N_{\text{m}} + N_{\text{f}} = N$  but let the sub-bands be selected by each BS independently. The general trend is similar to that in the orthogonal case, but the difference lies in that the curves show less variability with

**Table 1** System parameters

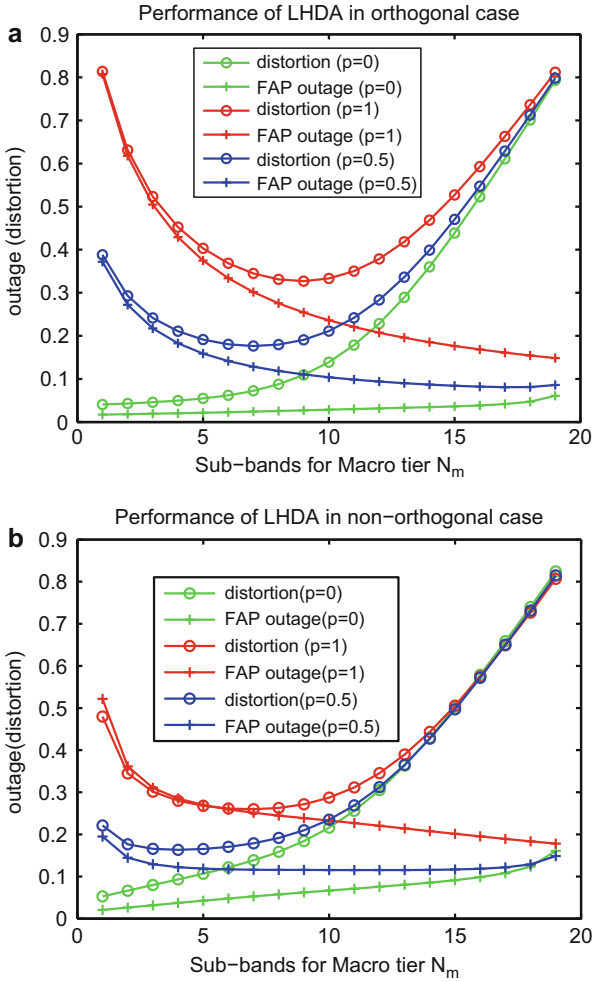
Symbol	Description	Typical value
$N$	Number of sub-bands	20
$W$	Bandwidth of a sub-band (MHz)	5
$P_{\text{m}}$	MBS transmit power per sub-band (dBm)	39
$P_{\text{f}}$	FAP transmit power per sub-band (dBm)	13
$\sigma^2$	Noise power (dBm)	-104
$\lambda_{\text{mb}}$	MBS intensity ( $\text{m}^{-2}$ )	1E-5
$\lambda_{\text{fb}}$	FAP intensity ( $\text{m}^{-2}$ )	5E-5
$\lambda_{\text{mu}}$	Macro UE intensity ( $\text{m}^{-2}$ )	2E-4
$\lambda_{\text{fu}}$	Femto UE intensity in coverage ( $\text{m}^{-2}$ )	8E-3
$R_{\text{f}}$	Coverage radius of FAP (m)	20
$\alpha$	Path loss exponent	4
$R_{\text{B}}$	Rate for the BL transmission (Mbps)	0.5
$R_{\text{E}}$	Rate for the EL transmission (Mbps)	4.5



**Fig. 4** Performances of LD in both orthogonal and non-orthogonal cases

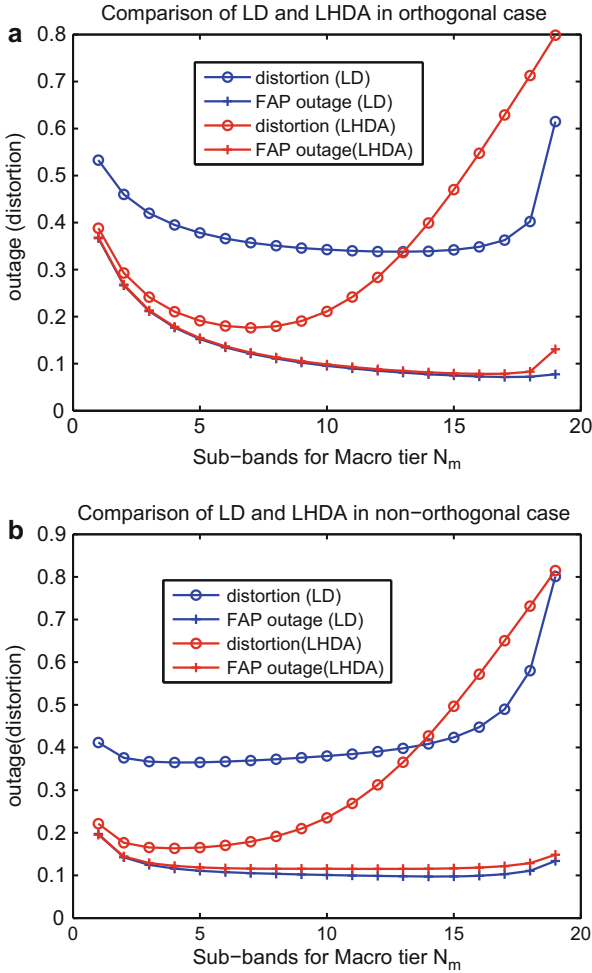
$N_m$  (except for those values near to  $N$ ). The reason for such a practically desirable insensitivity is due to the lessened tension between the resources for macro tiers and femto tiers from randomized sub-band selection.

Note that if  $p$  is large, the femto UE tends to connect to an MBS to receive the BL; the outage probability increases, and the HD probability decreases, i.e., the performance deteriorates. However, since an MBS can provide continuous coverage while an FAP cannot, if a femto UE is moving, then it may prefer to connect to an MBS to receive the BL, which prevents frequent handover between femto cells and enables uninterrupted reception of the BL video.



**Fig. 5** Performances of LHDA in both orthogonal and non-orthogonal cases

Figure 5a displays the performance of LHDA transmission in the orthogonal case. The outage probability for macro UE is the same as that in LD transmission, so we just neglect it in LHDA transmission. Since the frequency resource allocated to the macro tier increases, the resource for the femto tier decreases. The outage probability for the femto UE connected to the FAP (corresponding to  $p = 0$ ) to receive the BL increases, while the outage probability for the femto UE connected to the MBS (corresponding to  $p = 1$ ) to receive the BL decreases. The case where  $p = 0.5$  shows a tradeoff of these two extreme cases: the outage probability for femto UE first decreases and then slightly increases when the allocated resource for the FAP is small. When  $N_m$  is small, the performance of the macro tier is poor, and



**Fig. 6** Comparisons between LD and LHDA in both orthogonal and non-orthogonal cases

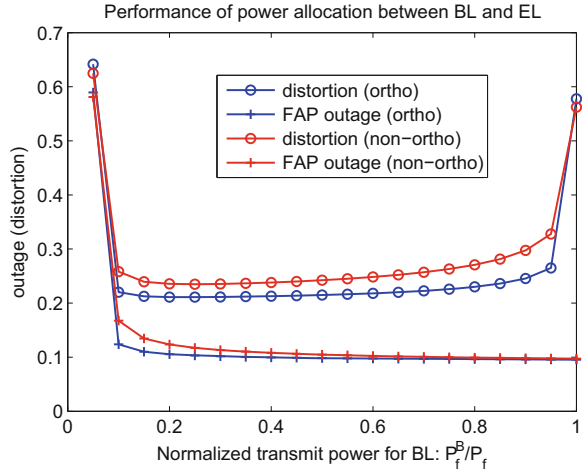
thus the distortion for the UE connected to the MBS to receive the BL is large. When increasing  $N_m$ , the performance of the macro tier becomes good, while that of the femto tier is poor.

Figure 5b displays the performance of LHDA transmission in the non-orthogonal case. The general trends of the curves of the outage and the average distortion are almost the same as that of Fig. 5a. The difference lies in that the outage probability is lower in the non-orthogonal case than that in the orthogonal case when  $N_m$  is small.

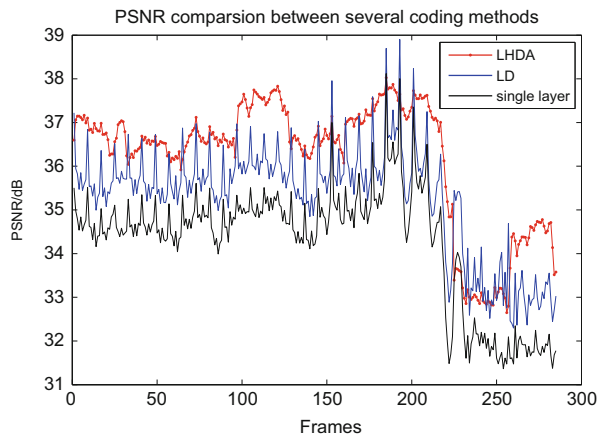
Figure 6 displays the comparison between LD transmission and LHDA transmission. Since the comparisons for different  $p$  are more or less the same, we set  $p = 0.5$



**Fig. 7** Power allocation between the BL and the EL in FAPs for LHDA transmission

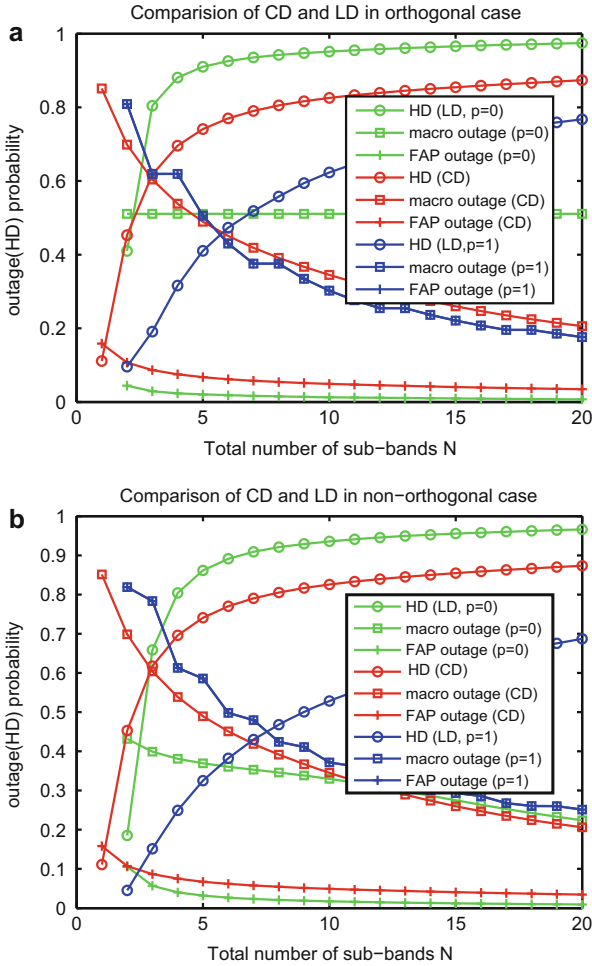


**Fig. 8** Comparisons of LHDA and LD and single-layer coding video transmission



as an example. In both orthogonal and non-orthogonal cases, LHDA outperforms LD when the proportion of frequency resource allocated to the femto tier exceeds a certain threshold, for example, 35% (i.e.,  $N_f \geq 7$ ) in the current deployment, as the outage probability is slightly increasing while the average distortion is obviously decreasing when  $N_m$  is small. The reason is that analog transmission avoids the cliff effect and offers the continuous quality scalability.

Figure 7 displays the power allocation between the digital BL and the analog EL for LHDA transmission. If the power allocated to the BL is increasing, the outage probability decreases monotonously and then approaches stable as the network is interference limited. With  $P_f^B$  increasing, the distortion for the BL is sharply decreasing, while the distortion for the EL is increasing. Thus, the total distortion firstly decreases owing to superior transmission of the BL and then increases owing to inferior transmission of the EL. Because of the tradeoff between the transmissions



**Fig. 9** Comparisons of CD and LD in both orthogonal and non-orthogonal cases

of the BL and the EL, the average distortion varies little when the power allocation ratio  $P_f^B / P_f$  lies in a wide range; thus the power allocation is robust.

QoE reflects the user experience and satisfaction, and its evaluation can be performed via subjective tests with the help of a panel of users. The mean opinion score (MOS) which reflects the quality perceived by the observers is dependent on human observation as well as time demanding and costly, while objective video metrics are mathematical-based metrics which are easy and fair to evaluate. The peak signal-to-noise ratio (PSNR) is usually chosen as the objective video metric.

Figure 8 displays the comparisons of performance of LHDA and LD and single-layer coding video transmission in a practical scenario. A standard video called Foreman.yuv is chosen, then it is encoded through H.264/AVC to a single layer or is

encoded through H.264/AVC to two layers to exploit the spatial scalability. LHDA and LD are employed to transmit the layered video contents. The performance peak signal-to-noise ratio (PSNR) of all the 288 frames is compared for the three transmission schemes. The performance of LHDA is the best, since it achieves the highest PSNR, and the variation of PSNR with respect to different frame is stable. The performance of LD is better than single-layer coding video transmission.

Figure 9 displays the comparison between CD transmission and LD transmission with  $p = 0$  and  $p = 1$  under orthogonal and non-orthogonal spectrum allocation methods. Since different  $N_m$  result in different performance, here the optimal  $N_m$  is chosen to achieve the minimal outage probability given the total number of sub-bands is  $N$ . It can be found that the HD probability of CD is between that of LD with  $p = 0$  and  $p = 1$ . Meanwhile the outage probabilities of macro UE and femto UE are nearly minimal.

---

## Conclusion and Future Directions

In order to overcome the challenges of the mobile networks induced by the rapidly growing video traffic, efficient networking strategy as well as providing optimal QoE is an urgent task for mobile operators. In this work, we proposed an analytical framework for user-cognizant scalable video transmission, which exploits the interplay among user demand, video source characteristic, and networking.

Specifically, three scalable transmission schemes are presented, i.e., LD, LHDA, and CD, which are shown to be an effective means for providing differentiated services for users. Through the analysis and comparison of system performance metrics, i.e., outage probability, HD probability, and average distortion, under different spectrum allocation methods, it is observed that:

- (1) Compared to the traditional non-scalable video transmission, our schemes can adaptively provide basic or high-definition video.
- (2) The frequency resource should be elaborately allocated between tiers to achieve good performance, and the choice of orthogonal and non-orthogonal spectrum allocation methods for LD and LHDA depend on the system configuration.
- (3) The hybrid digital-analog transmission can further improve the system performance by reducing video distortion and providing continuous quality scalability of high-definition video, and the performance is quite insensitive to the power allocation between the digital BL and the analog EL.
- (4) Cooperative video transmission can achieve relatively high HD probability with a little increase of outage probability.

To further enhance the networking performance and user QoE, the work can be extended to incorporate wireless caching and rateless codes. The heterogeneity of video quality and video popularity can be exploited to optimize the caching and transmission schemes [40]. Since the UE may simultaneously connect to the MBS and the FAP, rateless codes such as fountain codes [41, 42] can be employed

to reduce the complexity of synchronization of these two links. Moreover, rate adaptation and shaping strategies can also be studied with some background traffic, such as file downloading and over-the-top (OTT) services.

---

## References

1. Schwarz H, Marpe D, Wiegand T (2007) Overview of the scalable video coding extension of the H. 264/AVC standard. *IEEE Trans Circuits Syst Video Technol* 17(9):1103–1120
2. Sullivan GJ, Boyce JM, Chen Y, Ohm JR (2013) Standardized extensions of high efficiency video coding (HEVC). *IEEE J Sel Top Signal Process* 7(6):1001–1016
3. Chandrasekhar V, Andrews JG, Gatherer A (2008) Femtocell networks: a survey. *IEEE Commun Mag* 46(9):59–67
4. Ko C-H, Wei H-Y (2011) On-demand resource-sharing mechanism design in two-tier OFDMA femtocell networks. *IEEE Trans Veh Technol* 60(3):1059–1071
5. Schierl T, Stockhammer T, Wiegand T (2007) Mobile video transmission using scalable video coding. *IEEE Trans Circuits Syst Video Technol* 17(9):1204–1217
6. Bocus MZ, Coon JP, Canagarajah CN, Armour S, Doufexi A, McGeehan JP (2012) Per-subcarrier antenna selection for H. 264 MGS/CGS video transmission over cognitive radio networks. *IEEE Trans Veh Technol* 61(3):1060–1073
7. Radhakrishnan R, Nayak A (2012) Cross layer design for efficient video streaming over LTE using scalable video coding. In: *Proceedings of IEEE International Conference on Communications*, pp 6509–6513
8. Gupta V, Somayazulu S, Himayat N, Verma H, Bisht M, Nandwani V (2012) Design challenges in transmitting scalable video over multi-radio networks. In: *Proceedings of IEEE Globecom Workshops*, pp 46–51
9. Poularakis K, Iosifidis G, Argyriou A, Tassioulas L (2014) Video delivery over heterogeneous cellular networks: optimizing cost and performance. In: *Proceedings of IEEE INFOCOM*, pp 1078–1086
10. Chen C, Zhu X, De Veciana G, Bovik AC, Heath RW (2013) Rate adaptation and admission control for video transmission with subjective quality constraints. *IEEE J Sel Top Signal Process* 9(1):22–36
11. Thakolsri S, Khan S, Steinbach E, Kellerer W (2009) QoE-driven cross-layer optimization for high speed downlink packet access. *J Commun* 4(9):669–680
12. Fu B, Staehle D, Kunzmann G, Steinbach E, Kellerer W (2015) QoE-based SVC layer dropping in LTE networks using content-aware layer priorities. *ACM Trans Multimedia Comput Commun Appl* 12(1):1–23
13. Jakubczak S, Katabi D (2011) A cross-layer design for scalable mobile video. In: *Proceedings of ACM Proceedings of Annual International Conference on Mobile Computing and Networking*, pp 289–300
14. Gao Y, Tuncel E (2010) New hybrid digital/analog schemes for transmission of a Gaussian source over a Gaussian channel. *IEEE Trans Inf Theory* 56(12):6014–6019
15. Minero P, Lim SH, Kim Y-H (2015) A unified approach to hybrid coding. *IEEE Trans Inf Theory* 61(4):1509–1523
16. Yu L, Li H, Li W (2014) Wireless scalable video coding using a hybrid digital-analog scheme. *IEEE Trans Circuits Syst Video Technol* 24(2):331–345
17. Altieri A, Vega LR, Galarza CG, Piantanida P (2011) Cooperative strategies for interference-limited wireless networks. In: *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp 1623–1627
18. Lee J, Kim Y, Lee H, Ng BL, Mazzarese D, Liu J, Xiao W, Zhou Y (2012) Coordinated multi-point transmission and reception in lte-advanced systems. *IEEE Commun Mag* 50(11):44–50

19. Nigam G, Minero P, Haenggi M (2013) Coordinated multipoint in heterogeneous networks: a stochastic geometry approach. In: IEEE Globecom Workshops, pp 145–150
20. Nigam G, Minero P, Haenggi M (2014) Coordinated multipoint joint transmission in heterogeneous networks. *IEEE Trans Commun* 62(11):4134–4146
21. Bang I, Kim SH, Kim SM, Sung DK (2012) Energy-efficient subchannel allocation scheme based on adaptive base station cooperation in downlink cellular networks. In: Wireless Communications and Networking Conference (WCNC), pp 2434–2439
22. Zhang X, Shen XS, Xie L-L (2014) Joint subcarrier and power allocation for cooperative communications in LTE-advanced networks. *IEEE Trans Wirel Commun* 13(2):658–668
23. Kosmanos D, Argyriou A, Liu Y, Tassioulas L, Ci S (2015) A cooperative protocol for video streaming in dense small cell wireless relay networks. *Signal Process Image Commun* 31: 151–160
24. Chan CC, Hanly SV (2001) Calculating the outage probability in a CDMA network with spatial Poisson traffic. *IEEE Trans Veh Technol* 50(1):183–204
25. Haenggi M, Andrews JG, Baccelli F, Dousse O, Franceschetti M (2009) Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE J Sel Areas Commun* 27(7):1029–1046
26. ElSawy H, Hossain E, Haenggi M (2013) Stochastic geometry for modeling, analysis, and design of multi-tier and cognitive cellular wireless networks: a survey. *IEEE Commun Surv Tutorials* 15(3):996–1019
27. Andrews JG, Baccelli F, Ganti RK (2011) A tractable approach to coverage and rate in cellular networks. *IEEE Trans Commun* 59(11):3122–3134
28. Dhillon HS, Ganti RK, Baccelli F, Andrews JG (2012) Modeling and analysis of K-tier downlink heterogeneous cellular networks. *IEEE J Sel Areas Commun* 30(3):550–560
29. Baccelli F, Blaszczyzyn B (2009) Stochastic geometry and wireless networks: volume 1 – theory, vol 1. Now Publishers Inc, Boston
30. Haenggi M (2012) Stochastic geometry for wireless networks. Cambridge University Press, Cambridge
31. Wu L, Zhong Y, Zhang W, Haenggi M (2015) Scalable transmission over heterogeneous networks. In: Proceedings of International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp 459–466
32. Wu L, Zhong Y, Zhang W, Haenggi M (2016) Scalable transmission over heterogeneous network: a stochastic geometry analysis. *IEEE Trans Veh Technol* 66(2):1845–1859
33. Ferenc J-S, Néda Z (2007) On the size distribution of Poisson Voronoi cells. *Phys A Stat Mech Appl* 385(2):518–526
34. Zhong Y, Zhang W (2013) Multi-channel hybrid access femtocells: a stochastic geometric analysis. *IEEE Trans Commun* 61(7):3016–3026
35. Cheung WC, Quek TQ, Kountouris M (2012) Throughput optimization, spectrum allocation, and access control in two-tier femtocell networks. *IEEE J Sel Areas Commun* 30(3):561–574
36. Prabhakaran VM, Puri R, Ramchandran K (2011) Hybrid digital-analog codes for source-channel broadcast of Gaussian sources over Gaussian channels. *IEEE Trans Inf Theory* 57(7):4573–4588
37. Kochman Y, Zamir R (2011) Analog matching of colored sources to colored channels. *IEEE Trans Inf Theory* 57(6):3180–3195
38. Wildemeersch M, Quek TQ, Kountouris M, Rabbachin A, Slump CH (2014) Successive interference cancellation in heterogeneous networks. *IEEE Trans Commun* 62(12):4440–4453
39. Xu X, Gunduz D, Erkip E, Wang Y (2005) Layered cooperative source and channel coding. In: Proceedings of IEEE International Conference on Communications, pp 1200–1204
40. Wu L, Zhang W (2016) Caching-based scalable video transmission over cellular networks. *IEEE Commun Lett* 20(6):1156–1159
41. Luby M (2002) LT codes. In: Symposium on Foundations of Computer Science, pp 271–280
42. Shokrollahi A (2006) Raptor codes. *IEEE/ACM Trans Netw* 14(SI):2551–2567