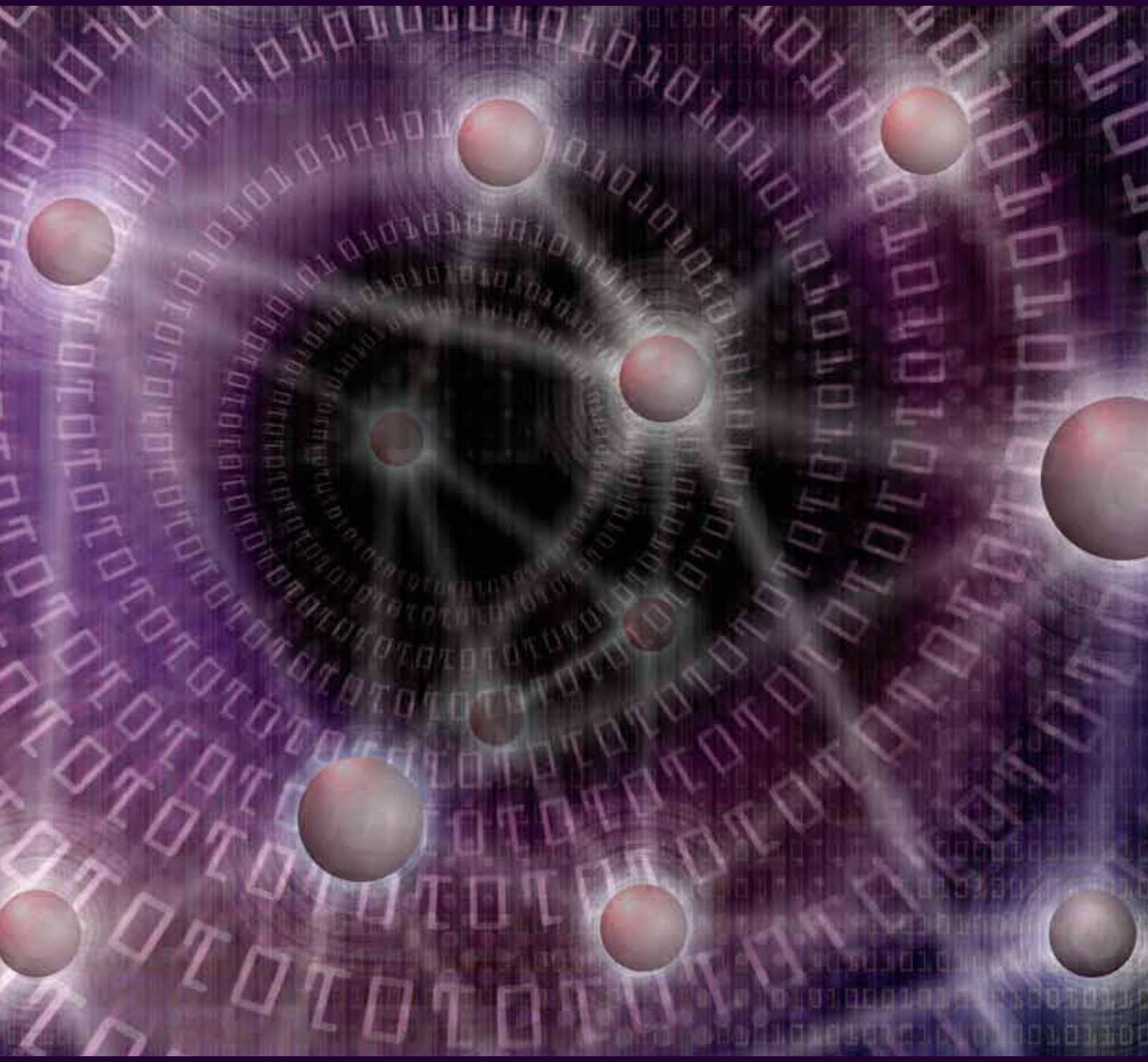
EURASIP Journal on Wireless Communications and Networking

3GPP LTE and LTE Advanced

Guest Editors: Bruno Clerckx, Angel Lozano, Stefania Sesia, Cornelius van Rensburg, and Constantinos B. Papadias



3GPP LTE and LTE Advanced

EURASIP Journal on
Wireless Communications and Networking

3GPP LTE and LTE Advanced

Guest Editors: Bruno Clerckx, Angel Lozano, Stefania Sesia, Cornelius van Rensburg, and Constantinos B. Papadias

Contents

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 472124, 3 pages
doi:10.1155/2009/472124

Editorial

3GPP LTE and LTE-Advanced

Bruno Clerckx,[1] Angel Lozano,[2] Stefania Sesia,[3] Cornelius van Rensburg,[4] and Constantinos B. Papadias[5]

[1] *Communication Laboratory, Samsung Advanced Institute of Technology, Samsung Electronics, Yongin-Si 446-712, South Korea*
[2] *Department of Information and Communication Technologies, Universitat Pompeu Fabra, 08002 Barcelona, Spain*
[3] *ST-NXPWireless - SW/FW/Algorithm Group, European Telecommunications Standard Institute (ETSI), Mobile Competence Center, 06921 Sophia Antipolis Cedex, France*
[4] *Wireless R&D Laboratory, Huawei Technologies, Plano, TX 75075, USA*
[5] *Broadband Wireless and Sensor Networks, Athens Information Technology, P.O. Box 68, Markopoulo Avenue, Peania 19002, Athens, Greece*

Correspondence should be addressed to Bruno Clerckx, bruno.clerckx@gmail.com

Received 17 September 2009; Accepted 17 September 2009

A new cellular standard, termed Long-Term Evolution (LTE), also referred to as Evolved UMTS Terrestrial Radio Access (E-UTRA), is currently being defined to replace the UMTS third-generation system. LTE-Advanced, in turn, refers to the most advanced version of LTE that was initiated about one year ago.

LTE represents a radical new step forward for the wireless industry, targeting order-of-magnitude increases in bit rates with respect to its predecessors by means of wider bandwidths and improved spectral efficiency. Beyond the improvement in bit rates, LTE aims to provide a highly efficient, low-latency, packet-optimized radio access technology offering enhanced spectrum flexibility.

The LTE design presents radical differences at every layer. Like many other communication technologies (e.g., digital video and audio broadcasting, DSL, wireless LANs), the physical layer uses OFDM waveforms in order to avoid the intersymbol interference that typically arises in high bandwidth systems. In terms of radio access, CDMA has given away to time and frequency multiple access. One differentiating aspect of the LTE standard is that from the onset, MIMO is an integral component, and not an add-on feature. At the network layer, a flatter architecture is being defined that represents the transition from the existing UTRA network, which combines circuit- and packet-switching, to an all-IP system.

The objective of this special issue is to disseminate new advances in both the physical and medium access control layers that are applicable to the LTE and LTE-Advanced technologies. Out of 23 submissions, we selected 13 papers for inclusion in this special issue. Papers have been classified in 6 categories. The first category (papers 1 and 2) consists of two tutorials and provides an introduction to the LTE system. The second category (papers 3 to 5) focuses on MIMO issues including Single-User MIMO (SU-MIMO) and Multi-User MIMO (MU-MIMO). The third category (paper 6) addresses the design of a practical scheduler for LTE. The fourth category (papers 7 to 10) focuses on multiple access (uplink) and specifically on Single-Carrier FDMA (SC-FDMA). The fifth category (papers 11 and 12) opens the way to LTE-Advanced and on relay architectures. The last category investigates the applicability of the 3GPP LTE interface to satellite transmission.

We hope that this excellent collection of papers will help the interested readers to identify a number of key challenges and opportunities that lie within the LTE and LTE-Advanced cellular standards, contributing not only to a better understanding of these systems but eventually also to the incorporation of techniques that will further boost the performance of the corresponding deployed networks.

A more detailed description of each category and the corresponding papers is as follows.

Two tutorial papers introduce the LTE system. The paper by Martin-Sacristan et al. provides a broad introduction to LTE, describing the core functionalities and the system requirements. It briefly introduces LTE-Advanced requirements and technologies currently under study. Lee et al. provide in their invited paper a tutorial on MIMO technologies. Being active players in the standardization of LTE and LTE-Advanced, they go beyond a simple description of MIMO and provide insightful explanations as to why these features have been adopted.

Three papers address MIMO issues. The first two papers focus on MU-MIMO while the last one focuses on SU-MIMO.

The paper by Shu et al. proposes an optimal MU-MIMO linear precoding scheme with linear (MMSE) detection based on particle swarm optimization. It shows that such a scheme significantly outperforms other well-known approaches (e.g., precoding schemes based on channel block diagonalization). Combined with the use of demodulation reference signals, the proposed scheme may be a good candidate for MU-MIMO in 3GPP LTE-Advanced.

The paper by Komulainen proposes some linear uplink (UL) MU-MIMO beamforming schemes for cellular TDD systems. The UL scheme is designed to coexist with downlink (DL) MU-MIMO block zero-forcing by coordinated transmit-receive processing. It relies on the channel state information of the zero-forced DL channels to precode UL transmission. The proposed scheme is shown to outperform SU-MIMO transmission combined with user selection as well as UL antenna selection. The proposed concept may be a promising candidate for 3GPP LTE-Advanced.

The paper by Yang examines the performance of various kinds of antenna selection schemes in MIMO systems with limited size antenna arrays and nonnegligible mutual coupling. It is shown that the performance of antenna selection is closely related to the antenna spacing and that appropriate selection schemes are required when mutual coupling is taken into account. Results indicate that soft selection always outperforms hard selection. This study provides novel insight into the deployment of antenna selection in both 3GPP LTE and LTE-Advanced.

One paper addresses the design of a scheduler for 3GPP LTE downlink. In particular, the paper by Sadiq et al. proposes a complete and practical scheduler with low computational complexity by integrating state-of-the-art techniques regarding resource allocation, fast algorithms, and scheduling. Simulations covering both PHY and MAC layers are performed and demonstrate the various design tradeoffs to be made in the selection of the queue-and-channel-aware QoS scheduling policies to cope with the diverse mix of traffic of an LTE downlink.

Performance of 3GPP LTE uplink multiple access based on SC-FDMA is discussed in the following four papers. Wu et al. compares the information rate achieved by SC-FDMA and OFDMA with linear frequency-domain equalizers. Based on some geometrical interpretation, it is shown that the loss incurred by SC-FDMA can be mitigated by exploiting multiuser diversity and spatial diversity. 3GPP LTE uplink performance evaluations confirm those findings.

The paper by Svensson et al. proposes a novel uplink multiple access scheme denoted as B-IFDMA (Block Interleaved Frequency Division Multiple Access). In the presence of realistic channel estimation at the receiver and imperfect channel state information at the transmitter, B-IFDMA is shown to outperform the currently adopted SC-FDMA method. Additional benefits in terms of amplifier and energy efficiency, robustness to frequency offsets, and Doppler spreads are detailed.

The paper by Assimi et al. investigates the design of Hybrid Automatic Repeat reQuest (HARQ) for single-carrier transmission as in 3GPP LTE UL. Here, transmit diversity techniques are introduced for HARQ retransmissions; this is because in slow fading environments, retransmitting identically the same packet does not provide additional diversity gain. The authors investigate the performance of both cyclic-frequency-shift-diversity and bit-interleaving-diversity using theoretical analysis and simulations. The choice of a specific diversity scheme is shown to depend on the desired performance/complexity tradeoff and the system parameters.

Karakaya et al. further address the problem of sensitivity of SC-FDMA to frequency offsets. A novel Kalman filter-based method is introduced, enabling the tracking of the channel taps in the time domain, to mitigate the intercarrier interference under high Doppler spreads. This method is evaluated under various scenarios to assess the impact of the design parameters on the performance.

The following two papers introduce new technologies based on relay for LTE-Advanced.

The paper by Teyeb et al. proposes to enhance the LTE architecture by performing dynamic relaying while maintaining backward compatibility with LTE Release 8. A flexible, efficient, and self-organizing multihop architecture is introduced where relays can be linked to base stations on an "as need" basis rather than in a fixed manner.

The paper by Peters et al. focuses on the problem of the extra interference added by the presence of relays, which is currently mostly overlooked. They also analyze the performance of various kinds of half-duplex relay strategies in interference-limited environments. The performance benefits of those strategies are discussed as a function of deployment scenarios and the system parameters.

The last paper by Bastia et al. studies the applicability of 3GPP LTE to satellite transmissions. With the introduction of several new features such as inter-TTI interleaving techniques that exploit the existing HARQ structure of LTE, Peak-to-Average Power Ratio (PAPR) reduction techniques, and preamble sequence design for random access, it is shown that the existing terrestrial air interface can be adopted for transmission over satellite links.

Acknowledgments

We would like to thank the authors for contributing to this special issue and to the reviewers for providing fast feedback

and constructive remarks that helped improve the quality of the manuscripts. We also thank the Editor-in-Chief and the editorial office for their support through the entire editing process.

Bruno Clerckx
Angel Lozano
Stefania Sesia
Cornelius van Rensburg
Constantinos B. Papadias

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 354089, 10 pages
doi:10.1155/2009/354089

Research Article

On the Way towards Fourth-Generation Mobile: 3GPP LTE and LTE-Advanced

David Martín-Sacristán, Jose F. Monserrat, Jorge Cabrejas-Peñuelas, Daniel Calabuig, Salvador Garrigas, and Narcís Cardona

iTEAM Research Institute, Universidad Politecnica de Valencia, Camino de Vera S/N, 46022 Valencia, Spain

Correspondence should be addressed to Jose F. Monserrat, jomondel@iteam.upv.es

Received 31 January 2009; Accepted 18 June 2009

Recommended by Claude Oestges

Long-Term Evolution (LTE) is the new standard recently specified by the 3GPP on the way towards fourth-generation mobile. This paper presents the main technical features of this standard as well as its performance in terms of peak bit rate and average cell throughput, among others. LTE entails a big technological improvement as compared with the previous 3G standard. However, this paper also demonstrates that LTE performance does not fulfil the technical requirements established by ITU-R to classify one radio access technology as a member of the IMT-Advanced family of standards. Thus, this paper describes the procedure followed by the 3GPP to address these challenging requirements. Through the design and optimization of new radio access techniques and a further evolution of the system, the 3GPP is laying down the foundations of the future LTE-Advanced standard, the 3GPP candidate for 4G. This paper offers a brief insight into these technological trends.

1. Introduction

In the last years, technology evolution in mobile communications is mainly motivated by three relevant agents: (1) the market globalization and liberalization and the increasing competence among vendors and operators coming from this new framework, (2) the popularization of IEEE 802 wireless technologies within the mobile communications sector and, finally, (3) the exponential increase in the demand for advanced telecommunication services.

Concerning the last item, the envisaged applications to be supported by current and future cellular systems include Voice over IP (VoIP), videoconference, push-to-talk over cellular (PoC), multimedia messaging, multiplayer games, audio and video streaming, content download of ring tones, video clips, Virtual Private Network (VPN) connections, web browsing, email access, File Transfer Protocol (FTP). All these applications can be classified in several ways based on the Quality of Service (QoS) treatment that they require. Some of them are real-time and delay-sensitive, like voice and videoconference, while some others require integrity, high data-rate, and are sensitive to latency (like VPN and FTP). The simultaneous support of applications with different QoS requirements is one of the most important challenges that cellular systems are facing. At the same time, the spectrum scarcity makes that new wideband cellular systems are designed with very high spectral efficiency.

It is precisely that this increasing market demand and its enormous economic benefits, together with the new challenges that come with the requirements in higher spectral efficiency and services aggregation, raised the need to allocate new frequency channels to mobile communications systems. That is why the ITU-R WP 8F started in October 2005 the definition of the future Fourth-Generation Mobile (4G), also known as International Mobile Telecommunications (IMTs) Advanced, following the same model of global standardization used with the Third Generation, IMT-2000. The objective of this initiative is to specify a set of requirements in terms of transmission capacity and quality of service, in such a way that if a certain technology fulfils all these requirements it is included by the ITU in the IMT-Advanced set of standards. This inclusion firstly endorses technologies and motivates operators to invest in them, but furthermore it allows these standards to make use of the frequency bands

specially designated for IMT-Advanced, what entails a great motivation for mobile operators to increase their offered services and transmission capacity.

The race towards IMT-Advanced was officially started in March 2008, when a Circular Letter was distributed asking for the submission of new technology proposals [1]. Previous to this official call, the 3rd Partnership Project (3GPP) established the Long Term Evolution (LTE) standardization activity as an ongoing task to build up a framework for the evolution of the 3GPP radio technologies, concretely UMTS, towards 4G. The 3GPP divided this work into two phases: the former concerns the completion of the first LTE standard (Release 8), whereas the latter intends to adapt LTE to the requirements of 4G through the specification of a new technology called LTE-Advanced (Release 9 and 10). Following this plan, in December 2008 3GPP approved the specifications of LTE Release 8 which encompasses the Evolved UTRAN (E-UTRAN) and the Evolved Packet Core (EPC). Otherwise, the LTE-Advanced Study Item was launched in May 2008, expecting its completion in October 2009 according to the ITU-R schedule for the IMT-Advanced process. In the meantime, research community has been called for the performance assessment of the definitive LTE Release 8 standard.

Actually, several papers deal with the performance evaluation of LTE. However, up to date this assessment has been partially done because of one of these two reasons. First, some of these works only focused on the physical layer, leaving out the retransmission processes and error correction [2–4]. System level analysis needs MAC layer performance information and cannot be carried out with only a physical layer characterization. Second, other papers assessing the performance of LTE radio access network assumed ideal channel estimation, which results in an optimistic estimation of LTE capacity [5–7].

This paper describes the main characteristics of LTE Release 8 and evaluates LTE link level performance considering a transmission chain fully compliant with LTE Release 8 and including realistic HARQ and turbo-decoding. Besides, the capacity of LTE systems is analyzed in terms of maximum achievable throughput and cell capacity distribution in a conventional scenario. These studies allow having a rough idea on the benefits and capabilities of the new standard. Finally, this paper offers an overview of the current research trends followed by 3GPP in the definition process of LTE-Advanced thus foreseeing the main characteristics of next generation mobile.

2. LTE

3GPP Long Term Evolution is the name given to the new standard developed by 3GPP to cope with the increasing throughput requirements of the market. LTE is the next step in the evolution of 2G and 3G systems and also in the provisioning of quality levels similar to those of current wired networks.

3GPP RAN working groups started LTE/EPC standardization in December 2004 with a feasibility study for an evolved UTRAN and for the all IP-based EPC. This is known as the Study Item phase. In December 2007 all LTE functional specifications were finished. Besides, EPC functional specifications reached major milestones for interworking with 3GPP and CDMA networks. In 2008 3GPP working groups were running to finish all protocol and performance specifications, being these tasks completed in December 2008 hence ending Release 8.

2.1. LTE Requirements. 3GPP collected in [8] the requirements that an evolved UTRAN should meet. Some of the requirements are defined in an absolute manner while other requirements are defined in relation to UTRA performance. It is worth to mention that for the UTRA baseline it is considered the use of Release 6 HSDPA with a $\mathbf{1} \times \mathbf{1}$ multiantenna scheme for the downlink and Release 6 HSUPA with a $\mathbf{1} \times \mathbf{2}$ multiantenna scheme in uplink. For the sake of comparison, in LTE it is considered transmission using up to $\mathbf{2} \times \mathbf{2}$ antennas in downlink and up to $\mathbf{1} \times \mathbf{2}$ antennas in uplink.

Among others, LTE design targets are the following.

(i) The system should support peak data rates of 100 Mbps in downlink and 50 Mbps in uplink within a 20 MHz bandwidth or, equivalently, spectral efficiency values of 5 bps/Hz and 2.5 bps/Hz, respectively. Baseline considers 2 antennas in UE for downlink and 1 antenna in UE for uplink.

(ii) Downlink and uplink user throughput per MHz at the 5% point of the CDF, 2 to 3 times Release 6 HSPA.

(iii) Downlink averaged user throughput per MHz, 3 to 4 times Release 6 HSDPA. Uplink averaged user throughput per MHz, 2 to 3 times Release 6 Enhanced Uplink.

(iv) Spectrum efficiency 3 to 4 times Release 6 HSDPA in downlink and 2 to 3 times Release 6 HSUPA in uplink, in a loaded network.

(v) Mobility up to 350 km/h.

(vi) Spectrum flexibility, seamless coexistence with previous technologies and reduced complexity and cost of the overall system.

2.2. LTE Release 8 Technical Overview. To meet these requirements, a combination of a new system architecture together with an enhanced radio access technology was incorporated in the specifications.

2.2.1. Architecture. There are different types of functions in a cellular network. Based on them, network can be split into two parts: a radio access network part and a core network part. Functions like modulation, header compression and handover belong to the access network, whereas other functions like charging or mobility management are part of the core network. In case of LTE, the radio access network is E-UTRAN and the core network EPC.

Radio Access Network. The radio access network of LTE is called E-UTRAN and one of its main features is that

all services, including real-time, will be supported over shared packet channels. This approach will achieve increased spectral efficiency which will turn into higher system capacity with respect to current UMTS and HSPA. An important consequence of using packet access for all services is the better integration among all multimedia services and among wireless and fixed services.

The main philosophy behind LTE is minimizing the number of nodes. Therefore the developers opted for a single-node architecture. The new base station is more complicated than the Node B in WCDMA/HSPA radio access networks, and is consequently called eNB (Enhanced Node B). The eNBs have all necessary functionalities for LTE radio access network including the functions related to radio resource management.

Core Network. The new core network is a radical evolution of the one of third generation systems and it only covers the packet-switched domain. Therefore it has a new name: Evolved Packet Core.

Following the same philosophy as for the E-UTRAN, the number of nodes is reduced. EPC divides user data flows into the control and the data planes. A specific node is defined for each plane plus the generic gateway that connects the LTE network to the internet and other systems. The EPC comprises several functional entities.

(i) The MME (Mobility Management Entity): is responsible for the control plane functions related to subscriber and session management.

(ii) The Serving Gateway: is the anchor point of the packet data interface towards E-UTRAN. Moreover, it acts as the routing node towards other 3GPP technologies.

(iii) The PDN Gateway (Packet Data Network): is the termination point for sessions towards the external packet data network. It is also the router to the Internet.

(iv) The PCRF (Policy and Charging Rules Function): controls the tariff making and the IP Multimedia Subsystem (IMS) configuration of each user.

The overall structure of LTE is shown in Figure 1.

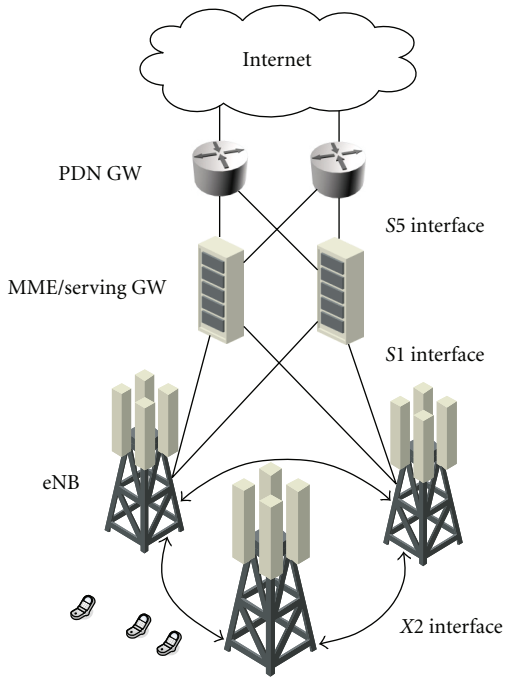


Figure 1: LTE Release 8 architecture.

2.2.2. Radio Access Fundamentals. The most important technologies included in the new radio access network are Orthogonal Frequency Division Multiplexing (OFDM), multidimensional (time, frequency) dynamic resource allocation and link adaptation, Multiple Input Multiple Output (MIMO) transmission, turbo coding and hybrid Automatic Repeat reQuest (ARQ) with soft combining. These technologies are shortly explained in the following paragraphs.

OFDM. Orthogonal Frequency Division Multiplexing is a kind of multicarrier transmission technique with a relatively large number of subcarriers. OFDM offers a lot of advantages. First of all, by using a multiple carrier transmission technique, the symbol time can be made substantially longer than the channel delay spread, which reduces significantly or even removes the intersymbol interference (ISI). In other words, OFDM provides a high robustness against frequency selective fading. Secondly, due to its specific structure, OFDM allows for low-complexity implementation by means of Fast Fourier Transform (FFT) processing. Thirdly, the access to the frequency domain (OFDMA) implies a high degree of freedom to the scheduler. Finally, it offers spectrum flexibility which facilitates a smooth evolution from already existing radio access technologies to LTE.

In the FDD mode of LTE each OFDM symbol is transmitted over subcarriers of 15 or 7.5 kHz. One subframe lasts 1 ms, divided in two 0.5 ms slots, and contains several consecutive OFDM symbols (14 and 12 for the 15 and 7.5 kHz modes, resp.).

In the uplink, Single Carrier Frequency Division Multiple Access (SC-FDMA) is used rather than OFDM. SC-FDMA is also known as DFT-spread OFDM modulation. Basically, SC-FDMA is identical to OFDM unless an initial FFT is applied before the OFDM modulation. The objective of such modification is to reduce the peak to average power ratio, thus decreasing the power consumption in the user terminals.

Multidimensional Dynamic Resource Allocation and Link Adaptation. In LTE, both uplink and downlink transmission schemes can assign smaller, nonoverlapping frequency bands to the different users, offering frequency division multiple access (FDMA). This assignment can be dynamically adjusted in time and is referred to as scheduling. Accordingly, the LTE resources can be represented as a time-frequency grid. The minor element of this grid is called resource

element and consists of one subcarrier during an OFDM symbol. However, the minor LTE resource allocation unit is the resource block that consists of 12 subcarriers during one slot.

Link adaptation is closely related to scheduling and deals with how to set the transmission parameters of a radio link to handle variations of the radio-link quality. This is achieved in LTE through adaptive channel coding and adaptive modulation. Specifically, in LTE available modulations are QPSK, 16QAM and 64QAM, whilst coding rate can take values from a lower edge of around 0.07 up to 0.93.

MIMO. One of the most important means to achieve the high data rate objectives for LTE is multiple antenna transmission. In LTE downlink it is supported one, two or four transmit antennas in the eNB and one, two or four receive antennas in the UE. Multiple antennas can be used in different ways: to obtain additional transmit/receive diversity or to get spatial multiplexing increasing the data rate by creating several parallel channels if conditions allow to. Nevertheless, in LTE uplink although one, two or four receive antennas are allowed in the eNB, only one transmitting antenna is allowed in the UE. Therefore, multiple antennas can be only used to obtain receive diversity.

Turbo Coding. In order to correct bit errors, introduced by channel variations and noise, channel coding is utilized. In case of the LTE downlink shared channel (DL-SCH) a turbo encoder with rate 1/3 is used, followed by a rate matching to adapt the coding rate to the desired level. In each subframe of 1 ms, one or two (with multicodeword MIMO) codewords can be coded and transmitted.

Hybrid ARQ with Soft Combining. Hybrid ARQ with soft combining is a technique that deals with the retransmission of data in case of errors. In an ARQ scheme, the receiver uses an error-detecting code to check if the received packet contains errors or not. The transmitter is informed by a NACK or ACK respectively. In case of a NACK, the packet is retransmitted.

A combination of forward error correction (FEC) and ARQ is known as hybrid ARQ. Most practical hybrid ARQ schemes are built around a CRC code for error detection and a turbocode for error correction, as it is the case of LTE.

In hybrid ARQ with soft combining, the erroneously received packet is stored in a buffer and later combined with the retransmission(s) to obtain a single packet that is more reliable than its constituents. In LTE full incremental redundancy (IR) is applied, which means that the retransmitted packets are typically not identical with the first transmission but carry complementary information.

2.3. Analysis of LTE Performance. Different methods can be used to assess the performance of a mobile technology. Each method is best suited for a particular kind of performance assessment. For instance, analytical methods or inspections are valid to evaluate peak data rates or peak spectral efficiencies. However, a deeper performance analysis requires the usage of simulation. Simulators are usually divided in two classes: link level simulators and system level simulators. Link level simulators are used to emulate the transmission of information from a unique transmitter to a unique receiver modeling the physical layer with high precision. They include models for coding/decoding, MIMO processing, scrambling, modulation, channel, channel estimation and equalization, and so forth. System level simulators emulate the operation of a network with a number of cells and several users per cell. In this kind of simulators, higher level functions are included for call admission control, scheduling, power control, and so forth, while link to system level models is used to facilitate the emulation of each radio link. This section presents some results obtained from both types of simulators.

In the course of the LTE standardization process, the 3GPP conducted several deep evaluations of the developing technology to ensure the achievement of requirements. With this aim, a feasibility study for E-UTRA and E-UTRAN was carried out in the 3GPP. Reference framework for the performance analysis is set by two documents [9, 10], to ensure the comparability of the different results. Mean LTE performance results obtained by the 3GPP partners are included in [11] where the results are also compared to the requirements. Results shown in that document are a summary of those in [12, 13] that collect the results of all the partners. In this assessment the used scenarios are similar to those used by the 3GPP to allow comparability of results.

This assessment allows getting an insight into to which extent LTE implies a revolution in comparison with UMTS. As shown in next section, LTE results demonstrate that this technology is quite close to the requirements established for the Fourth-Generation mobile, although further improvements are expected in LTE-Advanced.

2.3.1. Peak Spectral Efficiency. The peak spectral efficiency is the highest theoretical data rate assignable to a single mobile user divided by the allocated bandwidth. The highest data rate is calculated as the received data bits assuming error-free conditions and excluding radio resources that are used for control issues and guard bands. At the end, the radio access technology is classified as more or less powerful according to the achievable efficiency what makes this measurement perfect for comparative purposes.

Assuming a transmission bandwidth of 20 MHz the maximum achievable rates in downlink are: 91.2 Mbps for SIMO 1×2, 172.8 Mbps for MIMO 2×2 and 326.4 Mbps for MIMO 4×4. The resulting peak spectral efficiencies are 4.56, 8.64 and 16.32 b/s/Hz for the considered multiantenna schemes. These values have been calculated taking into account realistic overhead due to the reference signals and assuming that control signals overhead is equal to one OFDM symbol in each subframe. In uplink with SIMO 1×2 the maximum achievable rate is 86.4 Mbps with a transmission bandwidth of 20 MHz. Thus, the peak spectral efficiency is 4.32 b/s/Hz. These values have been calculated assuming that two OFDM symbols are occupied by reference signals. Both in downlink and uplink calculations 64QAM is the considered modulation and code rate is assumed to be 1.

The calculated peak spectral efficiencies of LTE are depicted in Figure 2 for both downlink and uplink together with the efficiencies of UMTS Release 6, that is, including HSDPA and HSUPA. From this peak spectrum efficiency it can be seen that LTE with 20 MHz meets and exceeds the 100 Mbps downlink and 50 Mbps uplink initial targets. Besides, the comparison with UMTS demonstrates that LTE is a major step forward in mobile radio communications. With these achievable data rates mobile systems will give a greater user experience with the capability of supporting more demanding applications.

2.3.2. LTE Link Level Performance. Based on link level simulations it can be assessed the relation between effective throughput (correctly received bits per time unit) and signal-to-noise plus interference ratio (SINR). Simulations assessed for this paper used 10 MHz of bandwidth for both downlink and uplink. This bandwidth is equivalent to 50 LTE resource blocks. The evaluation was focused on the performance experienced by a pedestrian user and hence the user mobility model used was the extended pedestrian A model [14] with a Doppler frequency of 5 Hz. The central frequency has been set to 2.5 GHz, the most likely band for initial LTE deployment. The set of modulation and coding schemes has been selected from the CQI table included in LTE specifications [15]. This set was selected by 3GPP to cover the LTE SINR dynamic margin with approximately 2 dB steps between consecutive curves. A distinction from other studies is that channel estimation was realistically calculated at the receivers. In order to exploit the multiantenna configuration at the receiver side, minimum mean-square error (MMSE) equalization was considered. The remaining parameters considered in the simulations are summarized in Table 1.

Concerning LTE downlink, different multiantenna configurations were modeled including SIMO 1×2, MIMO 2×2 and MIMO 4×4. Simulated MIMO scheme followed the open loop spatial multiplexing scheme as specified by the 3GPP [16], the number of codewords was 2 and the number of layers was equal to the number of transmit antennas, that is, 2 and 4. Additionally, the multiple channels among antennas were supposed uncorrelated. Control channel and signals overhead were taken into account and hence the first two OFDM symbols in each subframe were reserved for control channels. Besides, reference signals were emulated in detail, although neither broadcast information nor synchronization signals overhead was considered.

In the uplink, two different multiantenna configurations were simulated: SIMO 1×2 and SIMO 1×4. The multiple channels among antennas were supposed uncorrelated too. Nowadays, the LTE standard does not allow MIMO in uplink so that MIMO schemes were not simulated. Therefore, as established in the 3GPP specifications [17], only one codeword was considered. Moreover, 12 of the 14 available SC-FDMA symbols in a subframe were occupied by codified data since the other 2 were reserved for reference signals needed for the channel estimation at the receiver.

Taking into account these assumptions and parameters, a set of simulations was performed whose results are shown in Figure 3 for LTE downlink and in Figure 4 for LTE uplink. In both figures it can be observed that the maximum throughputs are not equal to the peak throughputs previously calculated. The reason is threefold: the used bandwidth is not 20 MHz but 10 MHz, the highest coding rate used is 0.93 instead of 1 and downlink control signals overhead is assumed to be 2 OFDM symbols instead of 1.

In LTE downlink, according to the results shown in Figure 3, MIMO 4×4 scheme provides a clearly better performance than the other schemes for almost all the useful SINR margin. Nevertheless, MIMO 2×2 scheme does not provide an important performance improvement until SINR reaches a value of 15 dB. Also, it can be observed that improvement factor in peak throughput due to MIMO schemes is far from being equal to the number of antennas (2 or 4). Instead, peak throughput is multiplied by 1.7 and 3.6 in MIMO 2×2 and MIMO 4×4 respectively. This is basically due to the higher quantity of reference signals needed in the MIMO schemes.

In LTE uplink, there is not any peak throughput gain when using more receiver antennas. But a nonnegligible SINR gain can be achieved. This gain is about 5 dB for a throughput of 20 Mbps. Note that in SIMO 1×4 maximum rate is achieved 10 dB before than in SIMO 1×2.

2.3.3. LTE System Level Performance. LTE performance analysis at system level requires the definition of system level statistics. The cell spectral efficiency and the cell edge user spectral efficiency are the more important ones. Given a multiuser/multicell scenario, the cell spectral efficiency is defined as the aggregate throughput of all users (the number of correctly received bits over a certain period of time) normalized by the overall cell bandwidth and divided by the number of cells. In the same scenario, the cell edge user spectral efficiency is the 5% point of CDF of the user throughput normalized with the overall cell bandwidth.

In order to calculate these values in the downlink, a dynamic system level simulator has been used. The main parameters of the considered scenario are shown in Table 1. The scenario is similar to the "Case 1" scenario in [9]. The main differences in this assessment are that the channel has been implemented using a tapped delay line model and a low correlation among channels has been assumed. Specifically, an ETU channel has been used [14]. The scheduler operation follows the proposal of [18] where scheduling algorithm is divided in two parts: one for the time domain and another for the frequency domain. For both domains a proportional fair approach has been used.

Following the proposed approach, average cell spectral efficiency in downlink was obtained yielding 1.52 bps/Hz/cell for SIMO 1×2, 1.70 bps/Hz/cell for MIMO 2×2 and 2.50 bps/Hz/cell for MIMO 4×4. The cell edge user spectral efficiencies are 0.02 bps/Hz/user, 0.03 bps/Hz/user and 0.05 bps/Hz/user, for the same antenna configurations. Note that the LTE values for the uplink have been extracted from the results presented by the 3GPP partners in [12], since the downlink values obtained in this assessment fit with 3GPP results. Since LTE requirements were defined as

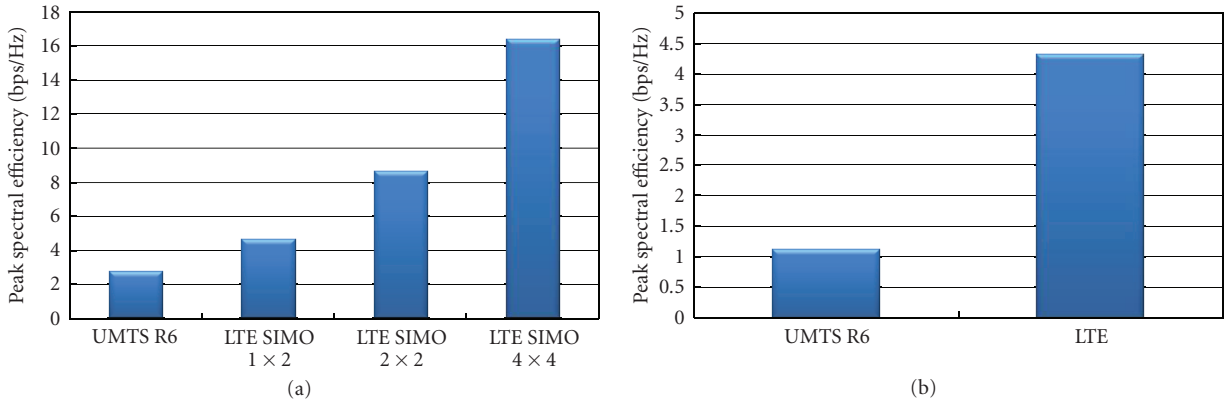


Figure 2: LTE peak spectral efficiencies in downlink (a) and uplink (b).

Table 1: Simulation parameters.

Common parameters		
Bandwidth		10 MHz (50 RB)
Channel		Tapped delay line: EPA with 5 Hz Doppler frequency at link level, ETU at system level
Central frequency		2.5 GHz
MCS		CQI 1–15
Multiantenna schemes	DL	SIMO 1×2, MIMO $2 \times 2/4 \times 4$
	UL	SIMO $1 \times 2/1 \times 4$
Control channels overhead	DL	2 OFDM symbols per subframe
	UL	Not considered
System level parameters		
Inter site distance (ISD)		500 m
Cell deployment		3-sector cells, reuse 1
Pathloss		$130.2 + 37.6 \log 10\,(\mathrm{d(km)})$ dB
Shadowing		lognormal, $\sigma = 8$ dB
eNB transmission power		46 dBm
Noise spectral density		−174 dBm/Hz
Scheduler		Proportional Fair in time and frequency domains up to 5 UEs is scheduled per subframe
Mobility		Users moving at 30 km/h

relative to HSPA performance, Table 2 includes HSPA figures extracted also from [12, 13]. After direct inspection, it can be concluded that most of the requirements specified by 3GPP are fulfilled by the current Release 8 version of LTE.

3. LTE-Advanced and the Fourth-Generation Mobile

The process of defining the future IMT-Advanced family was started with a Circular Letter issued by ITU-R calling for submission of candidate Radio Interface Technologies (RITs) and fvcandidate sets of Radio Interface Technologies (SRITs) for IMT-Advanced [1]. However, all documents available in that moment concerning IMT-Advanced did not specify any new technical details about the properties of future 4G systems. Instead, they just reference the Recommendation M.1645 [19], in which the objectives of the future development of IMT-Advanced family were barely defined: to reach 100 Mb/s for mobile access and up to 1 Gb/s for nomadic wireless access. Unfortunately, it was not until November 2008 when the requirements related to technical performance for IMT-Advanced candidate radio interfaces were described [20].

Just after receiving the Circular Letter, the 3GPP organized a workshop on IMT-Advanced where the following decisions were made.

(i) LTE-Advanced will be an evolution of LTE. Therefore LTE-Advanced must be backward compatible with LTE Release 8.

(ii) LTE-Advanced requirements will meet or even exceed IMT-Advanced requirements following the ITU-R agenda.

(iii) LTE-Advanced should support significantly increased instantaneous peak data rates in order to reach ITU requirements. Primary focus should be on low mobility users. Moreover, it is required a further improvement of cell edge data rates.

Table 2: LTE requirements related to technical performance.

		Requirements LTE TR 25.913	LTE simulation results
Peak data rate (Gbps)		0.1	0.172 (2 × 2) 0.326 (4 × 4)
Latency		C-Plane < 100 ms U-Plane < 5 ms	—
Peak spectral efficiency (bps/Hz)	DL	5 (1 × 2)	4.56 (1 × 2) 8.64 (2 × 2) 16.32 (4 × 4)
	UL	2.5 (1 × 2)	4.32 (1 × 2)
Average spectral efficiency (bps/Hz/cell)	DL	EUTRA (2 × 2) 3-4 times HSDPA R6 {0.53}	1.52 (1 × 2) 1.70 (2 × 2) 2.50 (4 × 4)
	UL	EUTRA (1 × 2) 2-3 times HSUPA R6 (1 × 2) {0.332}	0.73 (1 × 2)
Cell edge user spectral efficiency (bps/Hz/cell/user)	DL	EUTRA (2 × 2) 2-3 times HSDPA R6 {0.02}	0.02 (1 × 2) 0.04 (2 × 2) 0.05 (4 × 4)
	UL	EUTRA (1 × 2) 2-3 times HSUPA R6 (1 × 2) {0.009}	0.02 (1 × 2)
Mobility		Up to 350 km/h	30 km/h
Bandwidth		Up to 20 MHz	10 MHz

Table 3: IMT-Advanced requirements related to LTE-Advanced requirements.

		Requirement ITU-R M.2134	Requirements LTE-A TR 36.913
Peak data rate (Gbps)		1	1-(DL) 0.5-(UL)
Latency		C-Plane < 100 ms U-Plane < 10 ms	C-Plane < 50 ms U-Plane < 5 ms
Peak spectral efficiency (bps/Hz)	DL	15 (4 × 4)	30 (8 × 8)
	UL	6.75 (2 × 4)	15 (4 × 4)
Cell spectral efficiency (bps/Hz/cell)	DL	2.2 (4 × 2)	2.4 (2 × 2) 2.6 (4 × 2) 3.7 (4 × 4)
	UL	1.4 (2 × 4)	1.2 (1 × 2) 2.0 (2 × 4)
Cell edge user spectral efficiency (bps/Hz/cell/user)	DL	0.06 (4 × 2)	0.07 (2 × 2) 0.09 (4 × 2) 0.12 (4 × 4)
	UL	0.03 (2 × 4)	0.04 (1 × 2) 0.07 (2 × 4)
Mobility		Up to 350 km/h	Up to 350 km/h
Bandwidth		>40 MHz	Up to 100 MHz

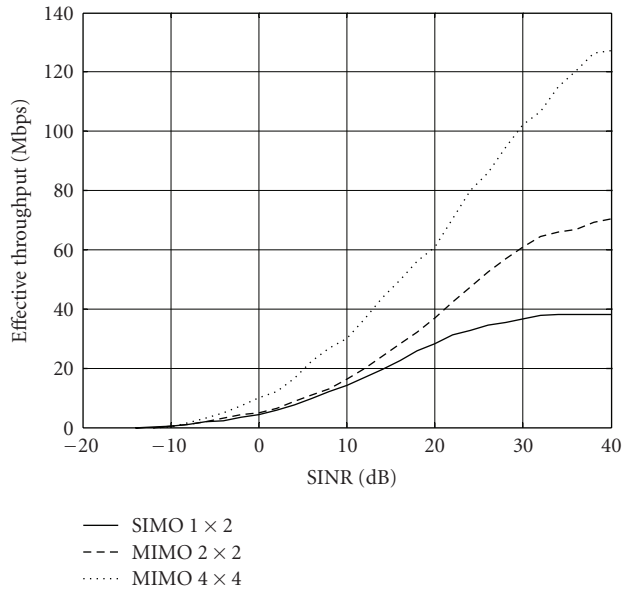


Figure 3: Link level evaluation of throughput versus SINR in LTE downlink.

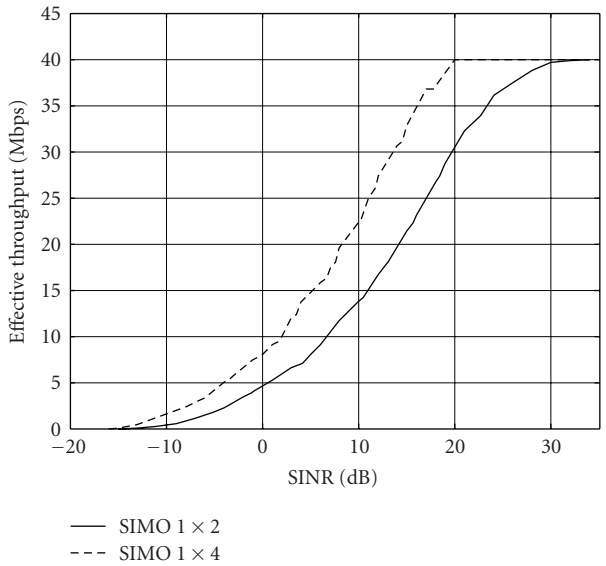


Figure 4: Link level evaluation of throughput versus SINR in LTE uplink.

With these clear objectives, and without knowing the final technical requirements yet, 3GPP defined a bullets list with the first requirements for LTE-Advanced and some technical proposals. Besides, it was decided to officially gather and approve them in the technical report TR 36.913 [21]. The remaining of this section deals with both aspects: requirements and technical proposals for LTE-Advanced.

3.1. LTE-Advanced Requirements. The requirement specification list was also included in TR 36.913. Although it is expected a list extension, these are some of the current agreements on the requirements for LTE-Advanced [21].

(i) Peak data rate of 1 Gbps for downlink (DL) and 500 Mbps for uplink (UL).

(ii) Regarding latency, in the C-plane the transition time from Idle to Connected should be lower than 50 ms. In the active state, a dormant user should take less than 10 ms to get synchronized and the scheduler should reduce the U-plane latency at maximum.

(iii) The system should support downlink peak spectral efficiency up to 30 bps/Hz and uplink peak spectral efficiency of 15 bps/Hz with an antenna configuration of 8×8 or less in DL and 4×4 or less in UL.

(iv) The 3GPP defined a base coverage urban scenario with intersite distance of 500 m and pedestrian users. Assuming this scenario, average user spectral efficiency in DL must be 2.4 bps/Hz/cell with MIMO 2×2, 2.6 bps/Hz/cell with MIMO 4×2 and 3.7 bps/Hz/cell with MIMO 4×4, whereas in UL the target average spectral efficiency is 1.2 bps/Hz/cell and 2.0 bps/Hz/cell with SIMO 1×2 and MIMO 2×4, respectively.

(v) In the same scenario with 10 users, cell edge user spectral efficiency will be 0.07 bps/Hz/cell/user in DL 2×2, 0.09 in DL 4×2 and 0.12 in DL 4×4. In the UL, this cell edge user spectral efficiency must be 0.04 bps/Hz/cell/user with SIMO 1×2 and 0.07 with MIMO 2×4.

(vi) The mobility and coverage requirements are identical to LTE Release 8. There are only differences with indoor deployments that need additional care in LTE-Advanced.

(vii) In terms of spectrum flexibility, the LTE-Advanced system will support scalable bandwidth and spectrum aggregation with transmission bandwidths up to 100 MHz in DL and UL.

(viii) LTE-Advanced must guarantee backward compatibility and interworking with LTE and with other 3GPP legacy systems.

Table 3 summarizes the list of requirements established by ITU-R and 3GPP allowing a direct comparison among 4G and LTE-Advanced. According to this table, it can be concluded that LTE-Advanced is being designed to be a strong candidate for next 4G, since it fulfils or even exceeds all IMT-Advanced requirements.

3.2. LTE-Advanced Technical Proposals. LTE Release 8 can already fulfill some of the requirements specified for IMT-Advanced systems. However, it is also clear that there are more challenging requirements under discussion in the 3GPP, which would need novel radio access techniques and system evolution. The 3GPP working groups, mainly RAN1 working on the physical layer, are currently evaluating some techniques to enhance LTE Release 8 performance. This section offers an overview of some of these proposals.

Support of Wider Bandwidth. A significant underlying feature of LTE-Advanced will be the flexible spectrum usage.

The framework for the LTE-Advanced air-interface technology is mostly determined by the use of wider bandwidths, potentially even up to 100 MHz, noncontiguous spectrum deployments, also referred to as spectrum aggregation, and a need for flexible spectrum usage.

In general OFDM provides a simple means to increase bandwidth: adding additional subcarriers. Due to the discontinuous spectrum reserved for IMT-Advanced, the available bandwidth might also be fragmented. Therefore, the user equipments should be able to filter, process and decode such a large variable bandwidth. The increased decoding complexity is one of the major challenges of this wider bandwidth.

Concerning the resource allocation in the eNB and the backward compatibility, minimum changes in the specifications will be required if scheduling, MIMO, Link Adaptation and HARQ are performed over groups of carriers of 20 MHz. For instance, a user receiving information in 100 MHz bandwidth will need 5 receiver chains, one per each 20 MHz block.

Coordinated Multiple Point Transmission and Reception. Coordinated multi point transmission and reception are considered for LTE-Advanced as one of the most promising techniques to improve data rates, hence increasing average cell throughput. It consists in coordinating the transmission and reception of signal from/to one UE in several geographically distributed points. So far, the discussions have focused on classifying the different alternatives and identifying their constraints. Potential impact on specifications comprises three areas: feedback and measurement mechanisms from the UE, preprocessing schemes and reference signal design.

Relaying Functionality. Relaying can be afforded from three different levels of complexity. The simplest one is the Layer 1 relaying, that is, the usage of repeaters. Repeaters receive the signal, amplify it and retransmit the information thus covering black holes inside cells. Terminals can make use of the repeated and direct signals. However, in order to combine constructively both signals there should be a small delay, less than the cyclic prefix, in their reception.

In Layer 2 relaying the relay node has the capability of controlling at least part of the RRM functionality. In some slots the relay node acts as a user terminal being in the subsequent slot a base station transmitting to some users located close to the relay.

Finally, Layer 3 relaying is conceived to use the LTE radio access in the backhaul wireless connecting one eNB with another eNB that behaves as a central hub. This anchor eNB routes the packets between the wired and wireless backhaul, acting like an IP router.

Enhanced Multiple-Input Multiple-Output Transmission. Another significant element of the LTE-Advanced technology framework is MIMO, as in theory it offers a simple way to increase the spectral efficiency. The combination of higher order MIMO transmission, beamforming or MultiUser (MU) MIMO is envisaged as one of the key technologies for LTE-Advanced.

In case of spectrum aggregation, the antenna correlation may be different in each spectrum segment given a fixed antenna configuration. Therefore, in LTE-Advanced one channel element may encompass both low correlation and high correlation scenarios simultaneously. Since MU-MIMO is more appropriated for high correlation scenarios than Single-User (SU) MIMO, to fully utilize the characteristics of different scattering scenarios both SU-MIMO and MU-MIMO should be simultaneously used.

4. Conclusions

LTE has been designed as a future technology to cope with next user requirements. In this paper two complete LTE Release 8 link and system level simulators have been presented together with several performance results. Based on these results, this paper concludes that LTE will offer peak rates of more than 150 Mbps in the downlink and 40 Mbps in the uplink with 10 MHz bandwidth. Besides, in the downlink the minimum average throughput will be around 30 Mbps, which represents a quite significant improvement in the cellular systems performance. As compared with current cellular systems, LTE entails an enhancement of more than six times the performance of HSDPA/HSUPA.

This paper has also given an initial insight into the new technical proposals currently under discussion in the framework of 3GPP. This analysis allows those who are interested in wireless communications to get aligned with the research community towards the definition and optimization of next Fourth-Generation mobile.

Acknowledgments

Part of this work has been performed in the framework of the CELTIC Project CP5-026 WINNER+. This work was supported by the Spanish Ministry of Science under the Project TEC2008-06817-C02-01/TEC.

References

[1] ITU-R Circular Letter 5/LCCE/2, "Invitation for submission of proposals for candidate radio interface technologies for the terrestrial components of the radio interface(s) for IMT-Advanced and invitation to participate in their subsequent evaluation," March 2008.

[2] J. J. Sánchez, D. Morales-Jiménez, G. Gómez, and J. T. Enbrambasaguas, "Physical layer performance of long term evolution cellular technology," in *Proceedings of the 16th IST Mobile and Wireless Communications Summit*, pp. 1–5, Budapest, Hungary, July 2007.

[3] A. Osman and A. Mohammed, "Performance evaluation of a low-complexity OFDM UMTS-LTE system," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2142–2146, Singapore, May 2008.

[4] K. Manolakis, A. Ibing, and V. Jungnickel, "Performance evaluation of a 3GPP LTE terminal receiver," in *Proceedings of the 14th European Wireless Conference (EW '08)*, pp. 1–5, Prague, Czech Republic, June 2008.

[5] C. Spiegel, J. Berkmann, Z. Bai, T. Scholand, and C. Drewes, "MIMO schemes in UTRA LTE, a comparison," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2228–2232, Singapore, May 2008.

[6] N. Wei, A. Pokhariyal, C. Rom, et al., "Baseline E-UTRA downlink spectral efficiency evaluation," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '06)*, pp. 2131–2135, Montreal, Canada, September 2006.

[7] P. Mogensen, W. Na, I. Z. Kovács, et al., "LTE capacity compared to the shannon bound," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 1234–1238, Dublin, Ireland, April 2007.

[8] 3GPP TR 25.913, "Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN)," v.8.0.0, December 2008.

[9] 3GPP TR 25.814, "Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)," September 2006.

[10] 3GPP R1-070674, "LTE physical layer framework for performance evaluation," February 2007.

[11] 3GPP TR 25.912, "Feasibility Study for evolved Universal Terrestrial Radio Access (UTRA) and Universal Terrestrial Radio Access Network (UTRAN)," v.8.0.0, December 2008.

[12] 3GPP R1-072261, "LTE Performance Evaluation—Uplink Summary," May 2007.

[13] 3GPP R1-072578, "Summary of Downlink Performance Evaluation," May 2007.

[14] 3GPP TS 36.101, "User Equipment (UE) radio transmission and reception," v.8.5.1, March 2008.

[15] 3GPP TS 36.213, "Physical layer procedures," v.8.6.0, March 2009.

[16] 3GPP TS 36.211, "Long Term Evolution physical layer; General description," v.8.6.0, March 2009.

[17] 3GPP TS 36.212, "Multiplexing and channel coding," v.8.6.0, March 2009.

[18] G. Monghal, K. I. Pedersen, I. Z. Kovács, and P. E. Mogensen, "QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution," in *Proceedings of the 67th IEEE Vehicular Technology Conference (VTC '08)*, pp. 2532–2536, May 2008.

[19] ITU-R Recommendation M.1645, "Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000," June 2003.

[20] ITU-R Report M.2134, "Requirements related to technical performance for IMT-Advanced radio interface(s)," November 2008.

[21] 3GPP TR 36.913, "Requirements for Further Advancements for E-UTRA," v.8.0.1, March 2009.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 302092, 10 pages
doi:10.1155/2009/302092

Review Article

MIMO Technologies in 3GPP LTE and LTE-Advanced

Juho Lee,[1] Jin-Kyu Han,[1] and Jianzhong (Charlie) Zhang[2]

[1] *Digital Media & Communications R&D Center, Samsung Electronics, 416, Maetan-3dong, Yeongtong-gu, Suwon-si 443-742, South Korea*
[2] *Wireless System Lab, Samsung Telecom America, 1301 E. Lookout Drive, Richardson, TX 75082, USA*

Correspondence should be addressed to Juho Lee, juho95.lee@samsung.com

Received 13 February 2009; Accepted 31 May 2009

Recommended by Angel Lozano

3rd Generation Partnership Project (3GPP) has recently completed the specification of the Long Term Evolution (LTE) standard. Majority of the world's operators and vendors are already committed to LTE deployments and developments, making LTE the market leader in the upcoming evolution to 4G wireless communication systems. Multiple input multiple output (MIMO) technologies introduced in LTE such as spatial multiplexing, transmit diversity, and beamforming are key components for providing higher peak rate at a better system efficiency, which are essential for supporting future broadband data service over wireless links. Further extension of LTE MIMO technologies is being studied under the 3GPP study item "LTE-Advanced" to meet the requirement of IMT-Advanced set by International Telecommunication Union Radiocommunication Sector (ITU-R). In this paper, we introduce various MIMO technologies employed in LTE and provide a brief overview on the MIMO technologies currently discussed in the LTE-Advanced forum.

1. Introduction

As multimedia communications become increasingly popular, mobile communications are expected to reliably support high data rate transmissions. Multiple input multiple output (MIMO) has been treated as an emerging technology to meet the demand for higher data rate and better cell coverage even without increasing average transmit power or frequency bandwidth, since it was proved that MIMO structure successfully constructs multiple spatial layers where multiple data streams are delivered on a given frequency-time resource and linearly increases the channel capacity [1–8]. Lots of recently specified wireless communications standards are ready to support MIMO technologies.

3rd Generation Partnership Project (3GPP) has recently specified an Orthogonal Frequency Division Multiplexing (OFDM) based technology, Evolved Universal Terrestrial Radio Access (E-UTRA), for support of wireless broadband data service up to 300 Mbps in the downlink and 75 Mbps in the uplink [9]. (E-UTRA is also known as LTE in the wireless industry.) In Long Term Evolution (LTE), MIMO technologies have been widely used to improve downlink peak rate, cell coverage, as well as average cell throughput.

To achieve this diverse set of objectives, LTE adopted various MIMO technologies including transmit diversity, single user (SU)-MIMO, multiuser (MU)-MIMO, closed-loop rank-1 precoding, and dedicated beamforming [10–13]. The SU-MIMO scheme is specified for the configuration with two or four transmit antennas in the downlink, which supports transmission of multiple spatial layers with up to four layers to a given User Equipment (UE). The transmit diversity scheme is specified for the configuration with two or four transmit antennas in the downlink, and with two transmit antennas in the uplink. The MU-MIMO scheme allows allocation of different spatial layers to different users in the same time-frequency resource, and is supported in both uplink and downlink. The closed-loop rank-1 precoding scheme is used to improve data coverage utilizing SU-MIMO technology based on the cell-specific common reference signal while introducing a control signal message that has lower overhead. The dedicated beamforming scheme is used for data coverage extension when the data demodulation based on dedicated reference signal is supported by the UE.

A study item called "LTE-Advanced" has recently started in 3GPP, with the goal of providing a competitive IMT-Advanced candidate proposal with accompanying

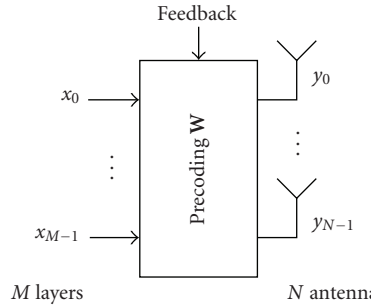


Figure 1: Closed-loop spatial multiplexing with N antennas and M layers.

Table 1: Precoding codebook for transmission on two antennas.

Codebook index	Number of layers M	
	1	2
0	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}$	—
1	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$	$1/2\begin{bmatrix}1&1\\1&-1\end{bmatrix}$
2	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\j\end{bmatrix}$	$1/2\begin{bmatrix}1&1\\j&-j\end{bmatrix}$
3	$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\-j\end{bmatrix}$	—

self-evaluation results for the submission to International Telecommunication Union Radiocommunication Sector (ITU-R) in October 2009 [14]. In LTE-Advanced, the existing SU-MIMO technologies are extended to support configuration with up to eight transmit antennas in the downlink, and up to four transmit antennas in the uplink. In addition, multicell coordinated multipoint transmission (COMP) are also under active discussion and evaluation [15].

In this paper, we introduce various MIMO technologies employed in LTE, and provide a brief overview on the MIMO technologies currently discussed in the LTE-Advanced forum.

2. Downlink SU-MIMO in LTE

The SU-MIMO scheme is applied to the Physical Downlink Shared Channel (PDSCH), which is the physical layer channel that carries the information data from the network to the UE. With SU-MIMO spatial multiplexing, the LTE system provides a peak rate of 150 Mbps for two transmit antennas and 300 Mbps for four transmit antennas [16]. There are two operation modes in SU-MIMO spatial multiplexing: the closed-loop spatial multiplexing mode and the open-loop spatial multiplexing mode.

In the closed-loop spatial multiplexing mode, the base station (also known as eNodeB) applies the spatial domain precoding on the transmitted signal taking into account the precoding matrix indicator (PMI) reported by the UE so that the transmitted signal matches with the spatial channel experienced by the UE. The closed-loop spatial multiplexing with M layers and N transmit antennas ($N \geq M$) is illustrated in Figure 1. To support the closed-loop spatial multiplexing in the downlink, the UE needs to feedback the rank indicator (RI), the PMI, and the channel quality indicator (CQI) in the uplink. The RI indicates the number of spatial layers that can be supported by the current channel experienced at the UE. The eNodeB may decide the transmission rank, M, taking into account the RI reported by the UE as well as other factors such as traffic pattern, available transmission power, etc. The CQI feedback indicates a combination of modulation scheme and channel coding rate that the eNodeB should use to ensure that the block error probability experienced at the UE will not exceed 10%.

The precoding operation for the closed-loop spatial multiplexing is defined by

$$\mathbf{y} = \mathbf{W}\mathbf{x}, \tag{1}$$

where $\mathbf{y} = [y_0, \ldots, y_{N-1}]^T$, $\mathrm{y_n}$ denotes the complex symbol transmitted on the nth antenna, $\mathbf{x} = [x_0, \ldots, x_{M-1}]^T$, $\mathrm{x_m}$ denotes the modulation symbol transmitted on the mth layer, and $\mathbf{W}$ denotes the $N \times M$ precoding matrix. For transmission on two antennas, the precoding matrix $\mathbf{W}$ is selected from Table 1, where each column vector is in the form of $[\,1 \; e^{j(\theta+k\pi)}\,]^T$ multiplied by a scaling factor. For transmission on four antennas, the precoding matrix $\mathbf{W}$ is selected from Table 2, where $\mathbf{W}_i^{\{c_1 \ldots c_m\}}$ denotes the matrix defined by the columns $c_1, \ldots, c_m$ of the matrix $\mathbf{W}_i = \mathbf{I}_{4\times 4} - 2\mathbf{u}_i\mathbf{u}_i^H/\mathbf{u}_i^H\mathbf{u}_i$. Design of the precoding for four transmit antennas is based on the Householder transformation [17] to reduce the computational complexity at the UE as well as the design complexity for finding out suitable precoding matrices due to its structure. Note that the downlink reference signal is common for all UEs belonging to the cell and hence is not precoded by $\mathbf{W}$. The UE receives the information from the eNodeB on what precoding matrix is used, which is utilized by the UE for demodulating the data.

The precoding codebook was designed to satisfy the following properties.

Constant modulus. All physical transmit antennas keep the same transmit power level regardless which precoding matrix is used to maximize the power amplifier utilization efficiency.

Nested property. Each precoding matrix in a higher rank subcodebook can find at least one precoding matrix in a lower rank subcodebook, which is a submatrix of the higher rank precoding matrix. This property is included to ensure proper performance when the eNodeB overrides the RI report and decides on a transmission rank that is lower than the channel rank reported in the RI. This property also helps reduce the CQI calculation complexity as some of the precoding matrix selection procedures can be shared among precoding matrices designed for different transmission ranks. For example, in Table 2, for the precoding matrix with codebook index 0 in M = 3 subcodebook, $\mathbf{W}_0^{\{124\}}/\sqrt{3}$, we can find a precoding matrix with codebook index 0 in $M = 2$ subcodebook, $\mathbf{W}_0^{\{14\}}/\sqrt{2}$, which is a submatrix of $\mathbf{W}_0^{\{124\}}/\sqrt{3}$ (up to power scaling).

Table 2: Precoding codebook for transmission on four antennas.

Codebook index	$\mathbf{u}_i$	Number of layers M			
		1	2	3	4
0	$\mathbf{u}_0 = [1\ -1\ -1\ -1]^T$	$\mathbf{W}_0^{\{1\}}$	$\mathbf{W}_0^{\{14\}}/\sqrt{2}$	$\mathbf{W}_0^{\{124\}}/\sqrt{3}$	$\mathbf{W}_0^{\{1234\}}/2$
1	$\mathbf{u}_1 = [1\ -j\ 1\ j]^T$	$\mathbf{W}_1^{\{1\}}$	$\mathbf{W}_1^{\{12\}}/\sqrt{2}$	$\mathbf{W}_1^{\{123\}}/\sqrt{3}$	$\mathbf{W}_1^{\{1234\}}/2$
2	$\mathbf{u}_2 = [1\ 1\ -1\ 1]^T$	$\mathbf{W}_2^{\{1\}}$	$\mathbf{W}_2^{\{12\}}/\sqrt{2}$	$\mathbf{W}_2^{\{123\}}/\sqrt{3}$	$\mathbf{W}_2^{\{3214\}}/2$
3	$\mathbf{u}_3 = [1\ j\ 1\ -j]^T$	$\mathbf{W}_3^{\{1\}}$	$\mathbf{W}_3^{\{12\}}/\sqrt{2}$	$\mathbf{W}_3^{\{123\}}/\sqrt{3}$	$\mathbf{W}_3^{\{3214\}}/2$
4	$\mathbf{u}_4 = [1\ (-1-j)/\sqrt{2}\ -j\ (1-j)/\sqrt{2}]^T$	$\mathbf{W}_4^{\{1\}}$	$\mathbf{W}_4^{\{14\}}/\sqrt{2}$	$\mathbf{W}_4^{\{124\}}/\sqrt{3}$	$\mathbf{W}_4^{\{1234\}}/2$
5	$\mathbf{u}_5 = [1\ (1-j)/\sqrt{2}\ j\ (-1-j)/\sqrt{2}]^T$	$\mathbf{W}_5^{\{1\}}$	$\mathbf{W}_5^{\{14\}}/\sqrt{2}$	$\mathbf{W}_5^{\{124\}}/\sqrt{3}$	$\mathbf{W}_5^{\{1234\}}/2$
6	$\mathbf{u}_6 = [1\ (1+j)/\sqrt{2}\ -j\ (-1+j)/\sqrt{2}]^T$	$\mathbf{W}_6^{\{1\}}$	$\mathbf{W}_6^{\{13\}}/\sqrt{2}$	$\mathbf{W}_6^{\{134\}}/\sqrt{3}$	$\mathbf{W}_6^{\{1324\}}/2$
7	$\mathbf{u}_7 = [1\ (-1+j)/\sqrt{2}\ j\ (1+j)/\sqrt{2}]^T$	$\mathbf{W}_7^{\{1\}}$	$\mathbf{W}_7^{\{13\}}/\sqrt{2}$	$\mathbf{W}_7^{\{134\}}/\sqrt{3}$	$\mathbf{W}_7^{\{1324\}}/2$
8	$\mathbf{u}_8 = [1\ -1\ 1\ 1]^T$	$\mathbf{W}_8^{\{1\}}$	$\mathbf{W}_8^{\{12\}}/\sqrt{2}$	$\mathbf{W}_8^{\{124\}}/\sqrt{3}$	$\mathbf{W}_8^{\{1234\}}/2$
9	$\mathbf{u}_9 = [1\ -j\ -1\ -j]^T$	$\mathbf{W}_9^{\{1\}}$	$\mathbf{W}_9^{\{14\}}/\sqrt{2}$	$\mathbf{W}_9^{\{134\}}/\sqrt{3}$	$\mathbf{W}_9^{\{1234\}}/2$
10	$\mathbf{u}_{10} = [1\ 1\ 1\ -1]^T$	$\mathbf{W}_{10}^{\{1\}}$	$\mathbf{W}_{10}^{\{13\}}/\sqrt{2}$	$\mathbf{W}_{10}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{10}^{\{1324\}}/2$
11	$\mathbf{u}_{11} = [1\ j\ -1\ j]^T$	$\mathbf{W}_{11}^{\{1\}}$	$\mathbf{W}_{11}^{\{13\}}/\sqrt{2}$	$\mathbf{W}_{11}^{\{134\}}/\sqrt{3}$	$\mathbf{W}_{11}^{\{1324\}}/2$
12	$\mathbf{u}_{12} = [1\ -1\ -1\ 1]^T$	$\mathbf{W}_{12}^{\{1\}}$	$\mathbf{W}_{12}^{\{12\}}/\sqrt{2}$	$\mathbf{W}_{12}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{12}^{\{1234\}}/2$
13	$\mathbf{u}_{13} = [1\ -1\ 1\ -1]^T$	$\mathbf{W}_{13}^{\{1\}}$	$\mathbf{W}_{13}^{\{13\}}/\sqrt{2}$	$\mathbf{W}_{13}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{13}^{\{1324\}}/2$
14	$\mathbf{u}_{14} = [1\ 1\ -1\ -1]^T$	$\mathbf{W}_{14}^{\{1\}}$	$\mathbf{W}_{14}^{\{13\}}/\sqrt{2}$	$\mathbf{W}_{14}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{14}^{\{3214\}}/2$
15	$\mathbf{u}_{15} = [1\ 1\ 1\ 1]^T$	$\mathbf{W}_{15}^{\{1\}}$	$\mathbf{W}_{15}^{\{12\}}/\sqrt{2}$	$\mathbf{W}_{15}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{15}^{\{1234\}}/2$

Constrained alphabet. In case of two transmit antennas, constructing the precoding matrices by only using QPSK alphabet $\{\pm 1, \pm j\}$ except the scaling factor of $1/\sqrt{2}$ or 1/2 avoids the need for matrix multiplication in applying the precoder to the channel matrix without significant loss in precoding performance. In case of four transmit antennas, on the other hand, the QPSK alphabet constraint turns out to be a limiting factor in achieving additional spectral efficiency gain from additional antennas. It is also noted that if the Minimum Mean Squared Error (MMSE) receiver is assumed, most of the computational burden for CQI calculation comes from the matrix multiplication required to inverse the instantaneous covariance matrix and not from the multiplication of the channel matrix and the precoder. Taking into account these aspects, 8-PSK alphabet $\{\pm 1, \pm j, \pm(1+j)/\sqrt{2}, \pm(-1+j)/\sqrt{2}\}$ is used for the elements of vector $\mathbf{u}_i$ as a tradeoff between the computational complexity and the achievable performance of the codebook designed for four transmit antenna.

For the spatial multiplexing, multiple codewords may be mapped to multiple layers depending on the transmission rank scheduled by the eNodeB. In the LTE downlink, hybrid automatic repeat request (HARQ) process is operated for each codeword. Each HARQ process requires an ACK/NAK feedback signaling on uplink. To reduce the uplink feedback overhead, only up to two codewords are transmitted even though more than two layers can be transmitted on downlink in a given subframe, giving rise to the need of defining a rule for mapping a codeword to its layers. In LTE, codewords are mapped to layers according to Table 3, where $d_k(i)$ denotes the ith modulation symbol of the kth codeword, $x_l(i)$ denotes the ith modulation symbol of the lth layer, S_{layer} denotes the number of modulation symbols of each layer, and S_{k} denotes the number of modulation symbols of the kth codeword. If there is one layer, there is one codeword. If there are two layers, the basic mode of operation is to carry a codeword for each layer. The case of transmitting a single codeword using two layers is only applicable for the eNodeB having four transmit antennas when its initial transmission contained two codewords and a codeword mapped onto two layers needs to be retransmitted. In case of three-layer transmission, the first layer carry the first codeword while the second and the third layers carries the second codeword, in which case the second codeword has two times modulation symbols than the first one. When four layers are scheduled, two codewords are transmitted, each of which is transmitted using two layers. As can be seen in Table 3, the modulation symbols of a codeword are equally split into two layers when the codeword is mapped to two layers.

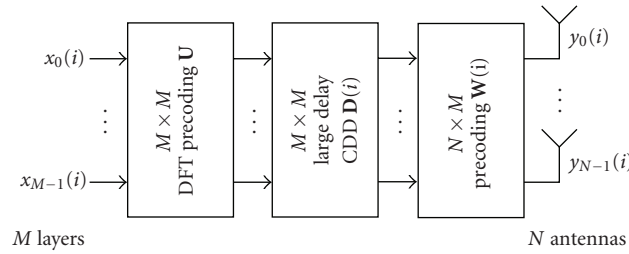


Figure 2: Open-loop spatial multiplexing with N antennas and M layers.

For the closed-loop spatial multiplexing, the eNodeB sends the scheduled UE the information about what precoding matrix is used as a part of downlink control information, using a three-bit information field for two transmit antennas and a six-bit information field for four transmit antennas. This information field is denoted transmit precoding matrix indication (TPMI). To support frequency-selective

Table 3: Codeword-to-layer mapping for spatial multiplexing.

Number of layers	Number of codewords	Codeword-to-layer mapping $i = 0, 1, \ldots, S_{layer} - 1$	
1	1	$x_0(i) = d_0(i)$	$S_{layer} = S_0$
2	2	$x_0(i) = d_0(i)$ $x_1(i) = d_1(i)$	$S_{layer} = S_0 = S_1$
2	1	$x_0(i) = d_0(2i)$ $x_1(i) = d_0(2i+1)$	$S_{layer} = S_0/2$
3	2	$x_0(i) = d_0(i)$ $x_1(i) = d_1(2i)$ $x_2(i) = d_1(2i+1)$	$S_{layer} = S_0 = S_1/2$
4	2	$x_0(i) = d_0(2i)$ $x_1(i) = d_0(2i+1)$ $x_2(i) = d_1(2i)$ $x_3(i) = d_1(2i+1)$	$S_{layer} = S_0/2 = S_1/2$

Table 4: DFT precoding matrix $\mathbf{U}$.

Number of layers M	M × M matrix $\mathbf{U}$
2	$\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & e^{-j2\pi/2} \end{bmatrix}$
3	$\frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{-j2\pi/3} & e^{-j4\pi/3} \\ 1 & e^{-j4\pi/3} & e^{-j8\pi/3} \end{bmatrix}$
4	$\frac{1}{2}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & e^{-j2\pi/4} & e^{-j4\pi/4} & e^{-j6\pi/4} \\ 1 & e^{-j4\pi/4} & e^{-j8\pi/4} & e^{-j12\pi/4} \\ 1 & e^{-j6\pi/4} & e^{-j12\pi/4} & e^{-j18\pi/4} \end{bmatrix}$

Table 5: Large delay CDD matrix $\mathbf{D}(i)$.

Number of layers M	$\mathbf{D}(i)$
2	$\begin{bmatrix} 1 & 0 \\ 0 & e^{-j2\pi i/2} \end{bmatrix}$
3	$\begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-j2\pi i/3} & 0 \\ 0 & 0 & e^{-j4\pi i/3} \end{bmatrix}$
4	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-j2\pi i/4} & 0 & 0 \\ 0 & 0 & e^{-j4\pi i/4} & 0 \\ 0 & 0 & 0 & e^{-j6\pi i/4} \end{bmatrix}$

precoding without excessive downlink signaling overhead, the TPMI can also indicate that the precoding matrices reported in the most recent PMI report from the scheduled UE are used for their corresponding frequency resources. If the TPMI indicates a precoding matrix, the indicated precoding matrix is applied to all frequency resources allocated. In order to cope with the situation that the spatial multiplexing is not possible due to channel variation, the eNodeB can instantaneously schedule downlink transmission using the transmit diversity even though the UE has been configured to be in the spatial multiplexing mode. Use of the transmit diversity is indicated by TPMI.

Table 6: Precoding matrix $\mathbf{C}_k (k = 1, 2, 3, 4)$ for the open-loop spatial multiplexing.

	Number of layers M		
	2	3	4
$\mathbf{C}_1$	$\mathbf{W}_{12}^{\{12\}}/\sqrt{2}$	$\mathbf{W}_{12}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{12}^{\{1234\}}/2$
$\mathbf{C}_2$	$\mathbf{W}_{13}^{\{13\}}/\sqrt{2}$	$\mathbf{W}_{13}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{13}^{\{1324\}}/2$
$\mathbf{C}_3$	$\mathbf{W}_{14}^{\{13\}}/\sqrt{2}$	$\mathbf{W}_{14}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{14}^{\{3214\}}/2$
$\mathbf{C}_4$	$\mathbf{W}_{15}^{\{12\}}/\sqrt{2}$	$\mathbf{W}_{15}^{\{123\}}/\sqrt{3}$	$\mathbf{W}_{15}^{\{1234\}}/2$

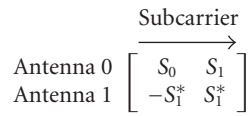


Figure 3: SFBC with two transmit antennas on downlink.

The open-loop spatial multiplexing may be operated when reliable PMI feedback is not available at the eNodeB, for example, when the UE speed is not slow enough or when the feedback overhead on uplink is too high. The open-loop spatial multiplexing with M layers and N transmit antennas ($N \geq M$) is illustrated in Figure 2. The feedback consists of the RI and the CQI in open-loop spatial multiplexing. In contrast to the closed-loop spatial multiplexing, the eNodeB only determines the transmission rank and a fixed set of precoding matrices are applied cyclically across all the scheduled subcarriers in the frequency domain.

The precoding for the open-loop spatial multiplexing mode is defined by

$$\mathbf{y}(i) = \mathbf{W}(i)\mathbf{D}(i)\mathbf{U}\mathbf{x}(i), \tag{2}$$

where $\mathbf{y}(i) = [y_0(i), \ldots, y_{N-1}(i)]^T$, $y_n(i)$ denotes the ith complex symbol transmitted on the nth antenna, the precoding matrix $\mathbf{W}(i)$ is of size $N \times M$, $\mathbf{x}(i) = [x_0(i), \ldots, x_{M-1}(i)]^T$, $x_m(i)$ denotes the ith modulation symbol transmitted on the mth layer, and the DFT precoding matrix $\mathbf{U}$ of size $M \times M$ and the matrix $\mathbf{D}(i)$ of size $M \times M$ supporting the large delay cyclic delay diversity (CDD) are defined in Tables 4 and 5, respectively. When multiple layers are transmitted,

Table 7: Feedback of wideband CQI and subband PMI via PUSCH.

Field	Bit width			
	2 antennas		4 antennas	
	Rank = 1	Rank = 2	Rank = 1	Rank > 1
Rank indicator	1		2	
Wideband CQI for codeword 0	4	4	4	4
Wideband CQI for codeword 1	0	4	0	4
Precoding matrix indication	$2K$	K	$4K$	$4K$

$$\begin{array}{c} \\ \text{Antenna 0} \\ \text{Antenna 1} \\ \text{Antenna 2} \\ \text{Antenna 3} \end{array} \begin{array}{c} \xrightarrow{\text{Subcarrier}} \\ \begin{bmatrix} S_0 & S_1 & 0 & 0 \\ 0 & 0 & S_2 & S_3 \\ -S_1^* & S_0^* & 0 & 0 \\ 0 & 0 & -S_3^* & S_2^* \end{bmatrix} \end{array}$$

Figure 4: SFBC + FSTD with four transmit antennas on downlink.

$\mathbf{D}(i)\mathbf{U}$ effectively makes the modulation symbols of a single codeword are mapped onto different layers for each i in a cyclic manner with period M as the index i increases so that a codeword can experience all the transmitted layers. For two transmit antennas, $\mathbf{W}(i)$ is given by

$$\mathbf{W}(i) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3}$$

For four transmit antennas, to have additional robustness against the spatial channel characteristics, a set of precoding matrices are assigned to $\mathbf{W}(i)$ cyclically as the index i increases according to the followings:

$$\mathbf{W}(i) = \mathbf{C}_k \tag{4}$$

where the index k is given by $k = (\lfloor i/M \rfloor \bmod 4) + 1$ and $\mathbf{C}_k(k = 1, 2, 3, 4)$ for different rank values is given in Table 6. It is noted that in the open-loop spatial multiplexing mode, the transmit diversity scheme is applied when the transmission rank is set to one.

3. Transmit Diversity in LTE

For LTE downlink, the transmit diversity schemes can be applied to all the physical channels such as PDSCH, Physical Broadcast Channel (PBCH), Physical Control Format Indicator Channel (PCFICH), Physical Downlink Control Channel (PDCCH), and Physical Hybrid ARQ Indicator Channel (PHICH) while the other MIMO schemes are only applicable to PDSCH.

A UE can recognize the number of transmit antennas at eNodeB among $\{1, 2, 4\}$ by blindly decoding PBCH, since there is no explicit signaling for it. Note that no transmit diversity scheme applied to the primary and secondary synchronization signals [11] is specified in LTE. Once the number of transmit antennas at eNodeB is detected, a specific transmit diversity scheme applicable to the other physical downlink channels is determined.

Transmit diversity schemes defined for LTE downlink are illustrated in Figures 3, 4, and 5. The space-frequency block code (SFBC) as shown in Figure 3 is used if the eNodeB has two transmit antennas. For the eNodeB with four transmit antennas, a combination of the SFBC and the frequency-switched transmit diversity (FSTD) as shown in Figure 4 is used to provide robustness against the correlation between channels from different transmit antennas and for easier UE receiver implementation. The transmit diversity scheme shown in Figure 4 can be used for all downlink channels other than PHICH. The transmit diversity scheme used for PHICH is shown in Figure 5. In this scheme, four different ACK/NAK bits are multiplexed using orthogonal codes with spreading factor of four over a group of four subcarriers and the resulting group is repeated three times in the frequency domain to achieve frequency diversity gain. To maintain the orthogonality between different codes in each repetition of four subcarriers, antenna switching is not applied within each repetition. Instead, the set of antennas changes across different repetitions as shown in Figure 5. When there are multiple PHICHs transmitted, using type 1 or type 2 alternatively for different PHICHs would be helpful to keep uniform power distribution over eNodeB transmit antennas.

A UE is configured for a transmission scheme such as transmit diversity, SU-MIMO, MU-MIMO, closed-loop rank-1 precoding, and dedicated beamforming when the eNodeB employs multiple transmit antennas. When the eNodeB tries to change the transmission scheme, it may not be possible to transmit the required control message using the configured transmission scheme, for example, SU-MIMO, for the indication of transmission scheme change, since the channel condition is not favorable any longer to the configured transmission scheme. For reliable change of the transmission scheme regardless which transmission scheme is configured for the UE, the transmit diversity can always be used for delivering the required control message to the UE. Hence, the UE shall always try to receive such control message sent using the transmit diversity.

For uplink, the transmit antenna selection diversity for the UE with two transmit antennas is specified. In case of the closed-loop transmit antenna selection, the eNodeB selects the antenna to be used for uplink transmission and communicate this selection to the UE using the downlink control message. For the open-loop transmit antenna selection, the UE autonomously selects the transmit antenna to be used for transmission without eNodeB's intervention. Note that

Type1:

Subcarrier →

Antenna 0	S_0	S_1	S_2	S_3
Antenna 1	0	0	0	0
Antenna 2	$-S_1^*$	S_0^*	$-S_3^*$	S_2^*
Antenna 3	0	0	0	0

1st repetition

Antenna 0	0	0	0	0
Antenna 1	S_0	S_1	S_2	S_3
Antenna 2	0	0	0	0
Antenna 3	$-S_1^*$	S_0^*	$-S_3^*$	S_2^*

2nd repetition

Antenna 0	S_0	S_1	S_2	S_3
Antenna 1	0	0	0	0
Antenna 2	$-S_1^*$	S_0^*	$-S_3^*$	S_2^*
Antenna 3	0	0	0	0

3rd repetition

(a)

Type2:

Subcarrier →

Antenna 0	0	0	0	0
Antenna 1	S_0	S_1	S_2	S_3
Antenna 2	0	0	0	0
Antenna 3	$-S_1^*$	S_0^*	$-S_3^*$	S_2^*

1st repetition

Antenna 0	S_0	S_1	S_2	S_3
Antenna 1	0	0	0	0
Antenna 2	$-S_1^*$	S_0^*	$-S_3^*$	S_2^*
Antenna 3	0	0	0	0

2nd repetition

Antenna 0	0	0	0	0
Antenna 1	S_0	S_1	S_2	S_3
Antenna 2	0	0	0	0
Antenna 3	$-S_1^*$	S_0^*	$-S_3^*$	S_2^*

3rd repetition

(b)

Figure 5: Modified SFBC + FSTD for PHICH with four transmit antennas on downlink.

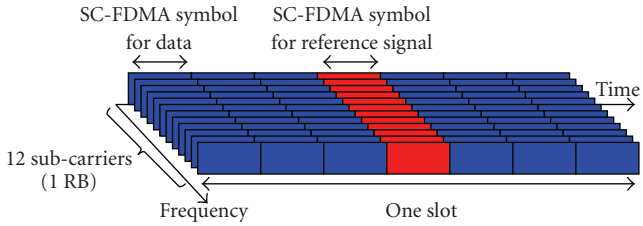


Figure 6: Multiplexing of data and reference signal on uplink.

SFBC-type transmit diversity scheme is not employed for the LTE uplink to avoid additional cost required to implement two power amplifiers at the UE.

4. Closed-loop Rank-1 Precoding in LTE

In the closed-loop rank-1 precoding mode, the eNodeB operates the closed-loop SU-MIMO scheme based on the cell-specific common reference signal with the limitation of selecting a rank-1 precoding matrix for transmission to a UE among the ones defined in Table 1 for two transmit antennas and Table 2 for four transmit antennas to improve data coverage without relying on the UE-specific reference signal. Since the transmission rank is fixed to one in this mode, the related downlink control signaling overhead is smaller than the case of operating the closed-loop SU-MIMO scheme, of which control signaling allows full freedom of selecting the transmission rank among all possible rank values.

5. MU-MIMO in LTE

MU-MIMO scheme is supported in both the uplink and downlink of the LTE standard. In the uplink, the eNodeB can always schedule more than one UEs to transmit in the same time-frequency resource, which forms a MU-MIMO transmission configuration. However, in order for the eNodeB to be able to correctly differentiate and demodulate these UEs' signals, eNodeB needs to assign orthogonal reference signals for these UEs scheduled for the MU-MIMO transmission. Figure 6 shows the uplink slot structure where the reference signal is transmitted using the fourth symbol and the data is transmitted using the others. For a given slot and subframe in each cell, a Zadoff-Chu sequence [11] is defined as the base sequence for uplink reference signals. The cyclically shifted versions of a given Zadoff-Chu sequence form an orthogonal set of sequences. Each UE scheduled for MU-MIMO transmission is assigned a distinctive cyclic shift value, and the UE combines this cyclic shift value with the knowledge of the base Zadoff-Chu sequence to form a reference signal sequence that is orthogonal to other UEs' reference signal sequences. It is noted that the cyclic shift value is always contained in the control signaling, which the UE has to receive for data transmission on uplink, regardless whether the MU-MIMO is operated or not.

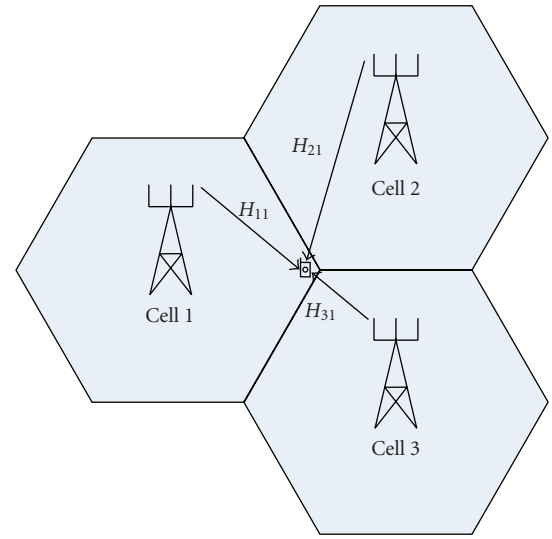


Figure 7: Coordinated multipoint transmission in the downlink.

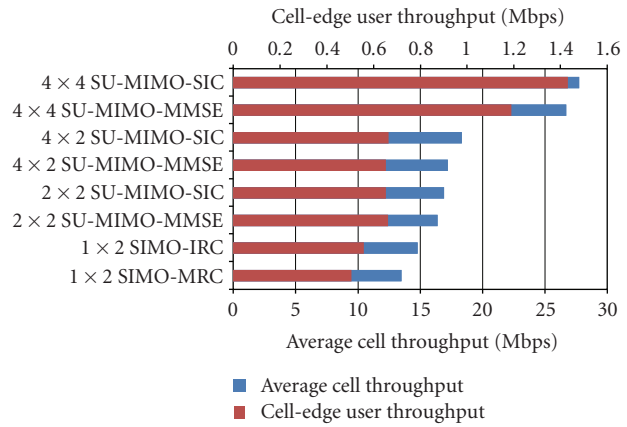


Figure 8: LTE downlink SU-MIMO performance.

In the downlink, if a UE is configured to be in the MU-MIMO transmission mode, only rank-1 transmission can be scheduled to the UE. The eNodeB can schedule multiple UEs, which are configured to be in the MU-MIMO transmission mode, in the same time-frequency resource using different rank-1 precoding matrices from Table 1 for two transmit antennas and Table 2 for four transmit antennas. Note that the UE receives only the information about its own precoding matrix. The scheduled UE then decodes the information data utilizing the common reference signal together with the precoding information obtained from the control signaling. The UE generates the PMI/CQI feedback without any knowledge about other simultaneously scheduled UEs. Hence, there could be mismatch between the UE's CQI report and the actual CQI experienced due to lack of knowledge of interference caused by another UEs scheduled simultaneously. In LTE, for support of receiving higher-order modulation signal such as 16QAM and 64QAM without causing too much complexity in the UE, the transmit power level for each UE is configured in a long-term manner. The per-UE preconfigured power level is hard to maintain in MU-MIMO transmission mode, since eNodeB power amplifier has to support multiple UEs scheduled on the same time-frequency resource. A 1-bit signaling is therefore introduced to indicate whether there is 3 dB power reduction with respect to the per-UE configured power level if a UE is configured in the MU-MIMO transmission mode.

6. Dedicated Downlink Beamforming in LTE

Dedicated beamforming is supported for improving data coverage when the UE supports data demodulation using the UE-specific reference signal. The eNodeB generates a beam using the array of antenna elements (e.g. array of 8 antenna elements), and then applies the same precoding to both the data payload and the UE-specific reference signal with this beam. It is noted that the UE-specific reference signal is transmitted in a way such that its time-frequency location does not overlap with the cell-specific reference signal.

7. Uplink Feedback in LTE

The uplink feedback for support of downlink data transmission consists of the RI, the PMI, and the CQI. The RI indicates the number of layers, which can be accommodated by the current spatial channel experienced at the UE. It was observed in the LTE evaluation that the frequency-selective RI reporting did not provide significant performance benefit, and therefore only one wideband RI is reported for the whole bandwidth. On the contrary, the reporting of PMI and CQI can be either wideband or frequency-selective. The PMI is calculated conditioned on the associated RI, and the CQI is calculated conditioned on the associated RI and PMI. For RI = 1, only one CQI is reported for each reporting unit in frequency, which could be either wideband or subband in the case of frequency-selective report. For RI > 1, CQI for each codeword is reported for the closed-loop spatial multiplexing as different codewords experience different layers, while only one CQI is reported for the open-loop spatial multiplexing as a codeword experiences all layers. The PMI indicates the preferred precoding candidate for the corresponding frequency unit, for example, a particular subband or the whole frequency bandwidth, and is selected from the possible precoding candidates of Table 1 for the case of two transmit antennas and Table 2 for the case of four transmit antennas according to the RI. The CQI indicates the combination of the maximum information data size and the modulation scheme among QPSK, 16QAM, and 64QAM, which can provide block error rate not exceeding 0.1 assuming that the reported rank and the reported precoding matrix are applied in the time-frequency resource. With this definition of CQI, PMI, and RI, the UE can report the maximum data size that it can receive and demodulate, taking into account its receiver ability.

In case of the frequency-selective PMI/CQI reporting, the UE reports a PMI/CQI for each subband. For the non-frequency-selective wideband PMI/CQI reporting, the UE reports a single wideband PMI/CQI corresponding to the whole bandwidth. In the frequency-selective reporting mode, the subband CQI is reported as a differential value with respect to the wideband CQI in order to reduce the signaling overhead. When the frequency-selective CQI reporting is configured, the subband CQIs as well as the wideband CQI is reported, and the wideband CQI serves as the baseline for recovering the downlink channel condition in the whole band.

The frequency-selective report naturally results in large signaling overhead. In the cases where the uplink overhead is a limiting factor, the eNodeB can also configure non-frequency-selective CQI/PMI reports. To cope with various channel conditions and various antenna configurations while keeping the signaling overhead at appropriate level, various feedback modes are specified concerning on the frequency selectivity of the CQI and the PMI reports [13].

The physical channels that can be used for the uplink feedback signaling are Physical Uplink Control Channel (PUCCH) and Physical Uplink Shared Channel (PUSCH). Feedback via PUSCH is used to accommodate large amount

Table 8: Feedback of subband differential CQI and wideband PMI via PUSCH.

Field	Bit width			
	2 antennas		4 antennas	
	Rank = 1	Rank = 2	Rank = 1	Rank > 1
Rank indicator	1		2	
Wideband CQI for codeword 0	4	4	4	4
Subband differential CQI for codeword 0	$2K$	$2K$	$2K$	$2K$
Wideband CQI for codeword 1	0	4	0	4
Subband differential CQI for codeword 1	0	$2K$	0	$2K$
Precoding matrix indication	2	1	4	4

Table 9: Feedback of wideband CQI/PMI via PUCCH.

Field	Bit width			
	2 antennas		4 antennas	
	Rank = 1	Rank = 2	Rank = 1	Rank > 1
Wideband CQI	4	4	4	4
Spatial differential CQI	0	3	0	3
Precoding matrix indication	2	1	4	4

of feedback information in a single reporting instance, since PUSCH is designed for carrying large-size information data packets. For example, the reporting on PUSCH can include the RI, the wideband CQI per codeword, and the PMI for each subband as shown in Table 7, where K is the number of subbands. Note that the rank indicator is separately encoded from the other fields, using one bit for the case of two antennas and two bits for the case of four antennas. The bit width of the other fields is determined according to the rank indicator. Another example is the reporting of the RI, the wideband PMI, the wideband CQI per codeword, and the subband differential CQI for each subband per codeword as shown in Table 8. It is noted that simultaneous reporting of the subband differential CQI and the subband PMI is not supported due to excessive signaling overhead.

PUCCH, on the other hand, is designed for transmission of small amount of control signaling information. Hence, for PUCCH reporting modes, separate time instances are used for reporting the RI, the wideband CQI/PMI, and the subband CQI of the subband selected by the UE. Examples of the wideband CQI/PMI feedback and the subband CQI feedback via PUCCH are shown in Tables 9 and 10, respectively. The spatial differential CQI represents the difference between the CQIs of two codewords and is defined for both the wideband CQI and the subband CQI. In case of the subband CQI feedback, each reporting instance corresponds to a group of subbands, from which the UE selects the subband with best CQI. Due to these characteristics, the eNodeB may configure periodic reporting of limited amount of feedback information on PUCCH, while triggering the UE to report large amount of detailed feedback information on PUSCH if accurate channel information is needed for transmission of large amount of data in the downlink.

8. MIMO Schemes in LTE-Advanced

In order to support downlink peak spectrum efficiency of 30 bps/Hz and uplink peak spectrum efficiency of 15 bps/Hz according to LTE-Advanced requirement [14], the spatial multiplexing with antenna configuration of 8×8 for downlink transmission and 4×4 for uplink transmission is being investigated. Here $N \times N$ denotes a configuration of N transmit antennas and N receive antennas.

In addition to meeting the peak spectrum efficiency, further improvement of the average cell throughput as well as the cell edge performance is also an important aspect of the LTE-advanced study. Coordinated multipoint transmission/reception is a candidate technology where antennas of multiple cell sites are utilized in a way such that the transmit/receive antennas of the serving cell as well as the neighboring cells can contribute in improving quality of the received signal at the UE/eNodeB, as well as in reducing the cochannel interferences from neighboring cells. The application of coordinated multipoint transmission is illustrated in Figure 7 for a downlink transmission scenario. An example scheme is to form a beam to the scheduled UE by using the transmit antennas of the cells 1, 2, and 3, where each cell transmits the same data to the scheduled UE and the UE-specific reference signal is used for support of demodulation at the UE. For the cells 1, 2, and 3 to jointly form the transmit signal matching to the composite channel experienced by the UE, it may be necessary to provide feedback representing the downlink spatial channel of each cell without any preassumption on operation at the eNodeB transmitter and the UE receiver. The feedback for explicit representation of the spatial channel of each cell may naturally require much larger overhead than the feedback defined in LTE such as CQI, PMI, and RI.

Table 10: Feedback of subband CQI via PUCCH.

Field	Bit width			
	2 antennas		4 antennas	
	Rank = 1	Rank = 2	Rank = 1	Rank > 1
Subband CQI	4	4	4	4
Spatial differential CQI	0	3	0	3
Subband label	1 or 2	1 or 2	1 or 2	1 or 2

Other areas for further investigation in LTE-Advanced are for example as follows:

(i) further enhancement of downlink MU-MIMO to improve the system throughout beyond what was achieved in LTE;

(ii) introduction of uplink transmit diversity utilizing up to four transmit antennas;

(iii) extension of downlink transmit diversity to eight transmit antennas;

9. Performance Evaluation

Figure 8 shows the LTE downlink SU-MIMO performance in terms of average cell throughput and cell-edge user throughput obtained by system level simulation, where $N \times L$ represents the configuration of N eNodeB transmit antennas and L UE receive antennas. Simulation parameters and assumptions follow the guideline provided in [2]. Linear antenna arrays with antenna spacing 10λ are assumed for eNodeB transmit antennas, where λ denotes the wavelength of the carrier frequency. The results in Figure 8 are for Case 1 representing interference-limited small urban macrocell environments, where the carrier frequency is 2 GHz, the inter-site distance is 500 m, the bandwidth is 10 MHz, and the UE speed is 3 km/h. The 2-tier cell layout with 57 cells in total was considered and 10 users were dropped per cell.

For single input multiple output (SIMO), two linear receiver methods are used: maximal ratio combining (MRC) and interference rejection combining (IRC). For SU-MIMO, MMSE and MMSE with successive interference cancellation (MMSE-SIC) receivers are used. Since IRC and SIC are designed to reduce or cancel the interference, they outperform MRC and MMSE, respectively, which are designed considering the desired signals only.

The simulation results for the antenna configurations of 2×2 and 4×4 show how much the spatial multiplexing scheme introduced in LTE improves the system level performance. Comparing MIMO-SIC with SIMO-IRC, we observe 14% and 87% gain from 2×2 and 4×4, respectively, in terms of average cell throughput.

In the antenna configuration of 4×2, up to 2 layers can be constructed using the precoding schemes designed for 4 transmit antennas. Comparing the performance between the antenna configurations of 4×2 and 2×2, we can observe the precoding gain with more transmit antennas. The results in Figure 8 show that 4×2 configuration provides 8.2% gain over 2×2 configuration in terms of average cell throughput.

It is also observed from the results in Figure 8 that the MIMO technologies introduced in LTE successfully improve the cell-edge user throughputs.

10. Conclusion

In this paper, we introduced MIMO features of LTE, which are downlink SU-MIMO, transmit diversity, closed-loop rank-1 precoding, MU-MIMO, dedicated beamforming, and further described technical backgrounds for specifying those technologies. Uplink feedback mechanisms for support of downlink MIMO technologies were also described to provide better understanding about LTE system operation. In addition, the MIMO schemes being studied for LTE-Advanced were briefly described.

References

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1995.

[2] 3GPP, TR 25.814, "Physical layer aspects for evolved Universal Terrestrial Radio Access (Release 7)".

[3] G. Foschini and M. J. Gans, "On the limits of wireless communications in a fading environment when usingmultiple atnennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–355, 1998.

[4] D. Agrawal, V. Tarokh, A. Naguib, and N. Seshadri, "Space-time coded OFDM for high data-rate wireless communication over wideband channels," in *Proceedings of the 48th IEEE Vehicular Technology Conference (VTC '98)*, vol. 3, pp. 2232–2236, Ottawa, Canada, 1998.

[5] H. El Gamal and M. O. Damen, "An algebraic number theoretic framework for space-time coding," in *Proceedings of IEEE International Symposium on Information Theory*, p. 132, Lausanne, Switzerland, June-July 2002.

[6] B. Hassibi and B. M. Hochwald, "High-rate codes that are linear in space and time," *IEEE Transactions on Information Theory*, vol. 48, no. 7, pp. 1804–1824, 2002.

[7] N. Al-Dhahir and A. H. H. Sayed, "The finite-length multi-input multi-output MMSE-DFE," *IEEE Transactions on Signal Processing*, vol. 48, no. 10, pp. 2921–2936, 2000.

[8] H. Bölcskei, D. Gesbert, and A. J. Paulraj, "On the capacity of OFDM-based spatial multiplexing systems," *IEEE Transactions on Communications*, vol. 50, no. 2, pp. 225–234, 2002.

[9] http://www.3gpp.org.

[10] 3GPP, TS 36.201, "Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Physical Layer-General Description (Release 8)".

[11] 3GPP, TS 36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 8)".

[12] 3GPP, TS 36.212, "Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding (Release 8)".
[13] 3GPP, TS 36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 8)".
[14] 3GPP, TR 36.913, "Requirements for Further Advancements for E-UTRA (LTE-Advanced) (Release 8)".
[15] 3GPP, TR 36.814, "Further Advancements for E-UTRA; Physical Layer Aspects".
[16] 3GPP, TS 36.306, "Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio access capabilities (Release 8)".
[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md, USA, 1996.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 197682, 10 pages
doi:10.1155/2009/197682

Research Article

Optimal Multiuser MIMO Linear Precoding with LMMSE Receiver

Fang Shu, Wu Gang, and Li Shao-Qian

National Key Lab of Communication, University of Electronic Science and Technology of China, Chengdu 611731, China

Correspondence should be addressed to Fang Shu, fangshu@uestc.edu.cn

Received 21 January 2009; Revised 11 May 2009; Accepted 19 June 2009

Recommended by Cornelius van Rensburg

The adoption of multiple antennas both at the transmitter and the receiver will explore additional spatial resources to provide substantial gain in system throughput with the spatial division multiple access (SDMA) technique. Optimal multiuser MIMO linear precoding is considered as a key issue in the area of multiuser MIMO research. The challenge in such multiuser system is designing the precoding vector to maximize the system capacity. An optimal multiuser MIMO linear precoding scheme with LMMSE detection based on particle swarm optimization is proposed in this paper. The proposed scheme aims to maximize the system capacity of multiuser MIMO system with linear precoding and linear detection. This paper explores a simplified function to solve the optimal problem. With the adoption of particle swarm optimization algorithm, the optimal linear precoding vector could be easily searched according to the simplified function. The proposed scheme provides significant performance improvement comparing to the multiuser MIMO linear precoding scheme based on channel block diagonalization method.

1. Introduction

In recent years, with the increasing demand of transmitting high data rates, the (Multiple-Input Multiple-Output) MIMO technique, a potential method to achieve high capacity has attracted enormous interest [1, 2]. When multiple antennas are equipped at both base stations (BSs) and mobile stations (MSs), the space dimension can be exploited for scheduling multi-user transmission besides time and frequency dimension. Therefore, the traditional MIMO technique focused on point-to-point single-user MIMO (SU-MIMO) has been extended to the point-to-multipoint multi-user MIMO (MU-MIMO) technique [3, 4]. It has been shown that time division multiple access (TDMA) systems can not achieve sum rate capacity of MU-MIMO system of broadcast channel (BC) [5] while MU-MIMO with spatial division multiple access (SDMA) could, where one BS communicates with several MSs within the same time slot and the same frequency band [6, 7]. MU-MIMO based on SDMA improves system capacity taking advantage of multi-user diversity and precanceling of multi-user interference at the transmitter.

Traditional MIMO technique focuses on point-to-point transmission as the STBC technique based on space-time coding and the VBLAST technique based on spatial multiplexing. The former one can efficiently combat channel fading while its spectral efficiency is low [8, 9]. The latter one could transmit parallel data streams, but its performance will be degraded under spatial correlated channel [10, 11]. When the MU-MIMO technique is adopted, both the multi-user diversity gain to improve the BER performance and the spatial multiplexing gain to increase the system capacity will be obtained. Since the receive antennas are distributed among several users, the spatial correlation will effect less on multi-user MIMO system. Besides, because the multi-user MIMO technique utilizes precoding at the transmit side to precanceling the cochannel interference (CCI), so the complexity of the receiver can be significantly simplified. However multi-user CCI becomes one of the main obstacles to improve MU-MIMO performance. The challenge is that the receiving antennas that are associated with different users are typically unable to coordinate with each other. By mitigating or ideally completely eliminating CCI, the BS exploits the channel state information (CSI) available at the

transmitter to cancel the CCI at the transmitter. It is essential to have CSI at the BS since it allows joint processing of all users' signals which results in a significant performance improvement and increased data rates.

The sum capacity in a multiuser MIMO broadcast channel is defined as the maximum aggregation of all the users' data rates. For Gaussion MIMO broadcast channels (BCs), it was proven in [12] that Dirty Paper Coding (DPC) can achieve the capacity region. The optimal precoding of multi-user MIMO is based on dirty paper coding (DPC) theory with the nonlinear precoding method. DPC theory proves that when a transmitter has advance knowledge of the interference, it could design a code to compensate for it. It is developed by Costa which can eliminate the interference by iterative precoding at the transmitter and achieve the broadcast MIMO channel capacity [13, 14]. The famous Tomlinson-Harashima precoding (THP) is the non-linear precoding based on DPC theory. It is first developed by Tomlinson [15] and Miyakawa and Harashima [16] independently and then has become the Tomlinson-Harashima precoding (THP) [17–20] to combat the multi-user cochannel interference (CCI) with non-linear precoding. Although THP performs well in a multi-user MIMO scenario, deploying it in real-time systems is difficult because of its high complexity of the precoding at the transmitter. Many suboptimal MU-MIMO linear precoding techniques have emerged recently, such as the channel inversion method [21] and the block diagonalization (BD) method [22–24]. Channel inversion method [25] employs some traditional MIMO detection criterions, such as the Zero Forcing (ZF) and Minimum Mean Squared Error (MMSE), precoding at the transmitter to suppress the CCI Channel inversion method based on ZF can suppress CCI completely; however it may lead to noise amplification since the precoding vectors are not normalized. Channel inversion method based on MMSE compromises the noise and the CCI, and outperforms ZF algorithm, but it still cannot obtain good performance. BD method decomposes a multi-user MIMO channel into multiple single user MIMO channels in parallel to completely cancel the CCI by making use of the null space. With BD, each users precoding matrix lies in the null space of all other users channels, and the CCI could be completely canceled. The generated null space vectors are normalized vectors, which could avoid the noise amplification problem efficiently. So BD method performs much better than channel inversion method. However, since BD method just aims to cancel the CCI and suppress the noise, its precoding gain is not optimized.

It is obvious that the CCI, the noise, and the precoding gain are the factors affecting on the performance of the preprocessing MU-MIMO. The above linear precoding methods just take one factor into account without entirely consideration. A rate maximization linear precoding method is proposed in [26]. This method aims to maximize the sum rate of the MU-MIMO system with linear preprocessing. However, the optimized function in [26] is too complex to compute. In this paper, we solve the optimal linear precoding with linear MMSE receiver problem in a more simplified way.

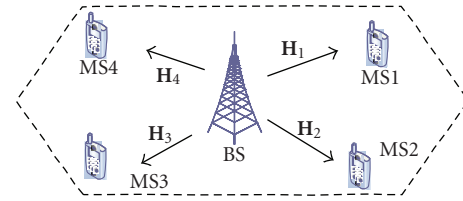


Figure 1: The configuration of MU-MIMO system

An optimal MU-MIMO linear precoding scheme with linear MMSE receiver based on particle swarm optimization (PSO) is proposed in this paper. PSO algorithm has been used in many complex optimization tasks, especially in solving the optimization of continuous space [27, 28]. In this paper, PSO is firstly introduced into MIMO research to solve some optimization issues. The adoption of PSO to MIMO system provides a new method to solve the MIMO processing problem. In this paper, we first analyze the optimal linear precoding vector with linear MMSE receiver and establish a simplified function to measure the optimal linear precoding problem. Then, we employ the novel PSO algorithm to search the optimal linear precoding vector according to the simplified function. The proposed scheme obtains significant MU-MIMO system capacity and outperforms the channel block diagonalization method.

This paper is organized into seven parts. The system model of MU-MIMO is given in Section 2. Then the analysis of optimal linear precoding with linear MMSE receiver is given in Section 3. The particle swarm optimization algorithm is given in Section 4. In Section 5, the proposed optimal linear precoding MU-MIMO scheme with LMMSE detection based on particle swarm optimization is introduced. In Section 6, the simulation results and comparisons are given. Conclusions are drawn in the last section. The channel block diagonalization algorithm is given in the appendix.

2. System Model of MU-MIMO

The MU-MIMO system could transmit data streams of multiple users of the same cellular at the same time and the same frequency resources as Figure 1 shows.

We consider an MU-MIMO system with one BS and K MS, where the BS is equipped with M antennas and each MS with N antennas, as shown in Figure 2. The point-to multipoint MU-MIMO system is employed in downlink transmission.

Because MU-MIMO aims to transmit data streams of multiple-users at the same time and frequency resources, we discuss the algorithm at single-carrier, for each subcarrier of the multicarrier system, and it is processed as same as the single-carrier case. Since OFDM technique deals the frequency selective fading as flat fading, we model the channel as the flat fading MIMO channel:

$$\mathbf{H}_k = \begin{bmatrix} h_{1,1} & \cdots & h_{1,M} \\ \vdots & \ddots & \vdots \\ h_{N,1} & \cdots & h_{N,M} \end{bmatrix} \quad (1)$$

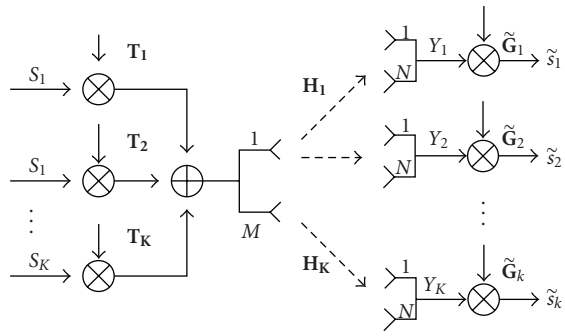


Figure 2: The block diagram of MU-MIMO system

where $\mathbf{H}_k$ is the MIMO channel matrix of user k. $h_{i,j}$ indicates the channel impulse response coupling the jth transmit antenna to the ith receive antenna. Its amplitude obeys independent and identically Rayleigh-distribution.

Data streams of K ($K \le M$) users are precoded by their precoding vectors $\mathbf{T}_k$ ($k = 1 \cdots K$) before transmission. $\mathbf{T}_k$ is the $M \times 1$ normalized precoding vector for user k with $\mathbf{T}_k^H \mathbf{T}_k = 1$. The received signal at the kth user is

$$\begin{aligned} \mathbf{y}_k &= \mathbf{H}_k \mathbf{T}_k \sqrt{p_k} s_k + \mathbf{H}_k \sum_{i=1, i\neq k}^{K} \mathbf{T}_i \sqrt{p_i} s_i + \mathbf{n}_k \\ \sum_{k=1}^{K} p_k &= p_0 \end{aligned} \tag{2}$$

where y_k is the received signal of user k. The elements of additive noise $\mathbf{n}_k$ obey distribution $CN(0, N_0)$ that are spatially and temporarily white. p_k is the transmit signal power of the kth data stream, and p_0 is the total transmit power.

The received signal at the kth user can also be expressed as

$$\begin{aligned} \mathbf{y}_k &= \mathbf{H}_k \mathbf{W} \mathbf{s} + \mathbf{n}_k \\ \mathbf{W} &= [\mathbf{T}_1 \quad \mathbf{T}_2 \quad \cdots \quad \mathbf{T}_K] \\ \mathbf{s} &= \left[\sqrt{p_1} s_1 \quad \sqrt{p_2} s_2 \quad \cdots \quad \sqrt{p_K} s_K\right]^T \end{aligned} \tag{3}$$

where $\mathbf{s}$ is the transmitted symbol vector with K data streams, $\mathbf{W}$ is the precoding matrix with K precoding vectors, and $[\cdot]^T$ denotes the matrix transposition:

$$\widetilde{\mathbf{H}}_k = \mathbf{H}_k \mathbf{W} \tag{4}$$

The channel matrix $\widetilde{\mathbf{H}}_k$ can be assumed as the virtual channel matrix of user k after precoding. At the receiver, a linear receiver $\widetilde{\mathbf{G}}_k$ is exploited to detect the transmit signal for the user k. The detected signal of the kth user is

$$\widehat{s}_k = \widetilde{\mathbf{G}}_k \mathbf{y}_k. \tag{5}$$

The linear receiver $\widetilde{\mathbf{G}}_k$ can be designed by ZF or MMSE criteria, and linear MMSE will obtain better performance. In order to simplify the analysis, the power allocation is assumed as equal $\beta = p_k/N_0 = p_0/KN_0$, and linear MMSE MIMO detection is used in this paper as

$$\widetilde{\mathbf{G}}_k = \widetilde{\mathbf{h}}_k^H \left(\widetilde{\mathbf{H}}_k \widetilde{\mathbf{H}}_k^H + \beta \mathbf{I}_N\right)^{-1}, \tag{6}$$

where $(\cdot)^{-1}$ indicates the inverse of the matrix, $(\cdot)^H$ denotes the matrix conjugation transposition, and $\mathbf{I}_N$ is the $N \times N$ identity matrix:

$$\widetilde{\mathbf{h}}_k = \mathbf{H}_k \mathbf{T}_k = \left[\widetilde{\mathbf{H}}_k\right]_k = [\mathbf{H}_k \mathbf{W}]_k, \tag{7}$$

where $[\cdot]_k$ denotes the kth column of the matrix. Then the detected SINR for the user k with the linear detection is

$$\begin{aligned} \mathrm{SINR}_k &= \frac{\beta \left|\widetilde{\mathbf{G}}_k \mathbf{H}_k \mathbf{T}_k\right|^2}{\sum_{i=1, i\neq k}^{K} \beta \left|\widetilde{\mathbf{G}}_k \mathbf{H}_k \mathbf{T}_i\right|^2 + \left\|\widetilde{\mathbf{G}}_k\right\|_2^2} \\ &= \frac{\beta \left|\widetilde{\mathbf{G}}_k \mathbf{H}_k \mathbf{T}_k\right|^2 / \left\|\widetilde{\mathbf{G}}_k\right\|_2^2}{\sum_{i=1, i\neq k}^{K} \beta \left|\widetilde{\mathbf{G}}_k \mathbf{H}_k \mathbf{T}_i\right|^2 / \left\|\widetilde{\mathbf{G}}_k\right\|_2^2 + 1}, \end{aligned} \tag{8}$$

where $\|\cdot\|_2$ denotes the matrix two-norm.

Because the nonnormalized precoding vector will amplify the noise at the receiver, the precoding vectors $\mathbf{T}_k$ are assumed to be normalized as follow:

$$\|\mathbf{T}_k\|^2 = 1 \tag{9}$$

for $k = 1, \ldots, K$.

3. Optimal Multiuser MIMO Linear Precoding

We assume that the MIMO channel matrices $\mathbf{H}_k$ ($k = 1, \ldots, K$) are available at the BS. It can be achieved either by channel reciprocity characteristics in time-division-duplex (TDD) mode or by feedback in frequency-division-duplex (FDD) mode. And the channel matrix $\mathbf{H}_k$ is known at the receiver k through channel estimation. We just discuss the equal power allocation case in this paper. The optimal power allocation is achieved through water-filling according to the SINR of each user.

The MIMO channel of user k can be decomposed by the singular value decomposition (SVD) as

$$\mathbf{H}_k = \mathbf{U}_k \mathbf{\Sigma}_k \mathbf{V}_k^H. \tag{10}$$

If we apply $\mathbf{T}_k = [\mathbf{V}_k]_1$ to precode for user k, it obtains the maximal precoding gain as follow.

Lemma 1. *One has*

$$\|\mathbf{H}_k \mathbf{T}_k\|_2 = \left\|\mathbf{H}_k [\mathbf{V}_k]_1\right\|_2 = \lambda_k^{\max}, \tag{11}$$

where $[\mathbf{V}_k]_1$ *denotes the first column of* $\mathbf{V}_k$, *and* $\lambda_k^{\max}$ *denotes the maximal singular value of* $\mathbf{H}_k$.

Proof. One has

$$\begin{aligned}\|\mathbf{H}_k\mathbf{T}_k\|_2 &= \sqrt{(\mathbf{H}_k\mathbf{T}_k)^H(\mathbf{H}_k\mathbf{T}_k)}\\ &= \sqrt{(\mathbf{H}_k[\mathbf{V}_k]_1)^H(\mathbf{H}_k[\mathbf{V}_k]_1)}\\ &= \sqrt{\left(\lambda_k^{\max}[\mathbf{U}_k]_1\right)^H\left(\lambda_k^{\max}[\mathbf{U}_k]_1\right)}.\end{aligned} \tag{12}$$

So

$$\|\mathbf{H}_k\mathbf{T}_k\|_2 = \lambda_k^{\max}, \tag{13}$$

where $[\mathbf{U}_k]_1$ denotes the first column of unitary matrix $\mathbf{U}_k$. □

Thus, precoding with the singular vector corresponding to the maximal singular value is an initial thought to obtain good performance. However, if the singular vector is directly used at the transmitter as the precoding vector, the CCI will be large, and the performance will be degraded severely. Only for the special case that the MIMO channel among all these users are orthogonal that the CCI will be zero if we directly use the singular vector of each user as its precoding vector. But in realistic case, the transmit users' channels are always nonorthogonal, and so the singular vector could not be utilized directly. We have drawn some analysis as follow.

(1)*Ideal channel case.* The ideal channel case is that the MIMO channels of transmitting users' are orthogonal. There is

$$\left|([\mathbf{V}_k]_1)^H[\mathbf{V}_i]_1\right| = 0 \quad (i \neq k) \tag{14}$$

If we apply $\mathbf{T}_k = [\mathbf{V}_k]_1$ to precode for user k, the maximal precoding gain will be obtained as (13) shows, and the CCI will turn to zero as follow.

Lemma 2. *One has*

$$\frac{\sum_{i=1,i\neq k}^{K}\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2}{\left\|\widetilde{\mathbf{G}}_k\right\|_2^2} = 0. \tag{15}$$

Proof. One has

$$\sum_{i=1,i\neq k}^{K}\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2 = \sum_{i=1,i\neq k}^{K}\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k[\mathbf{V}_i]_1\right|^2, \tag{16}$$

$$\widetilde{\mathbf{G}}_k = (\mathbf{H}_k\mathbf{T}_k)^H\left(\widetilde{\mathbf{H}}_k\widetilde{\mathbf{H}}_k^H + \frac{1}{\beta}\mathbf{I}_M\right)^{-1}, \tag{17}$$

$$\mathbf{H}_k\mathbf{T}_k = \mathbf{H}_k[\mathbf{V}_k]_1 = \lambda_k^{\max}[\mathbf{U}_k]_1, \tag{18}$$

$$\widetilde{\mathbf{G}}_k = \left(\lambda_k^{\max}[\mathbf{U}_k]_1\right)^H\left((\mathbf{H}_k\mathbf{W})(\mathbf{H}_k\mathbf{W})^H + \frac{1}{\beta}\mathbf{I}_M\right)^{-1}, \tag{19}$$

$$(\mathbf{H}_k\mathbf{W})(\mathbf{H}_k\mathbf{W})^H = \mathbf{U}_k\Sigma_k\mathbf{V}_k^H\mathbf{W}\mathbf{W}^H\mathbf{V}_k\Sigma_k\mathbf{U}_k^H. \tag{20}$$

Because we assume that $|[\mathbf{V}_k]_1^H[\mathbf{V}_i]_1| = 0$ $(i \neq k)$, and $[\mathbf{V}_k]_1$ is the unit vector with $|[\mathbf{V}_k]_1^H[\mathbf{V}_k]_1| = 1$, so $\mathbf{W} = [[\mathbf{V}_1]_1, [\mathbf{V}_2]_1, \ldots, [\mathbf{V}_K]_1]$ is an unitary matrix with

$$\begin{aligned}&\mathbf{W}\mathbf{W}^H = \mathbf{I}_M,\\ &\widetilde{\mathbf{G}}_k = (\mathbf{H}_k\mathbf{T}_k)^H\left(\widetilde{\mathbf{H}}_k\widetilde{\mathbf{H}}_k^H + \frac{1}{\beta}\mathbf{I}_M\right)^{-1}\\ &= \left(\lambda_k^{\max}[\mathbf{U}_k]_1\right)^H\left(\mathbf{U}_k\Sigma_k^2\mathbf{U}_k^H + \frac{1}{\beta}\mathbf{I}_M\right)^{-1}\\ &= \lambda_k^{\max}[\mathbf{U}_k]_1^H\left(\mathbf{U}_k\left(\Sigma_k^2 + \frac{1}{\beta}\mathbf{I}_M\right)\mathbf{U}_k^H\right)^{-1}.\end{aligned} \tag{21}$$

Since $([\mathbf{U}_k]_1)^H(\mathbf{U}_k^H)^{-1} = [1, 0, \ldots, 0]$, so

$$\begin{aligned}\widetilde{\mathbf{G}}_k &= \lambda_k^{\max}[\mathbf{U}_k]_1^H\left(\mathbf{U}_k^H\right)^{-1}\left(\Sigma_k^2 + \frac{1}{\beta}\mathbf{I}_M\right)^{-1}\mathbf{U}_k^{-1}\\ &= \frac{\lambda_k^{\max}}{(\lambda_k^{\max})^2 + 1/\beta}\left[\mathbf{U}_k^H\right]^1\\ &= \frac{\lambda_k^{\max}}{(\lambda_k^{\max})^2 + 1/\beta}[\mathbf{U}_k]_1^H,\end{aligned} \tag{22}$$

where $[\mathbf{U}_k^H]^1$ denotes the first row of $\mathbf{U}_k^H$.

Also

$$\begin{aligned}&\sum_{i=1,i\neq k}^{K}\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2\\ &= \sum_{i=1,i\neq k}^{K}\left|\frac{\lambda_k^{\max}}{(\lambda_k^{\max})^2 + 1/\beta}[\mathbf{U}_k]_1^H\mathbf{H}_k\mathbf{T}_i\right|^2\\ &= \sum_{i=1,i\neq k}^{K}\left|\frac{\lambda_k^{\max}}{(\lambda_k^{\max})^2 + 1/\beta}[\mathbf{U}_k]_1^H\mathbf{U}_k\Sigma_k\mathbf{V}_k^H[\mathbf{V}_i]_1\right|^2.\end{aligned} \tag{23}$$

Since $|([\mathbf{V}_k]_1)^H[\mathbf{V}_i]_1| = 0$, so $\mathbf{V}_k^H[\mathbf{V}_i]_1 = [0, a_1, \ldots, a_{K-1}]^T$. Combining $([\mathbf{U}_k]_1)^H\mathbf{U}_k = [1, 0, \ldots, 0]$, there is

$$\sum_{i=1,i\neq k}^{K}\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2 = \sum_{i=1,i\neq k}^{K}\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k[\mathbf{V}_i]_1\right|^2 = 0. \tag{24}$$

so

$$\frac{\sum_{i=1,i\neq k}^{K}\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2}{\left\|\widetilde{\mathbf{G}}_k\right\|_2^2} = 0. \tag{25}$$

□

After linear MMSE detection at the receiver, user k obtains the maximal SINR as follows.

Lemma 3. *One has*

$$SINR_k = \frac{\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k\right|^2}{\left\|\widetilde{\mathbf{G}}_k\right\|_2^2} = \beta\left(\lambda_k^{\max}\right)^2. \tag{26}$$

Proof. One has

$$\mathrm{SINR}_k = \frac{\beta(\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k)^H\left(\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k\right)}{\left|\widetilde{\mathbf{G}}_k\widetilde{\mathbf{G}}_k^H\right|}. \tag{27}$$

According to (13) and (22)

$$\begin{aligned}
&\left(\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k\right)^H\left(\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k\right)\\
&\quad=\left(\frac{\lambda_k^{\max}}{\left(\lambda_k^{\max}\right)^2+1/\beta}[\mathbf{U}_k]_1^H\lambda_k^{\max}[\mathbf{U}_k]_1\right)^H\\
&\qquad\times\left(\frac{\lambda_k^{\max}}{\left(\lambda_k^{\max}\right)^2+1/\beta}[\mathbf{U}_k]_1^H\lambda_k^{\max}[\mathbf{U}_k]_1\right)\\
&\quad=\left(\frac{\left(\lambda_k^{\max}\right)^2}{\left(\lambda_k^{\max}\right)^2+1/\beta}\right)^2\\
&\left|\widetilde{\mathbf{G}}_k\widetilde{\mathbf{G}}_k^H\right|=\left(\frac{\lambda_k^{\max}}{\left(\lambda_k^{\max}\right)^2+1/\beta}\right)^2\\
&\mathrm{SINR}_k=\beta\left(\lambda_k^{\max}\right)^2.
\end{aligned} \tag{28}$$

□

(2)*Ill channel case* The ill channel case is that all these transmitting users' channels are highly correlated. There is

$$\left|([\mathbf{V}_k]_1)^H[\mathbf{V}_i]_1\right|=1\quad(i\neq k). \tag{29}$$

If we still apply $\mathbf{T}_k=[\mathbf{V}_k]_1$ to precode for user k, the multiuser CCI will be very large, and the system performance will be degraded severely. The SINR after MMSE detection with equal power allocation for user k is as follows.

Lemma 4. *One has*

$$SINR_k=\frac{\beta\left(\lambda_k^{\max}\right)^2}{\sum_{i=1,i\neq k}^{K}\beta(\lambda_k^{\max})^2+1}. \tag{30}$$

Proof. Since we have proven that when $\mathbf{T}_k=[\mathbf{V}_k]_1$ to precode for user k, then $\beta|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k|^2/\|\widetilde{\mathbf{G}}_k\|_2^2=\beta(\lambda_k^{\max})^2$, so

$$\begin{aligned}
\mathrm{SINR}_k&=\frac{\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k\right|^2/\left\|\widetilde{\mathbf{G}}_k\right\|_2^2}{\sum_{i=1,i\neq k}^{K}\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2/\left\|\widetilde{\mathbf{G}}_k\right\|_2^2+1}\\
&=\frac{\beta\left(\lambda_k^{\max}\right)^2}{\sum_{i=1,i\neq k}^{K}\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2/\left\|\widetilde{\mathbf{G}}_k\right\|_2^2+1}.
\end{aligned} \tag{31}$$

According to (19)

$$\begin{aligned}
\widetilde{\mathbf{G}}_k&=\left(\lambda_k^{\max}[\mathbf{U}_k]_1\right)^H\\
&\quad\times\left(\mathbf{U}_k\Sigma_k\mathbf{V}_k^H\mathbf{W}\mathbf{W}^H\mathbf{V}_k\Sigma_k\mathbf{U}_k^H+\frac{1}{\beta}\mathbf{I}_M\right)^{-1}.
\end{aligned} \tag{32}$$

Since we assume that $|[\mathbf{V}_k]_1^H[\mathbf{V}_i]_1|=1$ $(i\neq k)$, so $\mathbf{W}=[[\mathbf{V}_1]_1,[\mathbf{V}_2]_1,\ldots,[\mathbf{V}_K]_1]$ is

$$\mathbf{W}=\begin{bmatrix}1&\cdots&1\\ \vdots&\ddots&\vdots\\ 1&\cdots&1\end{bmatrix}. \tag{33}$$

Let the diagonal matrix $\overline{\Sigma}_k=\Sigma_k\mathbf{V}_k^H\mathbf{W}\mathbf{W}^H\mathbf{V}_k\Sigma_k$, and so there is

$$\widetilde{\mathbf{G}}_k=\left(\lambda_k^{\max}[\mathbf{U}_k]_1\right)^H\left(\mathbf{U}_k\left(\overline{\Sigma}_k+\frac{1}{\beta}\mathbf{I}_M\right)\mathbf{U}_k^H\right)^{-1}. \tag{34}$$

Since $([\mathbf{U}_k]_1)^H(\mathbf{U}_k^H)^{-1}=[1,0,\ldots,0]$, so

$$\begin{aligned}
\widetilde{\mathbf{G}}_k&=\frac{\lambda_k^{\max}}{\overline{\lambda}_k+1/\beta}\left[\mathbf{U}_k^H\right]^1\\
&=\frac{\lambda_k^{\max}}{\overline{\lambda}_k+1/\beta}[\mathbf{U}_k]_1^H,
\end{aligned} \tag{35}$$

where $\overline{\lambda}_k$ indicates the first diagonal element of the diagonal matrix $\overline{\Sigma}_k$. So there is

$$\begin{aligned}
\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i&=\frac{\lambda_k^{\max}}{\overline{\lambda}_k+1/\beta}[\mathbf{U}_k]_1^H\mathbf{H}_k\mathbf{T}_i\\
&=\frac{\lambda_k^{\max}}{\overline{\lambda}_k+1/\beta}[\mathbf{U}_k]_1^H\mathbf{U}_k\Sigma_k\mathbf{V}_k^H[\mathbf{V}_i]_1.
\end{aligned} \tag{36}$$

Since $|([\mathbf{V}_k]_1)^H[\mathbf{V}_i]_1|=1$, so $\mathbf{V}_k^H[\mathbf{V}_i]_1=[1,a_1,\ldots,a_{K-1}]^T$. Combining $([\mathbf{U}_k]_1)^H\mathbf{U}_k=[1,0,\ldots,0]$, there is

$$\begin{aligned}
\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_i\right|^2&=\left(\frac{\left(\lambda_k^{\max}\right)^2}{\overline{\lambda}_k+1/\beta}\right)^2,\\
\left\|\widetilde{\mathbf{G}}_k\right\|_2^2&=\left|\widetilde{\mathbf{G}}_k\widetilde{\mathbf{G}}_k^H\right|=\left(\frac{\lambda_k^{\max}}{\overline{\lambda}_k+1/\beta}\right)^2,\\
\frac{\sum_{i=1,i\neq k}^{K}\beta\left|\widetilde{\mathbf{G}}_k\mathbf{H}_k\mathbf{T}_k\right|^2}{\left\|\widetilde{\mathbf{G}}_k\right\|_2^2}&=\sum_{i=1,i\neq k}^{K}\beta\left(\lambda_k^{\max}\right)^2.
\end{aligned} \tag{37}$$

So the SINR for user k is

$$SINR_k=\frac{\beta\left(\lambda_k^{\max}\right)^2}{\sum_{i=1,i\neq k}^{K}\beta\left(\lambda_k^{\max}\right)^2+1} \tag{38}$$

□

(3) *Practical case.* The practical case is that the transmitting users' channels are neither orthogonal nor ill. There is

$$\begin{aligned}
&\left|([\mathbf{V}_k]_1)^H[\mathbf{V}_i]_1\right|\neq 0\quad(i\neq k)\\
&\left|([\mathbf{V}_k]_1)^H[\mathbf{V}_i]_1\right|\neq 1\quad(i\neq k).
\end{aligned} \tag{39}$$

The practical case is usually in realistic environment. If we apply $\mathbf{T}_k$ $(\mathbf{T}_k\neq[\mathbf{V}_k]_1)$ to precode for user k, then

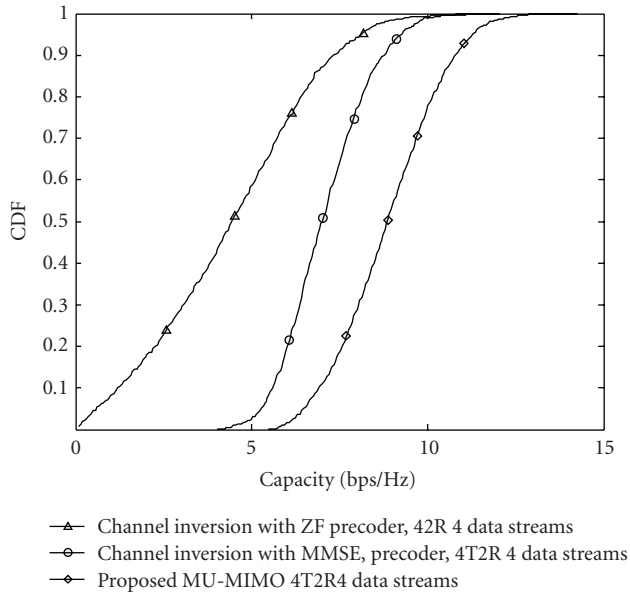


Figure 3: The system capacity CDF comparison of the two schemes.

$\xi_k = |\mathbf{T}_k^H[\mathbf{V}_k]_1|$can be the parameter to measure the precoding gain, and $\rho_i = |\mathbf{T}_k^H[\mathbf{V}_i]_1|$ can be the parameter to measure the CCI. The SINR for user k according to the above analysis can be approximated denoted as

$$\begin{aligned}\mathrm{SINR}_k &\approx \frac{\beta\left(\lambda_k^{\max}\xi_k\right)^2}{\sum_{i\neq k,i=1}^{K}\beta\left(\lambda_k^{\max}\rho_i\right)^2+1}\\ &= \frac{\left|\lambda_k^{\max}(\mathbf{T}_k^H[\mathbf{V}_k]_1)\right|^2}{\sum_{i\neq k,i=1}^{K}\left|\lambda_k^{\max}(\mathbf{T}_k^H[\mathbf{V}_i]_1)\right|^2+1/\beta}.\end{aligned} \tag{40}$$

The system capacity is related to SINR of the transmit users $k, (k = 1,\ldots,K)$. So in order to obtain the system capacity, we should obtain the SINR_k. Thus, when the optimal precoding vector is obtained by the PSO algorithm, the system capacity could be calculated by (41).

The system capacity of the MU-MIMO system can be indicated as

$$C_{\mathrm{MU}} = \sum_{k=1}^{K}\log_2(1+\mathrm{SINR}_k). \tag{41}$$

We aim to maximize the system capacity of the MU-MIMO system in this paper. The optimal MU-MIMO linear precoding vector for the MU-MIMO system is the vector that can maximize the SINR at each receiver as

$$\overline{\mathbf{T}}_k = \arg\max_{\mathbf{T}_k\in\mathbf{U}} \sum_{k=1}^{K}\log_2\left(1+\frac{\left|\lambda_k^{\max}(\mathbf{T}_k^H[\mathbf{V}_k]_1)\right|^2}{\sum_{i\neq k,i=1}^{K}\left|\lambda_k^{\max}(\mathbf{T}_k^H[\mathbf{V}_i]_1)\right|^2+1/\beta}\right) \tag{42}$$

where $\mathbf{U}$ denotes the unitary vector that $\mathbf{U}^H\mathbf{U} = \mathbf{I}$. From the above equation, it is clear that if we want to maximize the system capacity of MU-MIMO, then the SINR of each user should be maximized. The SINR of user k is associated with three parameters as the singular vector correspond to the maximal singular value of all users and the noise.

4. The Particle Swarm Optimization Algorithm

Particle swarm optimization algorithm was originally proposed by Kennedy and Eberhart[27] in 1995. It searches the optimal problem solution through cooperation and competition among the individuals of population.

Imagine a swarm of bees in a field. Their goal is to find in the field the location with the highest density of flowers. Without any prior knowledge of the field, the bees begin in random locations with random velocities looking for flowers. Each bee can remember the location that is found the most flowers and somehow knows the locations where the other bees found an abundance of flowers. Torn between returning to the location where it had personally found the most flowers, or exploring the location reported by others to have the most flowers, the ambivalent bee accelerates in both directions to fly somewhere between the two points. There is a function or method to evaluate the goodness of a position as the fitness function. Along the way, a bee might find a place with a higher concentration of flowers than it had found previously. Constantly, they are checking the concentration of flowers and hoping to find out the absolute highest concentration of flowers.

Suppose that the size of swarm and the dimension of search space are C and D ,respectively. Each individual in the swarm is referred to as a particle. The location and velocity of particle i $(i = 1,\ldots,C)$ are represented as the vector $\mathbf{x}_i = [x_{i1}, x_{i1},\ldots,x_{iD}]^T$ and $\mathbf{v}_i = [v_{i1}, v_{i2},\ldots,v_{iD}]^T$. Each bee remembers the location where it personally encountered the most flowers which is denoted as $\mathbf{P}_i = [p_{i1}, p_{i2},\ldots,p_{iD}]^T$, which is the flight experience of the particle itself. The highest concentration of flowers discovered by the entire swarm is denoted as $\mathbf{P}_g = [p_{g1}, p_{g2},\ldots,p_{gD}]^T$, which is the flight experience of all particles. Each particle is searching for the best location according to $\mathbf{P}_i$ and $\mathbf{P}_g$. The particle i updates its location and velocity according to the following two formulas [27]:

$$\begin{aligned}v_{id}^{t+1} &= wv_{id}^t + c_1\varphi_1\left(p_{id}^t - x_{id}^t\right) + c_2\varphi_2\left(p_{gd}^t - x_{id}^t\right)\\ x_{id}^{t+1} &= x_{id}^t + v_{id}^{t+1}\end{aligned} \tag{43}$$

where t is the current iteration number; v_{id}^t and $x_{id}^t + 1$ denote the velocity and location of the particle i in the dth dimensional direction. p_{id}^t is the individual best location of particle i in the dth dimensional direction, p_{gd}^t is the population best location in the dth dimensional direction. φ_1 and φ_2 are the random numbers between 0 and 1, c_1 and c_2 are the learning factors, and w is the inertia factor. Learning factors determine the relative "pull" of $\mathbf{P}_i^t$ and $\mathbf{P}_g^t$ that usually content $c_1 = c_2 = 2$. Inertia factor determines to what extent the particle remains along its original course unaffected by the pull of $\mathbf{P}_g^t$ and $\mathbf{P}_i^t$ that is usually between 0 and 1. After this process is carried out for each particle in the swarm, the

process is repeated until reaching the maximal iteration or the termination criteria are met.

5. The Optimal Linear Precoding Multiuser MIMO with LMMSE Detection Based on Particle Swarm Optimization

With the adoption of PSO algorithm and the simplified function (40), the optimal linear precoding vector $\overline{\mathbf{T}}_k$ ($k = 1, \ldots, K$) could be easily searched.

The proposed optimal MU-MIMO linear precoding scheme based on PSO algorithm will search the optimal precoding vector for each user following 6 steps.

(1) The BS obtains $\lambda_k^{\max}$, $[\mathbf{V}_k]_1$ and β of each user.

(2) The BS employs the PSO algorithm to search the optimal linear precoding vector for each user. For user k, the PSO algorithm sets the maximal iteration number I and a group of M dimensional particles with the initial velocity $\mathbf{v}_{i,k}^1 = [v_{i1,k}^1, v_{i2,k}^1, \ldots, v_{iM,k}^1]^T$ and the initial location $\mathbf{x}_{i,k}^1 = [x_{i1,k}^1, x_{i2,k}^1, \ldots, x_{iM,k}^1]^T$ for particle i ($i = 1, \ldots, C$). In order to accelerate the searching process, the initial location $\mathbf{x}_{i,k}^1$ could be initialized as $[\mathbf{V}_k]_1$, while the initial velocity $\mathbf{v}_{i,k}^1$ could be produced randomly. The real and imaginary parts of the initial velocity obey a normal distribution with mean zero and standard deviation one.

(3) The BS begins to search with the initial location $\mathbf{x}_{i,k}^1$ and velocity $\mathbf{v}_{i,k}^1$. The goodness of the location is measured by the following equation:

$$f_{i,k}^t = \frac{\left|\lambda_k^{\max}[(\mathbf{x}_{i,k}^t)^H[\mathbf{V}_k]_1]\right|^2}{\sum_{j=1,j\neq k}^K \left|\lambda_k^{\max}[(\mathbf{x}_{i,k}^t)^H[\mathbf{V}_j]_1]\right|^2 + 1/\beta}, \quad (44)$$

where the fitness function $f_{i,k}^t$ indicates the obtained SINR for user k precoded by $\mathbf{x}_{i,k}^t$. The PSO algorithm finds $\mathbf{P}_{i,k}^t$ and $\mathbf{P}_{g,k}^t$ that are individual best location and population best location measured by (44) for the next iteration. $\mathbf{P}_{i,k}^t$ denotes the individual best location which means the best location of particle i at the tth iteration of the kth user. $\mathbf{P}_{g,k}^t$ denotes the population best location which means the best location of all particles at the tth iteration of the kth user.

(4) For the tth iteration, the algorithm finds a $\mathbf{P}_{i,k}^t$ and a $\mathbf{P}_{g,k}^t$. The location and velocity for each particle will be updated according to (43) for the next iteration. In order to obtain the normalized optimal precoding vector to suppress the noise, the location $\mathbf{x}_{i,k}^t$ should be normalized in each iteration.

(5) When reaching the maximal iteration number I, the algorithm stops, and $\mathbf{P}_{g,k}^I$ is the obtained optimal precoding vector for user k.

(6) For an MU-MIMO system with K users, the scheme will search the precoding vectors according to the above criteria for each user.

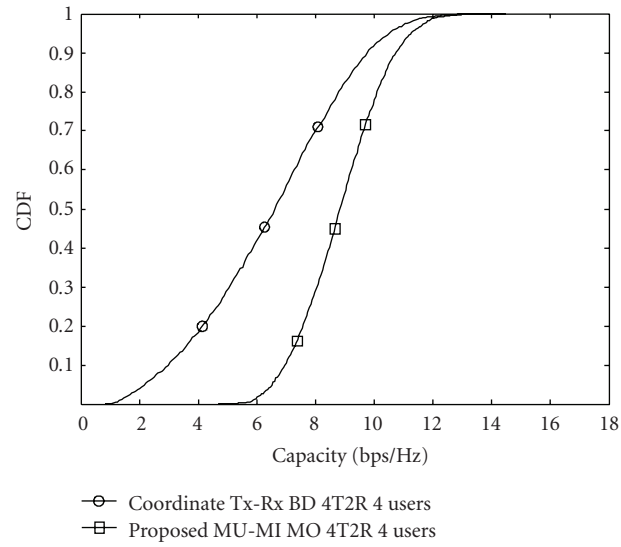


Figure 4: The system capacity CDF comparison of the two schemes.

6. Simulation Results

We simulated the proposed MU-MIMO scheme, the BD algorithm in [22] (Coordinate Tx-Rx BD), and the channel inversion algorithm in [25] in this paper to compare their performance under the same simulation environment.

Figure 3 is the system capacity comparison of the cumulative distribution function (CDF) of the channel inversion algorithm with ZF precoder and MMSE precoder and the proposed MU-MIMO algorithm when $M = 4, N = 2$, $p_0/N_0 = 5\,dB$ with equation power allocation and MMSE detection at the receiver. For channel inversion method, the BS transmits 4 date streams and 2 users simultaneously with 2 date stream for each user. For the proposed MU-MIMO, the BS transmit 4 data streams and 4 users simultaneously with 1 data stream for each user.

Figure 4 is the system capacity comparison of the CDF of the coordinated Tx-Rx BD algorithm and the proposed MU-MIMO algorithm when $M = 4$, $K = 4$, $N = 2$, $p_0/N_0 = 5\,dB$ with equation power allocation and MMSE detection at the receiver.

Figure 5 is the system capacity comparison of the CDF of the coordinated Tx-Rx BD algorithm and the proposed MU-MIMO algorithm when $M = 4$, $K = 4$, $N = 4$, $p_0/N_0 = 5\,dB$ with equation power allocation and MMSE detection at the receiver. Both the simulation results of the proposed MU-MIMO scheme with PSO algorithm from Figure 3 to Figure 5 are based on the PSO parameters with the particle number $C = 20$ and the iteration number $I = 20$. It could be seen that the proposed MU-MIMO scheme can effectively increase the system capacity compared to the BD algorithm and channel inversion algorithm.

Figure 6 is the average BER performance of the proposed MU-MIMO scheme and the coordinated Tx-Rx BD algorithm with $M = 4, K = 4, N = 4$. Figure 7 is the average BER performance of the proposed MU-MIMO scheme and the coordinated Tx-Rx BD algorithm with $M = 4, K = 4, N = 2$.

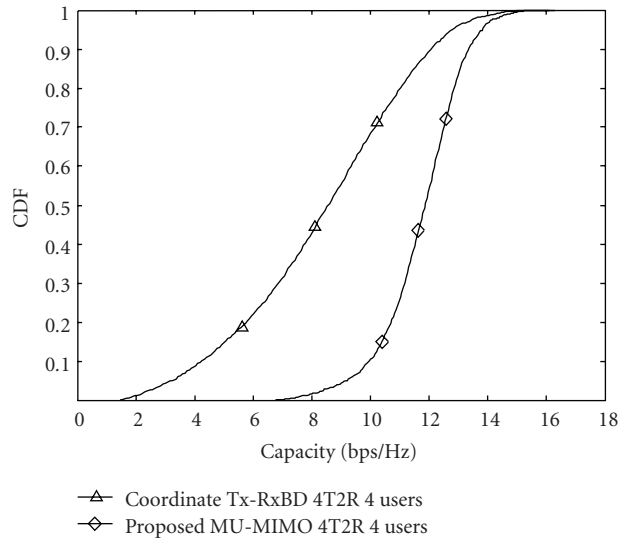


Figure 5: The system capacity CDF comparison of the two schemes.

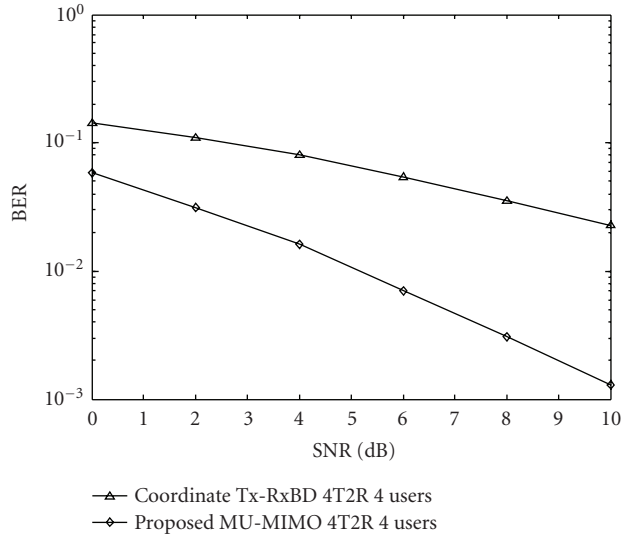


Figure 6: The BER comparison of the two schemes.

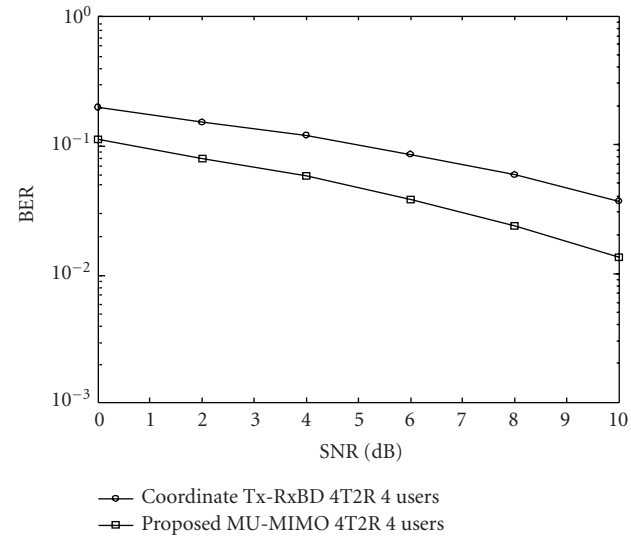


Figure 7: The BER comparison of the two schemes.

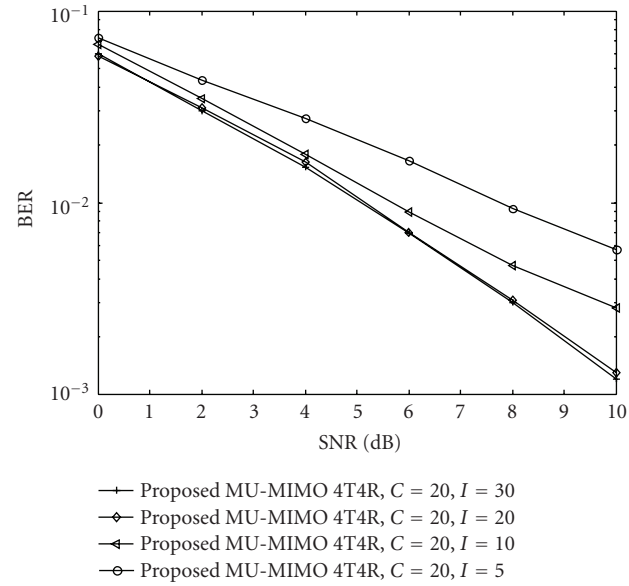


Figure 8: The BER comparison of the two schemes with different C and I.

Both the schemes adopt equal power allocation, MMSE detection, QPSK, and no channel coding. The proposed MU-MIMO scheme, with PSO algorithm from Figures 6 and 7 are based on the PSO parameters with the particle number $C = 20$ and the iteration number $I = 20$.

From the simulation results, it is clear that the proposed MU-MIMO linear precoding with LMMSE detection based on particle swarm optimization scheme outperforms the BD algorithm and the channel inversion algorithm. The reason lies in that the BD algorithm just aims to utilize the normalized precoding vector to cancel the CCI and suppress the noise. The channel inversion algorithm also aims to suppress CCI and noise. So the users' transmit signal covariance matrices of these schemes are generally not optimal that are caused by the inferior precoding gain. The proposed MU-MIMO optimal linear precoding scheme aims to find the optimal precoding vector to maximize each users' SINR at each receiver to improve the total system capacity.

Figure 8 shows the BER performance of the proposed MU-MIMO scheme with the same particle size and different iteration size when $M = 4$, $K = 4$, $N = 4$. It adopts equal power allocation, MMSE detection, QPSK, and no channel coding. The particle number C is 20, and the iteration number scales from 5 to 30. We could see that when the iteration number is small, the proposed scheme could not obtain the best performance. With the increase of the iteration number, more performance as well as the computational complexity will increase too. However, when the iteration number is

larger than 20 for this case, the algorithm could not obtain more performance gain. Generally, for different case, the best iteration number is different. The iteration number is related to the transmit antenna number M at the BS and the transmit user number K ($K \leq M$). With the increasing of M or K, the iteration number should increase in order to obtain the best performance.

7. Conclusion

This paper solves the optimal linear precoding problem with LMMSE detection for MU-MIMO system in downlink transmission. A simplified optimal function is proposed and proved to maximize the system capacity. With the adoption of the particle swarm optimization algorithm, the optimal linear precoding vector with LMMSE detection for each user could be searched. The proposed scheme can obtain significant system capacity improvement compared to the multi-user MIMO scheme based on channel block digonolization under the same simulation environment.

Appendix

Coordinated Tx-Rx BD Algorithm

Coordinated Tx-Rx BD algorithm is the improved BD algorithm. It could solve the antenna constraint problem in traditional BD algorithm and extends the BD algorithm to arbitrary antenna configuration. For a coordinated Tx-Rx BD algorithm with M transmit antennas at the BS, N receive antennas at the MS, and K users to be transmitted simultaneously, the algorithm follows 6 steps.

(1) For $j = 1, \ldots, K$, compute the SVD

$$\mathbf{H}_j = \mathbf{U}_j \mathbf{\Sigma}_j \mathbf{V}_j^H. \tag{A.1}$$

(2) Determine m_j, which is the number of subchannels for each user. In order to compare the two schemes fairly, $m_j = 1$ for each user.

(3) For $j = 1, \ldots, K$, let $\mathbf{A}_j$ be the first m_j columns of $\mathbf{U}_j$. Calculate $\overline{\mathbf{H}}_j = \mathbf{A}_j^H \mathbf{H}_j$.

$$\widetilde{\mathbf{H}}_j = \left[\overline{\mathbf{H}}_1^T \quad \cdots \quad \overline{\mathbf{H}}_{j-1}^T \quad \overline{\mathbf{H}}_{j+1}^T \quad \cdots \quad \overline{\mathbf{H}}_K^T\right]^T \tag{A.2}$$

(4) For $j = 1, \ldots, K$, compute $\widetilde{\mathbf{V}}_j^{(0)}$, the right null space of $\widetilde{\mathbf{H}}_j$ as

$$\widetilde{\mathbf{H}}_j = \widetilde{\mathbf{U}}_j \widetilde{\mathbf{\Sigma}}_j \left[\widetilde{\mathbf{V}}_j^{(1)} \quad \widetilde{\mathbf{V}}_j^{(0)}\right]^H, \tag{A.3}$$

where $\widetilde{\mathbf{V}}_j^{(1)}$ holds the first L_j right singular vectors, $\widetilde{\mathbf{V}}_j^{(0)}$ holds the last $N - L_j$ right singular vectors and $L_j = \operatorname{rank}(\mathbf{H}_j)$.

(5) Compute the SVD

$$\mathbf{H}_j \widetilde{\mathbf{V}}_j^{(0)} = \mathbf{U}_j \begin{bmatrix} \mathbf{\Sigma}_j & 0 \\ 0 & 0 \end{bmatrix} \left[\mathbf{V}_j^{(1)} \quad \mathbf{V}_j^{(0)}\right]^H. \tag{A.4}$$

(6) The precoding matrix $\mathbf{W}$ for the transmit users with average power allocation is

$$\mathbf{W} = \left[\widetilde{\mathbf{V}}_1^{(0)} \mathbf{V}_1^{(1)} \quad \widetilde{\mathbf{V}}_2^{(0)} \mathbf{V}_2^{(1)} \quad \cdots \quad \widetilde{\mathbf{V}}_K^{(0)} \mathbf{V}_K^{(1)}\right]. \tag{A.5}$$

Acknowledgments

The project was supported by the National Natural Science Foundation of China (60702073) and the Key Laboratory of Universal Wireless Communications Lab. (Beijing University of Posts and Telecommunications), Ministry of Education, China.

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[2] A. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, "Capacity limits of MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 5, pp. 684–702, 2003.

[3] Q. H. Spencer, C. B. Peel, A. L. Swindlehurst, and M. Haardt, "An introduction to the multi-user MIMO downlink," *IEEE Communications Magazine*, vol. 42, no. 10, pp. 60–67, 2004.

[4] C. Windpassinger, R. F. H. Fischer, T. Vencel, and J. B. Huber, "Precoding in multiantenna and multiuser communications," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1305–1316, 2004.

[5] N. Jindal and A. Goldsmith, "Dirty-paper coding versus TDMA for MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1783–1794, 2005.

[6] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[7] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2658–2668, 2003.

[8] E. A. Jorswieck and A. Sezgin, "Impact of spatial correlation on the performance of orthogonal space-time block codes," *IEEE Communications Letters*, vol. 8, no. 1, pp. 21–23, 2004.

[9] S. N. Diggavi, "On achievable performance of spatial diversity fading channels," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 308–325, 2001.

[10] P. Tarasak, H. Minn, and V. K. Bhargava, "Linear prediction receiver for differential space-time block codes with spatial correlation," *IEEE Communications Letters*, vol. 7, no. 11, pp. 543–545, 2003.

[11] H. Bölcskei, D. Gesbert, and A. J. Paulraj, "On the capacity of OFDM-based spatial multiplexing systems," *IEEE Transactions on Communications*, vol. 50, no. 2, pp. 225–234, 2002.

[12] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[13] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian MIMO broadcast channel," in *Proceedings of IEEE International Symposium on Information Theory*, p. 174, June 2004.

[14] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna Gaussian broadcast channel," *IEEE Transactions on Information Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[15] M. Tomlinson, "New automatic equaliser employing modulo arithmetic," *Electronics Letters*, vol. 7, no. 5-6, pp. 138–139, 1971.

[16] H. Miyakawa and H. Harashima, "Matched-transmission techniuqe for channels with intersymbol interference," *IEEE Transactions on Communications*, vol. 20, pp. 774–779, 1972.

[17] V. Stankovic and M. Haardt, "Successive optimization Tomlinson-Harashima precoding (SO THP) for multi-user MIMO systems," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '05)*, vol. 3, pp. 1117–1120, March 2005.

[18] W. Miao, L. Xiao, Y. Li, S. Zhou, and J. Wang, "Joint stream-wise THP transceiver disign for the multiuser MIMO downlink," in *Proceedings of IEEE Wireless Communications & Networking Conference (WCNC '08)*, pp. 330–334, 2008.

[19] K. Takeda, H. Tomeba, and F. Adachi, "BER performance of joint THP/pre-DFE," in *Proceedings of IEEE Vehicular Technology Conference (VTC '08)*, pp. 1016–1020, 2008.

[20] P. L. Athanasios, "Tomlinson-Harashima precoding with partial channel knowlegde," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 5–9, 2005.

[21] T. Haustein, C. von Helmolt, E. Jorswieck, V. Jungnickel, and V. Pohl, "Performance of MIMO systems with channel inversion," in *Proceedings of the 55th IEEE Vehicular Technology Conference (VTC '02)*, vol. 1, pp. 35–39, 2002.

[22] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[23] L.-U. Choi and R. D. Murch, "A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach," *IEEE Transactions on Wireless Communications*, vol. 3, no. 1, pp. 20–24, 2004.

[24] K.-K. Wong, R. D. Murch, and K. B. Letaief, "A joint-channel diagonalization for multiuser MIMO antenna systems," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 773–786, 2003.

[25] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—part I: channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, 2005.

[26] M. Stojnic, H. Vikalo, and B. Hassibi, "Rate maximization in multi-antenna broadcast channels with linear preprocessing," in *Proceedings of IEEE Global Telecommunications Conference (GLOBECOM '04)*, vol. 6, pp. 3957–3961, November 2004.

[27] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, vol. 5, pp. 1942–1948, 1995.

[28] J. Robinson and Y. Rahmat-Samii, "Particle swarm optimization in electromagnetics," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 2, pp. 397–407, 2004.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 894726, 11 pages
doi:10.1155/2009/894726

Research Article

Downlink Assisted Uplink Zero Forcing for TDD Multiuser MIMO Systems

Petri Komulainen, Antti Tölli, Matti Latva-aho, and Markku Juntti

Centre for Wireless Communications, University of Oulu, P.O. Box 4500, 90014 Oulu, Finland

Correspondence should be addressed to Petri Komulainen, petri.komulainen@ee.oulu.fi

Received 1 February 2009; Revised 11 May 2009; Accepted 19 July 2009

Recommended by Bruno Clerckx

This paper proposes practical coordinated linear transmit-receive processing schemes for the uplink (UL) of multiuser multiple-input multiple-output (MIMO) systems in the time division duplex (TDD) mode. The base station (BS) computes the transmission parameters in a centralized manner and employs downlink (DL) pilot signals to convey the information of the beam selection and beamformers to be used by the terminals. When coexisting with the DL transmit-receive zero forcing, the precoded DL demodulation pilots can be reused for UL beam allocation so that no additional pilot overhead is required. Furthermore, the locally available channel state information (CSI) of the effective MIMO channel is sufficient for the terminals to perform transmit power and rate allocation independently. In order to reduce the UL pilot overhead as well, we propose reusing the precoded UL demodulation pilots in turn for partial CSI sounding. The achievable sum rate of the system is evaluated in time-varying fading channels and with channel estimation. According to the results, the proposed UL transmission strategy provides increased rates compared to single-user MIMO transmission combined with user selection as well as to UL antenna selection transmission, without being sensitive to CSI uncertainty.

1. Introduction

In order to attain all the capacity gains available in multiple-input multiple-output (MIMO) communication systems, channel state information in the transmitter (CSIT) should be utilized. CSIT is available in time division duplex (TDD) systems, provided that the channel does not change significantly between the receive and transmit periods. Due to the channel reciprocity, the receiving node can estimate the state of the channel during one frame, and use that knowledge for the purposes of MIMO transmission in the next one. CSI can be estimated from pilot symbols that are known to the receiver. The pilots are also necessary for performing coherent demodulation in the receiver side. In order to keep the pilot overhead as low as possible, it is desirable that the same pilot symbols are a useful reference for both reception and transmission.

In a cellular multiuser MIMO system, the downlink (DL) comprises a broadcast channel (BC), whereas the uplink (UL) is a multiple access channel (MAC). The channel reciprocity leads into duality properties between the BC and MAC [1, 2]. When designing the user multiplexing strategy for a MIMO system, both directions need to be taken into account together. A distinctive difference between the base station (BS) and the user terminals is that the BS can have the CSI of the channels to all the terminals, while the terminals only have access to the CSI of their individual radio channels. Thus, the BS is capable to centralized processing to attain space division multiple access (SDMA). On the other hand, the terminals can attempt SDMA like transmission only based on the information contained in the signal received in the DL.

TDD is one of the modes included in the cellular 3GPP Long-Term Evolution (LTE) standard, and it is best applicable to urban, local area or office deployments, where the transmit powers, mobile speeds, and the channel propagation delays are relatively low. The TDD mode can well facilitate advanced multiuser MIMO DL transmission methods, if the terminals provide CSI to the BS by transmitting channel sounding pilots in the UL [3]. The motivation of this

paper is to study the DL transmission, and to propose a practical matching UL beamforming method for improving the capacity of the cellular system. The underlying assumption is that both the DL and the UL employ orthogonal frequency division multiplexing (OFDM), where the frequency-time resource blocks experience essentially flat fading.

Zero forcing DL transmission by a multiantenna BS provides SDMA in which intracell multiuser interference is nulled. For single-antenna terminals, zero forcing (ZF) is achieved simply by channel inversion in the transmitter [4]. Coordinated transmit-receive processing with block diagonalization (BD) is a zero forcing SDMA scheme that supports also multiantenna user terminals [5]. It decouples the MIMO channels of different users so that precoding based on singular value decomposition (SVD) can be carried out individually for each user. Our preferred transmit-receive solution is obtained when the terminals employ conventional maximal ratio receivers (MRCs) as suggested in [6]. In that case, the ZF solution can be found via an iterative algorithm that was proposed in [7], and further studied in [8]. While corresponding general closed form solutions have not been presented, in [9] it was derived for a two-user case and in [10] the solutions for a three-user setup were studied.

It is beneficial to combine multiuser beamforming with greedy beam selection [11]. In the context of multiuser MIMO DL with coordinated transmit-receive processing, greedy beam selection was studied in [12, 13].

In a time-varying fading radio channel the CSI obtained during the TDD receive frame is already partially outdated when the transmit frame starts. Therefore, the CSI contains a lag error that has a decremental impact on the system performance. The effect of delayed CSI in case of single-user MIMO communications was studied in [14], and in case of DL multiuser MIMO systems in [15]. In addition to the lag error, the effect of noisy CSI estimation on multiuser multiple antenna systems was analyzed in [16].

Based on the principles of DL multiuser transmit-receive zero forcing and beam selection, in this paper, we propose a corresponding communication strategy for the UL. In [17], we presented a similar approach based DL BD by transmit processing only. While in that simple form of BD, the number of antennas in the BS must always be equal to or larger than the aggregate number of antennas in the user terminals [5], the strategy described here can support more general antenna setups and resource allocation methods. We also evaluate by simulations the impact of imperfect CSI estimation as well as lag error on the achievable rates in the system.

While the algorithms for multiuser processing and beam selection are known from literature, the main contribution of our work consists of two novel signaling concepts. The first concept is to convey the UL beamforming parameters to the terminals by means of DL pilot signals. The second concept is to append the UL demodulation pilot signal with additional pilot beams so that the combined signal serves as a full CSI sounding pilot. While the both new techniques can be applied in TDD systems separately, we introduce them as features supporting a combined uplink-downlink strategy with reduced pilot overhead. As a result, the precoded pilot symbols are sufficient in both UL and DL to satisfy the needs of both transmission and reception.

The paper is organized as follows. In Section 2, the generic uplink-downlink multiuser MIMO system model is described. Section 3 summarizes the ideas of coordinated transmit-receive processing and beam selection. Section 4 presents the details of the proposed uplink-downlink beamforming scheme, and in Section 5, numerical capacity analysis results are given. Finally, Section 6 concludes the paper.

2. System Model

We consider a MIMO system with one base station having N_{B} antenna elements, and K user terminals with N_{U} antenna elements each. Furthermore, we assume the users are symbol synchronous, and that each user $k \in \{1, 2, \ldots, K\}$ is allocated with $L_k \leq N$ data streams in both UL and DL, where $N = \min(N_{\mathrm{B}}, N_{\mathrm{U}})$. We denote the set of active, that is, scheduled users as $\mathcal{K} = \{k \mid L_k > 0\}$.

The complex DL MIMO signal received by the terminal of user k at symbol interval n can be written as

$$\mathbf{x}_k^{\mathrm{d}}(n) = \mathbf{H}_k \sum_{i \in \mathcal{K}} \mathbf{M}_i^{\mathrm{d}} \mathbf{A}_i^{\mathrm{d}} \mathbf{b}_i^{\mathrm{d}}(n) + \mathbf{z}_k^{\mathrm{d}}(n), \tag{1}$$

where $\mathbf{H}_k \in \mathbb{C}^{N_{\mathrm{U}} \times N_{\mathrm{B}}}$ is the channel matrix, $\mathbf{M}_k^{\mathrm{d}} = [\mathbf{m}_{k,1}^{\mathrm{d}} \cdots \mathbf{m}_{k,L_k}^{\mathrm{d}}] \in \mathbb{C}^{N_{\mathrm{B}} \times L_k}$ is the DL transmit precoder matrix with unit norm column vectors, $\mathbf{A}_k^{\mathrm{d}} = \mathrm{diag}(\sqrt{p_{k,1}^{\mathrm{d}}}, \ldots, \sqrt{p_{k,L_k}^{\mathrm{d}}})$ is the real-valued diagonal transmit amplitude matrix, $\mathbf{b}_k^{\mathrm{d}}(n) \in \mathbb{C}^{L_k \times 1}$ is the data symbol vector, and $\mathbf{z}_k^{\mathrm{d}}(n) \in \mathbb{C}^{N_{\mathrm{U}} \times 1}$ is a white Gaussian noise vector with variance N_0 per element. Similarly, the UL signal received by the BS becomes

$$\mathbf{x}^{\mathrm{u}}(n) = \sum_{i \in \mathcal{K}} \mathbf{H}_i^{\mathrm{T}} \mathbf{M}_i^{\mathrm{u}} \mathbf{A}_i^{\mathrm{u}} \mathbf{b}_i^{\mathrm{u}}(n) + \mathbf{z}^{\mathrm{u}}(n), \tag{2}$$

where $\mathbf{M}_k^{\mathrm{u}} = [\mathbf{m}_{k,1}^{\mathrm{u}} \cdots \mathbf{m}_{k,L_k}^{\mathrm{u}}] \in \mathbb{C}^{N_{\mathrm{U}} \times L_k}$ is the UL transmit precoder matrix with unit norm column vectors, and $\mathbf{A}_k^{\mathrm{u}} = \mathrm{diag}(\sqrt{p_{k,1}^{\mathrm{u}}}, \ldots, \sqrt{p_{k,L_k}^{\mathrm{u}}})$ is the diagonal transmit amplitude matrix. Here, $()^{\mathrm{T}}$ denotes matrix transpose, and for complex conjugation and conjugate transposition, notations $()^*$ and $()^{\mathrm{H}}$ are used, respectively. The signal model is free from intersymbol interference; this can be realized, for example, by OFDM.

For the purposes of spatial processing, we write the singular value decomposition of the individual MIMO channel of user k as

$$\mathbf{H}_k = \mathbf{U}_k \mathbf{\Lambda}_k \mathbf{V}_k^{\mathrm{H}}, \tag{3}$$

where the matrices $\mathbf{U}_k = [\mathbf{u}_{k,1} \cdots \mathbf{u}_{k,N}] \in \mathbb{C}^{N_{\mathrm{U}} \times N}$, $\mathbf{V}_k = [\mathbf{v}_{k,1} \cdots \mathbf{v}_{k,N}] \in \mathbb{C}^{N_{\mathrm{B}} \times N}$, and $\mathbf{\Lambda}_k = \mathrm{diag}(\lambda_{k,1}, \ldots, \lambda_{k,N})$ contain, respectively, the left and right singular vectors and singular values in nonascending order, corresponding to the nonzero eigenmodes. Note that we excluded the null space from the decomposition. In physical channels, the number of nonzero singular values is typically N.

We also define generic linear receivers $\mathbf{W}_k^{\mathrm{d}} = [\mathbf{w}_{k,1}^{\mathrm{d}} \cdots \mathbf{w}_{k,L_k}^{\mathrm{d}}] \in \mathbb{C}^{N_{\mathrm{U}} \times L_k}$ and $\mathbf{W}_k^{\mathrm{u}} = [\mathbf{w}_{k,1}^{\mathrm{u}} \cdots \mathbf{w}_{k,L_k}^{\mathrm{u}}] \in \mathbb{C}^{N_{\mathrm{B}} \times L_k}$. Depending on the transmit precoders and receivers, signal-to-interference-plus-noise ratio (SINR) can be calculated for each stream [8]. Assuming the data streams are uncorrelated, SINR for stream s of user k in UL direction is

$$\gamma_{k,s}^{\mathrm{u}} = \frac{p_{k,s}^{\mathrm{u}} \left| \mathbf{w}_{k,s}^{\mathrm{u}\ \mathrm{H}} \mathbf{H}_k^{\mathrm{H}} \mathbf{m}_{k,s}^{\mathrm{u}} \right|^2}{\sum_{(i,j) \neq (k,s)} p_{i,j}^{\mathrm{u}} \left| \mathbf{w}_{k,s}^{\mathrm{u}\ \mathrm{H}} \mathbf{H}_i^{\mathrm{H}} \mathbf{m}_{i,j}^{\mathrm{u}} \right|^2 + N_0 \left\| \mathbf{w}_{k,s}^{\mathrm{u}} \right\|^2}, \quad (4)$$

and similarly

$$\gamma_{k,s}^{\mathrm{d}} = \frac{p_{k,s}^{\mathrm{d}} \left| \mathbf{w}_{k,s}^{\mathrm{d}\ \mathrm{H}} \mathbf{H}_k \mathbf{m}_{k,s}^{\mathrm{d}} \right|^2}{\sum_{(i,j) \neq (k,s)} p_{i,j}^{\mathrm{d}} \left| \mathbf{w}_{k,s}^{\mathrm{d}\ \mathrm{H}} \mathbf{H}_k \mathbf{m}_{i,j}^{\mathrm{d}} \right|^2 + N_0 \left\| \mathbf{w}_{k,s}^{\mathrm{d}} \right\|^2} \quad (5)$$

in DL. Furthermore, by assuming Gaussian symbol alphabets, the mutual information between the transmitted sequence and decision statistics per stream becomes

$$R_{k,s} = \log_2 (1 + \gamma_{k,s}) \text{ bits/s/Hz}, \quad (6)$$

which is also an upper bound for the achievable data rate.

3. Coordinated Transmit-Receive Processing

Coordinated transmit-receive processing by block diagonalization is a known method for DL zero forcing [5]. It can support any number of antennas in the BS and the terminals as well as flexible beam allocation. The DL signal processing chain is depicted in Figure 1(a). Let $\mathbf{F}_k \in \mathbb{C}^{N_{\mathrm{U}} \times L_k}$ be an orthonormal receiver processor matrix for user k. The zero forcing criterion between users can be expressed as

$$\mathbf{F}_k^{\mathrm{H}} \mathbf{H}_k \mathbf{C}_i = \mathbf{0}, \quad i \neq k, \quad (7)$$

which implies that the receiver finishes up the zero forcing by rejecting the residual interference seen in the receiver antennas. To enable this, the interference must lie in the $(N_{\mathrm{U}} - L_k)$-dimensional subspace orthogonal to the columns of $\mathbf{F}_k$. The task of the transmit processor $\mathbf{C}_k$ is to ensure this property.

The effective single-user MIMO DL channels are further decomposed into L_k parallel channels as

$$\overline{\mathbf{H}}_k = \mathbf{F}_k^{\mathrm{H}} \mathbf{H}_k \mathbf{C}_k = \overline{\mathbf{U}}_k \overline{\mathbf{\Lambda}}_k \overline{\mathbf{V}}_k^{\mathrm{H}}, \quad (8)$$

where $\overline{\mathbf{\Lambda}}_k = \mathrm{diag}(\overline{\lambda}_{k,1}, \ldots, \overline{\lambda}_{k,L_k})$, in order to apply SVD precoding so that the DL precoding matrix for user k is $\mathbf{M}_k^{\mathrm{d}} = \mathbf{C}_k \overline{\mathbf{V}}_k$ and the corresponding receiver $\mathbf{W}_k^{\mathrm{d}} = \mathbf{F}_k \overline{\mathbf{U}}_k$.

The multiuser MIMO system is effectively decoupled into a set of single-user MIMO links. Thus, power and rate allocation can be decoupled from the precoder design, and conventional coding and modulation methods can be applied. The achievable system sum rate becomes

$$R_{\mathrm{sum}} = \sum_{k,s} \log_2 \left(1 + \frac{p_{k,s} \overline{\lambda}_{k,s}^2}{N_0} \right), \quad (9)$$

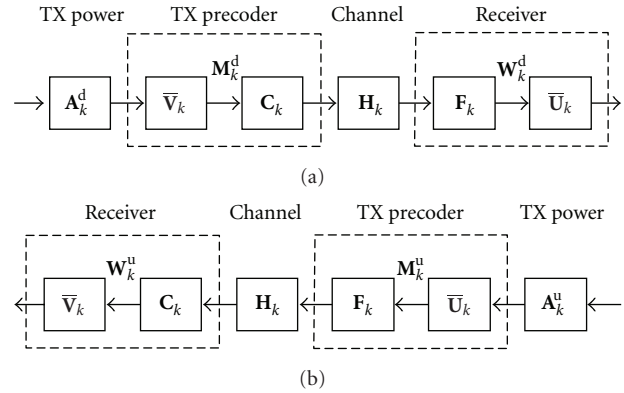


Figure 1: Ideal signal processing chain for multiuser zero forcing: (a) downlink, (b) uplink.

where $p_{k,s}$ is the transmit power allocated to the eigenmode s of user k.

In the coordinated transmit-receive processing, the BS computes all the transmitters and corresponding receivers in a centralized manner, based on the CSI of the selected users. In this section, the processing is described with the assumption that the channel matrices $\mathbf{H}_k$ are known. In Section 4 we explain how the UL pilot responses of our proposed strategy can be applied as a reference instead.

3.1. Closed-Form ZF Solution. The solution for (7) is not unique, as the receive processors $\mathbf{F}_k$ can be selected in multiple ways. One simple choice is to choose the column vectors associated to the strongest singular values from matrix $\mathbf{U}_k$ in (3) as suggested in [5]. Let $\mathbf{U}_k^{(1)} = [\mathbf{u}_{k,1} \cdots \mathbf{u}_{k,L_k}] \in \mathbb{C}^{N_{\mathrm{U}} \times L_k}$ contain the L_k selected left singular vectors and $\mathbf{V}_k^{(1)} = [\mathbf{v}_{k,1} \cdots \mathbf{v}_{k,L_k}] \in \mathbb{C}^{N_{\mathrm{B}} \times L_k}$ the corresponding right singular vectors. The zero forcing criterion becomes $\mathbf{U}_k^{(1)\mathrm{H}} \mathbf{H}_k \mathbf{C}_i = \mathbf{0}$, which can be shown to be equivalent to $\mathbf{V}_k^{(1)\mathrm{H}} \mathbf{C}_i = \mathbf{0}$.

The decomposition (8) lends itself for the purposes of UL transmission as well, as the effective UL MIMO channel is a transposed version of the DL so that $\overline{\mathbf{H}}_k^{\mathrm{T}} = \mathbf{C}_k^{\mathrm{T}} \mathbf{H}_k^{\mathrm{T}} \mathbf{F}_k^* = \overline{\mathbf{V}}_k^* \overline{\mathbf{\Lambda}}_k \overline{\mathbf{U}}_k^{\mathrm{T}}$. Thus our proposed UL signal processing chain is ideally a reversed version of the DL so that the receivers become transmitters and vice versa, as shown in Figure 1(b). Consequently, the zero forcing criterion in the UL is equivalent to (7), that is, $\mathbf{C}_i^{\mathrm{T}} \mathbf{H}_k^{\mathrm{T}} \mathbf{F}_k^* = \mathbf{0}$, $i \neq k$. Since in both directions the eigenmodes of the effective MIMO channels are the same, and as the interference is nulled both ways, for each user the UL and DL are essentially equal. The achievable rates differ only if different transmit powers are applied or if the background noise levels seen by the BS and the terminal are different.

3.2. Iterative ZF Solution. The iterative solution for (7) has two desirable properties. Firstly, the performance in terms of achievable rates compared to the closed form solution is improved. Secondly, the optimal receivers in user terminals are filters matched to the received stream responses so that

ideally, the terminal side needs not actively estimate and suppress interference.

In the iterative algorithm the processors $\mathbf{F}_k$ are initialized by matrix $\mathbf{U}_k^{(1)}$, and then the transmitter $\mathbf{C}_k$ and receiver $\mathbf{F}_k$ processors for each user are optimized successively until orthogonality between the users is achieved [7, 8]. After convergence, the received DL stream responses dedicated to user k are $\mathbf{H}_k\mathbf{C}_k\overline{\mathbf{V}}_k = \mathbf{F}_k\overline{\mathbf{\Lambda}}_k$, which implies that the final zero forcing receiver matrix is a set of matched filters.

In our simulations, in the case of $N_{\mathrm{B}} = 4$, $N_{\mathrm{U}} = 2$, and $K = 4$, the iterative algorithm converged on the average in less than five iterations. Our stopping condition of the algorithm required that the sum of the absolute values of all cross terms $\mathbf{F}_k^{\mathrm{H}}\mathbf{H}_k\mathbf{M}_i^{\mathrm{d}}$ must be less than 10^{-4}.

3.3. Greedy Beam Selection. Greedy beam selection is a process of allocating beams to the users based on their individual channel conditions and spatial compatibility [11]. In the context of the multiuser MIMO system and zero forcing, beam selection has been studied in [12, 13]. The algorithm consecutively selects at most $N_{\mathrm{S}} = \min(KN_{\mathrm{U}}, N_{\mathrm{B}})$ eigenbeams from the total set of $K \cdot \min(N_{\mathrm{U}}, N_{\mathrm{B}})$ to be allocated. Number N_{S} indicates the number of degrees of freedom available in the system.

First, the strongest eigenbeam, that is, the one with the largest singular value $\lambda_{k,s}$ among all users is selected. Subsequently, on each step of the selection process, the beam having the largest component orthogonal to the previously selected beams is chosen as

$$(k,s) = \arg\max_{k,s}\left[\left(\mathbf{I} - \mathbf{S}\left(\mathbf{S}^{\mathrm{H}}\mathbf{S}\right)^{-1}\mathbf{S}^{\mathrm{H}}\right)\lambda_{k,s}\mathbf{v}_{k,s}\right], \tag{10}$$

where matrix $\mathbf{S}$ contains as columns all the right singular vectors $\mathbf{v}_{k,s}$ corresponding to the previously selected eigenbeams. Note that the L_k eigenbeams selected for user k are not necessarily the strongest, since weaker beams may be preferred due to their better spatial compatibility properties.

The selection process stops if the calculated capacity of the system is reduced compared to the previously selected beam set. Thus, there may be fewer active streams in the system than there are degrees of freedom. In this paper, the stopping condition is always calculated based on the closed-form zero forcing solution in order to avoid multiple zero forcing iteration rounds.

The role of the beam selection is to make the problem of zero forcing relatively easy, by ensuring that the selected eigenbeams are nearly orthogonal so that the zero forcing loss remains acceptable. The stopping condition of the selection has a similar effect, as the algorithm rather stops than chooses more linearly dependent eigenbeams.

A straightforward simplification to the multiple access protocol can be introduced by restricting the maximum number of beams per user to be one, that is, $L_k \leq 1$. Especially when the number of users is high, the effect of the restriction on the system throughput is minor. However, by allowing multiple data streams per user, higher user peak data rates can be provided.

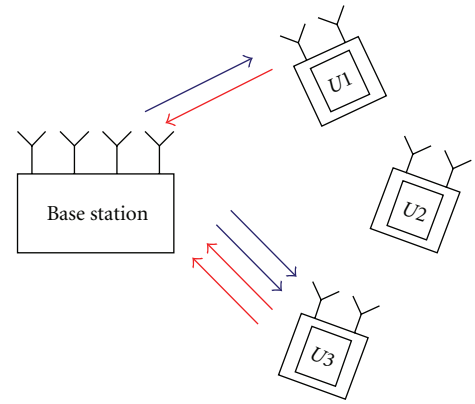


Figure 2: Example of uplink-downlink beam selection.

In our proposed strategy, the same beam set is selected both for UL and DL. An example outcome of the selection is depicted in Figure 2.

4. Uplink-Downlink Beamforming Strategy

The main contribution of this paper consists of two novel concepts. The first concept is to convey the uplink (UL) beamforming parameters to the terminals by means of downlink (DL) pilot signals. The second one is to append the UL demodulation pilot signal with additional pilot beams so that the combined signal serves as a CSI sounding pilot. While the both new techniques can be applied in TDD systems separately, we introduce them as features supporting a combined uplink-downlink strategy with reduced pilot overhead.

Most of the intelligence as well as the computational complexity of the proposed strategy lie in the base station (BS) that carries out the multiuser processing, including beam selection and precoding. On the other hand, the terminals essentially perform single-user MIMO processing in conjunction with interference suppression.

4.1. Signaling for Uplink Beamforming. The resource allocation and pilot signaling in TDD mode are in general open research problems and standardization issues. Due to the TDD channel reciprocity, the need for CSI quantization can be avoided unlike in the FDD mode. Thus, in principle, TDD can support more advanced spatial signal processing methods than FDD. However, reasonable pilot signal overhead is still required, and due to estimation errors CSI is not perfect. In order to facilitate fast advanced centralized processing in the BS, antenna-specific UL CSI sounding pilots are needed [3]. These pilots enable any form of multiuser MIMO precoding in the DL.

The use of the CSI sounding pilot enables centralized control also for the UL transmissions, as full multiuser CSI is gathered by the BS. A problem to solve is how to signal the desired UL beamforming parameters to the terminals. We propose to use beam allocation pilot signals to declare the desired UL transmit precoders. In conjunction with

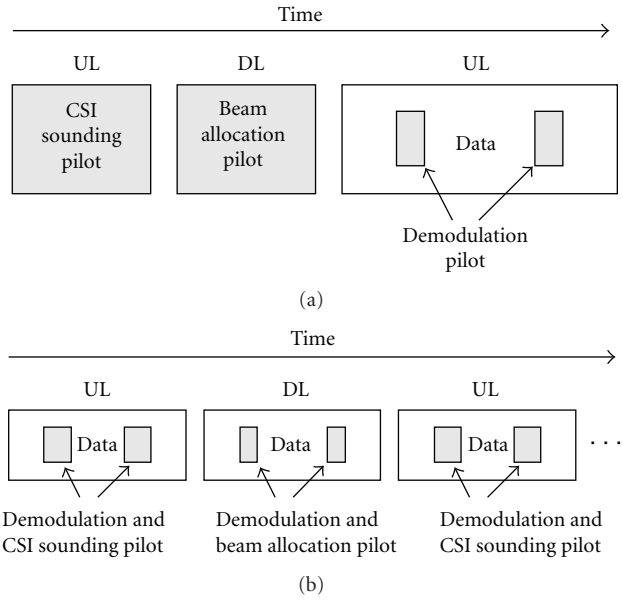


Figure 3: Simplified TDD frame and pilot structure needed for (a) UL beamforming, (b) UL/DL beamforming.

zero forcing multiplexing, and assuming knowledge of the background noise level at the receiving end, each terminal may then locally decide on the power control, modulation and coding of its UL data streams, without the need for the BS to communicate this to the terminal. In order to facilitate reception at the BS, the UL data includes embedded demodulation pilot symbols. The signaling sequence is depicted in Figure 3(a).

A more conventional signaling choice for the BS is to distribute quantized information, indicating desired UL precoders chosen from a predefined codebook. Due to the limited size of the codebook, perfect orthogonality between the users' effective channels cannot be ensured. Thus, in order to guarantee the UL decoding result, user-specific transmit power and rate parameters should be communicated as well. Comparison of the two schemes is presented in Table 1. In the simplest case, the quantized signaling can support UL antenna selection transmission, where the BS chooses a subset of terminal antennas that each simultaneously transmits one independent unprecoded data stream. This method is used as a benchmark in the simulations.

One more obvious method to facilitate UL precoding is to employ a DL common pilot so that each terminal can form beams based on the knowledge of its individual MIMO channel. However, this mode does not easily allow centralized multiuser control, and the resulting UL beams may end up undecodable if they are not spatially compatible.

4.2. Combined Uplink-Downlink Signaling. When applying multiuser MIMO precoding in the DL, the DL demodulation pilots may be reused as beam allocation pilots as shown in Figure 3(b). In this approach, the same spatial beams are active in both directions, and the need for specific DL signaling of the desired UL precoders is removed. On the other hand, the UL demodulation pilots can be reused for partial CSI sounding. By adding parallel pilot beams, full CSI sounding can be achieved, as described in the following subsection. As a result, the amount of required specific CSI sounding pilot overhead is reduced.

Table 1: UL MU beamforming approaches.

Method	UL signaling	DL signaling	Power and rate control
Unquantized precoding	CSI sounding pilot	Beam allocation pilots	May be locally decided by terminal
Quantized precoding	CSI sounding pilot	Precoder indexes and rate parameters	Signalled by BS

For example, in our simulation setup with $K = 4$, $N_{\mathrm{B}} = 4$, and $N_{\mathrm{U}} = 2$, coupling of the UL and DL beamforming halves the required DL pilot overhead. At the same time, the UL pilot overhead is reduced approximately by one third.

Obviously, the combined strategy sets constraints to the overall resource allocation of the system, as the same frequency resource blocks are assumed to be allocated to the same users in both UL and DL. Therefore, the concept is at its most efficient when the offered data traffic loads in both directions are approximately equal. In the system level, the possible asymmetry of the traffic can be treated in time domain, for example, by allocating more time frames to the DL than UL. Furthermore, the concept of reusing the demodulation pilot signals for CSI sounding and beam allocation can be utilized whenever the receive frame is close enough to the corresponding transmit frame. In other times, separate sounding pilots need to be employed.

4.3. Pilot Responses. Pilot symbols transmitted with beamforming via the same precoders as data are necessary in order to facilitate coherent demodulation. However, unlike data, we propose that the pilots have equal power allocation per stream. This way the channel gains can be correctly observed from the received signal without getting mixed with the amplitude adjustment caused by power allocation, and the pilot responses can be utilized for the purpose of transmit precoding as well.

For CSI sounding, it is necessary that the UL pilots of each user fully span the N_{U}-dimensional transmit signal space even when the number of data streams L_k is lower than N_{U}. Therefore, we propose appending the L_k UL pilot streams associated with the allocated data streams by another $N_{\mathrm{U}} - L_k$ pilot streams. Thus, the unitary pilot precoder matrix becomes

$$\overline{\mathbf{M}}_k^{\mathrm{u}} = \left[\mathbf{M}_k^{\mathrm{u}} \;\; \widetilde{\mathbf{M}}_k^{\mathrm{u}}\right] \in \mathbb{C}^{N_{\mathrm{U}} \times N_{\mathrm{U}}}, \tag{11}$$

where $\mathbf{M}_k^{\mathrm{u}} \in \mathbb{C}^{N_{\mathrm{U}} \times L_k}$ is the data precoder matrix, and $\widetilde{\mathbf{M}}_k^{\mathrm{u}} \in \mathbb{C}^{N_{\mathrm{U}} \times (N_{\mathrm{U}} - L_k)}$ contains the precoders for the additional pilot streams. On the other hand, in the DL it suffices to transmit just as many pilot streams as there are data streams.

Due to pilot precoding, neither the BS nor the terminals have explicit knowledge of channel matrices $\mathbf{H}_k$ but only the pilot responses. Excluding the transmit power and noise, the pilot responses are

$$\begin{aligned}\mathbf{R}_{k,i}^{\mathrm{d}} &= \mathbf{H}_k\mathbf{M}_i^{\mathrm{d}} \in \mathbb{C}^{N_{\mathrm{U}}\times L_i},\\ \mathbf{R}_k^{\mathrm{u}} &= \mathbf{H}_k^{\mathrm{T}}\mathbf{M}_k^{\mathrm{u}} \in \mathbb{C}^{N_{\mathrm{B}}\times L_k},\\ \overline{\mathbf{R}}_k^{\mathrm{u}} &= \mathbf{H}_k^{\mathrm{T}}\overline{\mathbf{M}}_k^{\mathrm{u}} \in \mathbb{C}^{N_{\mathrm{B}}\times N_{\mathrm{U}}}\end{aligned} \tag{12}$$

for DL and UL, respectively. In the DL, $\mathbf{R}_{k,i}^{\mathrm{d}}$ denotes the response seen by user k of the signal transmitted to user i.

The number of required pilot streams in UL is $K \cdot N_{\mathrm{U}}$ and increases with the number of simultaneous users, whereas for DL N_{B} pilot streams always suffice. Thus, the UL limits the practical number of users to be included in the same spatial processing group.

4.4. Base Station Processing. Section 3 described how the coordinated transmit-receive processing and beam selection are carried out by the BS, based on the knowledge of the MIMO channels $\mathbf{H}_k$. However, the same computations can be realized by replacing the channel matrices with the UL pilot responses $\overline{\mathbf{R}}_k^{\mathrm{uT}} = \overline{\mathbf{M}}_k^{\mathrm{uT}}\mathbf{H}_k \in \mathbb{C}^{N_{\mathrm{U}}\times N_{\mathrm{B}}}$ as well, since the right singular vectors (3), forming the transmit signal space, and the corresponding singular values are invariant to the multiplication by the unitary pilot precoder matrix. As a result, the BS obtains the same set of transmit precoders and powers as when applying the channel matrices directly. On the other hand, the set of receiver processors the algorithm assumes will be different.

Let $\widetilde{\mathbf{F}}_k \in \mathbb{C}^{N_{\mathrm{U}}\times L_k}$ be the orthonormal receiver processor matrices and $\mathbf{C}_k$ the orthonormal transmit processor matrices, $k \in \mathcal{K}$, given by the zero forcing algorithm—closed-form or iterative—at the BS after applying the UL pilot responses as a reference. These processors satisfy, instead of (7), the condition

$$\widetilde{\mathbf{F}}_k^{\mathrm{H}}\overline{\mathbf{R}}_k^{\mathrm{uT}}\mathbf{C}_i = \widetilde{\mathbf{F}}_k^{\mathrm{H}}\overline{\mathbf{M}}_k^{\mathrm{uT}}\mathbf{H}_k\mathbf{C}_i = \mathbf{0}, \quad i \neq k. \tag{13}$$

Furthermore, let $\mathbf{F}_k \in \mathbb{C}^{N_{\mathrm{U}}\times L_k}$ be the receiver processor the user terminal k applies in order to reject multiuser interference. This processor must satisfy $\mathbf{F}_k^{\mathrm{H}}\mathbf{H}_k\mathbf{C}_i = \mathbf{0}$, $i \neq k$. By comparing to (13) we can see that $\mathbf{F}_k = \overline{\mathbf{M}}_k^{\mathrm{u}*}\widetilde{\mathbf{F}}_k$ is the valid orthonormal zero forcing processor at the terminal.

The underlying assumption in the transmit-receive zero forcing strategy is that the receivers employed both in the DL and the UL are zero forcing detectors. However, the actual receiver side may construct other more advanced or robust detectors in order to improve performance. In addition to zero forcing (ZF), linear minimum mean square error (LMMSE) detectors are considered here. Both receiver types can be formulated for arbitrary transmit precoders and channel responses. Let us stack the UL stream responses and transmit amplitudes into large matrices $\mathbf{R}^{\mathrm{u}} = [\mathbf{R}_1^{\mathrm{u}} \cdots \mathbf{R}_K^{\mathrm{u}}] \in \mathbb{C}^{N_{\mathrm{B}}\times L}$ and $\mathbf{A}^{\mathrm{u}} = \mathrm{diag}(\mathbf{A}_1^{\mathrm{u}}, \ldots, \mathbf{A}_K^{\mathrm{u}})$, respectively, where $L = \sum_k L_k$ is the total number of streams to be detected. The ZF and LMMSE UL multiuser receivers become

$$\mathbf{W}_{\mathrm{ZF}}^{\mathrm{u}} = \mathbf{R}^{\mathrm{u}}\left(\mathbf{R}^{\mathrm{uH}}\mathbf{R}^{\mathrm{u}}\right)^{-1}, \tag{14}$$

$$\mathbf{W}_{\mathrm{MMSE}}^{\mathrm{u}} = \left(\mathbf{R}^{\mathrm{u}}\mathbf{A}^{\mathrm{u}}(\mathbf{R}^{\mathrm{u}}\mathbf{A}^{\mathrm{u}})^{\mathrm{H}} + N_0\mathbf{I}\right)^{-1}\mathbf{R}^{\mathrm{u}}, \tag{15}$$

respectively. Here, the user-specific receivers are stacked in the large result matrix as $\mathbf{W}^{\mathrm{u}} = [\mathbf{W}_1^{\mathrm{u}} \cdots \mathbf{W}_K^{\mathrm{u}}] \in \mathbb{C}^{N_{\mathrm{B}}\times L}$. Note that for our proposed UL precoding, the ZF receiver is ideally equivalent to the corresponding DL precoder $\mathbf{C}_k\overline{\mathbf{V}}_k$. In practice, however, due to estimation errors, channel time-variations and other nonidealities, the receiver must always rely on the received stream responses.

4.5. Terminal Processing. In the DL, the total number of allocated streams is usually larger than the number of receiver antennas in one terminal, that is, $N_{\mathrm{U}} < L$. Therefore, the terminal may not be able to perfectly cancel interference if the DL precoding was not perfect, and in this case the strict ZF receiver may be replaced with the least norm (LN) receiver. Let us again stack the stream responses into a large matrix $\mathbf{R}_k^{\mathrm{d}} = [\mathbf{R}_{k,1}^{\mathrm{d}} \cdots \mathbf{R}_{k,K}^{\mathrm{d}}] \in \mathbb{C}^{N_{\mathrm{U}}\times L}$ so that the user-specific ZF/LN receiver can be expressed as

$$\begin{aligned}\mathbf{W}_{k,\mathrm{ZF/LN}}^{\mathrm{d}} &= \mathbf{R}_{k,k}^{\mathrm{d}}\left(\mathbf{R}_k^{\mathrm{dH}}\mathbf{R}_k^{\mathrm{d}}\right)^{-1}, \quad N_{\mathrm{U}} > L,\\ \mathbf{W}_{k,\mathrm{ZF/LN}}^{\mathrm{d}} &= \left(\mathbf{R}_k^{\mathrm{d}}\mathbf{R}_k^{\mathrm{dH}}\right)^{-1}\mathbf{R}_{k,k}^{\mathrm{d}}, \quad N_{\mathrm{U}} \leq L.\end{aligned} \tag{16}$$

Note that in the case of the proposed DL precoding, ideally the ZF/LN receiver results in a true ZF receiver, even when $N_{\mathrm{U}} < L$. Furthermore, we formulate the LMMSE receiver as

$$\mathbf{W}_{k,\mathrm{MMSE}}^{\mathrm{d}} = \left(\mathbf{R}_k^{\mathrm{d}}\mathbf{A}^{\mathrm{d}}(\mathbf{R}_k^{\mathrm{d}}\mathbf{A}^{\mathrm{d}})^{\mathrm{H}} + N_0\mathbf{I}\right)^{-1}\mathbf{R}_{k,k}^{\mathrm{d}}, \tag{17}$$

where $\mathbf{A}^{\mathrm{d}} = \mathrm{diag}(\mathbf{A}_1^{\mathrm{d}}, \ldots, \mathbf{A}_K^{\mathrm{d}})$. For the iterative zero forcing transmit-receive processing, in an ideal case, both the ZF/LN and the LMMSE receiver are equivalent to the matched filter (MF) $\mathbf{W}_{k,\mathrm{MF}} = \mathbf{R}_{k,k}^{\mathrm{d}}$.

The transmit precoding for the UL relies on the locally available CSI of the effective MIMO channel and the reversal of the DL signal processing chain. The receive beamformers can be used in turn as transmit precoders. Let $\mathbf{R}_{k,k}^{\mathrm{d}} = [\mathbf{r}_{k,1}^{\mathrm{d}} \cdots \mathbf{r}_{k,L_k}^{\mathrm{d}}]$ be the received DL response matrix of user k, and $[\mathbf{w}_{k,1}^{\mathrm{d}} \cdots \mathbf{w}_{k,L_k}^{\mathrm{d}}]$ the corresponding ZF/LN receiver matrix in the case of ideal DL precoding. The UL precoders are obtained by normalizing $\mathbf{m}_{k,s}^{\mathrm{u}} = \mathbf{w}_{k,s}^{\mathrm{d}*}/\|\mathbf{w}_{k,s}^{\mathrm{d}}\|$, for $s = 1, \ldots, L_k$. As a result, the gains of the effective single user MIMO channel can be observed from $\overline{\lambda}_{k,s} = \mathbf{m}_{k,s}^{\mathrm{uT}}\mathbf{r}_{k,s}^{\mathrm{d}}$, for $s = 1, \ldots, L_k$, so that the terminal can perform UL transmit power allocation by maximizing

$$R_k^{\mathrm{u}} = \sum_{s=1}^{L_k}\log_2\left(1 + \frac{p_{k,s}}{N_0}\left|\mathbf{m}_{k,s}^{\mathrm{uT}}\mathbf{r}_{k,s}^{\mathrm{d}}\right|^2\right), \tag{18}$$

while applying the individual power constraint $\sum_s p_{k,s} = P_k$.

However, if the DL precoding was not ideal, or the terminal receiver is formulated based on estimated channel, the receive beamformers of user k do not necessarily remain orthogonal to each other. A conceptually straightforward way to orthonormalize the receive beamformers, and to simultaneously obtain the additional $N_U - L_k$ UL pilot precoders, is to perform full SVD as $\mathbf{W}_k^d = \dot{\mathbf{U}}_k \dot{\mathbf{\Lambda}}_k \dot{\mathbf{V}}_k^H$, and to set $\overline{\mathbf{M}}_k^u = \dot{\mathbf{U}}_k^* \in \mathbb{C}^{N_U \times N_U}$, where the first L_k columns correspond to the data streams. This method was used in the simulations of this paper.

It is worth noting that even when the terminal employs the LMMSE receiver, in the closed-form transmission mode, the transmit precoders are still calculated based on the ZF/LN receivers. In the iterative zero forcing mode, when operating with estimated CSI, it turned out that the MF receiver is the best reference for UL precoding, even though as a receiver ZF/LN performs better.

4.6. CSI Uncertainty. The treatment in the previous sections considered error-free CSI. In practice the beam selection, transmit precoding, and receiving have to be carried out based on noisy channel responses experienced during the latest received frame prior to transmission. In a time-varying channel this results in a lag error in transmit CSI. As a result, the orthogonality between users and streams in DL is partially lost. Also in the UL, the channel reciprocity is reduced. In the receiver side, the pilot reference is timely and correct so that both the desired signal and interference responses can be estimated and utilized without lag error.

We assume that the pilot symbol sequences associated with different streams and users are all mutually orthogonal, which accommodates interference free channel or pilot response estimation. For zero forcing transmit and receive processing, the estimation of the pilot responses $\overline{\mathbf{R}}_k^u$ and $\mathbf{R}_{k,i}^d$ is adequate. On the other hand, in order to construct LMMSE receivers, the spatial signal covariance or the transmit amplitudes $\mathbf{A}_k^u$ and $\mathbf{A}_k^d$ need to be known or estimated. For our simulations, the estimation of signal covariance is carried out as described in [17].

In the following, we exclude the user indexes and discuss how different error sources accumulate to the performance of the proposed system. The performance depends on the transmit precoders and receiver filters as indicated by (4) and (5). The choice of the unitary UL pilot precoder matrix $\overline{\mathbf{M}}^u$ has no effect on the DL precoding, whereas the DL pilot precoders affect the UL data precoding. The precoders are formed based on estimated pilot responses, so that

$$\begin{aligned} \widehat{\mathbf{M}}^d(n) &= f_B\left(\widehat{\overline{\mathbf{R}}}^u(n-1)\right), \\ \widehat{\mathbf{M}}^u(n) &= f_U\left(\widehat{\mathbf{R}}^d(n-1)\right), \end{aligned} \tag{19}$$

where n is the frame index, and f_B and f_U denote the precoding algorithms running in the BS and in the terminals, respectively. Let $\mathbf{D}$ be the channel lag error so that $\mathbf{H}(n-1) = \mathbf{H}(n) + \mathbf{D}(n)$. By denoting estimation noise $\mathbf{E}$, the estimates in BS become

$$\widehat{\overline{\mathbf{R}}}^u(n-1) = (\mathbf{H}(n) + \mathbf{D}(n))^T \overline{\mathbf{M}}^u + \mathbf{E}^u(n-1) \tag{20}$$

and in the terminal side

$$\widehat{\mathbf{R}}^d(n-1) = (\mathbf{H}(n) + \mathbf{D}(n))\widehat{\mathbf{M}}^d(n-1) + \mathbf{E}^d(n-1), \tag{21}$$

which indicates that the error sources seen in both UL and DL accumulate to affect the UL transmission.

5. Numerical Results

Different multiuser MIMO scenarios were simulated in frequency flat fading with Jakes' Doppler spectrum and uncorrelated channels between antennas. We denote the Doppler spread $D_S = 2f_d$ where f_d is the maximum Doppler shift. The equal length UL and DL TDD frames of duration T_{frame} follow each other consecutively as illustrated in Figure 3(b). Each simulation comprises 20 000 randomly generated, independent channel process bursts of several frames. The channel coefficients remain constant over each frame. System signal-to-noise-ratio SNR was set to 10 dB, and it is defined as $\sum_k P_k/N_0$. All the methods compared employ the same sum transmit power.

In order to compare the effect of spatial processing between DL and UL, we apply here the same power constraints in both directions. This is a reasonable assumption in office deployments or femto-cells, where the base station does not employ significantly higher transmit powers compared to the mobile devices. As a result, the supported rates in the UL and DL are ideally equal. In our simple and primitively fair allocation rule, each user is granted with a share of the total transmit power, proportional to the number of beams it was allocated. That is,

$$P_k = \frac{P \cdot L_k}{\sum_i L_i}, \tag{22}$$

where P is the total transmitted power in the cell.

One of the simulated benchmark methods is the UL antenna selection transmission, where the BS chooses a subset of terminal antennas that simultaneously transmit one independent unprecoded data stream each. Here, the greedy selection algorithm (10) is applied so that the channel singular vectors are replaced by channel vectors, that is, by rows from matrices $\mathbf{H}_k$. Thus, centralized multiuser control is exercised in order to ensure the spatial compatibility of the concurrent transmissions. Equal transmit power per antenna is allocated, and multiple data streams per user are allowed. While antenna selection is simpler compared to the UL beamforming, it offers no reduction to the required pilot overhead, since the UL CSI sounding pilots are still needed for reference.

Another comparison scheme is the single-user MIMO transmission, "best-user SVD", where the user with the strongest MIMO channel is always chosen for single-user MIMO transmission by SVD precoding. In that frame, the transmit power of the cell is allocated to one user.

Figure 4 shows the sum rate performance of the different schemes versus the number of users K, in conjunction with greedy beam selection and perfect CSI in static channel ($D_S = 0$) for $N_B = 4$ BS antennas and $N_U = 2$ terminal antennas. As can be seen, the iterative ZF solution

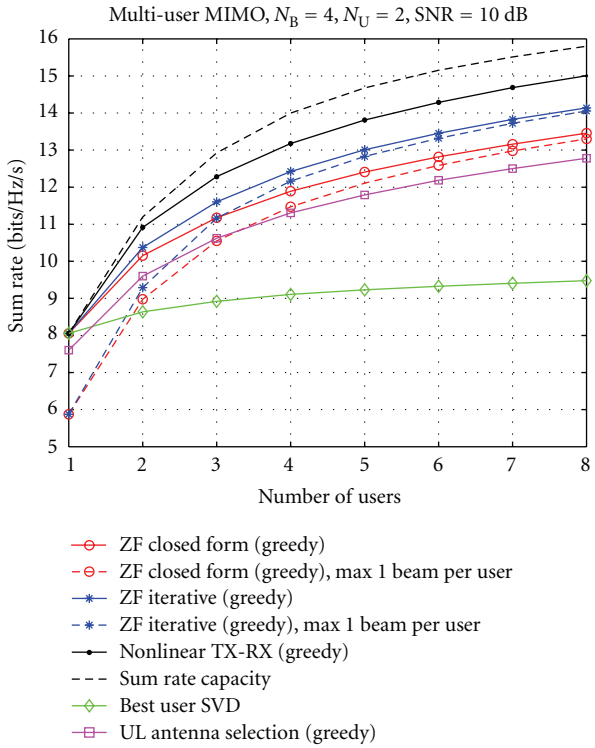


FIGURE 4: Average sum rate versus number of users, with ideal CSI, $N_B = 4$, $N_U = 2$, $D_S = 0$.

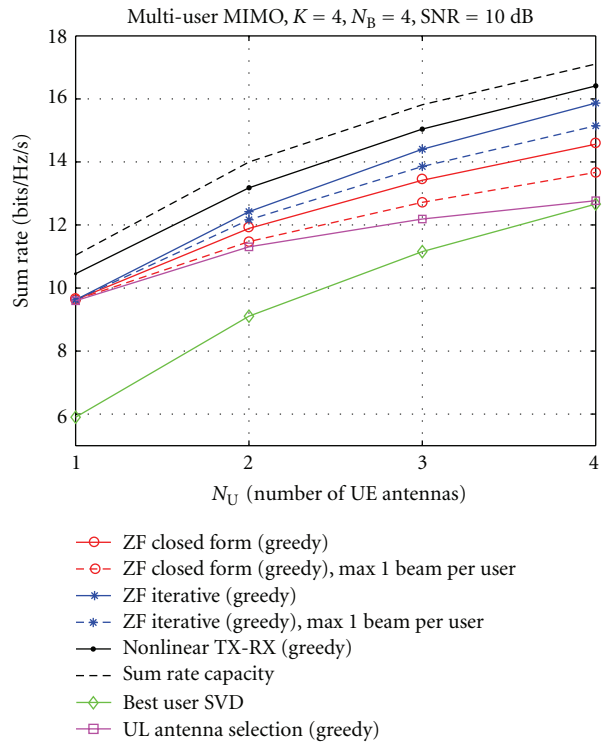


FIGURE 5: Average sum rate versus number of terminal antennas, with ideal CSI, $N_B = 4$, $N_U = 2$, $D_S = 0$.

always outperforms the closed-form solution. Furthermore, as the number of users grows, the loss from restricting the maximum number of beams per user to be one is reduced. Here the comparison curve "nonlinear TX-RX" refers to the capacity figures obtained by iterative waterfilling for the greedy beam allocation and with the power constraint (22). The difference to the ZF curves represents the capacity loss induced when restricting transmit-receive processing to be linear. The sum rate capacity shown in the figure is the sum rate achievable with the sum power constraint [18]. As can be seen, the single-user MIMO transmission is inefficient in the sense that it cannot utilize more than N_U out of the N_B potential spatial degrees of freedom available. On the other hand, the UL antenna selection shows competitive performance, and it benefits from multiuser diversity as much as the beamforming methods. The only difference is caused by the absence of beamforming gain.

The effect of the number of terminal antennas N_U when $K = 4$, is illustrated in Figure 5. With a higher number of antennas, all the beamforming methods benefit from the increased beamforming gain, while the advantage seen by the antenna selection is more limited. For the compared methods, CDFs of the sum rates for the special case $K = 4$ and $N_U = 2$ are depicted in Figure 6.

Figure 7 illustrates the effect of temporal fading and lag error of transmit CSI on the UL and DL schemes in a network of four users and with ZF receivers. As can be seen, DL is more sensitive to the lag error than the UL. The antenna selection is affected as well, as the selection is based on outdated observations, and the spatial compatibility of the antennas is reduced.

Figure 8 depicts the effect of noisy channel estimation in static channel for $N_B = 4$, $N_U = 2$, and $K = 4$. The achievable rates are shown versus pilot sum SNR $= N_{\text{pilot}} P_{\text{pilot}}/N_0$, where N_{pilot} is the number of pilot symbols per frame, and P_{pilot} is the total pilot power in both UL and DL. In the DL, the power is equally divided between the $\sum_k L_k$ pilot streams, while in the UL the power is divided between $K \cdot N_U$ pilot streams. The rates are averages over data fields only so that the fractional rate loss caused by the pilot overhead is not included. In Figure 8(a) the CSIR is assumed ideal so that all receivers operate on perfect channel knowledge, whereas the CSIT is noisy so that the transmit beamformers become imperfect. In the UL, the CSIT uncertainty accumulates from the estimation of both CSI sounding and the following beam allocation. For the antenna selection, the only source of error is the CSI sounding step. As can be seen, the iterative ZF method in UL outperforms the comparison schemes with any pilot SNR value. Figure 8(b) shows the accumulated effect of CSIT and CSIR uncertainty. As can be seen, the UL reception suffers more than DL from the reduced receiver performance, and the multiuser strategies suffer more than the single-user case. In the simulation setup, this is partially caused by the fact that UL pilot power has been distributed between the demodulation and additional CSI sounding pilots, which is inefficient from the receiver point of view.

In the previous figures, zero forcing receivers were assumed for all the schemes. Especially in the UL, it is

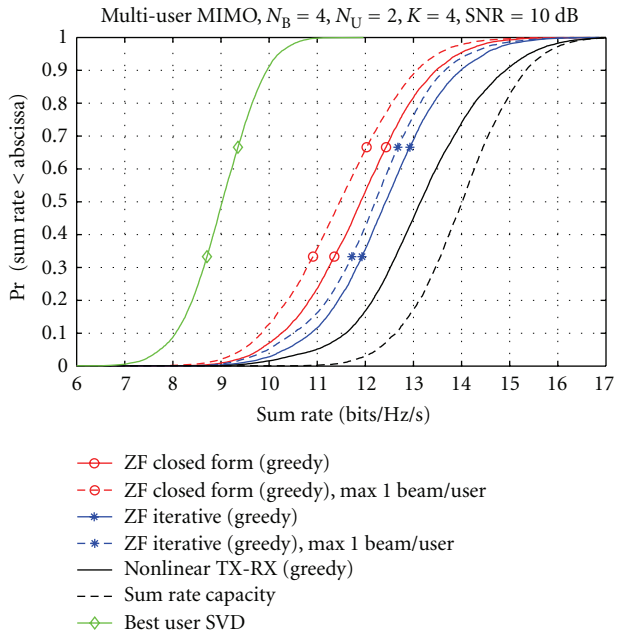


FIGURE 6: CDF of sum rate, with ideal CSI, $N_B = 4$, $N_U = 2$, $K = 4$, $D_S = 0$.

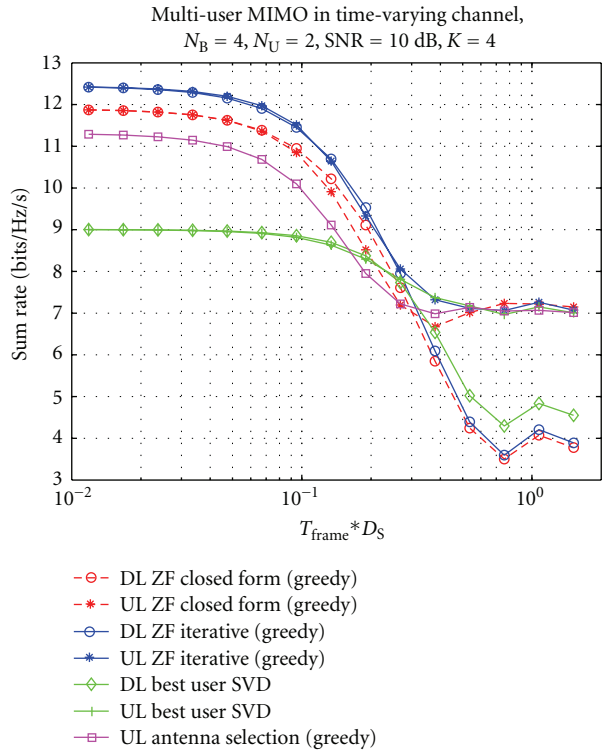


FIGURE 7: Average sum rate in time-varying channel, with noise-free CSI and ZF receivers, $N_B = 4$, $N_U = 2$, $K = 4$.

reasonable to assume that more advanced receiver structures are employed. Figure 9 compares the sum rate performance of ZF, LMMSE and optimal nonlinear receivers in the BS with perfect CSIR. As can be seen, the benefit to beamforming is minor, and to antenna selection moderate. For comparison, nonprecoded UL transmission with user selection was simulated as well. In this scenario, the BS always selects two out of four terminals with the strongest MIMO channels, to transmit two nonprecoded data streams each. As there is no control over the spatial compatibility of the transmitted signals, the significance of the receiver structure is dramatic.

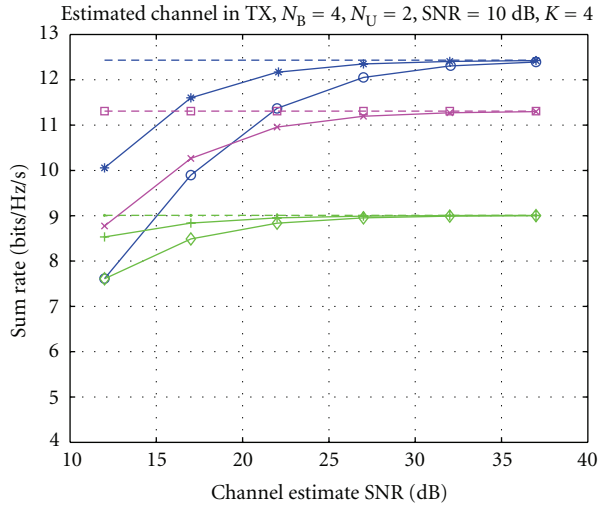
6. Conclusion

We have presented practical linear coordinated transmit-receive zero forcing schemes for the uplink of cellular multiuser MIMO systems in the TDD mode. Beam selection is an integral part of the strategy, as it helps to avoid excessive zero forcing loss while achieving gain from multiuser diversity. The BS computes the transmission parameters in a centralized manner and employs DL pilot signals to convey the information of the beam selection and beamformers to be used by the terminals. When coexisting with the DL transmit-receive zero forcing, the precoded DL demodulation pilots can be reused for UL beam allocation so that no additional pilot overhead is required. In order to reduce the UL pilot overhead as well, we proposed reusing the precoded UL demodulation pilots in turn for partial CSI sounding. As a result, only the precoded pilot symbols are needed in both UL and DL to satisfy the needs of both transmission and reception. The system is readily scalable, since any combination of base station and terminal antenna array setups can be supported.

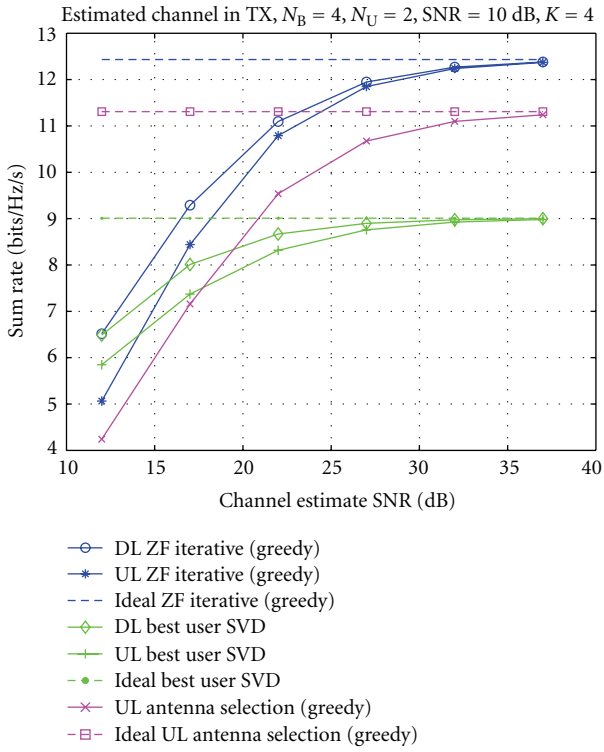
In zero forcing, the multiuser MIMO channel is decoupled into noninterfering parallel channels by linear processing. Thus, the strategy lends itself to straightforward power and rate allocation as well as coding and modulation. Furthermore, the system works well with suboptimal linear receivers that can be easily constructed based on simple CSI estimation tasks. The use of more complex nonlinear successive interference cancellers or turbo receivers is not necessary, which further increases the robustness of the system, as the possible error propagation between the users' signals is avoided.

We evaluated the performance of the strategy in time-varying fading channels and with CSI estimation. The largest gains from multiuser MIMO communication are obtained when the fading is slow, and when the quality of CSIT at the BS is good. It is worth noting that UL beamforming is not sensitive to the quality of CSIT at the terminals, and even the simple antenna selection transmission performs adequately in multiuser environments. Obviously, the benefit of beamforming grows with the number of terminal antenna elements.

From the results we conclude that multistream precoding also in the UL is in practice feasible, robust and beneficial from the system capacity point of view. Due to its practical nature, the proposed concept is a promising candidate for the evolution steps of future cellular systems such as 3GPP LTE.



(a)



(b)

Figure 8: Average sum rate, with noisy CSI and ZF receivers, $N_B = 4$, $N_U = 2$, $K = 4$, $D_S = 0$: (a) estimated CSIT and ideal CSIR, (b) estimated CSIT and estimated CSIR.

The uplink-downlink beamforming concept is at its most efficient when the offered data traffic loads in both directions are approximately equal. The possible asymmetry of the traffic can be treated in time domain, for example, by allocating longer time frames to the DL than UL. In the extreme case, UL beamforming can be decoupled from the DL data transmission completely. In this case, the BS would merely arrange the UL multiuser transmission by communicating the beam selection to the terminals via DL pilots.

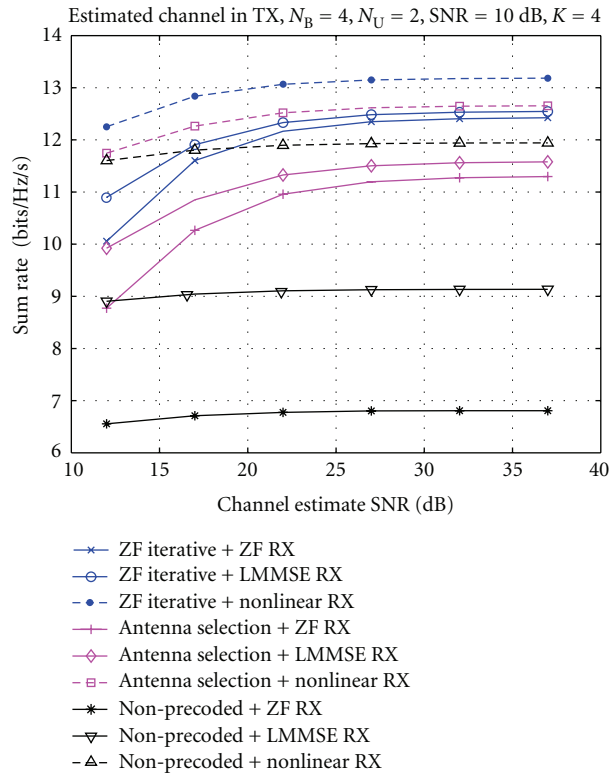


Figure 9: Uplink average sum rate, with noisy CSIT and different receivers, $N_B = 4$, $N_U = 2$, $K = 4$, $D_S = 0$.

Acknowledgments

This work has been supported by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia, Nokia Siemens Networks, Elektrobit and Tauno Tönning Foundation. This work has been performed in part in the framework of the CELTIC Project CP5-026 WINNER+. The authors would like to acknowledge the contributions of their colleagues.

References

[1] P. Viswanath and D. N. C. Tse, "Sum capacity of the vector Gaussian broadcast channel and uplink-downlink duality," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 1912–1921, 2003.

[2] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 50, no. 5, pp. 768–783, 2004.

[3] IST-4-027756 WINNER II, "D3.4.1 The WINNER II air interface: refined spatial-temporal processing solutions," October 2006.

[4] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication—part I: channel inversion and regularization," *IEEE Transactions on Communications*, vol. 53, no. 1, pp. 195–202, 2005.

[5] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser MIMO channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[6] K.-K. Wong, R. D. Murch, and K. B. Letaief, "A joint-channel diagonalization for multiuser MIMO antenna systems," *IEEE Transactions on Wireless Communications*, vol. 2, no. 4, pp. 773–786, 2003.

[7] B. Farhang-Boroujeny, Q. Spencer, and L. Swindlehurst, "Layering techniques for space-time communication in multi-user networks," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '03)*, vol. 2, pp. 1339–1343, Orlando, Fla, USA, October 2003.

[8] A. Tölli, M. Codreanu, and M. Juntti, "Cooperative MIMO-OFDM cellular system with soft handover between distributed base station antennas," *IEEE Transactions on Wireless Communications*, vol. 7, no. 4, pp. 1428–1440, 2008.

[9] C.-B. Chae, D. Mazzarese, N. Jindal, and R. W. Heath Jr., "Coordinated beamforming with limited feedback in the MIMO broadcast channel," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 8, pp. 1505–1515, 2008.

[10] C.-B. Chae, S. Kim, and R. W. Heath Jr., "Linear network coordinated beamforming for cell-boundary users," in *Proceedings of IEEE Workshop on Signal Processing Advances in Wireless Communications*, pp. 534–538, Perugia, Italy, June 2009.

[11] G. Dimic and N. D. Sidiropoulos, "On downlink beamforming with greedy user selection: performance analysis and a simple new algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 10, pp. 3857–3868, 2005.

[12] A. Tölli and M. Juntti, "Scheduling for multiuser MIMO downlink with linear processing," in *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, pp. 156–160, Berlin, Germany, September 2005.

[13] F. Boccardi and H. Huang, "A near-optimum technique using linear precoding for the MIMO broadcast channel," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 3, pp. 17–20, Honolulu, Hawaii, USA, April 2007.

[14] G. Lebrun, J. Gao, and M. Faulkner, "MIMO transmission over a time-varying channel using SVD," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 757–764, 2005.

[15] K. Zhang and Z. Niu, "MIMO broadcast transmission with outdated channel state information," in *Proceedings of Asia-Pacific Conference on Communications (APCC '06)*, pp. 1–5, Buson, Korea, August 2006.

[16] D. Samardzija and N. Mandayam, "Impact of pilot design on achievable data rates in multiple antenna multiuser TDD systems," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 7, pp. 1370–1379, 2007.

[17] P. Komulainen, M. Latva-Aho, and M. Juntti, "Block diagonalization for multiuser MIMO TDD downlink and uplink in time-varying channel," in *Proceedings of International ITG Workshop on Smart Antennas*, pp. 74–81, Darmstadt, Germany, February 2008.

[18] N. Jindal, W. Rhee, S. Vishwanath, S. Jafar, and A. Goldsmith, "Sum power iterative water-filling for multi-antenna Gaussian broadcast channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1570–1580, 2005.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 739828, 11 pages
doi:10.1155/2009/739828

Research Article

Antenna Selection for MIMO Systems with Closely Spaced Antennas

Yang Yang,[1] Rick S. Blum,[1] and Sana Sfar[2]

[1] *Department of Electrical and Computer Engineering, Lehigh University, 19 Memorial Drive West, Bethlehem, PA 18015, USA*
[2] *CTO Office, InterDigital Communications, LLC, 781 Third Avenues, King of Prussia, PA 19406, USA*

Correspondence should be addressed to Yang Yang, yay204@lehigh.edu

Received 1 February 2009; Revised 18 May 2009; Accepted 28 June 2009

Recommended by Angel Lozano

Physical size limitations in user equipment may force multiple antennas to be spaced closely, and this generates a considerable amount of mutual coupling between antenna elements whose effect cannot be neglected. Thus, the design and deployment of antenna selection schemes appropriate for next generation wireless standards such as 3GPP long term evolution (LTE) and LTE advanced needs to take these practical implementation issues into account. In this paper, we consider multiple-input multipleoutput (MIMO) systems where antenna elements are placed side by side in a limited-size linear array, and we examine the performance of some typical antenna selection approaches in such systems and under various scenarios of antenna spacing and mutual coupling. These antenna selection schemes range from the conventional hard selection method where only part of the antennas are active, to some newly proposed methods where all the antennas are used, which are categorized as soft selection. For the cases we consider, our results indicate that, given the presence of mutual coupling, soft selection can always achieve superior performance as compared to hard selection, and the interelement spacing is closely related to the effectiveness of antenna selection. Our work further reveals that, when the effect of mutual coupling is concerned, it is still possible to achieve better spectral efficiency by placing a few more than necessary antenna elements in user equipment and applying an appropriate antenna selection approach than plainly implementing the conventional MIMO system without antenna selection.

1. Introduction

The multiple-input multiple-output (MIMO) architecture has been demonstrated to be an effective means to boost the capacity of wireless communication systems [1], and has evolved to become an inherent component of various wireless standards, including the next-generation cellular systems 3GPP long term evolution (LTE) and LTE advanced. For example, the use of a MIMO scheme was proposed in the LTE standard, with possibly up to four antennas at the mobile side, and four antennas at the cell site [2]. In MIMO systems, antenna arrays can be exploited in two different ways, which are [3]: diversity transmission and spatial multiplexing. However, in either case, one main problem involved in the implementation of MIMO systems is the increased complexity, and thus the cost. Even though the cost for additional antenna elements is minimal, the radio frequency (RF) elements required by each antenna, which perform the microwave/baseband frequency translation, analog-to-digital conversion, and so forth, are usually costly.

These complexity and cost concerns with MIMO have motivated the recent popularity of antenna selection (AS)—an attractive technique which can alleviate the hardware complexity, and at the same time capture most of the advantages of MIMO systems. In fact, for its low user equipment (UE) complexity, AS (transmit) is currently being considered as a baseline of the single-user transmit diversity techniques in the LTE uplink which is a MIMO single carrier frequency division multiple access (SC-FDMA) system [4]. Further, when it comes to the RF processing manner, AS can be categorized into two groups: (1) hard selection, where only part of the antennas are active and the selection is implemented in the RF domain by means of a set of switches (e.g., [5–7]); (2) soft selection, where all the antennas are active and a certain form of transformation is performed

in the RF domain upon the received signals across all the antennas (e.g., [8–10]).

A considerable amount of research efforts have been dedicated to the investigation of AS, and have solidly demonstrated the theoretical benefits of AS (see [3] for a tutorial treatment). However, previous works largely ignore the hardware implementation issues related to AS. For instance, the physical size of UE such as mobile terminals and mobile personal assistants, are usually small and invariable, and the space allocated for an antenna array is limited. Such limitation makes the close spacing between antenna elements a necessity, inevitably leading to mutual coupling [11], and correlated signals. These issues have caught the interest of some researchers, and the capacity of conventional MIMO systems (without AS) under the described limitations and circumstances was investigated, among others, in [12–17]. To give an example, the study in [12] shows that as the number of receive antenna elements increases in a fixed-length array, the system capacity firstly increases to saturate shortly after the mutual coupling reaches a certain level of severeness; and drops after that.

Form factors of UE limit the performance promised by MIMO systems, and can further affect the proper functionality of AS schemes. These practical implementation issues merit our attention when designing and deploying AS schemes for the 3GPP LTE and LTE advanced technologies. There exist some interesting works, such as [18, 19] which consider AS in size sensitive wireless devices to improve the system performance. But in general, results, conclusions, and ideas on the critical implementation aspects of AS in MIMO systems still remain fragmented. In this paper, through electromagnetic modeling of the antenna array and theoretical analysis, we propose a comprehensive study of the performance of AS, to seek more effective implementation of AS in size sensitive UE employing MIMO where both mutual coupling and spatial correlation have a strong impact. In this process, besides the hybrid selection [5–7], a conventional yet popular hard AS approach, we are particularly interested in examining the performance potential of some typical soft AS schemes, including the FFT-based selection [9] and the phase-shift-based selection [10], that are very appealing but seem not to have attracted much attention so far. At the meantime, we also intend to identify the operational regimes of these representative AS schemes in the compact antenna array MIMO system. For the cases we consider, we find that in the presence of mutual coupling, soft AS can always achieve superior performance as compared to hard AS. Moreover, effectiveness of these AS schemes is closely related to the interelement spacing. For example, hard AS works well only when the interelement spacing is no less than a half wavelength.

Additionally, another goal of our study is to address a simple yet very practical question which deals with the cost-performance tradeoff in implementation: as far as mutual coupling is concerned, can we achieve better spectral efficiency by placing a few more than necessary antenna elements in size sensitive UE and applying a certain adequate AS approach than plainly implementing the conventional MIMO system without AS? Further, if the answer is yes, how would we decide the number of antenna elements for placement and the AS method for deployment? Our work will provide answers to the above questions, and it turns out the solution is closely related to identifying the saturation point of the spectral efficiency.

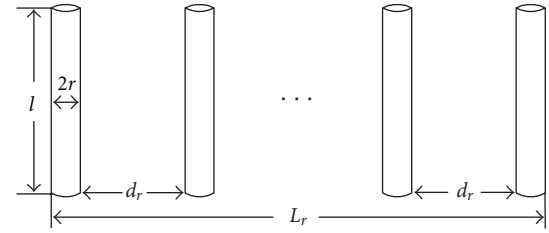


Figure 1: Dipole elements in a side-by-side configuration (receiver antenna array as an example).

This paper is organized as follows. In Section 2, we introduce the network model for the compact MIMO system and characterize the input-output relationship by taking into account the influence of mutual coupling. In Section 3, we describe the hard and soft AS schemes that will be used in our study, and also estimate their computational complexity. In Section 4, we present the simulation results. We discuss our main findings in Section 5, and finally conclude this paper in Section 6.

2. Network Model for Compact MIMO

We consider a MIMO system with M transmit and N receive antennas (M, $N > 1$). We assume antenna elements are placed in a side-by-side configuration along a fixed length at each terminal (transmitter and receiver), as shown in Figure 1. Other types of antenna configuration are also possible, for example, circular arrays [11]. But it is noted that, the side-by-side arrangement exhibits larger mutual coupling effects since the antennas are placed in the direction of maximum radiation [11, page 474]. Thus, the side-by-side configuration is more suitable to our study. We define L_t and L_r as the aperture lengths for transmitter and receiver sides, respectively. In particular, we are more interested in the case that L_r is fixed and small, which corresponds to the space limitation of the UE. We denote l as the dipole length, r as the dipole radius, and d_r (d_t) as the side-by-side distance between the adjacent dipoles at the receiver (transmitter) side. Thus, we have $d_r = L_r/(N-1)$ and $d_t = L_t/(M-1)$.

A simplified network model (as compared to [13, 14], e.g.) for transmitter and receiver sides is depicted in Figure 2. Figure 3 illustrates a direct conversion receiver that connects the output signals in Figure 2, where LNA denotes the low-noise amplifier, LO denotes the local oscillator, and ADC denotes the analog-to-digital converter. For the ease of the following analysis, we assume that in the circuit setup, all the antenna elements at the receiver side are grounded through the load impedance Z_{Li}, $i = 1, \ldots, N$ (cf. Figure 2), regardless of whether they will be selected or not. In fact, $Z_{L_i}, i = 1, \ldots, N$ constitute a simple matching circuit. Such matching circuit is necessary as it can enhance the efficiency of power

transfer from the generator to the load [20, Chapter 11]. We also assume that the input impedance of each LNA in Figure 3 which is located very close to the antenna element to amplify weak received signals, is high enough such that it has little measurable effect on the receive array's output voltages. This assumption is necessary to facilitate the analysis of the network model. However, it is also very reasonable because this ensures that the input of the amplifier will neither overload the source of the signal nor reduce the strength of the signal by a substantial amount [21].

Let us firstly consider the transmitter side, which can be regarded as a coupled M port network with M terminals. We define $\mathbf{i} = [i_1, \ldots, i_M]^T$ and $\mathbf{v}_t = [v_{t1}, \ldots, v_{tM}]^T$ as the vectors of terminal currents and voltages, respectively, and they are related through

$$\mathbf{v}_t = \mathbf{Z}_T \mathbf{i}, \tag{1}$$

where $\mathbf{Z}_T$ denotes the impedance matrix at the transmitter side. The (p,q)-th entry of $\mathbf{Z}_T(p,q)$, when $p \neq q$, denotes the mutual impedance between two antenna elements, and is given by [20, Chapter 21.2]:

$$\mathbf{Z}_T(p,q) = \frac{j\eta}{4\pi \sin^2(kl/2)} \int_{-l/2}^{l/2} F(z)\, dz, \tag{2}$$

where

$$F(z) = \left[\frac{e^{-jkR_1}}{R_1} + \frac{e^{-jkR_2}}{R_2} - 2\cos\left(\frac{kl}{2}\right) \cdot \frac{e^{-jkR_0}}{R_0} \right] \cdot \sin\left[k\left(\frac{l}{2} - |z| \right) \right]. \tag{3}$$

In the above expression, η denotes the characteristic impedance of the propagation medium, and can be calculated by $\eta = \sqrt{\mu/\epsilon}$, where μ and ϵ denote permittivity and permeability of the medium, respectively. Likewise, k denotes the propagation wavenumber of an electromagnetic wave propagating in a dielectric conducting medium, and can be computed through $k = \omega\sqrt{\mu\epsilon}$, where ω is the angular frequency. Finally R_0, R_1 and R_2 are defined as

$$R_0 = \sqrt{\frac{(p-q)^2 d_t^2}{(M-1)^2} + z^2},$$
$$R_1 = \sqrt{\frac{(p-q)^2 d_t^2}{(M-1)^2} + \left(z - \frac{l}{2}\right)^2}, \tag{4}$$
$$R_2 = \sqrt{\frac{(p-q)^2 d_t^2}{(M-1)^2} + \left(z + \frac{l}{2}\right)^2}.$$

When $p = q$, $\mathbf{Z}_T(p,q)$ is the self-impedance of a single antenna element, and can also be obtained from (2) by simply redefining R_0, R_1 and R_2 as follows:

$$R_0 = \sqrt{r^2 + z^2},$$
$$R_1 = \sqrt{r^2 + \left(z - \frac{l}{2}\right)^2}, \tag{5}$$
$$R_2 = \sqrt{r^2 + \left(z + \frac{l}{2}\right)^2}.$$

Thus, the self-impedance for an antenna element with $l = 0.5\lambda$ and $r = 0.001\lambda$ for example, is approximately

$$\mathbf{Z}_T(p,p) = 73.08 + 42.21j\,\Omega. \tag{6}$$

Further, let us consider an example that $M = 5$ antenna elements of such type are equally spaced over a linear array of length $L_t = 2\lambda$. The impedance matrix $\mathbf{Z}_T$ is given by

$$\mathbf{Z}_T = \begin{pmatrix} 73.08+42.21j & -12.52-29.91j & 4.01+17.73j & -1.89-12.30j & 1.08+9.36j \\ -12.52-29.91j & 73.08+42.21j & -12.52-29.91j & 4.01+17.73j & -1.89-12.30j \\ 4.01+17.73j & -12.52-29.91j & 73.08+42.21j & -12.52-29.91j & 4.01+17.73j \\ -1.89-12.30j & 4.01+17.73j & -12.52-29.91j & 73.08+42.21j & -12.52-29.91j \\ 1.08+9.36j & -1.89-12.30j & 4.01+17.73j & -12.52-29.91j & 73.08+42.21j \end{pmatrix}. \tag{7}$$

For $i = 1, \ldots, M$, the terminal voltage v_{ti} can be related to the source voltage x_i via the source impedance Z_{si} by $v_{ti} = x_i - Z_{si} i_i$. Define $\mathbf{Z}_S = \mathrm{diag}\{Z_{s1}, \ldots, Z_{sM}\}$, and $\mathbf{x} = [x_1, \ldots, x_M]$. Then, from Figure 2, we can obtain the following results: $\mathbf{v}_t = \mathbf{x} - \mathbf{Z}_S \mathbf{i}$ and $\mathbf{v}_t = \mathbf{Z}_T \mathbf{i}$. Therefore, the relationship between terminal voltages $\mathbf{v}_t$ and source voltages $\mathbf{x}$ can be written in matrix form as $\mathbf{v}_t = \mathbf{Z}_T(\mathbf{Z}_T + \mathbf{Z}_S)^{-1}\mathbf{x}$. Similar to [12], we choose $Z_{si} = \mathbf{Z}_T^*(i,i)$, which roughly corresponds to a conjugate match in the presence of mild coupling. In the case of uncoupling in the transmitter side, $\mathbf{Z}_T$ is diagonal, and its diagonal elements are all the same. Consequently, $\mathbf{Z}_T(\mathbf{Z}_T + \mathbf{Z}_S)^{-1}$ is also diagonal, and its diagonal element can be denoted as $\delta_T = \mathbf{Z}_T(1,1)/[\mathbf{Z}_T(1,1) + \mathbf{Z}_S(1,1)]$. To accommodate the special case of zero mutual coupling where $\mathbf{v}_t$ is equal to $\mathbf{x}$, in our model we modify the relationship between $\mathbf{v}_t$ and $\mathbf{x}$ into

$$\mathbf{v}_t = \mathbf{W}_T \mathbf{x}, \tag{8}$$

where $\mathbf{W}_T = \delta_T^{-1} \mathbf{Z}_T(\mathbf{Z}_T + \mathbf{Z}_S)^{-1}$.

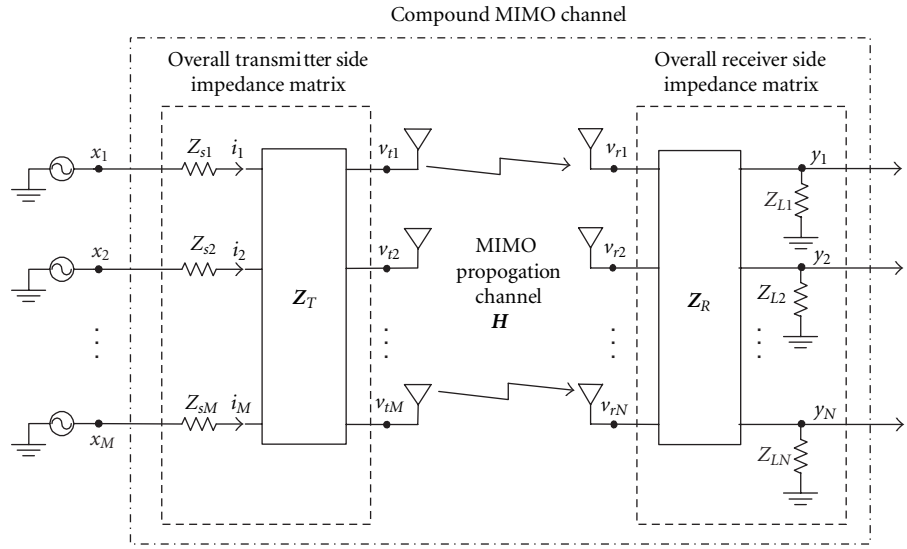


Figure 2: Network model for a (M, N) compact MIMO system.

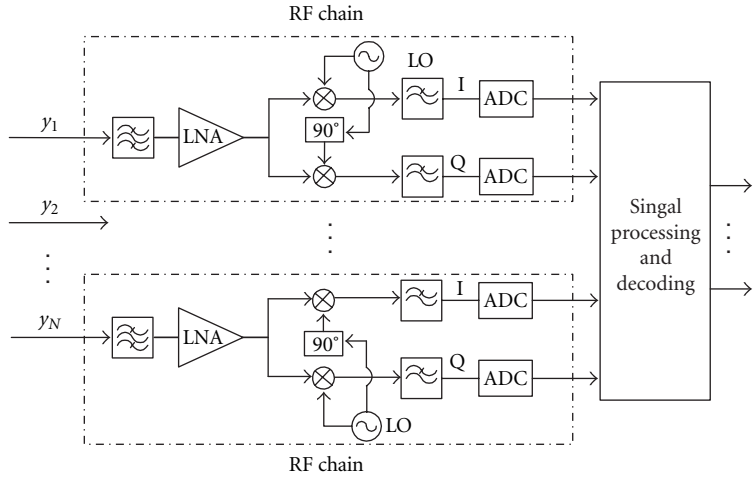


Figure 3: RF chains at the receiver side.

Denote $\mathbf{v}_r = [v_{r1}, \ldots, v_{rN}]$ as the vector of open circuited voltages induced across the receiver side antenna array, and $\mathbf{y} = [y_1, \ldots, y_N]$ as the voltage vector across the output of the receive array. Since we assumed high-input impedance of these LNAs, a similar network analysis can be carried out at the receiver side and will yield

$$\mathbf{y} = \mathbf{W}_R \mathbf{v}_r, \tag{9}$$

where $\mathbf{W}_R = \delta_R^{-1} \mathbf{Z}_L (\mathbf{Z}_R + \mathbf{Z}_L)^{-1}$. $\mathbf{Z}_R$ is the mutual impedance matrix at the receiver side, and $\mathbf{Z}_L$ is a diagonal matrix with its (i,i)th entry given by $\mathbf{Z}_L(i,i) = Z_{Li} = [\mathbf{Z}_R(i,i)]^*$, $i = 1, \ldots, N$. δ_R is given by $\delta_R = [\mathbf{Z}_R(1,1)]^* / \{\mathbf{Z}_R(1,1) + [\mathbf{Z}_R(1,1)]^*\}$. It is noted that the approximate conjugate match [12] is also assumed at the receiver side, so that the load impedance matrix $\mathbf{Z}_L$ is diagonal with its entry given by $\mathbf{Z}_R^*(i,i)$, for $i = 1, \ldots, N$.

In frequency-selective fading channels, the effectiveness of AS is considerably reduced [3], which in turn makes it difficult to observe the effect of mutual coupling. Therefore, we focus our attention solely on flat fading MIMO channels. The radiated signal $\mathbf{v}_t$ is related to the received signal $\mathbf{v}_r$ through

$$\mathbf{v}_r = \mathbf{H} \mathbf{v}_t, \tag{10}$$

where $\mathbf{H}$ is a $N \times M$ complex Gaussian matrix with correlated entries. To account for the spatial correlation effect and the Rayleigh fading, we adopt the *Kronecker model* [22, 23]. This model uses an assumption that the correlation matrix, obtained as $\boldsymbol{\Psi} = \mathrm{E}\{\mathrm{vec}(\mathbf{H})\ \mathrm{vec}(\mathbf{H})^H\}$ with $\mathrm{vec}(\mathbf{H})$ being the operator stacking the matrix $\mathbf{H}$ into a vector columnwise, can be written as a Kronecker product, that is, $\boldsymbol{\Psi} = \boldsymbol{\Psi}_R \otimes \boldsymbol{\Psi}_T$, where $\boldsymbol{\Psi}_R$ and $\boldsymbol{\Psi}_T$ are respectively, the receive and transmit correlation matrices, and $\otimes$ denotes the Kronecker product. This implies that the joint transmit and receive angle power spectrum can be written as a product of two independent

angle power spectrum at the transmitter and receiver. Thus, the correlated channel matrix $\mathbf{H}$ can be expressed as

$$\mathbf{H} = \mathbf{\Psi}_R^{1/2}\mathbf{H}_w\mathbf{\Psi}_T^{1/2}, \tag{11}$$

where $\mathbf{H}_w$ is a $N \times M$ matrix whose entries are independent identically distributed (i.i.d) circular symmetric complex Gaussian random variables with zero mean and unit variance. The (i, j)-th entry of $\mathbf{\Psi}_R$ or $\mathbf{\Psi}_T$ is given by $J_0(2\pi d_{ij}/\lambda)$ [24], where J_0 is the zeroth order Bessel function of the first kind, and d_{ij} denotes the distance between the i,j-th antenna elements.

Therefore, based on (8)–(11), the output signal vector $\mathbf{y}$ at the receiver can be expressed in terms of the input signal $\mathbf{x}$ at the transmitter through

$$\mathbf{y} = \mathbf{W}_R\mathbf{\Psi}_R^{1/2}\mathbf{H}_w\mathbf{\Psi}_T^{1/2}\mathbf{W}_T\mathbf{x} + \mathbf{n} = \mathcal{H}\mathbf{x} + \mathbf{n}, \tag{12}$$

where $\mathcal{H} = \mathbf{W}_R\mathbf{\Psi}_R^{1/2}\mathbf{H}_w\mathbf{\Psi}_T^{1/2}\mathbf{W}_T$ can be regarded as a *compound channel matrix* which takes into account both the Rayleigh fading in wireless channels and the mutual coupling effect at both transmitter and receiver sides, and $\mathbf{n}$ is the thermal noise. For simplicity, we assume uncorrelated noise at the receiving antenna element ports. For the case where correlated noise is considered, readers are referred to [16, 17].

3. Hard and Soft AS for Compact MIMO

We describe here some typical hard and soft AS schemes that we will investigate, assuming the compact antenna array MIMO system described in Section 2. For hard AS, we focus only on the hybrid selection method [5–7]. For soft AS, we study two typical schemes: the FFT-based selection [9] which embeds fast Fourier transform (FFT) operations in the RF chains, and the phase-shift-based selection [10] which uses variable phase shifters adapted to the channel coefficients in the RF chains. For simplicity, we only consider AS at the receiver side with n_R antennas being chosen out of the N available ones, and we focus on a spatial multiplexing transmission.

We assume that the propagation channel is flat fading and quasistatic, and is known at the receiver. We also assume that the power is uniformly allocated across all the M transmit antennas, that is, $\mathrm{E}\{\mathbf{x}\mathbf{x}^H\} = P_0\mathbf{I}_M/M$. We denote the noise power as σ_n^2, and the nominal signal-to-noise ratio (SNR) as $\rho = P_0/\sigma_n^2$. Then assuming some codes that approach the Shannon limit quite closely are used, the spectral efficiency (in bits/s/Hz) of this (M, N) full-complexity (FC) compact MIMO system without AS could be calculated through [1]

$$C_{\mathrm{FC}}(M,N) = \log_2\left\{\det\left[\mathbf{I}_M + \frac{\rho}{M}\mathcal{H}^H\mathcal{H}\right]\right\}. \tag{13}$$

It is worth noting that the length limits of transmit and receive arrays, L_t and L_r, enter into the compound channel matrix $\mathcal{H}$ in a very complicated way. It is thus difficult to find a close-form analytical relationship between $C_{\mathrm{FC}}(M, N)$ and L_t (L_r). Consequently, using Monte Carlo simulations to evaluate the performance of spectral efficiency becomes a necessity.

To avoid detailed system configurations and to make the performance comparison as general and as consistent as possible, we only use the spectral efficiency as the performance of interest. Moreover, all these AS schemes we study here are merely to optimize the spectral efficiency, not other metrics. Since each channel realization renders a spectral efficiency value, the ergodic spectral efficiency and the cumulative distribution function (CDF) of the spectral efficiency will be both meaningful. We will then consider them as performance measures for our study.

3.1. Hybrid Selection. This selection scheme belongs to the conventional hard selection, where n_R out of N receive antennas are chosen by means of a set of switches in the RF domain (e.g., [5–7]). Figure 4(a) illustrates the architecture of the hybrid selection at the receiver side. As all the antenna elements at the receiver side are presumed grounded through the load impedance Z_{Li}, $i = 1, \ldots, N$, the mutual coupling effect will be always present at the receiver side. However, this can facilitate the channel estimation and allow us to extract rows from $\mathcal{H}$ for subset selection. Otherwise, the mutual coupling effect will vary with respect to the selected antenna subsets. For convenience, we define $\mathbf{S}$ as the $n_R \times N$ selection matrix, which extracts n_R rows from $\mathcal{H}$ that are associated with the selected subset of antennas. We further define $\mathcal{S}$ as the collection of all possible selection matrices, whose cardinality is given by $|\mathcal{S}| = \binom{N}{n_R}$. Thus, the system with hybrid selection delivers a spectral efficiency of

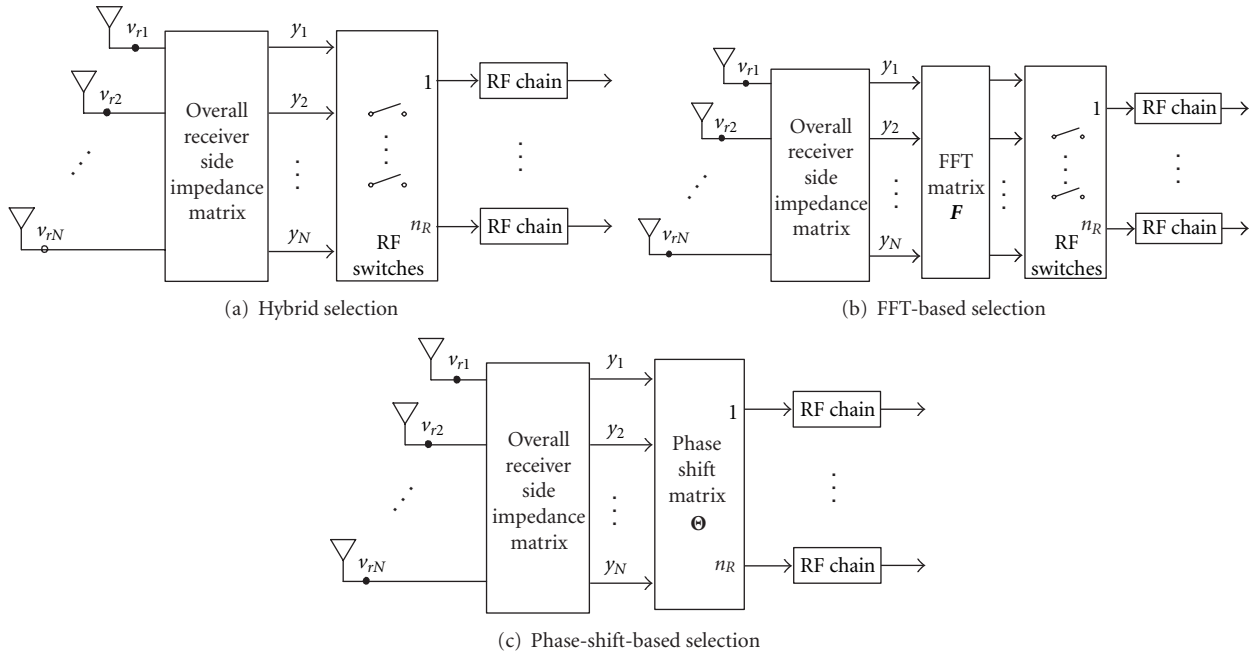
$$C_{\mathrm{HS}} = \max_{\mathbf{S}\in\mathcal{S}} \log_2\left\{\det\left[\mathbf{I}_M + \frac{\rho}{M}(\mathbf{S}\mathcal{H})^H(\mathbf{S}\mathcal{H})\right]\right\}. \tag{14}$$

Optimal selection that leads to C_{HS} requires an exhaustive search over all $\binom{N}{n_R}$ subsets of $\mathcal{S}$, which is evident by (14). Note that

$$\det\left[\mathbf{I}_M + \frac{\rho}{M}(\mathbf{S}\mathcal{H})^H(\mathbf{S}\mathcal{H})\right] = \det\left[\mathbf{I}_{n_R} + \frac{\rho}{M}(\mathbf{S}\mathcal{H})(\mathbf{S}\mathcal{H})^H\right]. \tag{15}$$

Then, the matrix multiplication in (14) has a complexity of $O(n_R M \cdot \min(n_R, M))$. Calculating the matrix determinant in (14) requires a complexity of $O((\min(n_R, M))^3)$. Thus, we can conclude that optimal selection requires about $O(|\mathcal{S}| \cdot n_R M \cdot \min(n_R, M))$ complex additions/multiplications. This estimated complexity for optimal selection can be deemed as an upper bound of the complexity of any hybrid AS scheme, since there exist some suboptimal but reduced complexity algorithms, such as the incremental selection and the decremental selection algorithms in [7].

3.2. FFT-based Selection. As for this soft selection scheme (e.g., [9]), a N-point FFT transformation (phase-shift only) is performed in the RF domain firstly, as shown in Figure 4(b), where information across all the receive antennas will be utilized. After that, a hybrid-selection-like scheme is applied to extract n_R out of N information streams.



(a) Hybrid selection

(b) FFT-based selection

(c) Phase-shift-based selection

Figure 4: AS at the receiver side for spatial multiplexing transimssions.

We denote $\mathbf{F}$ as the $N \times N$ unitary FFT matrix with its (k,l) th entry given by:

$$\mathbf{F}(k,l) = \frac{1}{\sqrt{N}} \exp\left\{\frac{-j2\pi(k-1)(l-1)}{N}\right\}, \quad \forall k,l \in [1,N]. \tag{16}$$

Accordingly, this system delivers a spectral efficiency of

$$C_{\text{FFTS}} = \max_{\mathbf{S}\in\mathcal{S}} \log_2\left\{\det\left[\mathbf{I}_M + \frac{\rho}{M}(\mathbf{SF}\mathcal{H})^H(\mathbf{SF}\mathcal{H})\right]\right\}. \tag{17}$$

The only difference between (14) and (17) is the N-point FFT transformation. Such FFT transformation requires a computational complexity of $O(MN\log N)$. If we assume $N\log N \leq n_R \cdot \min(n_R, M)$, then the computational complexity of optimal selection that achieves C_{FFTS} can be estimated as $O(|\mathcal{S}| \cdot n_R M \cdot \min(n_R, M))$, which is the worst-case complexity.

3.3. Phase-Shift Based Selection. This is another type of soft selection scheme (e.g., [10]) that we consider throughout this study. Its architecture is illustrated in Figure 4(c). Let us denote $\boldsymbol{\Theta}$ as one $n_R \times N$ matrix whose elements are nonzero and restricted to be pure phase-shifters, that we will fully define in what follows. There exists some other work such as [25] that also considers the use of tunable phase shifters to increase the total capacity of MIMO systems. However, in Figure 4(c), the matrix $\boldsymbol{\Theta}$ that performs phase-shift implementation in the RF domain essentially serves as a N-to-n_R switch with n_R output streams. Additionally, unlike the FFT matrix, $\boldsymbol{\Theta}$ might not be unitary, and hence the resulting noise can be colored. Finally, this system's spectral efficiency can be calculated by [10]

$$C_{\text{PSS}} = \max_{\boldsymbol{\Theta}} \log_2\left\{\det\left[\mathbf{I}_M + \frac{\rho}{M}(\boldsymbol{\Theta}\mathcal{H})^H\left(\boldsymbol{\Theta}\boldsymbol{\Theta}^H\right)^{-1}(\boldsymbol{\Theta}\mathcal{H})\right]\right\}. \tag{18}$$

Let us define the singular value decomposition (SVD) of $\mathcal{H}$ as $\mathcal{H} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^H$, where $\mathbf{U}$ and $\mathbf{V}$ are $N \times N$, $M \times M$ unitary matrices representing the left and right singular vector spaces of $\mathcal{H}$, respectively; $\boldsymbol{\Lambda}$ is a nonnegative and diagonal matrix, consisting of all the singular values of $\mathcal{H}$. In particular, we denote $\lambda_{\mathcal{H},i}$ as the ith largest singular value of $\mathcal{H}$, and $\mathbf{u}_{\mathcal{H},i}$ as the left singular vector of $\mathcal{H}$ associated with $\lambda_{\mathcal{H},i}$. Thus one solution to the phase shift matrix $\boldsymbol{\Theta}$ can be expressed as [10, Theorem 2]:

$$\boldsymbol{\Theta} = \exp\left\{j \times \text{angle}\left\{\left[\mathbf{u}_{\mathcal{H},1}, \ldots, \mathbf{u}_{\mathcal{H},n_R}\right]^H\right\}\right\} \tag{19}$$

where angle$\{\cdot\}$ gives the phase angles, in radians, of a matrix with complex elements, exp$\{\cdot\}$ denotes the element-by-element exponential of a matrix.

The overall cost for calculating the SVD of $\mathcal{H}$ is around $O(MN \cdot \min(M,N))$ [26, Lecture 31]. Computing the matrix multiplication in (19) requires a complexity around the order of $O(MNn_R)$. The matrix determinant has an order of complexity of $O((\min(n_R, M))^3)$. Therefore, the phase-shift-based selection requires around $O(MN \cdot \max(n_R, M))$ complex additions/multiplications.

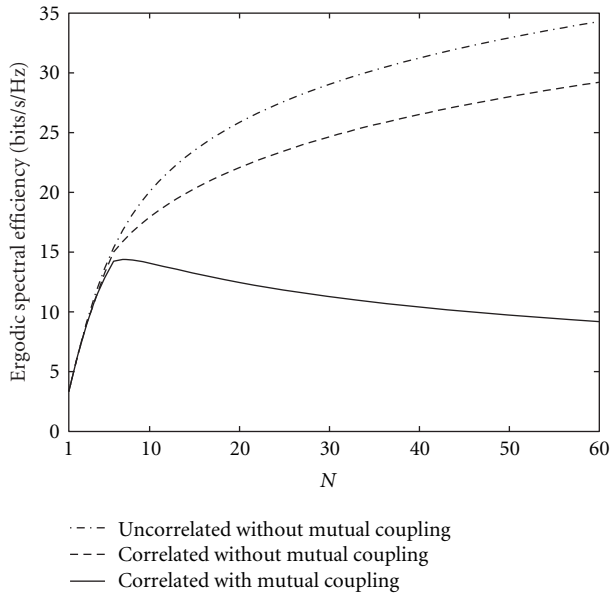


Figure 5: Ergodic spectral efficiency of a compact MIMO system ($M = 5$) with mutual coupling at both transmitter and receiver sides.

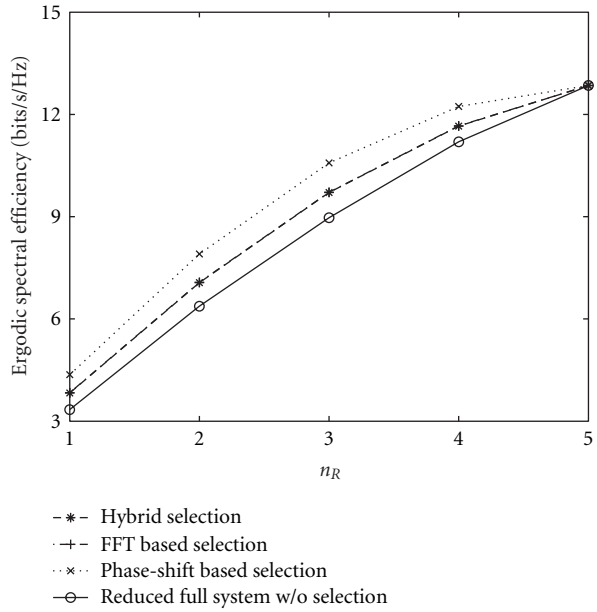


Figure 6: Ergodic spectral efficiency of a compact MIMO system with AS, where $M = 5$ and $N = 5$.

4. Simulations

Our simulations focus on the case when AS is implemented only on the receiver side, but mutual coupling and spatial correlation are accounted for at both terminals. However, in order to examine the mutual coupling effect on AS at the receiving antenna array, we further assume $M = 5$ equally-spaced antennas at the transmitter array, and the interelement spacing d_t is fixed at 10λ. This large spacing is chosen to make the mutual coupling effect negligible at the transmitting terminal. For the receiver terminal, we fix the array length L_r at 2λ. We choose $l = 0.5\lambda$ and $r = 0.001\lambda$ for all the dipole elements. Each component in the impedance matrices $\mathbf{Z}_T$ and $\mathbf{Z}_R$ is computed through (2) which analytically expresses the self and mutual impedance of dipole elements in a side-by-side configuration. Finally, we fix the nominal SNR at $\rho = 10$ dB.

As algorithm efficiency is not a focus in this paper, for both hybrid and FFT-based selection methods, we use the exhaustive search approach to find the best antenna subset. For the phase-shift-based selection, we compute the phase shift matrix $\mathbf{\Theta}$ through (19) given each $\mathcal{H}$. For each scenario of interest, we generate 5×10^4 random channel realizations, and study the performance in terms of the ergodic spectral efficiency and the CDF of the spectral efficiency.

4.1. Ergodic Spectral Efficiency of Compact MIMO. In Figure 5 we plot the ergodic spectral efficiency of a compact MIMO system for various N. The solid line in Figure 5 depicts the ergodic spectral efficiency when mutual coupling and spatial correlation is considered at both terminals. Also for the purpose of comparison, we include a dashed line which denotes the performance when only spatial correlation is considered at both sides, and a dash-dot line which corresponds to the case when only the simplest i.i.d Gaussian propagation channel is assumed in the system. It is clearly seen that mutual coupling in the compact MIMO system seriously decreases the system's spectral efficiency. Moreover, in accord with the observation in [12], our results also indicate that as the number of receive antenna elements increases, the spectral efficiency will firstly increase, but after reaching the maximum value (approximately around $N = 8$ in Figure 5), further increase in N would result in a decrease of the achieved spectral efficiency. It is also worth noting that when $N = 5$, the interelement spacing at the receiver side, d_r, is equal to $\lambda/2$, which probably is the most widely adopted interelement spacing in practice. Thus, results in Figure 5 basically indicate that, by adding a few more elements and squeezing the interelement spacing down from $\lambda/2$, it is possible to achieve some increase in the spectral efficiency, even in the presence of mutual coupling. But it is also observed that such increase is limited and relatively slow as compared to the spatial-correlated only case, and the spectral efficiency will saturate very shortly.

4.2. Ergodic Spectral Efficiency of Compact MIMO with AS. To study the performance of the ergodic spectral efficiency with regard to the number of selected antennas n_R for a compact MIMO system using AS, we consider three typical scenarios, namely, $N = 5$ in Figure 6, $N = 8$ in Figure 7, and $N = 12$ in Figure 8. In each figure, we plot the performance of the hybrid selection, FFT-based selection, and phase-shift-based selection. Additionally, we also depict in each figure the ergodic spectral efficiency of the reduced full-complexity (RFC) MIMO, denoted as $C_{\text{RFC}}(M, n_R)$, where only n_R receive antennas are distributed in the linear array and no AS is deployed. In Figure 6, it is observed that

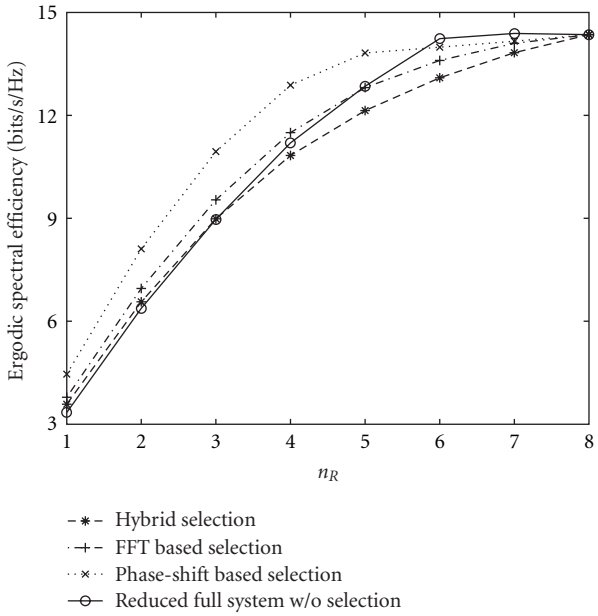


Figure 7: Ergodic spectral efficiency of a compact MIMO system with AS, where $M = 5$ and $N = 8$.

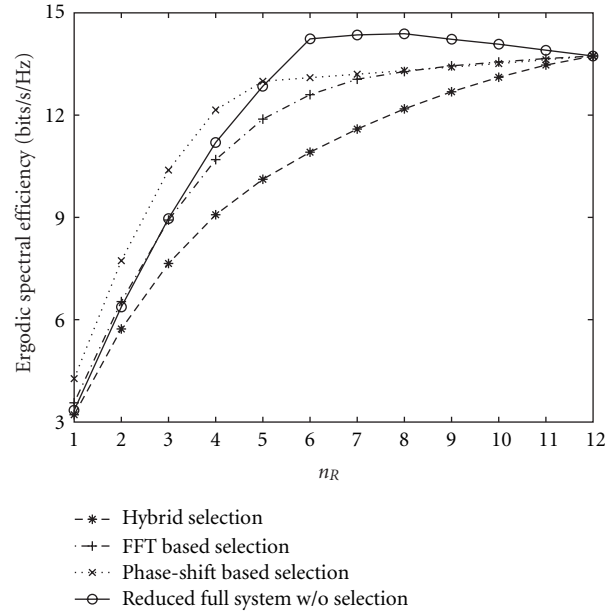


Figure 8: Ergodic spectral efficiency of a compact MIMO system with antenna selection, where $M = 5$ and $N = 12$.

$C_{\mathrm{FC}}(M = 5, N = 5) > C_{\mathrm{PSS}} > C_{\mathrm{FFTS}} = C_{\mathrm{HS}} > C_{\mathrm{RFC}}(M = 5, n_R)$, which in particular indicates the following:

(1) Soft AS always performs no worse than hard AS. The phase-shift-based selection performs strictly better than the FFT-based selection.

(2) With the same number of RF chains, the system with AS performs strictly better than the RFC system.

Interestingly, these conclusions that hold for this compact antenna array case are also generally true for MIMO systems without considering the mutual coupling effect (e.g., [10]). But a cross-reference to the results in Figure 5 can help understand this phenomenon. In Figure 5, it is shown that when N increases from 1 to 5, the ergodic spectral efficiency of the compact MIMO system behaves nearly the same as that of MIMO systems without considering the mutual coupling effect. Therefore, it appears natural that when AS is applied to the compact MIMO system with $N \leq 5$, similar conclusions can be obtained. It is also interesting that the FFT-based selection performs almost exactly the same as the hybrid selection.

Next we increase the number of placed antenna elements to $N = 8$, at which the compact MIMO system achieves the highest spectral efficiency (cf. Figure 5). We observe some different results in Figure 7, which are $C_{\mathrm{FC}}(M = 5, N = 8) > C_{\mathrm{PSS}} > C_{\mathrm{FFTS}} > C_{\mathrm{HS}}$. These results tell the following.

(1) Soft AS always outperforms hard AS. The phase-shift-based selection delivers the best performance among all these three AS schemes.

(2) The phase-shift-based selection performs better than the RFC system when $n_R \leq 5$. The FFT-based selection performs better than the RFC system when $n_R \leq 4$. The advantage of using the hybrid selection is very limited.

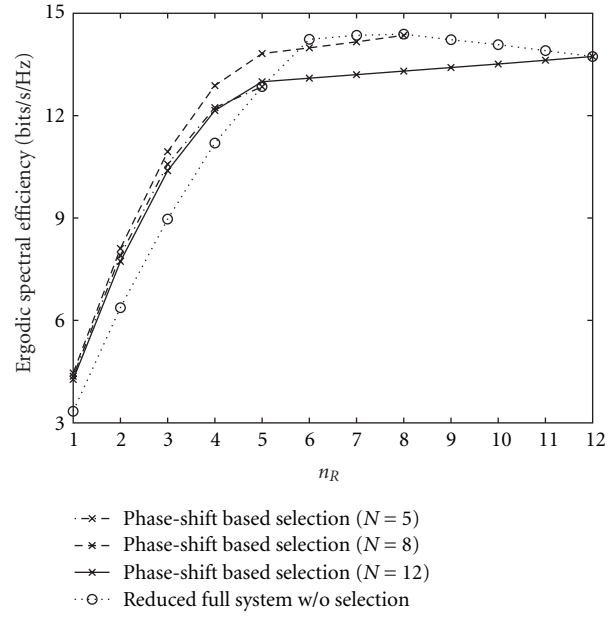


Figure 9: Ergodic spectral efficiency of a compact MIMO system with the phase-shift-based selection, where $M = 5$.

We further increase the number of antennas to $N = 12$. Now the mutual coupling effect becomes more severe, and different conclusions are demonstrated in Figure 8. It is observed that $C_{\mathrm{FC}}(M = 5, N = 12) > C_{\mathrm{PSS}} \geq C_{\mathrm{FFTS}} > C_{\mathrm{HS}}$, indicating the following.

(1) Soft AS performs strictly better than hard AS.

(2) The phase-shift-based selection performs better than the FFT-based selection when $n_R < 8$. After that, there is not much performance difference between them.

Also, similar to what we have observed in Figures 6 and 7, in terms of the ergodic spectral efficiency, none of the systems with AS outperforms the FC system with N receive antennas (and thus N RF chains). However, as for the RFC system with only n_R antennas (and thus n_R RF chains), in Figure 8 we observe the following.

(1) The RFC system always performs better than the hybrid selection. The hybrid selection seems futile in this case.

(2) The phase-shift-based selection performs better than the RFC system when $n_R < 5$. The benefit of the FFT-based selection is very limited, and it seems not worth implementing.

This indicates that due to the strong impact of mutual coupling in this compact MIMO system, only the phase-shift-based selection is still effective, but only for a limited range of numbers of the available RF chains. More specifically, when $n_R < 5$ it is best to use the phase-shift-based selection, otherwise the RFC system with n_R antennas when $5 \leq n_R < 8$. Further increase in the number of RF chains, however, will not lead to a corresponding increase in the spectral efficiency, as demonstrated in Figure 5.

For the purpose of comparison, we also plot the ergodic spectral efficiency of the phase-shift-based selection scheme in Figure 9, by extracting the corresponding curves from Figures 6–8. We find that by placing a few more antenna elements in the limited space so that the interelement spacing is less than $\lambda/2$, for example, $N = 8$ in Figure 9, the phase-shift-based selection approach can help boost the system spectral efficiency through selecting the best elements. In fact, the achieved performance is better than that of the conventional MIMO system without AS. This basically answers the question we posed in Section 1 that is related to the cost-performance tradeoff in implementation. However, further squeezing the interelement spacing will decrease the performance and bring no performance gain, as can be seen from the case of $N = 12$ in Figure 9.

4.3. Spectral Efficiency CDF of Compact MIMO with AS. In Figure 10, we investigate the CDF of the spectral efficiency for compact MIMO systems with $N = 8$. We consider the case of $n_R = 4$ in (Figure 10(a)) and $n_R = 6$ in (Figure 10(b)). We use dotted lines to denote the compact MIMO systems with AS, and dark solid lines for the FC compact MIMO systems (without AS). We also depict the spectral efficiency CDF-curves of the RFC systems of $N = 4$ and $N = 6$ in Figures 10(a) and 10(b), respectively in gray solid lines. As can be seen in Figure 10(a), soft AS schemes, that is, the phase-shift-based and FFT-based AS methods, perform pretty well as expected, but the hybrid selection performs even worse than the RFC system with $N = n_R = 4$ without AS. When we increase n_R to 6, as shown

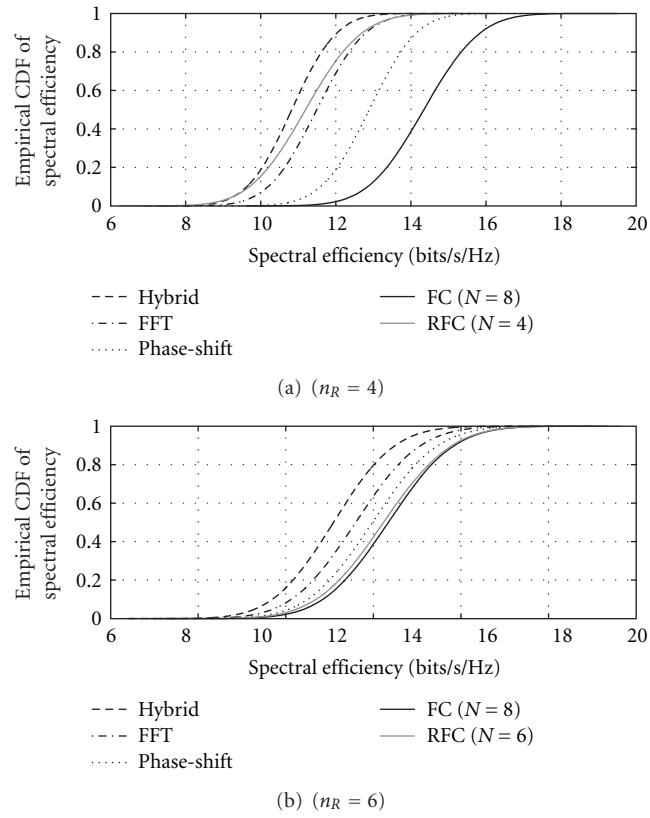


Figure 10: Empirical CDF of the spectral efficiency of a compact MIMO system with AS. $M = 5$ and $N = 8$.

in Figure 10(b), the performance difference between hard and soft AS schemes, or between the phase-shift-based and the FFT-based selection methods, is quite small. But none of these systems with various AS schemes outperform the RFC system of $N = n_R = 6$ without AS, which is consistent with what we have observed in Figure 7.

These results clearly indicate that when the mutual coupling effect becomes severe, the advantage of using AS can be greatly reduced, which however, is usually very pronounced in MIMO systems where only spatial correlation is considered at both terminals, as shown for example in Figure 11. On the other hand, it is also found that the spectral efficiency of a RFC system without AS, which is usually the lower bound spectral efficiency to that of MIMO systems with AS (as illustrated by an example of Figure 11), can become even superior to the counterpart when mutual coupling is taken into account (as shown in Figure 10 for instance). However, it should be noted that this phenomenon is closely related to the network model that we adopt in Section 2. In such model, we have assumed that all the antenna elements are grounded through the impedance $Z_{L_i}, i = 1, \ldots, N$, regardless of whether they will be selected or not. Thus, for MIMO systems with N receive elements and with a certain AS scheme, the mutual coupling impact at the receiver side comes from all these N elements, and is stronger than that of a RFC system with only n_R receive elements.

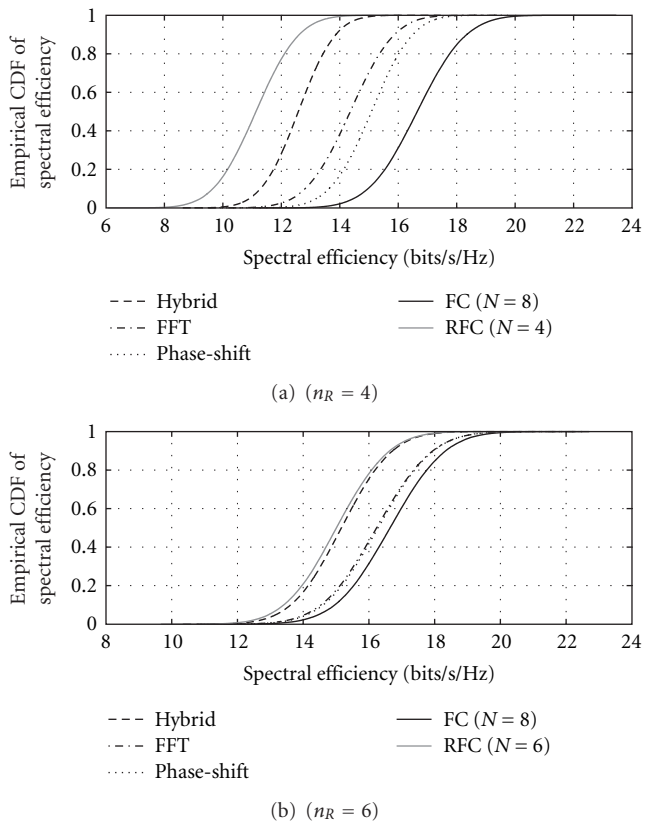


Figure 11: Empirical CDF of the spectral efficiency of a compact MIMO system with AS. $M = 5$, $N = 8$, and mutual coupling is not considered.

5. Discussions

In our study, we also test different scenarios by varying the length of linear array L_r, for example, we choose $L_r = 3\lambda$, 4λ, and so forth. For brevity, we leave out these simulation results here, but summarize our main findings as follows.

Suppose the ergodic spectral efficiency of a compact antenna array MIMO system saturates at N_{sat}. Our simulation results (e.g., Figure 5) indicate that

$$N_{\text{sat}} > \left\lfloor \frac{2L_r}{\lambda} + 1 \right\rfloor, \tag{20}$$

where $\lfloor \cdot \rfloor$ rounds the number inside to the nearest integer less than or equal to it. We also have $n_R < N$ for the sake of deploying AS. Our simulations reveal that the interelement spacing is closely related to the functionality of AS schemes. For the cases we study, the conclusion is the following:

(1) When $d_r \geq \lambda/2$, both soft and hard selection methods are effective, but the selection gains vary with respect to n_R. Particularly, the phase-shift-based selection delivers the best performance among these tested schemes. Performance of the FFT-based selection and the hybrid selection appears undistinguishable.

(2) When $d_r < \lambda/2$, there exist two situations:

(a) When $n_R \leq \lfloor 2L_r/\lambda + 1 \rfloor$, the selection gain of the phase-shift-based selection still appears pronounced, but tends to become smaller when n_R approaches $\lfloor 2L_r/\lambda + 1 \rfloor$. The advantage of using the FFT-based selection is quite limited. The hybrid selection seems rather futile.

(b) When $\lfloor 2L_r/\lambda + 1 \rfloor < n_R < N_{\text{sat}}$, neither soft nor hard selection seems effective. This suggests that AS might be unnecessary. Instead, we can simply use a RFC system with n_R RF chains by equally distributing the elements over the limited space.

It is noted that in all these cases we examine, soft AS always has a superior performance over hard selection. This is because soft selection tends to use all the information available, while hard selection loses some additional information by selecting only a subset of the antenna elements. Our simulation results also suggest that, if hard selection is to be used, it is necessary to maintain $d_r \geq \lambda/2$. Otherwise, the strong mutual coupling effect could render this approach useless. Further, if the best selection gain for system spectral efficiency is desired, one can place N_{sat} or so elements along the limited-length linear array, use $\lfloor 2L_r/\lambda + 1 \rfloor$ or less RF chains, and apply the phase-shift-based selection method. Therefore, it becomes crucial to identify the saturation point N_{sat}. This in turn requires the electromagnetic modeling of the antenna array that can take into account the mutual coupling effect.

6. Conclusion

In this paper, we proposed a study of some typical hard and soft AS methods for MIMO systems with closely spaced antennas. We assumed antenna elements are placed linearly in a side-by-side fashion, and we examined the mutual coupling effect through electromagnetic modeling of the antenna array and theoretical analysis. Our results indicate that, when the interelement spacing is larger or equal to one half wavelength, selection gains of these tested soft and hard AS schemes will be very pronounced. However, when the number of antennas to be placed becomes larger and the interelement spacing becomes smaller than a half wavelength, only the phase-shift-based selection remains effective and this is only true for a limited number of available RF chains. The same conclusions however, are not observed for the case of hard selection. Thus it seems necessary to maintain the interelement spacing no less than one half wavelength when the hard selection method is desired. On the other hand, if the best selection gain for system spectral efficiency is desired, one can employ a certain number of elements for which the compact MIMO system attains its maximum ergodic spectral efficiency, use $\lfloor 2L_r/\lambda + 1 \rfloor$ or less RF chains, and deploy the phase-shift-based selection method. This essentially indicates, if the cost-performance tradeoff in implementation is concerned, by placing a few more than necessary antenna elements so that the system spectral efficiency reaches saturation and deploying the phase-shift-based selection approach, we can achieve better

performance in terms of system spectral efficiency than the conventional MIMO system without AS. Overall, our study provides novel insight into the deployment of AS in future generation wireless systems, including the 3GPP LTE and LTE advanced technologies.

Acknowledgments

This material is based on research supported by the Air Force Research Laboratory under agreement FA9550-09-1-0576, by the National Science Foundation under Grant CCF-0829958, and by the U.S. Army Research Office under Grant W911NF-08-1-0449. The authors would like to thank Dr. Dmitry Chizhik and Dr. Dragan Samardzija of Bell Laboratories, Alcatel-Lucent for the helpful discussions on the modeling and implementation issues.

References

[1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, 1998.

[2] Overview of 3GPP Release 8 V0.0.3, November 2008, http://www.3gpp.org/Release-8.

[3] A. F. Molisch and M. Z. Win, "MIMO systems with antenna selection," *IEEE Microwave Magazine*, vol. 5, no. 1, pp. 46–56, 2004.

[4] J. Kotecha, "LTE:MIMO Techniques in 3GPP-LTE," Freescale Semiconductor, June 2008.

[5] M. Z. Win and J. H. Winters, "Analysis of hybrid selection/maximal-ratio combining in Rayleigh fading," *IEEE Transactions on Communications*, vol. 47, no. 12, pp. 1773–1776, 1999.

[6] A. F. Molisch, M. Z. Win, and J. H. Winter, "Reduced-complexity transmit/receive-diversity systems," *IEEE Transactions on Signal Processing*, vol. 51, no. 11, pp. 2729–2738, 2003.

[7] A. Gorokhov, D. A. Gore, and A. J. Paulraj, "Receive antenna selection for MIMO spatial multiplexing: theory and algorithms," *IEEE Transactions on Signal Processing*, vol. 51, no. 11, pp. 2796–2807, 2003.

[8] L. Collin, O. Berder, P. Rostaing, and G. Burel, "Soft vs. hard antenna selection based on the minimum distance for MIMO systems," in *Proceedings of the 38th Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1369–1373, Grove, Calif, USA, November 2002.

[9] A. F. Molisch and X. Zhang, "FFT-based hybrid antenna selection schemes for spatially correlated MIMO channels," *IEEE Communications Letters*, vol. 8, no. 1, pp. 36–38, 2004.

[10] X. Zhang, A. F. Molisch, and S.-Y. Kung, "Variable-phase-shift-based RF-baseband codesign for MIMO antenna selection," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4091–4103, 2005.

[11] C. A. Balanis, *Antenna Theory: Analysis and Design*, John Wiley & Sons, New York, NY, USA, 3rd edition, 2005.

[12] R. Janaswamy, "Effect of element mutual coupling on the capacity of fixed length linear arrays," *IEEE Antennas and Wireless Propagation Letters*, vol. 1, pp. 157–160, 2002.

[13] C. Waldschmidt, S. Schulteis, and W. Wiesbeck, "Complete RF system model for analysis of compact MIMO arrays," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 3, pp. 579–586, 2004.

[14] J. W. Wallace and M. A. Jensen, "Mutual coupling in MIMO wireless systems: a rigorous network theory analysis," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1317–1325, 2004.

[15] A. A. Abouda and S. G. Häggman, "Effect of mutual coupling on capacity of MIMO wireless channels in high SNR scenario," *Progress in Electromagnetics Research*, vol. 65, pp. 27–40, 2006.

[16] M. J. Gans, "Channel capacity between antenna arrays—part I: sky noise dominates," *IEEE Transactions on Communications*, vol. 54, no. 9, pp. 1586–1592, 2006.

[17] M. J. Gans, "Channel capacity between antenna arrays—part II: amplifier noise dominates," *IEEE Transactions on Communications*, vol. 54, no. 11, pp. 1983–1992, 2006.

[18] Z. Xu, S. Sfar, and R. S. Blum, "On the importance of modeling the mutual coupling for antenna selection for closely-spaced arrays," in *Proceedings of IEEE Conference on Information Sciences and Systems (CISS '06)*, pp. 1351–1355, Princeton, NJ, USA, March 2006.

[19] D. Lu, D. K. C. So, and A. K. Brown, "Receive antenna selection scheme for V-BLAST with mutual coupling in correlated channels," in *Proceedings of IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '08)*, pp. 1–5, Cannes, France, September 2008.

[20] S. J. Orfanidis, *Electromagnetic Waves and Antennas*, Rutgers University, Piscataway, NJ, USA, 2004.

[21] J. Hewes, "Impedance and reactance," The Electronics Club, February 2008, http://www.kpsec.freeuk.com/imped.htm.

[22] J. P. Kermoal, L. Schumacher, K. I. Pedersen, P. E. Mogensen, and F. Frederiksen, "A stochastic MIMO radio channel model with experimental validation," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 6, pp. 1211–1226, 2002.

[23] C. Oestges, B. Clerckx, D. Vanhoenacker-Janvier, and A. J. Paulraj, "Impact of fading correlations on MIMO communication systems in geometry-based statistical channel models," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1112–1120, 2005.

[24] D.-S. Shiu, G. J. Foschini, M. J. Gans, and J. M. Kahn, "Fading correlation and its effect on the capacity of multielement antenna systems," *IEEE Transactions on Communications*, vol. 48, no. 3, pp. 502–513, 2000.

[25] Y. Nakaya, T. Toda, S. Hara, J.-I. Takada, and Y. Oishi, "Incorporation of RF-adaptive array antenna into MIMO receivers," in *Proceedings of IEEE Topical Conference on Wireless Communication Technology (WCT '03)*, pp. 297–298, Honolulu, Hawaii, USA, October 2003.

[26] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, Pa, USA, 1997.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 510617, 18 pages
doi:10.1155/2009/510617

Research Article

Downlink Scheduling for Multiclass Traffic in LTE

Bilal Sadiq,[1] Ritesh Madan,[2] and Ashwin Sampath[2]

[1] *Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, The University of Texas at Austin, 1 University Station C0803, Austin, TX 78712-0240, USA*
[2] *Qualcomm Flarion Technologies, 500 Somerset Corporate Blvd, Bridgewater, NJ 08807, USA*

Correspondence should be addressed to Bilal Sadiq, sadiq@ece.utexas.edu

Received 16 February 2009; Revised 26 June 2009; Accepted 30 July 2009

Recommended by Cornelius van Rensburg

We present a design of a complete and practical scheduler for the 3GPP Long Term Evolution (LTE) downlink by integrating recent results on resource allocation, fast computational algorithms, and scheduling. Our scheduler has low computational complexity. We define the computational architecture and describe the exact computations that need to be done at each time step (1 milliseconds). Our computational framework is very general, and can be used to implement a wide variety of scheduling rules. For LTE, we provide quantitative performance results for our scheduler for full buffer, streaming video (with loose delay constraints), and live video (with tight delay constraints). Simulations are performed by selectively abstracting the PHY layer, accurately modeling the MAC layer, and following established network evaluation methods. The numerical results demonstrate that queue- and channel-aware QoS schedulers can and should be used in an LTE downlink to offer QoS to a diverse mix of traffic, including delay-sensitive flows. Through these results and via theoretical analysis, we illustrate the various design tradeoffs that need to be made in the selection of a specific queue-and-channel-aware scheduling policy. Moreover, the numerical results show that in many scenarios *strict prioritization* across traffic classes is suboptimal.

1. Introduction

The 3GPP standards' body has completed definition of the first release of the Long Term Evolution (LTE) system. LTE is an Orthogonal Frequency Division Multiple Access (OFDMA) system, which specifies data rates as high as 300 Mbps in 20 MHz of bandwidth. LTE can be operated as a purely scheduled system (on the shared data channel) in that all traffic including delay-sensitive services (e.g., VoIP or SIP signaling, see, e.g., [1, 2]) needs to be scheduled. Therefore, scheduler should be considered as a key element of the larger system design.

The fine granularity (180 KHz Resource Block times 1 millisecond Transmission Time Interval) afforded by LTE allows for packing efficiency and exploitation of time/frequency channel selectivity through opportunistic scheduling, thus enabling higher user throughputs. However, unlike what is typically the case in wired systems, more capacity does not easily translate to better user-perceived QoS for delay sensitive flows (VoIP, video-conferencing, stream video, etc.) in an opportunistic system. This is because a QoS scheduler has to carefully tradeoff *maximization of total transmission rate* versus *balancing of various QoS metrics (e.g., packet delays) across users*. In other words, one may need to sometimes schedule users whose delays/queues are becoming large but whose current channel is not the most favorable; see Section 2.1 for a review and discussion of results on best effort and QoS scheduling. Therefore, in this paper, we investigate the case for using queue- and channel-aware schedulers (see [3–5]) in an LTE downlink to deliver QoS requirements for a mix of traffic types.

We consider a very general scheduling framework, where each *flow* through its QoS class identifier (see Section 3.2) is mapped to a set of QoS parameters as required by the scheduler—the mapping can be changed to yield a different prioritization of flows; this requires no change in the computational framework. We make the following main contributions in this paper.

(i) We extend much existing work on *single*-user queue- and channel-aware schedulers (i.e., schedulers which pick a single user to transmit to in each scheduling interval) to multiuser ones for wideband systems. We also develop a computational architecture which allows for efficient computation of the scheduling policies in such a setting. The computational complexity of our scheduler is essentially $O(n)$ for n users—this complexity is amortized over multiple time steps.

(ii) Through analysis and numerical results for different traffic models, we illustrate the various design choices (e.g., the specifics of the tradeoff mentioned earlier in this section) that need to be made while selecting a scheduling policy. We demonstrate that queue- and channel-aware schedulers lead to significant performance improvements for LTE. Such schedulers not only increase the system capacity in terms of the number of QoS flows that can be supported but also reduce resource utilization. Our simulation methodology is based on established network evaluation methodologies. We accurately model the LTE MAC layer, and selectively abstract the PHY layer.

While we focus on LTE in this paper, we note that the computational framework and the insights gained via the numerical studies can be extended to other orthogonal division frequency multiple access (OFDMA) technologies such as Worldwide Interoperability for Microwave Access (WiMax) and Ultra Mobile Broadband (UMB).

The rest of the paper is organized as follows. In Section 2, we provide a representative (but by no means complete) sample of results in literature and relate some of our contributions to the existing work. We also discuss in greater detail the key analytical results on wireless scheduling, and in doing so, make a case for considering queue- and channel-aware schedulers for both delay sensitive and best effort flows. The system model—LTE scheduling framework and how various functionalities can and have been used—is presented in Section 3. Having done that, the detailed scheduler design and implementation using fast computational algorithms is presented in Section 4. Details of simulation setup—the PHY layer abstraction, network deployment models, and traffic models—are presented in Section 5. Simulations demonstrating the performance of the scheduler for various traffic types, namely, best effort, video-conferencing, and streaming video, are presented in Section 6. Finally, Section 7 concludes the paper.

2. Scheduling in Wireless Systems: Prior Work and Discussion

Resource allocation in wireless networks is fundamentally different than that in wired networks due to the time-varying nature of the wireless channel [6]. There has been much prior work on scheduling policies in wireless networks to allocate resources among different flows based on the channels they see and the flow state; see, for example, the excellent overview articles [6, 7], and the references therein.

Much prior work in this area can be divided into two categories: scheduling for Elastic (non-real-time) flows, and that for real-time flows.

Scheduling for Elastic (Non-Real-Time) Flows. The end-user experience for an elastic flow is modeled by a concave increasing utility function of the *average* rate experienced by the flow [8]. The proportional fair algorithm (see, e.g., [9]), where all the resources are allocated to the flow with the maximum ratio of instantaneous spectral efficiency (which depends on the channel gain) to the average rate, has been analyzed in [10–14]. Roughly speaking, this algorithm maximizes the sum (over flows) of the log of long-run average rates allocated to the flows. For OFDMA-based systems, resource allocation algorithms which focus on maximizing sum rate (without fairness or minimum rate guarantees) include [15–19]. Efficient computational algorithms for maximizing the sum of general concave utility functions of the current and/or average rate were obtained recently in [20].

Scheduling for Real-Time Flows. Real-time flows are typically modeled by independent (of service) random packet arrival processes into their respective queues, and where packets have a delay target, for example, a maximum-delay deadline. A *stabilizing* scheduling policy in this setting is one which ensures that the queue lengths do not grow without bound. Stabilizing policies for different wireless network models have been characterized in, for example, [3–5, 21–23]. Under all stabilizing policies, even though the average rate seen by a flow is equal to its mean arrival rate, still the (distribution of) packet delay can be very different under different policies [6]; it is for the same reason that in order to meet the packet delay/QoS requirement of a real-time flow, it is *not* sufficient to only guarantee the allocation of *at least a minimum average rate* to the flow. Analytical results regarding the queue (or packet delay) distribution under the schedulers proposed in [3–5] were recently obtained in [24–26], and are discussed in the following subsection. For the case where packets are dropped if their delay exceeds the deadline, the scheduling policy in [27] minimizes the percentage of packets lost. Work on providing throughput guarantees for real-time flows includes [28, 29], and references therein.

The policies to schedule a mixture of elastic and real-time flows (with delay deadlines of the order of a second) have been considered in [30] for narrowband systems, and in [31] for wideband OFDMA systems where the latter assumes that the statistics of the packet arrival process of the real-time flows along with the channel statistics are known. The scheduling policy in [31] is *persistent* and only provides an average rate guarantee to the real-time flows, which, as pointed out earlier, is generally not sufficient to guarantee the packet delay targets. By contrast with the above two, in this paper we investigate whether, given the faster MAC turn-around times and larger bandwidths of LTE systems, the queue- and channel-aware scheduler can and should be used for real-time flows with delay deadlines of few tens of milliseconds. (The answer is yes.)

There is an extensive body of work that uses some of the above results in the design of scheduling policies for

LTE specifically. The papers that investigate issues similar to those dealt with in this paper include [32–35]. In [32], it is shown that adaptive reuse can be beneficial when there is mix of VoIP and data flows, and VoIP is given strictly higher priority. A scheduling policy with strict priority across classes was also studied by [34]. Within a class, the proposed scheduling policy computes the resource allocation "chunk-by-chunk" leading to a high computational complexity; the computational complexity of such schedulers can in fact be reduced significantly by using the fast computational algorithms presented in this paper. The work in [33] showed that strict prioritization for session initiation protocol (SIP) packets over other packets can lead to better performance. While strict prioritization for low rate flows such as SIP may be feasible, we show that in general it can lead to greatly sub-optimal resource utilization. Specifically, we design scheduling policies where the priority of a class of flows in not *strict* but rather *opportunistic*. The work in [35] studies a scheduling policy that gives equal priority to all QoS packets until their delay gets close to the deadline; when the packet delays get close to the deadline, the scheduling priority of such packets is increased. In fact, this policy can be seen as belonging to a wider class of queue- and channel-aware schedulers which *smoothly* partition the queue or delay state space in regions where channel conditions are given a higher weight and regions where the delay deadlines are given a higher weight. This is made precise in the following subsection.

Scheduling policies specifically for voice over internet protocol (VoIP) have been studied in, for example, [36–38]. Policies for full buffer traffic have been studied in, for example, [2, 39–44]; many of these papers focus on modifications to the proportional fair algorithm. A packing algorithm to deal with the constraints on resource assignment due to single-carrier FDMA on the uplink was studied in [45]. Fractional power control and admission control for the uplink have been studied in [46, 47], respectively.

2.1. Discussion. To motivate and put into context the simulations presented in this paper, here we summarize some of the key analytical results in the area of opportunistic scheduling. Through this section, it will suffice to picture a fixed number, N, of users sharing a wireless channel. Each user's data arrives to a queue as a random stream where it awaits transmission/service. The wireless channel is time-varying in that the transmission rates supported for each user vary randomly over time. A scheduling rule in this context selects a *single* user/queue to receive service in every scheduling instant. However, most of the single-user schedulers can be extended to multiuser versions (for wideband systems) with some effort; in Section 4.2 we present the extensions for the ones used in this paper.

Among many others mentioned in the previous section, the work in [48] considers opportunistic scheduling in a setting where users' queues are *infinitely backlogged* (this full buffer setting is typically used to model elastic or best effort flows). They identify channel-aware opportunistic scheduling policies, which maximize the sum throughput (or, more generally, sum of any concave utility function of user throughput) under various types of fairness constraints. For example, let $\overline{x}_i$ denote the average rate offered to user i over a long run (assuming the average exists, which does under stationary channels and scheduling rules) and any weights $\alpha_i > 0$ be given, then a scheduler which maximizes $\sum_i \alpha_i \overline{x}_i$ is given like this: in any scheduling instant, if the users' time-varying channel spectral efficiencies take value $\mathbf{K} \equiv (K_i : 1 \le i \le N)$ (where K_i is the spectral efficiency of ith user's channel and is computed from its CQI), schedule a user i^* satisfying

$$i^*(\mathbf{K}) \in \arg\max_{1\le i\le N} \alpha_i K_i. \tag{1}$$

Setting $1/\alpha_i$ equal to either the exponentially filtered average of allocated rate (see $x_i(t)$ in (6)) or the long-run average of spectral efficiency, denoted by $\overline{K}_i$, yields two versions of proportional fair (PF) scheduling. With $\alpha_i = 1/\overline{K}_i$ in the above scheduler, define for later use $\overline{x}_i^{\mathrm{PF}} \equiv \mathbb{E}[K_i 1_{\{i^*(\mathbf{K})=i\}}]$, where expectation is with respect to *random* $\mathbf{K}$ having the same distribution as the time-varying channel spectral efficiencies. The missing element in these works is the impact of queueing dynamics, which certainly cannot be ignored for QoS flows like voice, live and streaming video, and so forth.

Once queueing dynamics are introduced, the opportunistic schedulers that are both queue- and channel-aware can and should be considered. Queue-awareness can be incorporated in a scheduler by, for example, replacing the fixed vector $\boldsymbol{\alpha} \equiv (\alpha_i : 1 \le i \le N)$ in (1) with a vector field $\boldsymbol{\alpha}(\cdot)$ on the state space of queue (or delay). That is, at any time when users' queues are in state $\mathbf{q} \equiv (q_i : 1 \le i \le N)$ and their channel spectral efficiencies are $\mathbf{K} \equiv (K_i : 1 \le i \le N)$, schedule a user i^* satisfying

$$i^*(\mathbf{q}, \mathbf{K}) \in \arg\max_{1\le i\le N} \alpha_i(\mathbf{q}) K_i. \tag{2}$$

Queue length q_i can be replaced/combined with head-of-line delay, w_i. We enumerate a few reasons why queue- and channel-aware schedulers should be considered.

(a) Opportunistic schedulers which are solely channel-aware may not even be stable (i.e., keep the users' queues bounded), unless chosen carefully, for example, using prior knowledge of mean arrival rates into the users' queues. See, for example, [49] which shows the instability of PF scheduling.

(b) There are queue- and channel-aware schedulers that are *throughput-optimal*, that is, they ensure the queues' stability without any knowledge of arrival and channel statistics if indeed stability can be achieved under any other scheduler. Examples are MaxWeight [3], Exponential (Exp) rule [4], and Log rule [5], which have the same form as (2). Moreover, necessary and sufficient conditions on $\boldsymbol{\alpha}(\cdot)$ for the scheduler in (1) to be throughput optimal have also been shown [50, 51].

(c) Throughput optimal schedulers, along with virtual token queues, can be used to offer minimum rate guarantees or maximize utility functions of user throughput under rate constraints [30, 52].

(d) Even if stability of the queues were not a concern, still it is imperative for a QoS scheduler to be both channel- and queue-aware: in order to meet QoS requirements, one may need to sometimes schedule users whose delays/queues are becoming large but whose current channel is not the most favorable.

(e) The work in [53] shows that under a *constant* load, scheduling algorithms that are oblivious to queue state will incur an average delay that grows at least linearly in number of users, whereas, channel- and queue-aware schedulers can achieve an average delay that is independent of the number of users.

Throughput optimal schedulers MaxWeight, Exp rule, and Log rule are defined as follows: when users' queues are in state $\mathbf{q}$ and their channel spectral efficiencies are $\mathbf{K} \equiv (K_i : 1 \le i \le N)$, schedulers MaxWeight, Exp, and Log rule serve a user i^*_{MW}, i^*_{EXP}, and i^*_{LOG}, respectively, that is given by

$$\begin{aligned} i^*_{\mathrm{MW}}(\mathbf{q},\mathbf{K}) &\in \arg\max_{1\le i\le N} \quad b_i q_i^{\beta} \times K_i, \\ i^*_{\mathrm{EXP}}(\mathbf{q},\mathbf{K}) &\in \arg\max_{1\le i\le N} \quad b_i \exp\left(\frac{a_i q_i}{c+\left((1/N)\sum_j a_j q_j\right)^{\eta}}\right)\times K_i, \\ i^*_{\mathrm{LOG}}(\mathbf{q},\mathbf{K}) &\in \arg\max_{1\le i\le N} \quad b_i \log(c + a_i q_i) \times K_i, \end{aligned} \tag{3}$$

for any fixed positive b_i's, a_i's, β, c, and $0 < \eta < 1$, and augmented with any fixed tie-breaking rule. Queue length q_i can be replaced with head-of-line delay, w_i, to obtain the delay-driven version of each scheduler.

As hinted at by the aforementioned (d), a key challenge in designing a queue- and channel-aware scheduler, that is, choosing the vector field $\boldsymbol{\alpha}(\cdot)$, is determining an optimal tradeoff between *maximizing current transmission rate* (being opportunistic now) versus *balancing unequal queues/delays* (enhancing subsequent user diversity to enable future high rate opportunities, ensuring fairness amongst users, and delivering QoS requirements.) Key optimality properties (beyond and more interesting than stability) can be understood from the way a scheduler makes this tradeoff. Next, we examine how the three throughput optimal schedulers mentioned earlier make this tradeoff, and relate it to the known asymptotics of queues/delays under these schedulers.

It can be seen that by setting $b_i = 1/\overline{K}_i$ for each i in (3), all three schedulers reduce to PF when queue lengths of all users are equal or *fairly close*. However, "fairly close" is interpreted differently by each scheduler. To define this more formally, assume that users' channels are stationary random processes and let

$$\overline{x}_i^{\mathrm{EXP}}(\mathbf{q}) \equiv \mathbb{E}\left[K_i 1_{\{i^*_{\mathrm{EXP}}(\mathbf{q},\mathbf{K})=i\}}\right] \quad i \in \{1,\ldots,N\} \tag{4}$$

(with $\overline{x}_i^{\mathrm{MW}}(\mathbf{q}), \overline{x}_i^{\mathrm{LOG}}(\mathbf{q})$ defined similarly) where the expectation is with respect to *random* $\mathbf{K}$ having the same distribution as the time-varying channel spectral efficiencies. Then, in a stable queueing system under EXP rule, $\overline{x}_i^{\mathrm{EXP}}(\mathbf{q})$ is the average rate seen by the ith user, conditional on queues being in state $\mathbf{q}$. For an $N = 2$ user system and parameters $a_1 = a_2$ in (3), Figure 1 illustrates the shape of the set

$$\mathcal{S}^{\mathrm{PF}}_{\mathrm{EXP}} = \left\{\mathbf{q} \ge 0 : \overline{\mathbf{x}}^{\mathrm{EXP}}(\mathbf{q}) = \overline{\mathbf{x}}^{\mathrm{PF}}\right\}, \tag{5}$$

that is, the partition of the queue state space where average rate of all users under Exp rule is the same as the average rate under PF; (sets $\mathcal{S}^{\mathrm{PF}}_{\mathrm{MW}}$, $\mathcal{S}^{\mathrm{PF}}_{\mathrm{LOG}}$ defined similarly). With line $\{\mathbf{q} : q_1 = q_2\}$ as an axis, the partition $\mathcal{S}^{\mathrm{PF}}_{\mathrm{MW}}$ is a cone, the partition $\mathcal{S}^{\mathrm{PF}}_{\mathrm{EXP}}$ is cylinder (with gradually increasing radius), and partition $\mathcal{S}^{\mathrm{PF}}_{\mathrm{LOG}}$ is shaped like a French horn [5].

As the queues move out of the partitions $\mathcal{S}^{\mathrm{PF}}_{(\cdot)}$ due to an increase in q_1 and/or decrease in q_2, the rate allocation changes in favor of q_1, that is, each scheduler moves away from being proportional fair in order to *balance* unequal queues (or delays). If q_1 continues to increase and/or q_2 decrease, each scheduler will eventually schedule only user 1 (whenever $K_1 \neq 0$): the partition where MaxWeight, Exp rule, and Log rule schedule only the ith queue (whenever $K_i \neq 0$) is, respectively, illustrated by $\mathcal{S}^i_{\mathrm{MW}}$, $\mathcal{S}^i_{\mathrm{EXP}}$, and $\mathcal{S}^i_{\mathrm{LOG}}$ on Figure 1.

The exact shape of each partition in terms of width, curvature of boundaries, and so forth, depends on the parameters in (3) and on the finite set that $\mathbf{K}$ takes values in (defined by all the available MCSs). However, the shapes of partitions do not depend on the distribution of random $\mathbf{K}$ [26]. So these shapes are what an engineer will implicitly or explicitly design (by choosing a vector field $\boldsymbol{\alpha}(\cdot)$ or changing parameters in (3)) in view of the QoS and rate requirements of users.

Beyond a visual description of partitions as a cone, cylinder, French horn, and so forth, the following mathematical description with useful insights can be given [5]: for any $\mathbf{q} > 0$ and scalar $s > 0$ and with b_i's as in (3):

(i) $\sum_{i=1}^N b_i \overline{x}_i^{\mathrm{MW}}(s\mathbf{q})$ is constant in s,

(ii) $\sum_{i=1}^N b_i \overline{x}_i^{\mathrm{EXP}}(s\mathbf{q})$ is decreasing in s, and in the limit $s \to \infty$, only the longest queue(s) are scheduled (as long as their channels are nonzero),

(iii) $\sum_{i=1}^N b_i \overline{x}_i^{\mathrm{LOG}}(s\mathbf{q})$ is increasing in s, and in the limit $s \to \infty$, the sum is the maximum possible. For example, with each b_i set to $1/\overline{K}_i$ in (3), $\lim_{s\to\infty} \overline{\mathbf{x}}^{\mathrm{LOG}}(s\mathbf{q}) = \overline{\mathbf{x}}^{\mathrm{PF}}$. This property is called radial sum-rate monotonicity (RSM).

Therefore, as the queues grow linearly, (i.e., scaled up by a constant), Log rule (or any scheduler satisfying RSM) schedules in a manner that de-emphasizes *queue-balancing* in favor of increasing the total weighted *service rate* (with respect to weight vector $\mathbf{b}$); whereas, the Exp rule schedules in a manner that emphasizes *queue-balancing* at the cost of total weighted *service rate*. Then, it is shown in [25] that Exp rule minimizes the asymptotic probability of *max*-queue, $\max_i a_i q_i(t)$, overflow (or, more precisely, the asymptotic exponential decay rate of max-queue distribution). Similarly, Log rule has been shown [26] to minimize the asymptotic probability of *sum*-queue, $\sum_i b_i q_i(t)$, overflow.

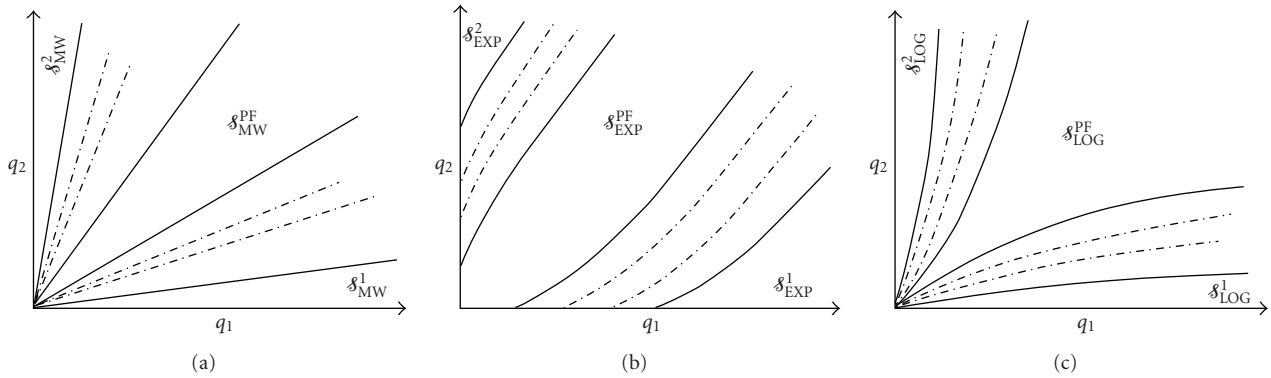


Figure 1: Partitions of queue state-space under (a) MaxWeight, (b) Exp, and (c) Log rules.

2.1.1. Use of Queue- and Channel-Aware Schedulers for Elastic Traffic. Throughput optimal schedulers, like Exp and Log rules, can also be used for scheduling elastic flows which are often modeled as full/infinitely backlogged buffers instead of dynamic queues with random arrivals that are independent of service rate. This is done by using virtual token queues that are fed by deterministic arrivals at a constant rate λ_i, and making scheduling decisions based on the virtual queues [30, 52]. If token rates λ_i are feasible (i.e., lie within the opportunistic capacity region associated with the channel), then each user i will be offered an average rate $\overline{x}_i \geq \lambda_i$. Moreover, if token rates λ_i are not feasible, then recent asymptotic analysis of Exp [25] and Log [26] rules show that the average rates $(\overline{x}_i : 1 \leq i \leq N)$ have the following interesting and desirable properties.

(i) Under Log rule, $\sum_{i=1}^{N} b_i \overline{x}_i$ is maximized subject to $\overline{x}_i \leq \lambda_i$. That is, Log rule splits users in two sets, for one set of users $\overline{x}_i = \lambda_i$, whereas for the other $\overline{x}_i < \lambda_i$, and the sets are chosen such that the total weighted rate $\sum_{i=1}^{N} b_i \overline{x}_i$ is maximized.

(ii) Under Exp rule, variable $d > 0$ is minimized subject to $\lambda_i - \overline{x}_i \leq d/a_i$. That is, either each user's average rate $\overline{x}_i$ is decremented by d/a_i (compared to its required rate λ_i), or decremented to 0 (i.e., $\overline{x}_i = 0$) if the required rate λ_i is already less than d/a_i.

LTE is a *purely scheduled* system in that all traffic with diverse QoS requirements needs to be scheduled. LTE supports sufficiently short turn-around latency allowing for some opportunistic scheduling even for delay sensitive traffic (with delay tolerance of few tens of milliseconds). In this lies the motivation for simulations presented in Section 6 where we make the case that indeed queue- and channel-aware schedulers can be successfully used for delay sensitive traffic to increase the number of users that can be supported, as well as reduce the resource utilization under a given load.

3. System Model

3.1. Terminology. We introduce the following standard 3GPP terminology to be used in the rest of the document:

(i) *slot*: basic unit of time, 0.5 millisecond,

(ii) *subframe*: unit of time, 1 millisecond; resources are assigned at subframe granularity,

(iii) *eNB*: evolved Node B, refers the base station,

(iv) *UE*, user equipment, refers to the mobile,

(v) *PDCCH*: physical downlink control channel, physical resources in time and frequency used to transmit control information from eNB to UE,

(vi) *PDSCH*: physical downlink shared channel, physical resources in time and frequency used to transmit data from eNB to UE,

(vii) *CQI*: channel quality indicator, measure of the signal to noise ratio (SINR) at the UE when eNB transmits at a reference power, fed back repeatedly from the UE to the eNB.

3.2. LTE Downlink Scheduling Framework. LTE is an OFDM system where spectral resources are divided in both time and frequency. A *resource block (RB)* consists of 180 kHz of bandwidth for a time duration of 1 millisecond. (Strict definition of a *physical resource block* in LTE is 180 KHz for 0.5 millisecond (slot), but for the purpose of the simulation this definition is adequate.) Thus, spectral resource allocation to different users on the downlink can be changed every 1 millisecond (subframe) at a granularity of 180 kHz. If hopping for frequency diversity is enabled, then hopping takes place at 0.5 millisecond point of the subframe (called slot). We use B to denote the total number of resource blocks in a single subframe.

LTE features a Hybrid-ARQ mechanism based on incremental redundancy. A transport block (consisting of data bytes to be transmitted in a subframe) is encoded using a rate 1/3 Turbo encoder and, depending on the CQI feedback, assigned RBs, and modulation, the encoded transport block is rate-matched appropriately to match the code rate supported by the indicated CQI. With each subsequent retransmission, additional coded bits can be sent reducing the effective code rate and/or improving the SINR. Though LTE allows the retransmission to be made at a different

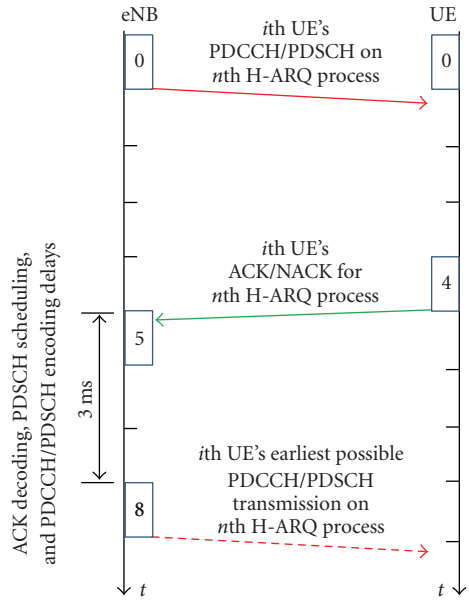


Figure 2: Downlink scheduling time-line and computational delays. At time 0 the eNB assigns resources for a first transmission to UE i; the assignment is carried over PDCCH while the actual data is sent over PDSCH, both in subframe 0. ACK/NACK information to convey whether the first transmission was decoded successfully is fed back to the eNB by the UE in subframe 5. Subframe 8 is the earliest possible time when a retransmission (if needed) for this packet can occur.

modulation scheme compared to the first transmission, this flexibility is not exploited in this paper.

Thus, in each subframe t, the scheduler grants spectral resources to users (UEs) for either fresh transmissions, or to continue past transmissions (retransmissions). We assume that each re-transmission of a packet occurs 8 ms (i.e., 8 subframes) after the previous transmission—packets are rescheduled for retransmission until they are successfully decoded at the UE, or the maximum (six) retransmissions have occurred. (LTE allows asynchronous HARQ retransmissions which means that retransmissions can occur any time after the ACK/NACK is received from the UE. In this paper, we do not exploit this flexibility and operate HARQ synchronously. Retransmissions occur with a delay in multiples of 8 ms.) For a new transmission, a modulation and coding scheme (MCS) is determined by a *rate prediction algorithm* which takes into account the most recent CQI report for the UE, and the past history of success/failure of transmissions to this UE—the rate prediction algorithm is explained in Section 4.1.

The control resources (PDCCH) to convey scheduling grants to the users are time-multiplexed with the resources to transmit data (PDSCH) over the downlink. In particular, each subframe is divided into 14 symbols, of which up to three symbols at the start of the subframe can be used for control signalling. We do not model the *details* of the control channel signalling, but we do model the overhead associated with this signalling. Specifically, we assume that out of $S = 14$ symbols every subframe, S_{cont} symbols are used for control signalling. We also model the computational delays as illustrated in Figure 2.

Downlink scheduling decisions can be made on the basis of the following information for each user.

(i) *QoS Class Identifier (QCI)*. In the LTE architecture downlink data flows from a Packet Gateway (called PDN GW) to eNB and then to the UE (user). The PDN GW to eNB is an IP link and the eNB to UE is over the wireless link. When the logical link from the bearer to the UE is set up (called a bearer), a QoS Class Identifier (QCI) is specified. This defines whether the bearer is guaranteed bit-rate or not, target delay and loss requirements, and so forth. The eNB translates the QCI attributes into logical channel attributes for the air-interface and the scheduler acts in accordance with those attributes. (We use the term user and logical channel interchangeably in this paper as we only state the results with one logical channel per user.)

(ii) *CQI*. The channel quality indicator (CQI) reports are generated by the UE and fed back to the eNB in quantized form periodically, but with a certain delay. These reports contain the value of the signal-to-noise and -interference ratio (SINR) measured by the user. We denote by $\gamma_i(t)$ the most recent wideband CQI value received by the eNB at or before time t for user i. The LTE system allows several reporting options for both wideband (over the system bandwidth) and subband (narrower than the system bandwidth) CQI, with the latter allowing exploitation of frequency selective fading.

(iii) *Buffer State*. The buffer state refers to the state of the users' buffers, representing the data available for scheduling. We assume that for each user i, the queue length in (the beginning of) subframe t, denoted by $q_i(t)$ bits, and the delay of each packet in the queue, with $w_i(t)$ ms denoting the delay of head-of-line packet, is available at the scheduler.

(iv) *Phy ACK/NACK*. At time t, ACK/NACK for all transmissions scheduled in subframe $(t - 8)$ are known to the scheduler.

(v) *Resource Allocation History*: Scheduling decisions can also be based on scheduling decisions in the past. For example, if a user was allocated multiple RBs over the past few subframes, then its priority at the current subframe may be reduced (even though ACKs/NACKs are still pending). A commonly used approach is to maintain the average rate, $x_i(t)$ at which a user is served. The average rate is updated at every time t using an exponential filter as follows:

$$x_i(t) = (1 - \tau_i)x_i(t-1) + \tau_i r_i(t), \quad t = 1, 2, \ldots, \tag{6}$$

where $r_i(t)$ is the rate allocated to the ith user at time t, and $\tau_i \in (0, 1)$ is a user specific constant; we refer to $1/\tau_i$ as *time-constant* for (rate averaging for) user i.

4. Scheduler Design for LTE

For each subframe t, the scheduler first assigns power and resource blocks to retransmissions for packets which were not decoded successfully at time $(t-8)$; the modulation and coding scheme for a retransmission is kept the same as for the previous transmission. The remaining power and spectral resources are distributed among the remaining users for transmissions of new packets. Specifically, each assignment consists of the following:

(i) the identity of the user for which the assignment is made,

(ii) the number of RBs assigned,

(iii) the transmission power for each RB,

(iv) the modulation and coding scheme for packet transmission.

In this paper, we present the schedulers and fast computational algorithms for the case where power is distributed uniformly across RBs and only the wideband CQI is being reported. However, the schedulers can be extended to case where one or both of the above restrictions are removed. More specifically, each scheduler is described as a solution to an optimization problem, where the optimization problem can be readily extended to the case where one or both of the above restrictions are removed. Moreover, fast computational algorithms to solve these more complex optimization problems are presented in [20]. Finally, we note that while we model the overhead for the control channel PDCCH, we do not study algorithms for control channel format selection.

We break the scheduling algorithm into two parts.

(a) *Rate Prediction*. The rate prediction algorithm maps (based on past history of transmissions for a UE) the CQI reports to a modulation and coding scheme that targets successful decoding in a specified number of transmissions of a packet. Even though a UE repeatedly sends CQI reports to the eNB, still rate prediction is essential in order to account for the uncertainty in the channel gain to the UE. This uncertainty arises due the following reasons:

(i) wireless channels are time-varying,

(ii) CQI is quantized to 4 bits and the quantized value may be too pessimistic (or optimistic),

(iii) CQI reports received by the eNB from a UE may be based on the channel state a few subframes earlier,

(iv) multiple retransmissions of a packet through H-ARQ may be desired to take advantage of the time diversity, where the channel can vary across the retransmissions.

(b) *Resource Assignment*. Given an achievable spectral efficiency as determined by the rate prediction algorithm, the resource allocation for new transmissions is determined as a solution of a constrained optimization problem. The optimization problem depends on the scheduling policy (proportional fair, Exponential rule, etc.).

4.1. Rate Prediction. Rate prediction is the task of determining and adapting to channel conditions, the mapping of reported CQI to the selected transport format. We start with a baseline mapping (subsequently denoted by f) that is optimal under AWGN channel. That is to say, assuming the channel gain is known and *static*, we optimize transport format for a fixed number of resources, such that the data packet is transmitted successfully to the UE in any targeted number of transmissions. The baseline mapping that is optimal for a static channel may no longer be so for a fading channel because the channel gain from an eNB to a UE can vary from one H-ARQ transmission to the next. Hence, the selection of the transport format has to take into account this uncertainty or variation in channel gains. One method of doing this is to use a *link margin* or *backoff factor*, that is adapted in a closed loop for each link individually, to adjust the transport format from that of the baseline.

Specifically, if ith user's CQI is $\gamma_i(t)$, the user is allocated $b_i(t)$ RBs at time t, and has a termination target (for successful decoding of the packet at the UE) of T_i H-ARQ transmissions, then let $f(\gamma_i(t), b_i(t), T_i)$ denote the maximum number of bits that can be transmitted over a static AWGN channel with SINR $\gamma_i(t)$. Then for a fading channel, we select the number of bits as

$$f(\gamma_i(t) - \delta_i(t), b_i(t), T_i), \tag{7}$$

where $\delta_i(t)$ is the backoff factor. The spectral efficiency (in bps/RB) for user i is then given by

$$K_i(t) = \frac{f(\gamma_i(t) - \delta_i(t), B, T_i)}{T_i B}. \tag{8}$$

The *backoff value* $\delta_i(t)$ is adapted in a closed loop manner as described in what follows. If the ith user's transmission is indeed decoded correctly in (or under) the targeted number of transmissions, T_i, then δ_i is decremented (to at most $\delta_{\text{MIN}} = -15$ dB) by some fixed small ε (dB), that is,

$$\delta_i(t+1) = \max(\delta_i(t) - \varepsilon, \delta_{\text{MIN}}). \tag{9}$$

If, however, the transmission is decoded in more than T_i number of transmissions (or not decoded at all), then δ_i is incremented (to at most $\delta_{\text{MAX}} = 15$ dB) by $s\varepsilon$ for some fixed $s \geq 1$, that is,

$$\delta_i(t+1) = \min(\delta_i(t) + s\varepsilon, \delta_{\text{MAX}}). \tag{10}$$

We note that the above rate prediction algorithm is fairly standard and has been studied in detail in [54].

For best effort flows, T_i is not fixed over time: it is set to 3 unless (i) $\gamma_i(t)$ is so high that setting T_i to a lower value results in more than 20% increase in spectral efficiency $K_i(t)$ (in which case T_i is chosen to maximize $K_i(t)$), (ii) $\gamma_i(t)$ is too low for $T_i = 3$ to be feasible (in which case T_i is set to the smallest feasible value). This allows for a high granularity in picking a spectral efficiency as well as for taking advantage of time diversity. For delay sensitive flows, T_i is always set to the smallest feasible value in order to minimize the latency incurred due to retransmissions of a packet.

4.2. Scheduling Policies. In this subsection, we describe the schedulers used for simulation results presented in Section 6, whereas, the fast computational algorithms for these schedulers are presented in the following subsection. Best effort flows are scheduled using a *utility maximizing* scheduler, whereas, QoS flows are scheduled using Exp rule, Log rule, or Earliest-Deadline-First (EDF). An efficient computational architecture to compute the resource allocation corresponding to a subset of these policies is presented in the following subsection.

4.2.1. Utility Maximizing Scheduler for Best Effort. Recall that $x_i(\cdot)$ denotes the exponentially filtered average rate of user i, that is,

$$x_i(t+1) = \tau_i K_i(t) b_i(t) + (1-\tau_i) x_i(t), \tag{11}$$

where $K_i(t)$ is defined in (8), $\tau_i \in (0,1)$ is a parameter, $b_i(t)$ is the number of RBs allocated to user i in subframe t, and $x_i(0) = 0$. We set $\tau_i = 1/500$ for all users (i.e., the time constant of the exponential filter for rate averaging is $1/\tau_i = 500$ subframe). Moreover, let $U_i : \mathbb{R}_+ \to \mathbb{R}$ be a concave continuously differentiable utility function (of average rate x_i) associated with user i. We consider functions U_i such that, for $x_i \in (0,\infty)$, we have

$$\frac{d}{dx_i} U_i(x_i) = \frac{1}{x_i^{1-\alpha}}, \tag{12}$$

for some fixed $\alpha \in (-\infty, 1]$; for example, $U_i(x_i) = \log(x_i)$ for $\alpha = 0$. Then in any subframe t, the utility maximizing scheduler allocates RBs $\mathbf{b}(t) = (b_i(t) : 1 \le i \le N)$, where N is the number of users) in order to maximize

$$\sum_{i=1}^{N} U_i(\tau_i K_i(t) b_i(t) + (1-\tau_i) x_i(t)). \tag{13}$$

We note the following points.

(a) As $\alpha \to 0$, the scheduler reduces to a proportional fair scheduler. Specifically, this scheduler will allocate the next fraction of available bandwidth resource to a user with maximum $K_i(t)/x_i(t)$.

(b) As $\alpha \to 1$, this scheduler reduces to max sum-rate scheduler.

(c) As $\alpha \to -\infty$, it reduces to the max-min fair scheduler, that is, it maximizes the minimum average rate.

4.2.2. Delay-Driven Log and Exp Rules. Log and Exp rules used in simulations are similar to the ones introduced in Section 2.1 (see (3)), however, instead of scheduling, one user in every scheduling instant, we can now schedule one user in every RB in the current subframe. So the scheduler makes scheduling decisions one RB at a time, and updates queues and the buffer state (e.g., head-of-line delay) after each assignment.

We use the delay-driven version of these rules. Let $w_i(t)$ denote the wait time of the head-of-line packet in ith user's queue at eNB in subframe t. Then under Log rule, in any subframe t,:

(i) the next available RB is allocated to a user $i^*(t)$ satisfying

$$i^*(t) \in \arg\max_{1\le i\le N} b_i \log(c + a_i w_i(t)) \times K_i(t), \tag{14}$$

with ties broken in favor of the user with smallest index,

(ii) $q_{i^*}(t)$ is decremented and $w_{i^*}(t)$ is updated based on the new buffer state. This is done before the scheduler computes the optimal user for the next RB.

Parameters b_i are set to $1/\mathbb{E}[K_i]$, $c = 1.1$, and $a_i = 5/d_i$ where d_i is the 99th percentile delay target of the ith user's flow. Recall the set $\mathcal{S}^{\mathrm{PF}}_{\mathrm{LOG}}$ from Section 2.1, that is, the partition of state space of delay (or queue) where Log rule and PF take the same scheduling decision. Then the magnitude of vector $\mathbf{a} \equiv (a_i : 1 \le i \le N)$ sets the *width* of this partition about the axis $\{\mathbf{q} \ge 0 : a_i q_i = a_j q_j\}$.

Exp rule is defined similarly, with (14) appropriately modified to,

$$i^*(t) \in \arg\max_{1\le i\le N} b_i \exp\left(\frac{a_i w_i(t)}{1+\sqrt{(1/N)\sum_j w_j(t)}}\right) \times K_i(t). \tag{15}$$

Parameters b_i are set to $1/\mathbb{E}[K_i]$ and a_i to either $6/d_i$ (in Section 6.2) or $10/d_i$, (see [30] for setting Exp rule parameters; typically $a_i \in [5/d_i, 10/d_i]$ gives good performance). Just as in the case of Log rule, magnitude of vector $\mathbf{a} \equiv (a_i : 1 \le i \le N)$ sets the *width* of partition $\mathcal{S}^{\mathrm{PF}}_{\mathrm{EXP}}$ about the axis $\{\mathbf{q} \ge 0 : a_i q_i = a_j q_j\}$.

4.2.3. Earliest-Deadline-First Scheduler. This is a queue-aware nonopportunistic scheduler which, in each subframe t, allocates the next available RB to a user $i^*(t) \in \arg\min_{1\le i\le N}(d_i - w_i(t))$, and then updates $w_{i^*}(t)$ just as in the case of Log and Exp rule.

4.3. Efficient Computation of RB Allocation under Various Schedulers. We now describe an efficient computational framework to compute the bandwidth allocations for each subframe under *utility maximization*, *queue-driven Log*, and *queue-driven MaxWeight* scheduling policies. We also show how this framework can be used to compute an approximate version of the delay-driven versions.

We first consider a generic optimization problem over the number of resource blocks, $b_i(t)$, allocated to each user i.:

$$\begin{aligned} &\text{maximize} \quad \sum_{i=1}^{N} g_i(K_i(t) b_i(t)), \\ &\text{subject to} \quad \mathbf{1}^T \mathbf{b}(t) \le B, \quad \mathbf{b}(t) \ge 0, \quad \mathbf{b}(t) \le \mathbf{b}^{\max}(t), \end{aligned} \tag{16}$$

where $g_i : \mathbb{R}_+ \mapsto \mathbb{R}$ are concave increasing functions. We ignore the constraints that $b_i(t)$'s are integers—LTE offers

high enough resource granularity, that is, with appropriate rounding techniques the loss in optimality is negligible. The maximum bandwidth that can be allocated to user i at time t is given by

$$b_i^{\max}(t) = \frac{q_i(t)}{K_i(t)}. \tag{17}$$

Using an appropriate definition of $g_i : \mathbb{R}_+ \mapsto \mathbb{R}$, the computation of different scheduling policies can be formulated as the aforementioned optimization problem as follows.

(i) *Utility Maximization*. Here, we define $g_i(y)$ as

$$g_i(y) = U_i((1-\alpha_i)x_i(t) + \alpha_i y), \quad \forall\, y \in R_+, \tag{18}$$

where we recall that $x_i(t)$ is the average rate allocated to user i as computed by an exponential filter at time t (see (11)).

(ii) *Queue-Driven Log Rule*. For all $y \in R_+$,

$$g_i(y) = -b_i\left(\left(q_i - y + \frac{c}{a_i}\right)\log(c + a_i(q_i - y)) - (q_i - y)\right). \tag{19}$$

(iii) *Queue-Driven MaxWeight Rule*. In this case, g_i is defined as

$$g_i(y) = -b_i(q_i(t) - y)^2, \quad \forall\, y \in R_+. \tag{20}$$

The delay-based versions of Log rule and MaxWeight can also be computed by first approximating those as queue-based rules like this: let $\widehat{\lambda}_i \equiv q_i(t)/w_i(t)$, that is, the average arrival rate over the wait time of the head of line packet. Then $w_i(t)$ in delay-based rules can be substituted with $q_i(t)/\widehat{\lambda}_i$.

Define the projection operator over $\mathbb{R}$ as

$$\mathcal{P}_{[a,b]}(y) = \max(\min(y, b), 0), \quad a, b \in \mathbb{R}. \tag{21}$$

This operator projects a real variable over the interval $[a, b]$.

Necessary and sufficient conditions for $\mathbf{b}(t)$ to be optimal are given by [20]

$$b_i(t) = \mathcal{P}_{[0,b_i^{\max}]}\left(\frac{1}{K_i(t)}g_i'^{-1}\left(\frac{\lambda}{K_i(t)}\right)\right), \qquad \mathbf{1}^T\mathbf{b}(t) = B, \quad \lambda > 0. \tag{22}$$

The following *bisection search* on λ can be used to solve the aforementioned problem [20]:

Given $\lambda^{\min} = 0$, $\lambda^{\max} = K_1(t)g\prime(K_1(t)B)$, tolerance ϵ.

Repeat

(a) *Bisect*. $\lambda = (\lambda^{\min} + \lambda^{\max})/2$.

(b) *Bandwidth Allocation*. Compute

$$b_i(t) = \mathcal{P}_{[0,b_i^{\max}]}\left(\frac{1}{K_i(t)}g_i'^{-1}\left(\frac{\lambda}{K_i(t)}\right)\right). \tag{23}$$

(c) *Stopping Criterion*. **quit** if $\lambda^{\max} - \lambda^{\min} < \epsilon$.

(d) *Update*. If $\mathbf{1}^T\mathbf{b}(t) < B$, $\lambda^{\max} = \lambda$, else $\lambda^{\min} = \lambda$.

In practice, about 10 iterations are sufficient to obtain a solution for an accuracy required for scheduling in LTE. An exact complexity analysis, and the choice of the tolerance ϵ to compute a solution within a certain bound of the optimal objective function are possible [20].

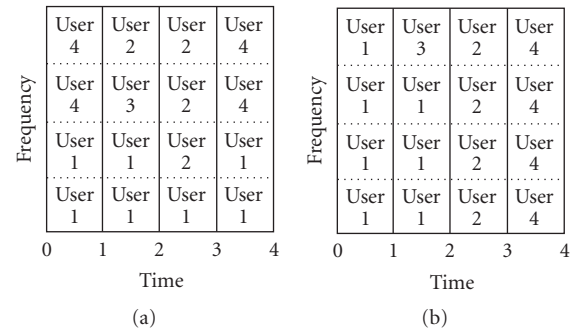


Figure 3: *Equivalent* schedules, (a) requires 9 grants versus 5 required by (b).

4.4. Further Reduction of Computation by Optimizing over a Horizon. The computational burden of above algorithms (especially for large N and B) can be reduced further by solving the convex optimization for a horizon of a few subframes rather than for each subframe. Specifically, we run the convex optimization and compute the optimal RB allocation to each user—called a user's RB target—over a horizon of a few subframes (say, 8). Then in each subsequent subframe till the next time the optimization is run, we allocate RBs by only doing the following computations (to fully exploit any CQI variation over the horizon).

(i) Update of QoS metric of users, that is, $x_i(t)$, $q_i(t)$, and/or $w_i(t)$, based on RB assignments in each subframe (as they are made).

(ii) Update of spectral efficiency $K_i(t)$ (for users for which a CQI report was received in the previous subframe).

(iii) Update of users' priority, that is, dU_i/db_i at $b_i = 0$, once the above two updates have been made.

(iv) RBs are first allocated to the highest priority user till its target is met. If some RBs remain available, they are assigned to next highest priority user, and so on. Any degenerate cases, like data buffers or control resources running out are handled such that as many as possible number of RBs are assigned in each subframe.

Remark 1. Beside reducing computational burden, solving the optimization for a horizon has an added advantage of reducing the required control signalling. This is because the a user's RB-target-over-a-horizon can now be allocated all at once in one subframe (or in a fewer number of subframes) rather than allocating only a few RBs per subframe over the duration of a horizon. For example, Figure 3 shows two schedules for a hypothetical 4-RBs-by-4-subframes scheduling problem; the two schedules are equivalent in terms of number of RBs assigned to each user. The schedule on the left is computed one subframe at a time, whereas the schedule on the right is computed using the method described earlier. That is, first, users' RB targets are computed once over the 4-RB-by-4-subframe horizon (by solving the convex program), then in each subsequent subframe, RBs

Table 1: Simulation Parameters.

Parameter	Value	Comments
Number of eNBs (3 sectored)	19	19 eNBs in a hexagonal pattern, each with 3 cells and wrap-around was used for full-buffer simulations and to generate the geometry (average SINR) distributions for the QoS simulations
Propagation Model (BTS Ant Ht = 32 m, MS = 1.5 m)	$28.6 + 35 \log 10(d)$ dB,d in meters	Modified Hata Urban Prop. Model @1.9 GHz (COST 231 ([59])). Modified means that pathloss is reduced by 3 dB in comparison to COST 231. This is a standard assumption (see, e.g., [58]).
Minimum separation between eNB and UE	35 meters	—
Log-Normal Shadowing	Standard Deviation = 8.9 dB	This shadowing is constant for each UE in each simulation run. The same shadowing amount will be used for all the sector antennas of a BS to a given UE. The correlation coefficient between the eNB's Tx antennas and a given UE and the eNB's RX antennas and a given UE is 1.
Shadowing correlation across cells in an eNB	1	—
Shadowing correlation across eNBs to a UE	0.5	—
Number of transmit antennas	1	—
Number of receive antennas	1	—
Number of resource blocks	64	This number slightly exceeds the 10 MHz bandwidth and was selected since powers of 2 are convenient when hopping is introduced. It does not change the conclusions about the schedulers. The reader can scale the numbers down to infer exact 10 MHz bandwidth performance.
Number of OFDM symbols per subframe	14	This is for normal cyclic prefix (CP). Of the 14, the first 3 are assigned to control transmissions (PDCCH, PCFICH and PHICH)
eNB transmit power per cell	20 Watts (43 dBm)	—
Thermal Noise density	−174 dBm/Hz	—
eNB and UE antenna gains	0 dBi	—
Site-to-site distance	2.0 km	—
HARQ	Synchronous, non-adaptive, incremental redundancy	—

(according to the computed targets) are allocated to the highest priority UE(s). Resultantly, the latter schedule has an advantage of requiring only 5 downlink grants on PDCCH versus 9 required by the former.

5. Simulation Framework

5.1. Network and Deployment Model. The deployment and channel models are mostly taken from the work in [55–58] and the relevant parameters are repeated here in Table 1. For the full-buffer simulation results, two-tiers (19 eNBs, 57 cells) with wrap-around was simulated with users in each eNB modeled explicitly. To save on simulation time, for the results with QoS traffic (e.g., streaming video or video conferencing) a two-step process was followed. First, the two-tier (19 eNBs, 57 cell) scenario was simulated under the assumption that all eNBs were transmitting at full power on the downlink (full loading). This was used to generate the distribution of SINRs (geometries) seen by UEs on the downlink, resulting from pathloss and shadowing. Wrap-around of cells as outlined in [58] was followed to avoid edge effects. Second, the center-cell alone was simulated with data traffic and schedulers, with each UE's SINR being drawn from the distribution calculated in the first step. Fast fading (time and frequency selective) was then generated for each UE to determine the instantaneous (per subframe) SINR.

For short-term fading, delay spread, and power-delay profile models from [57] are used. The Doppler spectrum

is the classic U-shaped power spectrum that results from Jakes/Clarke's model. The UE speed simulated was 3 km/h. The effect of channel estimation error was accounted for by applying a channel specific backoff factor (such as α term in the PHY abstraction modeling section), determined through link-level simulations.

5.2. Physical Layer Modeling. System simulations are conducted over a large number of cells/sectors and large number of users. As such, characterizing the channel, the physical layer waveform and/or exact decoding process at short timescales becomes prohibitive in terms of computation and simulation time. Yet, a reasonably accurate behavioral model of the physical layer performance is critically important in obtaining the correct system level performance representation and in tuning MAC/RLC algorithms (such as the scheduler). Link level performance is typically characterized by packet-error-rate (PER) versus long-term average SINR curves, where the latter is computed over all channel realizations. Such a curve is not very useful to use in system level simulations as several critical aspects such as user and channel sensitive rate scheduling, hybrid-ARQ and link adaptation are dependent on the short-term average channel. In some instances, the benefits of MIMO and spatial beamforming would also not be captured (e.g., those schemes often involve dynamic feedback of the spatial channel and subsequent adaptation of antenna weights in accordance), as those too are dependent on the short-term channel realization. Furthermore, one aspect of the system simulation is to allow the tuning of algorithms such as rate prediction, power control, and so forth, and therefore, the dynamic nature of physical layer performance is important to capture in the system simulation. A number of different approaches have been proposed and evaluated in the past (see [60] and references therein for a good summary). In most instances, an effective SINR that captures the channel and interference occurrences over all resource elements used in transmission of the encoded packet, is defined. [60, Equation (1)] generically defines effective SINR as follows:

$$\mathrm{SINR}_{\mathrm{eff}} = \alpha_1 I^{-1}\left(\frac{1}{P}\sum_{i=1}^{P} I\left(\frac{\mathrm{SINR}_j}{\alpha_j}\right)\right), \tag{24}$$

where P represents the number of resource elements (time-frequency resources) used over the packet transmission thus far, j is the index over the resource elements, SINR_j represents the signal-to-interference and noise ratio on jth resource element, and $I(\cdot)$ is function that is specific to the model. Note that if hybrid-ARQ is used, then the summation term should include all the H-ARQ transmissions and associated resources. The factors α_1 and α_j allow adaptation of the model to the characteristics of modulation and coding used as well as any adjustments for coded packet length relative to a baseline curve. In this paper, we use $\alpha_1 = \alpha_j = 1$ for all j. However, after calculating the effective SINR as described earlier, adjustments for packet size and channel estimation error are applied. These adjustments are computed using extensive link-level simulations for various fading channels and packet sizes. For the most part, the sensitivity to packet size is very minor and vanishes for packet sizes larger than around 500 bits. The work in [60] lists a few examples for the choice of $I(\cdot)$ as follows:

$$\begin{aligned} I(x) &= \log_2(1+x), \\ I(x) &= \exp(-x), \\ I(x) &= I_m(x). \end{aligned} \tag{25}$$

The first expression represents the unconstrained Gaussian channel capacity, the second is an exponential approximation called (Effective Exponential SINR metric) and the last expression uses I_m the mutual information at an SINR x, when modulation alphabet size of m is used. The last method, called Mutual Information Effective SINR Metric (MIESM), is widely used and is the method we will use in this paper. Once we compute the effective SINR per the above expression, then we look up the AWGN PER versus SINR curve corresponding to that modulation, code rate, and packet size to determine the probability of error. A binary random variable with that probability is then drawn and a corresponding error event is generated.

Few additional points are noteworthy, described as follows.

(i) Even though the aforementioned expressions are indexed by a resource element, in LTE, a resource element represents 1 sub-carrier (15 KHz) over 1 OFDM symbol (approximately 70 microseconds). This represents too fine a granularity and would slow down the simulation. Therefore, we use 1 resource block (180 KHz) over 1 subframe (1 millisecond) as the basic unit for generating the SINR in the simulation. Note that these values would lead to negligible, if any, loss in representation accuracy for practical delay spreads and Dopplers.

(ii) Look-up table is used to calculate the mutual information indexed by SINR and modulation type. The LTE downlink uses 3 modulation types: QPSK, 16-QAM, and 64-QAM.

(iii) We do not currently model modulation order adaptation on retransmissions.

(iv) As suggested in [60], a single parameter $\alpha_1 = \alpha_j = \beta$ for all j is used. In particular, a value of unity is used as mentioned earlier, with adjustments for channel estimation error and transport block size.

For CQI reporting, the effective SINR is calculated in a manner similar to the above, using LTE reference signals and the constrained capacity. The effective SINR is quantized to a 4-bit CQI value and fed back to the eNB. The table is generated from link curves in accordance with the block-error rate requirements of the LTE specification.

5.3. Traffic Models. The traffic models used for various simulations in Section 6 are, namely, full-buffer, streaming video, and live video. In full-buffer model, as the name suggests, each user's queue at eNB is assumed to always have infinite number of bits.

5.3.1. Streaming Video Model. Streaming video model is borrowed from [61], we summarize it here. Exactly 8 video packets arrive in a frame length of 100 milliseconds. Then the first arrival time from the beginning of a frame, as well as the seven subsequent interarrival times are independently drawn from a Pareto distribution with exponent 1.2 and truncated to [2.5 milliseconds, 12.5 milliseconds]. Moreover, packet sizes are independently drawn from a truncated Pareto distribution with exponent 0.8. The truncation depends on the desired mean rate, for example, [30, 350] bytes for a mean rate of 90 kbps.

5.3.2. Live Video Model. Live video is modeled as an ON-OFF Markov process. When in ON state, a packet of fixed size is generated every 20 ms. The transition probabilities are such that half the time the process is in ON state. Moreover, mean dwelling time in either state is 2 seconds. Then the parameter which controls the mean rate of a live video flow is the packet size, for example, 1 kilobyte for a mean rate of 200 kbps. This model is similar to the VoIP model in [61] but with higher rate due to bigger packet sizes.

6. Simulation Results

In this section, we present the results of a simulation-based evaluation of opportunistic schedulers described in Section 4.2, and discuss the key insights into scheduler design. Three sets of results are presented, each considering a different model for the arrival traffic into the users' queues at eNBs. The three traffic models are, namely, saturated queues at the eNB, multirate streaming video, and a mix of streaming and live video; the three sets of results are discussed in what follows.

6.1. Queues at eNB Are Saturated. We start by presenting the results for the case where users' queues at the eNBs are saturated (or infinitely backlogged); these results provide a good comparison and calibration against other published studies.

6.1.1. Model. The network deployment model is as described in Section 5.1, with 57 cells (3 per eNB) and 20 users per cell. Figure 4 shows the empirical CDF of users' geometry, that is, users' SINR induced by the path-loss/shadowing model when all eNBs are transmitting at full power. Each user's queue at eNB is assumed to be infinitely backlogged, and the transmissions are scheduled according to a utility maximizing best effort scheduler described earlier in Section 4.2. Moreover, to limit the computational burden, the scheduler solves the underlying convex optimization problem once in every 8 subframes over a horizon of 8 milliseconds. Then in each subsequent subframe, the scheduler combines this solution with the current CQI and average rate to compute a schedule, as described in Section 4.4.

6.1.2. Results and Discussion. The performance measures of interest are the average cell throughput (i.e., cell throughput averaged across all 57 cells) and the distribution of individual

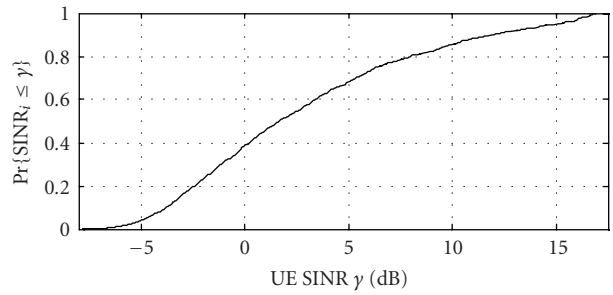


Figure 4: CDF of users' *geometry*.

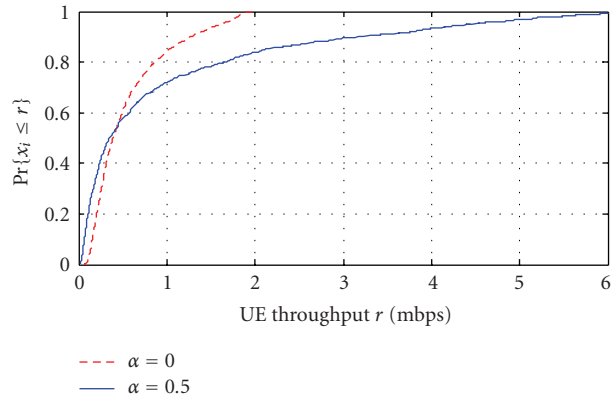


Figure 5: Empirical CDFs of average user throughput under best effort scheduler with $\alpha = 0$ (PF) and $\alpha = 0.5$.

Table 2: Fairness versus throughput tradeoff achieved by varying α.

α	Cell thrupt.	5 %-tile throughput	95%-tile throughput
0	1.02 bps/Hz/Cell	134 kbps	1.62 mbps
0.5	1.82 bps/Hz/Cell	48 kbps	4.42 mbps

users' throughput (i.e., time average of each user's rate) under various best effort schedulers, that is, as α associated with the utility function varies (see Section 4.2). Recall that $\alpha \to -\infty$ reduces to max-min fair scheduling, $\alpha = 0$ to PF scheduling, and $\alpha = 1$ to max-rate scheduling. Figure 5 shows the empirical CDFs of users' throughput (*rate CDF* for short) for the two cases, $\alpha = 0$ and $\alpha = 0.5$, and Table 2 gives the respective cell throughput as well as the 5 and the 95 percentile read from the two rate CDFs. Clearly, users' throughput under the scheduler with $\alpha = 0$ is more *fair* than users' throughput under the scheduler with $\alpha = 0.5$, however, this fairness comes at the cost of 44% drop in the average cell throughput (see Table 2). Moreover, from the cross-over point of the two CDFs in Figure 5 and the percentiles in Table 2, as α is increased from 0 to 0.5, about half the users see a higher throughput (e.g., 3 times higher around the 95 percentile) at the cost of the other half seeing a lower throughput (e.g., 3 times lower around the 5 percentile). Similarly other tradeoffs between fairness and cell throughput can be obtained by varying α, or by engineering other utility functions with desired slopes.

6.1.3. Future Work. It is clear that rate CDFs in Figure 5 are optimal in that these cannot be dominated by the rate CDFs under any other scheduler (i.e., throughput of a user can only be improved at the cost of that of another). While the above simulation shows that the rate CDF can be controlled to a good degree by varying the utility function, still other more interesting scheduling objectives are, for example,

(i) deliver at least a minimum average rate $\hat{x}_i$ to each user i, or

(ii) maximize a Utility function under minimum and maximum rate constraints.

Both these objectives can be met by devising appropriate utility functions that sharply increase at the minimum rate constraint and saturate at the maximum rate constraint. However, as briefly discussed in Subsection 2.1, these objectives can also be met using queue- and channel-aware schedulers augmented with *virtual* token queues. Such schedulers have been shown to offer greater control over the rate CDF [30, 52]. It would be interesting to obtain throughput numbers under these latter scheduling frameworks too.

6.2. Multirate Streaming Video

6.2.1. Model. The deployment model is as described in Section 5.1, with only 1 cell having 20 users. Therefore, the SNRs (induced by the path-loss and shadowing models) of the 20 users have the *same* empirical CDF as the SINR CDF of users in a multicell system (see Figure 4). Let $\overline{\gamma}_i$ denote the SNR (induced by the path-loss and shadowing models) of user i. We index the users in increasing order of $\overline{\gamma}_i$, that is, we have $\overline{\gamma}_1 < \overline{\gamma}_2 < \cdots < \overline{\gamma}_{20}$.

The ith user's queue at eNB is fed by a video stream (see Section 5.3) with mean rate λ_i, and the transmissions are scheduled according to EDF, Log, or Exp rules described in Section 4.2. The parameters for each scheduler are fixed for a (soft) 99 percentile packet delay target of 250 milliseconds. We present results for two different operational scenarios.

(a) *Load is 0.50 bps/Hz:* $\lambda_i = 90$ kbps for $i \in \{1, \ldots, 6\}$ and $\lambda_i = 360$ kbps for $i \in \{7, \ldots, 20\}$. That is, the mean rate of the video stream for the six lowest SNR users is 90 kbps, whereas, the mean rate of the video stream for the remaining fourteen users is 360 kbps.

(b) *Load is 0.64 bps/Hz*: $\lambda_i = 360$ kbps for all users $i \in \{1, \ldots, 20\}$.

Figure 6 gives the plot of λ_i for system load given in (a) and λ_i for system load given in (b) versus $\overline{\gamma}_i$ for each user $i \in \{1, \ldots, 20\}$. In order to better picture the system load, let us define the theoretical throughput $\overline{x}_i$ that each user $i \in \{1, 2, \ldots, 20\}$ will see over an AWGN channel under equal resource splitting and saturated queues, that is, $\overline{x}_i \equiv (S_{\text{data}}/S)(BW/20) \times \log(1 + \overline{\gamma}_i)$; (we note that this is roughly equal to the throughput users see under PF scheduling assuming infinitely back logged queues as in Section 6.1, that is, the gain due to opportunistic PF scheduling evens out

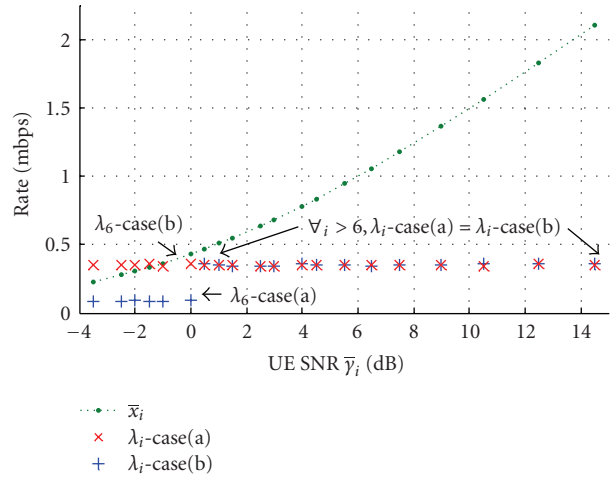


Figure 6: Mean arrival rates into the queues at eNB for operational scenarios (a) and (b) versus users' SNRs induced by path-loss model.

the loss due to the errors and delays in CQI reports as well as errors in rate prediction). Figure 6 also gives a plot of $\overline{x}_i$ versus $\overline{\gamma}_i$ for $i \in \{1, \ldots, 20\}$. For example, for the 6th user, rate λ_6-case(a) $= 90$ kbps ≈ 0.09 mbps, rate λ_6-case(b) $= 360$ kbps ≈ 0.35 mbps and rate $\overline{x}_6 = 0.43$ mbps are plotted against SNR $\overline{\gamma}_6 = 0$ dB.

6.2.2. Results and Discussion. Recall that the EDF scheduler is not throughput optimal nor opportunistic. However, in the case (a) above, each λ_i is chosen small enough for EDF scheduler to be stable; this, of course, does not guarantee that EDF will meet the QoS target of having the 99 percentile packet delay of less than 250 milliseconds. (The vector $(\lambda_i - \text{case(a)} : 1 \leq i \leq 20)$ can be shown to lie in the capacity region achievable under non-opportunistic schedulers.) In fact, the mean and the 99 percentile packet delays of all users under EDF scheduler turn out to be around 670 milliseconds and 1325 milliseconds, respectively. However, under the opportunistic Log and Exp schedulers, all users comfortably meet their delay targets: Figure 7 shows the mean and 99 percentile packet delays of each user and overall system under Log and Exp schedulers. The delay target of 250 milliseconds is about ten times the channel coherence time and we see that for a reasonable system load, opportunistic scheduling greatly increases the number of QoS flows that can be admitted; (flows with tighter delay constraints are considered in the following subsection).

The results get more favorable to the Log rule as the system load increases to that mentioned in case (b) above (see Figure 7). QoS degrades more gracefully under the Log rule, in that 1 user under the LOG rule versus 19 under the Exp rule miss the soft delay target of 250 milliseconds. However, Exp rule still maintains a lower delay spread *across* users than the Log rule. Clearly, the Exp rule's strong bias toward balancing delays is excessively compromising the realized throughput, and eventually the mean delays and tails for almost all users. Although Exp rule asymptotically

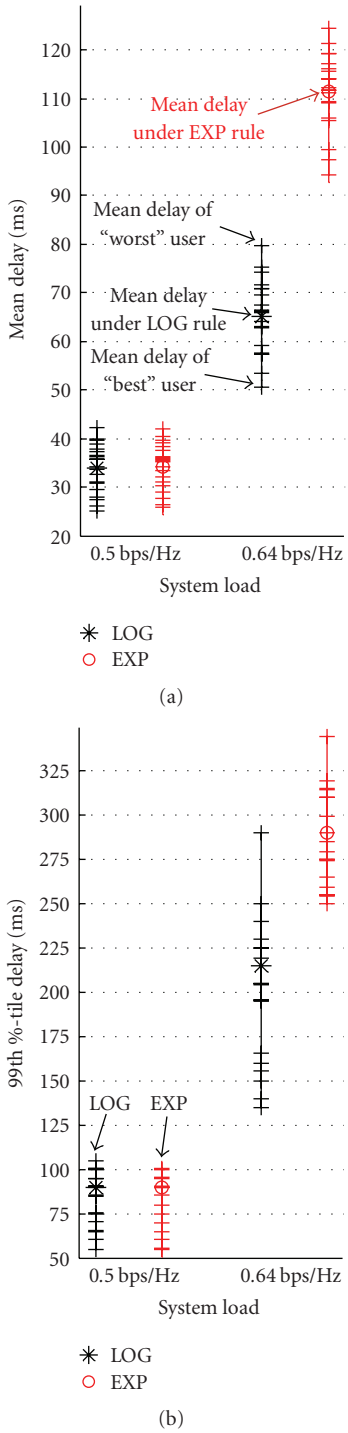


FIGURE 7: Users' and overall (left) mean delays and (right) 99 percentile delays under LOG and EXP rules for two different system loads. Each (+)-tick represents a user's delay and legend markers represent overall delays.

minimizes the exponential decay rate of the max-queue distribution irrespective of the values of parameters a_i, the pre-exponent must also be playing a role in determining the systems performance. The actual performance over the region of interest (not the theoretical asymptotic tail) achieved by the Exp rule is more sensitive to the values a_i. The RSM property of the Log rule naturally calibrates the scheduler to increased load. So unless parameters can be carefully tuned to possibly changing loads and unpredictable channel capacities, the Log rule appears to be more robust a scheduling policy. Intuitively, this is what one would expect from optimizing for the average/overall versus worst case asymptotic tail (see Section 2.1).

Suppose the aforementioned simulations also had best effort flows which were scheduled only using the resources spared by the streaming video flows. In that case, it is desirable for a QoS scheduler to meet the delay targets of streaming flows by utilizing fewer resources. Table 3 gives the resource utilization, that is, average number of RBs allocated to streaming flows per subframe, under each scheduler considered earlier. So, for example, borrowing the cell throughput figure of 1.02 bps/Hz for PF scheduling from Table 2, the total throughput seen by the best effort flows in case (a) can be expected to be about 2 mbps under the LOG rule which is about 7% higher than that expected under the Exp rule.

6.3. Mix of Live and Streaming Video

6.3.1. Model. Except for the traffic model, the system is identical to the one described earlier, that is, the streaming video simulation. The traffic model is as follows. As before, the users are indexed in increasing order of SNR $\overline{\gamma}_i$. Then the queue at the eNB of each (odd) user $i \in \{1, 3, \ldots, 19\}$ is fed by a streaming video source (see Section 5.3), whereas the queue for each (even) user $j \in \{2, 4, \ldots, 20\}$ is fed by a live video source. Video rates of each user are described later with the results. The 99 percentile delay target for live video flows is 80 milliseconds, whereas the target for streaming is 250 milliseconds as before. Transmissions are scheduled according to Log and Exp rules in two different manners.

(i) *Strict Priority Given to Live Video Flows.* Live video flows are scheduled first (according to Log and Exp rules with parameters set according to the delay target of 80 ms), if any RBs are left over after scheduling the live video flows, those are allocated to the streaming flows (again using Log and Exp rules with parameters set according to the delay target of 250 milliseconds). This scheduling method will be referred to as priority-Exp and priority-Log rules.

(ii) *All Flows Compete for Resources.* Live video flows are not prioritized in order of scheduling. Setting of scheduler parameters is described later with the results. Since resources are completely shared by the two classes of flows, this scheduling method will be referred to as complete-sharing, and written as cs-Exp and cs-Log rule for short.

6.3.2. Results and Discussion. We first determine by trail the highest arrival rate that all live video flows can be set to while still meeting the delay targets under both the priority-Exp and priority-Log rules. This turns out be around 200 kbps. The detailed results from this trial are not shown, however, we present the following interesting observation:

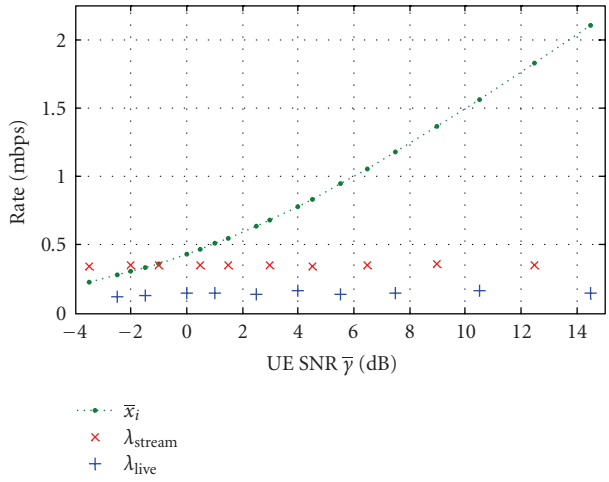


Figure 8: Users' SNRs induced by path-loss model, throughput expected under PF scheduling, and mean rates of live and streaming video flows.

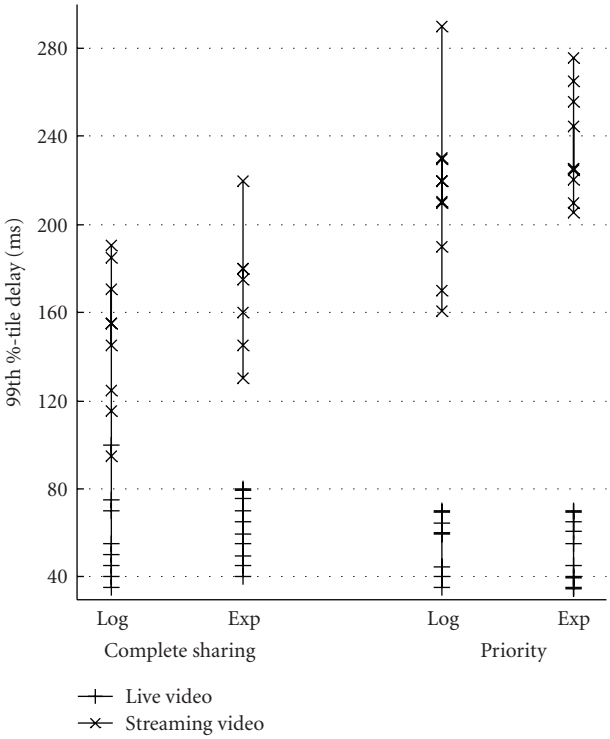


Figure 9: 99 percentile delays under (left two curves) cs-Log and cs-Exp rules, (right two curves) priority-Log and priority-Exp rules. Each (+)-tick represents a live video user's delay and each (x)-tick represents streaming video user's delay.

even though the channel is loaded to its capacity under the priority-Exp and priority Log rules when all live video flows are set to 200 kbps, we find that the system can still admit up to 10 streaming video users (5 higher SNR users at rate 360 kbps and 5 lower SNR users at 90 kbps) under priority-Exp and priority-Log rules while meeting their delay targets of 250 milliseconds. This is because the capacity (in terms of number of users that can be supported) of a time-varying channel is constrained by the delay targets: the longer the delay targets, the greater the opportunity to wait for a good channel thus exploiting opportunistic gain.

Since we want to mix live and streaming flows (the former with a much tighter delay deadline than the latter) and investigate the pros and cons of *priority* versus *complete-sharing* scheduling, we set the arrival rate of each live video flow to $\lambda_{\text{live}} = 150$ kbps (instead of a maximum possible of 200 kbps) to make the problem interesting. That is, for each user $i \in \{2, 4, \ldots, 20\}$ we have $\lambda_i = 150$ kbps. Next, we set the arrival rate for all streaming video flows to $\lambda_{\text{strm}} = 360$ kbps, that is, for each user $i \in \{1, 3, \ldots, 19\}$ we have $\lambda_i = 360$ kbps. Figure 8 illustrates λ_i and $\overline{x}_i$ versus $\overline{\gamma}_i$ for each user $i \in \{1, 2, \ldots, 20\}$.

Figure 9 shows the 99 percentile delays seen by both live and streaming video flows under priority-Log and priority-Exp rule (right two curves). Under these priority schedulers, while all live video users clearly meet their delay targets, 1 streaming video user under priority-Log while 3 under priority-Exp rule miss their soft delay targets of 250 milliseconds: the resources left over after scheduling live video users prove too scarce to meet the delay targets of all streaming video users. The question naturally arises: will delay performance improve if, instead of strictly prioritizing live video users, the users are *opportunistically* prioritized by using the parameters in each scheduler and letting all users compete for resources?

We find that when scheduler parameters for each user are set according to their delay targets, both cs-Log and cs-Exp rules comfortably meet the delay targets for streaming video users but fail for three or four live video users by up to 30 milliseconds (these results are not plotted). This is not desirable since streaming video delay targets are soft and if a scheduler must degrade performance a little, it should pick a streaming video user for that. While the priority schedulers were giving insufficient resources to the streaming flows, the complete-sharing schedulers are giving insufficient resources to the live flows.

Both cs-Log and cs-Exp rules can be made to give higher priority to the live video users by, for example, setting the parameters of live video users for a delay target of lower than 80 ms. Indeed, when the scheduler parameters of live video users are set according to the delay target of 50 milliseconds for Exp rule and 10 milliseconds for Log rule, all users meet their delay targets under cs-Exp rule, whereas, all but 1 live video users do under cs-Log rule (see Figure 9, left two curves). Table 4 gives resource utilization under each scheduler and shows that cs-Log rule makes available the most resources for any best effort users in the system, although by a small margin.

We conclude that although complete-sharing scheduling involves more complexity (due to the need for correctly setting relative priority of different classes), it not only reduces system utilization but it also improves system capacity in terms of number of users that can be supported. Infact, in a slightly different setting, [62] quantifies the capacity gains due to a candidate complete-sharing scheduler presented

Table 3: Resource utilization under various schedulers and system loads.

Scheduler	Utility under 0.50 bps/Hz	Utility under 0.64 bps/Hz
LOG	52.3 RBs/subframe	63.8 RBs/subframe
EXP	53.1 RBs/subframe	63.9 RBs/subframe
EDF	63.9 RBs/subframe	—

Table 4: Resource utilization under various schedulers.

Scheduler	Priority scheduling	CS scheduling
LOG	59.8 RBs/subframe	59.6 RBs/subframe
EXP	60.2 RBs/subframe	59.8 RBs/subframe

therein, with the caveat that indeed as QoS requirements on real-time traffic become tighter, the opportunistic gain due to complete-sharing diminishes as, eventually, one would need to simply give strict priority to real-time traffic. While call setup/SIP traffic cannot be treated as having the same priority as, say, streaming video (see [1]), our simulations show that perhaps cs-Exp or cs-Log scheduler can used to appropriately prioritize the SIP traffic.

7. Conclusions

LTE is a purely scheduled system that allows dynamic scheduling for diverse traffic types including delay-sensitive flows. By leveraging recent results on resource allocation and scheduling, we design a practical LTE downlink scheduler and characterized its performance for three traffic scenarios, namely, full-buffer, streaming video (loose delay constraint), and mixed streaming and live video (tight delay constraint). We show that the proposed utility maximizing scheduler offers good control over the rate CDF for the full buffer case. Similarly, we show that Exp and Log rules can support a mix of QoS traffic while increasing system capacity in terms of number of users that can be supported and, at the same time, reducing resource utilization.

Having evaluated various scheduling policies with a simpler (although complete) design, future work includes the implementation of other interesting features offered by LTE specifications, for example, asynchronous and adaptive HARQ for downlink, power shaping, and frequency-selective scheduling. Moreover, new scheduling policies will be considered, for example, one that resembles Exp rule when sum-delay is small but resembles Log rule when sum-delay is large (see Figure 1) can perhaps keep the delay spread small across users while still offering graceful degradation of service when system load increases (due to changes in traffic or wireless channel.)

Acknowledgments

This work was performed while B. Sadiq was at Qualcomm Flarion Technologies. This research was supported in part by NSF grant CNS-0721532. The authors thank Shelley Gu, Shailesh Patil, Sundeep Rangan, Niranjan Ratnakar, and Siddharth Ray for their help in developing the LTE system simulation infrastructure for studying scheduling algorithms. The authors also thank Raja Bachu for many discussions on the LTE specification.

References

[1] M. Wernersson, S. Wanstedt, and P. Synnergren, "Effects of QOS scheduling strategies on performance of mixed services over LTE," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, September 2007.

[2] A. Pokhariyal, G. Monghal, K. I. Pedersen, et al., "Frequency domain packet scheduling under fractional load for the UTRAN LTE downlink," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 699–703, April 2007.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probability in the Engineering and Informational Sciences*, vol. 18, no. 2, pp. 191–217, 2004.

[4] S. Shakkottai and A. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," in *Analytic Methods in Applied Probability*, vol. 207 of *American Mathematical Society Translations, Series 2, A Volume in Memory of F. Karpelevich*, pp. 185–202, American Mathematical Society, Providence, RI, USA, 2002.

[5] B. Sadiq, S. J. Baek, and G. de Veciana, "Delay-optimal opportunistic scheduling and approximations: the Log rule," in *Proceedings of the 27th Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '09)*, pp. 1–9, April 2009.

[6] S. Shakkottai and T. Rappaport, "Research challenges in wireless networks: a technical overview," in *Proceedings of the 5th International Symposium on Wireless Personal Multimedia Communications (WPMC '02)*, vol. 1, pp. 12–18, October 2002.

[7] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[8] F. P. Kelly, A. K. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, 1998.

[9] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge, UK, 2005.

[10] J. M. Holtzman, "Asymptotic analysis of proportional fair algorithm," in *Proceedings of the 12th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '01)*, vol. 2, pp. F33–F37, San Diego, Calif, USA, September-October 2001.

[11] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," in *Proceedings of the 22nd Annual Joint Conference on the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 321–331, San Francisco, Calif, USA, March-April 2003.

[12] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Transactions on Wireless Communications*, vol. 3, no. 4, pp. 1250–1259, 2004.

[13] J. Huang, V. G. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," in *Proceedings of the Conference on Information Sciences and Systems (CISS '06)*, 2006.

[14] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, 2005.

[15] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 2, pp. 171–178, 2003.

[16] L. M. C. Hoo, B. Halder, J. Tellado, and J. M. Cioffi, "Multiuser transmit optimization for multicarrier broadcast channels: asymptotic FDMA capacity region and algorithms," *IEEE Transactions on Communications*, vol. 52, no. 6, pp. 922–930, 2004.

[17] Y. J. Zhang and K. B. Letaief, "Multiuser adaptive subcarrier-and-bit allocation with adaptive cell selection for ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1566–1575, 2004.

[18] H. Yin and H. Liu, "An efficient multiuser loading algor-ithm for OFDM-based broadband wireless systems," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM '00)*, 2000.

[19] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for OFDMA downlink systems," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT '06)*, pp. 1394–1398, July 2006.

[20] R. Madan, S. P. Boyd, and S. Lall, "Fast algorithms for resource allocation in cellular networks," to appear in *IEEE/ACM Transactions on Networking*.

[21] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," in *Proceedings of the 29th IEEE Conference on Decision and Control (CDC '90)*, vol. 4, pp. 2130–2132, December 1990.

[22] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power and server allocation in a multi-beam satellite with time varying channels," in *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '02)*, vol. 3, pp. 1451–1460, 2002.

[23] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," in *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '03)*, vol. 1, pp. 745–755, March-April 2003.

[24] V. J. Venkataramanan and X. Lin, "On wireless scheduling algorithms for minimizing the queue-overflow probability," submitted to *IEEE/ACM Transactions on Networking*.

[25] A. L. Stolyar, "Large deviations of queues sharing a randomly time-varying server," *Queueing Systems*, vol. 59, no. 1, pp. 1–35, 2008.

[26] B. Sadiq and G. de Veciana, "Large deviation sum-queue optimality of a radial sum-rate monotone opportunistic scheduler," submitted to *IEEE Transactions on Information Theory*.

[27] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *Wireless Networks*, vol. 8, no. 1, pp. 13–26, 2002.

[28] N. Chen and S. Jordan, "Throughput in processor-sharing queues," *IEEE Transactions on Automatic Control*, vol. 52, no. 2, pp. 299–305, 2007.

[29] N. Chen and S. Jordan, "Downlink scheduling with probabilistic guarantees on short-term average throughputs," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '08)*, pp. 1865–1870, Las Vegas, Nev, USA, March-April 2008.

[30] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proceedings of the 17th International Teletraffic Congress (ITC '01)*, 2001.

[31] R. Agarwal, V. Majjigi, R. Vannithamby, and J. M. Cioffi, "Efficient scheduling for heterogeneous services in OFDMA downlink," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '07)*, pp. 3235–3239, November 2007.

[32] M. Lerida, *Adaptive radio resource management for VoIP and data traffic in 3GPP LTE networks*, M.S. thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2008.

[33] M. Wernersson, S. Wänstedt, and P. Synnergren, "Effects of QOS scheduling strategies on performance of mixed services over LTE," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, Athens, Greece, September 2007.

[34] M. Gidlund and J.-C. Laneri, "Scheduling algorithms for 3GPP longterm evolution systems: from a quality of service perspective," in *Proceedings of the 10th IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '08)*, pp. 114–117, August 2008.

[35] H. Lei, M. Yu, A. Zhao, Y. Chang, and D. Yang, "Adaptive connection admission control algorithm for LTE systems," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2336–2340, May 2008.

[36] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM '07)*, pp. 2861–2864, September 2007.

[37] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "MAC scheduling scheme for VoIP traffic service in 3G LTE," in *Proceedings of the 66th IEEE Vehicular Technology Conference (VTC '07)*, pp. 1441–1445, Baltimore, Md, USA, September-October 2007.

[38] H. Wang and D. Jiang, "Performance comparison of control-less scheduling policies for VoIP in LTE UL," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '08)*, pp. 2497–2501, Las Vegas, Nev, USA, March-April 2008.

[39] A. Pokhariyal, G. Monghal, K. I. Pedersen, et al., "Frequency domain packet scheduling under fractional load for the UTRAN LTE downlink," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 699–703, Dublin, Ireland, April 2007.

[40] X. Ning, Z. Ting, W. Ying, and Z. Ping, "A MC-GMR scheduler for shared data channel in 3GPP LTE system," in *Proceedings of the 64th IEEE Vehicular Technology Conference (VTC '06)*, pp. 1–5, Montreal, Canada, September 2006.

[41] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio, "Dynamic packet scheduling performance in UTRA long term evolution downlink," in *Proceedings of the 3rd International Symposium on Wireless Pervasive Computing (ISWPC '08)*, pp. 308–313, May 2008.

[42] M. Assaad and A. Mourad, "New frequency-time scheduling algorithms for 3GPP/LTE-like OFDMA air interface in the downlink," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 1964–1969, May 2008.

[43] K. C. Beh, S. Armour, and A. Doufexi, "Joint time-frequency domain proportional fair scheduler with HARQ for 3GPP LTE systems," in *Proceedings of the 68th IEEE Vehicular Technology Conference (VTC '08)*, pp. 1–5, September 2008.

[44] M. Al-Rawi, R. Jantti, J. Torsner, and M. Sagfors, "Opportunistic uplink scheduling for 3G LTE systems," in *Proceedings of the 4th International Conference on Innovations in Information Technology (IIT '07)*, pp. 705–709, November 2007.

[45] F. D. Calabrese, P. H. Michaelsen, C. Rosa, et al., "Search-tree based uplink channel aware packet scheduling for UTRAN LTE," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 1949–1953, May 2008.

[46] C. U. Castellanos, D. L. Villa, C. Rosa, et al., "Performance of uplink fractional power control in UTRAN LTE," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2517–2521, May 2008.

[47] M. Anas, C. Rosa, F. D. Calabrese, P. H. Michaelsen, K. I. Pedersen, and P. E. Mogensen, "QoS-aware single cell admission control for UTRAN LTE uplink," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2487–2491, May 2008.

[48] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451–474, 2003.

[49] M. Andrews, "Instability of the proportional fair scheduling algorithm for HDR," *IEEE Transactions on Wireless Communications*, vol. 3, no. 5, pp. 1422–1426, 2004.

[50] A. Eryilmaz, R. Srikant, and J. R. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Transactions on Networking*, vol. 13, no. 2, pp. 411–424, 2005.

[51] C. Zhou and G. Wunder, "General stability conditions in wireless broadcast channels," in *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 675–682, September 2008.

[52] M. Andrews, L. Qian, and A. Stolyar, "Optimal utility based multi-user throughput allocation subject to throughput constraints," in *Proceedings of the 24th IEEE Annual Joint Conference of the Computer and Communications Societies (INFOCOM '05)*, vol. 4, pp. 2415–2424, March 2005.

[53] M. J. Neely, "Order optimal delay for opportunistic scheduling in multiuser wireless uplinks and downlinks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 5, pp. 1188–1199, 2008.

[54] M. Yavuz and D. Paranchych, "Adaptive rate control in high data rate wireless networks," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '03)*, vol. 2, pp. 866–871, March 2003.

[55] 3GPP TR 25.814, "Physical layer aspects for evolved universal terrestrial radio access (UTRA)," http://www.3gpp.org/.

[56] 3GPP TR 25.848, "Physical layer aspects of utra high speed downlink packet access," http://www.3gpp.org/.

[57] 3GPP 25.896, "Feasibility study for enhanced uplink for UTRA FDD," http://www.3gpp.org/.

[58] 3GPP2 C.R1002-0 Version 1.0, "CDMA2000 evaluation methodology," http://www.3gpp2.org/.

[59] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice-Hall, Upper Saddle River, NJ, USA, 2002.

[60] K. Brueninghaus, D. Astélyt, T. Salzer, et al., "Link performance models for system level simulations of broadband radio access systems," in *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '05)*, vol. 4, pp. 2306–2311, 2005.

[61] "NGMN radio access performance evaluation methodology," 2008.

[62] S. Patil and G. de Veciana, "Managing resources and quality of service in heterogeneous wireless systems exploiting opportunism," *IEEE/ACM Transactions on Networking*, vol. 15, no. 5, pp. 1046–1058, 2007.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 957159, 11 pages
doi:10.1155/2009/957159

Research Article

On the Information Rate of Single-Carrier FDMA Using Linear Frequency Domain Equalization and Its Application for 3GPP-LTE Uplink

Hanguang Wu,[1] Thomas Haustein (EURASIP Member),[2] and Peter Adam Hoeher[3]

[1] *COO RTP PT Radio System Technology, Nokia Siemens Networks, St. Martin Street 76, 81617 Munich, Germany*
[2] *Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany*
[3] *Faculty of Engineering, University of Kiel, Kaiserstraße 2, 24143 Kiel, Germany*

Correspondence should be addressed to Hanguang Wu, wuhanguang@gmail.com

Received 31 January 2009; Revised 25 May 2009; Accepted 19 July 2009

Recommended by Bruno Clerckx

This paper compares the information rate achieved by SC-FDMA (single-carrier frequency-division multiple access) and OFDMA (orthogonal frequency-division multiple access), where a linear frequency-domain equalizer is assumed to combat frequency selective channels in both systems. Both the single user case and the multiple user case are considered. We prove analytically that there exists a rate loss in SC-FDMA compared to OFDMA if decoding is performed independently among the received data blocks for frequency selective channels. We also provide a geometrical interpretation of the achievable information rate in SC-FDMA systems and point out explicitly the relation to the well-known waterfilling procedure in OFDMA systems. The geometrical interpretation gives an insight into the cause of the rate loss and its impact on the achievable rate performance. Furthermore, motivated by this interpretation we point out and show that such a loss can be mitigated by exploiting multiuser diversity and spatial diversity in multi-user systems with multiple receive antennas. In particular, the performance is evaluated in 3GPP-LTE uplink scenarios.

1. Introduction

In high data rate wideband wireless communication systems, OFDM (orthogonal frequency-division multiplexing) and SC-FDE (single-carrier system with frequency domain equalization), are recognized as two popular techniques to combat the frequency selectivity of the channel. Both techniques use block transmission and employ a cyclic prefix at the transmitter which ensures orthogonality and enables efficient implementation of the system using the fast Fourier transform (FFT) and one tap scalar equalization per subcarrier at the receiver. There has been a long discussion on a comparison between OFDM and SC-FDE concerning different aspects [1–3]. In order to accommodate multiple users in the system, OFDM can be straightforward extended to a multiaccess scheme called OFDMA, where each user is assigned a different set of subcarriers. However, an extension to an SC-FDE based multiaccess scheme is not obvious and it has been developed only recently, called single-carrier FDMA (SC-FDMA) [4]. (A single-carrier waveform can only be obtained for some specific sub-carrier mapping constraints. In this paper we do not restrict ourself to these constraints but refer SC-FDMA to as DFT-precoded OFDMA with arbitrary sub-carrier mapping.) SC-FDMA can be viewed as a special OFDMA system with the user's signal pre-encoded by discrete Fourier transform (DFT), hence also known as DFT-precoded OFDMA or DFT-spread OFDMA. One prominent advantage of SC-FDMA over OFDMA is the lower PAPR (peak-to-average power ratio) of the transmit waveform for low-order modulations like QPSK and BPSK, which benefits the mobile users in terms of power efficiency [5]. Due to this advantage, recently SC-FDMA has been agreed on to be used for 3GPP LTE uplink transmission [6]. (LTE (Long Term Evolution) is the evolution of the 3G mobile network standard UMTS (Universal Mobile Telecommunications System) defined by the 3rd Generation

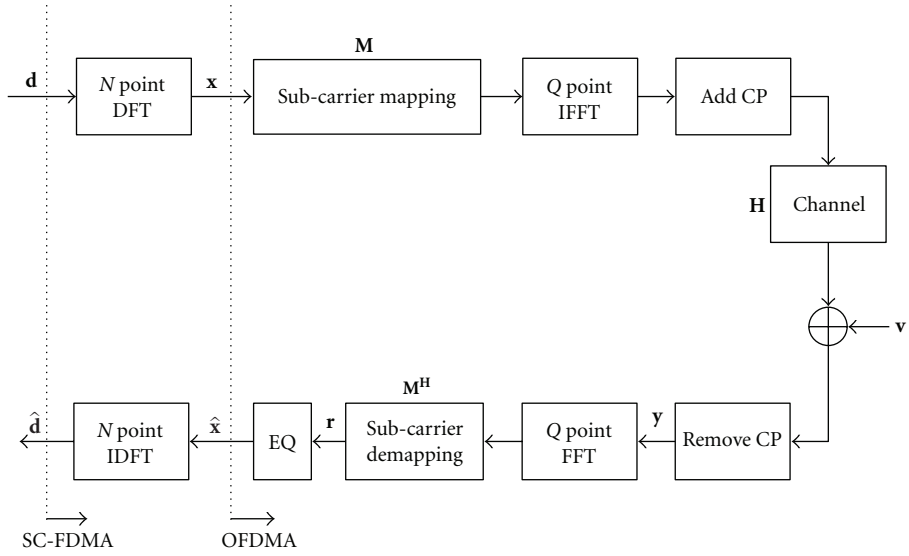


Figure 1: Block diagram of SC-FDMA systems and its relation to OFDMA systems.

Partnership Project (3GPP).) In order to obtain a PAPR comparable to the conventional single carrier waveform in the SC-FDMA transmitter, sub-carriers assigned to a specific user should be adjacent to each other [7] or equidistantly distributed over the entire bandwidth [8], where the former is usually referred to as localized mapping and the latter distributed mapping.

This paper investigates the achievable information rate using SC-FDMA in the uplink. We present a framework for analytical comparison between the achievable rate in SC-FDMA and that in OFDMA. In particular, we compare the rate based on a widely used transmission structure in both systems, where equal power allocation (meaning a flat power spectral density mask) is used for the transmitted signal of each user, and linear frequency domain equalization is employed at the receiver.

The fact that OFDMA decomposes the frequency-selective channel into parallel AWGN sub-channels suggests a separate coding for each sub-channel without losing channel capacity, where independent near-capacity-achieving AWGN codes can be used for each sub-channel and accordingly the received signal is decoded independently among the sub-channels. This communication structure is of high interest both in communication theory and in practice, since near-capacity-achieving codes (e.g., LDPC and Turbo codes) have been well studied for the AWGN channel. We show that although SC-FDMA can be viewed as a collection of virtual Gaussian sub-channels, these sub-channels are correlated; hence separate coding and decoding for each of them is not sufficient to achieve channel capacity. We further investigate the achievable rate in SC-FDMA if a separate capacity-achieving AWGN code for each sub-channel is used subject to equal power allocation of the transmitted signal. The special case that all the sub-carriers are exclusively utilized by a single user, that is, SC-FDE, is investigated in [3], and it is shown that the SC-FDE rate is always lower than the OFDM rate in frequency selective channels. However, an insight into the cause of the rate loss and its impact on the performance was not given. Such an insight is of interest and importance to design appropriate transmission strategies in SC-FDMA systems, where a number of sub-carriers and multi-users or possibly multiple antennas are involved. In this paper, based on the property of the circular matrix we derive a framework of rate analysis for SC-FDMA and OFDMA, which is a generalization of the result in [3], and it allows for the calculation of the achievable rate using arbitrary sub-carrier assignment methods in both the single user system and the multi-user system subject to individual power constraints of the users. We analyze the cause of the rate loss and its impact on the achievable rate as well as provide the geometrical interpretation of the achievable rate in SC-FDMA. Moreover, we reveal an interesting relation between the geometrical interpretation and the well-known waterfilling procedure in OFDMA systems. More importantly, motivated by this geometrical interpretation we show that such a loss can be mitigated by exploiting multi-user diversity and spatial diversity in the multi-user system with multiple receive antennas, which is usually available in mobile systems nowadays.

The paper is organized as follows. In Section 2 we introduce the system model and the information rate for OFDMA and SC-FDMA. In Section 3 we derive the SC-FDMA rate result and provide its geometrical interpretation assuming equal power allocation without joint decoding. Then we extend and discuss the SC-FDMA rate result for the multi-user case and for multi-antenna systems in Section 4. Simulation results are given in Section 5, and conclusions are drawn in Section 6.

2. System Model and Information Rate

Consider the SC-FDMA uplink transmission scheme depicted in Figure 1. The only difference from OFDMA is

the addition of the N point DFT at the transmitter and the N point IDFT at the receiver. The transmitted signal block $\mathbf{d} = [d_0, \ldots, d_{N-1}]^T$ of size N spreads onto the N sub-carriers selected by the sub-carrier mapping method. In other words, the transmitted signal vector is pre-encoded by DFT before going to the OFDMA modulator. For OFDMA transmission, a specific set of sub-carriers is assigned to the user through the sub-carrier mapping stage. Then multi-carrier modulation is performed via a Q point IFFT ($Q > N$), and a cyclic prefix (CP) longer than the maximum channel delay is inserted to avoid interblock interference. The frequency selective channel can be represented by a tap delay line model with the tap vector $\mathbf{h} = [h_0, h_1, \ldots, h_L]^T$ and the additive white Gaussian noise (AWGN) $\mathbf{v} \sim \mathcal{N}(0, N_0)$. At the receiver, the CP is removed and a Q point FFT is performed. A demapping procedure consisting of the spectral mask of the desired user is then applied, followed by zero forcing equalization which involves a scalar channel inversion per sub-carrier. For SC-FDMA, the equalized signal is further transformed to the time domain using an N point IDFT where decoding and detection take place.

In the following, we first briefly review the achievable sum rate in the OFDMA system and then show the sum rate relationship between OFDMA and SC-FDMA. We assume in the uplink that the users' channels are perfectly measured by the base station (BS), where the resource allocation algorithm takes place and its decision is then sent to the users via a signalling channel in the downlink. For simplicity, we start with the single-user single-input single-output system and then extend it to the multi-user case with multiple antennas at the BS. For convenience, the following notations are employed throughout the paper. $\mathbf{F_N}$ is the $N \times N$ Fourier matrix with the (n,k)th entry $[\mathbf{F_N}]_{n,k} = (1/\sqrt{N})e^{-j2\pi nk/N}$, and $\mathbf{F_N^H}$ denotes the inverse Fourier matrix. Further on, the assignment of data symbols x_n to specific sub-carriers is described by the $Q \times N$ sub-carrier mapping matrix $\mathbf{M}$ with the entry

$$m_{q,n} = \begin{cases} 1, & \text{if the } n\text{th data is assigned to the } q\text{th sub-carrier} \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

$0 \le q \le Q-1,\ 0 \le n \le N-1.$

2.1. OFDMA Rate. After CP removal at the receiver, the received block can be written as

$$\mathbf{y} = \mathbf{H}\mathbf{F_Q^H}\mathbf{M}\mathbf{x} + \mathbf{v}, \tag{2}$$

where $\mathbf{x} = [x_0, \ldots, x_{N-1}]$ is the transmitted block of the OFDMA system, and $\mathbf{H}$ is a $Q \times Q$ circulant matrix with the first column $\mathbf{h} = [h_0, \ldots, h_{L-1}, 0, \ldots, 0]^T$. The following discussion makes use of the important properties of circulant matrices given in the appendices (Facts 1 and 2). Performing multi-carrier demodulation using FFT and sub-carrier demapping using $\mathbf{M^H}$, we obtain the received block

$$\mathbf{r} = \mathbf{M^H}\mathbf{F_Q}\mathbf{y} = \mathbf{M^H}\mathbf{F_Q}\mathbf{H}\mathbf{F_Q^H}\mathbf{M}\mathbf{x} + \mathbf{M^H}\mathbf{F_Q}\mathbf{v} \tag{3}$$

$$= \mathbf{M^H}\mathbf{D}\mathbf{M}\mathbf{x} + \mathbf{M^H}\mathbf{F_Q}\mathbf{v} \tag{4}$$

$$= \mathbf{M^H}\mathbf{M}\boldsymbol{\Lambda}\mathbf{x} + \mathbf{M^H}\mathbf{F_Q}\mathbf{v} \tag{5}$$

$$= \boldsymbol{\Lambda}\mathbf{x} + \boldsymbol{\eta}, \tag{6}$$

where Fact 1 (see Appendix A) is used from step (3) to (4) and $\mathbf{D} = \mathbf{F_Q}\mathbf{H}\mathbf{F_Q^H} = \mathrm{diag}\{\widetilde{\mathbf{h}}\}$ with the diagonal entries being the frequency response of the channel. The step (4) to (5) follows from the equality

$$\mathbf{D}\mathbf{M} = \mathbf{M}\boldsymbol{\Lambda}, \tag{7}$$

where $\boldsymbol{\Lambda} = \mathrm{diag}\{\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_{N-1}\}$ is an $N \times N$ diagonal matrix with its diagonal entries being the channel frequency response at the selected sub-carriers of the user. This relationship can be readily verified since $\mathbf{M}$ has only a single nonzero unity entry per column, and this structure of $\mathbf{M}$ also leads to

$$\mathbf{M^H}\mathbf{M} = \mathbf{I_N}, \tag{8}$$

with which we arrive at step (6). The $N \times 1$ vector $\boldsymbol{\eta} = \mathbf{M^H}\mathbf{F_Q}\mathbf{v}$ is a linear transformation of $\mathbf{v}$, and hence it remains Gaussian whose covariance matrix is given by

$$E\{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathbf{H}}\} = E\left\{\mathbf{M^H}\mathbf{F_Q}\mathbf{v}\mathbf{v}^{\mathbf{H}}\mathbf{F_Q^H}\mathbf{M}\right\}$$

$$= \mathbf{M^H}\underbrace{\mathbf{F_Q}E\{\mathbf{v}\mathbf{v}^{\mathbf{H}}\}\mathbf{F_Q^H}}_{\mathrm{N_0I_Q}}\mathbf{M} \tag{9}$$

$$= N_0\mathbf{M^H}\mathbf{I_Q}\mathbf{M} \tag{10}$$

$$= N_0\mathbf{M^H}\mathbf{M}\mathbf{I_N} \tag{11}$$

$$= N_0\mathbf{I_N}, \tag{12}$$

where the step (9) to (10) follows from Fact 2 (see Appendix A), (10) to (11) follows from (7) since $\mathbf{I_Q}$ is also a diagonal matrix, and the step (11) to (12) results from (8). Therefore, $\boldsymbol{\eta}$ is a vector consisting of uncorrelated Gaussian noise samples. The frequency domain ZF equalizer is given by the inverse of the diagonal matrix $\boldsymbol{\Lambda}^{-1}$ which essentially preserves the mutual information provided that $\boldsymbol{\Lambda}$ is invertible. Here we assume that $\boldsymbol{\Lambda}$ is always invertible since the BS can avoid assigning sub-carriers with zero channel frequency response to the user. Due to the diagonal structure of $\boldsymbol{\Lambda}$ and independent noise samples of $\boldsymbol{\eta}$ (uncorrelated Gaussian samples are also independent), (6) can be viewed as the transmit signal components or the data symbols on the assigned sub-carriers propagating through independent Gaussian sub-channels with different gains. This structure suggests that coding can be done independently for each sub-channel to asymptotically achieve the channel capacity. The only loss is due to the cyclic prefix overhead relative to the transmit signal block length. The achievable sum rate of an

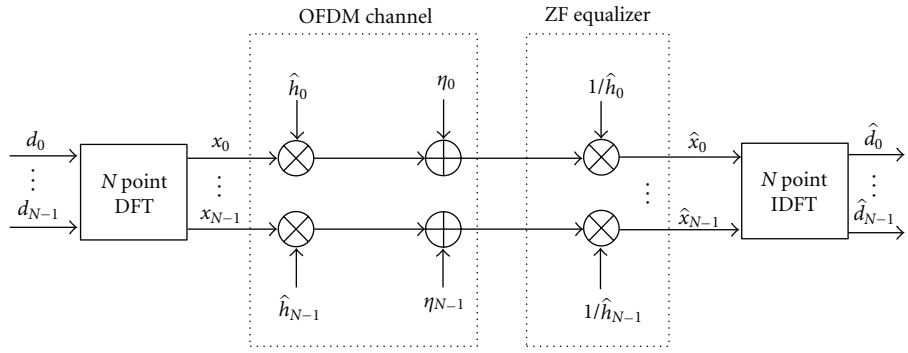


Figure 2: Equivalent block diagram of SC-FDMA systems.

OFDMA system can be calculated as the sum of the rates of the assigned sub-carriers, which is given by

$$C_{\text{OFDMA}} = \sum_{n=0}^{N-1} \log_2\left(1 + \frac{P_n \left|\hat{h}_n\right|^2}{N_0}\right), \tag{13}$$

where P_n is the power allocated to the nth sub-carrier. Note that the employment of a zero forcing (ZF) equalizer performing channel inversion for each sub-carrier preserves the capacity since the resulting signal-to-noise ratio (SNR) for each sub-carrier remains unchanged. To maximize the OFDMA rate subject to the total transmit power constraint P_{total}, the assignment of the transmit power to the n independent Gaussian sub-channels should follow the waterfilling principle, and so the optimal power P_n of the nth sub-carrier is given by

$$P_n = \max\left(0, \lambda - \frac{N_0}{\left|\hat{h}_n\right|^2}\right), \tag{14}$$

where the positive constant λ must be chosen in order to fulfill the total transmit power constraint

$$P_{\text{total}} = \text{tr}\{\mathbf{x}\mathbf{x}^{\mathbf{H}}\} = \sum_{n=0}^{N-1} \max\left(0, \lambda - \frac{N_0}{\left|\hat{h}_n\right|^2}\right), \tag{15}$$

where $\text{tr}\{\cdot\}$ stands for the trace of the argument. It should be noted that the waterfilling procedure implicitly selects the optimal sub-carriers out of the available sub-carriers in the system and assigns optimal transmit power to each of them. Therefore, it is possible that some sub-carriers are not used. In our model, the waterfilling procedure amounts to mapping $\mathbf{x}$ to the desired sub-carriers and at the same time constructing $\mathbf{x}$ having diagonal covariance matrix $\mathbf{R}_{\mathbf{x}} = \text{diag}\{P_0, P_1, \ldots, P_{N-1}\}$ with entries equal to the optimal power allocated to the desired sub-carriers.

2.2. SC-FDMA Rate. OFDMA converts the frequency selective channels into independent AWGN channels with different gains. Therefore, a block diagram of SC-FDMA can be equivalently regarded as applying DFT precoding for parallel AWGN channels and performing IDFT decoding after equalization as illustrated in Figure 2. The output of the IDFT can be derived as

$$\begin{aligned}\hat{\mathbf{d}} &= \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\mathbf{r} = \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\mathbf{x} + \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} \\ &= \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\mathbf{F}_{\mathbf{N}}\mathbf{d} + \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} \\ &= \mathbf{d} + \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} \\ &= \mathbf{d} + \hat{\boldsymbol{\eta}},\end{aligned} \tag{16}$$

where we denote $\hat{\boldsymbol{\eta}} = \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta}$ by the residual noise vector after ZF equalizer and IDFT. With (16) the transmit data components in SC-FDMA system can be viewed as propagating through virtual sub-channels distorted by the amount of noise given by $\hat{\boldsymbol{\eta}}$. Note that $\hat{\boldsymbol{\eta}}$ is a Gaussian vector due to the linear transformation but it is entries are generally correlated which we show in the following:

$$\begin{aligned}\mathbf{R}_{\hat{\eta}} &= E\left\{\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}^{\mathbf{H}}\right\} \\ &= E\left\{\mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^{\mathbf{H}}\boldsymbol{\Lambda}^{-\mathbf{H}}\mathbf{F}_{\mathbf{N}}\right\} \\ &= \mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}E\{\boldsymbol{\eta}\boldsymbol{\eta}^{\mathbf{H}}\}\boldsymbol{\Lambda}^{-\mathbf{H}}\mathbf{F}_{\mathbf{N}} \\ &= N_0\mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^{-\mathbf{H}}\mathbf{F}_{\mathbf{N}} \qquad (17) \\ &= N_0\underbrace{\mathbf{F}_{\mathbf{N}}^{\mathbf{H}}\underbrace{|\boldsymbol{\Lambda}|^{-2}}_{\text{diagonal}}\mathbf{F}_{\mathbf{N}}}_{\text{circulant}}, \qquad (18)\end{aligned}$$

where $|\cdot|$ is applied to $\boldsymbol{\Lambda}$ elementwise, and the step from (17) to (18) follows from the fact that $\boldsymbol{\Lambda}$ is a diagonal matrix. The matrix $|\boldsymbol{\Lambda}|^{-2}$ is hence also diagonal with the diagonal entries being the reciprocal of channel power gains of the assigned sub-carriers of the user, which are usually not equal in frequency selective channels. Hence $\mathbf{R}_{\hat{\eta}}$ is a circulant matrix according to Fact 2 (see Appendix A) with nonzero values on the off diagonal entries. Therefore, the residual noise on the virtual sub-channels is correlated and hence SC-FDMA does not have the same parallel AWGN sub-channel representation as OFDMA. However, note that the DFT at the SC-FDMA transmitter does not change the total transmit

power due to the property of the Fourier matrix $\mathbf{F^H F} = \mathbf{I}$, that is,

$$P_{\mathbf{x}} = \mathbf{x^H x} = \mathbf{d^H F^H F d} = \mathbf{d^H d} = P_{\mathbf{d}}. \tag{19}$$

The property of power conservation of the DFT precoder at the transmitter and invertibility of IDFT at the receiver leads to the conclusion that the mutual information is preserved. Hence, the mutual information between the transmit vector and post-detection vector $I(\mathbf{d}, \hat{\mathbf{d}})$ is equal to that of OFDMA $I(\mathbf{x}, \hat{\mathbf{x}})$. In other words, for any sub-carrier mapping and power allocation methods in OFDMA system, there exists a corresponding configuration in SC-FDMA which achieves the same rate as OFDMA. For example, suppose, for a given time invariant frequency selective channel, that $\mathbf{R_x}$ is the optimal covariance matrix given by the waterfilling solution in an OFDMA system. To obtain the same rate in an SC-FDMA system, the covariance matrix of the transmitted signal $\mathbf{R_d}$ can be designed as

$$\mathbf{R_d} = E\{\mathbf{dd^H}\} = E\{\mathbf{F^H x x^H F}\} = \underbrace{\mathbf{F^H} \underbrace{E\{\mathbf{xx^H}\}}_{\text{diag}\{\mathbf{p}\}} \mathbf{F}}_{\text{circulant}\{\hat{\mathbf{p}}\}}, \tag{20}$$

where in the last step we use Fact 2 (see Appendix A). Hence, $\mathbf{R_d}$ is a circulant matrix with the first column $\hat{\mathbf{p}} = (1/\sqrt{N})\mathbf{F^H p}$. Since both the covariance matrix of the transmitted signal and residual noise exhibit a circulant structure in an SC-FDMA system, correlation exists in both the transmitted symbols before DFT and the received symbols after IDFT. Such correlation complicates the code design problem in order to achieve the same rate as in OFDMA. This paper makes no attempt to design a proper coding scheme for SC-FDMA but we mention that SC-FDMA is not inferior to OFDMA regarding the achievable information rate from an information theoretical point of view. Instead, it can achieve the same rate as OFDMA if proper coding is employed. Note that the above statement implies using the same sub-carriers to convey information in both systems. Therefore, SC-FDMA and OFDMA are the same regarding the rate if they both use the same sub-carrier and the same corresponding power for each sub-carrier to convey information. However, in SC-FDMA coding and decoding should be applied across the transmitted and received signal components, respectively.

3. SC-FDMA Rate Using Equal Power Allocation without Joint Decoding

The waterfilling procedure discussed above is computationally complex which requires iterative sub-carrier and power allocation in the system. An efficient sub-optimal approach with reduced complexity is to use equal power allocation across a properly chosen subset of sub-carriers [9], which is shown to have very close performance to the waterfilling solution. In other words, this approach assumes $E\{\mathbf{xx^H}\} = (P_{\text{total}}/N)\mathbf{I} \triangleq P_e\mathbf{I}$ and designs a proper sub-carrier mapping matrix to approximate the waterfilling solution, where the number of used sub-channels N is also a design parameter. This approach can also be applied to an SC-FDMA system to approximate the waterfilling solution since DFT precoding and decoding are information lossless according to our discussion in Section 2. Note that DFT precoding does not change the equal power allocation property of the transmitted signal according to Fact 2 (see Appendix A), that is, $E\{\mathbf{dd^H}\} = E\{\mathbf{xx^H}\} = P_e\mathbf{I}$ ($P_{\mathbf{x}} = P_{\mathbf{d}} = P_{\text{total}}$). Therefore, to obtain the same rate as in OFDMA, coding does not need to be applied across transmitted signal components, and only correlation among the received signal components needs to be taken into account for decoding.

3.1. SC-FDMA Rate without Joint Decoding. We are interested to see what the achievable rate in SC-FDMA is if a capacity-achieving AWGN code is used for each transmitted component, which is decoded independently at the receiver. Under the above given condition, the achievable rate in SC-FDMA is the sum of the rate of each virtual subchannel for which we need to calculate the post-detection SNR, that is, the post-detection SNR of the nth virtual subchannel can be expressed as

$$\begin{aligned}
\gamma_{\text{SC-FDMA},n} &= \frac{P_e}{E\left\{\hat{\boldsymbol{\eta}}\hat{\boldsymbol{\eta}}^{\mathbf{H}}\right\}_{n,n}} = \frac{P_e}{N_0\left(\sum_{n=0}^{N-1}\left(1/\left|\hat{h}_n\right|^2\right)\right)/N} \\
&= \frac{NP_e}{N_0\left(\sum_{n=0}^{N-1}\left(1/\left|\hat{h}_n\right|^2\right)\right)} \\
&= \frac{HM\left(\hat{h}_n\right)\cdot P_e}{N_0} \qquad (21) \\
&= \gamma. \qquad (22)
\end{aligned}$$

In step (21) we denote $HM(|\hat{h}_n|^2) = N/\sum_{n=0}^{N-1}(1/|\hat{h}_n|^2)$ which is the harmonic mean of $|\hat{h}_n|^2, (n = 0, \ldots, N-1)$ by definition. In the last step we let $\gamma = (HM(|\hat{h}_n|^2)\cdot P_e)/N_0$ since the post-detection SNR is equal for all the virtual subchannels. Using Shannon's formula the achievable rate in SC-FDMA can be obtained as

$$\begin{aligned}
C_{\text{SC-FDMA}}^{\text{EP, Independent}} &= N\log_2(1+\gamma) \\
&= N\log_2\left(1 + \frac{HM\left(\left|\hat{h}_n\right|^2\right)\cdot P_e}{N_0}\right),
\end{aligned} \tag{23}$$

which is a function of the harmonic mean of the power gains at the assigned sub-carriers. Note that the result in [3] is a special case of (23) where all the available sub-carriers in the system are used by the user. It is perceivable that $C_{\text{SC-FDMA}}^{\text{EP,Independent}} \leq C_{\text{OFDM}}^{\text{EP}}$ because noise correlation between the received components is not exploited to recover

the signal. In the following, we will prove this inequality analytically. In order to prove

$$N\log_2\left(1+\frac{HM\left(\left|\hat{h}_n\right|^2\right)\cdot P_e}{N_0}\right)\le\sum_{n=0}^{N-1}\log_2\left(1+\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right), \tag{24}$$

it is equivalent to prove

$$\left(1+\frac{HM\left(\left|\hat{h}_n\right|^2\right)\cdot P_e}{N_0}\right)^N\le\prod_{n=0}^{N-1}\left(1+\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right), \tag{25}$$

since $\log_2(\cdot)$ is a monotonically increasing function. Because the term $(1+(|\hat{h}_n|^2P_e)/N_0)$ is positive and the geometric mean of positive values is not less than the harmonic mean, we have

$$\left(\prod_{n=0}^{N-1}\left(1+\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right)\right)^{1/N}\ge\frac{N}{\sum_{n=0}^{N-1}\left(1/\left(1+\left(\left|\hat{h}_n\right|^2P_e\right)/N_0\right)\right)}. \tag{26}$$

The Hoehn-Niven theorem [10] states the following: Let $HM(\cdot)$ be the harmonic mean and let $a_1,a_2,\ldots,a_m,x$ be the positive numbers, where the a_i's are not all equal, then

$$HM(x+a_1,x+a_2,\ldots,x+a_m)>x+HM(a_1,a_2,\ldots,a_m) \tag{27}$$

holds. If we let $a_n=(|\hat{h}_n|^2P_e)/N_0$, for all n and $x=1$, by applying (27) we have

$$\frac{N}{\sum_{n=0}^{N-1}\left(1/\left(1+\left(\left|\hat{h}_n\right|^2P_e\right)/N_0\right)\right)}>1+HM\left(\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right)=1+HM\left(\left|\hat{h}_n\right|^2\right)\frac{P_e}{N_0}, \tag{28}$$

where the last step follows from the fact that P_e/N_0 is a constant value so that it can be factored out of the $HM(\cdot)$ operation. Therefore, by applying the transitive property of inequality to (26) and (28) it follows that

$$\left(\prod_{n=0}^{N-1}\left(1+\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right)\right)^{1/N}>1+HM\left(\left|\hat{h}_n\right|^2\right)\frac{P_e}{N_0}, \tag{29}$$

and taking the Nth power on both sides of (29), we have

$$\left(1+\frac{HM\left(\left|\hat{h}_n\right|^2\right)\cdot P_e}{N_0}\right)^N<\prod_{n=0}^{N-1}\left(1+\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right). \tag{30}$$

By definition, it is easy to prove that if all the $|\hat{h}_n|^2, n=0,\ldots,N-1$ are equal, $HM(|\hat{h}_n|^2)=|\hat{h}_n|^2$ holds and thus

$$\left(1+\frac{HM\left(\left|\hat{h}_n\right|^2\right)\cdot P_e}{N_0}\right)^N=\prod_{n=0}^{N-1}\left(1+\frac{\left|\hat{h}_n\right|^2P_e}{N_0}\right) \tag{31}$$

holds, which corresponds to the case of frequency flat fading. Therefore, (24) holds in general.

The harmonic mean is sensitive to a single small value. $HM(|\hat{h}_n|^2)$ tends to be small if one of the values $|\hat{h}_n|^2$ is small. Therefore, the achievable sum rate in SC-FDMA depending on the harmonic mean of the power gain of the assigned sub-carriers would be sensitive to one single deep fade whose sub-carrier power gain is small. To give an intuitive impression how sensitive it is, we make use of the geometrical interpretation of the harmonic mean by Pappus of Alexandria [11] which is provided in Appendix B.

3.2. Relation to OFDMA. In the following, we will show that the achievable sum rate of SC-FDMA using equal power allocation without joint decoding is equivalent to that achieved by nonprecoded OFDMA system with equal gain power (EGP) allocation among the assigned sub-carriers. This conclusion will lead to our geometrical interpretation of the SC-FDMA system.

In an OFDMA system, the EGP allocation strategy pre-equalizes the transmitted signal so that all gains of the assigned sub-carriers are equal, that is,

$$P_n\frac{\left|\hat{h}_n\right|^2}{N_0}=\text{constant},\quad\forall n,\qquad\text{subject to }\sum_n P_n=P_{\text{total}}, \tag{32}$$

which requires the power allocated to the nth assigned sub-carrier $P_{eg,n}$ to be

$$P_{eg,n}=\frac{P_{\text{total}}}{\left|\hat{h}_n\right|^2\sum_{n=0}^{N-1}\left(1/\left|\hat{h}_n\right|^2\right)}. \tag{33}$$

Upon insertion of (33) into (13), the achievable sum rate using EGP can be calculated as

$$C_{\text{OFDMA}}^{\text{EGP}}=N\log_2\left(1+\frac{P_{\text{total}}}{\sum_{n=0}^{N-1}\left(1/\left|\hat{h}_n\right|^2\right)}\right)=N\log_2\left(1+\frac{HM\left(\left|\hat{h}_n\right|^2\right)\cdot P_e}{N_0}\right), \tag{34}$$

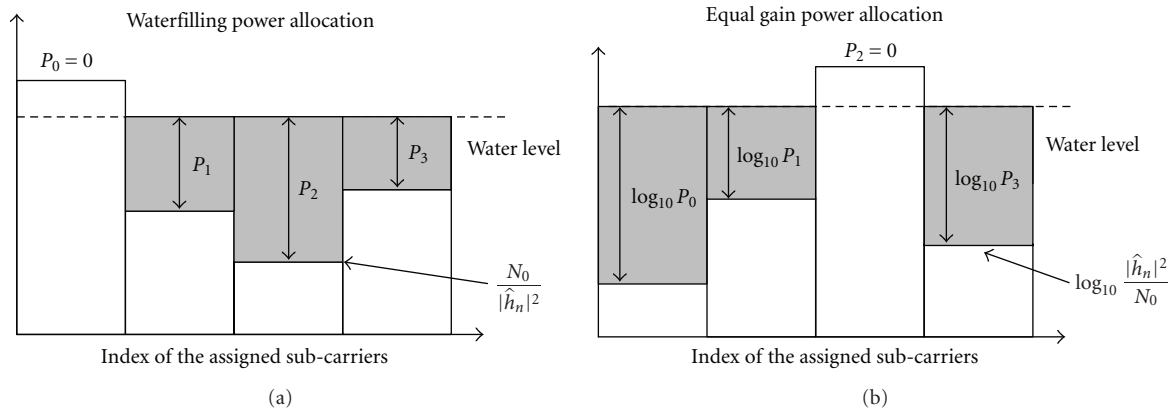


Figure 3: Comparison of geometrical interpretation between the waterfilling power allocation (a) and equal gain power allocation (b).

which is equal to $C_{\text{SC-FDMA}}^{\text{EP, Independent}}$ in (23), provided that both the SC-FDMA and OFDMA systems use the same assigned sub-carriers. This result leads to the conclusion that ZF equalized SC-FDMA with equal power allocation can be viewed as a nonprecoded OFDMA system performing EGP allocation among the assigned sub-carriers. It is worthy to point out that we find that EGP allocation shares a similar geometrical interpretation with waterfilling. This statement can be proven by applying logarithmic operation at both sides of the objective function of (32), which becomes

$$\begin{aligned} \log_{10} P_n + \log_{10}\left(\frac{\left|\hat{h}_n\right|^2}{N_0}\right) &= \log_{10}(\text{constant}) \\ &= \text{constant}, \quad \forall n \\ \text{subject to } \sum_n P_n &= P_{\text{total}}, \end{aligned} \tag{35}$$

where the objective function can be interpreted as shown in Figure 3(b): we can imagine that the quantity $\log_{10}(|\hat{h}_n|^2/N_0)$ is the bottom of a container and a fixed amount of water (power), P_{total}, is poured into the container. The water will then distribute inside the container to maintain a water level, denoted as constant in (35). Then the distance between the container bottom and the water level, that is, $\log_{10} P_n$, represents the power allocated to the nth assigned sub-carrier. Note that the waterfilling interpretation of EGP differs from the conventional waterfilling procedure of (14) in that firstly the container bottom is the inverse of that of the conventional waterfilling, and secondly the container bottom and the resulting power allocated to the individual sub-carrier should be measured in decibel. With the waterfilling interpretation of EGP it is possible to visualize how power is distributed among selected sub-carriers for ZF equalized SC-FDMA and also explain why putting power into weak sub-channels wastes so much capacity. Due to the inverse property of the container bottom, EGP allocates a larger portion of power to weaker sub-carriers and a smaller portion of power to stronger sub-carriers, which is opposite to the conventional waterfilling solution. Therefore, in order to achieve a higher data rate in SC-FDMA, it is important not to include weaker sub-carriers for communication because larger amount of power would be "wasted" in those sub-carriers. This observation suggests using strong sub-carriers for communication where an optimal sub-carrier allocation method, that is, optimal EGP allocation, is proposed in [12]. In frequency selective channels, such strong sub-carriers are usually not to be found adjacent to each other or equidistantly distributed over the entire bandwidth. Therefore, the sub-carrier mapping constraints to maintain the nice low PAPR for SC-FDMA has to be compromised if the optimal EGP allocation is applied. Within the scope of the work, we do not investigate such trade-off between the PAPR reduction and rate maximization. Instead, we will discuss in the following section that it is possible to obtain comparable rate performance as OFDMA and low PAPR as the single carrier waveform at the same time if multiple antennas are available at the BS.

4. Extension to Multiuser Case and Multiantenna Systems

The information rate analysis in Sections 2 and 3 assumes only one user in the system. However, the principle also holds for the multi-user case where each user's signal will be first individually precoded by DFT and then mapped to a different set of sub-carriers. It is known that in the multi-user OFDMA system, the maximum sum rate of all the users can be obtained by the multi-user waterfilling solution [13] where each user subject to an individual power constraint is assigned a different set of sub-carriers associated with a given power. Therefore, the information rate achieved in the system can be calculated as a sum of rate of each user, which can again be calculated similarly as in the single-user system. As a result, a multi-user SC-FDMA system can achieve the same rate as a multi-user OFDMA system since DFT and IDFT essentially preserve the mutual information of each user if the same resource allocation is assumed. If equal power allocation of the transmitted signal without joint decoding is assumed for each user, the system sum

rate $C_{\text{SC-FDMA,MU}}^{\text{EP,Independent}}$ of U users can be straightforward extended from (23), that is,

$$C_{\text{SC-FDMA,MU}}^{\text{EP, Independent}} = \sum_{u=1}^{U} N_u \log_2(1+\gamma_u) = \sum_{u=1}^{U} N_u \log_2\left(1+\frac{HM\left(\left|\hat{h}_{n,u}\right|^2\right)\cdot P_{n,u}}{N_0}\right), \tag{36}$$

where N_u is the length of the transmitted signal block of the uth user whose post-detection SNR is denoted as γ_u, $P_{n,u}$ is the power of the nth transmitted symbol of the uth user, and $\hat{h}_{n,u}$ is the channel frequency response at the nth assigned subcarrier of the uth user. The geometrical interpretation of the achievable sum rate in the multiuser SC-FDMA system can be straightforward interpreted as performing multiuser EGP allocation in the system, where each user, subject to a given transmit power constraint, performs EGP allocation in the assigned set of subcarriers. It can be proven that $C_{\text{SC-FDMA,MU}}^{\text{EP, Independent}} \le C_{\text{OFDMA,MU}}^{\text{EP}} = \sum_{u=1}^{U}\sum_{n=0}^{N_u-1}\log_2(1+(|\hat{h}_{n,u}|\cdot P_{n,u})/N_0)$ by summing up the rate of all the users, each of which obeys (24), where the equality occurs when the channel frequency response at the assigned sub-carriers of each user is equal; that is, each user experiences flat fading among the assigned subcarriers for communication but the channel power gains can be different for different users. Note that the optimal multi-user waterfilling solution tends to exploit multi-user diversity and schedule at any time and any subcarrier of the user with the highest sub-carrier power gain-to-noise ratio to transmit to the BS. Consequently, from the system point of view, only the relatively strong sub-carriers, possibly from different users, are selected and the relative weak ones are avoided. In other words, each user is only assigned a set of relative strong sub-carriers. It will be a good choice if the above sub-carrier allocation scheme is applied for each user in SC-FDMA systems, because it is essentially equivalent to performing EGP among the relative strong sub-carriers for each user. As the number of users increases, the weak sub-carriers can be more effectively avoided due to the multi-user diversity. As a result, the effective channel for each user becomes less frequency selective, and the rate loss in SC-FDMA compared to OFDMA becomes smaller. The same effect happens if the BS is equipped with multiple antennas to exploit the spatial diversity to harden the channels. For SC-FDMA with the localized mapping constraint or the equidistantly distributed mapping constraint, multi-user diversity may help to reduce the rate loss with respect to an OFDMA system but with less degrees of freedom because multi-user diversity cannot guarantee that good sub-carriers assigned to each user are adjacent to each other or equidistantly distributed in the entire bandwidth. In this case, spatial diversity is much more important because it can always reduce frequency selectivity of each user's channel by using, for example, a maximum ratio combiner (MRC) at the receiver. As a result, the user specific resource allocation has less influence on the achievable rate no matter which sub-carriers are selected by the users but only the number of sub-carriers assigned to each user is needed to be considered. Consequently, not only is the rate loss mitigated but also the multi-user resource scheduler is greatly simplified. As an additional advantage, SC-FDMA can offer lower PAPR than OFDMA with negligible rate loss.

Table 1: Parameter assumptions for simulation

Parameters	Assumption
Carrier frequency	2.0 GHz
Transmission bandwidth	1.25 MHz, 2.5 MHz, 5 MHz, 10 MHz, 15 MHz and 20 MHz
Subcarrier spacing	15 KHz
Number of subcarriers in the system	75, 150, 300, 600, 900 and 1200
Number of subcarriers per RB	12
Channel model	3GPP SCME urban macro [14]
Number of UEs	up to 6
Number of BSs	1
Antennas per UE	1
Antennas per BS	1, 2, 3
BS antenna spacing	10 wavelengths
UE velocity	10 m/s

5. Simulation Results

In this section, we evaluate the performance of SC-FDMA in terms of the average achievable rate in LTE uplink scenario according to Table 1, along with specific comparison with OFDMA. In the simulation, time slots are generated using the SCME "urban macro" channel model [14]. The total numbers of the available sub-carriers in the system are assumed to be 75, 150, 300, 600, 900, and 1200 with the same sub-carrier spacing of 15 KHz, which correspond to the 1.25 MHz, 2.5 MHz, 5 MHz, 10 MHz, 15 MHz, and 20 MHz bandwidth system defined in LTE, respectively. These sub-carriers are grouped in blocks of 12 adjacent sub-carriers, which are the minimum addressable resource unit in the frequency domain, also termed a resource block (RB). For simplicity, we assume that each RB experiences the same channel condition, and for simulation its channel frequency response is represented by the 6th sub-carrier of that RB. We further assume that the transmit power is equally divided in all the transmitted components and decoding performs independently among the received block. In all the simulations, the resulting achievable system sum rate is normalized by the corresponding system bandwidth; that is, system spectral efficiency (bits/s/Hz) is used as a metric for performance evaluation.

First we evaluate the impact of the used bandwidth on system spectral efficiency. We consider a single user system where all the available subcarriers in the system are occupied by the single user. Figure 4 compares the

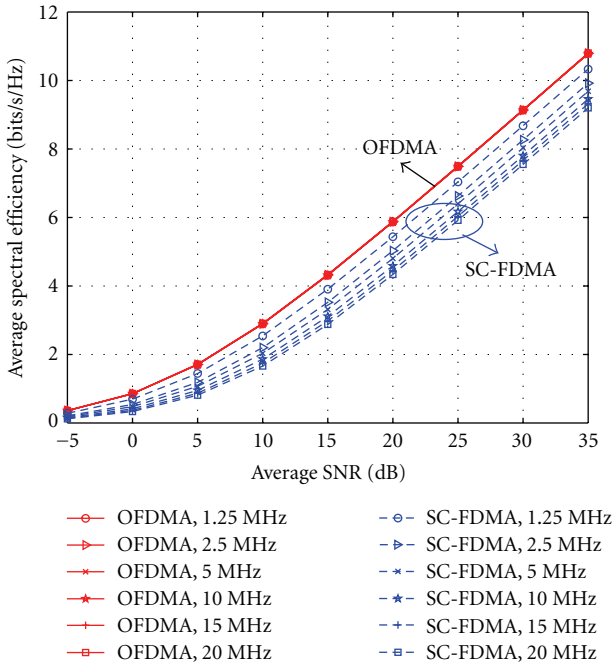


Figure 4: Comparison of the achievable information rate between OFDMA and SC-FDMA for different bandwidths under different average receive SNR conditions in the SCME "urban-macro" scenario with a single user in the system.

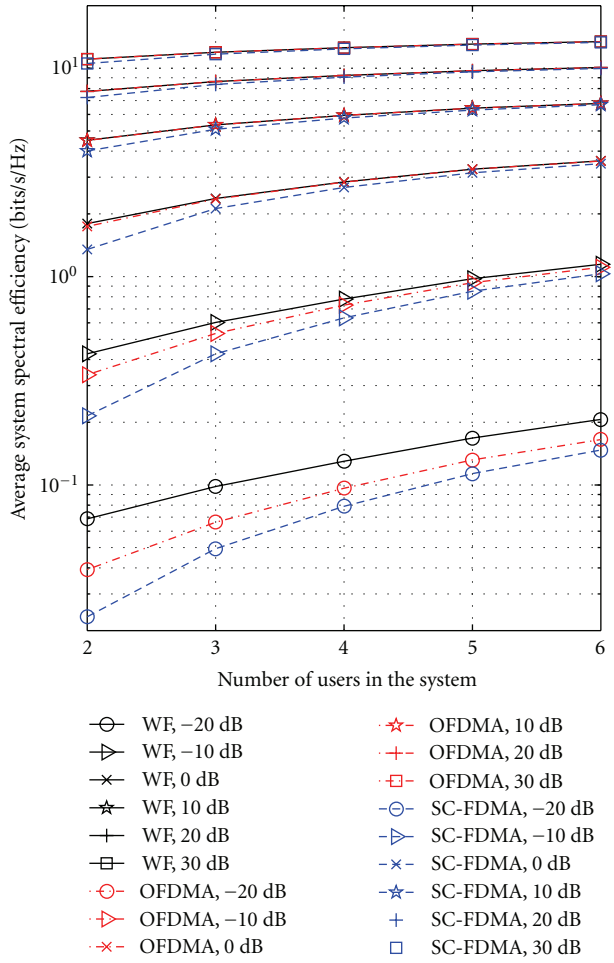


Figure 5: Comparison of the achievable system information rate between OFDMA and SC-FDMA for different numbers of users under different receive SNR conditions in SCME "urban-macro" scenario.

achievable average spectral efficiency between OFDMA and SC-FDMA for different transmission bandwidths under different average receive SNR conditions. It can be observed clearly that for the same average receive SNR, the average spectral efficiency for SC-FDMA is always smaller than that for OFDMA, which agrees very well with the analytical result presented in Section 3.1. Moreover, the achievable rate for OFDMA almost remains constant for different transmission bandwidths, while for SC-FDMA it decreases as the transmission bandwidth increases. This may due to the fact that as the transmission bandwidth increases and when it is much larger than the coherence bandwidth, each time slot consists of a similar number of weak subcarriers. Since the SC-FDMA rate is mainly constrained by channel deep fades (more power allocated for weak subcarriers and less power for good sub-carriers), having similar number of weak subcarriers for each time slot is less spectrally efficient than having more weak subcarriers for some time slots and less for the others, where the latter happens in the smaller bandwidth system with less frequency diversity. On the other hand, in the OFDMA system, transmit power is equally allocated in the used subcarriers; therefore, the achievable rate is insensitive to the distribution of the deep fades over different time slots.

Then we evaluate the impact of multi-user diversity on the system spectral efficiency. We assume that a number of users with the same transmit power constraints simultaneously communicate with the BS. Their path loss is compensated at the BS so that the average receive SNRs from all the users are the same, which varies from −20 dB to 30 dB in the 20 MHz bandwidth. First, the multi-user waterfilling (WF) algorithm [15] subject to the individual power constraint of the users is used to approximate the multi-user channel capacity, which gives a result close to the optimal power and subcarrier allocation solution for each user in the system. Then this subcarrier allocation solution which implicitly exploits multi-user diversity is adopted for simulations in both the OFDMA system and the SC-FDMA system but equal power allocation is used for the transmitted signal. Figure 5 plots the average system spectral efficiency over different numbers of users in both the SC-FDMA system and the OFDMA system under different receive SNR conditions. It can be seen that the average system spectral efficiency increases as the number of users in the system increases in both systems. Due to the multi-user diversity, the rate loss in SC-FDMA compared to OFDMA decreases as the number of users increases and it tends to disappear in high SNR conditions. It should be noted that the subcarrier allocation solution considered here is still suboptimal for both systems and a higher sum rate can be achieved in theory.

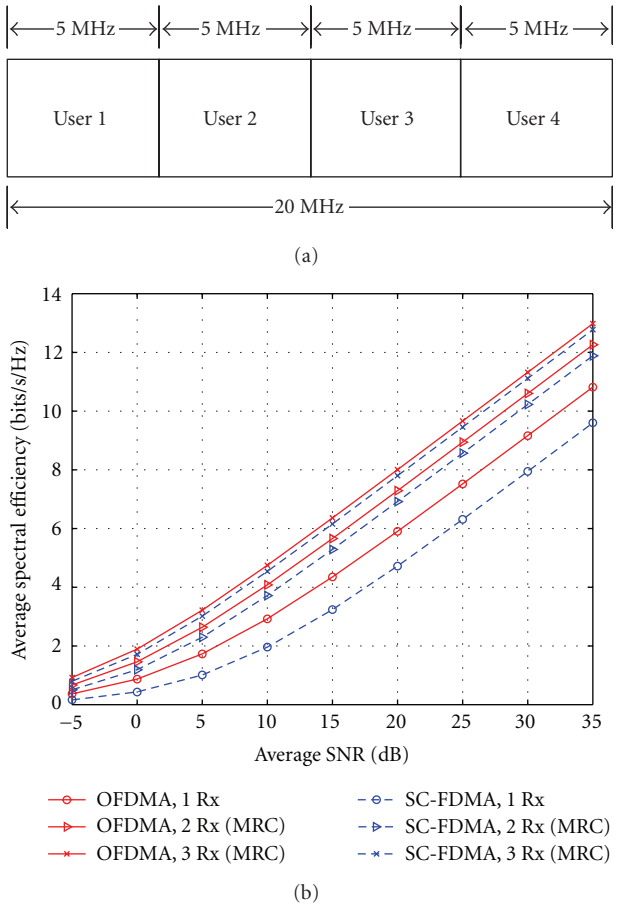


Figure 6: Comparison of the achievable information rate between OFDMA and SC-FDMA for different numbers of receive antennas in SCME "urban-macro" scenario. The system consists of 4 users with each occupying 5 MHz bandwidth.

Next, we evaluate the impact of spatial diversity on the system spectral efficiency. We consider that 4 users communicate simultaneously with the serving BS in the 20 MHz system, where each user occupies 5 MHz bandwidth as shown in the upper part of Figure 6. The number of receive antennas at the BS varies from 1 to 3. For multiple antennas, we assume that maximum ratio combining (MRC) is used in the frequency domain for both the SC-FDMA and OFDMA systems. It can be observed that as the number of receive antenna increases, the rate loss in SC-FDMA compared to OFDMA decreases significantly due to the channel hardening effect. Note that the simulation results have not taken into account the fact that SC-FDMA can further benefit from the lower PAPR property provided by the consecutive sub-carrier mapping for each user. Therefore, while being able to achieve a system sum rate very close to that in OFDMA, SC-FDMA has an additional lower PAPR advantage.

6. Conclusion

We have presented a framework for an analytical comparison between the achievable information rate in SC-FDMA and

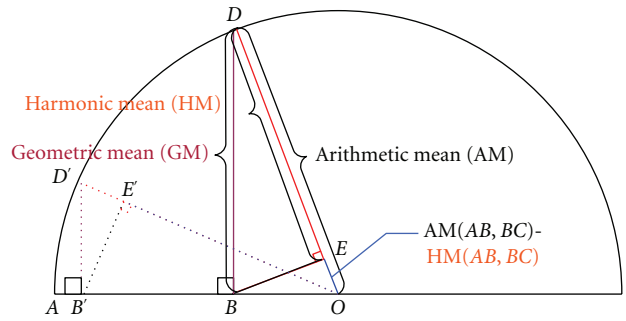


Figure 7: Geometrical interpretation of the harmonic mean, the arithmetic mean, and the geometric mean of $AB(AB')$ and $BC(B'C)$.

that in OFDMA. Ideally, SC-FDMA can achieve the same information rate as in OFDMA since DFT and IDFT are information lossless; however, proper coding across the transmitted signal components and decoding across the received signal components have to be used. We further investigated the achievable rate if independent capacity achieving AWGN codes is used and accordingly decoding is performed independently among the received components for SC-FDMA, assuming equal power allocation of the transmitted signal. A rate loss compared to OFDMA was analytically proven in the case of frequency selective channels, and the impact of the weak sub-carriers on the achievable rate was discussed. We also showed that the achievable rate in SC-FDMA can be interpreted as performing EGP allocation among the assigned sub-carriers in the nonprecoded OFDMA systems which has a similar geometrical interpretation with waterfilling. More importantly, it was pointed out and shown in 3GPP-LTE uplink scenario that the rate loss could be mitigated by exploiting multi-user diversity and spatial diversity. In particular, with spatial diversity we showed that while being able to achieve a system sum rate very close to that in OFDMA, SC-FDMA provides an additional lower PAPR advantage.

Appendices

A. Properties of the Circulant Matrix

Fact 1 ([16], Diagonalization of a circulant matrix). Denote $\mathbf{a}$ by the first column of a $Q \times Q$ circulant matrix $\mathbf{A}$ and $\mathrm{diag}\{\cdot\}$ by the diagonal matrix with the argument on the diagonal entries, then $\mathbf{A}$ can be diagonalized by pre- and postmultiplication with a Q-point FFT and IFFT matrices, that is, $\mathbf{F_Q A F_Q^H} = \mathbf{B} = \sqrt{Q}\,\mathrm{diag}\{\mathbf{F_Q a}\}$, where $\mathbf{B}$ is a $Q \times Q$ diagonal matrix with diagonal entries being a scale version of the Fourier transform of $\mathbf{a}$.

Fact 2. Because FFT and thus its matrix $\mathbf{F_Q}$ is invertible, it follows from Fact 1 that

$$\mathbf{A} = \underbrace{\mathbf{F_Q^H B F_Q}}_{\text{circulant }\{\mathbf{a}\}}, \tag{A.1}$$

where circulant $\{\mathbf{a}\}$ denotes a circulant matrix with the first column $\mathbf{a}$. Equation (A.1) means that a circulant matrix can be written as a multiplication of IFFT matrix, diagonal matrix, and FFT matrix. In particular, if and only if all entries of $\mathbf{B}$ are equal, then $\mathbf{A} = \mathbf{B}$ holds which is also a diagonal matrix with equal entries.

B. Geometrical Interpretation of the Harmonic Mean, the Arithmetic Mean, and the Geometric Mean

Suppose that we want to find out the harmonic mean of two values, represented by the length of AB and BC ($AB < BC$), respectively. By constructing a semicircle with radius $OC = (AB + BC)/2$ as depicted in Figure 7, the harmonic mean follows directly from the right-angled triangle DBO and DBE, where the length of DE equals the harmonic mean of AB and BC, denoted by $HM(AB,BC)$. It can be observed that $HM(AB,BC)$ is comparable to AB, the smaller value of the two. In order to show the sensitivity of harmonic mean to a small value, we can make AB much smaller than BC but keep $AB + BC$ fixed so that we can use the same semicircle to find their harmonic mean. Suppose that AB now becomes AB' and BC becomes $B'C$ as depicted in Figure 7, following the same way the harmonic mean of AB' and $B'C$, that is, $HM(AB',B'C)$ is given by $D'E'$. It can be seen that $D'E'$ is almost comparable to AB' which is the smaller value of the two. Therefore, the conclusion can be drawn that the harmonic mean is mainly constrained by the smaller value AB' although $B'C$ is much larger than AB'. For comparison purpose, the arithmetic mean and geometric mean of $AB(AB')$ and $BC(B'C)$, denoted by $DO(D'O)$ and $BD(B'D')$, respectively, are also drawn in Figure 7. It can easily be seen that if $AB = BC$, their harmonic mean is equal to their arithmetic mean and geometric mean.

References

[1] D. Falconer, S. L. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Communications Magazine*, vol. 40, no. 4, pp. 58–66, 2002.

[2] N. Wang and S. Blostein, "Comparison of CP-based single carrier and OFDMA with power allocation," *IEEE Transactions on Communications*, vol. 53, no. 3, pp. 391–394, 2005.

[3] T. Shi, S. Zhou, and Y. Yao, "Capacity of single carrier systems with frequency-domain equalization," in *Proceedings of the 6th IEEE Circuits and Systems Symposium on Emerging Technologies: Frontiers of Mobile and Wireless Communication (CAS '04)*, vol. 2, pp. 429–432, Shanghai, China, May-June 2004.

[4] H. G. Myung, "Introduction to single carrier FDMA," in *Proceedings of the 15th European Signal Processing Conference (EUSIPCO '07)*, Poznan, Poland, September 2007.

[5] H. Ekström, A. Furuskär, J. Karlsson, et al., "Technical solutions for the 3G long-term evolution," *IEEE Communications Magazine*, vol. 44, no. 3, pp. 38–45, 2006.

[6] 3GPP TR 25.814, "3GPP TSG RAN physical layer aspect for UTRA," v7.1.0.

[7] V. Jungnickel, T. Hindelang, T. Haustein, and W. Zirwas, "SC-FDMA waveform design, performance, power dynamics and evolution to MIMO," in *Proceedings of the IEEE International Conference on Portable Information Devices (PIDs '07)*, Orlando, Fla, USA, March 2007.

[8] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30–38, 2006.

[9] W. Yu and J. M. Cioffi, "Constant-power waterfilling: performance bound and low-complexity implementation," *IEEE Transactions on Communications*, vol. 54, no. 1, pp. 23–28, 2006.

[10] L. Hoehn and I. Niven, "Averages on the move," *Mathematics Magazine*, vol. 58, no. 3, pp. 151–156, 1985.

[11] S. Cuomo, *Pappus of Alexandria and the Mathematics of Late Antiquity*, Cambridge University Press, Cambridge, UK, 2000.

[12] H. Wu and T. Haustein, "Radio resource management for the multi-user uplink using DFT-precoded OFDM," in *Proceedings of the IEEE International Conference on Communications (ICC '08)*, pp. 4724–4728, Beijing, China, May 2008.

[13] R. S. Cheng and S. Verdú, "Gaussian multiaccess channels with ISI: capacity region and multiuser water-filling," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 773–785, 1993.

[14] D. S. Baum, J. Hansen, G. Del Galdo, M. Milojevic, J. Salo, and P. Kyösti, "An interim channel model for beyond-3G systems: extending the 3GPP spatial channel model (SCM)," in *Proceedings of the 61st IEEE Vehicular Technology Conference (VTC '05)*, vol. 2, pp. 3132–3136, Stockholm, Sweden, May-June 2005.

[15] C. Zeng, L. M. C. Hoo, and J. M. Cioffi, "Efficient water-filling algorithms for a Gaussian mutliaccess channel with ISI," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '00)*, Boston, Mass, USA, September 2000.

[16] G. H. Golub and C. F. van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, Md, USA, 3rd edition, 1996.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 720973, 18 pages
doi:10.1155/2009/720973

Research Article

Block Interleaved Frequency Division Multiple Access for Power Efficiency, Robustness, Flexibility, and Scalability

Tommy Svensson,[1] Tobias Frank,[2] Thomas Eriksson,[1] Daniel Aronsson,[3] Mikael Sternad (EURASIP Member),[3] and Anja Klein[2]

[1] *Department of Signals and Systems, Chalmers University of Technology, SE-412 96 Göteborg, Sweden*
[2] *Communications Engineering Laboratory, Technische Universität Darmstadt, 64283 Darmstadt, Germany*
[3] *Signals and Systems, Uppsala University, SE-751 21 Uppsala, Sweden*

Correspondence should be addressed to Tommy Svensson, tommy.svensson@chalmers.se

Received 1 February 2009; Revised 20 June 2009; Accepted 27 July 2009

Recommended by Cornelius van Rensburg

The multiple access solution in an IMT-Advanced mobile radio system has to meet challenging requirements such as high throughput, low delays, high flexibility, good robustness, low computational complexity, and a high power efficiency, especially in the uplink. In this paper, a novel multiple access scheme for uplinks denoted as B-IFDMA is presented. We show that this scheme is able to provide equal or better error rate performance than the Single-Carrier Frequency Division Multiple Access (SCFDMA) schemes IFDMA and LFDMA, when considering realistic channel estimation performance at the receiver and no reliable channel state information at the transmitter. We also show that B-IFDMA provides better amplifier efficiency than OFDMA and can provide better end-to-end energy efficiency than IFDMA and LFDMA. Moreover, the scheme shows a promisingly high robustness to frequency-offsets and Doppler spread. Thus, this scheme can be regarded as a promising solution for the uplink of future mobile radio systems.

1. Introduction

Future mobile communication systems need to efficiently support fully packet-based services with largely different requirements on data rates, ranging from a few kbps to hundreds of Mbps, and largely varying Quality of Service (QoS) requirements. The systems need to flexibly support deployment in various propagation scenarios ranging from isolated hot spots to wide area cellular, including support for high speed trains. In addition, they need to support deployment in various spectrum allocation scenarios with system bandwidths up to 100 MHz at a carrier frequency of several GHz, cf. [1–5]. These system requirements imply that the multiple access solution in an IMT-Advanced mobile radio system has many challenges to meet.

It has been shown feasible to implement a fully synchronous network, [6, 7]. Thus, resources can be allocated based on a chunk concept, where a chunk is a time-frequency resource unit. With multiple antennas, spatial reuse of chunks is enabled and denoted as chunk layers [2, 4, 5, 8, 9]. The chunk concept is adopted in 3GPP Long Term Evolution (LTE), where a chunk is denoted as Resource Block. The chunk size is chosen in such a way that it experiences essentially flat fading in its time-frequency extent, also in largely frequency selective channels and for users at vehicular speeds.

With channel quality information (CQI) available at the transmitter it is possible to adapt to the small-scale fading of the chunk resources, so-called frequency-adaptive (FA) transmission [9]. Adaptive Orthogonal Frequency Division Multiple Access (OFDMA) with a chunk-based Time Division Multiple Access (TDMA) component is such an FA multiple access scheme [9]. Adaptive TDMA/OFDMA can provide a large increase in the system capacity, also in presence of channel prediction errors due to gains in multiuser scheduling and chunk-wise link adaptation [10, 11]. This is very important for high cell load situations. FA transmission is best suited for scenarios with favorable

channel conditions such as high Signal to Interference and Noise Ratios (SINR), and reasonably low speeds [10]. FA is especially suited for transmission of rather large data volumes and high instantaneous data rates for low service latency. However, the FA scheme must be accompanied by a robust diversity based transmission mode, since FA transmission without reliable CQI can deteriorate.

The diversity-based scheme, here denoted as non-frequency-adaptive (NFA) transmission, should efficiently support users in all other usage scenarios, such as low SINR, high user equipment (UE) velocities, small and delay critical packet transfers, broadcasting that cannot benefit from a retransmission scheme, as well as for multicast transmission to multiple users with widely varying channels. In these scenarios a diversity based scheme has the potential to be more robust, more spectrally efficient and also more energy efficient.

Various relaying concepts are also considered in future wireless systems, [2–5]. However, multihop relaying increases the end-to-end delay in the Radio Access Network (RAN). Thus, an important requirement of the multiple access solution is to support a very low delay. This requirement also enables FA transmission at vehicular speeds even with a several GHz carrier frequency. It furthermore enables the use of retransmissions also for delay constrained services such as voice. However, such a low delay requirement implies a very short frame duration with very limited time diversity. Thus, the diversity for the NFA transmission scheme must come from the frequency domain and/or the spatial domain.

Below we summarize important requirements that we have identified for the NFA multiple access scheme.

(i) Robustness to small-scale fading without time diversity.

(ii) Tuneable degree of frequency-diversity.

(iii) Need to support high energy efficiency in the transmitters and the receivers.

(iv) Robustness to carrier frequency offsets and large Doppler spread.

(v) Support for widely varying packet sizes.

(vi) Enable efficient resource allocation.

(vii) Be of use for in-band control signals.

(viii) Enable efficient coexistence with adaptive TDMA/OFDMA.

(ix) Facilitate low complexity transmitter in UE.

To define a scheme that optimally fulfills all of these requirements at the same time is challenging, and a tradeoff is needed. In addition, the tradeoff would look different in different deployment and usage scenarios. Thus a flexible scheme is desirable that can be adjusted towards a good tradeoff in each scenario.

In this paper, we present a novel multiple access scheme denoted as Block-Interleaved Frequency Division Multiple Access (B-IFDMA), which is intended to fulfill the above requirements and also to provide a good tradeoff between them for NFA transmission in uplinks. We have briefly introduced the scheme in [12]. B-IFDMA is based on OFDMA. In B-IFDMA equidistantly frequency-separated blocks, each consisting of a few subcarriers, are allocated to each user. A Discrete Fourier Transform (DFT) precoding step is performed on each Orthogonal Frequency-Division Multiplexing (OFDM) symbol before transmission. In addition, a short TDMA component is introduced within the chunks. B-IFDMA is a generalization of DFT precoded OFDMA with interleaved subcarrier allocation, as described in [13], also denoted as Interleaved Frequency Division Multiple Access (IFDMA) in the original paper [14] or Single-Carrier Frequency Division Multiple Access (SC-FDMA) with distributed mapping [15, 16]. (Some authors distinguish between DFT precoded OFDMA and the original IFDMA scheme as the frequency domain generation and the time domain generation approaches, and regard them as different schemes with different performance by assuming that spectrum shaping is made in the corresponding domain. Here we regard the two schemes as equivalent.) B-IFDMA is also a generalization of Localized Frequency Division Multiple Access (LFDMA) [17], also denoted as Localized DFTS-OFDM or SC-FDMA with localized mapping, [15, 16]. In this paper we use the acronym IFDMA for SC-FDMA with distributed mapping and LFDMA for SC-FDMA with localized mapping. In contrast to IFDMA, B-IFDMA can assign adjacent subcarriers in the blocks, and in contrast to LFDMA multiple noncontiguous subcarriers can be assigned, see illustration in Figure 1.

The IFDMA scheme has been considered in the uplink of the LTE standard, but LFDMA was adopted [15, 16] in LTE Release 8. In LTE, with rather flat fading Resource Blocks (RBs), link adaptation and multiuser diversity gains can be obtained whenever reliable CQI is available. Some frequency-diversity collected over multiple slots can be obtained when needed through frequency-hopping, but at the cost of higher delay and delay jitter. To maintain a low RAN delay in a multihop relaying scenario, frequency hopping is less attractive.

Our evaluations in this paper of the B-IFDMA scheme towards the identified requirements for NFA transmission are focused on the *error rate performance*, *energy efficiency* and *robustness* of the scheme compared to OFDMA, IFDMA and LFDMA. We investigate the properties of the scheme under close to real conditions such as realistic pulse shaping and realistic power amplifiers, correlated Multiple Input Multiple Output (MIMO) mobile radio channels, realistic channel estimation performance under constraints set by a low pilot overhead loss and a realistic frame structure. Such a system is hard to analyze theoretically, but for application in IMT Advanced systems such a property analysis is of interest. Thus, the performance investigations in this paper are performed with simulations.

The investigations show that in an IMT-Advanced scenario, B-IFDMA provides equal or better error rate performance than the Single-Carrier Frequency Division Multiple Access (SC-FDMA) schemes IFDMA and LFDMA, when considering realistic channel estimation performance at the receiver and no reliable channel state information at the transmitter. We also show that B-IFDMA provides better

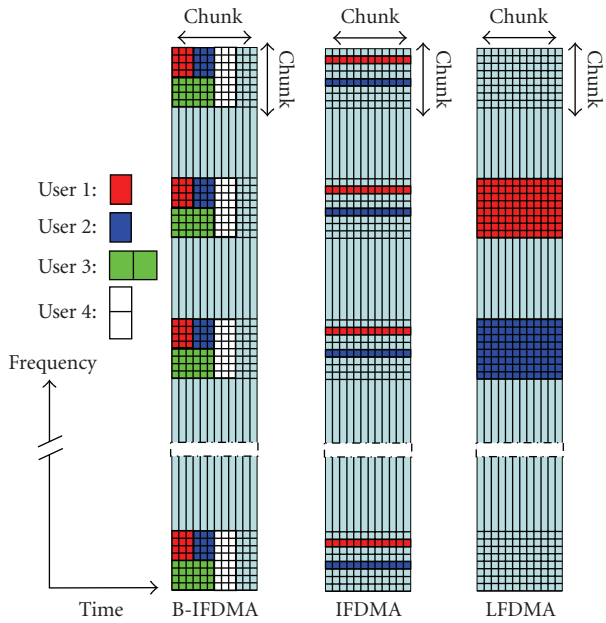


FIGURE 1: Illustration of B-IFDMA using $M = 4$ subcarriers and $N_t = 3$ OFDM symbols per subcarrier block within a time-frequency resource denoted as chunk. SC-FDMA with localized mapping (LFDMA) and SC-FDMA with distributed mapping (IFDMA) are shown for comparison. In B-IFDMA, high rate users are allocated more blocks within the chunks in either the time or the frequency direction. (A similar illustration is included in [12].)

amplifier efficiency than OFDMA and can provide better end-to-end energy efficiency than IFDMA and LFDMA. Moreover, the scheme shows a promisingly high robustness to frequency-offsets (CFOs) and Doppler spread (DS). Thus, this scheme can be regarded as a promising solution for the uplink of future mobile radio systems. (The B-IFDMA scheme has been adopted for the NFA uplink in the WINNER system concept [2, 4, 5]. A scheme similar to B-IFDMA denoted as Block Equidistant Frequency Division Multiple Access (B-EFDMA) has also been proposed for NFA downlinks [4, 5, 18]. The difference to B-IFDMA is that the DFT precoding step is not included, since the benefit of DFT precoding is lost in the multiple signal multiplexing in the downlink. The other benefits are similar as for B-IFDMA, including the possibility to time localize the transmission in the base station (BS) in low load situations, in order to save energy in both the BS and the UE. The B-EFDMA scheme has been adopted for the WINNER NFA downlink.)

This paper is organized as follows: we start in Section 2 with a detailed definition of B-IFDMA. Then, in Section 3 we investigate the error rate performance of B-IFDMA with perfect and nonperfect channel estimation at the receiver. These investigations show the capability of B-IFDMA to collect large diversity gains under realistic assumptions on channel estimation performance, also for rather low data rates, without using time-diversity. We proceed in Section 4 with the energy efficiency of B-IFDMA with respect to High Power Amplifier (HPA) performance and end-to-end energy efficiency. These investigations motivate the use of a DFT precoding step, and the integration of the TDMA component within the B-IFDMA scheme. These results also motivate the regular subcarrier allocation in B-IFDMA. In Section 5 we investigate the robustness of B-IFDMA to carrier frequency offsets and to Doppler spreads. These results show that B-IFDMA offers the possibility to combine robustness and provision of frequency diversity. In Section 6 we summarize our investigation results, and we comment on the suitability of B-IFDMA to meet our identified list of requirements above on the NFA uplink scheme. In Section 7 we conclude the paper.

2. System Model

As an introduction to B-IFDMA, the resource allocation for B-IFDMA is illustrated in Figure 1 along with IFDMA and LFDMA for comparison, assuming the Frequency Division Duplex (FDD) chunk size in [19]. The scheme is defined in detail in the subsequent sections.

2.1. Signal Definition. In this section, a transmitter signal model for B-IFDMA is given, following the block diagram in Figure 2. In the following, all signals are represented by their discrete time equivalents in the complex baseband. Upper case bold letters denote matrices and lower case bold letters denote column vectors. Further on, $(\cdot)^{\dagger}$ denotes the pseudo-inverse and $(\cdot)^{\mathrm{H}}$ the Hermitian of a matrix and $(\cdot)^{\mathrm{T}}$ the transpose of a vector or a matrix, respectively. Finally, $[\cdot]_{l,m}$ denotes the element of a matrix in the lth row and mth column.

An uplink transmission system with K users with user index k, $k = 0, \ldots, K-1$ is considered. Let $c_{\nu}^{(k)}$, $\nu \in \mathbb{Z}$, denote a sequence of data symbols of user k at symbol rate $1/T_s$ taken from the alphabet of an arbitrary bit mapping scheme applied after channel encoding and bit interleaving.

At first, the data symbols $c_{\nu}^{(k)}$ are grouped into data symbol vectors

$$\mathbf{d}_{\eta}^{(k)} = \left[d_{\eta,0}^{(k)}, \ldots, d_{\eta,Q-1}^{(k)}\right]^{\mathrm{T}} \tag{1}$$

with Q elements $d_{\eta,q}^{(k)} = c_{\eta\cdot Q+q}^{(k)}$, $q = 0, \ldots, Q-1$, $\eta \in \mathbb{Z}$. For sake of simplicity, throughout this section it is assumed that the number Q is the same for all users. However, note that for B-IFDMA also different numbers Q can be assigned to the users, cf. [20]. Each data symbol vector $\mathbf{d}_{\eta}^{(k)}$ is precoded by a DFT represented by a $Q \times Q$ matrix $\mathbf{F}_{\mathrm{Q}}$ with elements

$$\left[\mathbf{F}_{\mathrm{Q}}\right]_{p,q} = \frac{1}{\sqrt{Q}} \cdot e^{-j(2\pi/Q)pq}, \quad p, q = 0, \ldots, Q-1. \tag{2}$$

After DFT precoding, the Q elements of the vector $\mathbf{F}_{\mathrm{Q}} \cdot \mathbf{d}_{\eta}^{(k)}$ are mapped to a set of Q out of $N = K \cdot Q$ subcarriers available in the system. The mapping is performed in a block-interleaved manner. Let M denote the number of subcarriers in each subcarrier block, L denote the numbers of subcarrier

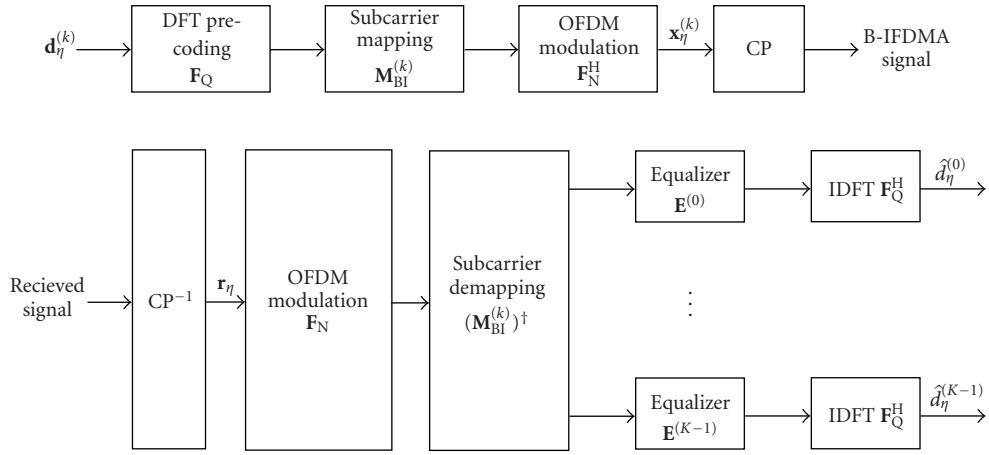


Figure 2: B-IFDMA transceiver, transmitter (top) and receiver (bottom). In case the same amount of resources are allocated per user, for each user k out of K uplink user terminals, Q out of N subcarriers are allocated by the subcarrier mapping matrix $\mathbf{M}_\mathrm{BI}^{(k)}$. The allocated subcarriers consist of L blocks, each containing M adjacent subcarriers.

blocks and let $Q = M \cdot L$. The block-interleaved mapping can be described by an $N \times Q$ matrix $\mathbf{M}_\mathrm{BI}^{(k)}$ with elements

$$\left[\mathbf{M}_\mathrm{BI}^{(k)}\right]_{n,q} = \begin{cases} 1, & n = l \cdot \dfrac{N}{L} + m + kM, \\ 0, & \text{else}, \end{cases} \tag{3}$$

where $l = 0, \ldots, L-1$, $m = 0, \ldots, M-1$, and $q = m + l \cdot M$. After subcarrier mapping, OFDM modulation is applied. The OFDM modulation is performed by an N-point Inverse DFT (IDFT) represented by matrix $\mathbf{F}_\mathrm{N}^\mathrm{H}$ with elements

$$\left[\mathbf{F}_\mathrm{N}^\mathrm{H}\right]_{n,\mu} = \frac{1}{\sqrt{N}} \cdot e^{j(2\pi/N)n\mu}, \quad n, \mu = 0, \ldots, N-1. \tag{4}$$

The ηth B-IFDMA-modulated data vector

$$\mathbf{x}_\eta^{(k)} = \left[x_{\eta,0}^{(k)}, \ldots, x_{\eta,N-1}^{(k)}\right]^\mathrm{T} \tag{5}$$

of user k with elements $x_{\eta,n}^{(k)}$, $n = 0, \ldots, N-1$, at sampling rate N/T_s is, thus, given by

$$\mathbf{x}_\eta^{(k)} = \mathbf{F}_\mathrm{N}^\mathrm{H} \cdot \mathbf{M}_\mathrm{BI}^{(k)} \cdot \mathbf{F}_\mathrm{Q} \cdot \mathbf{d}_\eta^{(k)}. \tag{6}$$

From (6), it follows that B-IFDMA can be considered as OFDMA with block-interleaved subcarrier allocation and DFT precoding of the data symbols before OFDMA modulation. For the special case $M = 1$, that is, for one subcarrier per block in the allocated OFDM symbols, B-IFDMA is equivalent to IFDMA [14, 21]. For the special case $L = 1$, that is, for one block of subcarriers, B-IFDMA is equivalent to LFDMA [17]. Thus, B-IFDMA can be understood as a generalization of these schemes. In the appendix we show that a B-IFDMA signal can be efficiently generated in the time domain, that is, without the DFT operation.

2.2. Receiver Structure. In the following a B-IFDMA receiver is described for an uplink scenario, following the block diagram in Figure 2. Let

$$\mathbf{h}_\eta^{(k)} = \left(h_{\eta,0}^{(k)}, \ldots, h_{\eta,L_p-1}^{(k)}, 0, \ldots, 0\right)^\mathrm{T} \tag{7}$$

denote the $N \times 1$ vector representation of a multipath channel of user k. Let further $h_{\eta,l}^{(k)}$, $l = 0, \ldots, L_p - 1$, denote the L_p nonzero channel coefficients at sampling rate N/T_s with $L_p \leq N$. Before transmission over the channel $\mathbf{h}_\eta^{(k)}$, a Cyclic Prefix (CP), with length at least $\mathrm{L_p} - 1$, is inserted in between consecutive modulated data vectors $\mathbf{x}_\eta^{(k)}$. At the receiver, the CP is removed before demodulation. For the time interval T required for transmission of vector $\mathbf{x}_\eta^{(k)}$ and the CP, the channel is assumed to be time invariant. Moreover, perfect time and frequency synchronization is assumed. Thus, with $\mathbf{H}^{(k)}$ denoting the circulant channel matrix with vector $\mathbf{h}_\eta^{(k)}$ in its first column [22], the ηth received signal vector $\mathbf{r}_\eta$ after removal of the CP is given by

$$\mathbf{r}_\eta = \sum_{k=0}^{K-1} \mathbf{H}_\eta^{(k)} \cdot \mathbf{x}_\eta^{(k)} + \mathbf{n}_\eta, \tag{8}$$

where

$$\mathbf{n}_\eta = \left(n_{\eta,0}, \ldots, n_{\eta,N-1}\right)^\mathrm{T} \tag{9}$$

denotes an Additional White Gaussian Noise (AWGN) vector with samples $n_{\eta,n}$, $n = 0, \ldots, N-1$ at sampling rate N/T_s.

At the receiver, after removal of the CP, an N-point DFT is applied to the received signal $\mathbf{r}_\eta$. Subsequently, the signal is user specifically demapped. After demapping, for each user k the impact of the channel is compensated by an equalizer and the DFT precoding is compensated by a Q-point IDFT. In the following, a Frequency Domain Equalizer (FDE) [23, 24] represented by a $Q \times Q$ diagonal matrix $\mathbf{E}^{(k)}$ is considered. Thus, at the receiver, estimates $\hat{\mathbf{d}}_\eta^{(k)}$ of the data symbol vectors $\mathbf{d}_\eta^{(k)}$ for user k are given by

$$\hat{\mathbf{d}}_\eta^{(k)} = \mathbf{F}_\mathrm{Q}^\mathrm{H} \cdot \mathbf{E}^{(k)} \cdot \left(\mathbf{M}_\mathrm{BI}^{(k)}\right)^\dagger \cdot \mathbf{F}_\mathrm{N} \cdot \mathbf{r}_\eta. \tag{10}$$

3. Error Rate Performance

In this section we investigate the error rate performance of B-IFDMA with various block sizes. The aim of these investigations is to show the capability of B-IFDMA to collect large diversity gains under realistic assumptions on channel estimation performance, also for rather low data rates, without using time-diversity. We start in Section 3.1 by investigating the diversity gains under the assumption of perfect channel estimation at the receiver. Then, in Section 3.2 we quantify the channel estimation performance for various B-IFDMA block sizes. With these performance results at hand, we proceed in Section 3.3 by discussing the tradeoff between these performance measures for different B-IFDMA block sizes, and we illustrate with quantitative examples.

3.1. Diversity Gains. As discussed in Section 1 robustness to small-scale fading based on frequency diversity and/or spatial diversity is needed to satisfy delay critical services, especially in bad channel conditions. Time diversity based schemes are less attractive in order to keep a short delay over the air interface. In this section, we investigate the uplink performance of B-IFDMA with Quadrature Phase Shift Keying (QPSK) modulated and Forward Error Correction (FEC) encoded transmission over a frequency-selective fading wide area mobile radio channel. We show results for single antenna transmission (SISO), two transmit antennas at the UE using Alamouti Space-Frequency Coding [25, 26] with one receive antenna (MISO, Alamouti) and for two transmit antennas at the UE using Alamouti Space-Frequency Coding with two receive antennas at the base station (BS) applying Maximum Ratio Combining (MIMO, Alamouti and MRC). Each OFDM symbol is formed as described in Section 2 and a joint FEC encoding and interleaving is performed over the used OFDM symbols in the chunk. All simulation assumptions are listed in Table 1.

The coherence time T_c and the coherence bandwidth B_c of the mobile radio channel play an important role. In the literature various different definitions for coherence time and coherence bandwidth are used, but in Table 1 they are calculated as follows. Let c_0, f_0, and v denote the speed of light, the carrier frequency and the velocity of a mobile station, respectively. Let further $f_{D,\max} = f_0 \cdot (v/c_0)$ denote the maximum Doppler frequency for this mobile station. The coherence time T_c can be defined as

$$T_c = \frac{1}{2 \cdot f_{D,\max}} = \frac{1}{B_D}, \tag{11}$$

where $B_D = 2 \cdot f_{D,\max}$ is the well-known Doppler bandwidth.

The coherence bandwidth B_c can be defined as

$$B_c = \frac{1}{\Delta\tau}, \tag{12}$$

where $\Delta\tau$ denotes the time difference between the first and the last received propagation path of the mobile radio channel, usually denoted as the delay spread of the channel.

Table 1: Simulation parameters.

Bandwidth	40 MHz
Total number of subcarriers	1024
Carrier frequency	3.7 GHz
Sampling rate	1/(25 ns)
Guard Interval	3.2 μs
Modulation	QPSK
Code	Convolutional code, rate 1/2
Code polynomials	133,171
Constraint length	6
Decoder	BCJR [27]
Interleaving	Random over 12 OFDM symbols
Channel	WINNER C2 Urban Macro-cell [28]
Scenario	Wide Area
Antenna distance	Tx: $\lambda/2$, Rx: 2λ
User velocity	50 km/h
Coherence bandwidth	550 kHz
Coherence time	2.9 ms
Channel estimation	Perfect

The Bit Error Rate (BER) performance of B-IFDMA for different numbers M of subcarriers per block is given in Figures 3, 4, and 5. Perfect channel estimation is assumed and the pilot symbol overhead required for channel estimation is not considered. In these figures the 3 dB antenna gain in the 2 times 2 MIMO cases is removed to simplify the comparison of the diversity gains in the different scenarios.

When the distance of the subcarrier blocks is large compared to the coherence bandwidth, they receive almost independent fading, and thus the frequency diversity is improved. For large numbers Q of subcarriers per user, the distance between the subcarrier blocks is reduced and, thus, the frequency diversity gains are decreased. Regarding the simulation results for MISO and MIMO transmission it can be concluded that even for B-IFDMA exploiting spatial diversity, the differences in frequency diversity are still considerable.

From Figures 3, 4, and 5 it can also be concluded that for a given data rate, that is, for a given number Q of subcarriers assigned to a user, the performance of B-IFDMA increases with decreasing number M of subcarriers per block. The reason for that is that for a given number Q with decreasing number M, the number of subcarrier blocks L increases. However, as discussed in Section 4.2 and illustrated in Figure 1, for a given average data rate per frame the number of blocks can be maintained by introducing a TDMA component with increased number of used subcarriers and a correspondingly smaller duty cycle within the chunk. In Figure 6 we can see that the diversity gain depends mainly on the number of blocks L. Hence the same robustness towards small-scale fading can be maintained also with time-localized transmission to take advantage of the gain in transceiver power efficiency as discussed later in Section 4.2.

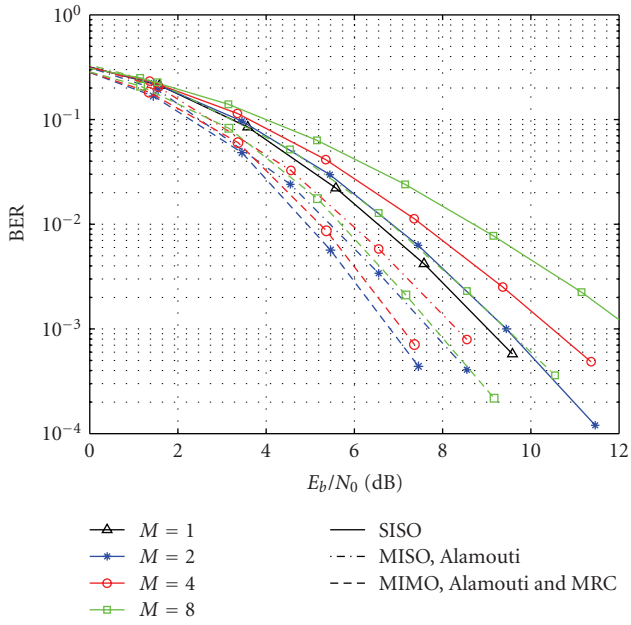


Figure 3: Coded performance for B-IFDMA with instantaneous data rate 1.11 Mbps, that is, $Q = 32$ subcarriers per user with normalized antenna gain.

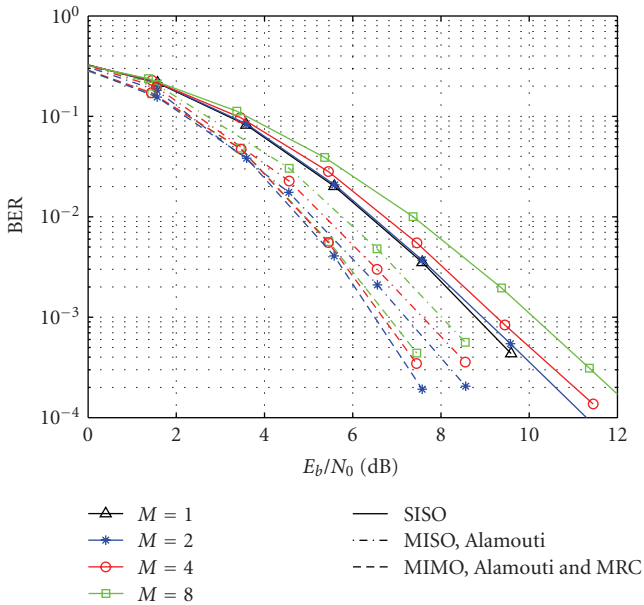


Figure 4: Coded performance for B-IFDMA with instantaneous data rate 2.22 Mbps, that is, $Q = 64$ subcarriers per user with normalized antenna gain.

3.2. Channel Estimation. In Section 3.1 we showed the simulated diversity gains for B-IFDMA with various parameterizations under the assumption of perfect channel estimation. However, in general the less correlation among the subcarriers the better diversity but also the less correlation to be used in the channel estimation scheme over the subcarrier blocks. In addition, with pilot-aided channel estimation it is important to keep the pilot overhead low. Thus, with a given pilot overhead, there is an inherent tradeoff to be made between attainable diversity gains and loss due to nonideal channel estimation performance. In this section, we first define in Section 3.2.1 what we mean by pilot overhead, and then in Section 3.2.2 we show the attainable performance of memory-based and memory-less pilot-aided channel estimation schemes for various B-IFDMA block sizes.

3.2.1. Pilot Overhead. In pilot-aided channel estimation [29–31], the complex gain of the OFDM subcarriers is estimated at the receiver based on known time-frequency pilot symbols (also denoted as reference symbols) placed within each block. The channel equalization and payload data detection/decoding is then based on inferred complex channel gains at the payload symbol locations.

With pilot aided channel estimation, there is a pilot overhead loss in both signal-to-noise ratio (SNR) degradation due to the energy put on the pilots and in spectral efficiency due to the channel symbols occupied by the pilot symbols. Below we assume that the pilot symbols are inserted as subcarrier channel symbols with the same energy as the data carrying channel symbols (i.e., no pilot boosting). In this case the SNR loss and the spectral efficiency loss are the same. Assuming that there are P pilots per block and the block size equals M subcarriers times N_t OFDM symbols, the pilot overhead loss becomes $P/(M \cdot N_t)$ and the SNR degradation $\log_{10}(M \cdot N_t/(M \cdot N_t - P))$ dB.

Below in Section 3.2.2 we discuss the suitable pilot schemes and corresponding channel estimation performance under the assumption of a constant pilot overhead loss of 1/12 for the different block sizes, that is, 8.3% loss in spectral efficiency and 0.38 dB in SNR degradation.

3.2.2. Block Size Effect on Channel Estimation. Because of the variation of the complex gain with frequency (due to the multipath propagation) and with time (due to mobility), the channel at payload positions will in general differ from that at the pilot positions. The coherence time and coherence bandwidth as defined in (11) and (12), respectively give an estimate of the order of the needed sampling interval in time and frequency for the mobile radio channel according to the sampling theorem [32]. However, the channel has to be estimated based on received noisy pilot symbols, and in a packet oriented system the channel resources needed per packet transmission are not very large. Hence, due to the limited number of noisy pilots available for channel estimation, an oversampling factor is typically needed, that is, a more dense pilot pattern means better estimation performance.

For the considered diversity-based transmission schemes, a problem is then encountered in uplinks: large blocks will have many embedded pilots and thus good possibilities for interpolation, which is more robust than extrapolation. But if the pilot overhead is to be held fixed, small blocks will contain only one or a few pilot symbols. This effect may partly or completely cancel the effect of frequency diversity.

Good channel estimation performance is achieved by mainly three different strategies.

(i) *Use pilots from adjacent blocks, to enable interpolation over frequency.* This strategy is possible and recommended in downlinks, but it cannot be used in uplinks, where adjacent blocks are either unused or used by other UEs. Blocks used by the UE itself are in general placed significant distances apart in frequency, with low inter-block channel correlation. They are therefore of limited use for channel estimation.

(ii) *Use pilots from previous blocks.* This can be done in general in downlinks. In uplinks, it becomes possible only if the UE uses the same blocks over multiple frames (persistent scheduling). In the investigation below, we illustrate the potential maximum estimation performance obtainable by using optimal Kalman smoothing that uses an unlimited amount of past payload symbols.

(iii) *Use also data symbols for channel estimation*, by iterative channel estimation. The pilot based channel estimate is then used as a first step. Decoded soft bits are then used in a second step to improve the channel estimates. Iterative channel estimation has been found to be beneficial for the IMT Advanced scenarios and pilot schemes, see [7, 33]. It improves upon pilot-based estimates by 1-2 dB in realistic cases. The almost constant offset makes it possible to roughly estimate the accuracy of iterative schemes if the accuracy of the initializing pilot-based estimate is known. We therefore focus here on pilot-based noniterative schemes.

The channel estimation performance is investigated below for two schemes:

(i) *Block Least Squares Estimation (Block-LSE)*: least squares estimation based on present but not past pilot data, also often called 2D-Wiener filtering [29, 30];

(ii) Kalman smoothing [34, 35], using present and past pilots from every second time-slot backwards in time. Blocks from odd numbered past time-slots are not used. In half-duplex FDD uplinks they would be used by other UEs. In Time Division Duplex (TDD) systems, they would be used for downlink transmissions. The time-slots (half frames) are assumed to have duration 12 OFDM symbols as in [4, 5, 36].

The block sizes used in the investigations and the related pilot positions are illustrated by Figure 7. The choice of these block sizes is related to the frame structure in the FDD mode of [4, 5]. In order to maintain a low radio access delay and to support also high speed trains, one slot (half frame) consists of only 12 OFDM symbols, [4, 5].

Here we consider uplinks, so neither method uses pilot information from subcarriers outside of the blocks. The results for the two estimation methods for the various block sizes are shown in Figure 8, for UE velocity 50 km/h at 3.7 GHz carrier frequency as well as all other parameters as in Table 1. Please refer to [37] for further details on the channel estimation methods and for additional results for other UE velocities and block sizes.

In [7] it has been shown that the effect of channel estimation errors on various decoder and detection algorithms in OFDM receivers can be well modelled by treating the estimation error as an additional white noise contribution

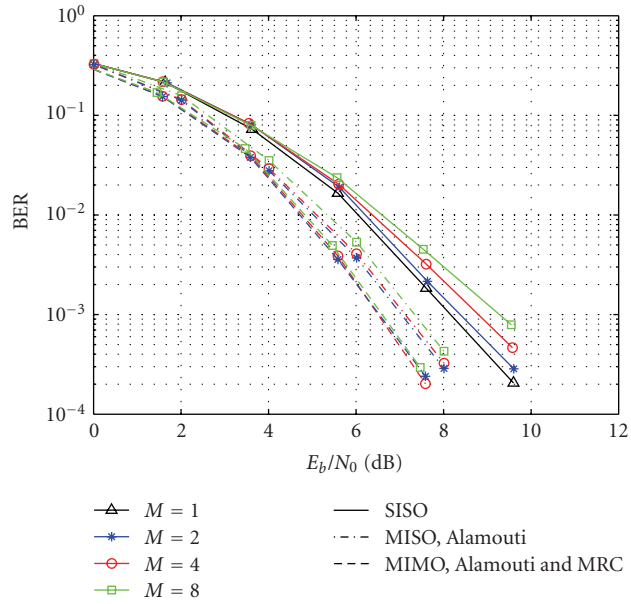


Figure 5: Coded performance for B-IFDMA with instantaneous data rate 4.44 Mbps, that is, $Q = 128$ subcarriers per user with normalized antenna gain.

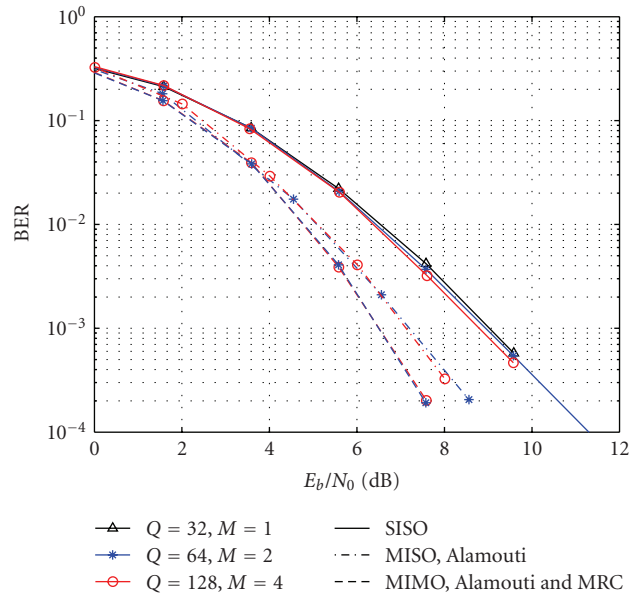


Figure 6: Coded performance for B-IFDMA with the same number of $L = 32$ blocks per user and with normalized antenna gain.

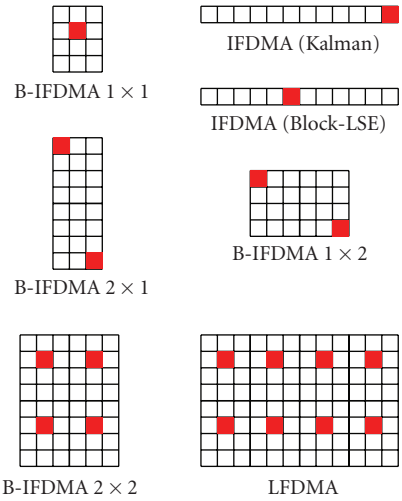


Figure 7: The pilot patterns used for the investigated block allocations that use combinations of a basic block of 4 subcarriers by-3-OFDM-symbols, with one pilot and 11 payload symbols (i.e., pilot overhead 1/12): B-IFDMA 1×1 ($M = 4$, $N_t = 3$), 1×2 ($M = 4$, $N_t = 6$), 2×1 ($M = 8$, $N_t = 3$), 2×2 ($M = 8$, $N_t = 6$), IFDMA ($M = 1$, $N_t = 12$), and LFDMA ($M = 8$, $N_t = 12$). Time axis is horizontal and frequency axis is vertical in this figure. The pilot positions within blocks have been determined by global optimization of the channel estimation performance of the Block-LSE (Wiener) method, and they differ from those specified for uplinks in [4, 5].

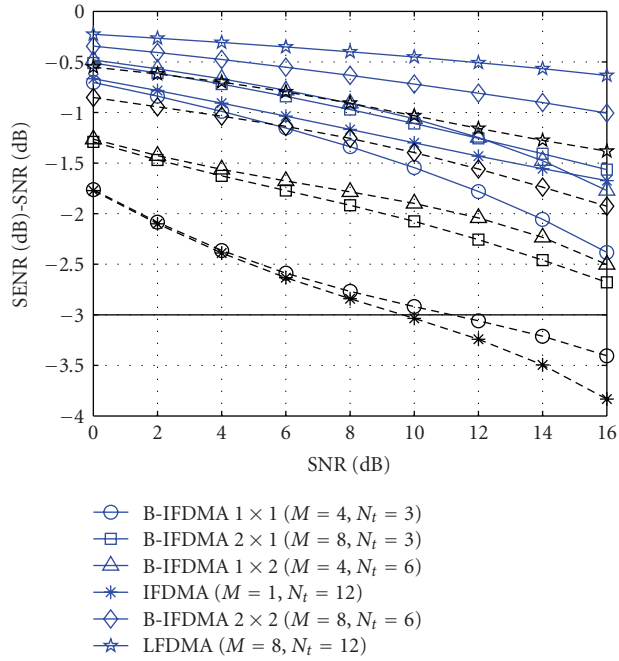


Figure 8: Performance degradation in dB due to imperfect channel estimation versus ideal SNR. The vertical axis shows the difference between actual perceived signal-to-estimation-error-plus noise ratio (SENR, in dB) and ideal SNR (in dB). The horizontal axis shows the ideal SNR, that is, assuming perfect channel state information. For example, the value -3 on the vertical axis means that a bit-error-rate curve generated in an idealized setting where perfect channel estimation is assumed should be displaced 3 dB to the right to correctly represent performance when the influence of channel estimation is taken into consideration. Solid curves represent (optimal) smoothed Kalman filter performance. Dashed curves represent Wiener filter performance, where no previous measurements are used by the estimator.

at the receiver, with a variance given by the estimation error variance. Therefore, in Figure 8 we show the channel estimation results in terms of SNR offset due to channel estimation errors at the receiver. This performance measure makes the results directly comparable to the SNR gains and losses due to different choices of number of subcarriers M per block in Figures 3, 4, and 5, as discussed further in Section 3.3.

It is evident that significant performance gains can be obtained by using Kalman smoothing which takes blocks in previous time-slots into account. Note that in the investigated case assuming half-duplex FDD, every second of the past timeslots cannot be used. The performance gain increases for slower UE velocities as shown in [37]. Full duplex FDD UEs would also benefit from the more dense slot and thus more dense pilot structure in time.

3.3. Performance Tradeoffs. By analyzing the results in Sections 3.1 and 3.2, we can quantify the tradeoff between frequency diversity gains and channel estimation performance for different B-IFDMA subcarrier block sizes. To this end, we adopt the parameters of the FDD wide area mode in the IMT Advanced capable system concept in [4, 5].

In Figure 9, we show such an example of combined diversity and channel estimation performance for the SISO case with Block-LSE channel estimation and $Q = 32$ subcarriers assigned per user. As seen, despite the better channel estimation with LFDMA, at this rather low number of Q; IFDMA and B-IFDMA are substantially better than LFDMA. The reason is the low frequency diversity obtained with the adjacent subcarriers in LFDMA. With increasing Q, B-IFDMA approaches IFDMA, and B-IFDMA becomes better than IFDMA when the diversity gains saturates in IFDMA. The reason for this is the better channel estimation performance for B-IFDMA, cf. Figure 8. In particular, making the same comparison as in Figure 9 but with $Q = 64$ subcarriers, B-IFDMA is better than IFDMA for both $M = 4$ and $M = 8$. At BER 10^{-3}, B-IFDMA with $M = 4$ is 0.5 dB better and B-IFDMA with $M = 8$ is 0.2 dB better than IFDMA. Note also that due to the block length $N_t = 6$ used in B-IFDMA, this performance is achieved with an average data rate over the chunk that is half compared to IFDMA and LFDMA, which is useful for transmission of small packets.

Below we exemplify the diversity versus channel estimation tradeoff for B-IFDMA, assuming different block lengths N_t for both Block-LSE and Kalman channel estimation. Since the pilot overhead is the same for all considered schemes, this loss is not included.

Example 1. Referring to Table 2, under the assumption that $Q = 32$ subcarriers are assigned to a user, we can see in

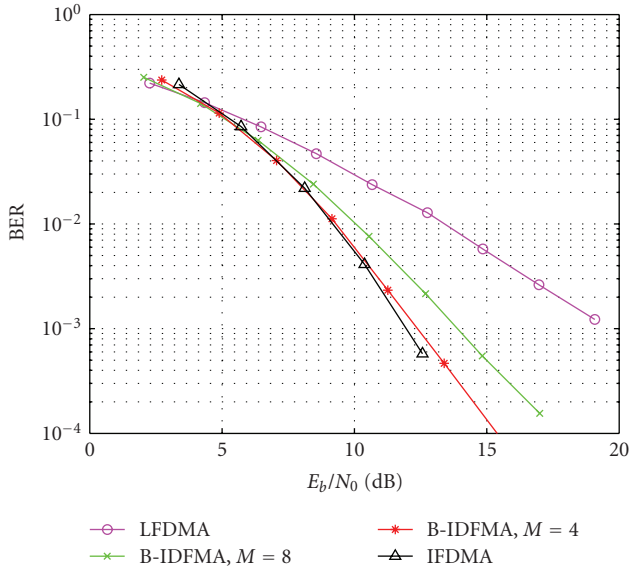


Figure 9: Coded SISO performance for B-IFDMA, IFDMA and LFDMA with nonperfect channel estimation, and $Q = 32$ subcarriers assigned per user. The Block-LSE channel estimation performance results from Figure 8 are used. B-IFDMA uses block of sizes ($M = 4$, $N_t = 6$) or ($M = 8$, $N_t = 6$).

Figure 3 that at BER 10^{-3} in the SISO case when going from $M = 8$ to $M = 4$ subcarriers per block, that is, changing from number of subcarrier blocks $L = 4$ to $L = 8$, there is a diversity gain of 1.9 dB, that is, a reduction in required SNR from around 12.4 to 10.5 dB. This gain should be compared to the loss in channel estimation performance in Figure 8 due to the fewer number of subcarriers per block. With block length $N_t = 3$, the channel estimation loss at the intermediate SNR 11 dB is -1.2 dB for $M = 8$ and -1.7 dB for $M = 4$ subcarriers per block with Kalman filtering. That is, there is an overall gain of $1.9 - 0.5 = 1.4$ dB including channel estimation for using $M = 4$ subcarriers compared to $M = 8$. With Block-LSE, the corresponding overall gain is $1.9 - 0.8 = 1.1$ dB. With the longer blocks having $N_t = 6$ (double mean data rate over the slot for a given number of blocks L) the corresponding gains when going from $M = 8$ to $M = 4$ are $1.9 - 0.4 = 1.5$ dB (Kalman) and $1.9 - 0.5 = 1.4$ dB (Block-LSE).

Example 2. In Table 2, we also show the corresponding case with $Q = 64$ subcarriers per user based on the results in Figures 4 and 8. Here the two cases with $M = 8$ and $M = 4$ subcarriers per block perform very similar, that is, the diversity gain with $L = 16$ blocks compared to $L = 8$ blocks is almost completely lost due to the worse channel estimation performance.

Similar tradeoff comparisons can be made for the MISO with Alamouti case and the MIMO with Alamouti and MRC case based on the diversity results in Figure 6 and the channel estimation performance results in Figure 8, since the results on channel estimation performance in Figure 8 are directly applicable to uplinks with multiple UE antennas. Pilots are then placed at different time-frequency positions for different antennas, and these positions are not used by payload data at the other antennas to limit interference. Therefore, the pilot overhead increases, but the channel estimation accuracy stays unchanged. Due to the additional spatial diversity gains, fewer blocks L are typically needed, down to $L = 2$ to 4.

4. Energy Efficiency

In this section we aim to quantify the end-to-end energy efficiency of B-IFDMA. The aim of these investigations is to motivate the use of a DFT precoding step, and the advantage of the TDMA component within the B-IFDMA scheme. To this end, we start in Section 4.1 by characterizing the envelope properties of B-IFDMA in terms of popular envelope variation metrics. These metrics are commonly used in the literature to characterize the signal envelope variations and to give an indication of the efficiency of a generic High Power Amplifier (HPA). These results also motivate the regular subcarrier allocation in B-IFDMA. In order to give a quantitative measure of the energy efficiency with a representative HPA, we continue in Section 4.2 by showing the HPA efficiency with different B-IFDMA parameterizations and different HPA operation modes for a real HPA. These investigations enable us to quantify the energy efficiency gains of DFT precoded schemes compared to OFDMA. In addition, they allow us to characterize the gains with time-localized transmission, and to quantify the end-to-end energy efficiency with various B-IFDMA parameterizations.

4.1. Envelope Properties. It is well known that for increasing envelope fluctuations of the transmit signal, the cost of a typical commercial HPA in the UE increases and the power efficiency decreases. Thus, especially in the uplink, the provision of low envelope fluctuations is important for the transmitted signal. In this section we investigate the envelope properties of B-IFDMA, and we predict the efficiency of the HPA based on an amplifier model. For that purpose, a signal model including oversampling, pulse shaping and windowing is assumed, all according to [38]. The oversampling factor is $S = 8$ and the pulse shaping filter is chosen such that an OFDM-like rectangular spectrum of the B-IFDMA signal is provided. Furthermore, a Raised-Cosine window with a roll-off region that is 5% of the symbol duration is applied.

In Figure 10, the envelope of the B-IFDMA transmit signal is investigated in terms of the well-known Peak-to-Average Power Ratio (PAPR) [39] for $N = 1024$ subcarriers in the system and $Q = 64$ subcarriers assigned to a user using QPSK modulation. As references, the PAPR of two signals are given that differ from the B-IFDMA in the following properties. The first signal does not use DFT precoding and the second signal uses a random allocation of the subcarrier blocks instead of a regular one. From Figure 10 it can be clearly seen that *both DFT precoding* and *regular allocation* of the subcarrier blocks is required in order to provide a low PAPR. B-IFDMA provides a mean PAPR that is

Table 2: Overall performance comparison of SISO B-IFDMA with $Q = 32$ or $Q = 64$ subcarriers per user and $M = 4$ or $M = 8$ subcarriers per block with $N = 1024$ subcarriers in the system.

Gain	$N_t = 3$		$N_t = 6$	
(dB)	Kalman	Block-LSE	Kalman	Block-LSE
B-IFDMA $Q = 32$, $M = 4$ versus $M = 8$ ($L = 8$ versus $L = 4$)				
Diversity, BER 10^{-3}	1.9	1.9	1.9	1.9
Channel est. 11 dB	−0.5	−0.8	−0.4	−0.5
Total	1.4	1.1	1.5	1.4
B-IFDMA $Q = 64$, $M = 4$ versus $M = 8$ ($L = 16$ versus $L = 8$)				
Diversity, BER 10^{-3}	0.7	0.7	0.7	0.7
Channel est. 10 dB	−0.5	−0.8	−0.4	−0.5
Total	0.2	−0.1	0.3	0.2

1.2–1.5 dB lower than the mean PAPR of the corresponding scheme without DFT precoding. Compared to a scheme with random allocation of the subcarrier blocks with DFT precoding, the PAPR gain of B-IFDMA is greater than 3 dB for a number $L = 64$ subcarrier blocks, that is, for the special case of IFDMA. The gain decreases to ≈0.7 dB for $L = 4$ subcarrier blocks. For $L = 2$, the regular and the random allocation of the subcarrier blocks are equivalent except for the distance of the subcarrier blocks and, thus, the mean PAPR is similar.

Figure 11 analyzes the envelope of the B-IFDMA transmit signal based on different metrics. In addition to the PAPR, the well-known Raw Cubic Metric (RCM) as defined in [40, equation (15)], which is related to the 3GPP Cubic Metric (CM) in [41], is regarded. The motivation for the CM and RCM are the fact that the primary cause of distortion is the third order nonlinearity of the amplifier gain characteristic. Moreover, the HPA power efficiency is predicted. For that purpose, a nonlinear amplifier is assumed that produces increased out-of-band radiation due to nonlinear distortions dependent on the envelope of the input signal. The power efficiency of the given HPA depends on the power back-off (BO) that is required to meet a given spectral mask for the transmit signal. Thus, for investigation of the impact of the envelope fluctuations on the power efficiency, also the required BO is analyzed. In the following, for the HPA, the well-known Rapp model [39] with Rapp-parameter $p = 2$ is used which represents the model of a power amplifier with high nonlinearities. The spectrum requirement mask is representative for IMT Advanced systems, and is given in [38].

The results for the different metrics are summarized in Figure 11. Again, $N = 1024$ subcarriers is assumed in the system, with $Q = 64$ subcarriers per user and QPSK modulation. A scheme without DFT precoding is regarded as a reference. It can be concluded that, regardless of the number L of subcarrier blocks, for B-IFDMA, the envelope fluctuations are significantly lower compared to the scheme without DFT precoding. The mean PAPR and the RCM have a minimum for $L = Q$ and $L = 1$, that is, for LFDMA and for IFDMA, where B-IFDMA can be interpreted as a single-carrier scheme and have a maximum for $L = 8$. However,

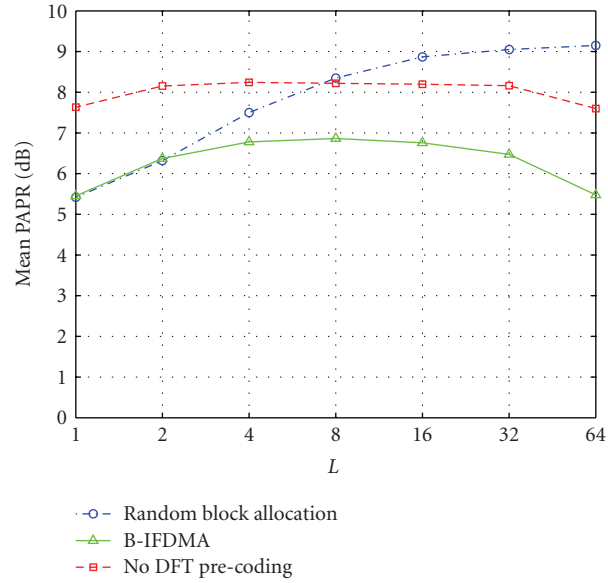


Figure 10: Mean PAPR of B-IFDMA transmit signals with $Q = 64$ as a function of number of blocks L compared to the corresponding schemes without DFT precoding and schemes with random allocation of the subcarrier blocks.

even at the maximum, the envelope fluctuations of B-IFDMA are considerably lower than for a corresponding scheme without DFT precoding. In difference to the mean PAPR and the RCM, the required BO increases with decreasing number L of subcarrier blocks. The reason for that is that in addition to the envelope of the signal also the shape of the spectrum changes and the side-lobes are increased. However, for the special case of $L = 1$, that is, for LFDMA, the side-lobes are significantly reduced. Thus, in this case, the spectral mask is less relevant, and results for $L = 1$ are omitted.

From Figure 11 it can be concluded that the effects shown in Figure 10 can be considered to be almost independent of the metric that is used. Thus, B-IFDMA can be considered to provide a higher power efficiency and lower envelope fluctuations compared to schemes without DFT precoding and without regular subcarrier allocation, respectively.

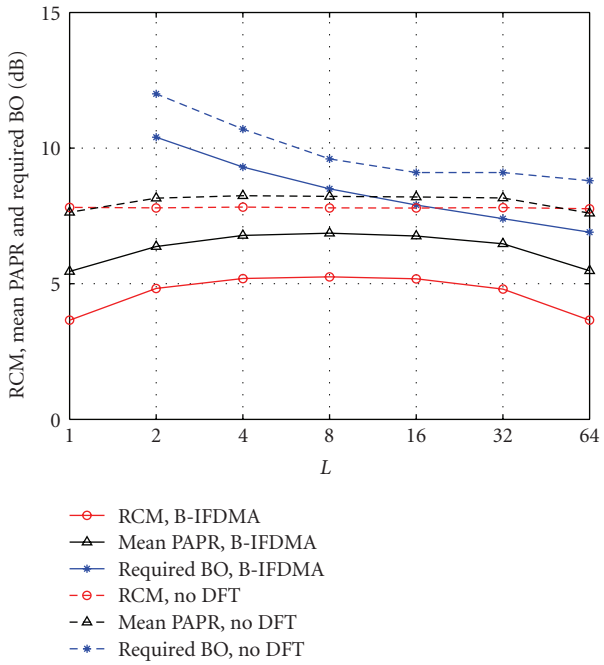


Figure 11: Results for the analysis of the envelope fluctuations of B-IFDMA transmit signals with $Q = 64$ (with DFT) compared to signals without DFT precoding for different numbers L of subcarriers per block using different metrics.

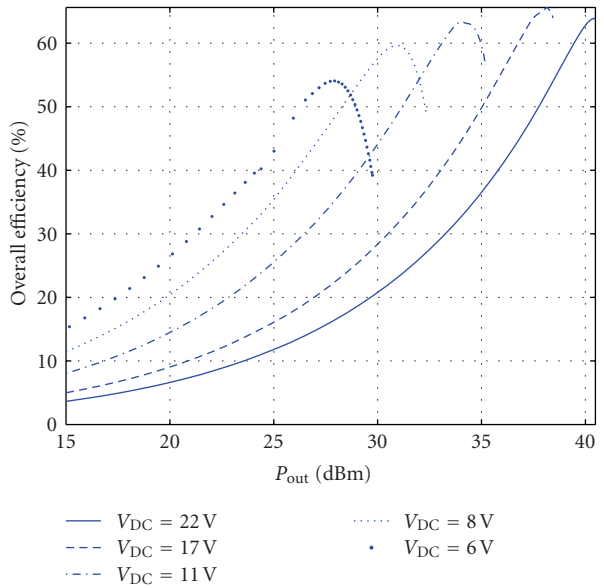


Figure 12: Illustration of overall efficiency as a function of output power for different amplifier drive voltages V_{DC} using the HPA in [42], where $V_{DC} = 22$ V is the highest possible drive voltage.

4.2. Gain with Time Localized Transmission. In Section 4.1, we characterized the envelope properties of B-IFDMA according to different metrics. In this section we make an analysis of the energy efficiency of B-IFDMA with a real amplifier. The aim of this investigation is to correlate the HPA energy efficiency with the prediction by the metrics in Section 4.1. The aim is also to show and quantify the gain by optimizing the operation point of the power amplifier, in order to motivate the benefit of the TDMA component within the B-IFDMA scheme.

4.2.1. HPA Efficiency. The efficiency of an HPA is best described by the *overall efficiency*, defined as

$$\eta_A = \frac{P_{Out}}{P_{DC} + P_{In}}, \tag{13}$$

where P_{In} is the power at the input of the HPA, and P_{Out} is the resulting output power. P_{DC} is the power at the DC input of the amplifier, computed as the product between the DC voltage and the DC current, $P_{DC} = V_{DC} \cdot I_{DC}$. For a given V_{DC}, the efficiency is a function mainly of the desired output power; the general trend is that the efficiency is higher for high output powers. However, by varying the drive voltage V_{DC}, the efficiency curve of the amplifier can be changed. In Figure 12, we illustrate the overall efficiency as a function of output power for different drive voltages V_{DC}, when the input signal is an *unmodulated carrier* signal, using the HPA in [42].

Figure 12 shows that the efficiency of the amplifier is highest when its output power is close to the maximum attainable output power, that is, when it is driven close to saturation. However, due to the signal dynamics and other system considerations such as power control, it is in general not possible to drive the amplifier in its most efficient mode at all times. In situations when we need a lower average output power, we can see that by lowering the (constant) drive voltage we can get an improved efficiency, but the overall efficiency is still lower than when its output power is close to the maximum attainable output power. (The optimal way of driving the amplifier would be to jointly vary the drive and Radio Frequency (RF) power for maximum efficiency [42], but this is not generally regarded as practical to implement today.)

We have evaluated the overall efficiency of several different amplifiers, when driven at different constant V_{DC} and with different *modulated* signals. The constant V_{DC} drive voltage has been chosen for maximum overall efficiency, and the following modulated input signals have been used: OFDM, IFDMA, LFDMA and B-IFDMA with different numbers of blocks (L) and pulse shaped as in Section 4.1.

To compute the overall efficiency of the HPA, we use the measured characteristics of the HPA in terms of required P_{DC} for a given P_{In} and desired P_{Out} of each signal sample, and then we perform a weighted averaging over the consumed and transmitted powers using the desired output power histograms for the modulated signal. Thus, ideal predistortion of the signals is assumed, and the operation point is chosen such that maximum 1% of the samples are above the saturation point, which is generally regarded as an acceptable level of signal distortion.The results are shown in Table 3.

As can be seen, we have the following.

(i) Due to the different amplitude distributions of these signals, they lead to different power efficiencies.

(ii) In accordance with the envelope metric results in Section 4.1, and compared to the TDMA-OFDM system, the various DFT-precoding based schemes perform better both with respect to the overall efficiency and with respect to the maximum output power (not shown in Table 3). This can potentially be used for increasing the cell size and/or larger data rates at a given path loss, provided regulations on maximum transmit power are not violated. The better HPA efficiency also implies less heat dissipation in the UE, which simplifies the design and can cut other supporting component costs.

(iii) The difference in efficiency of the various DFT precoded schemes is very small, including the B-IFDMA scheme. In particular, these differences are smaller than predicted by the envelope metric results in Section 4.1.

The results in Table 3 were obtained using the class E Laterally Diffused Metal Oxide Semiconductor (LDMOS) amplifier in [42]. However, to verify the qualitative conclusions we have also repeated the experiments with a class D LDMOS and a class E Gallium Nitride (GaN) amplifier. We have also studied other designs in the literature, for example, [43], and other classes of operation, such as class A, AB, or B. The overall conclusion is that the qualitative results are the same as above regardless of the amplifier.

4.2.2. Efficient HPA Operation. From the results in Table 3 we see that there is a gain to be made if the power amplifier as often as possible can operate close to its optimal operation point. However, in order to limit the Multiple Access Interference (MAI) from different users in a multicarrier based uplink due to imperfections in transmitter hardware, synchronization and Doppler spread, it is important to have some kind of power control to limit the difference in received power spectral density from different users. Thus, if all users were allocated the same number of subcarriers, with a constraint on maximum received power spectral density, there would be situations when the HPA has to operate at a low and suboptimal transmit power level. In these scenarios HPA efficiency would benefit from an increase in the number of allocated subcarriers in a given OFDM symbol, because then the UE could transmit during a shorter time, that is, on a lower number of OFDM symbols, for a given average data rate. One possibility to do this would be to decrease the subcarrier separation in an IFDMA scheme, but that would imply a larger channel estimation overhead due to the low correlation among the subcarriers as shown in Section 3.2. With a short frame duration, aiming at low delays, this overhead would be prohibitive. In addition, it would limit the possibility for coexistence with adaptive TDMA/OFDMA as discussed in [9].

In order to quantify the HPA efficiency with and without time localized transmission, we assume that the choices are to either (a) transmit at full power 25% of the time, and turn off the transmitter for 75% of the time, or (b) to transmit at 25% of full power all the time. Thus, in scenario (b) the amplifier is backed-off 6 dB compared to scenario (a). In a BS, the amplifier is usually optimized for a high output power, while in the UEs the amplifier works at a low power level most of the time, for example, 21–46 dBm for BSs and 21–24 dBm for UEs depending on the deployment scenario ranging from local area to wide area, [19].

As seen in Table 3 maximum overall efficiency is obtained when operating the HPA close to the maximum output power level. Thus scenario (a) leads to higher efficiency in all cases. For example, assume the same number of well separated blocks L. If the options are to use IFDMA in the full duration of a chunk with 32 subcarriers at a constant output power level of 18 dBm, we get an overall HPA efficiency of 29% (Table 3 row 2, column 5), whereas if we use B-IFDMA with $Q = 128$ subcarriers and $M = 4$ subcarriers per block (i.e., $L = 32$ blocks) one quarter of the frame duration and an instantaneous output power level of 24 dBm, we get an overall HPA efficiency of 41% (Table 3 row 5, column 4). The corresponding difference is smaller when the average transmit power is closer to the maximum efficiency. For example, at 24 dBm average transmit power the corresponding overall efficiencies are 43% (Table 3 row 2, column 4) for IFDMA and 45% (Table 3 row 5, column 3) for B-IFDMA. Thus, there is a large benefit to introduce a TDMA component in the B-IFDMA scheme in order to allow a shorter transmit duration than a full frame with a larger instantaneous data rate, except in case the required overall data rate is already close to the maximum supported. With time-localized transmission and reception, we also introduce the additional possibility of micro-sleep mode within scheduled frames. The feasibility and potential of micro-sleep mode is discussed in [12].

4.3. End-to-End Energy Efficiency. By combining the results in Sections 4.1 and 4.2 with the results in Section 3, we can quantify the end-to-end energy efficiency for different B-IFDMA parameterizations. Below we illustrate this tradeoff, by building on the examples in Section 3.3, assuming a target BER of 10^{-3}.

Example 3 (revisited). Similar to the mean PAPR results for $Q = 64$ subcarriers in Figure 10, the mean PAPR for B-IFDMA with $Q = 32$ subcarriers is very similar for $M = 4$ ($L = 8$) and $M = 8$ ($L = 4$). In addition, not shown in this paper, the mean PAPR values of B-IFDMA have been found to correlate well with the overall HPA efficiency values. Thus, with SISO using $Q = 32$ subcarriers, the scheme with $L = 8$ blocks with $M = 4$ subcarriers each seems to provide the best tradeoff also considering end-to-end energy efficiency.

Example 4 (revisited). The HPA efficiency for $Q = 64$ predicted by the mean PAPR as shown in Figure 10 is very similar also for $L = 16$ blocks compared to $L = 8$. Thus, also with respect to end-to-end energy efficiency the two cases with $M = 4$ and $M = 8$ subcarriers per block seem to perform very similar. If instead the HPA efficiency is predicted

Table 3: Overall efficiency η_A in % of the HPA in [42] with constant drive voltage operation with V_{DC} chosen for maximum overall efficiency for different input signals, all using QPSK symbol constellations and a system with $N = 1024$ subcarriers.

Const V_{DC}	Max efficiency	Max efficiency −6 dB	30 dBm	24 dBm	18 dBm
TDMA-	40% @	34% @	39% @	39% @	29% @
OFDM	26 dBm	20 dBm	30 dBm	24 dBm	18 dBm
IFDMA	49% @	43% @	49% @	43% @	29% @
$Q = 32$	30 dBm	24 dBm	30 dBm	24 dBm	18 dBm
B-IFDMA	47% @	38% @	45% @	41% @	29% @
$Q, M = 32, 4$	28 dBm	22 dBm	30 dBm	24 dBm	18 dBm
B-IFDMA	46% @	38% @	45% @	41% @	29% @
$Q, M = 64, 4$	28 dBm	22 dBm	30 dBm	24 dBm	18 dBm
B-IFDMA	46% @	38% @	45% @	41% @	29% @
$Q, M = 128, 4$	28 dBm	22 dBm	30 dBm	24 dBm	18 dBm
LFDMA	48% @	38% @	47% @	42% @	29% @
$Q = 32$	28 dBm	22 dBm	30 dBm	24 dBm	18 dBm

by the required power backoff to satisfy a spectrum mask, the results for required BO in Figure 11 apply. In this case, the end-to-end energy efficiency seems to be around 0.5 dB better with $L = 16$, that is, for $M = 4$ subcarriers per block.

In the next example, we now make a comparison between B-IFDMA and IFDMA with the same data rate, also taking the end-to-end energy efficiency into account.

Example 5. Let us compare the two options to use IFDMA in SISO at the *same data rate*, for example, B-IFDMA using $M = 1$, $L = 32$, and $N_t = 12$ having $Q = 32$ subcarriers for the user with B-IFDMA using $M = 4$, $L = 32$, and $N_t = 3$ having $Q = 128$ subcarriers. The data rate is the same since in both cases $M \cdot N_t \cdot L = 384$ symbols are transmitted per slot. To generate the same RF energy, IFDMA would operate with 6 dB less transmit power during 4 times longer duty cycle. Thus, this scenario is especially relevant for the case when the required uplink data rate is below the maximum achievable for the UE at the given channel conditions. Consider the diversity gains in Figure 6, the channel estimation performance in Figure 8 and the HPA efficiency for this case in Table 3. Using the corresponding HPA efficiency values as discussed in Section 4.2.2, the end-to-end energy efficiency comparison is shown in Table 4 for target BER 10^{-3}. As seen in Table 4, there is an overall gain for B-IFDMA with $M = 4$ over the IFDMA case. The gain is more than 2 dB at low output power levels, but there is a substantial gain also at operation closer to the maximum output power level. This gain is achieved without considering additional potential sleep mode gains enabled by the short blocks, as mentioned in Section 4.2.2.

Table 4: End-to-end energy efficiency comparison of SISO B-IFDMA using block size $M = 4$, $N_t = 3$ and $Q = 128$ subcarriers per user at two different HPA output power levels versus B-IFDMA using block size $M = 1$, $N_t = 12$ and $Q = 32$ subcarriers per user at −6 dBm lower HPA output power level. IFDMA uses 4 times longer blocks. $L = 32$ in both cases and there are $N = 1024$ subcarriers in the system.

Gain	30 dBm		24 dBm	
(dB)	Kalman	Block-LSE	Kalman	Block-LSE
B-IFDMA $M = 4$, $N_t = 3$, $Q = 128$ versus $M = 1$, $N_t = 12$, $Q = 32$				
Diversity, BER 10^{-3}	0.24	0.24	0.24	0.24
Channel est. 9 dB	−0.15	0.1	−0.15	0.1
HPA efficiency	0.2	0.2	2.07	2.07
Total	0.29	0.54	2.16	2.41

5. Robustness

In this section, the robustness of B-IFDMA to carrier frequency offsets (CFOs) and to Doppler spreads (DSs) is analyzed for the uplink dependent on the signal parameters. The aim of this investigation is to show that the block based subcarrier allocation in B-IFDMA enables the possibility to combine robustness and provision of frequency diversity.

In mobile radio applications, CFOs are typically caused by oscillator imperfections due to low cost hardware components or Doppler shifts due to the mobility of the users. The CFOs result in a shift of the spectra of the different users' signals. Hence, the orthogonality of the subcarriers is destroyed and inter-carrier interference (ICI) occurs.

In general, two types of ICI can be distinguished. Regarding a particular user's signal, on the one hand, due to the shift of the spectrum, interference between the subcarriers of this user occurs. In the following this is denoted as self-interference (SI). On the other hand, in addition interference between the subcarriers of different users occurs. This case is in the following denoted as multiple access interference (MAI).

The DS is caused by the fact that in a mobile radio channel typically the same signal is received from different propagation paths where each path suffers from a different Doppler shift. The superposition of differently shifted replicas of the same signal at the receiver leads to a spread of the subcarriers of the different users' signals. Consequently, also for DSs the orthogonality of the subcarriers is destroyed and

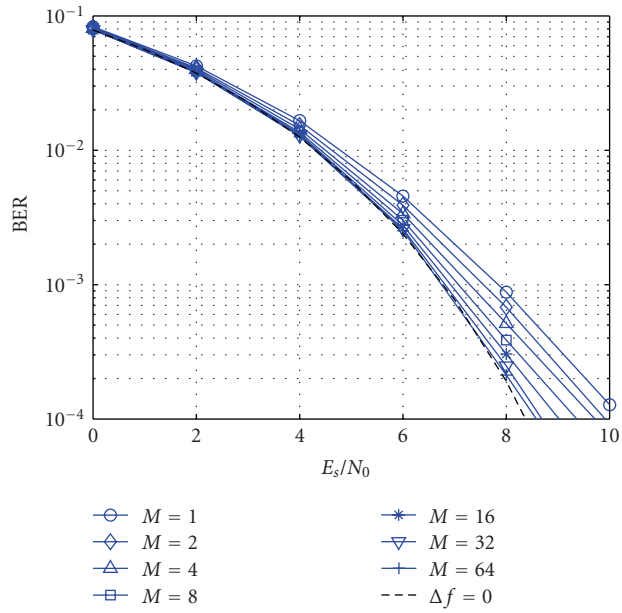


FIGURE 13: Performance for $\Delta\overline{f}_{\mathrm{CFO}}^{(k)} = 10\%$ maximum relative CFO for different numbers M of subcarriers per block, assuming $N = 1024$ subcarriers and $Q = 64$ subcarriers per user.

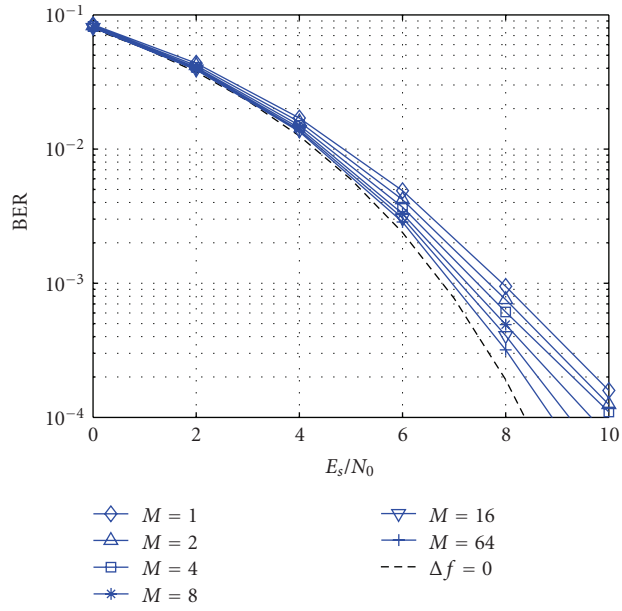


FIGURE 14: Performance for Doppler Spread with $\Delta\overline{f}_{\mathrm{CFO}}^{(k)} = 15\%$ relative carrier frequency offset per path for different numbers M of subcarriers per block, assuming $N = 1024$ subcarriers and $Q = 64$ subcarriers per user.

ICI occurs. Similar to the effects of CFOs, also for DSs two types of ICI, namely SI and MAI can be distinguished.

For uplink transmission, the CFOs and the DS for the received signals of different users are different. Thus, if CFOs and DS are known at the receiver, compensation of SI is possible, whereas compensation of MAI can only be obtained by application of joint detection techniques that require a high computational effort.

For the analysis of the robustness of B-IFDMA to CFOs, let

$$\Delta\overline{f}_{\mathrm{CFO}}^{(k)} = \frac{\Delta f_{\mathrm{CFO}}^{(k)}}{\Delta f} \tag{14}$$

denote the relative CFO of user k normalizing the CFO $\Delta f_{\mathrm{CFO}}^{(k)}$ of user k to the subcarrier bandwidth Δf. The relative CFO $\overline{f}_{\mathrm{CFO}}^{(k)}$ is modeled as a random variable that is uniformly distributed in $[-\Delta\overline{f}_{\mathrm{CFO,max}}^{(k)}, \Delta\overline{f}_{\mathrm{CFO,max}}^{(k)}]$ with $\Delta\overline{f}_{\mathrm{CFO,max}}^{(k)}$ denoting the maximum relative CFO of user k that occurs. The CFOs $f_{\mathrm{CFO}}^{(k)}$ are assumed to be known at the receiver. Thus, SI can be perfectly compensated by reversing the CFO $f_{\mathrm{CFO}}^{(k)}$. For the compensation of the MAI, joint detection techniques are required. For this investigation, it is assumed that, due to the high complexity, the compensation of MAI at the receiver is not feasible. In order to analyze the robustness to CFOs independently of the diversity effects, a B-IFDMA transmission over an AWGN channel is regarded.

For the analysis of the robustness of B-IFDMA to DSs, let

$$\Delta\overline{f}_{\mathrm{DS}}^{(k)} = \frac{\Delta f_{\mathrm{DS}}^{(k)}}{\Delta f} \tag{15}$$

denote the relative DS of user k normalizing the DS $\Delta f_{\mathrm{DS}}^{(k)}$ of user k to the subcarrier bandwidth Δf. Similar to the modelling of the CFOs, also the DS $\Delta\overline{f}_{\mathrm{DS}}^{(k)}$ is modelled as a random variable. Assuming that for the different propagation paths the angle of arrival is uniformly distributed in $[0, 2\pi]$, the relative Doppler shift is Jakes distributed in $[-\Delta\overline{f}_{\mathrm{D,max}}^{(k)}, \Delta\overline{f}_{\mathrm{D,max}}^{(k)}]$, where $\Delta\overline{f}_{\mathrm{D,max}}^{(k)}$ denotes the maximum Doppler shift normalized to the subcarrier bandwidth Δf. At the receiver, SI is combatted by application of a linear Minimum Mean Square Error (MMSE) receiver, cf. [44]. For the compensation of the MAI, again, joint detection techniques are required and it is assumed that, due to the high complexity, the compensation of MAI at the receiver is not feasible. In order to separate the effect of the Doppler spread from frequency selective fading effects that are also caused by multipath propagation, a Doppler spread channel according to [44] is regarded, that is, the different propagation paths are assumed to arrive at the receiver at the same time. The channel is distorted by AWGN and the received signals from the different propagation paths suffer from mutually independent relative Doppler shifts.

Figure 13 depicts the performance results without coding for the robustness investigations to CFOs assuming $N = 1024$ subcarriers in the system, $Q = 64$ subcarriers per user, $K = 16$ users and $\Delta\overline{f}_{\mathrm{CFO}}^{(k)} = 10\%$ for all users. From Figure 13 it can be concluded that, for B-IFDMA, the robustness to CFOs increases with an increasing number M of subcarriers per block. The reason for that is that the strongest intercarrier interference is caused by neighboring subcarriers. Thus, increasing the number of neighboring subcarriers belonging to the same user increases the robustness to MAI at the expense of additional SI that, however, can be compensated. Note, that already for low numbers M of

subcarriers per block the robustness of B-IFDMA to CFOs is significantly improved.

Figure 14 depicts the performance results for the robustness investigations to DS assuming the same parameters as in Figure 13 and $\Delta\overline{f}^{(k)}_{\mathrm{D,max}} = 15\%$ for all users. The value for $\Delta\overline{f}^{(k)}_{\mathrm{D,max}}$ represents the maximum relative Doppler shift for a system with $\Delta f = 10\,\mathrm{kHz}$ and a carrier frequency of 5 GHz with a user velocity of 315 km/h and, thus, represents a high mobility scenario, for example, for high speed trains.

From the results in Figure 14 it can be concluded that also the robustness of B-IFDMA to DS increases with an increasing number M of subcarriers per block. The reason is the same as for the improved robustness to CFOs. Again, already for small numbers M a significant robustness gain is provided. The increased robustness to CFOs and DSs of B-IFDMA with $M > 1$ makes B-IFDMA suitable for high speed users and systems with limited frequency synchronization.

6. Discussion of Results

In this section we summarize and comment on the investigations made in Sections 3, 4, and 5. In Section 1, we identified the following important requirements for the diversity based multiple access scheme, which should complement the FA multiple access scheme in an IMT Advanced capable system. Below we comment on the suitability of B-IFDMA to meet these requirements.

(i) *Robustness to small-scale fading without time diversity*. The results in Section 3 showed that B-IFDMA can be defined based on a rather few number of subcarriers per block also with realistic channel estimation performance. Thus, also at rather low data rates, that is, rather few subcarriers assigned to a user per slot, a large frequency diversity can be obtained.

(ii) *Tuneable degree of frequency-diversity*. As shown in Sections 3.3 and 4.3 rather small blocks, also with a sub-slot duration, can provide a good error rate performance. Thus, for a given data rate additional blocks can be allocated either well-separated in frequency to provide additional frequency-diversity, or adjacent in time or frequency (i.e., in same chunk cf. Figure 1), if enough diversity is already obtained from the frequency and/or spatial domain.

(iii) *Need to support high energy efficiency in the transmitters and the receivers*. We showed in Section 4 that B-IFDMA including pulse shaping can provide similar HPA efficiency as IFDMA and LFDMA, which is substantially better than for TDMA-OFDM without DFT precoding. In addition, whenever the UE is not power limited, in Section 4.3 a substantial overall gain of more than 2 dB is shown with time localized transmission in a fraction of the time slot, also when including realistic channel estimation performance and disregarding potential sleep mode gains. (These results are valid also for a downlink scenario using the B-EFDMA scheme.)

(iv) *Robustness to carrier frequency offsets and large Doppler spread*. This property was evaluated in Section 5 and the conclusion is that already for small numbers of subcarriers per block a significant robustness gain against carrier CFOs as well as against Doppler spreads is provided compared to IFDMA. This property could provide a significant gain in certain scenarios, like for deployment in frequency bands using a several GHz carrier frequency. Another scenario is to support high-speed trains, and/or to deploy a system with rather narrow subcarrier bandwidth.

(v) *Support for widely varying packet sizes*. Due to the good error rate performance of B-IFMDA with few number of subcarriers, transmission of rather small blocks perform well by using the TDMA component. Large packets can use the full chunk duration. The benefit of configuration flexibility motivates the introduction of a small basic block consisting of, for example, $M = 4 \times N_{\mathrm{t}} = 3$ (subcarriers $\times$ OFDM symbols) in B-IFDMA as a building block to enable adaptive block allocation in different deployment and usage scenarios, as illustrated in Figure 1, see [5, 45] for further discussion.

(vi) *Enable efficient resource allocation*. As discussed in [45] just a few different block allocations should be sufficient in a cell. The regular block allocation is beneficial for low addressing overhead, and it was shown in Section 4.1 to also be beneficial for lowering the envelope variations.

(vii) *Be of use for in-band control signals*. This is possible due to the efficient support for small packets as discussed above. In addition, the short TDMA component in B-IFDMA is useful to support precise timing of control messages for FA transmission, cf. [9].

(viii) *Enable efficient coexistence with adaptive TDMA/OFDMA*. Since B-IFDMA also is based on OFDM with the same parameters, these two schemes are compatible. With well-separated regular block allocations in frequency, adaptive TDMA/OFDMA resources can be interlaced as shown, for example, in [9, Figure 2].

(ix) *Facilitate low complexity transmitter in UEs*. The good envelope properties of B-IFDMA enables the use of a less complex HPA and predistortion unit. In the appendix we show that a B-IFDMA signal can be efficiently generated in the time domain, that is, without the DFT operation.

7. Conclusions

In this paper we have shown that B-IFDMA, which is a generalization of the SC-FDMA concept, is a power efficient, flexible and scalable multiple access scheme that can serve as a robust complement to future adaptive IMT Advanced capable wireless systems. Based on the included

investigations, we have shown that B-IFDMA is able to provide equal or better error rate performance than IFDMA and LFDMA, when considering realistic channel estimation performance at the receiver and no reliable channel state information at the transmitter, also for rather low data rates, and without using time diversity.

We also showed that B-IFDMA provides better amplifier efficiency than OFDMA and can provide better end-to-end energy efficiency than IFDMA and LFDMA. These investigations motivated the use of the DFT precoding step, and the integration of the TDMA component within the B-IFDMA scheme. These results also motivated the use of a regular subcarrier allocation in B-IFDMA.

Then, we showed that B-IFDMA offers the possibility to combine robustness against carrier frequency offsets and Doppler spread with provision of frequency diversity. This property could provide a significant gain in certain scenarios, like for deployment in frequency bands using a several GHz carrier frequency. Another scenario is to support high-speed trains, and/or to deploy a system with rather narrow subcarrier bandwidth.

Finally, we argued for that B-IFDMA has the capability to fulfill and provide a good tradeoff between the requirements envisioned for the robust transmission mode in the uplink of future IMT Advanced capable wireless systems.

Appendix

Time Domain Representation

In this appendix, the samples of the B-IFDMA time domain signal are analyzed. For sake of simplicity, throughout this appendix, the index η is omitted. Combining the precoding, the user specific block-interleaved subcarrier mapping and the OFDM modulation, the elements $x_n^{(k)}$, $n = 0, \ldots, N-1$, of the modulated data vector $\mathbf{x}^{(k)}$ in (6) can be written as

$$x_n^{(k)} = \sum_{\mu=0}^{M-1} d_{(n+\mu L) \bmod Q}^{(k)} \cdot \Theta_n^{(\mu,k)} \tag{A.1}$$

with

$$\Theta_n^{(\mu,k)} = \frac{L}{\sqrt{QN}} e^{j(2\pi/N)nkM} \sum_{m=0}^{M-1} e^{-j2\pi m(n/Q - n/N + \mu/M)} \tag{A.2}$$

for $\mu = 0, \ldots, M-1$. The derivation can be found in [46]. Equation (A.1) is illustrated in Figure 15.

The sequence $d_{(n+\mu L) \bmod Q}^{(k)}$ in (A.1) can be interpreted as a compression of the sequence of data symbols $d_q^{(k)}$, $q = 0, \ldots, Q-1$, in time by factor N/Q, a subsequent N/Q-fold repetition and, finally, a cyclic shift of the N elements of the resulting sequence by $n + \mu L$, as illustrated in [46, Figure 3]. Thus, B-IFDMA can be considered as a superposition of M single carrier signals weighted by different complex numbers $\Theta_n^{(\mu,k)}$.

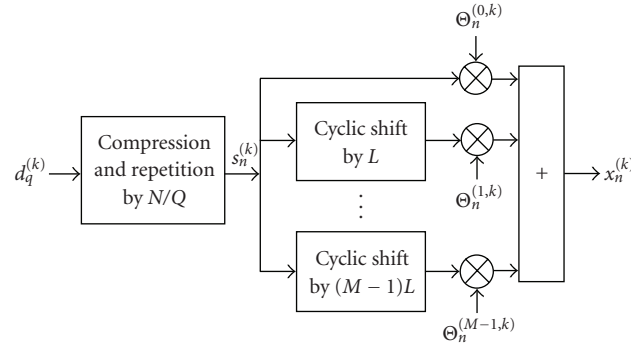


Figure 15: B-IFDMA time domain modulation.

The expression $e^{j(2\pi/N)kM}$ in (A.2) represents a user specific frequency shift by kM. For $M = 1$ the expression $\Theta_n^{(\mu,k)}$ reduces to

$$\Theta_n^{(\mu,k)} = \sqrt{\frac{Q}{N}} e^{j(2\pi/N)kn}, \tag{A.3}$$

and (A.1) simplifies to compression and repetition of the data symbols, subsequent user specific frequency shift and a normalization to $\sqrt{Q/N}$. This is equivalent to the generation of an IFDMA signal [14] which is a single carrier signal, that is, a serially modulated carrier. Note that also for the special case $M = Q$, an expression is obtained that is equivalent to a single carrier signal because for this case the expression from (A.1) is equivalent to the time domain samples of an LFDMA signal that are described in [47].

Since the coefficients by (A.2) are independent from the data symbols, they can be calculated offline. Thus, (A.1) also represents an alternative implementation for B-IFDMA modulation that does not require N and Q to be powers of 2 as it would be the case for an implementation of B-IFDMA modulation according to (6), if the Fast Fourier Transform (FFT) algorithm is used.

Similar to the modulation, also the B-IFDMA demodulation can be reformulated as follows. The elements $\rho_q^{(k)}$, $q = 0, \ldots, Q-1$ of the demodulated B-IFDMA signal

$$\rho^{(k)} = \mathbf{F}_Q^H \cdot \left(\mathbf{M}_{\mathrm{BI}}^{(k)}\right)^{\dagger} \cdot \mathbf{F}_N \cdot \mathbf{r} \tag{A.4}$$

can be expressed as

$$\rho_q^{(k)} = \sum_{\nu=0}^{N/L-1} r_{(q+\nu L) \bmod N} \cdot \Psi_{(q+\nu L) \bmod N}^{(\nu,k)} \tag{A.5}$$

with

$$\Psi_n^{(\nu,k)} = \left(\Theta_n^{(-\nu \bmod M,k)}\right)^*; \quad n = (q + \nu L) \bmod N. \tag{A.6}$$

The derivation of the demodulator can be found in [46] with an illustration in [46, Figure 4]. It can be regarded as a generalization of the demodulation for IFDMA described in [14].

Acknowledgments

This work has partly been performed in the framework of the IST project IST-4-027756 WINNER II, which was partly funded by the European Union. The authors would like to acknowledge the contributions of their colleagues in WINNER II, although the views expressed are those of the authors and do not necessarily represent the project.

References

[1] International Telecommunication Union Radiocommunication Sector ITU-R, "Recommendation ITU-R M1645, framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000," 2003.

[2] C. Wijing, K. Doppler, K. Kalliojärvi, et al., "WINNER II system concept: advanced radio technologies for future wireless systems," in *Proceedings of the ICT-Mobile Summit Conference*, June 2008.

[3] S. Parkvall, E. Dahlman, A. Furuskär, et al., "LTE-advanced evolving LTE towards IMT-advanced," in *Proceedings of the 68th IEEE Vehicular Technology Conference (VTC '08)*, pp. 1–5, Calgary, Canada, September 2008.

[4] G. Auer, M. Döttling, K. Doppler, et al., "D6.13.14 WINNER II system concept description," Tech. Rep. IST-4-027756 WINNER II, version 1.1, WINNER, January 2008.

[5] M. Döttling, W. Mohr, and A. Osseiran, *Radio Technologies and Concepts for IMT-Advanced*, John Wiley & Sons, New York, NY, USA, 2009.

[6] A. Tyrrell and G. Auer, "Imposing a reference timing onto firefly synchronization in wireless networks," in *Proceedings of the 65th IEEE Vehicular Technology Conference (VTC '07)*, pp. 222–226, Dublin, Ireland, April 2007.

[7] D. Falconer, B. Ng, C.-T. Lam, et al., "D2.3.3 WINNER II link level procedures for the WINNER system," Tech. Rep. IST-4-027756 WINNER II, version 1.00, WINNER, November 2007.

[8] M. Döttling, M. Sternad, G. Klang, J. von Häfen, and M. Olsson, "Integration of spatial processing in the WINNER B3G air interface design," in *Proceedings of the 63rd IEEE Vehicular Technology Conference (VTC '06)*, vol. 1, pp. 246–250, Melbourne, Australia, May 2006.

[9] M. Sternad, T. Svensson, T. Ottosson, A. Ahlén, A. Svensson, and A. Brunstrom, "Towards systems beyond 3G based on adaptive OFDMA transmission," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2432–2455, 2007.

[10] M. Sternad, S. Falahati, T. Svensson, and D. Aronsson, "Adaptive TDMA/OFDMA for wide-area coverage and vehicular velocities," in *Proceedings of the IST Mobile and Vehicular Communication Summit*, Dresden, Germany, June 2005.

[11] K. Safjan, J. Oszmianski, M. Döttling, and A. Bohdanowicz, "Frequency-domain link adaptation for wideband OFDMA systems," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '08)*, pp. 1703–1708, Las Vegas, Nev, USA, April 2008.

[12] T. Svensson, T. Franky, D. Falconer, et al., "B-IFDMA—a power efficient multiple access scheme for non-frequency-adaptive transmission," in *Proceedings of the 16th IST Mobile and Wireless Communications Summit*, Budapest, Hungary, July 2007.

[13] D. Galda, H. Rohling, E. Costa, H. Haas, and E. Schulz, "A low complexity transmitter structure for OFDM-FDMA uplink systems," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '02)*, vol. 4, pp. 1737–1741, Birmingham, UK, May 2002.

[14] U. Sorger, I. De Broeck, and M. Schnell, "IFDMA—a new spreadspectrum multiple-access scheme," in *Multi-Carrier Spread-Spectrum*, pp. 111–118, Kluwer Academic Publishers, Amsterdam, The Netherlands, 1997.

[15] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution: HSPA and LTE for Mobile Broadband*, Academic Press, New York, NY, USA, 2007.

[16] The 3rd Generation Partnership Project (3GPP), "Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)," Technical Specification Group Radio Access Network TR 25.814 v7.1.0, 3GPP, Sophia Antipolis Cedex, France, 2006.

[17] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30–38, 2006.

[18] S. Plass, T. Svensson, and A. Dammann, "Block-equidistant resource mapping in OFDM, MC-CDMA and SS-MC-MA," in *Proceedings of the 12th International OFDM Workshop*, Hamburg, Germany, August 2007.

[19] M. Döttling, et al., "D6.13.7 WINNER II test scenarios and calibration cases issue 2," Tech. Rep. IST-4-027756 WINNER II, version 1.00, WINNER, December 2006.

[20] T. Frank, A. Klein, E. Costa, and E. Schulz, "Interleaved orthogonal frequency division multiple access with variable data rates," in *Proceedings of the International OFDM Workshop*, pp. 179–183, Hamburg, Germany, August-September 2005.

[21] T. Frank, A. Klein, and E. Costa, "IFDMA: a scheme combining the advantages of OFDMA and CDMA," *IEEE Wireless Communications*, vol. 14, no. 3, pp. 9–17, 2007.

[22] Z. Wang and G. B. Giannakis, "Wireless multicarrier communications," *IEEE Signal Processing Magazine*, vol. 17, no. 3, pp. 29–48, 2000.

[23] H. Sari, G. Karam, and I. Jeanclaude, "Frequency domain equalization of mobile radio terrestrial broadcast channels," in *Proceedings of the IEEE Global Telecommunications Conference (GLOBECOM '94)*, pp. 1–5, San Francisco, Calif, USA, November-December 1994.

[24] D. Falconer, S. Ariyavisitakul, A. Benyamin-Seeyar, and B. Eidson, "Frequency domain equalization for single-carrier broadband wireless systems," *IEEE Communications Magazine*, vol. 40, pp. 58–66, 2002.

[25] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 8, pp. 1451–1458, 1998.

[26] H. Boelcskei and A. J. Paulraj, "Space-frequency coded broadband OFDM systems," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '00)*, vol. 1, pp. 1–6, Chicago, Ill, USA, September 2000.

[27] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, 1974.

[28] P. Kyösti, J. Meinilä, L. Hentila, et al., "D1.1.2 WINNER II channel models—part I channel models," Tech. Rep. IST-4-027756 WINNER II, version 1.2, WINNER, September 2007.

[29] G. Auer, "Analysis of pilot-symbol aided channel estimation for OFDM systems with multiple transmit antennas," in

Proceedings of the IEEE International Conference on Communications (ICC '04), vol. 6, pp. 3221–3225, Paris, France, June 2004.

[30] P. Hoeher, S. Kaiser, and P. Robertson, "Two-dimensional pilot-symbol aided channel estimation by Wiener filtering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 3, pp. 1845–1848, Munich, Germany, April 1997.

[31] O. Edfors, M. Sandell, J.-J. van de Beek, S. Wilson, and P. Börjesson, "OFDM channel estimation by singular value decomposition," *IEEE Transactions on Communications*, vol. 46, no. 7, pp. 931–939, 1998.

[32] K. Fazel and S. Kaiser, *Multi-Carrier Spread Spectrum Systems*, John Wiley & Sons, New York, NY, USA, 1st edition, 2003.

[33] C. T. Lam, D. Falconer, and F. Danilo-Lemoine, "Channel estimation for sub-chunk-based DFT-precoded OFDM systems," in *Proceedings of the 18th Meeting on Wireless World Research Forum (WWRF '07)*, Helsinki, Finland, June 2007.

[34] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice-Hall, Upper Saddle River, NJ, USA, 2000.

[35] D. Aronsson, *Channel estimation and prediction from a Bayesian perspective*, Licentiate thesis, Department of Signals and Systems, Uppsala University, Uppsala, Sweden, 2007.

[36] C. Wijting, K. Doppler, K. Kalliojärvi, et al., "Key technologies for IMT-advanced mobile communication systems," *IEEE Wireless Communications Magazine*, 2009.

[37] D. Aronsson, T. Svensson, and M. Sternad, "Performance evaluation of memory-less and Kalman-based channel estimation for OFDMA," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '09)*, Barcelona, Spain, April 2009.

[38] T. Frank, A. Klein, and T. Haustein, "A survey on the envleope fluctuations of DFT precoded OFDMA," in *Proceedings of the International Conference on Communications (ICC '08)*, Beijing, China, May 2008.

[39] R. van Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*, Artech House, Boston, Mass, USA, 1st edition, 2000.

[40] A. Skrzypczak, P. Siohan, and J.-P. Javaudin, "Power spectral density and cubic metric for the OFDM/OQAM modulation," in *Proceedings of the 6th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT '06)*, pp. 846–850, Vancouver, Canada, August 2006.

[41] The 3rd Generation Partnership Project (3GPP), "Comparison of PAR and cubic metric for powerderating," Technical Documents TDoc R4-040367, TSG RAN WG4 31, 3GPP, Sophia Antipolis Cedex, France, 2004.

[42] H. M. Nemati, C. Fager, U. Gustavsson, et al., "Characterization of switched-mode LDMOS and GaN power amplifiers for optimal use in polar transmitter architectures," in *Proceedings of the IEEE International Microwave Symposium*, Atlanta, Ga, USA, June 2008.

[43] F. Wang, D. Kimball, J. Popp, et al., "Wideband envelope elimination and restoration power amplifier with high efficiency wideband envelope amplifier for WLAN 802.11g applications," in *Proceedings of the Microwave Symposium Digest*, pp. 645–648, La Jolla, Calif, USA, June 2005.

[44] T. H. Eggen, A. B. Baggeroer, and J. C. Preisig, "Communication over doppler spread channels-part I: channel and receiver presentation," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 1, pp. 62–71, 2000.

[45] M. Sternad, T. Svensson, and M. Döttling, "Resource allocation and control signaling in the WINNER flexible MAC concept," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, Calgary, Canada, September 2008.

[46] T. Frank, A. Klein, and E. Costa, "An efficient implementation for block-IFDMA," in *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, Athens, Greece, September 2007.

[47] H. G. Myung, J. Lim, and D. J. Goodman, "Peak-to-average power ratio of single carrier FDMA signals with pulse shaping," in *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '06)*, Helsinki, Finland, September 2006.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 406028, 10 pages
doi:10.1155/2009/406028

Research Article

Diversity Techniques for Single-Carrier Packet Retransmissions over Frequency-Selective Channels

Abdel-Nasser Assimi, Charly Poulliat, and Inbar Fijalkow (EURASIP Member)

ETIS, CNRS, ENSEA, Cergy-Pontoise University, 6 avenue du Ponceau, 95000 Cergy-Pontoise, France

Correspondence should be addressed to Abdel-Nasser Assimi, abdelnasser.assimi@ensea.fr

Received 16 February 2009; Revised 16 June 2009; Accepted 16 August 2009

Recommended by Stefania Sesia

In data packet communication systems over multipath frequency-selective channels, hybrid automatic repeat request (HARQ) protocols are usually used in order to ensure data reliability. For single-carrier packet transmission in slow fading environment, an identical retransmission of the same packet, due to a decoding failure, does not fully exploit the available time diversity in retransmission-based HARQ protocols. In this paper, we compare two transmit diversity techniques, namely, cyclic frequency-shift diversity and bit-interleaving diversity. Both techniques can be integrated in the HARQ scheme in order to improve the performance of the joint detector. Their performance in terms of pairwise error probability is investigated using maximum likelihood detection and decoding. The impact of the channel memory and the modulation order on the performance gain is emphasized. In practice, we use low complexity linear filter-based equalization which can be efficiently implemented in the frequency domain. The use of iterative equalization and decoding is also considered. The performance gain in terms of frame error rate and data throughput is evaluated by numerical simulations.

1. Introduction

Single carrier with cyclic-prefix transmissions has recently gained a certain attention, especially after its adoption for the uplink in the 3GPP Long-Term-Evolution (LTE) standard [1]. Actually, single-carrier signaling provides a low peak-to-average power ratio (PAPR) compared to the orthogonal frequency division multiplexing (OFDM). Moreover, the insertion of a cyclic prefix allows simplified signal processing in the frequency domain at the receiver. Reliable data communication systems usually implement HARQ protocols [2] in order to combat errors introduced by the communication channel. This includes channel noise and intersymbol interference (ISI) resulting from multipath propagation in wireless channels. In order to reduce the effect of the ISI on the performance of the system, one could implement a sophisticated detection scheme at the receiver, such as a turboequalizer [3], for example, at the expense of increased receiver complexity. Another possibility is to use a simple linear equalizer with a low rate channel code in order to handle the residual interference remaining after equalization. The price to pay for this solution is reduced data throughput, even in good channel conditions.

In the context of HARQ protocols, joint equalization of multiple received copies of the same packet significantly enhances system performance, especially when there is channel diversity among subsequent HARQ transmissions. When a part of the available bandwidth falls in a deep fading, a decoding failure may occur and a retransmission request is made by the receiver. An identical retransmission of the same packet would suffer from the same problem if the channel remains unchanged. Combining both received packets provides some signal-to-noise ratio (SNR) gain resulting from noise averaging, but the interference power remains the same.

In order to enhance the joint detection performance, many transmit diversity schemes have been proposed for multiple HARQ transmissions. When channel state information at the transmitter (CSIT) is available, precoding (preequalization) techniques [4, 5] can be used at the transmitter in order to transform the frequency selective

channel into a flat channel. In [6], linear precoding filters are optimized for multiple HARQ transmissions. In general, linear filtering increases the PAPR of the transmitted signal, especially when the channel response contains a deep fading. Note that methods based on the availability of CSIT require an increased load on the feedback channel. In addition, these methods can be sensitive to channel mismatch and can not be applied when the channel changes rapidly from one transmission to the next.

For communication systems with very limited feedback channels, the CSIT assumption is not applicable. However, in the absence of CSIT, there are some useful techniques that enhance the system performance in slow time-varying channel conditions while keeping the system performance unchanged in fast changing channel conditions without the need for switching mechanisms. In the absence of CSIT, a phase-precoding scheme has been proposed in [7]. In this scheme, a periodic phase rotation pattern is applied for each HARQ transmission in order to decorrelate the ISI among the received copies of the same packet. This can be seen in the frequency domain as a frequency shift by more than the coherence bandwidth of the channel. The advantage of the phase-precoding transmit diversity scheme is the conservation of the power characteristics of the transmitted symbols. Hence, it does not increase the PAPR of the transmitted signal. Another transmit diversity scheme is the bit-interleaving diversity initially proposed in [8] for noncoded transmissions using iterative equalization at the receiver. This scheme outperforms joint equalization of identically interleaved transmissions but it has higher complexity. For coded transmissions, it has been found in [9] that the iterative equalization approach is not suitable for the bit-interleaving diversity. Performing separate equalization with joint decoding instead leads to a significant performance improvement and reduced complexity. In [10], a mapping diversity scheme was proposed for high-order modulations. This scheme results in an increased Euclidean distance separation between transmitted frames. The drawback of this method is to be limited to high-order modulations which makes it not applicable for BPSK or QPSK modulations.

In this paper, we compare two transmit diversity schemes: the cyclic frequency-shift diversity and the bit-interleaving diversity. The theoretical comparison is performed assuming optimal ML detection and decoding. Since the ML receiver is practically nonrealistic, an iterative receiver using a turboequalizer is considered in this paper in order to verify the theoretical results. However, the performance of a noniterative receiver is also evaluated for low complexity requirements.

The remaining of this paper is organized as follows. In Section 2, the system model for both diversity schemes is introduced. In Section 3, we investigate their respective performance using an optimal ML receiver. In Section 4, we present the corresponding receivers and investigate their respective complexity. In Section 5, we give some simulation results showing the advantages of each diversity scheme for different system parameters. Finally, conclusions are given in Section 6.

Notation. The following notations are used throughout this paper. Uppercase boldface letters ($\mathbf{A}$) denote matrices; lowercase boldface letters ($\mathbf{a}$) denote (column) vectors, and italics (a, A) denote scalars; an ensemble of elements is represented with calligraphic fonts ($\mathcal{A}$).

2. System Model

We consider the communication system model shown in Figure 1 using single carrier bit-interleaved coded modulation with multiple HARQ transmissions over a frequency selective channel.

A data packet $\mathbf{d}$, of KQ information bits including cyclic redundancy check (CRC) bits for error detection, is first encoded by a rate-K/N error correction code to obtain QN coded bits $\mathbf{c}$. The codeword $\mathbf{c}$ is stored at the transmitter in order to be retransmitted later if it is requested by the receiver due to a transmission error. Each branch in Figure 1 corresponds to a single transmission of the same packet. Thus, for $t = 1, 2, \ldots, T$, the tth branch corresponds to the tth (re)transmission of $\mathbf{c}$ according to the considered HARQ scheme.

For the first transmission of the coded packet, a bit-interleaver $\pi^{(1)}$ is applied on $\mathbf{c}$ in order to statistically decorrelate the encoded bits. The obtained coded and interleaved bits $\mathbf{c}^{(1)}$ are then mapped into a sequence of N symbols, denoted by $\mathbf{s}^{(1)}$, using a complex constellation alphabet $\mathcal{S}$ of size $|\mathcal{S}| = 2^Q$ symbols having unit average power. The modulated symbols are then processed by a channel precoder to generate the signal $\mathbf{x}^{(1)}$. In this paper, the channel precoder performs a simple cyclic frequency-shift (CFS) operation on the signal $\mathbf{s}^{(1)}$. Before the transmission of $\mathbf{x}^{(1)}$ over the propagation channel, a cyclic prefix (CP) of length P is inserted at the beginning of the packet in order to avoid interpacket interference and to facilitate the equalization in the frequency domain.

At the receiver side, if the packet is successfully decoded by the receiver, a positive acknowledgment (ACK) signal is returned to the transmitter through an error-free feedback channel with zero delay; otherwise a negative acknowledgment (NACK) signal is returned indicating a decoding failure. In the latter case, the transmitter responds by resending the same coded packet $\mathbf{c}$ but in a different way according to the considered transmit diversity scheme. If the packet is still in error after a maximum number $T_{\max}$ of allowable transmissions (the first transmission plus $T_{\max} - 1$ possible retransmissions), an error is declared and the packet is dropped out from the transmission buffer.

Note that this model corresponds to SC-FDMA transmission in LTE system when each user is allocated the entire system bandwidth as in time division multiplexing. However, the main results of this paper are still applicable when the same subcarriers are allocated to the user during all HARQ retransmissions by considering the equivalent channel response seen by the user's carriers. We define three transmission schemes.

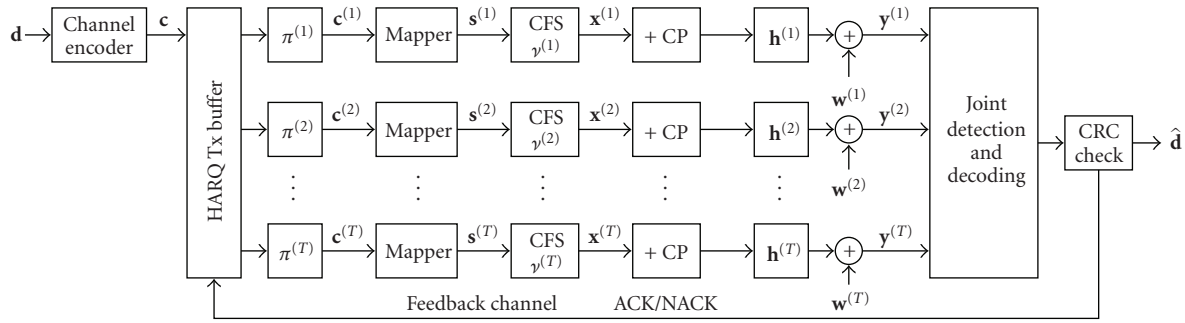


Figure 1: System model for single-carrier cyclic-prefix transmit diversity for HARQ retransmission protocols.

(a) Identical Transmissions (IT) Scheme. In this scheme, the same interleaver is used for all transmissions with no channel precoding. As stated in the introduction of this paper, the benefit of the IT-HARQ scheme in slow time-varying channels is the SNR gain due to noise averaging. This scheme is used as a reference in order to evaluate the gain introduced by the other diversity schemes.

(b) Bit-Interleaving Diversity (BID) Scheme. In this scheme, a different bit-interleaver is used for each retransmission with no channel precoding.

(c) Cyclic Frequency-Shift Diversity (CFSD) Scheme. In this scheme, the same interleaver is used for all transmissions but a different channel precoder is used for each transmission. The precoder cyclically shifts the transmitted signal in the frequency domain by the normalized frequency value $\nu^{(t)} = k/N$ for $k \in [0, N-1]$, where t denotes the HARQ transmission index. This operation can be performed in the time domain by

$$x^{(t)}(n) = e^{j2\pi n\nu^{(t)}} s^{(t)}(n), \tag{1}$$

for $n = 0, \ldots, N-1$.

The transmission channel is frequency-selective modeled by its equivalent complex-valued discrete-time finite impulse response of length L, denoted by $\mathbf{h}^{(t)} = (h^{(t)}(0), \ldots, h^{(t)}(L-1))$ assumed constant during the period of one packet transmission. Each channel tap is a zero mean complex random variable with a given variance which is determined from the power-delay profile of the channel. In addition, we assume that the channel response changes slowly from one transmission to the next. In our analysis, we consider the long-term static channel model where the channel remains the same for all HARQ transmissions of the same packet, but changes independently from packet to packet as in [11]. The independence assumption between channel responses from packet to packet may not be justified in practice, but it is adopted in this paper in order to evaluate the average system performance for all possible channel realizations from link to link. However, we keep the indexing of the channel response by the transmission index t for the sake of generality of the receiver structure. Moreover, we assume that the length of the cyclic prefix P is larger than the maximum delay spread $L_{\max}$. According to this model, the received sequence samples, denoted by $y^{(t)}(n)$, are given by

$$y^{(t)}(n) = \sum_{i=0}^{L-1} h^{(t)}(i) x^{(t)}(n-i) + w^{(t)}(n), \tag{2}$$

where $w^{(t)}(n)$ is an additive complex white Gaussian noise with variance σ_w^2 ($\sigma_w^2/2$ per real dimension).

We compare the achievable performance between the different transmission schemes under investigation assuming an optimal joint ML receiver with perfect channel state information at the receiver while no CSIT is assumed. A comparative analysis based on the average pairwise error probability (PEP) is presented in Section 3.

3. Error Probability Analysis

In order to compare the theoretical performance of the BID and the CFSD schemes, we consider an optimal ML receiver, and we compare the properties of the Euclidean distance distribution at the output of the frequency-selective channel for multiple transmissions.

Let $\mathbf{c}$ and $\hat{\mathbf{c}}$ be the transmitted and the estimated binary codewords after T transmissions. Let $\mathbf{x}_T = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)})$ and $\hat{\mathbf{x}}_T = (\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(T)})$ be the corresponding transmitted sequences. We define the error sequence between $\hat{\mathbf{x}}_T$ and $\mathbf{x}_T$ by $\mathbf{e}_T \triangleq (\mathbf{e}^{(1)}, \ldots, \mathbf{e}^{(T)}) = \hat{\mathbf{x}}_T - \mathbf{x}_T$. For a joint ML receiver, Forney has shown in [12] that the PEP between any pair of sequences is given as a function of the error sequence $\mathbf{e}_T$ between them by

$$P_2(\hat{\mathbf{c}}, \mathbf{c}) = Q\left(\sqrt{\frac{d_E^2(\mathbf{e}_T)}{4\sigma_w^2}}\right), \tag{3}$$

where $Q(\cdot)$ is the complementary distribution function of standard Gaussian, and d_E is the Euclidean distance between $\hat{\mathbf{x}}_T$ and $\mathbf{x}_T$ at the output of the noiseless channel. For a given set of channel realizations $\{\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(T)}\}$, the squared Euclidean distance d_E^2 can be evaluated as

$$d_E^2(\mathbf{e}_T) = \sum_{t=1}^{T} \sum_{n=0}^{N-1} \left| \sum_{i=0}^{L-1} h^{(t)}(i) e^{(t)}(n-i) \right|^2. \tag{4}$$

By developing the squared sum in (4) and performing some algebraic computations, we obtain

$$d_E^2(\mathbf{e}_T) = \sum_{t=1}^{T} \sum_{\ell=-L+1}^{L-1} R_\ell^\star\left(\mathbf{h}^{(t)}\right) R_\ell\left(\mathbf{e}^{(t)}\right), \tag{5}$$

where the superscript $(\cdot)^\star$ denotes the complex conjugate and $R_\ell(\cdot)$ is the deterministic periodic autocorrelation function for a lag ℓ, defined for an arbitrary complex sequence $\mathbf{x}$ of length N by $R_\ell(\mathbf{x}) \triangleq \sum_{n=0}^{N-1} x(n)x^\star(n-\ell)$ with $x(-n) = x(N-n)$. Expression (5) for the squared Euclidean distance is equivalent to that given by Forney in [12] using polynomial notations.

From (5), we note that the channel and the error sequence have a symmetrical effect on the Euclidean distance through their respective autocorrelation functions. By analogy to channel diversity, transmit diversity is a way to decrease the probability of error sequences leading to a low output Euclidean distance. In fact, the auto-correlation function of the error sequence $R_\ell(\mathbf{e}^{(t)})$ depends simultaneously on the Hamming weight of the binary error sequence, the interleaving, and the mapping scheme. Therefore, most of diversity techniques try to enhance the statistical distribution of d_E by modifying some system parameters such as the mapping [13], or by adding additional devices at the transmitter such as a binary precoder [8], for example.

For convenience, we denote the squared Euclidean distance by the new variable $\Delta_T \triangleq d_E^2(\mathbf{e}_T)$. We can rewrite (5) as the sum of two variables as follows:

$$\Delta_T = \Gamma_T + \Theta_T, \tag{6}$$

with

$$\Gamma_T \triangleq \sum_{t=1}^{T} R_0\left(\mathbf{h}^{(t)}\right) R_0\left(\mathbf{e}^{(t)}\right), \tag{7}$$

$$\Theta_T \triangleq 2\Re\left[\sum_{t=1}^{T} \sum_{\ell=1}^{L-1} R_\ell^\star\left(\mathbf{h}^{(t)}\right) R_\ell\left(\mathbf{e}^{(t)}\right)\right], \tag{8}$$

where $\Re[\cdot]$ denotes the real part. In (6), the first variable Γ_T takes positive real values reflecting the effect of the channel gain on the squared Euclidean distance, whereas the second variable Θ_T takes signed real values reflecting the fluctuation of the Euclidean distance due to the presence of the ISI. For an ISI-free channel, it is obvious that $\Theta_T = 0$ and the performance limit for channel equalization are only determined by the properties of Γ_T.

The PEP depends actually on the Hamming weight d of the binary error codeword between $\hat{\mathbf{c}}$ and $\mathbf{c}$. The average PEP over the space of all possible error sequences of a given Hamming weight d and all channel realizations depends on the statistical distribution of Δ_T over this probability space. Since its difficult in general to analytically derive the probability density function (pdf) of Δ_T, we compare different transmission schemes by comparing the main statistical properties of Δ_T for each scheme, that is, the mean and the variance. A higher mean value and/or a smaller variance indicates better error performance. First, we compare the limiting performance of both diversity schemes assuming perfect interference cancellation by the receiver, then we compare the ISI power between them.

3.1. Performance Limits. A lower bound on the PEP can be obtained by assuming that the ISI is completely removed by the receiver, that is, $\Theta_T = 0$ and $\Delta_T = \Gamma_T$. This is equivalent to packet transmission over an equivalent flat-fading channel with an equivalent squared gain of $\gamma^{(t)} = R_0(\mathbf{h}^{(t)}) = \|\mathbf{h}^{(t)}\|^2$. This bound is usually referred to as the matched filter lower bound (MFB). Assuming that the channel remains the same for all retransmissions $\mathbf{h}^{(t)} = \mathbf{h}$ and defining $\varepsilon^{(t)} = \|\mathbf{e}^{(t)}\|^2$, we can rewrite (7) as

$$\Gamma_T = \gamma \sum_{t=1}^{T} \varepsilon^{(t)}. \tag{9}$$

The variable $\varepsilon^{(t)}$ depends on the binary error pattern and the underlying modulation. For each diversity scheme, we will calculate the mean and the variance of Γ_T.

For the CFSD scheme, multiplying each symbol by a unit amplitude complex number does not change the amplitude of the error symbol. Therefore, the variables $\varepsilon^{(t)}$ are identical. Let μ_e and σ_e^2 be the mean and the variance of $\varepsilon^{(1)}$. Let μ_h and σ_h^2 be the mean and the variance of the squared channel gain γ. Using the independence between $\varepsilon^{(1)}$ and γ, we obtain the following expressions for the mean and the variance of Γ_T:

$$\mu_{\mathrm{CFSD}}(\Gamma_T) = T\mu_h\mu_e, \tag{10}$$

$$\sigma^2_{\mathrm{CFSD}}(\Gamma_T) = T^2\mu_e^2\sigma_h^2 + T^2\left(\mu_h^2 + \sigma_h^2\right)\sigma_e^2. \tag{11}$$

Consequently, the performance limits for the CFSD scheme are the same as for the IT scheme.

For the BID scheme, assuming independent interleavers, the variables $\varepsilon^{(t)}$ are i.i.d. random variables. In this case we obtain

$$\mu_{\mathrm{BID}}(\Gamma_T) = T\mu_h\mu_e, \tag{12}$$

$$\sigma^2_{\mathrm{BID}}(\Gamma_T) = T^2\mu_e^2\sigma_h^2 + T\left(\mu_h^2 + \sigma_h^2\right)\sigma_e^2. \tag{13}$$

For a given mapping scheme the computation of μ_e and σ_e^2 is shown in the appendix under the uniform interleaving assumption [14] which gives the average estimations over all possible deterministic random interleavers. Note that μ_e and σ_e depend on the Hamming weight d.

By comparing (11) with (13), we note that the second term in the variance expression for the CFSD scheme is reduced by a factor T for the BID scheme. This reflects the inherent modulation diversity of the BID scheme because error bits are located in different symbols at each retransmission. However, in some special cases such as BPSK and QPSK modulations with Gray mapping, $\varepsilon^{(t)}$ is invariant to bit-interleaving. Indeed, we have $\varepsilon^{(t)} = \alpha d$, where $\alpha = 4$ for BPSK and $\alpha = 2$ for QPSK. Consequently, we have $\sigma_e^2 = 0$, and both diversity schemes have the same performance limits as for the IT scheme in this case. By contrast, for a higher order modulation such as 16-QAM or 64-QAM, $\sigma_e^2 \neq 0$ and some variance reduction can be expected.

3.2. Intersymbol Interference Power. In this section, we show the effect of both diversity schemes on the interference power by evaluating the variance of the variable Θ_T. For the long-term static channel model, (8) can be written as

$$\Theta_T = 2\Re\left[\sum_{\ell=1}^{L-1} R_\ell^\star(\mathbf{h}) S_\ell\right], \tag{14}$$

where $S_\ell = \sum_{t=1}^{T} R_\ell(\mathbf{e}^{(t)})$. Assuming that the channel tap coefficients are independent with zero mean, this implies that $R_\ell^\star(\mathbf{h})$ are zero mean random variables and pairwise uncorrelated for different ℓ. Consequently, Θ_T is also a zero mean random variable. In addition, we assume that both the channel response and the error sequence have the same power per real dimension; the variance of Θ_T can be computed as

$$\sigma^2(\Theta_T) = \mathbb{E}\left(|\Theta_T|^2\right) = 2\sum_{\ell=1}^{L-1} \mathbb{E}\left(\left|R_\ell^\star(\mathbf{h})\right|^2\right)\mathbb{E}\left(|S_\ell|^2\right). \tag{15}$$

The difference between both transmit diversity schemes concerns the value of $\mathbb{E}(|S_\ell|^2)$. Thanks to the interleaver, we can assume that error symbols e_n in the transmitted packet are uncorrelated (but not independent due to the constraint on their total Hamming weight d). Consequently, the random variables $R_\ell(\mathbf{e}^{(t)})$ have a zero mean and pairwise uncorrelated for different ℓ. This yields

$$\mathbb{E}\left(|S_\ell|^2\right) = \sum_{t=1}^{T} \mathbb{E}\left(\left|R_\ell\left(\mathbf{e}^{(t)}\right)\right|^2\right). \tag{16}$$

Moreover, two error symbols e_i and e_j are conditionally independent to their respective Hamming weight k_i and k_j. Using all previous assumptions, it is straightforward to compute the variance of S_ℓ for both diversity schemes.

For the BID scheme we obtain

$$\mathbb{E}\left(|S_\ell|^2\right)_{\text{BID}} = \overline{\rho}_s(N-\ell)T, \tag{17}$$

where $\overline{\rho}_s = \mathbb{E}(|s_i|^2|s_j|^2)$ for $i \neq j$ which can be computed as indicated in the appendix.

For the CFSD scheme we obtain

$$\mathbb{E}\left(|S_\ell|^2\right)_{\text{CFSD}} = \overline{\rho}_s(N-\ell)\lambda_\ell, \tag{18}$$

where

$$\lambda_\ell = \left|\sum_{t=1}^{T} e^{-j2\pi\ell\nu^{(t)}}\right|^2. \tag{19}$$

We remark from (15) that the variance $\sigma^2(\Theta_T)$ depends on the power-delay profile of the channel. Since no CSIT is assumed, the optimal frequency-shift values are those that minimize the objective function $J_T = \sum_{\ell=1}^{L-1} \lambda_\ell^2$. As it is shown in [15], this function can achieve its absolute minimum value when

$$\lambda_\ell = T\frac{L-T}{L-1}, \quad \forall \ell, T < L. \tag{20}$$

This minimum value could be achieved by a proper choice of $\nu^{(t)}$ from the set $\{k/L : k = 0, \ldots, L-1\}$. For unknown channel length L, frequency shifts can be chosen as the maximum possible in order to take account for the shortest channel memory.

By comparing the value of $\mathbb{E}(|S_\ell|^2)$ for the BID scheme given in (17) with its value for the CFSD scheme given in (18), we note that the CFSD scheme leads to a smaller interference variance $\sigma^2(\Theta_T)$ because $\lambda_\ell < T$. In the particular case when $T = L$, we can have $\lambda_\ell = 0$, hence $\sigma^2(\Theta_T) = 0$ which means that the interference is completely cancelled by the CFSD scheme.

For large values of channel memory L, we have $\lambda_\ell \approx T$ and the difference between the two diversity schemes with regard to the ISI power becomes smaller. Note that for the IT scheme, we have $\mathbb{E}(|S_\ell|^2) = \rho_s(N-\ell)T^2$ which is obtained by setting $\nu^{(t)} = 0$ in (18).

In conclusion, the BID scheme has a better performance limit than the CFSD scheme for high-order modulations, but the CFSD scheme is more efficient in combating the interference for a short channel memory.

4. Iterative Receiver Structure

It is known that the performance of an optimal ML receiver can be approached by using an iterative equalization and decoding approach as in turboequalization. In this section we present the structure of the turboequalizer with integrated packet combining for both diversity schemes with the purpose of showing the performance-complexity tradeoff achieved by these diversity techniques.

4.1. Cyclic Frequency-Shift Diversity. The receiver structure for the CFSD scheme is shown in Figure 2. For each received frame $\mathbf{y}^{(t)}$, the CP is first removed and then a discrete Fourier transform (DFT) is applied in order to perform equalization in the frequency domain. In the following, the DFTs of signals are denoted by capital letters as a function of the normalized frequency ν. Thanks to the cyclic prefix insertion, the time-domain convolution becomes a simple multiplication in the frequency domain. The received frame can be written as

$$Y^{(t)}(\nu) = H^{(t)}(\nu)X^{(t)}(\nu) + W^{(t)}(\nu). \tag{21}$$

The inverse frequency shift is performed on $Y^{(t)}$ to obtain $Z^{(t)}$ which is given by

$$\begin{aligned} Z^{(t)}(\nu) &= Y^{(t)}(\nu - \nu_t) \\ &= H^{(t)}(\nu - \nu_t)S(\nu) + W^{(t)}(\nu - \nu_t) \\ &= \widetilde{H}^{(t)}(\nu)S(\nu) + \widetilde{W}^{(t)}(\nu). \end{aligned} \tag{22}$$

This gives the equivalent single-input multiple-output (SIMO) model for the CFSD scheme, where $\widetilde{H}^{(t)}$ is the equivalent channel and $\widetilde{W}^{(t)}$ is the equivalent noise. The signals $Z^{(t)}$ are then processed by a turboequalizer including two soft-input soft-output (SISO) modules which are connected

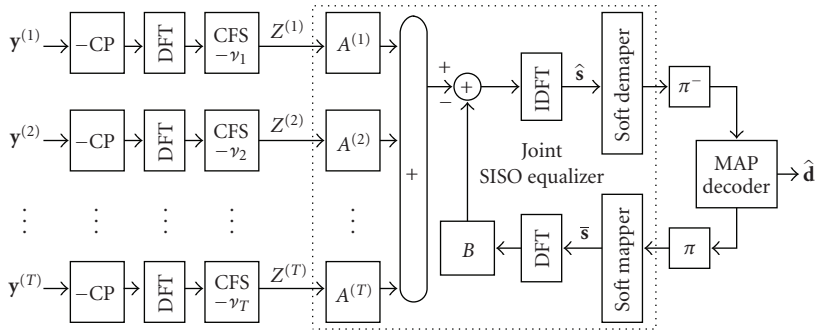


Figure 2: Iterative receiver structure for the CFSD scheme with joint equalization.

iteratively through the interleaver. One SISO module for joint MMSE equalization operating in the frequency domain and another SISO module for a maximum a posteriori (MAP) channel decoding [16] operating in the time domain. The joint MMSE equalizer includes multiple forward linear filters $A^{(t)}$ and a backward filter B. According to this structure, the linear estimate $\hat{\mathbf{s}}$ of $\mathbf{s}$ after T transmissions is given by

$$\hat{S} = \sum_{t=1}^{T} A^{(t)} Z^{(t)} - B\bar{S}. \tag{23}$$

Following the same analysis in [17, 18] and using the equivalent SIMO model, the derivation of the MMSE filters that minimize the mean square error $E[|\hat{s}(n) - s(n)|^2]$ is straightforward and leads to the following solution:

$$\begin{aligned} A^{(t)} &= \frac{\left(\widetilde{H}^{(t)}\right)^*}{\sigma_w^2 + \nu \sum_{t=1}^{T} \left|\widetilde{H}^{(t)}\right|^2}, \\ B &= \sum_{t=1}^{T} A^{(t)} \widetilde{H}^{(t)} - \mu, \\ \nu &= \frac{1}{N} \sum_{n=0}^{N-1} \mathrm{var}(\bar{s}(n)), \\ \mu &= \frac{1}{N} \sum_{k=0}^{N-1} \frac{H_T^2(k/N)}{\sigma_w^2 + \nu H_T^2(k/N)}, \end{aligned} \tag{24}$$

where H_T is the compound channel defined by its squared amplitude $H_T^2 \triangleq \sum_{t=1}^{T} |\widetilde{H}^{(t)}|^2$ and ν is reliability of the decoder feedback, where $\nu = 0$ indicates a perfect feedback, and $\nu = 1$ for no a priori. The output of the MMSE estimator can be written in the time domain after an IDFT using the Gaussian model for the estimated symbols as

$$\hat{s}(n) = \mu s(n) + \eta(n), \tag{25}$$

where η is a complex Gaussian noise with zero mean and variance $\sigma_\eta^2 = \mu(1 - \nu\mu)$. The output extrinsic a posteriori probabilities (APPs) are given by

$$\mathrm{APP}(s(n) = s \in \mathcal{S}) = K \exp\left(-\frac{|\hat{s}(n) - \mu s|^2}{\sigma_\eta^2}\right), \tag{26}$$

where K is a normalization factor in order to have a true probability mass function. The extrinsic log-likelihood ratios (LLRs) of the coded bits are then computed by soft demapping in order to decode the received frame by a MAP decoder after deinterleaving. For an iterative processing, the decoder's soft decisions in the form of extrinsic LLRs are interleaved and returned to the equalizer which, in turn, produces soft symbol decisions $\bar{\mathbf{s}}$ to be used as priory in the next iteration. Note that for separate detection and decoding, one can put the equalizer's soft input to zero ($\nu = 1$).

With regard to the system complexity, we see that the CFSD requires only N additional complex multiplications at the transmitter and a simple vector shift operation at the receiver. In addition, the complexity of the joint MMSE equalizer in the frequency domain is almost the same as for an MMSE equalizer with a single input. To show that, we note that the numerator of each forward filter is the matched filter to the channel which does not change with turboiterations. Hence, it is performed once per transmission. Since the denominator is common for all forward filters, the division can be performed after summation of the matched filters outputs. Consequently, for each new reception, the accumulated sum of the matched filters is updated and the same for the squared compound channel. Other operations are the same as for an equalizer with single input.

4.2. Bit-Interleaving Diversity. Joint equalization for the BID scheme is not possible because the transmitted symbols at each HARQ round are different. Therefore, we perform a postcombining at the bit level by adding the LLRs issued from all equalizers as shown in Figure 3. The structure of the SISO equalizer is similar to the joint equalizer presented for the CFSD scheme with only one single input.

Here, we need for each turboiteration two DFT operations and two interleaving operations per equalizer. Since there is T parallel equalizers in the BID scheme, the complexity of the receiver increases linearly with the number of transmissions. While in the CFSD scheme, there is one joint equalizer which requires only two DFTs and two interleaving operations per turbo-iteration independently of the number of transmissions. Therefore, the BID scheme has a larger complexity in comparison with the CFSD scheme if turbo-equalization is performed.

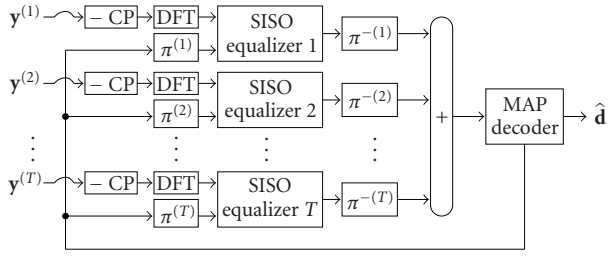


Figure 3: Iterative receiver structure for the BID scheme with separate equalization and LLR combining.

Table 1: Simulation parameters.

Parameter	Value
Frame length	$N = 516$ for QPSK, $N = 258$ for 16-QAM
Symbol rate	7.68 Msps
CP length	$P = 64$
Channel model	SCME urban macroscenario
Shaping filter	Raised cosine with roll off 0.23
Doppler	No Doppler

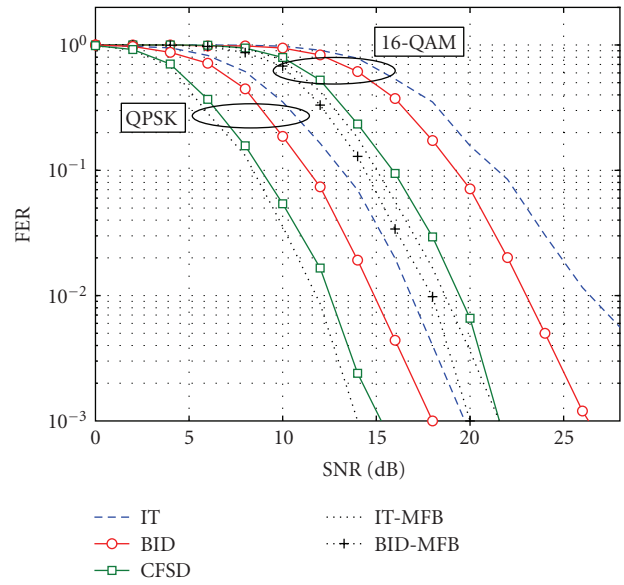


Figure 4: FER performance comparison between different transmission schemes for a non coded system using QPSK and 16-QAM modulations.

5. Results

In this section, we present some simulation results comparing the performance of the two transmit diversity schemes for different system configurations.

Simulations are performed using the 3GPP Spatial Channel Model Extended (SCME) of the European WINNER framework as specified in [19, 20] in the case of monoantenna transmission. This channel model is characterized by six nonzero taps with varying delays per link. For each transmitted packet, a random channel realization is generated and then used for all HARQ retransmissions of the packet. The system performance is evaluated in terms of FER versus the average SNR defined by $E_s/N_0 = 1/\sigma_w^2$. We assume that the maximum of HARQ transmissions is $T_{\max} = 4$. For the CFSD scheme, frequency-shift parameters are $\nu^{(1)} = 0$, $\nu^{(2)} = 1/2$, $\nu^{(3)} = 1/4$, and $\nu^{(4)} = 3/4$. All used interleavers are pseudorandom interleavers. Other simulation parameters inspired from the LTE standard [21] are listed in Table 1. Monte Carlo simulations are performed over a maximum of 5000 packets.

We first consider a noncoded transmission system in order to show the intrinsic gain for both diversity schemes compared to the identical transmission scheme. This corresponds to the system performance before channel decoding for coded systems. Figure 4 shows the FER performance versus the average SNR after the last HARQ round ($T = 4$) for QPSK and 16-QAM modulations.

We can observe the superiority of the CFSD scheme among all transmission schemes due to its best capability in interference mitigation. For QPSK modulation, we have SNR gain at FER = 10^{-2} of about 2 dB for the BID scheme and 4 dB for the CFSD scheme in comparison with the IT scheme. Note that the CFSD scheme is only at 0.4 dB of the MFB which is the same for all schemes. For 16-QAM modulation, the MFB for the BID scheme gives the best performance, but the better performance for the CFSD scheme is due to better performance of the joint equalization compared to the LLR combining used for the BID scheme. It is true that the used channel has a large channel memory which may attain more than 100 symbol periods, but it has a decreasing power-delay profile with most of the interference power originating from the less delayed paths. In this sense, the effective channel memory is not very large. This explains the larger interference reduction in the case of the CFSD scheme.

Now, we consider a coded system with a noniterative receiver including separate equalization and channel decoding without turboiteration. The performance of the noniterative receiver is obtained by performing one equalization step followed by one channel decoding step.

The channel code is the LTE turbocode of rate-1/3 using two identical constituent convolutional codes $(1, 15/13)_8$ with quadratic permutation polynomial internal interleaver of length $K = 344$ taken from [21, (Table 5.1.3-3)]. For simplicity, no trellis termination is performed for the component codes. The receiver performs one equalization step followed by one channel decoding step. The channel decoder itself performs a maximum of five internal iterations between the two internal convolutional decoders in the turbodecoder. Simulation results are given in Figure 5 for both QPSK and 16-QAM modulations. Using a powerful code, both diversity schemes have almost similar performances. We can observe that the performance of the BID scheme is still far from the corresponding MFB for 16-QAM modulation. Note that for high throughput requirements, bit-puncturing can be applied in order to increase the coding rate. For a higher coding rate, the performance gains of

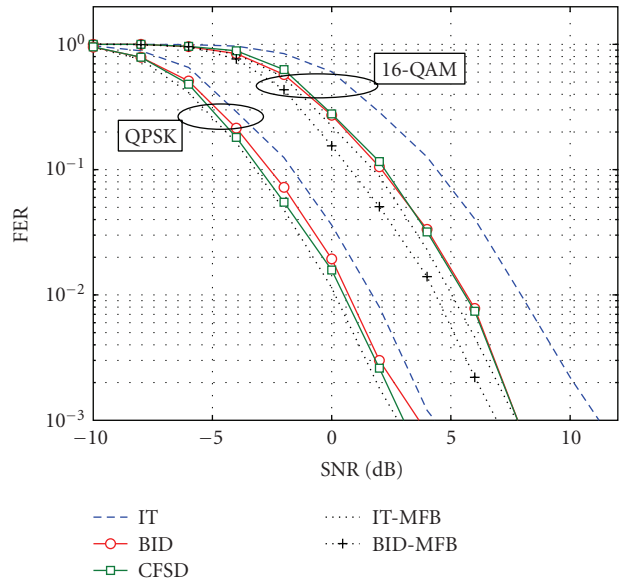


Figure 5: FER performance comparison between different transmission schemes for a coded system using a turbo-code for QPSK and 16-QAM modulations.

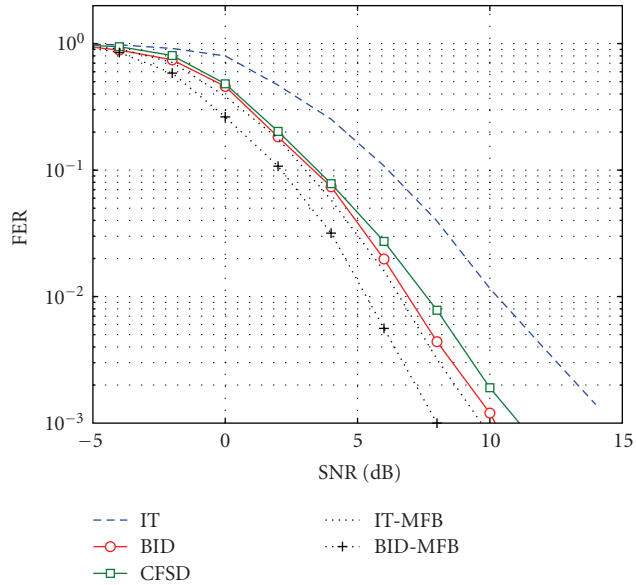


Figure 6: FER performance for different transmission schemes for a coded system with a rate-1/3 convolutional code using 16-QAM modulation and linear detection.

the proposed diversity schemes lay somewhere between the full rate case (rate 1/3) and the uncoded case. In order to close this gap, an iterative processing can be performed between the detector and the channel decoder. Due to the high complexity of the iterative processing using a turbocode, we use the LTE convolutional code of rate-1/3 whose generator polynomial is $(133, 171, 165)_8$. Here again, no trellis termination is performed for convolutional codes. Figure 6 shows the FER performance at the last HARQ round for separate detection and decoding, while Figure 7 shows the corresponding FER performance for a turbo-equalizer which performs a maximum of four turbo-iterations.

We note that for a linear receiver without turbo-iterations, the performance of both diversity schemes is almost the same. With a turbo-equalizer, the BID scheme outperforms the CFSD scheme unlike the noncoded system because the iterative receiver performs closely to the MFB which is better for the BID scheme.

In conclusion, we find that the CFSD is suitable for a linear receiver with separate equalization and decoding, especially for high rate channel coding. The BID scheme gives better performance with an iterative receiver at the expense of a higher system complexity.

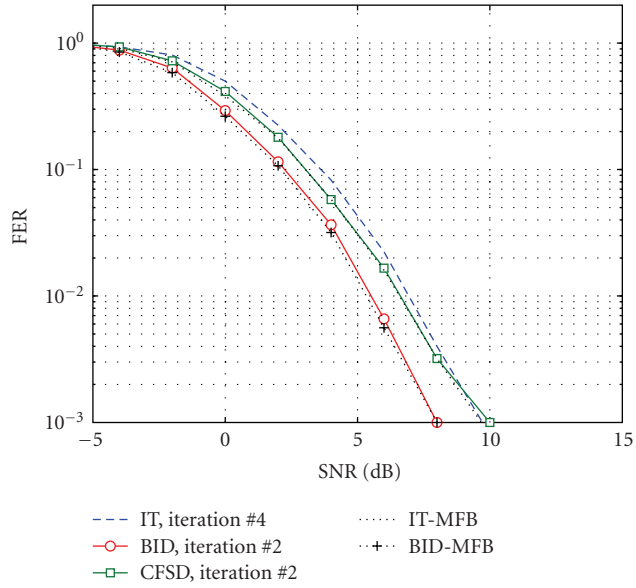


Figure 7: FER performance for different transmission schemes for a coded system with a rate-1/3 convolutional code using 16-QAM modulation and turbo-equalization.

6. Conclusions

We have presented and compared two transmit diversity schemes for multiple HARQ retransmission using single carrier signaling over frequency selective channels. Our theoretical analysis shows that the BID scheme has better performance limits than the CFSD scheme for high order modulation, but the CFSD scheme is more efficient in combating the ISI for channels with short memory. The CFSD is suitable for a linear receiver with separate equalization and decoding, while the BID scheme gives a better performance with an iterative receiver at the expense of a higher system complexity. These diversity schemes can be used in order to compensate for poor channel diversity in slow fading environment depending to the desired performance complexity tradeoff and the system parameters including the channel coding rate, the modulation order.

Appendix

Assuming uniform interleaving, the error symbols are considered as identically distributed but not independent due the constraint on the sum of their Hamming weights. However, any two error symbols are conditionally independent knowing their respective Hamming weights. The coded and interleaved packet contains NQ bits which are modulated to N symbols. The error packet contains d errors which are assumed uniformly distributed over the packet. The probability that a symbol e_n has a Hamming weight $d_H(e_n) = k$ is given by

$$\Pr(d_H(e_n) = k) = \frac{\binom{Q}{k}\binom{NQ-Q}{d-k}}{\binom{NQ}{d}}. \tag{A.1}$$

The average squared amplitude μ_e can be calculated as

$$\begin{aligned}\mu_e(d) &= \mathbb{E}\left[\|\mathbf{e}\|^2 \mid d\right] \\ &= N\sum_{k=1}^{Q} m_2(k)\Pr(d_H(e_n) = k) \\ &= N\binom{NQ}{d}^{-1}\sum_{k=1}^{Q}\binom{Q}{k}\binom{NQ-Q}{d-k}m_2(k),\end{aligned} \tag{A.2}$$

where $m_2(k) = \mathbb{E}[|e_n|^2 \mid k]$ for $k = 1, \ldots, Q$ is the conditional mean of $|e_n|^2$ giving its Hamming weight k.

The variance σ_e^2 can be similarly calculated as follows:

$$\sigma_e^2(d) = \mathbb{E}\left[\left(\|\mathbf{e}\|^2 - \mu_e\right)^2 \mid d\right] = \mathbb{E}\left[\|\mathbf{e}\|^4 \mid d\right] - \mu_e^2(d), \tag{A.3}$$

where

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{e}\|^4 \mid d\right] &= N\overline{m}_4(d) + N(N-1)\overline{\rho}_2(d), \\ \overline{m}_4(d) &= \mathbb{E}\left[|e_n|^4 \mid d\right] \\ &= \binom{NQ}{d}^{-1}\sum_{k=1}^{Q}\binom{Q}{k}\binom{NQ-Q}{d-k}m_4(k), \\ \overline{\rho}_2(d) &= \mathbb{E}\left[|e_{n_1}|^2|e_{n_2}|^2 \mid d\right] \\ &= \binom{NQ}{d}^{-1} \\ &\quad\times \sum_{\substack{k_1,k_2=1 \\ k_1+k_2\le d}}^{Q}\binom{Q}{k_1}\binom{Q}{k_2}\binom{NQ-2Q}{d-k_1-k_2}m_2(k_1)m_2(k_2),\end{aligned} \tag{A.4}$$

for $n_1 \neq n_2$, where $m_4(k) = \mathbb{E}[|e_n|^4 \mid k]$. The conditional moments m_2 and m_4 can be computed directly from the modulation and the mapping scheme.

Acknowledgment

This work was supported by the project "Urbanisme des Radiocommunications" of the Pôle de compétitivité SYSTEM@TIC.

References

[1] 3GPP Technical Specification Group Radio Access Network E-UTRA (Release 8), "LTE physical layer-general description," 3GPP TS 36.201 V8.3.0, March 2009, http://www.3gpp.org/ftp/Specs/html-info/36-series.htm.

[2] S. Lin, D. Costello Jr., and M. Miller, "Automatic-repeat-request error-control schemes," *IEEE Communications Magazine*, vol. 22, no. 12, pp. 5–17, 1984.

[3] C. Douillard, A. Picart, P. Didier, M. Jézéquel, C. Berrou, and A. Glavieux, "Iterative correction of intersymbol interference: turbo-equalization," *European Transactions on Telecommunications and Related Technologies*, vol. 6, no. 5, pp. 507–512, 1995.

[4] H. Harashima and H. Miyakawa, "Matched-transmission technique for channels with intersymbol interference," *IEEE Transactions on Communications*, vol. 20, no. 4, pp. 774–780, 1972.

[5] G. D. Forney Jr. and M. V. Eyuboglu, "Combined equalization and coding using precoding," *IEEE Communications Magazine*, vol. 29, no. 12, pp. 25–34, 1991.

[6] H. Samra, H. Sun, and Z. Ding, "Capacity and linear precoding for packet retransmissions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 3, pp. 541–544, 2005.

[7] A.-N. Assimi, C. Poulliat, I. Fijalkow, and D. Declercq, "Periodic Hadamard phase precoding for HARQ systems over intersymbol interference channels," in *Proceedings of the IEEE International Symposium on Spread Spectrum Techniques and Applications (ISSSTA '08)*, pp. 714–718, Bologna, Italy, 2008.

[8] D. N. Doan and K. R. Narayanan, "Iterative packet combining schemes for intersymbol interference channels," *IEEE Transactions on Communications*, vol. 50, no. 4, pp. 560–570, 2002.

[9] A.-N. Assimi, C. Poulliat, and I. Fijalkow, "Packet combining for turbo-diversity in HARQ systems with integrated turbo-equalization," in *Proceedings of the 5th International Symposium on Turbo Codes and Related Topics (TURBOCODING '08)*, pp. 61–66, Lausanne, Switzerland, 2008.

[10] H. Samra and Z. Ding, "Symbol mapping diversity in iterative decoding/demodulation of ARQ systems," in *Proceedings of IEEE International Conference on Communications (ICC '03)*, vol. 5, pp. 3585–3589, Anchorage, Alaska, USA, 2003.

[11] H. El Gamal, G. Caire, and M. O. Damen, "The MIMO ARQ channel: diversity-multiplexing-delay tradeoff," *IEEE Transactions on Information Theory*, vol. 52, no. 8, pp. 3601–3621, 2006.

[12] G. D. Forney Jr., "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Transactions on Information Theory*, vol. 18, no. 3, pp. 363–378, 1972.

[13] H. Samra, Z. Ding, and P. M. Hahn, "Optimal symbol mapping diversity for multiple packet transmissions," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 4, pp. 181–184, Hong Kong, 2003.

[14] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial concatenation of interleaved codes: performance analysis, design, and iterative decoding," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 909–926, 1998.

[15] P. Xia, S. Zhou, and G. B. Giannakis, "Achieving the Welch bound with difference sets," *IEEE Transactions on Information Theory*, vol. 51, no. 5, pp. 1900–1907, 2005.

[16] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Transactions on Information Theory*, vol. 20, no. 2, pp. 284–287, 1974.

[17] R. Visoz, A. O. Berthet, and S. Chtourou, "Frequency-domain block turbo-equalization for single-carrier transmission over MIMO broadband wireless channel," *IEEE Transactions on Communications*, vol. 54, no. 12, pp. 2144–2149, 2006.

[18] T. Ait-Idir, H. Chafnaji, and S. Saoudi, "Joint hybrid ARQ and iterative space-time equalization for coded transmission over the MIMO-ISI channel," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '08)*, pp. 622–627, Las Vegas, Nev, USA, March 2008.

[19] D. S. Baum, J. Hansen, and J. Salo, "An interim channel model for beyond-3G systems: extending the 3GPP spatial channel model (SCM)," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '05)*, vol. 5, pp. 3132–3136, Zurich, Switzerland, 2005.

[20] J. Salo, et al., "Matlab implementation of the 3GPP spatial channel model (3GPP TR 25.996)," 2005, http://www.ist-winner.org/3gpp_scm.html.

[21] 3GPP Technical Specification Group Radio Access Network E-UTRA (Release 8), "Base station (BS) radio transmission and reception," 3GPP TS 36.104 V8.5.0, March 2009, http://www.3gpp.org/ftp/Specs/html-info/36-series.htm.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 893751, 10 pages
doi:10.1155/2009/893751

Research Article

An Adaptive Channel Interpolator Based on Kalman Filter for LTE Uplink in High Doppler Spread Environments

Bahattin Karakaya,[1] Hüseyin Arslan,[2] and Hakan A. Çırpan[1]

[1] *Department of Electrical and Electronics Engineering, Istanbul University, Avcilar, 34320 Istanbul, Turkey*
[2] *Department of Electrical Engineering, University of South Florida, 4202 E. Fowler Avenue, ENB118, Tampa, FL 33620, USA*

Correspondence should be addressed to Bahattin Karakaya, bahattin@istanbul.edu.tr

Received 17 February 2009; Revised 5 June 2009; Accepted 27 July 2009

Recommended by Cornelius van Rensburg

Long-Term Evolution (LTE) systems will employ single carrier frequency division multiple access (SC-FDMA) for the uplink. Similar to the Orthogonal frequency-division multiple access (OFDMA) technology, SC-FDMA is sensitive to frequency offsets leading to intercarrier interference (ICI). In this paper, we propose a Kalman filter-based approach in order to mitigate ICI under high Doppler spread scenarios by tracking the variation of channel taps jointly in time domain for LTE uplink systems. Upon acquiring the estimates of channel taps from the Kalman tracker, we employ an interpolation algorithm based on polynomial fitting whose order is changed adaptively. The proposed method is evaluated under four different scenarios with different settings in order to reflect the impact of various critical parameters on the performance such as propagation environment, speed, and size of resource block (RB) assignments. Results are given along with discussions.

1. Introduction

3GPP Long-Term Evolution (LTE) aims at improving the Universal Mobile Telecommunication System (UMTS) mobile phone standard to cope with future requirements. The LTE project is not a standard itself, but it will result in the new evolved Release 8 of the UMTS standard, including most or all of the extensions and modifications of the UMTS system. Orthogonal frequency-division multiplexing (OFDM) is considered as the strongest candidate of the technology that will be deployed in LTE because of its advantages in lessening the severe effect of frequency selective fading. Since wideband channels experience frequency selectivity because of multipath effect single-carrier modulations necessitate the use of equalizers whose implementations are impractical due to their complexities. Therefore, OFDM is selected in order to overcome these drawbacks of single-carrier modulation techniques [1]. In OFDM, the entire signal bandwidth is divided into a number of narrower bands or orthogonal subcarriers, and signal is transmitted over those bands in parallel. This way, computationally complex intersymbol interference (ISI) equalization is avoided and channel estimation/equalization task becomes easier. However, orthogonal frequency-division multiple accessing (OFDMA) has a high peak-to-average power ratio (PAPR) because of very pronounced envelope fluctuations, which will decrease the power efficiency in user equipment (UE) and thus decrease the coverage efficiency in uplink for the low cost power amplifier (PA). Moreover, in the uplink, inevitable frequency offset error caused by different terminals that transmit simultaneously destroys the orthogonality of the transmissions leading to multiple access interference [2].

In the literature, various methods are proposed in order to alleviate the aforementioned problems and shortcomings. In order to keep the PAPR as low as possible, single carrier frequency-division multiple access (SC-FDMA) that combines single-carrier frequency-domain equalization (SC-FDE) system with FDMA scheme is introduced. SC-FDMA has many similarities to OFDMA in terms of throughput performance, spectral efficiency, immunity to multipath interference, and overall complexity. Furthermore, it can be regarded as discrete Fourier transform (DFT)—spread

OFDMA, where time domain data symbols are transformed into frequency-domain by a DFT before going through OFDMA modulation [2]. Therefore, air interface of Release 8 is being referred to as Evolved Universal Terrestrial Radio Access (E-UTRA) which is assumed to employ SC-FDMA for the uplink and OFDMA for the downlink [3].

To the best knowledge of authors, the very first papers addressing the channel estimation problem in the context of SC-FDMA are [4, 5] both of which consider time-invariant frequency-selective multipath channels, throughout an SC-FDMA symbol. In these papers, zeroforcing (ZF) or minimum mean squared error (MMSE) linear channel estimation methods have been proposed in frequency-domain although they all suffer from ICI, without proposing any cancellation method. Note that, since most of the next generation wireless network standards require transmission in high speed environments, time-variant frequency-selective multipath assumption should be considered rather than time-invariant frequency-selective multipath assumption. However, it is important to note that when the channel is time-variant, the subcarrier orthogonality is destroyed giving rise to ICI due to channel variation within an SC-FDMA symbol.

Even though they are not in SC-FDMA context, there are methods proposed in the literature dealing with ICI mitigation for OFDM-based systems [6–8]. In [6], receiver antenna diversity has been proposed; however, high normalized Doppler spread reduces the efficiency of this approach. In [7], a piece-wise linear approximation is proposed based on a comb-type pilot subcarrier allocation scheme in order to track the time-variations of the channel. In [8] Modified Kalman filter- (MKF-) based time-domain channel estimation approach for OFDM with fast fading channels has been investigated. The proposed receiver structure models the time-varying channel as an AR-process; tracks the channel with MKF; performs curve fitting, extrapolation and MMSE time domain equalizer. In [9], matched filter, LS and MMSE estimator that incorporate decision feedback low complexity time-domain channel estimation and detection techniques are presented for multicarrier signals in a fast and frequency-selective Rayleigh fading channel for OFDM systems. Moreover, polynomial interpolation approaches have been commonly used for channel estimation [10].

In this paper, we focus on a major challenge, namely, the SC-FDMA transmission over time-varying multipath fading channels in very high speed environments, which is regarded as one of the most difficult problems in 3GPP systems. Inspired by the conclusions in [6–9], the signal model in [9] is extended to SC-FDMA systems. A channel estimation algorithm based on Kalman filter and a polynomial curve fitting interpolator whose order is selected adaptively is proposed for LTE uplink systems which include time-varying channels in high speed environments. The variations of channel taps are tracked jointly by Kalman filter in time domain during training symbols. Since channel tap information is missing between the training symbols of two consecutive slots within a single subframe, an interpolation operation is performed to recover it. Hence, the interpolation is established by using a polynomial curve fitting that is based on linear model estimator. The contributions of this study are twofold. (i) The factors which affect the selection of the order of the polynomial curve fitting interpolator are identified; (ii) A procedure that is based on mean squared error (MSE) is developed in order to determine the optimum polynomial order values.

The remainder of the paper is organized as follows. Section 2 outlines the characteristics of the channel model considered along with a discussion that is related to sample-spaced and fractional-spaced channel impulse response concerns. In Section 3, LTE uplink system model is introduced and subcarrier mapping is discussed. In addition, the impact of ICI is formally described for SC-FDMA system. Section 4 provides the details of frequency-domain least squares channel estimation, Kalman filter tracking, and polynomial curve fitting interpolation along with the discussion regarding the selection of its order. Section 5 introduces simulation setups for various scenarios and presents corresponding performance results. Finally, in Section 6, concluding remarks are given along with possible future research directions.

2. Channel Model

The complex baseband representation of a wireless mobile time-variant channel impulse response (CIR) can be described by

$$h'(t,\tau) = \sum_i \alpha_i(t)\delta(t-\tau_i), \tag{1}$$

where $\alpha_i(t)$ is the time-variant complex tap coefficients of the ith path, and τ_i is the corresponding path delay. The fading channel coefficients $\alpha_i(t)$ are modeled as zero mean complex Gaussian random variables. Based on the Wide Sense Stationary Uncorrelated Scattering (WSSUS) assumption, the fading channel coefficients in different delay taps are statistically independent. In time domain, fading coefficients are correlated and have Doppler power spectrum density modeled as in [11] with the following autocorrelation function:

$$E\{\alpha_i(t_1)\alpha_i^*(t_2)\} = \sigma_{\alpha_i}^2 J_0\left(2\pi f_d T_s(t_2-t_1)\right), \tag{2}$$

where $\sigma_{\alpha_i}^2 = E\{|\alpha_i(t)|^2\}$ denotes the average power of the ith path channel coefficient, f_d is the maximum Doppler frequency in Hertz, and $(\cdot)^*$ represents the complex conjugate operation. The term $f_d T_s$ represents the normalized Doppler frequency; T_s is the sampling period. $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind.

Considering the effect of transmitter-receiver pair in a more generalized way, (1) can be written as follows [12]:

$$h(t,\tau) = h'(t,\tau) * c(\tau) = \sum_i \alpha_i(t)c(t-\tau_i), \tag{3}$$

where $*$ denotes convolution operation, and $c(\tau)$ is the aggregate impulse response of the transmitter-receiver pair,

which corresponds to the Nyquist filter. Continuous channel transfer function (CTF) can be obtained from (3) as follows:

$$\begin{aligned} H(t,f) &= \int_{-\infty}^{\infty} h(t,\tau)e^{-j2\pi f\tau}d\tau \\ &= C(f)\sum_{l}\alpha_l(t)e^{-j2\pi f\tau_l}, \end{aligned} \tag{4}$$

where $C(f)$ is the Fourier transform of impulse response, $c(\tau)$, of the transceiver pair. For LTE Uplink system of interest, which uses a sufficiently long cyclic prefix (CP) and adequate synchronization, the discrete subcarrier-related CTF can be expressed as

$$\begin{aligned} H[m,n] &\triangleq H(mT_s, n\Delta f) \\ &= C(n\Delta f)\sum_{i=0}^{L'-1}\alpha_i(mT_s)\exp\left(-\frac{j2\pi n\tau_i}{M}\right) \\ &= \sum_{l=0}^{L-1} h[m,l]e^{-j2\pi nl/M}, \end{aligned} \tag{5}$$

where

$$h[m,l] \triangleq h(mT_s, lT_s) = \sum_{i=0}^{L'-1}\alpha_i(mT_s)c(lT_s-\tau_i) \tag{6}$$

is the CIR which has sample-spaced delays at lT_s time instant. M denotes the number of SC-FDMA subcarriers, T_s denotes the base-band signal's sample duration, L' and L denote the number of fractionally-spaced channel paths and the number of equivalent sample spaced CIR taps, respectively. Note that because of the convolution with impulse response of the system, sample-spaced CIR (SS-CIR) has correlated nonzero taps compared to fractionally spaced CIR (FS-CIR). Due to the band limited property of the physical systems, SS-CIR cannot be implemented with limited number of components. One of the solutions to this problem is to truncate SS-CIR in such a way that most of its energy is preserved in the truncated part. In this study, truncation strategy is adopted in simulations. However, for the sake of completeness, in Figure 1, the impact of truncation strategy is illustrated for 3GPP rural area channel model for a bandwidth of 10 MHz. All of the steps prior to truncation operation, which are given in (1), (3), and (6), respectively, are given in this figure with appropriate labels.

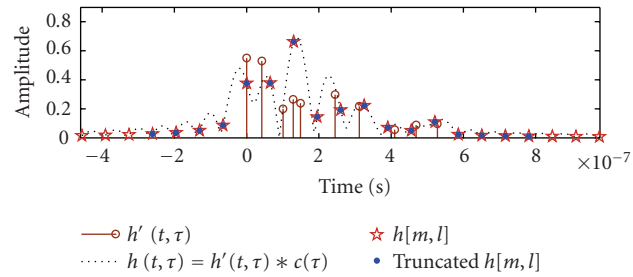


Figure 1: 3GPP rural area channel model for a bandwidth of 10 MHz. Note that all of the steps prior to truncation operation are illusturated with appropriate labels corresponding to (1), (3), and (6), respectively.

3. System Model

Figure 2 shows the discrete baseband equivalent system model. We assume an N-point DFT for spreading the pth users time domain signal $d[k]$ into frequency-domain:

$$D^{(p)}[\kappa] = \sum_{k=0}^{N-1} d^{(p)}[k]e^{-j2\pi k\kappa/N}. \tag{7}$$

After spreading, $D^{(p)}[\kappa]$ is mapped onto the nth subcarrier $S^{(p)}[n]$ as follows:

$$S^{(p)}[n] = \begin{cases} D^{(p)}[\kappa], & n \in \Gamma_N^{(p)}[\kappa], \\ 0, & n \in (\Phi - \Gamma_N[\kappa]), \end{cases} \tag{8}$$

where $\Gamma_N^{(p)}[\kappa]$ denotes N-element mapping set of pth user, Φ is a set of indices whose elements are $\{0,\ldots,M-1\}$ with $M > N$. The fundamental unit of spectrum for LTE uplink is a single subcarrier. A Resource block (RB) is composed of 12 adjacent subcarriers and forms the fundamental unit of resources to be assigned a single user as illustrated in Figure 3. Assigning adjacent RBs to a single user is called localized mapping which is the current working assumption in LTE [13]. Alternatively, if RBs are assigned apart, then, it is called distributed mapping, which is generally employed for frequency diversity [3] and possible candidate for LTE Advanced.

The transmitted single carrier signal at sample time m is given by

$$s^{(p)}[m] = \frac{1}{M}\sum_{n=0}^{M-1} S^{(p)}[n]e^{j2\pi mn/M}. \tag{9}$$

The received signal at base station can be expressed as

$$y[m] = \sum_{p=0}^{P-1}\sum_{l=0}^{L-1} h^{(p)}[m,l]s^{(p)}[m-l] + w[m], \tag{10}$$

where $h^{(p)}[m,l]$ is the sample spaced channel response of the lth path during the time sample m of pth user, L is the total number of paths of the frequency selective fading channel, and $w[m]$ is the additive white Gaussian noise (AWGN) with $\mathcal{N}(0,\sigma_w^2)$.

In this paper, we assume that there is only one user, $P = 1$, therefore (10) becomes

$$y[m] = \sum_{l=0}^{L-1} h[m,l]s[m-l] + w[m]. \tag{11}$$

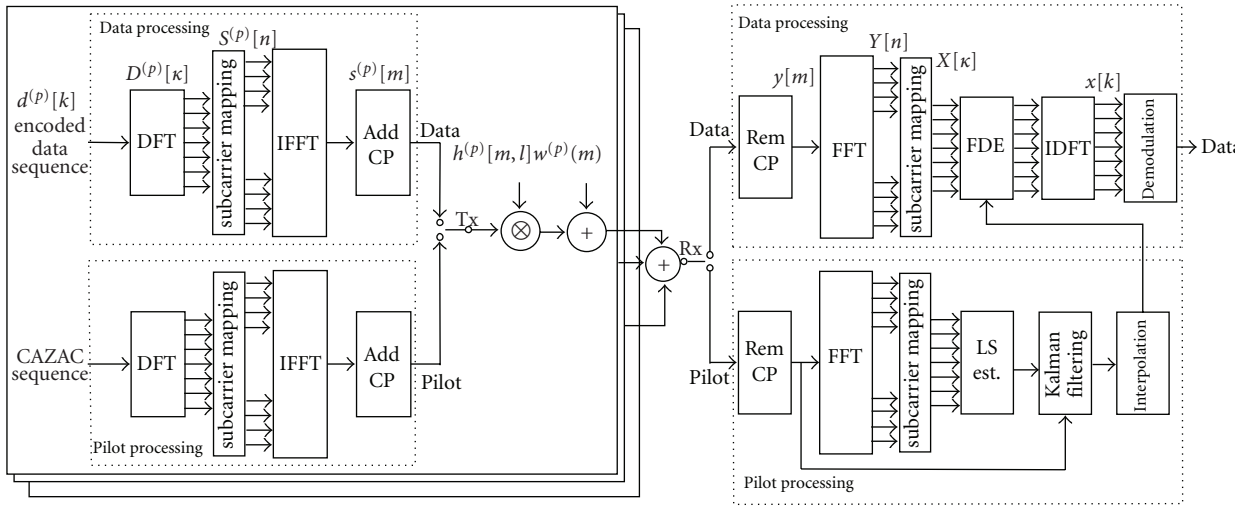


Figure 2: SC-FDMA transceiver system model.

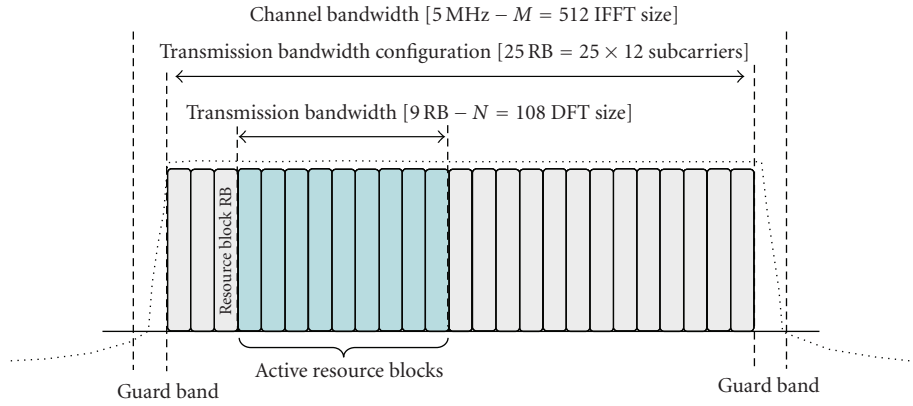


Figure 3: An example subcarrier mapping for a specific scenario.

By plugging (9) into (11), the received signal can be rewritten as follows:

$$y[m] = \frac{1}{M}\sum_{n=0}^{M-1} S[n]\sum_{l=0}^{L-1} h[m,l]e^{j(2\pi n(m-l)/M)} + w[m]. \quad (12)$$

When (5) is placed into (12), it yields:

$$y[m] = \frac{1}{M}\sum_{n=0}^{M-1} S[n]H[m,n]e^{j2\pi mn/M} + w[m]. \quad (13)$$

Thus FFT output at nth subcarrier can be expressed in the following form:

$$\begin{aligned} Y[n] &= \sum_{m=0}^{M-1} y[m]e^{-j2\pi mn/M} \\ &= S[n]H[n] + I[n] + W[n], \end{aligned} \quad (14)$$

where $H[n]$ represents frequency-domain channel response expressed as

$$H[n] = \frac{1}{M}\sum_{m=0}^{M-1} H[m,n], \quad (15)$$

and $I[n]$ is ICI caused by the time-varying nature of the channel given as

$$I[n] = \frac{1}{M}\sum_{i=0,i\neq n}^{M-1} S[i]\sum_{m=0}^{M-1} H[m,i]e^{j2\pi m(i-n)/M}, \quad (16)$$

and $W[n]$ represents Fourier transform of noise vector $w[m]$ as follows:

$$W[n] = \sum_{m=0}^{M-1} w[m]e^{-j2\pi mn/M}. \quad (17)$$

Because of the $I[n]$ term, there is an irreducible error floor even in the training sequences since pilot symbols are also corrupted by ICI. Time-varying channel destroys the orthogonality between subcarriers. Therefore, channel estimation should be performed before the FFT block. In order to compensate for the ICI, a high quality estimate of the CIR is required in the receiver. In this paper, the proposed channel estimation is performed in time domain, where time-varying-channel coefficients are tracked by Kalman filter within the training intervals. Variation of channel taps

during the data symbols between two consecutive pilots is obtained by interpolation.

We assume that equalization is performed in frequency-domain after the subcarrier demapping block. Data are obtained after the demapping described as

$$\begin{aligned} X[\kappa] &= Y[n], \quad \text{where } n \in \Gamma_N[\kappa] \\ &= D[\kappa]H[n] + I[n] + W[n]. \end{aligned} \tag{18}$$

4. Channel Estimation

4.1. Frequency-Domain Least Squares Estimation. In this study, frequency-domain least squares channel estimation is employed in order to find the initial values required by Kalman filter. Channel frequency response, which corresponds to used subcarriers, can be found by the following equation:

$$\hat{H}_0[n] = \begin{cases} \dfrac{X[\kappa]D_t^*[\kappa]}{|D_t[\kappa]|^2}, & n \in \Gamma_N[\kappa], \\ 0, & n \in (\Phi - \Gamma_N[\kappa]), \end{cases} \tag{19}$$

where $(\cdot)_0$ denotes the initial value, and $D_t[\kappa]$ is a training sequence known by the receiver. If (5) and (15) are considered together, yielding time average of time-varying frequency response over one SC-FDMA symbol is

$$\begin{aligned} H[n] &= \sum_{l=0}^{L-1} \frac{1}{M} \sum_{m=0}^{M-1} h[m,l] e^{-j2\pi nl/M} \\ &= \sum_{l=0}^{L-1} h[l] e^{-j2\pi nl/M}, \end{aligned} \tag{20}$$

where $h[l]$ is the time average of time-varying impulse response over one SC-FDMA symbol:

$$h[l] = \frac{1}{M} \sum_{m=0}^{M-1} h[m,l]. \tag{21}$$

It can be easily observed that in (22) and (20) the DFT pair will result in corresponding channel representations both in time and frequency-domains, respectively,

$$h[l] = \sum_{n=0}^{M-1} H[n] e^{-j2\pi nl/M}. \tag{22}$$

Hence, in order to initial values for Kalman filtering in time domain, we can write M-point IFFT of $\hat{H}_0[n]$ as

$$\hat{h}_0[l] = \frac{1}{M} \sum_{n \in \Gamma_N[\kappa]} \hat{H}_0[n] e^{j2\pi nl/M}, \quad l = 0, \ldots, M-1. \tag{23}$$

Recall that in (19) some of the subcarriers are left unused for a given user. It is also known that transform-domain techniques introduce CIR path leaks due to the suppression of unused subcarriers [14]. Besides, Kalman filter needs time-domain samples in order to initiate the tracking procedure. However, due to the aforementioned leakage problem, unused subcarriers for a given user will create inaccurate time-domain value. In the literature, the problem has been studied for a single user OFDM system in [15–17]. As mentioned before, leakage problem just affects the initialization of the algorithm therefore we do not focus on the leakage problem and in the subsequent subsection Kalman filtering is introduced along with this inherent leakage problem. By using sophisticated solutions for the leakage problem, initialization of the Kalman can also be improved.

4.2. Kalman Filtering. It was shown that time selective fading channel can be sufficiently approximated by using first-order autoregressive (AR) model. Time-varying channel taps can be modeled through the use of a first-order AR process in the vector form as follows [18, 19]:

$$\mathbf{h}[m+1] = \beta \mathbf{h}[m] + \mathbf{v}[m+1], \tag{24}$$

where $\mathbf{h}[m] = [h[m,0], \ldots, h[m,L-1]]$, which is also called process equation in Kalman filtering [20]. $\mathbf{v}[m]$ and $\beta \mathbf{I}_L$ are called process noise and state transition matrix, respectively. The correlation matrix of the process noise and the state transition matrix can be obtained through the Yule-Walker equation [21]

$$\begin{aligned} \mathbf{Q}[m] &= (1-\beta^2)\,\mathrm{diag}\left(\boldsymbol{\sigma}^2_{\mathbf{h}[m]}\right) \\ \beta &= J_0(2\pi f_d T_s), \end{aligned} \tag{25}$$

where $\boldsymbol{\sigma}^2_{\mathbf{h}[m]} = [\sigma^2_{h[m,0]}, \sigma^2_{h[m,1]}, \ldots, \sigma^2_{h[m,L-1]}]$ is the power delay profile of the channel. The equivalent of (11), which is a measurement equation in the state-space model of Kalman filter, can be shown in vector form as

$$y[m] = \mathbf{s}^T[m]\mathbf{h}[m] + w[m], \tag{26}$$

where $\mathbf{s}[m] = [s[m], s[m-1], \ldots, s[m-L+1]]^T$. The channel estimate $\hat{\mathbf{h}}[m+1]$ can be obtained by a set of recursions

$$\begin{aligned} e[m] &= y[m] - \hat{y}[m] = y[m] - \mathbf{s}^T[m]\hat{\mathbf{h}}[m], \\ \mathbf{K}[m] &= \beta \mathbf{P}[m]\mathbf{s}^*[m]\left(\sigma_w^2 + \mathbf{s}^T[m]\mathbf{P}[m]\mathbf{s}^*[m]\right)^{-1}, \end{aligned} \tag{27}$$

where $\mathbf{P}[m] = E\{(\mathbf{h}[m] - \hat{\mathbf{h}}[m])(\mathbf{h}[m] - \hat{\mathbf{h}}[m])^H\}$. The updating rule of recursion is as follows:

$$\begin{aligned} \hat{\mathbf{h}}[m+1] &= \beta \hat{\mathbf{h}}[m] + \mathbf{K}[m]e[m], \\ \mathbf{P}[m+1] &= \beta\left(\beta \mathbf{I} - \mathbf{K}[m]\mathbf{s}^T[m]\right)\mathbf{P}[m] + \mathbf{Q}[m+1]. \end{aligned} \tag{28}$$

4.3. Polynomial Curve Fitting Based on Linear Model Estimator and Order Selection. When the frame structure in Figure 4 is considered, one can easily notice that the channel tap information is missing in between the training symbols of two consecutive slots within a single subframe. The purpose of interpolation is to recover this missing information in between by employing a polynomial curve fitting based on

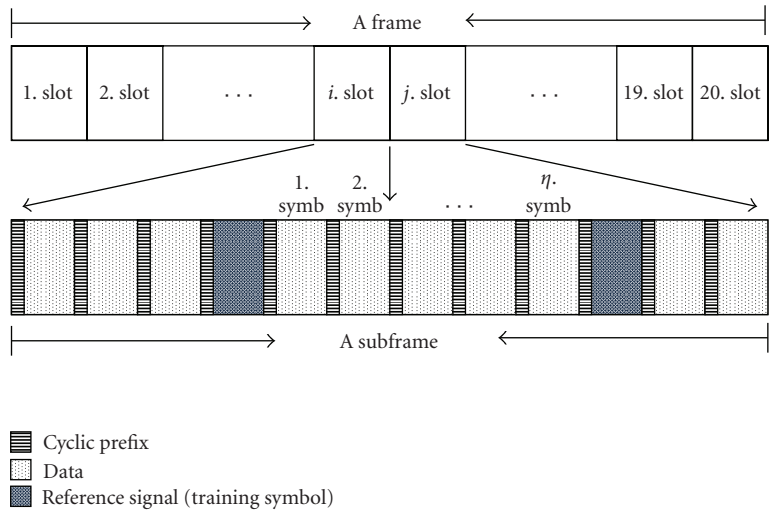


Figure 4: An LTE uplink type 1 frame structure with extended CP. In one slot there are six symbols for extended CP case whereas there are seven symbols for normal CP case [3].

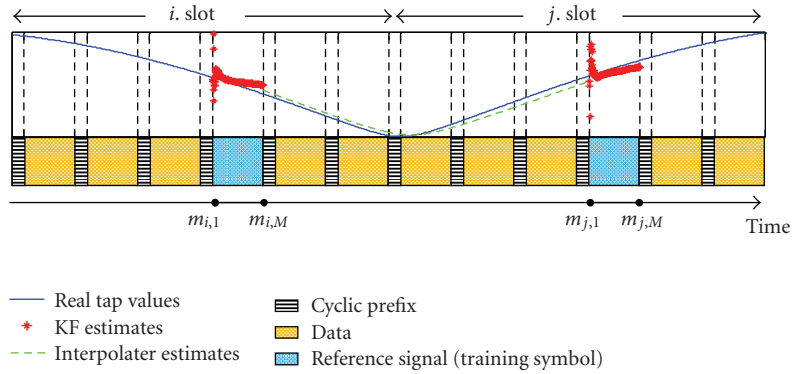


Figure 5: Kalman tracking and polynomial curve fitting procedure applied in consecutive slots with type 1 frame structure and extended CP size.

linear model estimator. Note that, in this study, it is assumed that within one training symbol duration the channel is time-variant. Kalman filter is employed in order to keep track of the changes within a single training symbol; therefore, these estimates are mandatory for interpolating the values in between because the channel might vary significantly from one training symbol to the next one. Curve fitting is established by estimating the coefficients of the polynomial of interest. In order to estimate the coefficients, in this study, the linear model estimator is applied to the channel tap estimates generated by Kalman tracker within training symbols; see Figure 5. The linear model considered here can be expressed in the following form [22]:

$$\Xi[l] = \Sigma \cdot \Theta[l], \tag{29}$$

where $\Xi[l] = [\hat{h}[m_{i,1}, l], \ldots, \hat{h}[m_{i,M}, l], \hat{h}[m_{j,1}, l], \ldots, \hat{h}[m_{j,M}, l]]^T$ is a $2M \times 1$ vector of observations supplied by lth path Kalman filter channel estimates, and $m_{i,a}$ and $m_{j,a}$ $a = 1, \ldots, M$ are time instants of training symbols. $\Sigma = [\mathbf{V}_i^T \mathbf{V}_j^T]^T$ is a known $2M \times \nu$ matrix which is constructed with two Vandermonde matrices $V_i(k, \mu) = m_{i,k}^{\mu-1}$, $V_j(k, \mu) = m_{j,k}^{\mu-1}$, $k = 1, \ldots, M$ and $\mu = 1, \ldots, \nu$. $\Theta[l] = [\theta_1[l], \ldots, \theta_\nu[l]]^T$ is a $\nu \times 1$ vector of polynomial coefficients to be estimated and ν is the order of the polynomial. In order to obtain the estimates, classical least-squares approach is employed as follows:

$$\hat{\Theta}[l] = \left(\Sigma^T \Sigma\right)^{-1} \Sigma^T \Xi[l]. \tag{30}$$

Based on the general description of the linear model and its estimator given in (29) and (30), respectively, the channel taps that are estimated with the aid of interpolation operation are given by

$$\overline{h}[m, l] = \sum_{\mu=1}^{\nu} \hat{\theta}_\mu[l] m^{\mu-1}, \quad m_{i,M} < m < m_{j,1}. \tag{31}$$

Up until this point, a general sketch of the linear model estimator is outlined. However, the most important parameter of the procedure defined (29) through (31), which is the order of the polynomial, has not been introduced yet.

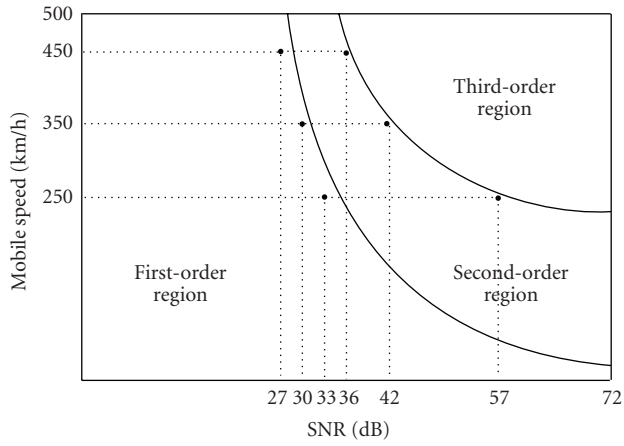


Figure 6: An example of polynomial curve fitting order selection chart based on SNR-mobile speed pair. This chart is calculated through the use of numerical methods for 3 MHz of bandwidth with fully assigned RBs to a single user in a rural area.

Selection of the order of the polynomial depends on many factors such as distance between training symbols in time, maximum Doppler shift, SNR, propagation environment including number of multipath components and delay spread, and so on. In other words, all of the parameters that affect the performance of the tracker and some of the structural factors (e.g., training symbol placements) have an influence on the order of the polynomial. In this study, to decide on the order of the polynomial, mean squared error (MSE) is selected to be the performance metric in the following manner:

$$\text{MSE} = \frac{1}{L}\sum_{l}\frac{1}{M}\sum_{m}\left|h[m,l]-\overline{h}[m,l]\right|^2. \tag{32}$$

Because the proposed method requires the order of the polynomial as an input, a special scenario in which Doppler, SNR, and propagation environment are taken into account while neglecting the impact of the rest of the aforementioned factors is investigated. The order information is obtained via steps (29) through (32) in a recursive fashion and recursion is terminated when the MSE reaches its minimum for a specific case. Figure 6 plots an instance of the output of this procedure which solely focuses on mobile speed-SNR pair. It is seen in this figure that low SNR values actually prevent the selection of higher orders due to the deteriorated tracker performance. However in realistic scenarios, channel parameters are not known exactly, prior knowledge on channel and its statistics can be used to form look-up table which contains optimum order values for various scenarios.

We now summarize the proposed method for LTE uplink systems.

Step 1. Initialization. Frequency-domain LS estimation to obtain initial tracking parameters for Kalman filter.

Step 2. Tracking. Jointly track CIR taps with Kalman filter employing training sybols.

Table 1: 3GPP channel models which are used in simulations.

Channel model	FS-CIR	1.4 MHz SS-CIR	3 MHz SS-CIR	5 MHz SS-CIR
TUx	20	13	17	25
RAx	10	10	11	13

Table 2: LTE uplink simulation parameters.

Parameters	1.4 MHz	3 MHz	5 MHz
Sampling frequency, f_s	1.92 MHz	3.84 MHz	7.68 MHz
FFT size, M	128	256	512
Maximum available subcarriers, N	72 (6 RB)	180 (15 RB)	300 (25 RB)
Extended CP	32	64	128

Step 3. Order decision. Decide the order of the polynomial from the look-up table (i.e., Figure 6).

Step 4. Coefficient Estimation. Compute the polynomial coefficients by applying least-squares approach (30) to the linear model (29) of Kalman estimates and Vandermonde matrix of corresponding time instants.

Step 5. Curve Fitting. Estimate the CIR taps from data symbols by using polynomial coefficients.

5. Simulation Results

In this section, computer simulation results are presented in order to evaluate the performance of the proposed channel estimation technique for LTE uplink systems. In simulations, the channel models given in [23] are used. Only typical urban (TUx) and rural area (RAx) models are taken into account. In addition to the default speed values, higher speed values are also considered in simulations. It is important to state one more time that there is a discrepancy between the number of channel taps given in [23] and simulated ones due to the reasons explained in Section 2. A comparison of these discrepancies with respect to different settings can be found in Table 1 by using the FS-CIR and SS-CIR notions.

A QPSK modulation format is employed. We consider type 1 frame structure, constant amplitude zero autocorrelation (CAZAC) pilot sequences, and extended CP size for LTE uplink [13]. As shown in Figure 4, frames have 20 slots, and each slot has six symbols. Fourth symbol in each slot is a pilot symbol, and the rest is data symbols. Critical parameters of simulation environments are given in Table 2. In each simulation loop, one frame (100 data symbols) is transmitted. In what follows, simulation scenarios are presented sequentially in detail.

Scenario 1. In this scenario, bandwidth is 1.4 MHz, all resource blocks are assigned to one user, and the channel environment is rural area so there are 10 taps to track. Two speed values are considered, namely, 60 Km/h and 120 Km/h, for UE. Simulation is run 500 times in order to obtain reliable

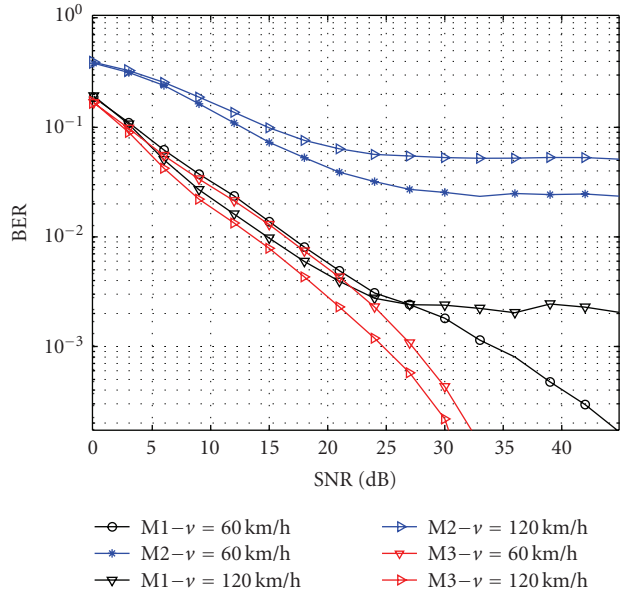


Figure 7: BER performance comparisons of methods for scenario 1. M1: the proposed method which is LS estimate is obtained from the pilots for the CFR used with the Kalman filter and then linear interpolation is used for symbols in between. M2: frequency-domain LS is used. M3: perfect channel state information is used.

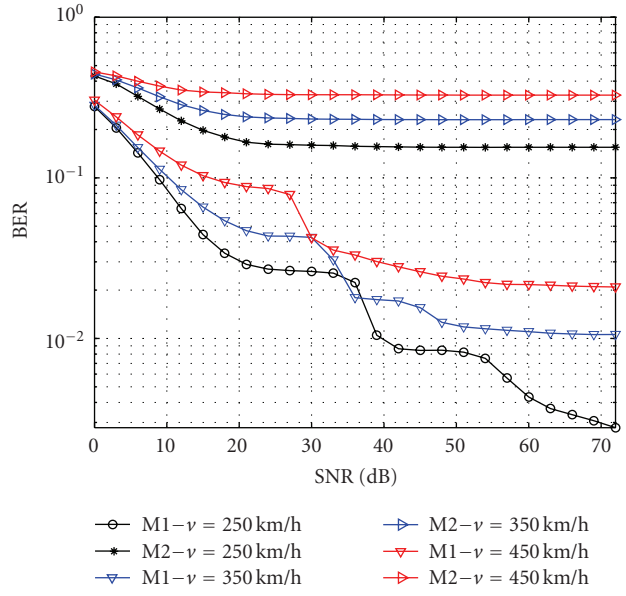


Figure 8: BER performance comparisons of different methods with respect to the method proposed which employs polynomial curve fitting whose orders are selected adaptively in senario 2 for different mobile speed values. Note that the performance of the method proposed exhibits a staircase-like behavior over the SNR values that correspond to the order shifts which can also be cross-checked with the points given in Figure 6. M1: the proposed method which is LS estimate is obtained from the pilots for the CFR used with the Kalman filter and then linear interpolation is used for symbols in between. M2: frequency-domain LS is used.

statistics. The results are plotted in Figure 7. The proposed method (Method 1—M1) is compared with two methods. In the first method, perfect CSI (Method 3—M3) is fed into the equalization process, whereas in the second one, which is outlined in Section 4.1, LS estimates (Method 2—M2) of CSI are used. It is worth mentioning that in the M2 the same channel frequency response (CFR) estimates are used until the next reference (training) symbol. As expected, M3 case provides the best performance among all. On the other hand, M2 performs the worst among all of the methods considered in this scenario, since it neither keeps track of the channel during data symbols nor takes the channel variation into account during the training sequence. Furthermore, during the training sequence, it just calculates the average CSI which is already contaminated by noise. Note that the performance of M1 is placed in between these two cases while its performance converges that of M3 case for low SNR values, whereas diverging it diverges for high SNR values. This is not surprising, because high SNR values allow one to observe the irreducible ICI error floor due to time-varying channel. Also note that for M3 case faster speed corresponds to better performance because when a proper detection technique is adopted, the time-varying nature of the channel can be exploited as a provider of time diversity [9].

Scenario 2. In this scenario, the impact of adaptive selection of the order of polynomial curve fitting on the performance of the method proposed is investigated with the following settings. Transmission bandwidth is 3 MHz, all resource blocks are assigned to one user, and the channel environment is rural area so there are 11 taps to track and the mobile speeds are 250 Km/h, 350 Km/h, and 450 Km/h. The proposed method (M1) and LS estimates (M2) which is aforementioned in Scenario 1 are compared to each other with respect to their bit error rate (BER) performances in Figure 8. It is worth noting that the performance of the proposed method improves by experiencing a staircase-like effect. This stems from changing the order of the polynomial curve fitting adaptively based on the results presented in Figure 6. In addition to comparative analysis, the MSE performance of the method proposed is also investigated in Figure 9. In conjunction with BER performances, as can be seen in both Figures 8 and 9, drastic drops in the performance curves occur in parallel to the corresponding mobile speed-SNR pairs given in Figure 6. It is very important to state that, the results presented in Figure 6 are peculiar to the setup considered here and calculated through the use of numerical methods, since its analysis is out of the scope of this study.

Scenario 3. Another important aspect of the problem considered here is to examine how the behavior of Kalman filter is affected by the accuracy of the initial value of channel taps. As discussed in Section 4.1, the structure of frequency spectrum of OFDM-based multicarrier systems causes a phenomenon called leakage problem [14] in transform domain methods. In the method proposed, leakage problem combined with LS estimation in frequency-domain leads to inaccurate initial value of channel taps to be fed into

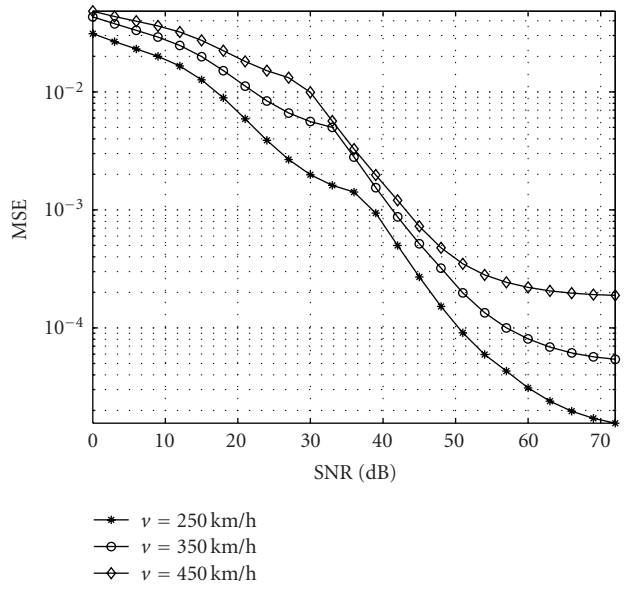


Figure 9: MSE performances of the method proposed which employs polynomial curve fitting whose orders are selected adaptively in Senario 2 for different mobile speed values. Note that the performance of the method proposed exhibits a staircase-like behavior over the SNR values that correspond to the order shifts which can also be cross-checked with the points given in Figure 6.

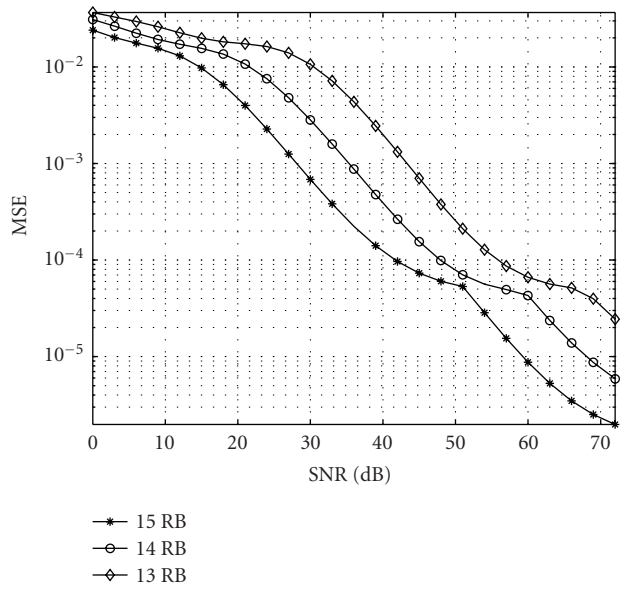


Figure 10: MSE performance comparisons for different resource block assignments to a single user in Scenario 3. Note that a decrease in number of assigned resource block worsens the performance stemming from the leakage problem.

Kalman tracker. In order to see how this leakage problem influences the MSE performance of the method proposed, another simulation setup is constructed with the following parameters. Transmission bandwidth is 3 MHz; different numbers of RBs are assigned one user each time in a typical rural area environment in which there are 11 taps

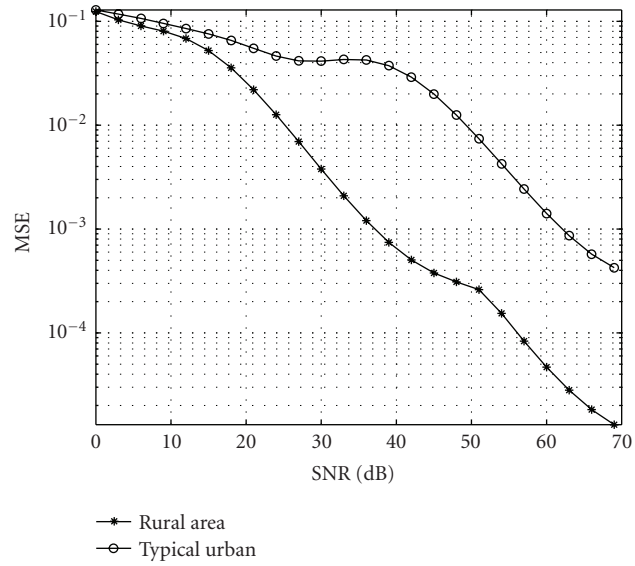


Figure 11: MSE performance comparisons for different propagation channel environments in Scenario 4.

to track for a fixed mobile speed of 120 Km/h. The results are given in Figure 10. In this figure, it is clearly observed that assigning less number of RBs gives rise to poorer performances compared to those of which are assigned more RBs. This stems from the fact that less number of RBs causes more leakage yielding worse accuracy in the initial values of channel taps in time domain.

Scenario 4. Finally, the overall impact of propagation environment is also investigated through the simulations. Two different setups, namely, rural and typical urban area environments, are considered with the following common parameters. Transmission bandwidth is 3 MHz, all RBs are assigned to one user, and the mobile speed is 120 Km/h. The results are plotted in Figure 11. It is clear that the performance is significantly dropped in a typical urban area compared to that in rural area because the number of channel taps in a typical urban is greater than that in rural area, as specified in Table 1. Since Kalman filter strives to track the taps jointly in time, having a larger number of channel taps yields worse performance, as expected.

6. Concluding Remarks and Future Directions

Future wireless communication systems such as LTE aim at very high data rates for high mobility scenarios. Since many of these systems have an OFDM-based physical layer, they are very sensitive to ICI. In this study, a channel estimation method is proposed for OFDM-based wireless systems that transmit only block-type pilots (training symbols). In the method proposed, Kalman filter is employed to obtain channel estimates during the training symbols. Next, polynomial curve fitting whose order is adjusted adaptively is applied in order to recover the time-variation of channel taps between training symbols within two consecutive slots in a single subframe. Results show that selecting the order

of the polynomial adaptively improves the BER performance significantly. However, as in most of the OFDM-based systems, the method proposed suffers from transform domain techniques as well, since they introduce CIR path leaks due to the suppression of unused subcarriers [14].

This study also reveals that selection of the order of the polynomial used in interpolation depends on many factors such as distance between training symbols in time, maximum Doppler shift, SNR, propagation environment including number of multipath components and delay spread, and so on. However, to the best knowledge of authors, there is no closed-form expression that takes all of the aforementioned factors into account and determines the optimum order value for the interpolation polynomial. In case deriving a closed-form expression is impossible or intractable, generating look-up tables which contain the optimum order values for various scenarios is essential.

The performance of the proposed approach directly related to Kalman filter performance. Specifically for more than one user case Kalman performance will be effected by initialization and the number of parameters to be tracked. Since unused subcarriers increase additional channel impulse response path leakage will degrade the performance of the initialization resulting in overall performance degradation in the proposed approach.

Acknowledgments

The authors would like to thank WCSP group members at USF for their insightful comments and helpful discussions. The authors would like to acknowledge the use of the services provided by Research Computing, University of South Florida. This work is supported in part by the Turkish Scientific and Technical Research Institute (TUBITAK) under Grant no. 108E054 and Research Fund of the Istanbul University under Projects UDP-2042/23012008, T-880/02062006. Part of the results of this paper is presented at the IEEE-WCNC, USA, March 31-April 3, 2008.

References

[1] R. van Nee and R. Prasad, *OFDM for Wireless Multimedia Communications*, Artech House, Norwood, Mass, USA, 2000.

[2] H. G. Myung, J. Lim, and D. J. Goodman, "Single carrier FDMA for uplink wireless transmission," *IEEE Vehicular Technology Magazine*, vol. 1, no. 3, pp. 30–38, 2006.

[3] 3GPP, TR 25.814, "Physical Layer Aspects for Evolved UTRA," http://www.3gpp.org/.

[4] D. Grieco, K. Pan, R. Olesen, and N. Shah, "Uplink single-user MIMO for 3GPP LTE," in *Proceedings of the 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC '07)*, pp. 1–5, Athens, Greece, September 2007.

[5] D.-H. Lee, S.-B. Im, and H.-J. Choi, "A novel pilot mapping method for channel-quality estimation in SC-FDMA system," in *Proceedings of Asia-Pacific Conference on Communications (APCC '07)*, pp. 307–310, 2007.

[6] M. Russell and G. L. Stuber, "Interchannel interference analysis of OFDM in a mobile environment," in *Proceedings of the 45th IEEE Vehicular Technology Conference (VTC '95)*, vol. 2, pp. 820–824, Chicago, Ill, USA, July 1995.

[7] Y. Mostofi and D. C. Cox, "ICI mitigation for pilot-aided OFDM mobile systems," *IEEE Transactions on Wireless Communications*, vol. 4, no. 2, pp. 765–774, 2005.

[8] K.-Y. Han, S.-W. Lee, J.-S. Lim, and K.-M. Sung, "Channel estimation for OFDM with fast fading channels by modified Kalman filter," *IEEE Transactions on Consumer Electronics*, vol. 50, no. 2, pp. 443–449, 2004.

[9] Y.-S. Choi, P. J. Voltz, and F. A. Cassara, "On channel estimation and detection for multicarrier signals in fast and selective Rayleigh fading channels," *IEEE Transactions on Communications*, vol. 49, no. 8, pp. 1375–1387, 2001.

[10] W. H. Chin, D. B. Ward, and A. G. Constantinides, "An algorithm for exploiting channel time selectivity in pilot-aided MIMO systems," *IET Communications*, vol. 1, no. 6, pp. 1267–1273, 2007.

[11] W. Jakes and D. Cox, *Microwave Mobile Communications*, Wiley-IEEE Press, New York, NY, USA, 1994.

[12] J. Akhtman and L. Hanzo, "Sample-spaced and fractionally-spaced cir estimation aided decision directed channel estimation for OFDM and MC-CDMA," in *Proceedings of the 62nd IEEE Vehicular Technology Conference (VTC '05)*, vol. 3, pp. 1916–1920, Dallas, Tex, USA, September 2005.

[13] 3GPP, TS 36.211 V8.6.0, "Physical Channels and Modulation," http://www.3gpp.org/.

[14] M. Ozdemir and H. Arslan, "Channel estimation for wireless ofdm systems," *IEEE Communications Surveys & Tutorials*, vol. 9, no. 2, pp. 18–48, 2007.

[15] K. Kwak, S. Lee, J. Kim, and D. Hong, "A new DFT-based channel estimation approach for OFDM with virtual subcarriers by leakage estimation," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, pp. 2004–2008, 2008.

[16] D. Li, F. Guo, G. Li, and L. Cai, "Enhanced DFT interpolation-based channel estimation for OFDM systems with virtual subcarriers," in *Proceedings of the IEEE 63rd Vehicular Technology Conference (VTC '06)*, vol. 4, pp. 1580–1584, Melbourne, Australia, May 2006.

[17] B. Yang, Z. Cao, and K. Letaief, "Analysis of low-complexity windowed DFT-based MMSE channel estimator for OFDM systems," *IEEE Transactions on Communications*, vol. 49, no. 11, pp. 1977–1987, 2001.

[18] L. M. Davis, L. B. Collings, and R. J. Evans, "Coupled estimators for equalization of fast-fading mobile channels," *IEEE Transactions on Communications*, vol. 46, no. 10, pp. 1262–1265, 1998.

[19] M. K. Tsatsanis, G. B. Giannakis, and G. Zhou, "Estimation and equalization of fading channels with random coefficients," *Signal Processing*, vol. 53, no. 2-3, pp. 211–229, 1996.

[20] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall Information And System Sciences Series, Prentice-Hall, Upper Saddle River, NJ, USA, 1996.

[21] B. Porat, *Digital Processing of Random Signals: Theory and Methods*, Prentice-Hall, Upper Saddle River, NJ, USA, 1994.

[22] S. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 1993.

[23] 3GPP TR 25.943 V6.0.0 (2004-12), "Deployment aspects (Release 6)," http://www.3gpp.org/.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 731317, 11 pages
doi:10.1155/2009/731317

Research Article

Dynamic Relaying in 3GPP LTE-Advanced Networks

Oumer Teyeb,[1] Vinh Van Phan,[2] Bernhard Raaf,[3] and Simone Redana[3]

[1] *Radio Access Technologies (RATE) Section, Department of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg Øst, Denmark*
[2] *Nokia Siemens Networks, COO Research Technology & Platform, Kaapelitie 4, 90630 Oulu, Finland*
[3] *Nokia Siemens Networks, COO Research Technology & Platform, St.-Martin-Strasse 76, 81541 Munich, Germany*

Correspondence should be addressed to Oumer Teyeb, oumer@es.aau.dk

Received 30 January 2009; Accepted 30 July 2009

Recommended by Constantinos B. Papadias

Relaying is one of the proposed technologies for LTE-Advanced networks. In order to enable a flexible and reliable relaying support, the currently adopted architectural structure of LTE networks has to be modified. In this paper, we extend the LTE architecture to enable dynamic relaying, while maintaining backward compatibility with LTE Release 8 user equipments, and without limiting the flexibility and reliability expected from relaying. With dynamic relaying, relays can be associated with base stations on a need basis rather than in a fixed manner which is based only on initial radio planning. Proposals are also given on how to further improve a relay enhanced LTE network by enabling multiple interfaces between the relay nodes and their controlling base stations, which can possibly be based on technologies different from LTE, so that load balancing can be realized. This load balancing can be either between different base stations or even between different networks.

1. Introduction

The use of radio relaying with the deployment of relay nodes (RNs) for coverage extension in cellular networks is not a new concept [1]. Apart from the main goal of coverage extension, enabling relaying in a cellular network can also help in the provisioning of high data rate coverage in high shadowing environments (e.g., indoors) and hotspots, reducing the deployment costs of cellular networks, prolonging the battery lifetime for user equipments (UEs), and generally saving power by reducing the overall transmission of cellular networks and enhancing cell capacity and effective throughput. Figure 1 shows the most typical usage scenarios for relaying.

Many of the earlier studies on relaying were rather theoretical and mainly concerned with information theoretic capacity limits. It is only recently that practical solutions have been proposed due to the maturity of cellular systems and the ever increasing demand for high data rate services [1–4]. After being carefully considered in prestandardization activities like the IST-WINNER project [2], relay enhanced systems are achieving the level of maturity needed in ongoing standardization activities. The best evidence of this maturity is the IEEE 802.16j standard specifying relaying for the mobile WiMAX (802.16e) systems [3, 4].

The 3rd Generation Partnership Project (3GPP) is currently finalizing the specification of UTRAN long-term evolution (LTE) Release 8. Discussions have already started regarding LTE-Advanced standardization, and relaying has been proposed as one of the key candidate features [5, 6]. However, for the sake of economic viability, LTE-Advanced is required to be as much backward compatible as possible with LTE Release 8. This is especially important from the UE side, as it will allow users to benefit from relaying with their Release 8 terminals. Due to the assumption of focusing this paper on LTE-Advanced we refer to a base station by the 3GPP term enhanced Node B (eNB).

Several kinds of relaying systems have been proposed, the most representative ones being simple repeaters that amplify and forward the received signal, decode and forward relays that decode the received signal and regenerate it, and relays that support the full functionalities of an eNB [7]. From a system level point of view, relaying can be performed either in a *conventional* or *cooperative/collaborative* fashion.

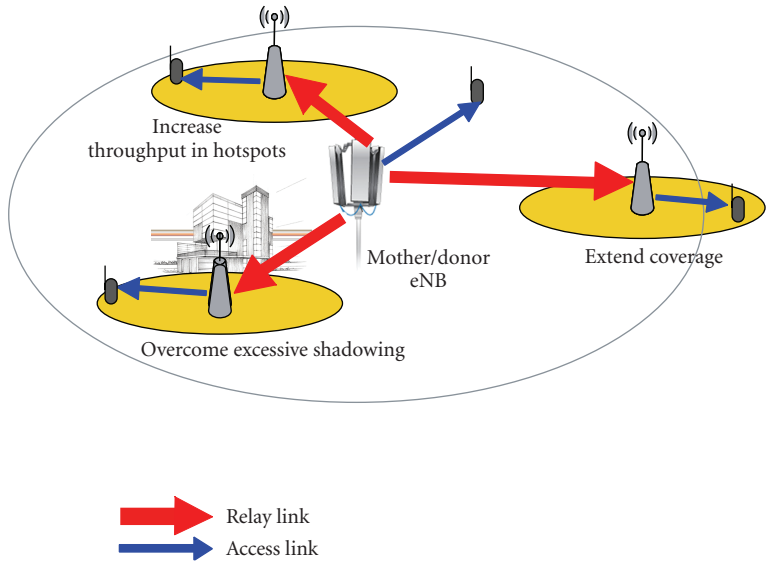


Figure 1: Examples of an LTE radio access network deployment with fixed relay nodes.

In conventional relaying, the UEs are receiving data either from the serving eNB or the RN. In collaborative relaying, on the other hand, the UEs can receive and combine the signals from several RNs and the eNB [8]. A conventional relaying scheme is assumed in this paper to support backwards compatibility as it is simple and practical for transitioning LTE into the realm of multihop systems.

Relaying can be realized at the different layers of the protocol stack. A simple amplifying and forwarding RN can be realized at the Layer 1 (L1) of the protocol stack where the RN is required to have only (some part of) the PHY layer. Layer 2 (L2) RNs, which include the protocol stack up to the Medium Access Control (MAC)/Radio Link Control (RLC) layers, enable the possibility of decentralized radio resource management (RRM). Layer 3 (L3) or higher layer RNs could almost be considered as wireless eNBs and support all the protocol layers of normal eNBs, except that they will not require an expensive backhaul as in a normal eNB (i.e., the backhaul between the RN and the eNB will be based on the LTE air interface instead of microwave or wired interface), and they are assumed to have low transmission power capabilities. Unless otherwise specified, L2 or L3 relays are assumed throughout this paper.

We consider a simple scenario where at most two hops are allowed. Such a scenario is most attractive from a practical perspective because the system complexity is strongly related to the number of hops. Throughout this paper, we refer by direct link to the connection between an eNB and a UE, by backhaul or relay link to the connection between an eNB and an RN, and by access link to the connection between an RN and a UE.

The aim of this paper is to present an architecture that will enable dynamic relay deployment in LTE networks in a backward compatible way from the UE's point of view. In particular, dynamic backhauling, multimode relay, and distributed relaying are components of the designed architecture and separately treated in the following sections. The rest of the paper is organized as follows. In Section 2, a simple extension of the basic LTE architecture to enable static deployment of fixed RNs is described. Section 3 extends this architecture to support dynamic backhauling. The possibility of using multiple air interfaces for optimal radio resource utilization is described in Section 4. Section 5 extends the multiple interfacing concepts in order to support the integration of relaying and home eNB into a single equipment. After discussing distributed relaying where a relay node can be served via several eNBs in Section 6, a conclusion is given.

2. Architecture for Static Relay Deployment

Figure 2 shows the architecture of an LTE network extended to support static relay deployment. The gateways are in charge of functions such as nonaccess stratum signalling, mobility, and bearer establishment/maintenance [9]. Due to the conventional nature of the relaying, a UE is connected either directly to an eNB or an RN, but not to both. All the traffic intended for a relayed UE is always routed to the controlling eNB (also referred to as "mother eNB" or "donor eNB") of the concerned RN by the gateways and then routed to the RN via the donor eNB.

As in Release 8 LTE, the eNBs control the radio resource management (RRM) of the system [9]. Additionally, they are also responsible for the configuration and controlling of the RNs and their resources, routing of traffic to the RNs, ensuring reliable communication links between the eNB and the UE by means of outer Automatic Repeat reQuest (ARQ), flow control to allow smaller (and cheaper) buffers in the RNs, and so forth. Thus, the gateways do not necessarily have to be aware of the existence of RNs in the system.

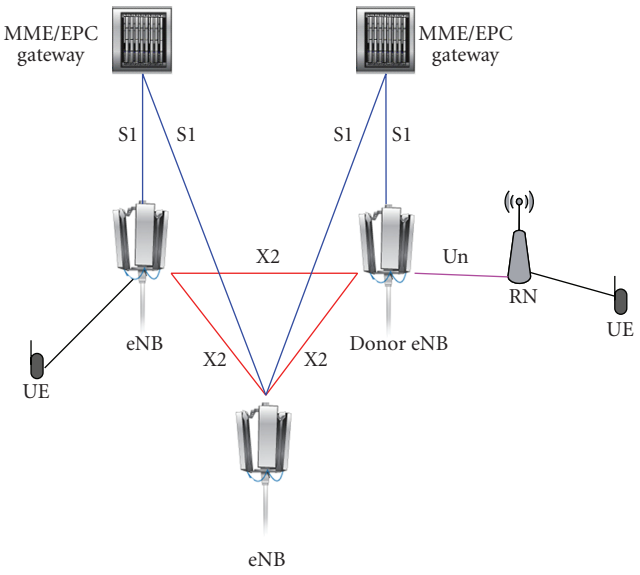


Figure 2: Architecture for relay enhanced LTE network with fixed RN deployment.

The most important task of the RN is to forward data between the eNB and the UE. It is supposed that standard LTE Release 8 UEs should be able to communicate via the RNs. An RN should thus be capable of broadcasting system level information in the same manner as the eNBs so as to appear as a normal eNB to the UEs. Due to this, the UE-RN interface should be the same as the Release 8 UE-eNB interface [9, 10]. The L3 RNs considered here also support decentralized RRM. That is, the RNs are responsible for scheduling packets on the radio interface using the resources that have been reserved to them by the eNB.

Since the link between the eNB and RN carries both the traffic to the UEs and the traffic needed to control the RNs, this interface will have different characteristics than the LTE Release 8 air interface. It is referred to as the *Un* interface and it can contain the functionalities of both the S1 interface (Gateway-eNB: for control information as well as for data transport), the X2 interface (eNB-eNB: for forwarding user data, similar to the case of handover forwarding in LTE Release 8), as well as the LTE air interface (eNB-UE: for control information as well as for data transport).

In the architecture described so far, the RNs are assumed to be deployed by the operator on certain locations, especially on hotspots and locations that are highly likely to suffer from coverage loss (cell edge and high shadowing areas), and each RN is associated with a donor eNB. However, such a static association limits the flexibility/efficiency of the system.

3. Dynamic Backhauling

Enabling dynamic backhauling of RNs is attractive for several reasons. The donor eNB may be overwhelmed by a high load within its cell, while a neighbouring cell is completely unloaded. Moreover, static association limits the system to support only stationary RNs, and thus mobile RNs (e.g., RNs attached to trains) cannot be used. Running a system with several lightly loaded cells is highly energy inefficient [11] and efficiency can be enhanced by powering off some of the eNBs or RNs in the system (e.g., during late night hours on weekdays) and concentrating the system load on few eNBs. If the eNBs that are to be powered off are relay enabled, it would also be beneficial to associate the RNs in these cells with other eNBs that are still active, instead of rendering them useless when their donor eNBs are powered off. Finally, dynamic backhauling, where the RNs can work in plug-and-play fashion, is a requirement in a Self-Organizing Network (SON), which is one of the important features demanded by cellular operators for future LTE releases [12].

In order to enable dynamic backhauling, a mechanism is needed for the RNs to discover relay-enabled eNBs that can act as their donor eNBs. The eNBs that support relaying can inform RNs about their relaying capability by including this information in the message blocks they broadcast regularly to the whole cell. This will not affect backward compatibility as the UEs can simply ignore this extra information. As an alternative, the RNs can query the eNBs to see if they support relaying using a new Radio Resource Control (RRC) protocol procedure. eNBs that do not support relaying will not recognize this message and will thus ignore it.

The information element for identifying relaying support that is sent by eNBs could include several entries such as the cell load, geographical locations, where the eNB is experiencing capacity/coverage problem and hence support by an RN is highly needed, supported mode of relaying (L1, L2, L3, etc...), energy saving settings, if any (e.g., when the eNB is scheduled to be powered off next).

When an RN is powered on, it is required that it has to be associated with an eNB before it can become fully operational. The reason for this is that it is not yet connected to the core network side and relaying a UE's connection is feasible only through the donor eNB. If relaying capability is discovered through broadcasting, the RN, when powered on, will listen to the different relay support broadcast messages of neighbouring eNBs and will select the one that satisfies acceptable criteria for donor eNB selection. On the other hand, if relaying capability information is to be acquired via a RRC request procedure, the RN has to send the request to the neighbouring eNBs that it can detect and then it will select the one that satisfies acceptable selection criteria.

Once the RN has identified the eNBs that support relaying, it can select its donor eNB based on several criteria. It can select the eNB with the best path gain (which in many cases will be the cell in which the RN is geographically located). Apart from the path gain, the interference and the load can also be taken into consideration to find the eNB that can provide the highest backhaul data rate to the RN. The one with the highest load can also be selected as the donor eNB, as it might probably need some load sharing. If the relaying capability information contains locations where coverage/capacity problems are being experienced by the eNBs, this can enable the RN to associate with the eNB that it can help optimally. Additionally, if energy saving settings are provided, it is also beneficial to select an eNB that is not scheduled to be powered off soon in order to

avoid unnecessary handovers. A combination of several of these criteria mentioned perviously can also be used in the decision process.

Although it is not within the scope of this paper, it is very informative to mention that the overall radio resource can be partitioned between the access links and the relayed links in an orthogonal fashion (also referred to as *Hard reuse*) where the relay and access links use different resources, or a nonorthogonal resource scheme (also referred to as *Soft reuse*) where the relay and access links share the radio resources to some extent. If multiple RNs are deployed within a cell, spatial separation between the access links of the RNs can guarantee a safe reuse of the resources. The partitioning can be done either in a fixed way where a subset of the radio resources is reserved for each RN or in a dynamic way where the needed resources are allocated per scheduling period. The resource splitting can be done either in a time division multiplexing (TDM) or frequency division multiplexing (FDM) manner. The splitting of resources (either in frequency or time domain) will require the use of two time or frequency slots (one for relay link and one for access link) instead of one slot for a direct connection. However, due to the high channel quality of the relay link and the access links (since it is UEs with the worst direct link quality to the eNB that end up being connected to the RN), there will be overall gain in the end-to-end throughput [13].

Dynamic backhauling implies the possibility of handing over an RN and all its associated UEs to another eNB. In order to support this, the handover mechanism of LTE specified in [9] has to be extended, as shown in Figure 3 [14]. Based on the measurement results it is getting from the RN, and also on other conditions such as energy saving settings and load-balancing communication from neighbouring cells, the donor eNB (source eNB) decides to hand over the RN to another eNB (target eNB). This handover decision is communicated to the target eNB. In this handover request message, the donor eNB summarizes the total resources required to accommodate the RN to be handed over including its associated users.

The donor eNB can at least indicate the overall (backhaul) traffic demands of the RN and its cell with necessary UE contexts in detail. It can also inform the target eNB the RN location and RN measurement reports about the RN-target eNB radio link. The target eNB can use this information to estimate the required radio resources to admit the RN.

If the required resources are available, which is checked by the admission control procedure in the target eNB, a handover request acknowledgement message is sent to the source eNB. The source eNB then sends a handover command to the RN (RRC connection reconfiguration including mobility control information), and from then on, the data destined to the RN is forwarded to the target eNB until the handover process is finalized.

Upon the reception of the handover command, the RN reacts differently depending on the scenario: the source and target eNBs have the same modes of operation (i.e., they both use the same duplexing mode, frame structure, etc.), and they are also synchronized with each other; or the two eNBs operate in different modes, or they are not synchronized.

In the first case, the RN can maintain its timing, and the UEs that it is serving do not have to change their timings. Thus, the UEs do not have to be aware of the RN handover, that is, the handover is transparent to the UEs, and as such the messages in the orange box in Figure 3 are not required. Thus in the first case, the RN detaches itself from the source and immediately starts synchronizing with the target when it gets the handover command.

The second case is more complex as the RN has to change its timing and possibly other parameters such as frame structure, cell ID, scrambling code, and reference signal structure. In this case the RN has to command the UEs to handover to the new cell, that is, the RN after the reconfiguration and timing changes. The messages inside the orange box in Figure 3 are thus required. The handover command has to be sent to the UEs before the reconfiguration of the RN, that is, before the RN synchronizes to the target eNB, because the RN will typically change its timing or other configurations at that time.

Once the RN has achieved L1/L2 synchronization with the target eNB (and in the second case, in addition to this, also when all the UEs have resynchronized with the RN), the RN sends a handover confirmation message (RRC connection reconfiguration complete) to the new donor eNB. This confirmation is a composite message that includes information about each UE that is being served through the RN. The new donor eNB then sends out a path switch request to the Mobility Management Entity (MME), which initiates a user plane update request to the serving gateway. User plane update is then performed by the serving gateway for each UE indicated in the composite handover confirmation message. A user plane update basically switches the downlink data path to the target eNB. The serving gateway then sends "end marker" packets to the source eNB, to indicate that the old path is not going to be used anymore for the concerned UEs. After the route update is performed, packets destined to the UEs served by the RN will be properly routed via the new donor eNB.

The source eNB is then advised that it can release the resources pertaining to the RN (Release RN-UE context, as the RN is seen as a special UE from the eNB's point of view), and the link between the source eNB and the RN is released. After the forwarding of the final packet in flight to the target eNB, the final resources are released by the source eNB and the handover is finalized.

It should be noted that the load-balancing handover described here is not as delay critical as a regular handover of a UE. Thus, enough time could be taken to negotiate and settle resource issues for the RN cell and its UEs. Upon receiving a handover command, the RN might have to take time to reconfigure its cell and UEs first, some UEs might need to be downgraded or even dropped due to lack of resources.

This need to downgrade or drop calls can be gathered from some indication about the available resources for the RN and its UEs in the new target cell, which can be included either in the HO request acknowledged command or an

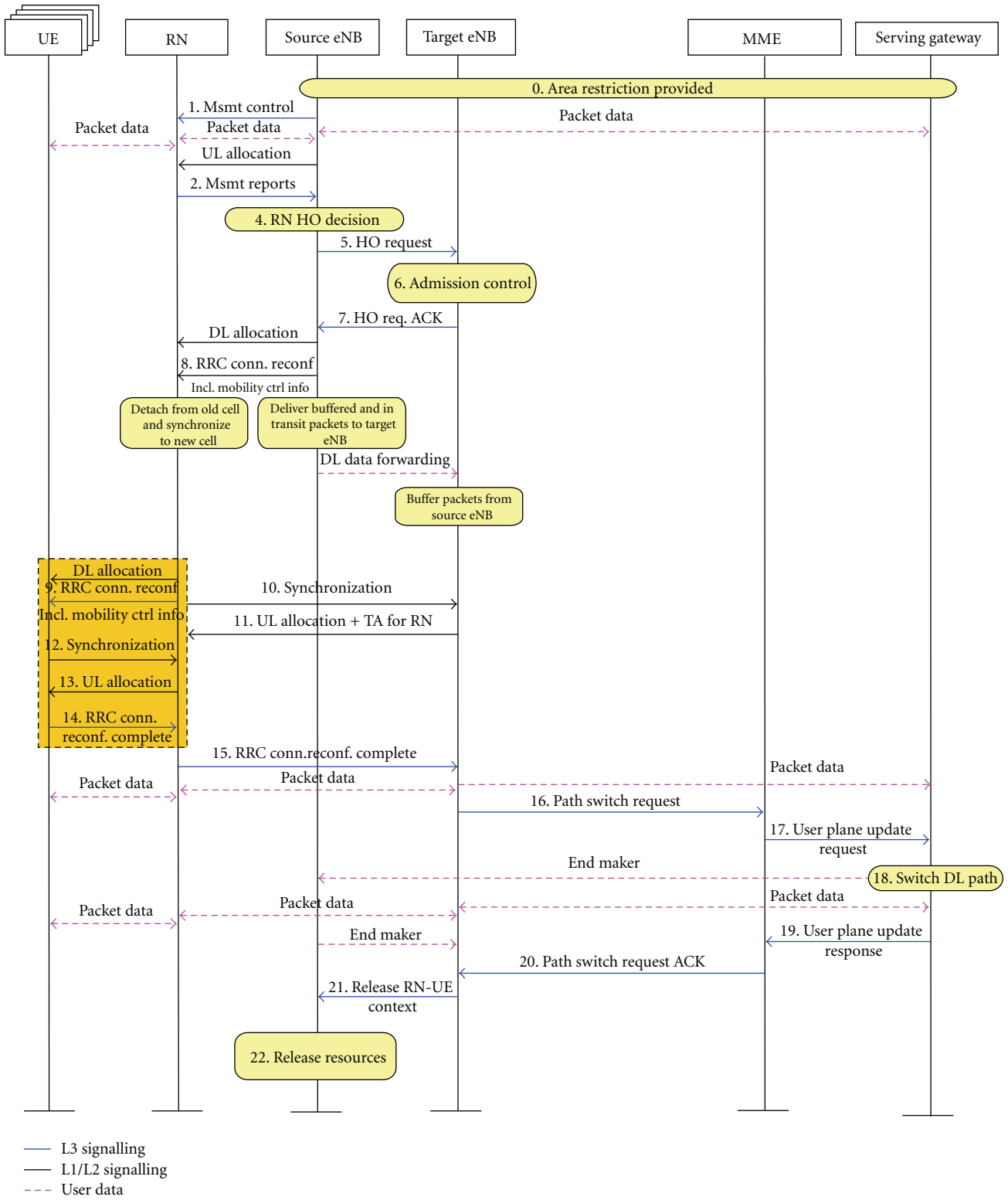


Figure 3: Handover of an RN and all its associated UEs in an LTE network that supports dynamic RN deployment.

additional signal sent from the target eNB. However, the RN might avoid these downgrades and call drops by initiating a separate handover of the concerned UEs to another RN or eNB. These additional handovers can be initiated either after the RN handover is finalized or even before that once the HO command is received. Doing these additional handovers before the RN handover is finalized will save the UEs from performing two consecutive handovers and thus it reduces the total handover delay. However, such pre-emptive handovers are not foolproof as resource reallocation and repartitioning after the RN handover is finalized might have been sufficient to provide the required quality for all the connections. Thus, the pre-emptive handovers should be initiated only when there is a very high disparity between the required resources of the relayed UEs and the available resources in the target cell.

Some UEs may have to be assigned to different resources, if the currently used resources collide with resources used for other purposes in the new eNB, for example, resources that the target eNB intends to use to communicate with the RN. It is possible to assign resources fully dynamically in LTE via fast scheduling and these allocations can be changed accordingly on the fly. On the other hand, for semipersistent scheduling, UEs are granted resources for a longer time interval and these grants need to be reconfigured for the handover. This can possibly happen even before handover initiation (i.e., before sending the handover command to the UE), immediately afterwards or after the re-establishment of the link.

Apart from dynamic association of RNs and eNBs, dynamic backhauling also allows the operator to activate and deactivate RNs in a certain area as needed in order to deploy extra capacity needed to satisfy peaks in users' demand, but at the same time save energy when it is not needed by simply powering off unnecessary RNs. Figure 4 illustrates the deactivation procedure.

Based on the measurements from the RN and load conditions in neighbouring cells, the donor eNB decides to deactivate the RN, as shown in Figure 4. The donor eNB sends out a deactivation command to the RN, and the RN initiates a handover procedure for all its users. Once the handover for each relayed UE is finalized, the RN deactivates itself and sends a deactivation confirmation message to the donor eNB.

The deactivation command from the donor eNB to the RN can contain parameters needed for future reactivation of the RN. This can include timer values such as a sleep interval during which the RN completely shuts down its transceiver, and also on-duration periods during which the RN will listen on a common control channel such as a paging channel to determine if the donor eNB is trying to reactivate it. This procedure can be done in a way similar to the Discontinuous Reception (DRX) procedure of LTE [9].

4. Multimode Relays

As one of the main driving forces behind the deployment of RNs is the low infrastructure cost, RNs in LTE-Advanced use in-band wireless links based on the LTE-Advanced air interface instead of using wired links or dedicated out-of-band (e.g., microwave) wireless links.

Though this is a cheap and simple solution, there might be a need to support out-of-band wireless links. This is especially significant in scenarios/durations where the radio resources are limited and not enough resources can be reserved to the backhaul link without compromising the quality of users that are directly being served by the eNB. Thus, it can be beneficial to use multiple interfaces in the backhaul link, not only to increase the system performance, but also to add robustness to the system by switching between the interfaces, or even sharing the load between the multiple interfaces when the need arises. The prevalence of overlay networks where multiple networks (e.g., GPRS, UMTS, HSPA, etc.) are available at a given site facilitates the possibility of switching/load-sharing.

Figure 5(a) shows the different interfaces in a relay enhanced LTE network. As can be seen in the figure, the backhaul link is using the LTE-Advanced air interface while the access link between the RN and the UE is using LTE Release 8 air interface for backward compatibility purposes. In Figure 5(b), we propose a new architecture where additional interfaces are available (Ib2 to IbN in the figure) in the backhaul link apart from the LTE-Advanced air interface.

Multimode RNs should coexist with single mode RNs (those that support only the LTE-Advanced interface), and as such, it is necessary for the eNBs to find out the interfacing options available at the RN. The exchange of the interfacing capability of the RN can be done either during the association of the eNB and the RN (i.e., when the RN is first activated or handed over to another eNB as described in Section 3), or using separate RRC *interface capability* messages after the association is finalized. This interface capability request will be ignored by single mode RNs as they are not aware of this procedure, and thus the eNB can safely assume that the concerned RN is a single mode RN if it does not get a response after a certain number of repetitions of the interface capability request.

The LTE-Advanced air interface is the default interface to be used between the donor eNB and the RN. However, once the association procedure is complete and the donor eNB is aware of the interfaces supported by the RN, the interface to be used can be modified according to system need. The interface selection decision is controlled by the eNB.

The optimal interface can be chosen based on several criteria. The interface that belongs to the network with the lowest load, or in other words, the network that can provide the highest capacity for the relay link is an obvious choice. However, the cheapest network, from the operator's operating expense (OPEX) point of view, or the network that optimizes certain resources like energy can also be chosen.

The main advantage of using multimode RNs is to dynamically modify the interface to be used for the backhaul link for optimum system performance. Thus, there is a need to deactivate the currently active interface and (re)activate another interface. The decision to change the interface is done in a similar fashion as in the case of the interface

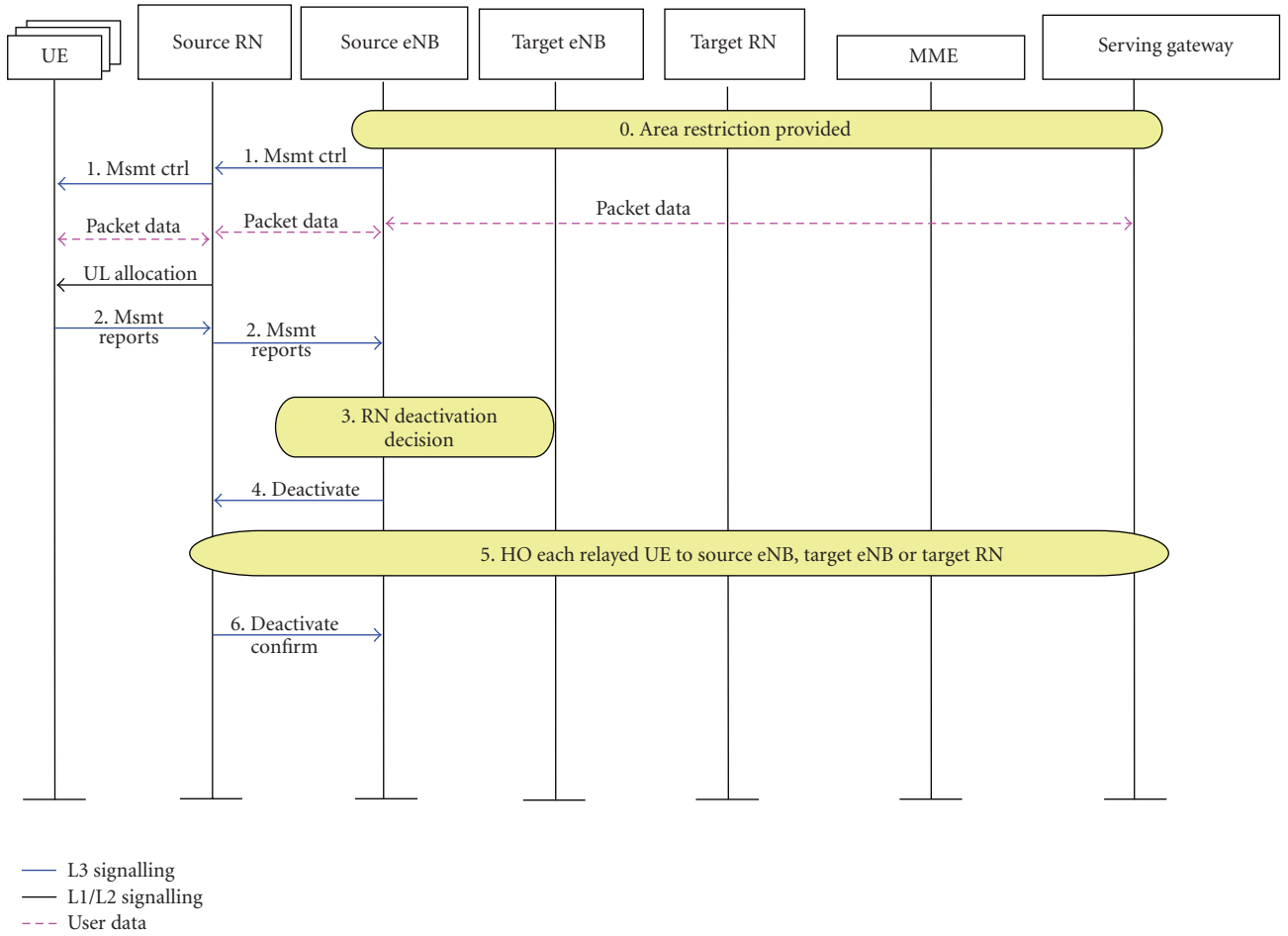


Figure 4: Deactivation of an RN in an LTE network that supports dynamic RN deployment.

selection procedure, that is, based on load, throughput, cost factors, or other reasons.

Based on the aforementioned factors, measurement results that it is getting from the RN, and other conditions such as load-balancing communication from neighbouring cells, the donor eNB decides to change the interface for the backhaul link. This decision is communicated to the RN and the RN deactivates the current interface and (re)activates the new one. The relayed UEs should not be aware of the interface changes.

Several factors have to be considered in order to make sure that no user data is lost and all active bearers that belong to relayed UEs are not disconnected. Adequate radio resources have to be allocated on the new network before the interface modification is initiated to ensure that the Quality of Service (QoS) of active bearers will still be satisfied. Also, any outstanding data on the old interface should be forwarded to the new interface, both in the donor eNB and RN, before the old interface is deactivated.

Though the possibility to switch between the different interfaces available to a multimode RN is important to transfer the connection to the network for optimal overall system utilization, there might be cases when one network is not able to provide all the needed resources for all the relayed UEs. Thus, it is essential to enable the simultaneous activation of multiple interfaces of a multimode RN.

The measurement reports from the RNs that may lead to interface activation/deactivation can be sent either periodically or triggered when certain thresholds related to the allowable load values on a given interface are reached. A combination of periodic and event based measurement reporting can also be used. For example, periodic reporting with a long reporting period can be used under normal conditions to minimize measurement overhead, but threshold-based triggering can override the periodicity and send the reports in order to avoid unnecessary delay in the interface activation/deactivation which can possibly lead to the downgrading or even the dropping of active bearers. Just as in the case of a simple UE handover between two eNBs, ping-pong effects can be prevented using hysteresis thresholds.

The two main reasons for activating multiple interfaces are either that a new bearer has to be established by a relayed UE and there are not enough resources for the backhaul link for this connection, or that a new bearer is to be established by a directly connected UE but there are not enough radio resources, unless some resources being used for the backhaul link are freed. In the first case, the new bearer will be

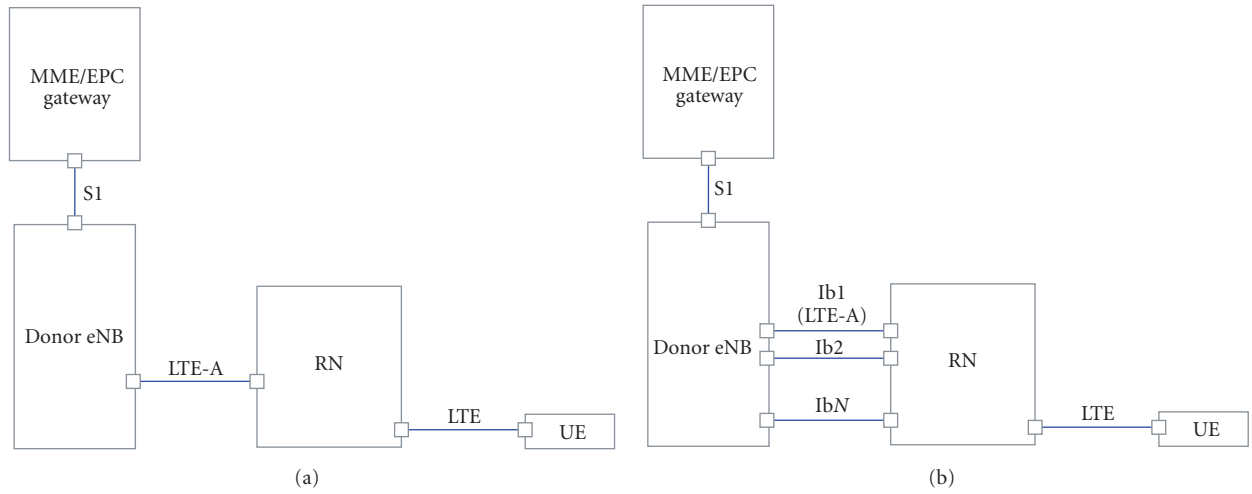


Figure 5: Interfaces between different network elements, (a) single mode RNs, (b) multimode RNs.

associated with the new interface and the rest of the bearers is not affected. In the second case, on the other hand, the donor eNB has to select a relayed bearer or even directly connected bearers which have to be transferred to the new network, so that the new direct bearer can be admitted to the cell. Measures have to be taken on the data of the selected bearers in order to avoid user data loss and dropped bearers.

5. Relay Node-Home Enhanced NB Integration

3GPP is currently standardizing home eNBs, also known as "femto-cells" [15]. Home eNBs are similar to WLAN access points and will be installed in residence and office buildings where there is already an access to Internet, for example, via a wired system. They will appear as normal eNBs to the UEs and they will access the core network of the operator via the Internet. Generic Access Network (GAN), also known as Unlicensed Mobile Access (UMA), is chosen by 3GPP as the way to provide the interfacing to the operator's core network through the Internet. Though home eNBs seem to be an attractive solution for nonreal time (NRT) services, there might be some real time (RT) services that have very strict QoS requirements that might not be met via the Internet connection (e.g., when there is congestion). In order to resolve this issue, we extend the concept of multimode RNs described in Section 4 to support also home eNB functionality.

Figure 6 shows a multimode RN enhanced to support home eNB functionality. As can be seen from the figure, there is a new interface, which we refer to as $S1_{RN}$, between the RN and the core network elements using DSL or cable, in a similar fashion as a home eNB. Note that $S1_{RN}$ does not necessarily have to be a wired interface, as long as it gives a direct connection to the Internet, which is then routed to the core network of the operator. Note that L3 or higher layer relays are required to enable this functionality as routing via the Internet is required.

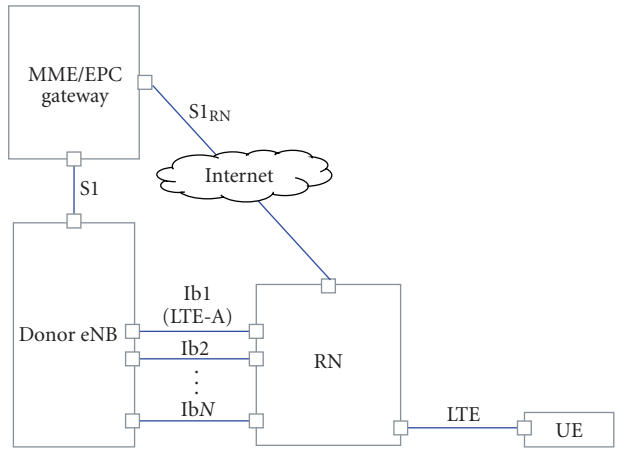


Figure 6: A multimode RN enhanced with home eNB functionality.

Such an RN that is equipped with a wired/wireless access to the Internet can act as a home eNB when the need arises, and can operate in several modes.

(1) It can use the Internet as an alternative interface to the LTE-Advanced interface; that is, the data for the relayed users will be routed from the eNB to the RN via the core network and the Internet, and the RN forwards it to the relayed UEs via the LTE air interface (and vice versa for uplink traffic).

(2) It can use the Internet for load balancing where some of the bearers will be supplied via the Internet as in the first case, and the rest is provided by LTE-Advanced or/and the other wireless interfaces in the backhaul.

(3) It can operate as a stand alone home eNB where the RN is directly connected via the $S1_{RN}$ interface to the core network.

The first two options require that the system has to be configured to support home eNBs (i.e., connection to the operator's network via the Internet) and a modification in the GAN in order to route the data back and forth via the eNB instead of the gateways. This means, while the RN is connected to the eNB logically, this connection is realized physically via the Internet and the GAN instead of a direct physical link between the RN and eNB that has been assumed so far in the previous sections.

The third option is basically the same as a home eNB operation, and the donor eNB does not have to be concerned with the data for the relayed UEs any more as they are transported directly to the gateway without the need to reach the donor eNB. However, it is still beneficial to maintain a connection between the donor eNB and the RN, which uses very few radio resources, in order to enable the RN to switch back to the "normal RN" mode when enough radio resources become available for the backhaul.

The deployment of RNs that can also simultaneously act as home eNBs will not only make the system more flexible by creating alternatives for load sharing and load switching, it also makes the system more robust to failures. That is, when the radio resources in one of the backhaul links of the RN are exhausted, there is still a way to transfer the load via the home eNB interface and vice versa.

A typical usage scenario for a home eNB enabled RN is to use the wireless backhaul connections for RT services with very strict QoS requirements while using the Internet for NRT services and for RT services with more relaxed QoS requirements. This is due to the fact that the operator has full control over the different (wireless) networks available for the backhaul link, but not on the Internet. Notable latency may be expected due to the longer path needed for the packets if the RN is operating as a home eNB for the first two options. For example in the uplink, in the home eNB case the path is UE-RN-GAN-Gateways while for the first and second options (in the home eNB mode), it will be UE-RN-GAN-donor eNB-GAN-Gateways. That is, option 1 and home eNB mode of option 2 are more suitable for NRT services or RT services with relaxed QoS requirements, while the normal relay mode of option 2 (and to some extent, option 3) is more suitable for RT services with strict QoS requirements.

6. Distributed Relaying

When we refer to relaying, especially in the context of relay enhanced LTE, the normal assumption is that there is a one-to-one association between RNs and eNBs (i.e., multiple RNs can be connected through an eNB, but an RN is connected only to one eNB). Though such an architecture, as shown in Figure 2, is a straightforward and simple solution to enable relaying in LTE, it might limit the system performance because the end-to-end performance of relayed UEs will be constrained by the capacity available on the backbone link between the donor eNB and the core network (i.e., the link that is accessible through the S1 interface). For example, even if there are sufficient radio resources for the relay link, the performance of relayed UEs can degrade if there is congestion

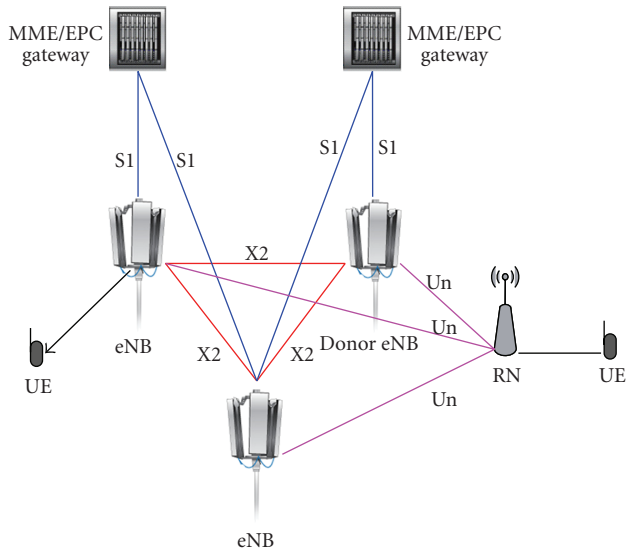


Figure 7: Architecture for enabling distributed relaying in LTE.

in the backbone. In practice, S1 links are expensive, and usually operators do not deploy enough capacity to support the maximum cell capacity offered by the air interface.

Apart from the backbone that can turn out to be a bottleneck, we can have insufficient resources on the relay link with the donor eNB while we have a lightly loaded neighbouring cell. In [12], it is proposed that neighbouring cells can communicate their load information via the X2 interface, which will then probably lead to the handover of some of the users to the neighbour cell. However, the S1 links of the two eNBs (assuming they use different S1 links) are not shared; that is, the load sharing is performed by handing over some of the users to the slightly loaded cell. Though handover to the lightly loaded cell is an option, it is not a totally flexible solution as back and forth handover of the relay and all its relayed UEs between eNBs can be an expensive procedure. Not only that, with handover we are only able to use the capacity of just one neighbour cell, instead of the sum of the available capacity in all the neighbouring cells.

In order to enable many-to-many connections between RNs and eNBs, the RNs can be connected to multiple eNBs through the Un interface, or one connection can be kept with the donor eNB and this eNB distributes the data to neighbouring eNBs via the X2 interface. Since the focus of this paper is on the radio access network, rather than the transport network, we will focus only on the first approach.

Figure 7 shows the architecture of a relay enhanced LTE system modified to support multiple Un connections. Originally, the RN is associated only with the donor eNB, and it remains so if the required QoS can be achieved for the relayed UEs. Then, due to congestion on the S1 link and/or unfavourable conditions on the Un radio link, the performance of the relayed UEs starts to degrade. At this point, based on the latest measurement reports by the RN regarding neighbouring eNBs, the donor eNB may contact suitable neighbouring eNBs to find out how much and which

resources they are willing to share in providing additional Un radio links for the RN. These include all the necessary radio network identifiers and radio configuration information for the RN to establish and communicate with and via the relevant neighbouring eNB(s).

Once an agreement is reached between the donor eNBs and its neighbours, the donor eNB may contact a central network controller (i.e., a new functionality introduced in, e.g., the gateway to support SONs) to set up corresponding coordinated S1 links with indicated neighbouring eNBs for the RN and request the RN to establish additional Un radio connection with the concerned eNBs according to the assigned configurations.

The RN will still have only one donor eNB that is responsible for network coordination and control but this donor eNB role can be switched between the different eNBs depending on the resources available in the different cells. The RN will be notified either by the donor eNB or by the new eNB(s) as to which resources it can use to communicate with the new eNB(s).

Thus, in the uplink, the RN from then onwards will send data to a given eNB if the resources used are those assigned for communicating with that eNB. The data is delivered via multiple S1 links to the gateway. The gateway may have to resequence the data arriving from the different eNBs before forwarding them to the destination, similar to the case of handover.

The downlink operation may require the gateway to distribute the data to multiple eNBs. The gateway can be informed which eNBs are connected to the RN and depending on the load of the S1 links of these involved eNBs, the gateway will route the data over the different S1 links. Data belonging to the same bearer may end up being routed via different S1 links, and thus there should be a mechanism on the RN to resequence the data flow. Though this way of data distribution is optimal from the usage of the S1 link, it is suboptimal as it does not consider the load on the relay link of the different eNBs. A periodic reporting of the relay link load of the different eNBs could thus be beneficial for the gateway in order to distribute the load reasonably.

Another way to distribute the data is to let the eNBs and the gateway collaborate to decide which bearers should be delivered through which eNBs, and the gateway routes the downlink data based on this agreement. This is simpler than the fully gateway controlled data distribution described above, as it does not require the resequencing of data that belong to the same bearer. However, it is not as flexible and might lead to suboptimal decisions.

In case multimode RNs are used, the distributed S1 comes handy when the available resources in one network are not enough for the backhaul link and some of the connections have to be transferred to another network (i.e., the RN will have active connections with several eNBs via different network interfaces).

The support of distributed S1 relaying described in this paper is completely transparent to the UEs. However, changes are required in the RN, eNB, and the gateways. For distributing S1 via several X2 links, actually no changes are required at the RN. The most significant change is the decision mechanism at the eNB, RN, and gateways as to where to route the data, and the resequencing of data arriving via several S1 interfaces at the gateway and data coming via several Un interfaces at the RN. The support of distributed S1 interface will not only make the system more robust to problems related to transport network under dimensioning, but it will also make the system more flexible by creating alternatives for load sharing and load switching.

7. Conclusions

Relaying is expected to play a pivotal role in LTE-Advanced networks, by helping to extend the coverage around cell edges and high shadowing environments and also increasing the capacity in hotspots. Thus, we have proposed a flexible architecture that will enable dynamic relaying in LTE networks, while still maintaining backward compatibility with LTE Release 8 user equipments. The dynamic backhauling configuration proposed in this paper paves the way to flexible, efficient, and self-optimizing multihop cellular networks. Operators do not have to put extensive effort in finding the most optimal locations for placing relay nodes through exhaustive radio planning as optimal eNB-RN associations can be made on the fly. This can lead to big reductions in the planning costs required for enabling relaying in future releases of LTE, enabling even end-users to be able to install relays as easily as WLAN access points. We have also proposed multimode relays that support several network interfaces. This will not only make the system more flexible by creating alternatives for load sharing/switching between different links; it also makes the system more robust to failures. That is, when the resources of one network are exhausted, there is still a possibility to transfer the load fully or partially to other networks, or even using a connection via the Internet, if the RN is enabled to support home eNB functionality. Finally, we have proposed a distributed relaying architecture with many-to-many connections between relay nodes and eNBs, to make the system more robust to problems related to transport network under dimensioning, and also enable load sharing between different cells.

References

[1] H. Yanikmoeroglu, "Fixed and mobile relaying technologies for cellular networks," in *Proceedings of the 2nd Workshop in Applications and Services in Wireless Networks (ASWN '02)*, pp. 75–81, Berlin, Germany, July 2002.

[2] "IST-WINNER Project," http://www.ist-winner.org/.

[3] IEEE STD.2007.4312731, "Draft Standard for Local and Metropolitan Area Networks–Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems : Multihop Relay Specification," 2007.

[4] IEEE P802.16j Base Line Document, "Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Multihop Relay Specification," 2007.

[5] P. E. Mogensen, et al., "LTE-Advanced: The path towards gigabit/s in wireless communications," Wireless Vitae Conference, May 2009.

[6] 3GPP TR 36.813 v0.1.1, "Further advancements for E-UTRA, physical layer aspects," November 2008.
[7] S. Valentin, et al., "Cooperative wireless networking beyond store-and-forward: perspectives for PHY and MAC design," in *Proceedings of the Wireless World Research Forum (WWRF '06)*, WG3 whitepaper, 2006.
[8] R. Pabst, B. H. Walke, D. C. Schultz, et al., "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Communications Magazine*, vol. 42, pp. 80–89, 2004.
[9] 3GPP TS 36.300 v8.7.0, "E-UTRA and E-UTRAN overall description : stage 2 (release 8)," December 2008.
[10] E. Dahlman, et al., *3G Evolution: HSPA and LTE for Mobile Broadband*, Academic Press, New York, NY, USA, 2007.
[11] T. Ruuska, "Adaptive power management in telecommunication network," US patent no. 6584330, June 2003.
[12] 3GPP TR 36.902, "Self-configuring and self-optimizing network use cases and solutions (release 9)," September 2008.
[13] T. Wirth, V. Venkatkumar, and T. Haustein, "LTE-Advanced relaying for outdoor range extension," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '09)*, Anchorage, Alaska, USA, September 2009.
[14] O. Teyeb, V. Van Phan, B. Raaf, and S. Redana, "Handover framework for relay enhanced LTE networks," in *Proceedings of the IEEE International Conference on Communications (ICC '09)*, Dresden, Germany, June 2009.
[15] 3GPP TS 25.467 V9.0.1, "UTRAN architecture for 3G Home Node B (HNB);Stage 2," Sep. 2009.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 618787, 14 pages
doi:10.1155/2009/618787

Research Article

Relay Architectures for 3GPP LTE-Advanced

Steven W. Peters, Ali Y. Panah, Kien T. Truong, and Robert W. Heath Jr.

Department of Electrical and Computer Engineering, The University of Texas at Austin, 1 University Station C0803, Austin, TX 78712-0240, USA

Correspondence should be addressed to Steven W. Peters, speters@mail.utexas.edu

Received 17 February 2009; Accepted 31 May 2009

Recommended by Angel Lozano

The Third Generation Partnership Project's Long Term Evolution-Advanced is considering relaying for cost-effective throughput enhancement and coverage extension. While analog repeaters have been used to enhance coverage in commercial cellular networks, the use of more sophisticated fixed relays is relatively new. The main challenge faced by relay deployments in cellular systems is overcoming the extra interference added by the presence of relays. Most prior work on relaying does not consider interference, however. This paper analyzes the performance of several emerging half-duplex relay strategies in interference-limited cellular systems: one-way, two-way, and shared relays. The performance of each strategy as a function of location, sectoring, and frequency reuse are compared with localized base station coordination. One-way relaying is shown to provide modest gains over single-hop cellular networks in some regimes. Shared relaying is shown to approach the gains of local base station coordination at reduced complexity, while two-way relaying further reduces complexity but only works well when the relay is close to the handset. Frequency reuse of one, where each sector uses the same spectrum, is shown to have the highest network throughput. Simulations with realistic channel models provide performance comparisons that reveal the importance of interference mitigation in multihop cellular networks.

1. Introduction

The Third Generation Partnership Program's Long-Term Evolution Advanced (3GPP-LTE-Advanced) group is developing a new standard for mobile broadband access that will meet the throughput and coverage requirements of a fourth generation cellular technology [1]. One of the main challenges faced by the developing standard is providing high throughput at the cell edge. Technologies like multiple input multiple output (MIMO), orthogonal frequency division multiplexing (OFDM), and advanced error control codes enhance per-link throughput but do not inherently mitigate the effects of interference. Cell edge performance is becoming more important as cellular systems employ higher bandwidths with the same amount of transmit power and use higher carrier frequencies with infrastructure designed for lower carrier frequencies [2]. One solution to improve coverage is the use of fixed relays, pieces of infrastructure without a wired backhaul connection, that relay messages between the base station (BS) and mobile stations (MSs) through multihop communication [3–11].

Many different relay transmission techniques have been developed over the past ten years. The simplest strategy (already deployed in commercial systems) is the analog repeater, which uses a combination of directional antennas and a power amplifier to repeat the transmit signal [12]. More advanced strategies use signal processing of the received signal. Amplify-and-forward relays apply linear transformation to the received signal [13–15] while decode-and-forward relays decode the signal then re-encode for transmission [16]. Other hybrid types of transmission are possible including the information-theoretic compress-and-forward [17] and the more practical demodulate-and-forward [18]. In research, relays are often assumed to be half-duplex (they can either send or receive but not at the same time) or full-duplex (can send and receive at the same time) [19]. While full-duplex relays are under investigation, practical systems are considering half-duplex relay operation,

which incur a rate penalty since they require two (or more timeslots) to relay a message. Two-way relays avoid the half-duplex assumption by using a form of analog network coding that allows two messages to be sent and received in two time-slots [20]. Relaying has been combined with multiple antennas in the MIMO relay channel [21, 22], and the multiuser MIMO relay [23]. Despite extensive work on relaying, prior work has not as extensively investigated the impact of interference as seen in cellular systems. One exception is [24], which utilizes resource allocation to avoid interference. Conversely, this paper considers exploiting the interference using increased spatial dimensions via extra antennas at the relay.

The first commercial wireless network to incorporate multihop communication was IEEE 802.16j [25]. Its architecture constrained the relays for being served by a single base station and allowed them to communicate in only one direction at a time (i.e., either uplink or downlink). From a design perspective, unfortunately, IEEE 802.16j had several restrictions that drastically limited its capability, for example, the transparent mode that supports relaying-ignorant mobile subscribers. Further, the relays were not designed to specifically mitigate interference. Consequently, LTE-advanced may consider more sophisticated relay strategies and thus may expect larger performance gains from the inclusion of relaying.

Investigation into the possible relaying architecture for LTE-Advanced has begun. The coverage and throughput gains for an OFDMA network have been numerically analyzed using both idealized terrain [26] and ray tracing software applied to particular urban areas [27, 28]. The types of relaying strategies considered in these papers were relatively simple, considering only one-way single-antenna decode-and-forward relaying. The general conclusion is that multihop relaying is a cost-efficient solution to achieving the systemwide goals of next generation OFDMA networks.

In this paper, we evaluate the benefits of several promising relaying strategies for 3GPP-LTE-Advanced. We consider three specific strategies including one-way relays, two-way relays, and shared relays. The one-way relay possesses only a single antenna and is deployed once in every sector. It performs a decode-and-forward operation and must aid the uplink and downlink using orthogonal resources. The shared relay concept was recently proposed in IEEE 802.16m [29] but is readily applicable to GPP. The idea is to place a multiple antenna relay at the intersection of two or more cells. The relay decodes the signals from the intersecting base stations using the multiple receive antennas to cancel interference and retransmits to multiple users using MIMO broadcast methods. The two-way relay, also called analog network coding [30] and bidirectional relaying [31], is a way of avoiding the half-duplex loss of one-way relays [32]. The key idea with the two-way relay is that both the base station and mobile station transmit to the relay at the same time in the first time slot. Then, in the second time slot, the relay rebroadcasts what it received to the base station and mobile station. Using channel state information and knowledge of their own messages, the base and mobile stations are able to decode information sent from the other party.

To study the performance of each relaying strategy we derive expressions for their achievable rate assuming Gaussian signaling. The rate expressions illustrate how other-sector and other-cell interferences impact performance and allow for efficient network simulation. For example, the analysis shows that two-way relaying has the potential for severe interference enhancement since (i) there are more sources of interference and (ii) it performs an amplify and forward that rebroadcasts the received interference. Shared relaying seems to offer the most resilience to interference since it exploits the MIMO MAC (multiple access) channel to decode three signals cochannel and the MIMO broadcast channel to deliver three interference-free signals. The direct path is neglected in each of the relaying scenarios as the area under consideration is mainly the cell edge.

To compare the performance of different relay strategies, we compare their performance using a system simulator. Channel models from the IEEE 802.16j specification [33] are used since they include models for fixed relays. The simulator places users in fixed locations in each sector and computes the sum rates derived in this paper assuming that the channel is fixed over the length of the packet. These rates are reasonable in that they are nearly achievable in real slow-fading systems with powerful coding and aggressive adaptive modulation. Comparing the performance of different relaying strategies in a single set of simulations provides extensive comparability that is not possible when comparing different references.

As a baseline for performance comparison we compare with several different cellular configurations including sectoring and frequency reuse. To be fair, we also compare with an emerging transmission technique known as base station coordination [34–37]. The idea is that by coordinating the transmission of multiple base stations, sharing data and channel state information, it is possible to eliminate interference by effectively having the multiple base stations act as one single transceiver. Several suboptimal strategies have been proposed to realize base station coordination such as coordinated resource allocation [38] or clustered coordination [39]. Such strategies have made base station coordination a viable technology for GPP that may be complementary to relaying or a more complex alternative.

The main conclusions of this paper are as follows. The one-way relay enhances capacity near the cell edge but is very limited by interference. The shared relay is able to remove much of the dominant interference and provides much of the gain of localized base station coordination, which gives the highest rates of the strategies compared in this paper. The two-way relay struggles to get any rate to the mobile-to-base station link unless the relay is very close to the mobile station because of interference from adjacent base stations. Further research into this area is warranted, however, by the success of the two-way relay in the downlink combined with its simplicity. In all cases, frequency reuse 1 (where each sector and each cell use the same spectrum) outperformed frequency reuse 6 (where the spectrum is divided into six bands, one for each sector).

The rest of this paper is organized as follows. Section 2 introduces the general cellular model considered in this paper. Section 3 discusses the one-way architecture as a baseline of comparison for the rest of the paper. Section 4 considers two-way relaying and derives the sum rate over a number of different CSI assumptions. Section 5 presents a transmission strategy for shared relaying and derives the sum rate. Section 6 discusses base station coordination over a limited area. Section 7 compares all of the presented strategies under different frequency reuse plans. Section 8 gives a discussion of the results from the previous section while Section 9 summarizes the main results in the paper and provides directions for future work.

This paper uses the following notation. The log refers to $\log_2$. Bold uppercase letters, such as $\mathbf{A}$, denote matrices, bold lowercase letters, such as $\mathbf{a}$, denote column vectors, and normal letters a denote scalars. The notation $\mathbf{A}^*$ denotes the Hermitian transpose of matrix $\mathbf{A}$. The letter $\mathbb{E}$ denotes expectation, $\min\{a, b\}$ denotes the minimum of a and b, $|a|$ is the magnitude of the complex number a, and $\|\mathbf{a}\|$ is the Euclidean norm of vector $\mathbf{a}$.

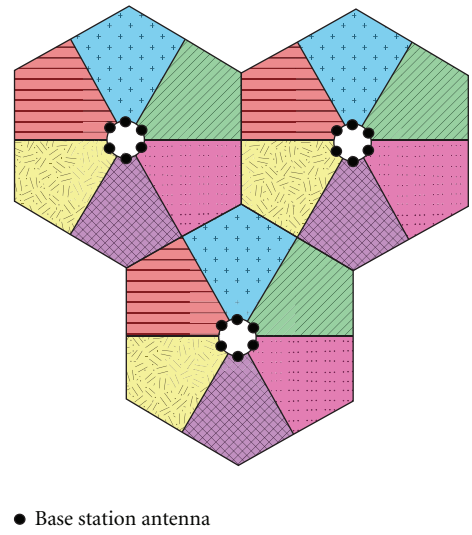


Figure 1: System model with 3 cells, each with 6 sectors. The analysis makes no assumption on the number of cells, and the frequency reuse pattern varies for the different architectures under consideration. This paper focuses on the triangular region in the center of the model.

2. System Model

In the analysis we consider an arbitrary hexagonal cellular network with at least three cells as shown in Figure 1; the simulations will include an extra tier of cells, providing two tiers of total interference (see Section 7 for details). The base stations are located in the center of each cell and consist of six directional antennas, each serving a different sector of the cell. The antenna patterns are those specified in the IEEE 802.16j channel models [33]. The channel is assumed static over the length of the packet, and perfect transmit CSI is assumed in each case to allow for comparison of capacity expressions. Thus, each cell has $S = 6$ sectors. The multiple access strategy in each sector is orthogonal such that each antenna is serving one user in any given time/frequency resource. We assume that the channels are narrowband in each time/frequency resource, constant over the length of a packet, and independent for each packet. This is known as the block fading model. These assumptions correspond to one ideal LTE OFDM subchannel and, although unrealistic in practice, are useful for deriving capacity equations that can be used for deciding the actual data rate and for simulations deriving an upper bound on throughput.

Most of the analysis in this paper will focus on downlink communication, but a similar analysis can be applied to the uplink in each case. In the one-way and shared relay cases, communication takes place in two orthogonal phases. In the first phase, the base station transmits while the relay receives (the mobile may or may not receive), and in the second phase the relay transmits while the mobile receives. There will be a capacity penalty due to the use of two phases to transmit the same information. We assume that the phases are synchronized so that the first phase and second phase occur simultaneously in all cells. In the two-way case, the base station and mobile stations both transmit in the first phase, while the relay transmits in the second phase, as will be explained in Section 4.

We consider different rates of frequency reuse. For a reuse of r, the spectrum is divided into r orthogonal bands where each one will be used in a regular pattern M/r times over an area covering M cells. We refer to this as $M \times r$ reuse. In this paper we will consider only 1×1 reuse and 1×6 reuse, and thus for simplicity we will henceforth drop the M from the notation and refer to only reuse r. In this case, mutual information will be scaled by $1/r$ to make fair comparisons. Different patterns of frequency reuse are used in different scenarios as shown in Figure 2. For shared relaying and base station coordination, the interfering sectors share the same frequency. For the one-way relay and the two-way relay, the interfering sectors use different frequencies. The analysis assumes that one user per sector has been arbitrarily scheduled, meaning that the exact scheduler is not considered since we are not analyzing multiuser diversity.

The system details of each specific architecture are explained in their respective sections. Specifically, we compare each transmission model with frequency reuse factors of 1 and 6. The one-way model consists of one single-antenna relay per sector serving only users in its sector. The shared relay is shared among three sectors in three adjacent cells (e.g., the sectors making up the center triangle in Figure 1), allowing it to serve users in each of those sectors. The two-way model consists of a single amplify-and-forward relay per sector and allows simultaneous uplink/downlink communication, removing the half-duplex loss of conventional relaying. Base station coordination assumes a lossless, zero-delay fiber link between adjacent sectors (the same ones serving the shared relay) and allows the base stations to cooperatively transmit in the downlink and receive in the uplink as if they were one large multiple-antenna transceiver.

Each of these models is discussed in the remainder of this paper.

Each hop of communication is assumed to use ideal coding and adaptive modulation so that mutual information may be used. This does not, however, guarantee that the end-to-end capacity is reached as the relays are performing a strictly suboptimal strategy (decode-and-forward for the shared and one-way relays, amplify-and-forward for the two-way relay). Other-sector and other-cell interference is assumed Gaussian and treated as noise unless specifically treated as in the shared relay case. All RF receive chains are assumed to have identical noise variance σ_N^2.

3. One-Way Relaying Model

In this section we introduce the one-way transmission model, which resembles IEEE 802.16j relaying. As with IEEE 802.16j, each relay has a single "parent" base station, creating a tree architecture. The relay, which decodes its receives signal, is thus a part of the cell its parent BS serves. Further, the uplink and downlink are divided orthogonally in time or frequency, depending on the duplexing method. Finally, the mobile station is unable to exploit the direct link. To simplify the analysis and ensure for fair comparison, we allow one single-antenna decode-and-forward relay per sector.

Assuming that all base stations transmit at the same time, frequency, and power, and that the cellular architecture is such that each cell sees the same interference (i.e., neglecting network edge effects), we can focus on a single sector of a single cell and avoid overuse of subscripts. As mentioned in Section 2, we assume an i.i.d. block fading model and can thus focus on the transmission of a single block of packets over which the channel is static. We also remove time indices of the symbols for ease of notation.

If the scheduled user is being served by the relay in its sector, the relay will receive

$$y_R = hs + \mathbf{h}_I^* \mathbf{s}_I + v_R, \tag{1}$$

where h is the BS-RS channel (transmit power is absorbed into h), s is the symbol transmitted by the BS (normalized so that $\mathbb{E}|s|^2 = 1$), $\mathbf{h}_I$ is the vector of channels between the relay and all interfering base stations (including intercell and intersector), $\mathbf{s}_I$ is the vector of transmitted symbols from all the interferers, and v_R is the additive white Gaussian noise observed at the relay with variance σ_N^2. The subscript I refers to interference, N refers to noise, and the subscript R denotes that the reception is at the relay.

Assuming that $\mathbf{h}_I^* \mathbf{s}_I$ is Gaussian with variance $\sigma_{h_I}^2$, then the relay can decode s with arbitrary reliability if s is drawn from a Gaussian codebook with rate

$$R_1 \leq \log\left(1 + \frac{|h|^2}{\sigma_{h_I}^2 + \sigma_N^2}\right). \tag{2}$$

(We assume no knowledge of $\mathbf{h}_I$ and thus each interfering term is unlikely to be truly Gaussian, although the sum over many interferers helps in this regard. This assumption is an ideality in order to treat the interference as noise and is made frequently in the literature. Further, the variance of the interference will change from block to block but will be constant over the packet.)

The relay then re-encodes s into x with rate R_2 and transmits x in the second phase of transmission. The mobile receives

$$y_M = gx + \mathbf{g}_I^* \mathbf{x}_I + v_M. \tag{3}$$

Here, g is the RS-MS channel (with absorbed transmit power as in the first hop), $\mathbf{g}_I$ is the vector of channels between the mobile and all interfering relays, and $\mathbf{x}_I$ is the vector of transmitted symbols from all the interferers in the second phase of transmission. As in the first hop, the interference is assumed to be Gaussian and has variance $\sigma_{g_I}^2$.

The mobile will theoretically be able to decode x with arbitrary reliability if it is drawn from a constellation with rate

$$R_2 \leq \log\left(1 + \frac{|g|^2}{\sigma_{g_I}^2 + \sigma_N^2}\right). \tag{4}$$

We assume that the normalized durations of two phases of transmission are t and $(1-t)$ with $0 \leq t \leq 1$. The capacity of the two-hop transmission is defined as the bottleneck of the two hops with the optimal time sharing as [40]

$$R = \min_{0 \leq t \leq 1} \{tR_1, (1-t)R_2\}. \tag{5}$$

Given R_1 and R_2, while tR_1 is an increasing function of t, $(1-t)R_2$ is decreasing with t. The time sharing is thus optimal when the two terms are equal, which results in the optimal time sharing $t^* = R_2/(R_1 + R_2)$. When using optimal time-sharing, the rate of the two-hop scenario is

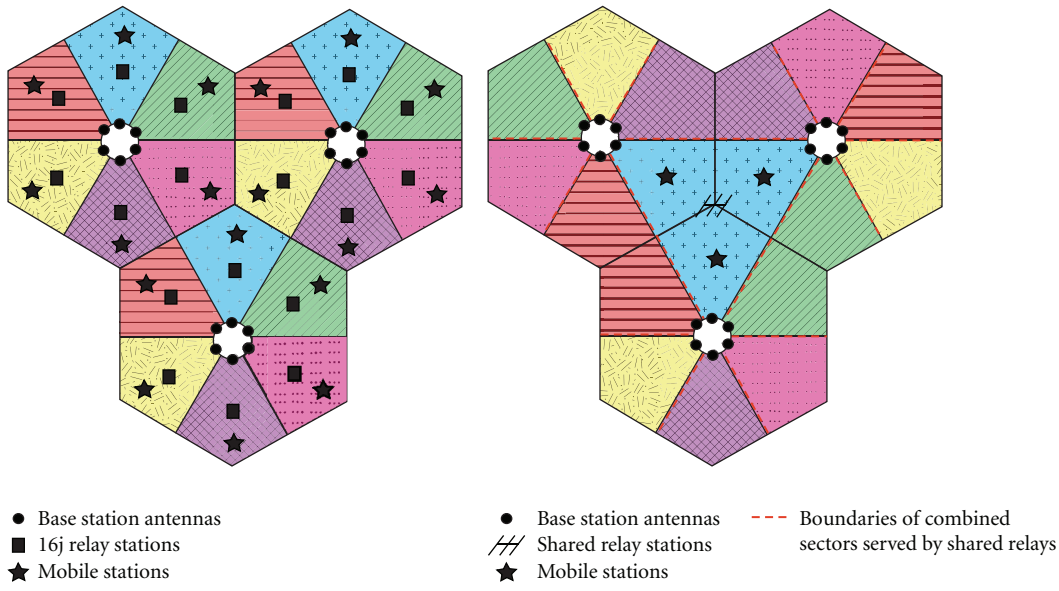
$$r_{\mathrm{OW,DL}} = \frac{R_1 R_2}{R_1 + R_2}. \tag{6}$$

Here, the subscripts OW and DL refer to one-way relaying and downlink transmission, respectively. Further, the letter r is used to refer to the rate of a single user rather than a sum of users.

The rate in (6) is the downlink rate of one user in one sector of the network. In the simulations of Section 7, we will focus on the sum rate over adjacent sectors, which will simply be the sum of (6) over those users. The main assumptions and parameters for the two-way model are given in Table 1.

4. Two-Way Relaying

Consider the cellular network model of Figure 3 where each cell is sectorized, and each sector has a single relay station (RS) serving a single mobile station (MS). There are an arbitrary number of cells in the network, and the base station (BS) in each cell is equipped with one antenna per sector. As in previous sections, we can assume a large number of cells to allow the analysis to focus on one arbitrary sector in one arbitrary cell. The objective then is to transmit the symbol (again dropping the time index as in previous sections) s_i from the ith BS to the ith MS and the symbol u_i from the



(a) Reuse pattern for one-way and two-way relaying (b) Reuse pattern for shared relaying and base station coordination

Figure 2: Frequency reuse patterns with reuse 6 for (a) one-way and two-way relaying and (b) shared relaying and base station coordination.

Table 1: System parameters for one-way relay model. The main differences between the one-way relay model and the shared relay are the number of antennas per relay, the relay transmit power, and the number of relays per sector. Since over a large network there will be approximately 3 times as many relays for the one-way model than the shared relay model, they are given 1/3 the transmission power and 1/3 the antennas.

BS TX power	P_{BS}
Relay TX power	$P_{\text{RS}}/3$
Antennas per BS (sector)	1
Antennas per relay	1
Relays per sector	1
Antennas per mobile	1
Relay location	2/3 cell radius from BS

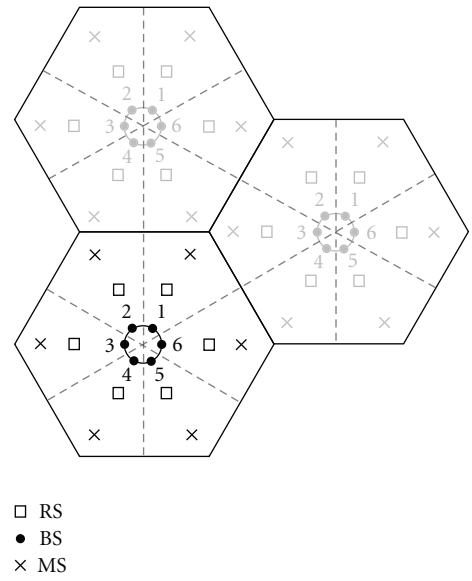


Figure 3: Base system model for two-way relaying. Each sector contains one single-antenna amplify-and-forward relay, and there is no coordination between cells. The sectors in a given cell may cooperate to decode the uplink signals from the users in the cell but do not cooperate in the downlink.

ith MS to the ith BS. The relays are designed to facilitate the downlink transmission of $\mathbf{s}$ and the uplink transmission of $\mathbf{u}$ (where $\mathbf{u} = [u_1 u_2 \cdots]^T$ is the vector of transmitted symbols from each mobile and similarly for $\mathbf{s}$ simultaneously over two time slots, avoiding the half-duplex loss of one-way relaying. We shall refer to this simultaneous uplink-downlink transmission as one complete transmission cycle.

In this section we consider the case where the relays are utilized as bidirectional terminals, a configuration also known as two-way relaying. Consider a single physical layer frame in IEEE 802.16j [25]. There are four distinct parts of the frame: (1) the base station transmits in the downlink, then (2) the relay transmits in the downlink, then (3) the mobile transmits in the uplink, and then (4) the relay transmits in the uplink. In two-way relaying this transmission cycle would be cut in half. That is, parts (1) and (3) could take place simultaneously in one segment of the frame, and parts (2) and (4) could take place simultaneously in the rest of the frame. During the first time slot (phase I) all information-generating nodes in the cell (BSs and MSs) transmit their signals to the relay. In the second time slot (phase II), and after proper processing, the RSs broadcast symbols from which the network nodes, that is, BSs and MSs, may extract their intended signals. This two-phase operation is shown in Figure 4.

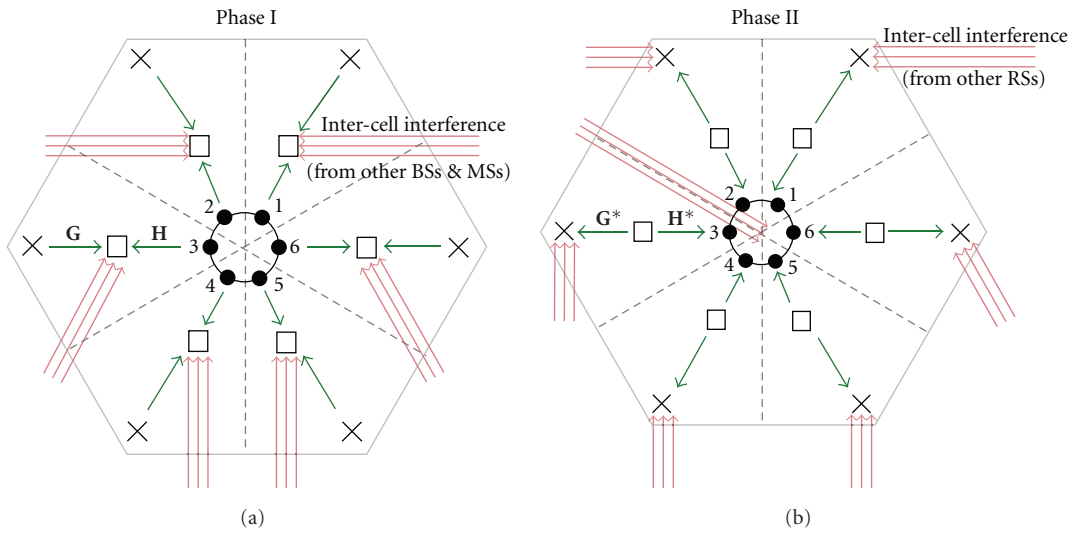


Figure 4: Two-way relaying operation in a single cell. In the first phase, all transceivers transmit except the relays. In the second phase, only relays transmit, and other transceivers are able to cancel the interference they caused in the first phase.

Phase I. We consider the signals from each relay in the sector since the base station can utilize all antennas in all sectors to decode the uplink. Using Gaussian codebooks, the BSs and MSs transmit $\mathbf{s}$ and $\mathbf{u}$, respectively. Denote by $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{G}}$ the channels from the base station array and mobile stations to the relays, respectively. The received signal at the relays in the cell of interest is then

$$\mathbf{y}_R = \mathbf{H}\mathbf{s} + \mathbf{G}\mathbf{u} + \mathbf{H}_{IC}\mathbf{s}_{IC} + \mathbf{G}_{IC}\mathbf{u}_{IC} + \mathbf{v}_R, \tag{7}$$

where for the reuse pattern of Figure 2, $\mathbf{H}$ and $\mathbf{G}$ contain only the diagonals of $\tilde{\mathbf{H}}$ and $\tilde{\mathbf{G}}$. $\mathbf{H}_{IC}$ is the channel from base stations serving other cells to each relay, $\mathbf{G}_{IC}$ is the channel from mobiles in other cells, and $\mathbf{v}_R$ is zero-mean additive white Gaussian noise at the relay with variance σ_N^2. The subscript IC refers to intersector interference, whereas (as in previous sections) the subscript R refers to the relay, and N refers to noise. Further, transmit powers have been absorbed into the channels as in previous sections. Finally, the channels $\mathbf{H}$ and $\mathbf{G}$ may have some zero entries depending on the frequency reuse factor of the network, but the analysis is general to any reuse factor.

Phase II. Under a nonregenerative assumption, the output of each RS is a scaled version of the input $\hat{\mathbf{y}}_R = \mathbf{\Gamma}\mathbf{y}_R$ where $\mathbf{\Gamma}$ is a diagonal matrix determined by the power constraint $\mathbb{E}\{\hat{\mathbf{y}}_R\hat{\mathbf{y}}_R^*\} = \mathbf{I}$ (since transmit powers are absorbed into the channels). Since we allow the BS antennas to cooperate in decoding the uplink, we analyze the entire received signal at the BS array:

$$\begin{aligned}\mathbf{y}_B &= \tilde{\mathbf{H}}^*\hat{\mathbf{y}}_R + \mathbf{W}_{IC}\hat{\mathbf{y}}_{R,IC} + \mathbf{v}_B \\ &= \tilde{\mathbf{H}}^*\mathbf{\Gamma}(\mathbf{H}\mathbf{s} + \mathbf{G}\mathbf{u} + \mathbf{H}_{IC}\mathbf{s}_{IC} + \mathbf{G}_{IC}\mathbf{u}_{IC} + \mathbf{v}_R) \\ &\quad + \mathbf{W}_{IC}\hat{\mathbf{y}}_{R,IC} + \mathbf{v}_B,\end{aligned} \tag{8}$$

where $\tilde{\mathbf{H}}$ was defined before, $\mathbf{W}_{IC}$ is the matrix channel from relays in other cells to the base station, $\hat{\mathbf{y}}_R$ is the amplified signal from all the relays in the cell, $\hat{\mathbf{y}}_{R,IC}$ is the amplified signal from relays in other cells, and the subscript B denotes that reception is at the base station. The spatial covariance of the interference and noise at the base station is then

$$\begin{aligned}\mathbf{R}_{IN} &= \mathbf{H}^*\mathbf{\Gamma}(\mathbf{H}_{IC}\mathbf{H}_{IC}^* + \mathbf{G}_{IC}\mathbf{G}_{IC}^* + \sigma_N^2\mathbf{I})\mathbf{\Gamma}\tilde{\mathbf{H}} \\ &\quad + \mathbf{W}_{IC}\hat{\mathbf{y}}_{R,IC}\hat{\mathbf{y}}_{R,IC}^*\mathbf{W}_{IC}^* + \sigma_N^2\mathbf{I}.\end{aligned} \tag{9}$$

Note that the term $\hat{\mathbf{y}}_{R,IC}$ has information about the Phase-I signals transmitted in the cell of interest even though it is an interference term. In fact, if the channels to nodes in other cells were estimated, these terms could be canceled. However, we will assume only in-cell channel state information in this paper. Since the base station can cancel the terms that explicitly contain $\mathbf{s}$, the uplink sum rate for the whole cell is

$$R_{\mathrm{TW,UL}} = \frac{1}{2}\log\left|\mathbf{I} + \mathbf{R}_{IN}^{-1}\tilde{\mathbf{H}}^*\mathbf{\Gamma}\mathbf{G}\mathbf{G}^*\mathbf{\Gamma}\tilde{\mathbf{H}}\right|, \tag{10}$$

where subscript TW denotes two-way relaying, and UL denotes the uplink. The rate for any given user can be computed from this using the multiple access rates as given in Section 5.

For the downlink, the users cannot cooperatively decode, and thus we can compute the rate for the user in the sector of interest. This user will receive

$$y_M = g\hat{y}_R + \mathbf{q}_{IS}^*\hat{\mathbf{y}}_{R,IS} + \mathbf{q}_{IC}^*\hat{\mathbf{y}}_{R,IC} + v_M, \tag{11}$$

where $\mathbf{q}_{IS}$ is the vector channel from the other-sector relays to the user, $\mathbf{q}_{IC}$ is the vector from other-cell relays to the user, and v_M is the noise with variance σ_N^2. Note that we distinguish between the channels between other-cell mobiles and the relays of interest $\mathbf{G}_{IC}$, and the channels between other-cell *relays* and the *mobile* of interest $\mathbf{q}_{IC}$. Note also that $\hat{\mathbf{y}}_{R,IS}$ and $\hat{\mathbf{y}}_{R,IC}$ have information about both the uplink and downlink signal. In particular, with the proper CSI, the

mobile could cancel its signal from $\hat{\mathbf{y}}_{R,IS}$ and similarly use what is available of the downlink signal in these terms to help decode; however, we will not assume this complexity in this paper. The interference variance is then

$$\sigma_I^2 = \left|\mathbf{q}_{IS}^* \hat{\mathbf{y}}_{R,IS}\right|^2 + \left|\mathbf{q}_{IC}^* \hat{\mathbf{y}}_{R,IC}\right|^2 + |g|^2 \|\mathbf{h}_I\|^2 + |g|^2 \|\mathbf{g}_I\|^2, \tag{12}$$

where $\mathbf{h}_I$ is the vector channel of interferers seen by the relay in Phase I (relative to the downlink transmitted symbol $\mathbf{s}$), and $\mathbf{g}_I$ is the channel of interferers seen by the relay in Phase I (relative to the uplink transmitted symbol $\mathbf{u}$). Thus, the downlink rate for this user is

$$r_{\mathrm{TW,DL}} = \frac{1}{2} \log\left(1 + \frac{|gh|^2}{\sigma_I^2 + \sigma_N^2}\right). \tag{13}$$

We use the notation r instead of R to refer to a single user rather than the sum over users.

The main assumptions and parameters for the two-way model are identical to those for the one-way model and are given in Table 1.

5. Shared Relaying

A *shared* relay is a relay that is the subordinate of multiple base stations—the base stations share the relay. As discussed in Section 3, IEEE 802.16j does not permit this architecture, but shared relaying has distinct advantages over the one-way model. The relay has KM antennas, where M is the number of base station antennas serving each sector, and K is the number of base stations sharing the relay. For simplicity in our analysis, $M = 1$, but the model is readily extendable to $M > 1$. Figure 5 shows a typical configuration for a shared relay under the general cellular model presented in Section 2. The relay is placed at the corner of three adjacent cells (hence $K = 3$, so that each base station has a sector pointing directly at the shared relay).

By placing many antennas at the shared relay, interference can be canceled in both hops of communication. The shared relay behaves as a coordination of many single-antenna relays and thus alleviates the need for coordination among base stations. As will be shown in Section 7, the shared relay achieves much of the capacity gain of base station coordination without the need for expensive information-passing between distributed base stations.

As in the one-way model, downlink communication occurs in two time slots (since we assume no base station coordination, even among sectors, the uplink analysis is identical to that of the downlink with lower transmit power at the mobile). In the first hop, the relay receives

$$\mathbf{y}_R = \sum_{k=1}^{K} \mathbf{h}_k s_k + \mathbf{H}_I \mathbf{s}_I + \mathbf{v}_R, \tag{14}$$

where $\mathbf{h}_k$ is the channel from the kth parent base station to the relay, s_k is the symbol transmitted by the kth base station (intended for the kth user being served by the shared relay), $\mathbf{H}_I$ is the matrix of channel coefficients from interfering base stations, $\mathbf{s}_I$ is the vector of symbols transmitted by the interferers, and $\mathbf{v}_R$ is spatially white zero-mean additive white Gaussian noise at the relay.

This first hop of communication is the MIMO multiple access channel, and its capacity can be achieved via multiuser detection at the relay. That is, no coordination is necessary among the base stations beyond frame synchronization. Assuming, without loss of generality, that the users are ordered relative to channel SNR (i.e., $\|\mathbf{h}_1\| > \|\mathbf{h}_2\| > \cdots > \|\mathbf{h}_K\|$), we will decode s_1 first, and so on, so that s_k is decoded in the midst of interference from only the $(k+1)$ through Kth streams (and the term $\mathbf{H}_I \mathbf{s}_I$ which is common to all streams). Then the mutual information for user k in the first hop is

$$R_{1k} = \log\left|\mathbf{I} + \mathbf{A}_k^{-1} \mathbf{R}_{I1}^{-1} \mathbf{h}_k \mathbf{h}_k^*\right|, \tag{15}$$

where $\mathbf{R}_{I1} = \mathbf{H}_I \mathbf{H}_I^* + \sigma_N^2 \mathbf{I}$ and $\mathbf{A}_k$ is defined recursively as

$$\begin{aligned} \mathbf{A}_k &= \mathbf{I} + \mathbf{A}_{k+1}^{-1} \mathbf{R}_{I1}^{-1} \mathbf{h}_{k+1} \mathbf{h}_{k+1}^*, \\ \mathbf{A}_K &= \mathbf{I}. \end{aligned} \tag{16}$$

Now that the relay has decoded the first hop, it can transmit the $\{s_k\}$ to the mobiles in the second hop at a different rate than the first hop. It thus re-encodes the $\{s_k\}$ into another vector $\{x_k\}$ at the highest rate the second hop can support. Note that this is the Gaussian MIMO broadcast channel, and its capacity can be achieved by performing an LQ factorization on the aggregate channel matrix, performing dirty paper coding on the interfering signals, and waterfilling over the signals [41]. The user receives only its signal from the relay, plus interference from the external interferers. This is modeled as

$$y_{M,k} = g_k x_k + \mathbf{g}_{I,k}^* \mathbf{x}_I + v_{M,k}, \tag{17}$$

where g_k is the effective channel after precoding, waterfilling, and dirty paper coding between the relay and the kth mobile station, $\mathbf{g}_{I,k}$ is the vector channel from all the interferers to the kth mobile, $\mathbf{x}_I$ is the transmitted vector at the interferers during the second hop, and $v_{M,k}$ is the additive white Gaussian noise at mobile k.

For user k the rate in the second hop is

$$R_{2k} = \log\left(1 + \frac{|g_k|^2}{\|\mathbf{g}_{I,k}\|^2 + \sigma_N^2}\right). \tag{18}$$

As in Section 3, we must optimize the time sharing between the two hops. In this case however, we have to optimize the sum rate and cannot optimize the rate for each user. The sum rate is

$$R_S = \max_{t \in [0,1]} \sum_{k=1}^{K} \min\{tR_{1k}, (1-t)R_{2k}\}. \tag{19}$$

Here we use the subscript S to denote shared relaying. The main assumptions and parameters for the shared model are given in Table 2.

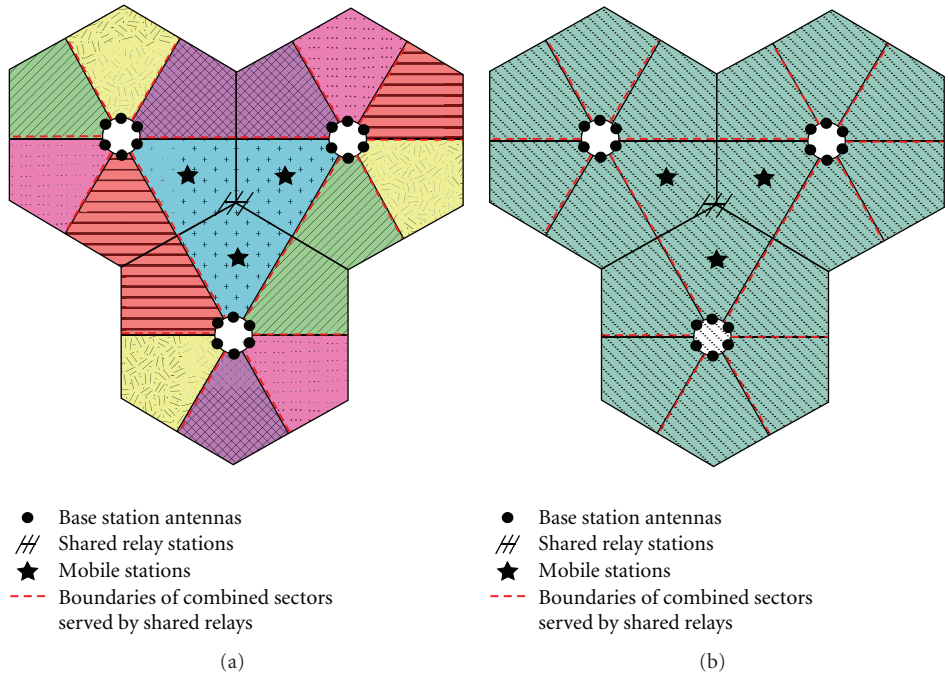


Figure 5: Models of systems using shared relays with (a) frequency reuse factor of 6 or (b) frequency reuse factor of 1.

Table 2: System parameters for shared relay model. The main differences between the shared relay model and the one-way relay are the number of antennas per relay, the relay transmit power, and the number of relays per sector. Since over a large network there will be approximately 3 times fewer relays for the shared model than the one-way relay model, shared relays are given 3 times the transmission power and 3 times the antennas.

BS TX power	P_{BS}
Relay TX power	P_{RS}
Antennas per BS (sector)	1
Antennas per relay	3
Relays per sector	1
Antennas per mobile	1
Relay location	cell radius from BS

6. Base Station Coordination

Base station coordination allows distributed base stations to act as a single multiantenna transmitter by sharing the data to be transmitted via a high-capacity low-delay wired backbone [34]. If all base stations can coordinate their transmissions to all scheduled users, then all interference can be removed. However, full coordination over a wide area is impractical because of the complexity of coordinated transmission, and so localized coordination has been investigated recently [42]. Here, to give an interesting comparison to the shared relay, we allow coordination of sectors pointing at each other at each of the corners of the cells, as shown in Figure 6. No relaying is performed under this architecture. We assume a sum power constraint for all the coordinated antennas. Although this assumption is not practical, the pooled power constraint is a very close approximation to the per-base power constraint, with much lower complexity in calculation [43, 44].

As this channel model is again the Gaussian MIMO broadcast channel, the user rates are similar to those achieved in the second hop of the shared relay transmission in Section 5. Mobile k receives

$$y = h_k s_k + \mathbf{h}_{I,k}^{*} \mathbf{s}_I + v_k, \tag{20}$$

where h_k is the effective channel gain from the base stations to the kth mobile after precoding, dirty paper coding, and waterfilling, s_k is the transmitted symbol intended for the kth mobile, $\mathbf{h}_{I,k}$ is the vector channel from the interferers to the kth mobile, $\mathbf{s}_I$ is the vector of symbols transmitted by the interferers, and v_k is the additive white Gaussian noise at the kth mobile. The rate for user k is thus

$$r_{k,BC} = \log\left(1 + \frac{|h_k|^2}{||\mathbf{h}_{I,k}||^{\mathbf{h}_{I,k}} + \sigma_N^2}\right). \tag{21}$$

Here we have used the subscript BC to denote base station coordination and the notation r instead of R to refer to a single user rather than the sum of users. The rate in (21) is the rate of K users in K sectors and is thus directly comparable to (19) assuming that the services areas are the same for the two cases. For the uplink, the rates are that for the MIMO multiple access channel (MIMO MAC), whose forms are identical to those for the downlink but for the proper uplink channel substituted for h_k and the interfering channels [45]. The base station parameters for this model are the same as previous models, and there are no relays included in this model.

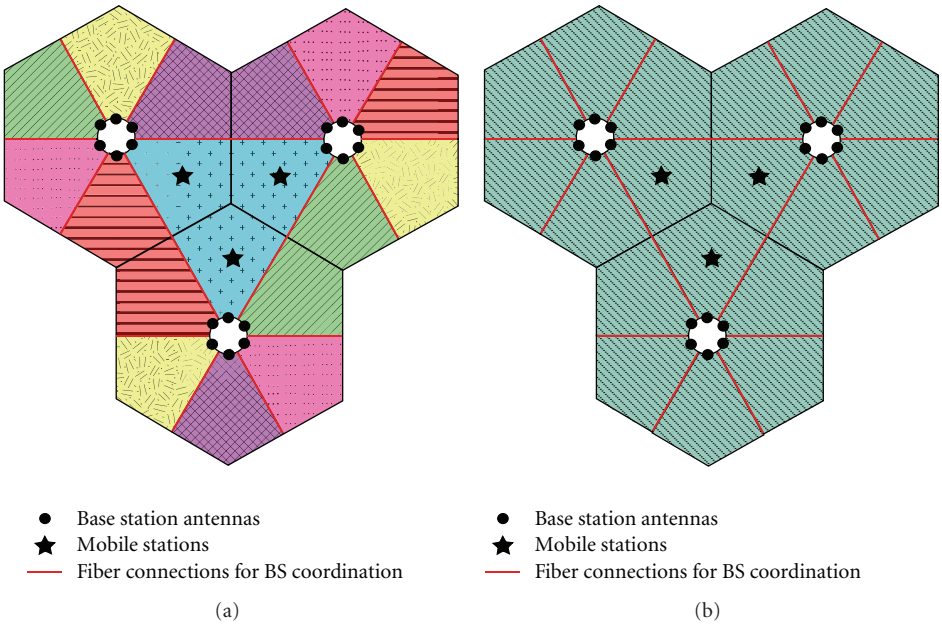


Figure 6: System models for base station coordination with (a) frequency reuse factor of 6 or (b) frequency reuse factor of 1.

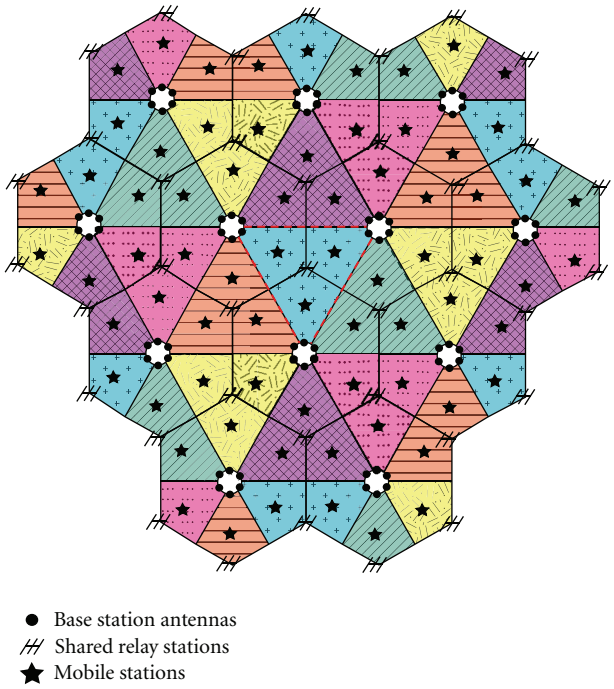


Figure 7: System model under consideration for the simulations presented in this paper. The focus is on the triangular area in the center of the network. This figure also shows the frequency reuse pattern for the shared relay and base station coordination under reuse factor 6.

7. Simulations

Each of the systems described in the previous four sections was tested under a system-level cellular network simulation. A layer of interfering cells was wrapped around the three main cells, as shown in Figure 7. These outer cells have the same architecture as the inner cells for the respective simulations. For instance, a network implementing the shared relay will contain a relay at each vertex of each hexagonal cell, as in Figure 7. Since the sectors making up the central triangle are our area of interest, there are actually two layers of interfering relays in this case.

Table 3: System parameters used for the simulations in this paper.

BS TX power	47 dBm
BS-RS channel model	IEEE 802.16j, Type H [33]
BS-MS channel model	IEEE 802.16j, Type E [33]
RS-MS channel model	IEEE 802.16j, Type E [33]
Number of Realizations	1000
Cell radius	876 m
Carrier frequency	2 GHz
Noise power	−144 dBW
Mobile height	1 m
Relay height	15 m
BS height	30 m
Propagation environment	Urban

The metric of comparison is the achievable sum rate (derived in each architecture's respective section) in the central triangle outlined in Figure 7. That is, the sum rate is the rate of the three users in the three sectors making up the central triangle in Figure 7, averaged over a number of fading and shadowing iterations. Since we have assumed arbitrary scheduling and orthogonal signaling inside each sector (corresponding to a single subchannel of the OFDM waveform), the sum rate is calculated over three users. The parameters of the simulation are given in Table 3.

The Type H channel model specifies a channel from a node transmitting from above the roofline to another node above the roofline. The fading is Rician with K-factor 4, the carrier frequency is 2 GHz, there is no shadowing, the relay height is 15 m, and the base station height is 30 m. For the Type E channel model, for the BS-MS and RS-MS links, the mobile is located 1 m above the ground, the street width is 12 m, the roof height is 15 m, and the distance between building centers is 60 m (based on an urban environment). The noise power is −144 dBW, corresponding to a 10 MHz channel.

Figure 8 shows the downlink sum rate for each of the architectures presented in this paper as a function of relay transmit power for reuse factors $r = 1, 6$. For each case, $r = 1$ outperforms $r = 6$ to varying degree. Base station coordination and conventional transmission are constant across the plot because no relays are included in these system models.

Base station coordination, unsurprisingly, gives the highest downlink sum rates, a roughly 119% increase over a conventional architecture with no relaying or coordination. More striking, however, is that shared relaying achieves approximately 60% of the gains of base station coordination. When comparing the two systems, it must be emphasized that shared relaying requires no coordination between its base stations beyond that needed for synchronization in the multiple access channel of the first hop. Its main disadvantage relative to coordination is the half-duplex loss and delay associated with decode-and-forward relaying. Note that for $r = 6$ the gains of shared relaying diminish relative to $r = 1$.

The one-way architecture only gives a roughly 15% increase in rate relative to a conventional system, whereas two-way relaying performs worse than conventional in the regime plotted in Figure 8. Here, the multiplexing gain of the two-way relay is not apparent because we are considering only the downlink.

Uplink sum rates are given in Figure 9. In this regime, conventional architectures (without power control, soft handoff, or multiuser diversity which have been abstracted out of the system) have extremely low uplink SINR, resulting in almost no rate. Two-way relaying performs similarly since the interference from nearby base stations is overwhelming the mobile device's signal unless the relay is extremely close to it (as will be discussed in the next section). The curves on this graph are flat partly because they are already in the interference-limited regime and partly because, in the case of relaying, the system is limited by the first hop, which is not a function of the relay transmit power.

In this regime, shared relaying achieves around 90% of the achievable rate of base station coordination due to the relay's ability to remove interference and its proximity to the cell edge. The half-duplex loss is much less severe in this case. One-way relaying achieves roughly 50% of the rates of base station coordination. As in the downlink case, frequency use factor $r = 1$ drastically outperforms $r = 6$ across the board.

Figure 10 shows the downlink sum rate of coordination, shared relaying, and a conventional system with no relaying or coordination throughout an entire sector. The figure is

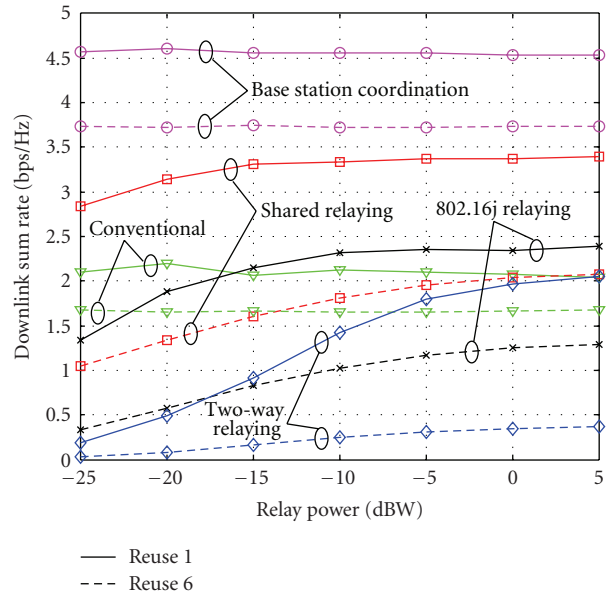


Figure 8: Downlink sum rates for each of the strategies presented in this paper as a function of the relay transmit power. The solid lines represent reuse factor 1, while the dotted lines represent reuse factor 6.

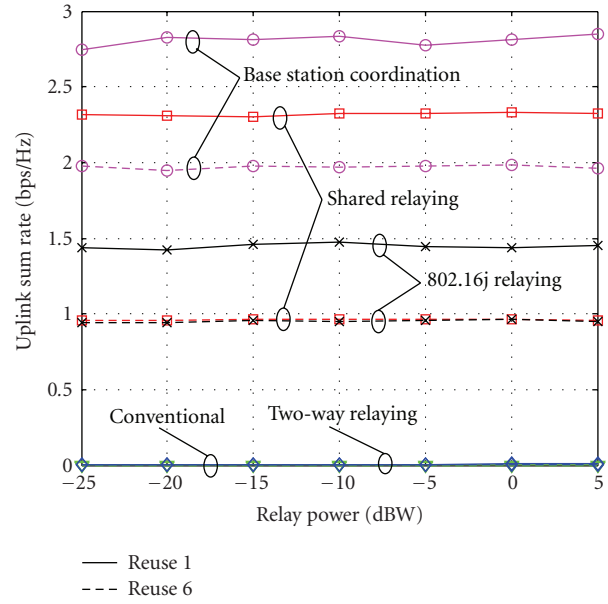


Figure 9: Uplink sum rates for each of the strategies presented in this paper as a function of the relay transmit power. The solid lines represent reuse factor 1, while the dotted lines represent reuse factor 6.

for frequency reuse factor 6 because the curves are more separated in this case. At around half-way between the base station and shared relay (which is located at the left-most corner of the sector), direct transmission becomes more desirable than relaying. By adapting between these two cases based on the position of the mobile station, the downlink rate

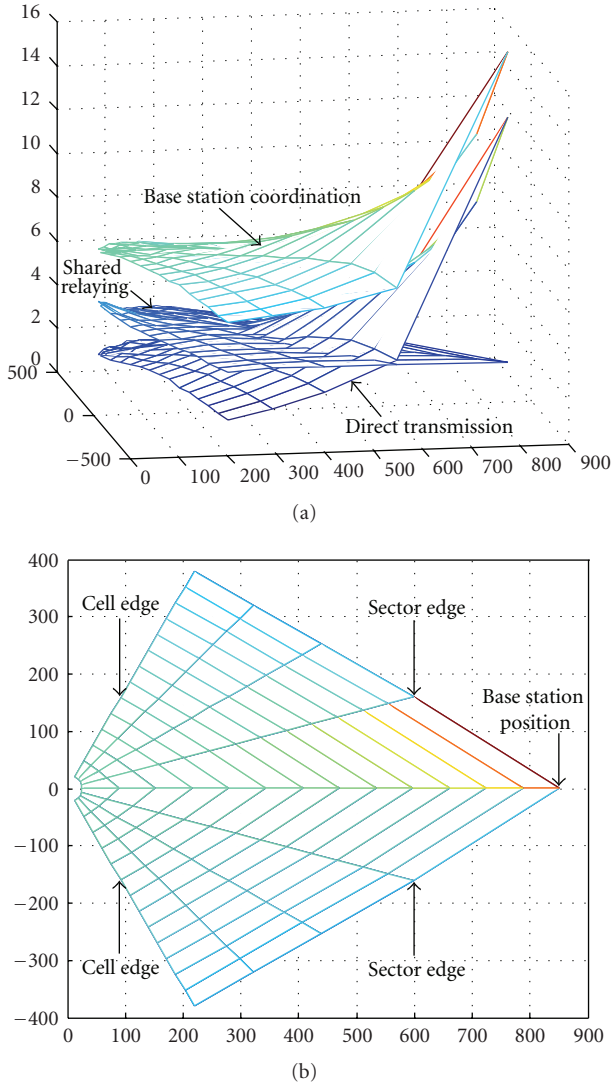


Figure 10: (a) Downlink sum rate in one sector versus mobile station position for base station coordination, shared relaying, and direct transmission. A reuse factor 6 is shown because the curves are more separated in this case. By adapting between shared relaying and direct transmission depending on user location, the rates of base station coordination can be approached. (b) The geometry of the sector, explaining the x- and y-axes of part (a).

approaches that of base station coordination over the entire cell.

The simulations of the this section give relative performance gains between different transmission strategies in a cellular network. This section describes the insight these simulations can give and summarizes the general conclusions we can draw from them beyond the relative performances. First, having a relay act as an interference-reducing station gets nearly the gains of having BS coordination over the same area. The reason this is not obvious is because of the half-duplex nature of the relay. This is made up for by the fact that the relay can be placed in an LOS position with the BS and is closer to the MS than the BS in the regime of interest. In more precise terms, the degrees of freedom lost in performing half-duplex relaying are almost made up for by practical considerations such as RS placement, all at a reduced complexity. The second conclusion we can draw is that two-way relaying is severely limited in the uplink unless the relay is extremely close to the mobile and does not in general compensate for the half-duplex loss of one-way relaying in the simulated regime. We will discuss practical ways of overcoming this problem in the next section.

8. Discussion

In the previous section, shared relaying was shown to be a simpler alternative to base station coordination. Further, by spatially removing local interference, the shared relay outperforms one-way relaying by over 80% in the downlink. By allowing the relay to be shared among multiple base stations, the shared relay avoids the BS coordination task of associating each mobile station with multiple base stations. We now briefly discuss some practical considerations for shared relaying.

8.1. Practical Shared Relaying. We have been assuming thus far that the shared relay is moderately complex. Since it serves 3 adjacent sectors, there will be 1/3 as many relays in the network than with the one-way model (neglecting the edge of the network). Thus, an increase in unit complexity is at least partially offset by a decrease in deployment cost relative to the one-way model.

The shared relay may also mitigate the need for coordinated scheduling between the sectors. If the shared relay is allowed to transmit its own control information, as in the nontransparent relay of IEEE 802.16j [25], it can achieve a large multiuser diversity gain across sectors without the need for the base stations to share information.

It may also make handoff easier by allowing for a buffer zone where which base station a mobile is associated with is unimportant. For example, consider a mobile station moving away from a base station and toward a shared relay. As it enters the relay's zone of service, it is now served by this relay but still associated with its original base station. As it continues past the relay and into the next cell, it is still served by the shared relay, which may signal to the original base station that it is time to handoff the mobile to the adjacent BS. So long as the handoff procedure is done before the mobile leaves the shared relay's zone of service, the mobile will stay connected to the network.

8.2. Improving Two-Way Relaying. Recall that Figure 9 showed that uplink rates for two-way relaying were practically zero. In this scenario, since the base stations and mobile stations are transmitting simultaneously, nearby base stations are drowning out the mobile stations. This can be mitigated by only performing two-way relaying for mobiles that are very near the relay. Figure 11 shows the uplink sum rate for various transmission strategies as a function of the mobile station distance from the base station. Conversely, Figure 12 shows the downlink sum rate for the

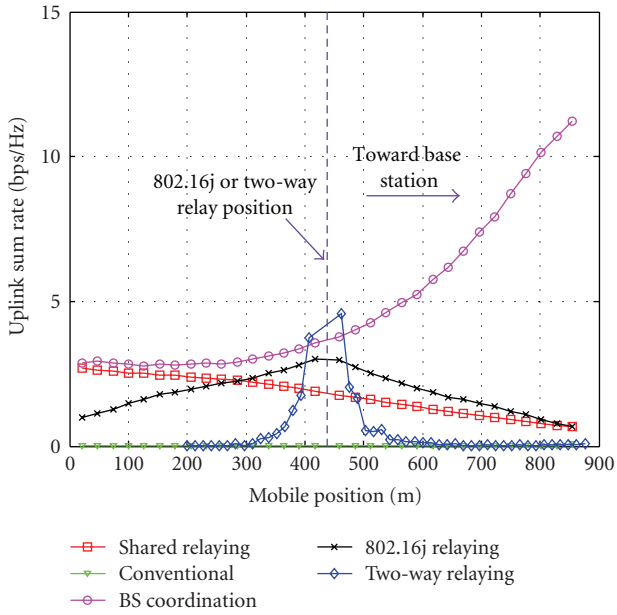


Figure 11: Uplink sum rate of two-way relaying and other strategies versus MS position relative to cell edge. The relay station is located 440 m from the base station.

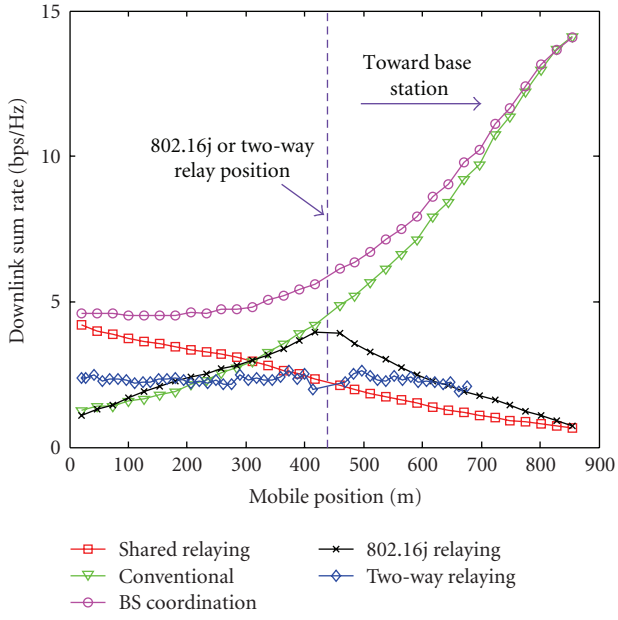


Figure 12: Downlink sum rate of two-way relaying and other strategies versus MS position relative to cell edge. The relay station is located 440 m from the base station.

same strategies. In this case, the relay station is located 440 m from the cell edge. The mobile's power begins to overshadow the adjacent BS interference power at around 100 m from the relay, and the sum rate quickly rises.

Two-way relaying aims to increase the sum uplink plus downlink rates relative to conventional relaying. However, in a mobile broadband cellular network, the uplink and downlink are inherently asymmetric, making this sum an inappropriate metric. For instance, to truly maximize the uplink plus downlink rate, one will simply allow the downlink to occur all the time.

Further, allowing adjacent base stations and mobile stations to transmit simultaneously is an inherently bad idea unless the receiver is located very close to the mobile. For example, if we allow the mobile to transmit at 23 dB below the base station power, and using simple free-space path loss, the relay would have to be approximately 36 times closer to the mobile than the nearest out-of-cell base station for a 0 dB SINR. Of course, this is a simple calculation intended only to show the nature of the problem.

One way of combating this is to use an antenna array at the relay to steer nulls toward the nearest base stations. This risks a mobile being in the same direction as the base station and being in the same null. Other strategies include conventional ways of avoiding interference in cellular systems such as power control and frequency reuse.

9. Conclusions and Future Work

We have analyzed and compared four cellular architectures for LTE-Advanced. While base station coordination between adjacent sectors in neighboring cells achieved the highest rates, it is also the most complex architecture. Sharing a multiantenna relay among the same sectors is a simpler way to achieve much of the gains of local interference mitigation but still has significant complexity within the relay itself. One-way relaying, where each relay is associated with only one base station, is unlikely to give substantial throughput gains near the cell edge because it does not directly treat interference, and two-way relaying overcomes the half-duplex loss of conventional relaying provided that the relay is extremely close to the mobile.

Future work will focus on more detailed design of shared relays, including scheduling, feedback, and dealing with mobility. Two-way relaying requires research for interference mitigation in the uplink. Finally, combining base station coordination and relaying is an emerging area that will be the subject of future research [46–50].

Acknowledgment

This work was supported by a gift from Huawei Technologies, Inc.

References

[1] S. Parkvall, E. Dahlman, A. Furuskar, et al., "LTE-advanced—evolving LTE towards IMT-advanced," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 1–5, September 2008.

[2] H. Yanikomeroglu, "Cellular multihop communications: infrastructure-based relay network architecture for 4G wireless systems," in *Proceedings of the 22nd Biennial Symposium on Communications*, Queen's University, Kingston, Canada, June 2004.

[3] R. Pabst, B. H. Walke, D. C. Schultz, et al., "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Communications Magazine*, vol. 42, no. 9, pp. 80–89, 2004.

[4] L. Le and E. Hossain, "Multihop cellular networks: potential gains, research challenges, and a resource allocation framework," *IEEE Communications Magazine*, vol. 45, no. 9, pp. 66–73, 2007.

[5] O. Oyman, N. J. Laneman, and S. Sandhu, "Multihop relaying for broadband wireless mesh networks: from theory to practice," *IEEE Communications Magazine*, vol. 45, no. 11, pp. 116–122, 2007.

[6] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: a new architecture for wireless communications," in *Proceedings of the 19th Annual Joint Conference of IEEE Computer and Communications Societies (INFOCOM '00)*, vol. 3, pp. 1273–1282, Tel Aviv, Israel, March 2000.

[7] H. Wu, C. Qiao, S. De, and O. Tonguz, "Integrated cellular and ad hoc relaying systems: iCAR," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 10, pp. 2105–2115, 2001.

[8] V. Sreng, H. Yanikomeroglu, and D. Falconer, "Coverage enhancement through two-hop relaying in cellular radio systems," in *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC '02)*, vol. 2, pp. 881–885, March 2002.

[9] A. Bletsas, A. Khisti, D. P. Reed, and A. Lippman, "A simple cooperative diversity method based on network path selection," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 659–672, 2006.

[10] M. Qin and R. S. Blum, "Capacity of wireless ad hoc networks with cooperative diversity: a warning on the interaction of relaying and multi-hop routing," in *Proceedings of the IEEE International Conference on Communications*, vol. 2, pp. 1128–1131, May 2005.

[11] G. Scutari, S. Barbarossa, and D. Ludovici, "Cooperation diversity in multihop wireless networks using opportunistic driven multiple access," in *Proceedings of the IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC '03)*, pp. 170–174, June 2003.

[12] E. H. Drucker, "Development and application of a cellular repeater," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '88)*, pp. 321–325, June 1988.

[13] O. Muñoz-Medina, J. Vidal, and A. Agustín, "Linear transceiver design in nonregenerative relays with channel state information," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2593–2604, 2007.

[14] X. Tang and Y. Hua, "Optimal design of non-regenerative MIMO wireless relays," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1398–1406, 2007.

[15] S. W. Peters and R. W. Heath Jr., "Nonregenerative MIMO relaying with optimal transmit antenna selection," *IEEE Signal Processing Letters*, vol. 15, pp. 421–424, 2008.

[16] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," *IEEE Transactions on Information Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[17] G. Kramer, M. Gastpar, and P. Gupta, "Cooperative strategies and capacity theorems for relay networks," *IEEE Transactions on Information Theory*, vol. 51, no. 9, pp. 3037–3063, 2005.

[18] D. Chen and J. N. Laneman, "Modulation and demodulation for cooperative diversity in wireless systems," *IEEE Transactions on Wireless Communications*, vol. 5, no. 7, pp. 1785–1794, 2006.

[19] K. Azarian, H. El Gamal, and P. Schniter, "On the achievable diversity-multiplexing tradeoff in half-duplex cooperative channels," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4152–4172, 2005.

[20] B. Rankov and A. Wittneben, "Spectral efficient signaling for half-duplex relay channels," in *Proceedings of the Asilomar Conference on Signals, Systems and Computers*, pp. 1066–1071, November 2005.

[21] B. Wang, J. Zhang, and A. Høst-Madsen, "On the capacity of MIMO relay channels," *IEEE Transactions on Information Theory*, vol. 51, no. 1, pp. 29–43, 2005.

[22] C. K. Lo, S. Vishwanath, and R. W. Heath Jr., "Rate bounds for MIMO relay channels," *Journal of Communications and Networks*, vol. 10, no. 2, pp. 194–203, 2008.

[23] C.-B. Chae, T. Tang, R. W. Heath Jr., and S. Cho, "MIMO relaying with linear processing for multiuser transmission in fixed relay networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 2, pp. 727–738, 2008.

[24] Y. Zhu and H. Zheng, "Understanding the impact of interference on collaborative relays," *IEEE Transactions on Mobile Computing*, vol. 7, no. 6, pp. 724–736, 2008.

[25] S. W. Peters and R. W. Heath Jr., "The future of WiMAX: multi-hop relaying with IEEE 802.16j," *IEEE Communications Magazine*, vol. 1, no. 47, 2009.

[26] K. Doppler, C. Wijting, and K. Valkealahti, "On the benefits of relays in a metropolitan area network," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '08)*, pp. 2301–2305, May 2008.

[27] R. Schoenen, W. Zirwas, and B. H. Walke, "Raising coverage and capacity using fixed relays in a realistic scenario," in *Proceedings of the 14th European Wireless Conference (EW '08)*, pp. 1–6, June 2008.

[28] R. Irmer and F. Diehm, "On coverage and capacity of relaying in LTE-advanced in example deployments," in *Proceedings of the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, pp. 1–5, September 2008.

[29] Y. Song, et al., "Relay station shared by multiple base stations for inter-cell interference mitigation," IEEE C802.16m-08/1436r1, November 2008.

[30] S. Katti, S. Gollakota, and D. Katabi, "Embracing wireless interference: analog network coding," in *Proceedings of the ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM '07)*, pp. 397–408, ACM, 2007.

[31] S. J. Kim, N. Devroye, P. Mitran, and V. Tarokh, "Comparison of bi-directional relaying protocols," in *Proceedings of the IEEE Sarnoff Symposium (SARNOFF '08)*, pp. 1–5, April 2008.

[32] A. S. Avestimehr, A. Sezgin, and D. N. C. Tse, "Approximate capacity of the two-way relay channel: a deterministic approach," in *Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1582–1589, September 2008.

[33] G. Senarath, et al., "Multi-hop relay system evaluation methodology (channel model and performance metric)," IEEE 802.16j-06/013r3, February 2007.

[34] G. J. Foschini, K. Karakayali, and R. A. Valenzuela, "Coordinating multiple antenna cellular networks to achieve enormous spectral efficiency," *IEE Proceedings: Communications*, vol. 153, no. 4, pp. 548–555, 2006.

[35] S. Jing, D. N. C. Tse, J. B. Soriaga, J. Hou, J. E. Smee, and R. Padovani, "Multicell downlink capacity with coordinated processing," *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, 2008.

[36] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Communications*, vol. 13, no. 4, pp. 56–61, 2006.

[37] H. Zhang and H. Dai, "Cochannel interference mitigation and cooperative processing in downlink multicell multiuser MIMO networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 2004, no. 2, pp. 222–235, 2004.

[38] D. Gesbert, S. G. Kiani, A. Gjendemsjø, and G. E. Øien, "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2393–2409, 2007.

[39] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath Jr., "Networked MIMO with clustered linear precoding," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, pp. 1910–1921, 2009.

[40] A. Chakrabarti, A. Sabharwal, and B. Aazhang, "Sensitivity of achievable rates for half-duplex relay channel," in *Proceedings of the 6th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC '05)*, pp. 970–974, June 2005.

[41] S. Shamai and B. M. Zaidel, "Enhancing the cellular downlink capacity via co-processing at the transmitting end," in *Proceedings of the IEEE Vehicular Technology Conference (VTC '01)*, vol. 3, pp. 1745–1749, 2001.

[42] H. Huang and M. Trivellato, "Performance of multiuser MIMO and network coordination in downlink cellular networks," in *Proceedings of the 6th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt '08)*, pp. 85–90, April 2008.

[43] S. A. Jafar and A. J. Goldsmith, "Transmitter optimization for multiple antenna cellular systems," in *Proceedings of the IEEE International Symposium on Information Theory*, p. 50, 2002.

[44] M. K. Karakayali, G. J. Foschini, R. A. Valenzuela, and R. D. Yates, "On the maximum common rate achievable in a coordinated network," in *Proceedings of the IEEE International Conference on Communications*, vol. 9, pp. 4333–4338, June 2006.

[45] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*, Cambridge University Press, Cambridge, UK, 2007.

[46] O. Simeone, O. Somekh, Y. Bar-Ness, and U. Spagnolini, "Throughput of low-power cellular systems with collaborative base stations and relaying," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 459–467, 2008.

[47] O. Somekh, O. Simeone, H. V. Poor, and S. Shamai, "Cellular systems with full-duplex amplify-and-forward relaying and cooperative base-stations," in *Proceedings of the IEEE International Symposium on Information Theory*, pp. 16–20, June 2007.

[48] O. Simeone, O. Somekh, Y. Bar-Ness, and U. Spagnolini, "Uplink throughput of TDMA cellular systems with multicell processing and amplify-and-forward cooperation between mobiles," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, pp. 2942–2951, 2007.

[49] O. Simeone, O. Somekh, Y. Bar-Ness, H. V. Poor, and S. Shamai, "Capacity of linear two-hop mesh networks with rate splitting, decode-and-forward relaying and cooperation," in *Proceedings of the Allerton Conference*, Monticello, Ill, USA, Septtember 2007.

[50] O. Somekh, O. Simeone, H. V. Poor, and S. Shamai, "Cellular systems with full-duplex compress-and-forward relaying and cooperative base stations," in *Proceedings of the IEEE International Symposium on Information Theory*, pp. 2086–2090, July 2008.

Hindawi Publishing Corporation
EURASIP Journal on Wireless Communications and Networking
Volume 2009, Article ID 989062, 13 pages
doi:10.1155/2009/989062

Research Article

LTE Adaptation for Mobile Broadband Satellite Networks

Francesco Bastia, Cecilia Bersani, Enzo Alberto Candreva, Stefano Cioni, Giovanni Emanuele Corazza, Massimo Neri, Claudio Palestini, Marco Papaleo, Stefano Rosati, and Alessandro Vanelli-Coralli

ARCES, University of Bologna, Via V. Toffano, 2/2, 40125 Bologna, Italy

Correspondence should be addressed to Stefano Cioni, scioni@arces.unibo.it

Received 31 January 2009; Revised 29 May 2009; Accepted 30 July 2009

Recommended by Constantinos B. Papadias

One of the key factors for the successful deployment of mobile satellite systems in 4G networks is the maximization of the technology commonalities with the terrestrial systems. An effective way of achieving this objective consists in considering the terrestrial radio interface as the baseline for the satellite radio interface. Since the 3GPP Long Term Evolution (LTE) standard will be one of the main players in the 4G scenario, along with other emerging technologies, such as mobile WiMAX; this paper analyzes the possible applicability of the 3GPP LTE interface to satellite transmission, presenting several enabling techniques for this adaptation. In particular, we propose the introduction of an inter-TTI interleaving technique that exploits the existing H-ARQ facilities provided by the LTE physical layer, the use of PAPR reduction techniques to increase the resilience of the OFDM waveform to non linear distortion, and the design of the sequences for Random Access, taking into account the requirements deriving from the large round trip times. The outcomes of this analysis show that, with the required proposed enablers, it is possible to reuse the existing terrestrial air interface to transmit over the satellite link.

1. Introduction and Motivation

Integrated terrestrial and satellite communication system is a paradigm that has been addressed for many years and that is at the fore front of the research and development activity within the satellite community. The recent development of the DVB-SH standard [1] for mobile broadcasting demonstrates that virtuous synergies can be introduced when terrestrial networks are complemented with a satellite component able to extend their service and coverage capabilities. A key aspect for the successful integration of the satellite and terrestrial components is the maximization of technological commonalities aimed at the exploitation of the economy of scale that derives from the vast market basis achievable by the integrated system. In order to replicate in 4G networks the success of the integrated mobile broadcasting systems, many initiatives are being carried out [2, 3] for the design of a satellite air interface that maximizes the commonalities with the 4G terrestrial air interface. These initiatives aim at introducing only those modifications that are strictly needed to deal with the satellite channel peculiarities, such, for example, nonlinear distortion introduced by the on-board power amplifiers, long round-trip propagation times, and reduced time diversity, while keeping everything else untouched. Specifically, it is important to highlight the different mobile channel propagation models between terrestrial and satellite environments. In fact, in terrestrial deployments, channel fades are typically both time and frequency selective, and are counteracted by the use of opportunistic scheduling solutions, which select for each user the time slots and the frequency bands where good channel conditions are experienced. On the other hand, satellite links are characterized by large round trip delay, which hinders the timeliness of the channel quality indicators and sounding signals, continuously exchanged between users and terrestrial base stations. Further, satellite channel fades are typically frequency-flat, due to the almost Line-of-Sight (LOS) nature of propagation in open area environments, thus alternative solutions have to be designed in order to increase the satellite link reliability.

In this framework, this paper investigates the adaptability of the 3GPP Long Term Evolution (LTE) standard [4] to the satellite scenarios. The 3GPP LTE standard is in fact gaining momentum and it is easily predictable to be one of the main players in the 4G scenario, along with other emerging technologies such as mobile WiMAX [5]. Thanks to this analysis, we propose the introduction of few technology enablers that allow the LTE air interface to be used on a satellite channel. In particular, we propose the following:

(i) an inter-TTI (Transmission Time Interval) interleaving technique that is able to break the channel correlation in slowly varying channels by exploiting the existing H-ARQ facilities provided by the LTE physical layer;

(ii) the introduction of PAPR reduction techniques to increase the resilience of the OFDM waveform to nonlinear distortions;

(iii) a specific design of the sequences for the random access scheme, taking into account the requirements deriving from large satellite round trip times.

In addition, with the aim of further enhancing the robustness to long channel fades, an Upper-Layer (UL) Forward Error Correction (FEC) technique is also proposed and compared with the inter-TTI technique.

According to market and business analysis [6], two application scenarios are considered: mobile broadcasting using linguistic beams with national coverage and two-way communications using multispot coverage with frequency reuse. Clearly, the service typologies paired with these two application scenarios have different requirements in terms of data rates, tolerable latency, and QoS. This has been taken into account into the air interface analysis.

2. GPP LTE: Main Features

The 3GPP LTE air interface is shortly summarized to ensure self-containment and to provide the perspective for the introduction of advanced solutions for the adaptation to satellite links, as described in Section 3.

The FEC technique adopted by LTE for processing the information data is a Turbo scheme using Parallel Concatenated Convolutional Code (PCCC) [7]. Two 8-state constituent encoders are foreseen and the resulting coding rate is 1/3. The LTE technical specifications provide several values for the input block size K_{TC} to the Turbo encoder, varying form $K_{\mathrm{TC}} = 40$ up to $K_{\mathrm{TC}} = 6144$. After channel encoding, the Circular Buffer (CB) and Rate Matching (RM) block allows to interleave, collect and select the three input streams coming from the Turbo encoder (systematic bits, parity sequence from encoder-1 and encoder-2), as depicted in Figure 1. The three input streams are processed with the following steps.

(1) Each of the three streams is interleaved separately by a sub-block interleaver.

(2) The interleaved systematic bits are written into the buffer in sequence, with the first bit of the interleaved systematic bit stream at the beginning of the buffer.

(3) The interleaved P1 and P2 streams are interlaced bit by bit. The interleaved and interlaced parity bit streams are written into the buffer in sequence, with the first bit of the stream next to the last bit of the interleaved systematic bit stream.

(4) Eight different Redundancy Versions (RVs) are defined, each of which specifies a starting bit index in the buffer. The transmitter reads a block of coded bits from the buffer, starting from the bit index specified by a chosen RV. For a desired code rate of operation, the number of coded bits N_{data} to be selected for transmission is calculated and passed to the RM block as an input. If the end of the buffer is reached and more coded bits are needed for transmission, the transmitter wraps around and continues at the beginning of the buffer, hence the term of "circular buffer." Therefore, puncturing, and repetition can be achieved using a single method.

The CB has an advantage in flexibility (in code rates achieved) and also granularity (in stream sizes). In LTE, the encoded and interleaved bits after the RM block are mapped into OFDM symbols. The time unit for arranging the rate matched bits is the Transmission Time Interval (TTI).

Throughout all LTE specifications, the size of various fields in the time domain is expressed as a number of time units, $T_s = 1/(15000 \times 2048)$ seconds. Both downlink and uplink transmissions are organized into radio frames with duration $T_f = 307200T_s = 10\,\mathrm{ms}$. In the following, the *Type-1* frame structure, applicable to both FDD and TDD interface, is considered. Each radio frame consists of 20 slots of length $T_{\mathrm{slot}} = 15360T_s = 0.5\,\mathrm{ms}$, numbered from 0 to 19. A sub-frame is defined as two consecutive slots, where sub-frame i consists of slots $2i$ and $2i + 1$. A TTI corresponds to one sub-frame.

In general, the baseband signal representing a downlink physical channel is built through the following steps:

(i) scrambling of coded bits in each of the code words to be transmitted on a physical channel;

(ii) modulation of scrambled bits to generate complex-valued modulation symbols;

(iii) mapping of the complex-valued modulation symbols onto one or several transmission layers;

(iv) pre-coding of the complex-valued modulation symbols on each layer for transmission on the antenna ports;

(v) mapping of complex-valued modulation symbols for each antenna port to resource elements;

(vi) generation of complex-valued time-domain OFDM signal for each antenna port.

These operations are depicted and summarized in Figure 2. The details and implementation aspects of each block can be extracted from [4]. The transmitted signal in each slot is mapped onto a resource grid of N_a active subcarriers (frequency domain) and N_{symb} OFDM symbols (time domain). The number of OFDM symbols in a slot, N_{symb}, depends on

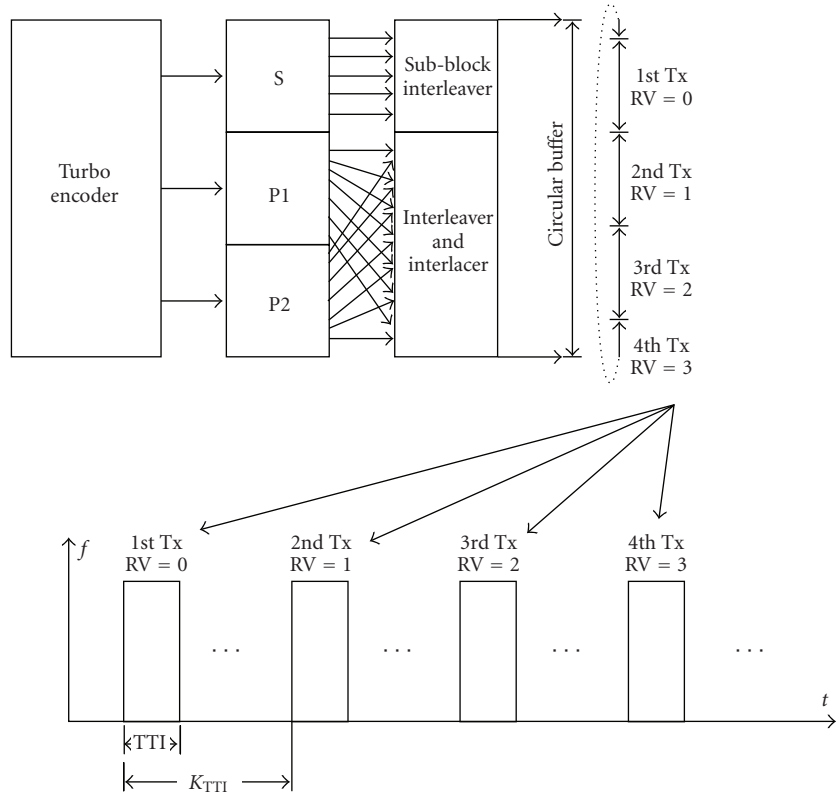


Figure 1: Rate matching and Virtual Circular Buffer.

the cyclic prefix length, N_{cp}, and the subcarrier spacing, Δf. In case of multiantenna transmission, there is one resource grid defined per antenna port. The size of the FFT/IFFT block, N_{FFT}, is equal to 2048 for $\Delta f = 15$ kHz and 4096 for $\Delta f = 7.5$ kHz. Finally, the time continuous signal of the generic ℓ-th OFDM symbol on the antenna port p can be written as

$$s_{\ell}^{(p)}(t) = \sum_{k=-\lfloor N_a/2 \rfloor}^{-1} a_{k+\lfloor N_a/2 \rfloor,\ell}^{(p)} e^{J2\pi k\Delta f(t-N_{cp}T_s)} + \sum_{k=1}^{\lceil N_a/2 \rceil} a_{k+\lfloor N_a/2 \rfloor-1,\ell}^{(p)} e^{J2\pi k\Delta f(t-N_{cp}T_s)} \quad (1)$$

for $0 \leq t \leq (N_{cp} + N_{FFT})T_s$ and where $a_{k,\ell}^{(p)}$ is a complex modulated symbol.

3. Adapting LTE to Satellite Links: Enablers

In the following sections, we propose and analyze some solutions to adapt the 3GPP LTE air interface to broadband satellite networks. These advanced techniques are applied to the transmitter or receiver side in order to enhance and maximize the system capacity in a mobile satellite environment.

3.1. Inter-TTI Interleaving. In this section, we propose an inter-TTI interleaving technique allowing to break channel correlation in slowly varying channels, achieved through the reuse of existing H-ARQ facilities provided by the physical layer of the LTE standard [8].

The LTE standard does not foresee time interleaving techniques outside a TTI [7]. Thus, since the physical layer codeword is mapped into one TTI, the maximum time diversity exploitable by the Turbo decoder is limited to one TTI (T_{TTI}). For low to medium terminal speeds, the channel coherence time is larger than T_{TTI}, thus fading events cannot be counteracted by physical layer channel coding. In order to cope with such a fading events, LTE exploits both "intelligent" scheduling algorithms based on the knowledge of channel coefficients both in the time and in the frequency dimension, and H-ARQ techniques. The former technique consists in exploiting the channel state information (CSI) in order to map data into sub-carriers characterized by high signal to noise ratio (good channel quality). Of course this technique shows great benefits when frequency diversity is present within the active subcarriers.

H-ARQ consists in the "cooperation" between FEC and ARQ protocols. In LTE, H-ARQ operation is performed by exploiting the virtual circular buffer described in Section 2. Orthogonal retransmissions can be obtained by setting the RV number in each retransmission, thus transmitting different patterns of bits within the same circular buffer. Of course, H-ARQ technique yields to great performance improvement when time correlation is present because retransmission can have a time separation greater than channel coherence time.

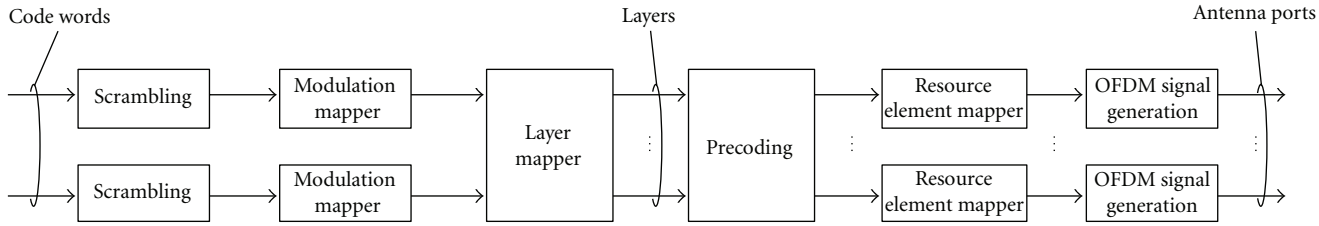


Figure 2: Overview of physical channel processing [4].

Unfortunately, neither of the aforementioned techniques can be directly applied to the satellite case due to the exceedingly large transmission delays, affecting both the reliability of the channel quality indicators and of the acknowledgements. Nevertheless, it is possible to devise a way to exploit the existing H-ARQ facilities adapting them to the satellite use. To this aim, we propose a novel forced retransmission technique, which basically consists in transmitting the bits carried in the same circular buffer within several TTIs, that acts as an inter-TTI interleaving. To do this, we can exploit the same mechanism as provided by the LTE technical specifications for the H-ARQ operations with circular buffer. For the explanation of this solution, the block diagram depicted in Figure 1 can be taken as reference. In this example, 4 retransmissions are obtained by using 4 different RVs, starting from 0 up to 3. Each of the 4 transmission bursts is mapped into different TTIs, spaced by $K_{\mathrm{TTI}} \cdot T_{\mathrm{TTI}}$. K_{TTI} is a key parameter because it determines the interleaving depth and it should be set according to channel conditions and latency requirements.

It is straightforward to derive the maximum time diversity achievable by adopting such as technique. Let R_{TTI} be the number of retransmissions needed to complete the transmission of a single circular buffer, L_{SUB} the number of OFDM symbols transmitted in each retransmission, and T_{SUB} the duration of L_{SUB} OFDM symbols. (The duration of the OFDM symbol T_{OFDM} is intended to be the sum of the useful symbol and cyclic prefix duration.) We have that a codeword is spread over total protection time $T_{\mathrm{TPT}} = K_{\mathrm{TTI}} \cdot (R_{\mathrm{TTI}} - 1) \cdot T_{\mathrm{TTI}} + T_{\mathrm{TTI}}$. Given the fact that the standard facilities are used, no additional complexity is introduced. The drawback involved with the use of such technique is the data rate reduction, brought about by the fact that one codeword is not transmitted in T_{TTI} but in T_{TPT}. A possible way to maintain the original data rate is to introduce in the terminals the capability of storing larger quantities of data, equivalent to the possibility to support multiple H-ARQ processes in terminals designed for terrestrial use. In this way, capacity and memory occupation grow linearly with the number of supported equivalent H-ARQ processes, and is upper bounded by the data rate of the original link without inter-TTI.

3.2. PAPR Reduction Techniques. The tails in Peak-to-Average Power Ratio (PAPR) distribution for OFDM signals are very significant, and this implies an detrimental source of distortion in a satellite scenario, where the on-board amplifier is driven near saturation. To have an idea of the cumulative distribution of PAPR, a Gaussian approximation can be used. With this approach, if OFDM symbols in time domain are assumed to be Gaussian distributed, their envelope can be modeled with a Rayleigh distribution. Thus, the cumulative distribution function of PAPR variable is

$$\mathbb{P}[\mathrm{PAPR} \leq \gamma] = (1 - \mathrm{e}^{-\gamma})^{N_{\mathrm{FFT}}}. \tag{2}$$

A more meaningful measure is given by the *complementary cumulative distribution function*, which gives the probability that PAPR exceeds a given value γ, and can be written as

$$\mathbb{P}[\mathrm{PAPR} \geq \gamma] = 1 - (1 - \mathrm{e}^{-\gamma})^{N_{\mathrm{FFT}}}. \tag{3}$$

As an example of using this simple approximation, which becomes increasingly tight increasing the FFT size, it is easy to check that a PAPR of 9 dB is exceeded with a probability of 0.5 assuming $N_{\mathrm{FFT}} = 2048$, while a PAPR of 12 dB is exceeded with a probability of $2.7 \cdot 10^{-4}$.

This argument motivates the use of a PAPR reduction technique, in order to lower the PAPR and drive the satellite amplifier with a lower back-off. Power efficiency is at a prime in satellite communications, and an eventual reduction of the back-off implies an improvement in the link budget and an eventual increase of the coverage area. Amongst all requisites for PAPR reduction techniques (see [9, 10] for a general overview), the compatibility with the LTE standard is still fundamental. Secondly, the receiver complexity must not be significantly increased. Furthermore, no degradation in BER will be tolerated, because it would require an increased power margin. Finally, the PAPR reduction method will cope with the severe distortion given by the satellite: even if the amplifier has an ideal pre-distortion apparatus on-board, it is operated near to its saturation, where a predistorter could not invert the flat HPA characteristic. The cascade of an ideal predistorter and the HPA is the so-called *ideal clipping* or *soft limiter*. In such a scenario, if the PAPR is lower than the IBO the signal will not be distorted, while if the PAPR is significantly higher the signal will be impaired by non-linear distortion. Thus, the PAPR reduction technique should offer a good PAPR decrease for almost all OFDM symbols, rather than a decrease which can be experienced with a very low probability.

Several techniques have been proposed in the literature, and even focusing on techniques which do not decrease the spectral efficiency, the adaptation to satellite scenario remains an issue: this is the case of *Tone Reservation* [11–13], the intermodulation products of satellite amplifier

prevent using this technique, while it is very popular in the wired scenario and when the amplifier is closer to its linear region. The *Selected Mapping* technique [14, 15], although easy and elegant, needs a *side information* at the receiver. The side information can be avoided, at expense of a significant computational complexity increase at the receiver. Companding techniques (see [10] and references therein) offer a dramatic reduction in PAPR and do not require complex processing. On the other hand, there is a noise enhancement, which turns out to be an important source of degradation at the very low SNRs used in satellite communications.

The *Active Constellation Extension* (ACE) technique [16] fulfills those requirements, moreover the power increase due to PAPR reduction is exploited efficiently, obtaining an additional margin against noise. The ACE approach is based on the possibility to dynamically extend the position of some constellation points in order to reduce the peaks of the time domain signal (due to a constructive sum of a subset of the frequency domain data) without increasing Error Rate: the points are distanced from the borders of their Voronoi regions. The extension is performed iteratively, according to the following procedure.

(1) Start with the frequency domain representation of a OFDM symbol.

(2) Convert into the time-domain signal, and clip all samples exceeding a given magnitude V_{clip}. If no sample is clipped, then exit.

(3) Reconvert into the frequency domain representation and restore all constellation points which have been moved towards the borders of their Voronoi regions.

(4) Go back to 2 until a fixed number of iteration is reached.

This algorithm is applied to data carriers only, excluding thus pilots, preamble/signalling and guard bands. In the performance evaluation of the algorithm, the amplitude clipping value is expressed in term of the corresponding PAPR, which is called *PAPR-Target* in the following.

The most critical point of this method is the choice of the clipping level V_{clip}: a large value for V_{clip} (which corresponds to an high *PAPR-Target*) will yield a negligible power increase and a poor convergence, since signal is unlikely to be clipped. On the opposite extreme, a very low clipping level will yield again a poor convergence and a negligible power increase. In fact, considering the above algorithm, almost all points will be moved by clipping in *step-2* and then restored by the constellation constraint enforcing in *step-3*. A compromise value, which will lead to a PAPR around 5 or 6 dB is advisable, yielding a good convergence and a slight energy increase, due to the effectiveness of the extension procedure. Although there are other ACE strategies [16], the solution presented here is attractive because it can be easily implemented both in hardware and software, as reported in [17].

3.3. Random Access Signal Detection. The Random Access Channel (RACH) is a contention-based channel for initial uplink transmission, that is, from mobile user to base station. While the Physical RACH (PRACH) procedures as defined in the 3G systems are mainly used to register the terminal after power-on to the network, in 4G networks, PRACH is in charge of dealing with new purposes and constraints. In an OFDM based system, in fact, orthogonal messages have to be sent, thus the major challenge in such a system is to maintain uplink orthogonality among users. Hence both frequency and time synchronization of the transmitted signals from the users are needed. A downlink broadcast signal can be sent to the users in order to allow a preliminary timing and frequency estimation by the mobile users, and, accordingly a timing and frequency adjustment in the return link. The remaining frequency misalignment is due to Doppler effects and cannot be estimated nor compensated. On the other hand, the fine timing estimation has to be performed by the base station when the signals coming from users are detected. Thus, the main goal of PRACH is to obtain fine time synchronization by informing the mobile users how to compensate for the round trip delay. After a successful random access procedure, in fact, the base station and the mobile user should be synchronized within a fraction of the uplink cyclic prefix. In this way, the subsequent uplink signals could be correctly decoded and would not interfere with other users connected to the network.

PRACH procedure in 4G systems consists in the transmission of a set of preambles, one for mobile user, in order to allocate different resources to different users. In order to reduce collision probability, in the LTE standard, Zadoff-Chu (ZC) sequences [18], known also as a Constant Amplitude Zero Autocorrelation (CAZAC) sequences, are used as signatures between different use, because of the good correlation properties. The ZC sequence obtained from the u-th root is defined by

$$x_u(n) = \exp^{-j(\pi un(n+1)/N_{\text{ZC}})} \quad 0 \le n \le N_{\text{ZC}} - 1, \tag{4}$$

where N_{ZC} is the preamble length in samples and it has been set to 839. ZC sequences present very good autocorrelation and cross-correlation properties that make them perfect candidates for the PRACH procedure. In fact, orthogonal preambles can be obtained cyclic rotating two sequences obtained with the same root, according to the scheme shown in Figure 3 and the expression

$$x_{u,v}(n) = x_u((n + C_v) \bmod N_{\text{ZC}}) \quad v = 0, 1, \ldots, \left\lfloor \frac{N_{\text{ZC}}}{N_{\text{CS}}} \right\rfloor - 1, \tag{5}$$

where N_{CS} is the number of cyclic shifts. It can be easily verified that the cross correlation function presents N_{CS} peaks and N_{CS} zero correlation zones. Figure 4(a) shows a magnification of the cross correlation function for different shifts considering $N_{\text{CS}} = 64$. It will be noted that there are $N_{\text{CS}} - 2$ zero correlation zones with length equal to 12 samples and the last zero correlation zone with 20 samples. Preambles obtained from different roots are no longer orthogonal but, nevertheless, they present good correlation properties.

Considering a 4G system via satellite, the number of users to be allocated in each cell depends on the system

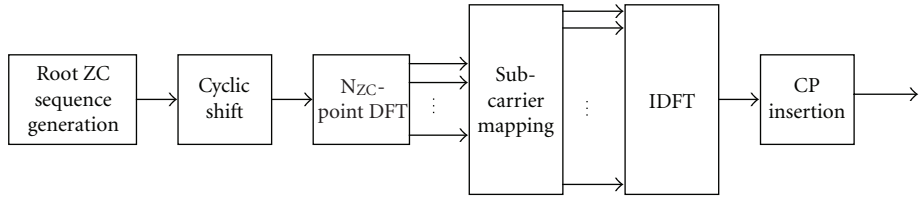
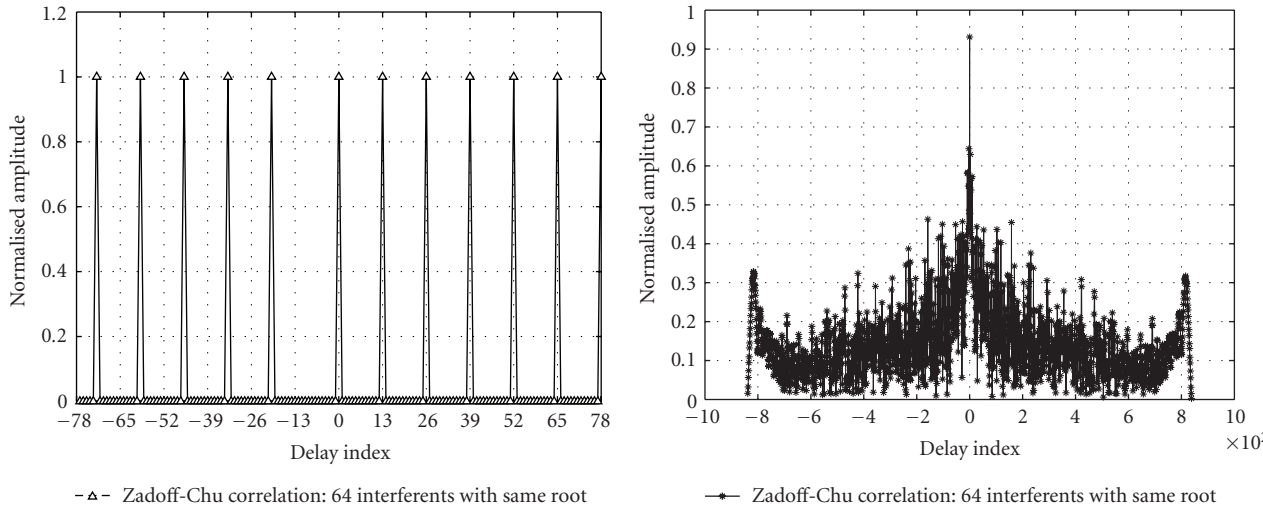


Figure 3: ZC generation in time domain processing.



(a) Correlation properties with 64 Zadoff-Chu interfering sequences with the same root and different cyclic shifts

(b) Correlation properties with 64 Zadoff-Chu interfering sequences with different roots

Figure 4: Detection properties in the presence of interferers.

Table 1: ZC allocation for GEO satellite scenario.

Cell Radius [km]	Number of root ZC sequences	Number of cyclic shift per root sequence
150 (Near polar arctic circle)	64	1
300 (Near polar arctic circle)	64	1
500 (Near polar arctic circle)	64	1
150 (Europe)	64	1
300 (Europe)	64	1
500 (Europe)	64	1
150 (Tropical)	32	2
300 (Tropical)	64	1
500 (Tropical)	64	1
150 (Equator)	2	32
150 (Equator)	8	8
150 (Equator)	16	4

design. The zero correlation zone of the preambles has to be larger then the maximum round trip propagation delay, depending on cell radius and multipath delay. The number of root ZC sequences and the number of cyclic shift sequences depend on cell radius and on the geographical position, and they are reported in Table 1 for GEO satellites. Note that the worst case corresponds to the presence of 64 sequences obtained from different roots. In this case, the satellite has to detect each sequence even between the interference from the others. Figure 4(b) shows the correlation function in a scenario like this, and it is worthwhile noting that the peak can once more be detected, also in the presence of 63 interferers. Detection performance in terms of Receiver Operating Characteristics (ROC), that is, Missed Detection Probability (P_{md}) as a function of False Alarm Probability (P_{fa}) have been reported for different numbers of interferers in Figure 5. It will be highlighted that the detection has been performed in the frequency domain and a Non-Coherent Post-Detection Integration (NCPDI) [19] scheme has been adopted. Finally, the results are shown in a AWGN scenario with a signal to noise ratio, E_s/N_0, equal to 0 dB.

4. Upper Layer FEC Analysis

In this section, we propose a UL-FEC technique working on top of the PHY layer. It is well known that channel coding can be performed at different layers of the protocol stack. Two are the main differences which arise when physical layer or upper layer coding is addressed: the symbols composing each codeword, and the channel affecting the transmitted codeword. Indeed, at physical layer the symbols involved in the coding process typically belong to the Galois Field of

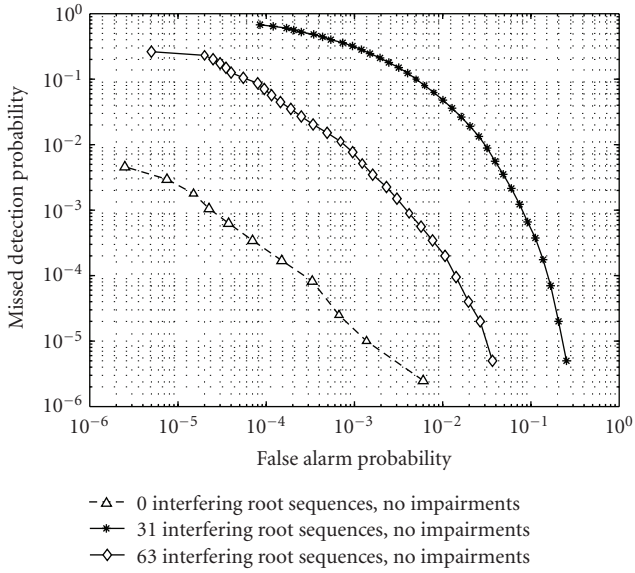


FIGURE 5: ROC in AWGN channel with $E_s/N_0 = 0.0$ dB without interference, and with interferers with different roots.

order m, $GF(m)$. Nevertheless, also non binary codes can be adopted. Working at upper layer each symbol composing the UL codeword can be made up of packets of bits, depending on the application level.

In order to build the UL-FEC technique on solid ground, the design and analysis has been carried out starting from the Multi Protocol Encapsulation Forward Error Correction Technique (MPE-FEC) adopted by the DVB-H standard [20], and successively enhanced and modified in the framework of the DVB-SH [1] standardization group. With respect to the MPE-FEC approach, the implementation of the UL-FEC technique for this framework has required to adapt the parameter setting to the LTE physical layer configurations. In the following, we adopt this terminology:

(i) k: the UL block length, that is the number of systematic symbols to be encoded by the UL encoder

(ii) n: the UL codeword length, that is the number of UL symbols produced by the UL encoder

(iii) k': the actual UL-FEC block length if zero-padding is applied

(iv) n': the actual UL-FEC codeword length if zero-padding and/or puncturing is applied

(v) N_{JCC}: number of jointly coded channels at physical layer

(vi) S_{JCC}: size of each channel in bytes

(vii) $S_{\text{UL-CRC}}$: size of the upper layer Cyclic Redundancy Check (CRC) in bytes

(viii) $S_{\text{PHY-CRC}}$: size of the physical layer CRC in bytes

(ix) K_{PHY}: physical layer block length in bytes.

As in MPE-FEC, we define the UL-FEC matrix as a matrix composed of a variable number of rows (n_of_rows) and n columns. Each entry of the matrix is an UL-symbol, that is, 1 byte. The first k columns represent the systematic part of the matrix and are filled with the systematic UL-symbols coming from the higher level. The last $n - k$ columns carry the redundancy data computed on the first k columns. It is worthwhile to notice that the n and k values depend on the selected UL code rate only, while n_of_rows is a parameter chosen accordingly to the physical layer configuration and is set by using the following formula: $n\_\text{of}\_\text{rows} = K_{\text{PHY}} - S_{\text{PHY-CRC}} - N_{\text{JCC}}S_{\text{UL-CRC}}$. As a consequence, the number of bytes available for each channel in a given UL-FEC matrix column is $S_{\text{JCC}} = n\_\text{of}\_\text{rows}/N_{\text{JCC}}$. With this configuration, the following operations must be sequentially performed.

(1) The information data coming from higher layer are written columns-wise in the systematic data part of the UL-FEC matrix.

(2) A Reed-Solomon (RS) encoding (n, k) is performed on each row producing the redundancy part of the UL-FEC matrix.

(3) The data are transmitted column-wise.

(4) An UL-CRC is appended after each group of S_{JCC} bytes.

(5) Each group of $K_{\text{PHY}} = N_{\text{JCC}}(S_{\text{JCC}} + S_{\text{UL-CRC}})$ bytes composes a physical layer information packet.

(6) The PHY-CRC is appended to each physical layer information packet according to the LTE specifications [7].

For sake of simplicity, we adopt the same RS mother code provided in [20], which is an RS(255,191). The code rate of this mother code is 3/4. Further code rates can be achieved by using padding or puncturing techniques. For instance, if a UL-FEC rate 1/2 is needed, zero-padding is performed over the last 127 columns of the systematic data part of the UL-FEC matrix, yielding to $k' = 64$ and $n' = 128$. The choice of this RS code allows fully compatibility with DVB-H networks.

It is important to note how the application of the CRC at UL and physical layer has an impact on the overall system performance. To better evaluate this impact, we distinguish to study cases:

(i) *Case-A*: only the PHY-CRC is considered ($S_{\text{UL-CRC}} = 0$). In this scenario, the receiver is not able to check the integrity of a single UL packet carried within the same physical layer information packets. This basically means that if error is detected in the physical layer information packet, all UL packets will be discarded;

(ii) *Case-B*: both PHY and UL CRC are applied.

It is quite obvious that *Case-B* outperforms *Case-A*. In fact, if only a small fraction of bits are wrong after physical layer decoding, *Case-B* is able to discard only the UL packets in which erroneous bits are present, while *Case-A* discards all N_{JCC} carried within the physical layer information packets. The price to pay is an increased overhead of *Case-B* with respect to *Case-A* due to the extra CRC bits appended.

At the receiver side, depending whether *Case-A* or *Case-B* is taken into account, CRC integrity must be performed at different levels. If the *Case-A* is considered, only the CRC at physical layer determines the data reliability; whereas in the *Case-B*, the PHY-CRC could be ignored and the data reliability is only determined by the UL-CRC. Then, the UL-FEC matrix is filled with the reliable data. In particular, for the *Case-A* an entire column is marked as reliable or not reliable, while in the *Case-B* the UL-FEC matrix columns could be partially reliable. Finally, the RS(n, k) decoding is performed on each row. If the number of reliable position in a row is at least k, the decoder is able to successfully decode the received information, and all unreliable positions are recovered.

The UL-FEC protection capability against burst of errors can be characterized by the so-called Maximum Tolerable Burst Length (MTBL) [21], which consists in the maximum time protection that the UL-FEC technique can provide. The MTBL depends on both UL-FEC parameters and PHY data rate. In our proposal one PHY information packet is mapped in one column of the UL-FEC matrix. Since we are dealing with MDS codes, the decoder is able to successfully decode if at least k' columns are correctly received in the UL-FEC matrix. Thus, the MTBL is simply given by the time taken by transmitting $n' - k'$ columns, that is, the duration of $n' - k'$ information packets. The MTBL can be increased by adopting a sliding encoding mechanism [22]. The sliding encoding is a UL interleaver mechanism: a UL-FEC encoder implementing sliding encoding selects the k' data columns from a window (SW) among the UL-FEC matrices and spreads the $n' - k'$ parity sections over the same window. Basically, the same effect could be obtained by first normally encoding SW frames and then interleaving sections among the encoded SW frames. The total protection time $\mathrm{TPT_{UL}}$ achievable at upper layer by means of such a technique is given by $\mathrm{TPT_{UL}} = n' \cdot \mathrm{SW} \cdot T_{\mathrm{TTI}}$.

5. Simulation Results

Here, we discuss separately the numerical results obtained by implementing the solutions presented in Section 3. The following general assumptions have been considered during the implementation of all techniques.

The LTE transmitted signal occupies 5 MHz of bandwidth, $N_a = 300$, located in S-band (central frequency $f_0 = 2$ GHz), the sub-carrier spacing is $\Delta f = 15$ kHz, and FFT/IFFT size is fixed to $N_{\mathrm{FFT}} = 2048$. The long cyclic prefix is assumed, $N_{\mathrm{cp}} = 512$, thus $N_{\mathrm{symb}} = 12$ OFDM symbols are transmitted in each TTI. The resulting OFDM symbols duration is $T_{\mathrm{ofdm}} = 83.33\,\mu$s, including the cyclic prefix duration of $T_{\mathrm{cp}} = 16.67\,\mu$s.

5.1. Inter-TTI Improvements. For evaluating the inter-TTI proposal, the turbo encoder is fed with 2496 information bits, while the circular buffer size is assumed to be 6300, thus resulting in an actual system code rate equal to $R \simeq 2/5$. All simulations have considered QPSK modulation.

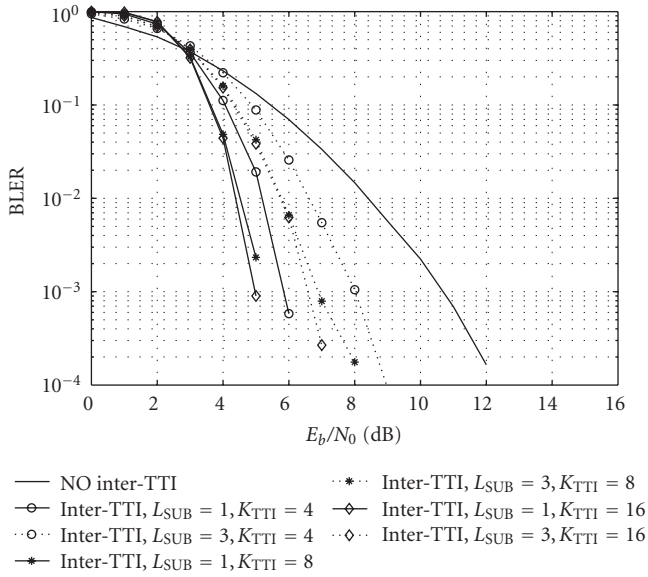


Figure 6: BLER versus E_b/N_0. Terminal speed is equal to 30 km/h.

Figure 6 shows the block error rate (BLER) performance versus E_b/N_0, with E_b being the energy per information bit and N_0 the one-sided noise power spectral density. The curves refer to a user terminal speed of 30 km/h. The solid line curves represent the cases in which the number of transmitted OFDM symbols for each retransmission (L_{SUB}) is 1, resulting in a total number of retransmissions $R_{\mathrm{TTI}} = 12$, while the dashed line curves depict the case with $L_{\mathrm{SUB}} = 3$ and $R_{\mathrm{TTI}} = 4$. In these configurations, we set the value of K_{TTI} such that the total protection time T_{TTI} is larger than the channel coherence time T_c, which for these simulations is about $T_c \simeq 9$ ms. (This is the coherence time of the small scale fluctuations, and it depends directly from the terminal speed and the central carrier frequency.) In particular, the simulated values K_{TTI} are 4, 8, 16. As it can be observed, the solid line curves always outperform the dashed line ones. This is easily explained considering the different diversity granularity: in the case of $L_{\mathrm{SUB}} = 1$, each OFDM symbol is transmitted in a separated TTI. Therefore, the codeword spanned over the 12 OFDM symbols composing the entire TTI can benefit of diversity degree equal to 12. On the other hand, if the case of $L_{\mathrm{SUB}} = 3$ is considered, the degree diversity is reduced to 4. It is worthwhile to note the large performance enhancement yielded by the adoption of the inter-TTI technique. For instance, looking at Figure 6, the performance gain at BLER $= 10^{-3}$ increases up to 6 dB in the case of $L_{\mathrm{SUB}} = 1$, and up to 4 dB considering $L_{\mathrm{SUB}} = 3$.

5.2. ACE Performance. The results of the ACE algorithm for PAPR reduction are discussed. First of all, the CCDF of PAPR distribution have been analyzed for verifying the effectiveness of the selected method.

Figures 7 and 8 show PAPR distribution for QPSK and 16QAM, respectively. As it can be seen, if the *PAPR-Target* is too low, the CCDF curve has a poor slope. Increasing the *PAPR-Target*, the curve is shifted left until

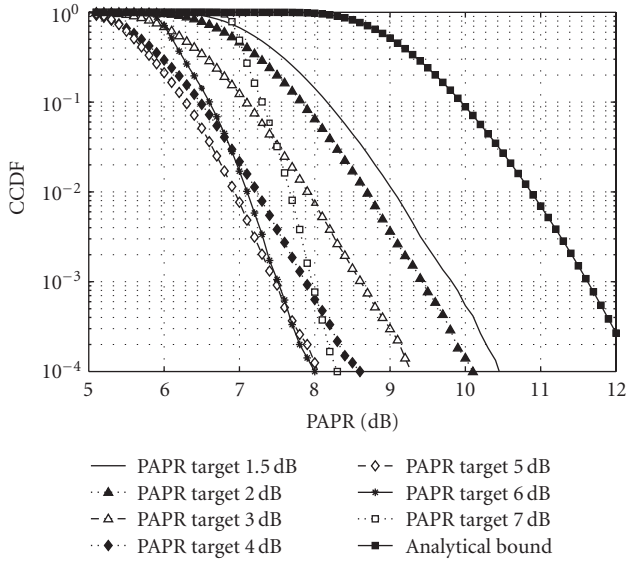


Figure 7: PAPR CCDF with QPSK modulation.

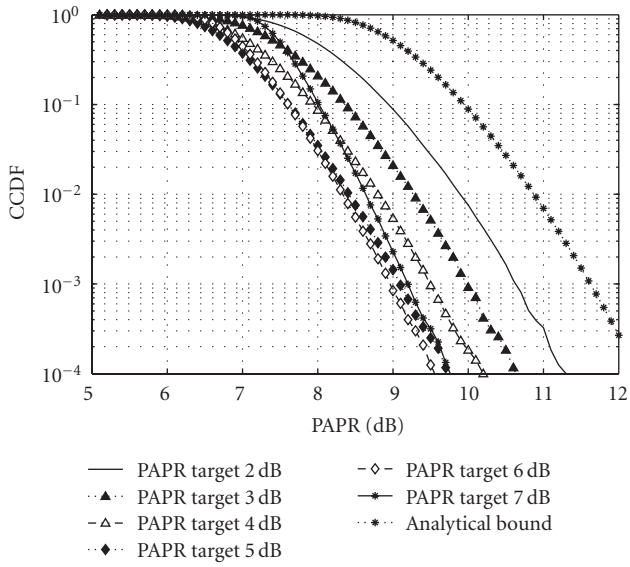


Figure 8: PAPR CCDF with 16QAM modulation.

a certain value, then the steepness increases and, if the *PAPR-Target* is furthermore increased, the curve is shifted right, maintaining the same steepness. This phenomenon is more evident for QPSK modulation rather than for 16QAM, and this difference can be explained considering that all QPSK constellation points can be moved in some directions by the ACE algorithm, while for 16QAM the inner points must be immediately restored, and the points on the edges have only one degree of freedom.

A more interesting figure of merit related to this PAPR reduction technique is the improvement in terms of bit error rate, which summarizes the impact of PAPR reduction on the end-to-end performance. Figure 9 shows the BER improvement in a frequency selective channel, with the

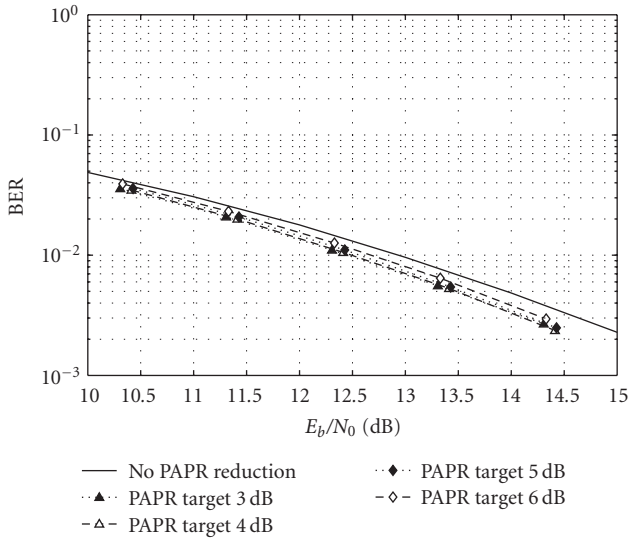


Figure 9: BER performance using PAPR techniques with 16QAM and code-rate = 3/5.

amplifier Input Back-Off (IBO) set to 3 dB. The 16QAM modulation is considered, the coding rate is $r = 3/5$, and the packet size is chosen equal to 7552 bits. As shown in Figure 9, there is a gain of almost 0.5 dB if the *PAPR-Target* is kept low; the gain is slightly lower if the *PAPR Target* is chosen in order to maximize the beneficial effects of ACE technique in terms of PAPR CCDF. This result can be justified by considering the worst-case conditions assumed in these simulations: the amplifier driven 3 dB far from saturation requires a PAPR value as low as possible, while the slight energy increase is conveniently exploited in such a severe fading channel environment.

5.3. Redundancy Split Analysis. A comparison between the UL-FEC and the inter-TTI interleaver technique is reported. In order to make a fair comparison between these two techniques, in the following we keep constant the overall spectral efficiency by distributing the redundancy between UL-FEC and physical layer. Figure 10 shows the numerical results obtained in the case of assuming the terminal speed equal to 3 km/h, and ideal channel estimation. The performance is measured in terms of BLER versus E_b/N_0. All reported curves have a spectral efficiency equal to 4/5 bit/s/Hz. In the inter TTI case, we have considered the coding rate $r = 2/5$ and QPSK modulation, and we have varied both the interleaver depth and the subframe size. On the other hand, the UL-FEC solution have been implemented by considering $r = 4/5$ with QPSK modulation at the physical layer, and the $(k' = 64, n' = 128)$ code at the upper layer. Since the considered UL-FEC protection spans over $n' = 128$, that corresponds to 128 ms, the most comparable protection time provided by the inter-TTI approach is obtained by adopting the parameters $K_{\text{TTI}} = 40$ and $L_{\text{SUB}} = 3$, which still guarantee orthogonal retrasmissions. In this case, the physical layer codeword spans over $K_{\text{TTI}} \cdot 4 = 160$ TTIs, that is, 160 ms. From the analysis of the results, we can state that on the

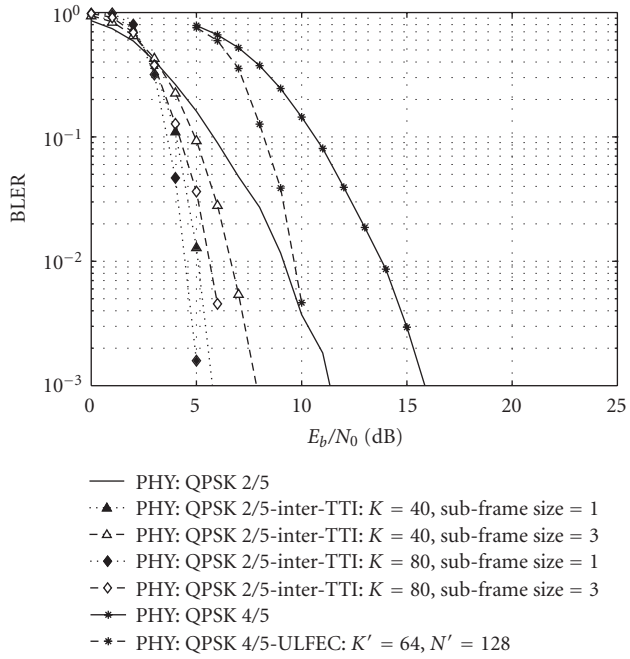


Figure 10: Comparison between Inter-TTI and UL-FEC techniques.

one hand, the inter TTI techniques outperforms the UL-FEC technique, which can be justified recalling that at physical layer the decoder can exploit soft information, thus achieving much better performance with respect to the hard decoding performed at upper layer. On the other hand, the inter-TTI technique requires a large memory buffer at the output of the base-band processor. A through complexity analysis must be carried out to this respect in order to understand the hardware feasibility of the assumption considered for the inter-TTI interleaving case.

5.4. End-to-End Performance Evaluation. In this section, the results obtained considering end-to-end simulations in realistic satellite propagation scenario are analyzed. To this aim, we have adopted the Land Mobile Satellite (LMS) channel model proposed in [23], which is based on measurement campaigns. This channel model is characterized by a three states Markov model. Each state describes different propagation conditions, that are line of sight, moderate shadowing conditions, and deep shadowing conditions. By suitably setting the Markov chain parameters, several environment can be modeled. In our analysis we have considered an elevation angle of 40 degrees and the following environments: open area [O], Suburban [S], Intermediate tree shadow [ITS], Heavy Tree Shadow [HTS]. Such environments are characterized by long fading events due to the superposition of shadowing effects. It is quite obvious that applying the proposed UL-FEC technique without any interleaver working at UL does not allow to cope with such channel impairments. Indeed, the MTBL achievable by adopting UL-FEC without sliding interleaving (SW = 1) is in the order of hundreds milliseconds. To increase the MTBL we adopt the sliding window encoding technique. As already mentioned, this technique basically consists in applying a block interleaver at UL.

In order to correctly evaluate the achievable performance of the proposed UL-FEC technique, we have fed the UL-FEC decoder with time series. Since the fading is frequency flat and for low to medium terminal speeds time selectivity is negligible with respect to the TTI duration (channel coherence time equal to 9 ms at 30 km/h, whereas TTI duration equal to 1 ms for LTE), we can assume that the SNR is constant within the whole TTI (both in frequency and in time). (Again, this fading coherence time is referred to the small scale fluctuations, while the large scale is taken into account in the LMS channel parameters.) Under these assumptions, the BLER time series can be generated using a simplified method, that does not require the actual simulation of the whole physical layer chain. The adopted procedure is depicted in Figure 11, and is made up by the following steps:

(1) perform AWGN simulations (including NL distortion), to obtain the function BLER versus E_b/N_0;

(2) generate the Perez Fontan channel coefficients, obtaining signal levels relative to LOS component;

(3) calculate the received C/N_0 value in LOS conditions;

(4) map the instantaneous C/N_0 value into E_b/N_0;

(5) generate the time series, producing a "1" (wrong block) or a "0" (correct block) according to the following algorithm: *if [uniform-random-variable < BLER ($E_b/N_0{}^*$)] then time-series-value = 1, else time-series-value = 0.*

In order to get a synthetic analysis of the results, we have assessed the Erroneous Seconds Ratio (ESR) criterion. The ESR was also considered by the DVB-SSP [24] group to be the most relevant performance parameter for the assessment of the impact on the video quality. In particular, we take into account the ESR5(20) criterion: ESR5(20) is fulfilled for a given time interval of 20 seconds if the percentage of erroneous seconds in the same time interval does not exceed 5%, which corresponds to a maximum of 1 erroneous second. The percentage of time satisfying the ESR5(20) criterion represents the "ESR5(20) fulfillment percentage." The conclusions of this analysis are summarized in Figure 12, where the achievable spectral efficiency is reported as a function of the C/N required to satisfy the ESR5(20) criterion at 90%. The spectral efficiency is computed considering the PHY configurations listed in Table 2. Notably, since usually in a LTE frame both information and control data are transmitted, we assumed that the equivalent of 1 OFDM symbol per TTI, that is, 1/12 of the TTI, is completely dedicated to the transmission of control data. As a consequence, the PHY spectral efficiency resulting from Table 2 has been reduced by a factor (11/12).

In Figure 12, each curve represents the performance of the QPSK constellation in a given scenario and for a given UL-FEC coding rate. The connected markers in each curve represent the corresponding PHY configurations in a given

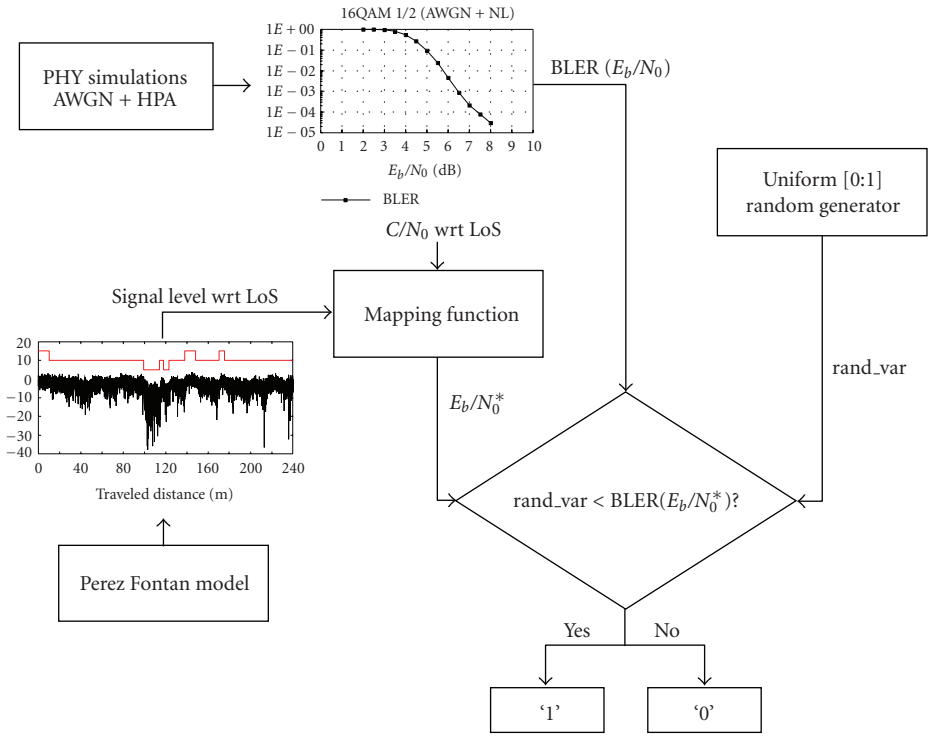


Figure 11: Block diagram of the procedure adopted for generating the time series.

Table 2: Adopted LTE Physical layer configurations.

Number of jointly coded channels/number of channel groups	Info-bits per packet	Allocated data carriers per sub-frame [RBs × OFDM symbols]	Modul.	FEC Code rate	Overall Bit Rate Channel
8/1	2496	3150 [25 × 12]	QPSK	2/5	2.50 Mb/s
16/1	4992	3150 [25 × 12]	QPSK	4/5	4.99 Mb/s
24/1	7552	3150 [25 × 12]	16QAM	3/5	7.49 Mb/s

scenario and for a given UL-FEC coding rate. Regarding the UL parameters, two configuration have been taken into account: rate 1/2 ($n' = 128$, $k' = 128$) and rate 3/4 ($n' = 191$, $k' = 255$). The adopted sliding window size has been set to SW = 101 for the rate 1/2 case, and 50 for the rate 3/4, yielding to a total protection time at UL equal to $\text{TPT}_{\text{UL}} = 12.928$ s, and $\text{TPT}_{\text{UL}} = 12.75$ s, respectively. Notably, for the 16QAM constellation, only one PHY FEC scheme has been considered. Interestingly, the lower UL-FEC protection, that is, 3/4, always outperforms, at the same total spectral efficiency, the higher UL-FEC protection, with the only exception of the Heavy Tree Shadow scenario. In that case, the extremely challenging propagation conditions calls in fact for a very strong protection along with a quite demanding link budget.

6. Conclusions and Recommendations

The adoption of the 3GPP LTE air interface to broadband satellite networks has been evaluated. The rationale for this choice was the maximization of the commonalities with the terrestrial air interfaces, so as to reduce both non-recurrent engineering and production costs, while easing interworking procedures. The selected numerologies for forward and reverse links are standard compatible. In this sense, the results produced are significant from the 3GPP point of view.

Regarding time domain fade mitigation techniques, one of the major findings consists in a way to obtain the above diversity in an almost standard compatible way. This is the inter-TTI technique, which has been shown to bring significant benefits without touching the physical layer definition.

PAPR reduction algorithms, coupled to predistortion techniques, are a novelty for OFDM transmission through a satellite. We have explored this architecture and our results show that the PAPR itself can be reduced by 2 to 4 dB (guaranteed at 99.9%), which translates into the possibility to reduce the OBO by about 0.7 dB and to gain about 0.5 dB in E_b/N_0 for typical quality of services. All in all, we can

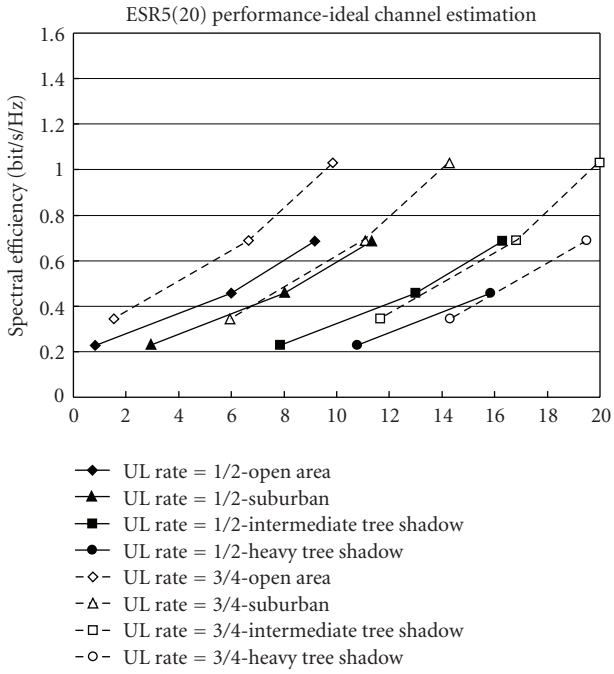


Figure 12: Overall (PHY + UL) spectral efficiency versus *C/N* for 90% ESR5(20).

expect a gain in total degradation around 1 dB, which is certainly not negligible.

Regarding frame acquisition procedures, they are quite specific for LTE air interface. The design of acquisition sequences for 3GPP LTE has been performed adapting it to the different requirements set by satellite transmission involving the use of large geographic beams.

Additionally, in order to further extend the link reliability over the satellite link, the use of UL-FEC techniques has been investigated. Simulation results clearly show that the UL-FEC technique is a very effective solution that can drastically improve the achievable block error rate and ESR5(20) performance.

In order to provide useful guidelines for the system design, the analysis of the optimum redundancy split between physical and upper layer coding has been performed. In this case, results show that in most cases it is beneficial to limit the protection at physical layer in order to ease channel estimation and to compensate the reduced performance through a stronger UL coding. The rationale behind this conclusion is that the UL-FEC benefits a larger time diversity thus performing significantly better than the physical layer coding in almost all scenarios.

Acknowledgment

This work is supported in part by the ESA contract no. 20194/06/NL/US, "Study of Satellite Role in 4G Mobile Networks."

References

[1] ETSI EN 302 583, "Digital video broadcasting (DVB); framing structure, channel coding and modulation for satellite services to handheld devices (SH) below 3 GHz," v1.1.1, March 2008.

[2] The Integral Satcom Initiative (ISI), "ISI strategic research agenda," FP7 Technology Platform, v1.1, January 2006, http://www.isi-initiative.eu.org/isi_joomla.

[3] ETSI TR 102 443, "Satellite earth stations and systems (SES); satellite component of UMTS/IMT-2000; evaluation of the OFDM as a satellite radio interface," v1.1.1, August 2008.

[4] 3GPP TS36.211, "3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Channels and Modulation (Release 8)," v8.2.0, March 2008.

[5] "IEEE standard for local and metropolitan area networks, part 16: air Interface for fixed and mobile broadband wireless access systems amendment 2: physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1," IEEE Computer Society and the IEEE Microwave Theory and Techniques Society, 20 February 2006.

[6] G. E. Corazza, P. Britten, I. Buret, et al., "Defining the role of satellite communications in 4G," in *Proceedings of the 8th World Wireless Congress on Fourth Generation Mobile Communications (WWC '07)*, pp. 60–64, San Francisco, Calif, USA, May 2007.

[7] 3GPP TS36.212, "3rd Generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access (E-UTRA); multiplexing and channel coding (release 8)," v8.2.0, March 2008.

[8] M. Papaleo, M. Neri, G. E. Corazza, and A. Vanelli-Coralli, "Using LTE in 4G satellite communications: increasing time diversity through forced retransmission," in *Proceedings of the 10th International Workshop on Signal Processing for Space Communications (SPSC '08)*, Rhodes Island, Greece, October 2008.

[9] S. H. Han and J. H. Lee, "An overview of peak-to-average power ratio reduction techniques for multicarrier transmission," *IEEE Wireless Communications*, vol. 12, no. 2, pp. 56–65, 2005.

[10] T. Jiang and Y. Wu, "An overview: peak-to-average power ratio reduction techniques for OFDM signals," *IEEE Transactions on Broadcasting*, vol. 54, no. 2, pp. 257–268, 2008.

[11] J. Tellado, *Peak to average power reduction for multicarrier modulation*, Ph.D. dissertation, Stanford University, Stanford, Calif, USA, 2000.

[12] B. S. Krongold and D. L. Jones, "An active-set approach for OFDM PAR reduction via tone reservation," *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 495–509, 2004.

[13] S. Janaaththanan, C. Kasparis, and B. G. Evans, "A gradient based algorithm for PAPR reduction of OFDM using tone reservation technique," in *Proceedings of the 67th IEEE Vehicular Technology Conference (VTC '08)*, pp. 2977–2980, Marina Bay, Singapore, May 2008.

[14] R. W. Bäuml, R. F. H. Fischer, and J. B. Huber, "Reducing the peak-to-average power ratio of multicarrier modulation by selected mapping," *Electronics Letters*, vol. 32, no. 22, pp. 2056–2057, 1996.

[15] M. Breiling, S. H. Mller-Weinfurtner, and J. B. Huber, "SML peak-power reduction without explicit side information," *IEEE Communications Letters*, vol. 5, no. 6, pp. 239–241, 2001.

[16] B. S. Krongold and D. L. Jones, "PAR reduction in OFDM via active constellation extension," *IEEE Transactions on Broadcasting*, vol. 49, no. 3, pp. 258–268, 2003.
[17] ETSI EN 302 755, "Digital video broadcasting (DVB); frame structure channel coding and modulation for a second generation digital terrestrial television broadcasting system (DVB-T2)," April 2008.
[18] D. C. Chu, "Polyphase codes with good periodic correlation properties," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 531–532, 1972.
[19] A. J. Viterbi, *CDMA Principles of Spread Spectrum Communications*, Addison-Wesley Wireless Communications Series, Addison-Wesley, Reading, Mass, USA, 2nd edition, 1995.
[20] ETSI TR 102 377, "Digital video broadcasting (DVB); DVB-H implementation guidelines," v1.2.1, November 2005.
[21] M. Papaleo, R. Firrincieli, S. Cioni, et al., "Link layer FEC in DVB-RCS: performance evaluation in nLoS conditions," in *Proceedings of the 67th IEEE Vehicular Technology Conference (VTC '08)*, pp. 2972–2976, Marina Bay, Singapore, May 2008.
[22] M. Papaleo, R. Firrincieli, G. E. Corazza, and A. Vanelli-Coralli, "On the application of MPE-FEC to mobile DVB-S2: performance evaluation in deep fading conditions," in *Proceedings of the International Workshop on Satellite and Space Communication (IWSSC '07)*, pp. 223–227, Salzburg, Austria, September 2007.
[23] F. Pérez-Fontán, M. Vázquez-Castro, C. E. Cabado, J. P. García, and E. Kubista, "Statistical modeling of the LMS channel," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 6, pp. 1549–1567, 2001.
[24] ETSI TM-SSP252-Revision 6 (2007-05), "Digital video broadcasting (DVB); DVB-SH implementation guidelines".