Frame Structure Design and Analysis for Millimeter Wave Cellular Systems

Sourjya Dutta, Student Member, IEEE, Marco Mezzavilla, Member, IEEE, Russell Ford, Student Member, IEEE, Menglei Zhang, Student Member, IEEE, Sundeep Rangan, Senior Member, IEEE, Michele Zorzi, Fellow, IEEE

Abstract—Millimeter-wave (mmWave) frequencies between 30 and 300 GHz have attracted considerable attention for fifth generation (5G) cellular communication as they offer orders of magnitude greater bandwidth than current cellular systems. However, the MAC layer may need to be significantly redesigned to support the highly directional transmissions, very low latencies and high peak rates inherent in mmWave communication. This paper analyzes two aspects of the mmWave MAC design. Firstly, this paper analyzes radio frame design options for a mmWave cellular aircell under the assumption of an LTE-like OFDM frame structure. Simple analytic formulae are derived to estimate the utilization as a function of key parameters such as the control periodicity, number of users, and channel and traffic statistics. It is found that certain flexible frame structures can offer dramatically improved utilization under various assumptions on the traffic pattern. Secondly, the beamforming choices are analyzed based on the control channel overhead that the link incurs. Analytical expressions for the overhead due to the physical layer control messages are derived as a function of control periodicity, signal-to-noise ratio, antenna gains and the frame design choice. It is shown that fully digital beamforming architectures offer significantly lower overhead compared to analog and hybrid beamforming under equivalent power budgets.

Index Terms—5G cellular systems, millimeter wave, frame structure, radio resource utilization, control overhead.

I. INTRODUCTION

The millimeter wave (mmWave) bands – roughly corresponding to frequencies above 10 GHz – have attracted considerable attention for next-generation cellular wireless systems [1]–[5]. These frequency bands offer orders of magnitude more spectrum than the congested bands in conventional UHF and microwave frequencies below 3 GHz. In addition, advances in CMOS RF circuits combined with the small wavelengths of mmWave frequencies enable large numbers of electrically steerable antenna elements to be placed in a picocellular access point or mobile terminal providing further gains via adaptive beamforming and spatial multiplexing. Preliminary capacity estimates demonstrate that this combination of massive bandwidth with large numbers of spatial degrees of freedom can enable orders of magnitude increases in capacity over current cellular systems [6], [7].

However, the design of mmWave cellular systems is still in its infancy with many outstanding design issues. While mmWave systems have been successfully used for satellite communications, cellular backhaul and wireless LANs, these applications consist generally of point-to-point links with limited mobility [8]–[12]. Cellular systems require additional mechanisms to support handover, track channel conditions, and coordinate interference and traffic between users, both within each cell and between neighboring cells.

In this work, we focus on one particularly important aspect of this design problem – namely the MAC-layer frame structure. By the frame structure we mean the time-frequency placement of all the relevant MAC-layer channels, including the data, assignments, acknowledgements (ACKs), and other control information to enable efficient use of the spectrum resources by the cell. While several groups have presented prototype designs [13]–[17], as we will see below, alternate design choices can obtain dramatically improved overhead and utilization.

The authors in [4] envision that 5G cellular systems will be highly heterogeneous and will require increased integration between different radio access technologies (RATs). Moreover, due to the propagation issues associated with mmWave links, it might also be necessary to concurrently use the 4G and 5G networks. Hence for our analysis, we assume a similar MAC layer structure as in 3GPP LTE [13]. Specifically, in the downlink, we assume that scheduling follows an identical hybrid automatic repeat request (H-ARQ) sequence as LTE with channel quality indicator (CQI) reports, downlink (DL) grants, downlink (DL) data and uplink (UL) ACK. Similarly, we assume the UL H-ARQ uses the LTE sequence of scheduling requests, UL grant, UL data and DL ACK. The frame structure design problem is then how to allocate the various control and data channels to meet latency, overhead and other requirements.

Similar to LTE downlink, we will also assume an orthogonal frequency division multiple access (OFDMA) waveform. OFDMA has the benefit of simple equalization and the ability to support orthogonal allocations in frequency and time. Whether OFDMA is the optimal choice for mmWave cellular remains to be determined. The purpose of this paper is to investigate the frame structure under an OFDMA assumption. However, most of our analysis abstracts out the details of the particular waveform and we thus believe that many of the concepts in our investigation can be applied to other systems as well.
A. Design Requirements and Challenges

The challenges in designing an efficient mmWave MAC-layer frame structure derive from the expectation that 5G cellular systems will support extremely high peak data rates with very low latency [5], [19]. In this work, we will consider the frame design under three key goals:

- **Ultra low-latency:** One of the most challenging goals is the desire to obtain round-trip (base to UE and back) airlink latencies of approximately 1 ms. Applications related to healthcare, logistics, automotive applications and mission-critical control will require this stringent bound on the latency [20], [21]. Moreover, as detailed in [22], real-time cyber physical experiences have similar latency requirements. This ultra low latency target is at least an order of magnitude faster than the minimum latency currently offered by 3GPP LTE (10 ms, see [23]). Of course, the actual latency of the system will also depend on hardware processing capabilities not within the scope of the study. However, for our purpose, we will consider a frame structure that can offer frequent opportunities for transmission of data and control to meet these targets.

- **Multiple users:** While 802.11 systems already offer sub-millisecond latencies, achieving similarly low latency in a multi-user cellular system is significantly more challenging. Cellular systems depend on careful scheduling between multiple users and cells to efficiently use the airlink and achieve high levels of spatial reuse. This scheduling demands significant control messaging. As we will see below, this control overhead grows with the number of users and one of the main objectives of this paper is to find efficient ways to accommodate multiple users and keep the ability to efficiently and rapidly allocate airlink resources.

- **Short bursty traffic:** One of the main attractions of mmWave is its ability to support multi-Gbps throughputs. Cellular communication systems will need to efficiently support these high data rates for both full buffer traffic and short bursty transmissions aggregated from multiple users. Short transmission bursts may be needed for RRC control messages, TCP ACKs, and applications that occasionally send short pieces of information. Based on our analysis, we argue that the frame structure design choices have significant impact on the utilization of the airlink in the presence of these short transmissions. In this work we show that frame design not only depends on latency and overhead metrics, but also is greatly influenced by the nature of the data traffic.

B. Contributions

The contributions of this work are as follows.

**System design:** We present two potential frame structure designs for a mmWave cellular MAC: (i) a fixed transmit time interval (TTI) structure similar to the one currently used in 3GPP LTE; and (ii) a dynamically scheduled TTI structure. The variable TTI structure was recently proposed in [24]. We investigate the details of the key downlink (DL) and uplink (UL) control channels.

**Design evaluation methodology:** We present a novel framework to evaluate the tradeoffs between two key performance criteria for the design options: (i) control overhead and (ii) resource utilization for small packets. The methodology is based on statistical distribution of SNRs and packet sizes and can thus be applied under a wide range of deployment and traffic assumptions.

**Effect of beamforming architectures:** Importantly, the proposed methodology elucidates the effect of different MIMO architectures on MAC layer design. Due to the wide bandwidths and large number of antenna elements in the mmWave range, it may not be possible from a power consumption perspective for the mobile receiver to obtain high rate digital samples from all antenna elements using fine quantization in the A/D conversion [25]. Most proposed designs perform analog beamforming (at RF or IF) prior to the A/D conversion [26]–[30]. A key limitation of these architectures is that they permit the mobile to “look” in only one or a small number of directions at a time. A key contribution of this work is to consider both analog and hybrid beamforming as well as a theoretical fully digital architecture that can look in all directions at once. To compensate for the high power consumption of the fully digital architecture, we consider a low-resolution design where each antenna is quantized at very low bit rates (say 2 to 3 bits per antenna) as proposed by [31].

**Design assessment in realistic cellular scenarios:** We use the model to assess various design options under realistic assumptions for next-generation cellular evolution. Our key findings reveal that (i) variable TTI structures can dramatically improve the utilization with small packets and also provide significant gains for latency; and (ii) fully digital architectures can significantly reduce the control utilization in settings with even moderate numbers of UEs per cell.

II. MIMO ARCHITECTURE MODELS

A. Transceiver Architectures

Before describing the channel structure, we need to consider the different beamforming (BF) capabilities available at the base station. As we will see, the MIMO processing assumptions will have a significant impact on the control overhead and latency.

In beamforming at conventional UHF and microwave frequencies, there is typically a separate RF chain and A/D conversion path for each antenna element. This architecture enables the most flexibility in that signals from the different antenna elements can be combined digitally as shown in Fig. [16] In the sequel, we will refer to this model as a fully digital architecture.

Unfortunately, in the mmWave range, due to the large number of antenna elements and wide bandwidths, it may not be possible from a power consumption perspective for the base station receiver to obtain high rate digital samples from all antenna elements [25]. Most proposed designs thus perform analog beamforming (at RF or IF) prior to the A/D conversion [26]–[30]. This model saves power by using only
a single A/D or D/A. However, the flexibility is reduced since the node can only beamform in one direction at a time. This model, shown in Fig. 1a, is called analog beamforming.

A hybrid between the digital and analog BF, shown in Fig. 1c, is called analog beamforming. It has to have K A/D or D/A conversion paths. Each of the K paths corresponds to a link between the BS and the UE called a stream. This is called hybrid beamforming and was proposed first in [29]. In the special case when K = 1, we obtain analog BF and when K = \( \frac{N}{\text{ant}} \), the number of antenna elements, we obtain the capability of fully digital BF. In hybrid BF, the base station can beamform and combine to K users at a time with the full antenna gain.

B. Beamforming Gains

To model the effect of directionality, let \( G_{BS} \) and \( G_{UE} \) denote the maximum directional gain achievable at the BS or UE when the beamforming is aligned along the BS-UE link. Due to reciprocity, we assume the same gain is achievable in either the transmit (TX) or receive (RX) direction. The exact directional gain will depend on the number of antenna elements, the multipath angular scattering and channel estimation accuracy, and may vary along different links. However, to simplify the analysis, we will assume that the maximum gains are constant across all links and given by

\[
G_{BS} = N_{\text{ant}}^{\text{ant}}_{BS}, \quad G_{UE} = N_{\text{ant}}^{\text{ant}}_{UE},
\]

where \( N_{\text{ant}}^{\text{ant}}_{BS} \) and \( N_{\text{ant}}^{\text{ant}}_{UE} \) are the number of antenna elements available at the BS and UE. The model in (1) holds exactly when the channel has a single angular path with no angular dispersion [32]. Studies of experimental data in [6] show that even in non line of sight (NLOS) channels with extensive scattering and long-term beamforming we can generally obtain a gain that is within 2 dB of this theoretical value, so we will make this assumption in the paper.

For fully digital BF, we will assume that the maximum gains can be obtained simultaneously for an arbitrary number of UEs since the BS can combine signals digitally. In contrast, for analog BF, it can only obtain the maximum directional gains for one user at a time. For hybrid BF with K streams, it can obtain the maximum gain\(^1\) for K users at a time. Note that, for hybrid BF, we have assumed that all antenna elements are available to all streams via splitters or combiners.

Even with analog and hybrid BF, it may be necessary for the BS to transmit to or receive from a number of users that exceeds the number of digital streams. In this case, the BS will not be able to obtain the full directional gain to the UE and must set the antenna pattern in a wide angle to transmit the signals to or receive them from all the UEs. We will let \( G_{\text{omni}}^{\text{BS}} \) denote the BS-side gain in this scenario. A conservative lower bound on this gain is given by

\[
G_{BS}^{\text{omni}} = K,
\]

where K is the number of digital streams. For analog BF, K = 1. In practice, the omni-directional gain may be larger if, for example, the UE are clustered angularly. Also, a recent work [33] has considered the problem of optimizing the K streams to receive from multiple UEs with known spatial patterns. This may increase the gain further. However, in the analysis, we will be conservative and always assume an omni-directional gain only as in (2).

C. Low Resolution Fully Digital BF

One possible solution to the high power consumption of fully digital architectures is to still have a full A/D conversion path for each antenna element, but to use very few bits per element – say 2 to 3 bits per I/Q dimension. Since the power

---

\(^1\)At the BS, both the signal power and the noise power are split into K parts, keeping the SNR constant. [33]
consumption of an A/D or D/A generally scales exponentially in the number of bits (i.e., linearly in the number of levels), using very low quantization resolutions can theoretically compensate for the large numbers of parallel A/D and D/A paths. This low resolution digital architecture has been considered in [31], [34], [35].

To model the effect of low quantization resolution, we will use an additive quantization noise model [36], as used in [37]. As described in [38], we can model the effect of quantization as a reduction in the SNR. Specifically, if a channel is received with an SNR of $\gamma$ per antenna, the SNR after quantization can be modeled as

$$\gamma' = \frac{\gamma}{1 + \alpha \gamma},$$

where $1/\alpha$ is an upper bound for the SNR, that depends on the number of bits and on the quantizer design.

### III. Frame Structure Models

Having described the MIMO architectures, we can now describe the frame structure and channelization options. We consider two basic models: fixed and variable TTI. In either structure, we will consider the same basic sequence of steps for UL and DL scheduling as currently performed in LTE, but with modifications to accommodate the multiple access and directionality constraints that characterize mmWave systems.

#### A. Fixed TTI

The time-frequency frame structure for this option is illustrated in Fig. 2. Time is divided into regular intervals with a fixed number of OFDM symbols per interval. This is the format used in the current LTE standard [18], where the subframes are 1 ms long and consist of either 12 or 14 symbols. A fixed TTI has also been considered for mmWave cellular systems in [1]. Following 3GPP terminology, we will call each interval a subframe. Each subframe is in turn divided into a control and a data portion.

The control portion is used for signaling various control messages such as ACKs, grants, channel quality indicator (CQI) report, etc. In Fig. 2 the control period is shown at the beginning of each subframe, but the precise location of the control channels is immaterial for the analysis in this paper. What is relevant is the total control resource usage, which we will detail in Section IV-C.

As standard in mmWave systems [3], we assume a dynamic time division duplex (TDD) system where the data portions in the subframes are allocated for UL or DL data in a dynamic manner. The top panel of Fig. 2 illustrates the sequence of events for scheduling and retransmission for transmitting DL data, while the bottom panel illustrates the corresponding timeline in the UL.

We assume, as in LTE, that frame times are aligned at the BS. The BS will perform continuous feedback timing control so that the UEs can advance their TX timing relative to their RX timing to ensure that signals are properly aligned at the BS. In a TDD system, this requires a small overhead for TX-RX switching. We discuss this overhead in detail in Section IV-C.

### DL Scheduling Timeline

We assume that the DL scheduling sequence follows the same basic steps as in LTE. To begin with, to facilitate DL scheduling, we assume that each UE in connected mode can be assigned dedicated uplink control resources similar to the Physical Uplink Control Channel (PUCCH) in LTE. These resources can be used to periodically transmit DL CQI reports where the mobile periodically reports the SNR and other channel characteristics for rate adaptation.

For a TDD system, the UL control channel can also be used for estimation of the DL beamforming. Note that the CQI reports will be continuously transmitted for all mobiles in connected mode, even when they are not scheduled data.

Based on the CQI and buffer states of the different UEs, the base station makes a scheduling decision in each subframe. For now, we will assume that the BS employs analog beamforming with a single phased array so that it can direct its transmissions during the data portion at only one UE at a time. Multiple UEs are supported via time division multiplexing across different subframes. Modifications in the event that the BS has the capability of transmitting to multiple UEs (via either hybrid or digital BF) are discussed in Section IV-C.

We assume that the scheduling decision, called a DL grant, is transmitted in a fixed time-frequency location within the data portion of the subframe and indicates the identity of the destination UE, the modulation and coding scheme (MCS) and other information to decode the data. Similar to LTE, we will assume that the system supports multi-process Hybrid Auto Repeat Request (HARQ) so the DL grant will also convey the HARQ process number, new data indicator and redundancy version so that, in case of retransmissions, the UE will know which prior transmissions to combine the current transmissions with. Aside from the DL grant, the remainder of the data portion can be used for the DL data. Since the DL grant and data are destined to the same UE, the BS can transmit both together in the same beamforming direction.

Each UE will attempt to decode the DL grant in every subframe. If the DL grant is decoded successfully and is for the UE, the UE will attempt to decode the remaining DL data portion based on the MCS and other information contained in the DL grant. Otherwise, the UE will ignore the remaining modulation symbols in the subframe.

A UE that has decoded the data successfully will transmit an UL ACK, with the absence of a transmission of an ACK indicating a negative ACK. As a simplification of LTE-TDD, we assume a simple structure where each subframe has dedicated resources for the UL ACK with a fixed delay between the DL data and the UL ACK. Specifically, we assume that the UL ACK for data transmitted in subframe $n$ will be sent in subframe $n + T_{\text{ack,DL}}$ where $T_{\text{ack,DL}}$ is the latency for the UE to decode the DL data. The BS will attempt to detect the ACK in the allocated time and, based on the presence or absence of the ACK, the BS can either transmit new data or transmit HARQ redundancy for the prior data. We will let $T_{\text{retx,DL}}$ be the minimum number of subframes for a retransmission, which shall include the time for decoding the ACK and scheduling.

#### UL scheduling timeline:

The UL scheduling timeline is shown in the bottom panel of Fig. 2 and again follows the
same basic sequence as in LTE. Specifically, after a packet arrives at a UE, we assume that the UE sends a scheduling request (SR) to the BS to indicate that it has data to transmit. Each UE must be allocated periodic resources to transmit the SR. As we will see in Section IV-C, since we have access to a large number of degrees of freedom in the mmWave context, we will consider multibit SRs, similar to the LTE full buffer status report (BSR). The multibit SR can enable a more fine grained report on the precise buffer levels at the UE.

After decoding the SRs, the BS makes a scheduling decision that allocates subframes to UEs in the UL direction. The UL grant is transmitted in the control period in some subframe and, similar to LTE, shall contain the MCS and UL power control information. A UE that decodes the UL grant will then transmit the data in the allocated subframe. We assume there is a fixed delay of $T_{\text{dat,UL}}$ between the UL grant and UL data transmission. The BS attempts to decode the UL data and, based on the decoding result, can allocate either a grant for new data or a HARQ retransmission.

B. Variable TTI

A shortcoming of the fixed TTI approach is that subframes may be highly underutilized when transmitting very small MAC PDUs. In the current LTE standard, this problem does not arise since small MAC PDUs can be allocated a very small portion of the bandwidth (as small as one resource block), with the remainder of the bandwidth being assigned to other UEs. However, for a mmWave base station that can only direct its beam to one user at a time, such FDMA scheduling is not possible, and thus the entire bandwidth must be allocated for the entire subframe.

The potential resource wastage is particularly dramatic for very wide bandwidth systems, as envisioned in the mmWave context. We will characterize this wastage precisely in Section IV-B.

To overcome this problem, we consider an alternate system with a variable length TTI, as proposed in [24]. In this option, the UL and DL grants can begin in any OFDM symbol and the grant has extra bits to signal the time duration of the data. Hence, the TTI times are dynamically scheduled instead of being fixed. As discussed in [24], the variable TTI sizes can also improve the delay for short packets.

The UL and DL timeline for this variable TTI structure is essentially identical to the fixed TTI structure in Fig. 2 except that the location and duration of the TTIs are also signaled in

---

3While FDMA scheduling would still be possible if the users are angularly close, this cannot be guaranteed in general.
the UL or DL grants. The time location for the ACKs could also be signaled in the grants, to enable the turnaround time to be a function of the UE or BS decoding capability.

Of course, many of the control signals such as scheduling requests and CQI reports need to be allocated to dedicated resources whose time locations would be fixed over relatively long periods (e.g., reallocation requires a higher level RRC reconfiguration). The dynamic UL and DL data transmissions would thus need to be allocated “around” these semi-statically fixed transmissions. One simple mechanism would be to assign UL and DL grants in sets of discontinuous time intervals, thereby allowing data TTIs to “skip” over any pre-assigned control signals. We will account for this extra signaling overhead in Section IV-C.

C. Enhancements with Hybrid or Digital Beamforming

Up to now, we have assumed that the BS can transmit and receive in only one direction at a time. This constraint arises when the BS performs analog beamforming (at RF or IF) with a phased array and a single A/D conversion path. As discussed in Section II however, the BS may be capable of hybrid or low-resolution fully digital beamforming, in which case it can transmit or receive in multiple directions at a time. In this case, the BS can perform either spatial division or frequency division multiple access (SDMA or FDMA) to serve multiple UEs simultaneously.

Thus, employing digital or hybrid beamforming for the transmission of PHY layer control messages will improve the overhead and latency of the overall system. As we find from our previous discussion, for both fixed and flexible TTI based designs the control symbols are more or less fixed. Hence the ability to transmit to or receive from multiple UEs at the same time will decrease the control overhead and allow concurrent processing of these messages.

For simplicity, in this work we will consider only FDMA when either hybrid or digital beamforming is available. We will see that even the simple use of FDMA dramatically reduces the control overhead. FDMA will imply that control signals to or from different UEs can be allocated in different orthogonal resources (such as different OFDM subcarriers) using beams directed individually to each UE. A BS with hybrid BF with \( K \) digital streams can multiplex up to \( K \) UEs with full directional gain. A fully digital architecture can support an arbitrary number of UEs, but is constrained by the number of frequency degrees of freedom.

For the transmission of PHY layer data we do not consider the use of digital or hybrid beamforming, as finely quantized multiplexing in the time domain can be achieved using the flexible TTI based design. Additionally, with a proper choice of scheduling algorithms (e.g., as discussed in [39]) the latency and resource utilization can be greatly improved without incurring the additional complexity associated with hybrid or digital BF.

IV. Design Analysis

A. Evaluation Model

SNR Distribution: We consider a single cell with \( N_{UE} \) UEs in connected state. We assume that the omni-directional DL and UL SNR for the \( i \)-th UE are given by

\[
\gamma_i^{DL} = \frac{P_{BS}}{L_i N_0 W_{tot}}, \quad \gamma_i^{UL} = \frac{P_{UE}}{L_i N_0 W_{tot}}
\]  

(4)

where \( P_{BS} \) and \( P_{UE} \) are the transmit powers of the BS and the UE; \( L_i \) is the path loss between the BS and the UE; \( N_0 \) is the thermal noise power spectral density (including the noise figure) and \( W_{tot} \) is the system bandwidth. The path loss values \( L_i \) are modeled as independent random variables with some distribution that depends on the path loss model and cell radius. In our evaluation section, we will use the mmWave statistical path loss model proposed in [9], but the framework can be applied to arbitrary distributions. Importantly, the SNRs in (4) are omni-directional SNRs in that they do not include the directional gain. Depending on the directionality of the transmissions, we will apply the gains as described in Section V-A.

Rounding: We will assume that all the channels must be an integer number of OFDM symbols. Hence, in many of our calculations for computing overhead and utilization, we will need to round up to the smallest number of integer symbols. Given an OFDM symbol period \( (T_{sym}) \), let

\[
Q(T) = T_{sym}\lceil(T/T_{sym})\rceil,
\]  

(5)

which is the value of time \( T \) rounded up to the nearest integer multiple of \( T_{sym} \).

B. Utilization

As discussed in Sec. III-C we consider analog beamforming based designs for the physical data channels (PDSCH or PUSCH). In order to quantify the utilization of the data channel we define the utilization factor \( \eta \) as the ratio between the minimum time required to transmit a particular data packet and the total time allocated for the corresponding transmission. In this section we analyze the difference between the fixed and the flexible TTI based designs in terms of the utilization factor \( \eta \).

Firstly, we consider a simple model with \( N \) users, each having one full buffer TCP flow. A full buffer flow implies that at every transmission opportunity, the radio resources available for data transmission are completely used.

Let us assume that we have \( N \) such flows in the DL, though the direction of the flow does not affect the analysis. For each of these \( N \) flows, the network expects TCP acknowledgments, which will be transmitted in the UL. TCP ACKs, unlike the data, have a small size. For example a 1500 byte TCP data can be acknowledged by a 74 byte ACK. Hence, the ACK will not need the maximum possible radio resource available and in fact, the ACKs can be treated as bursty data.

The number of TCP segments (\( S_N \)) that are transmitted in one time slot for the full buffer transmission is given as

\[
S_N = \frac{T_{TTI,max} \cdot \rho_{DL} W_{tot}}{L_{data}},
\]  

(6)

where \( L_{data} \) is the length of each segment in bits, \( \rho_{DL} \) is the spectral efficiency for downlink transmission and \( W_{tot} \) is the total bandwidth available for the mmWave system.
In the reverse direction, each of these $S_N$ segments will be acknowledged by one TCP ACK packet of length $L_{ack}$ which includes all packet overhead. Thus there are at most $S_N L_{ack}$ bits of data to be transmitted (we will have less data if ACKs are combined). The minimum time required to transmit the acknowledgements is

$$T_{min}^{ack} = \frac{S_N L_{ack}}{\rho_{UL} W_{tot}}, \quad (7)$$

where $\rho_{UL}$ is the UL spectral efficiency. The computation of the spectral efficiency is presented in the Appendix.

We define $T_{TTI,max}$ as a design specific constant which is the maximum TTI for a particular mmWave system. For the fixed TTI based design, all allocations made are for a duration of $T_{TTI,max}$ independent of the size of the data to be transmitted. Thus the utilization factor ($\eta_{fix}$) for the full buffer TCP flow is given by

$$\eta_{fix} = \frac{T_{TTI,max} + T_{min}^{ack}}{2T_{TTI,max}}. \quad (8)$$

A flexible TTI based system, on the other hand, will allocate a transmission time of $T_{TTI,max}$ in the downlink but on the uplink will allocate an integer number of symbols depending on the packet size. Thus the total time allocated for the transmission of a TCP ACKs is

$$T_{ack} = Q(T_{ack}^{min}). \quad (9)$$

The utilization factor for flexible TTI based design is,

$$\eta_{flex} = \frac{T_{TTI,max} + T_{ack}^{min}}{T_{TTI,max} + T_{ack}}. \quad (10)$$

In order to make more general predictions, we model the number of packets to be transmitted over the physical channel as a random variable that depends on the packet generation rate. The size of each of the packet transmitted during this period is modeled as a random variable $b$ with a known distribution. The minimum time required to transmit $b$ bits of data at the $i$th time slot is

$$T_i^{min} = \frac{b}{\rho W_{tot}}, \quad (11)$$

where $\rho$ is the spectral efficiency of the link and $W_{tot}$ is the system bandwidth.

Following the analysis for the full buffer model, we can write the time allocated to the bursty flows for both schemes as

$$T_i = \begin{cases} T_{TTI,max} : \text{Fixed} \\ Q(T_i^{min}) : \text{Flexible} \end{cases} \quad (12)$$

Unlike the analysis of the first traffic model, it must be noted that the $T_i$’s are random variables, as they are a function of $b$ and $\rho$. For the fixed TTI based design the expected value of the utilization factor calculated over a large time duration is given by

$$\eta_{fix} = \frac{1}{W_{tot}} \mathbb{E} \left[ \frac{1}{\rho} \frac{\mathbb{E}[b]}{T_{TTI,max}} \right]. \quad (13)$$

The utilization factor for the flexible scheme on the other hand is

$$\eta_{flex} = \frac{1}{W_{tot}} \mathbb{E} \left[ \frac{1}{\rho} \frac{\mathbb{E}[b]}{T_{TTI,max}} \right]. \quad (14)$$

C. Overhead

**Scheduling request:** Assume that there must be a dedicated opportunity for each UE to transmit a scheduling request (SR) at least once every $T_{per,SR}$ seconds. The value of $T_{per,SR}$ is one component of the UL delay. Let $b_{SR}$ be the number of bits in the SR and $\gamma_b$ be the minimum $E_b/N_0$ for the channel. When the BS has analog BF, there are two options to receive the SR. First, it can receive the SRs in a TDMA manner in which case it gets the full directional gain on both the BS and the UE side. Thus, the time to receive the SR for each UE will be at least

$$T_{SR} \geq Q \left( \frac{b_{SR} \gamma_b}{G_{BS} G_{UE} W_{tot} \gamma_{UL}} \right), \quad (15)$$

where $Q(\cdot)$ is the operator in [5] that rounds up to the nearest number of OFDM symbols. In [15], we have assumed that the time slot width for the UE is dimensioned for the worst-case UE and hence we have used the minimum UL SNR, $\gamma_{UL}$. For $N_{UE}$ UEs, the total overhead using TDMA for the SR is:

$$\alpha_{SR,TDMA} = \frac{N_{UE} T_{SR}}{T_{per,SR}} = \frac{N_{UE}}{T_{per,SR}} Q \left( \frac{b_{SR} \gamma_b}{G_{BS} G_{UE} W_{tot} \gamma_{UL}} \right). \quad (16)$$

An alternate option is for the UEs to transmit the SRs simultaneously on different frequency resources, and have the BS receive the SRs omni-directionally. Assuming that the number of degrees of freedom is sufficient (which is likely since the bandwidth is large), the overhead using this FDMA scheme is

$$\alpha_{SR,FDMA} = \frac{1}{T_{per,SR}} Q \left( \frac{b_{SR} \gamma_b}{G_{omni} G_{BS} G_{UE} W_{tot} \gamma_{min}^{UL}} \right), \quad (17)$$

where the BS uses the omni-directional gain. Finally, for digital BF, the BS can receive all the SR signals simultaneously while obtaining the directional gain, so the overhead fraction is

$$\alpha_{SR,Dig} = \frac{1}{T_{per,SR}} Q \left( \frac{b_{SR} \gamma_b}{G_{omni} G_{BS} G_{UE} W_{tot} \gamma_{min}^{UL}} \right). \quad (18)$$

The number of bits $b_{SR}$ will depend on the quantization resolution for the buffer status. More bits will be required in the variable TTI mode so that the BS can schedule the correct TTI size. Once the UE is scheduled, it can transmit a full buffer status report inband as in LTE. Additionally, in either option, we will need further bits for a CRC to prevent false alarms that would result in UL allocation wastage. We will describe this in Section [V].
UL CQI and other control: In LTE, in order to enable the BS to perform rate adaptation in the downlink, each UE continuously transmits channel quality indicator (CQI) reports on a dedicated uplink control channel. The dedicated UL control channel is also used for transmitting indications of the channel spatial rank for MIMO. The overhead for this channel can be computed identically to the SR report above. The resulting expressions will depend on $T_{	ext{per, CQI}}$, the periodicity of the CQI reports, $b_{\text{CQI}}$, the number of bits per report, and $\gamma_b$, CQI, the $E_b/N_0$ for the channel.

UL grants: The overhead taken by the UL grants depends critically on the MIMO architecture. If the BS has only analog BF, it can transmit in only one direction at a time, and therefore cannot transmit a UL grant and a transmission to another UE simultaneously. Thus, the UL grant must be transmitted by itself across the entire bandwidth as shown in Fig. 3a. If we let $b_g$ be the number of bits in the grant, and $\gamma_{bg}$ be the minimum $E_b/N_0$ for the channel.

$$T_{\text{ULG, ana}} = Q \left( \frac{b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \gamma_i^\text{DL}} \right).$$

(19)

Now, say that a fraction $p_{\text{UL}}$ of the TTI is allocated for the UL and $1 - p_{\text{UL}}$ for the DL. Since there is one UL grant for every TTI allocated to the UL, the overhead for the UL grants with analog BF will be

$$\alpha_{\text{ULG, ana}} = \frac{p_{\text{UL}}}{E[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ Q \left( \frac{b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \gamma_i^\text{DL}} \right) \right].$$

(20)

where the expectation in $\mathbb{E}[T_{\text{TTI}}]$ is taken over the TTI sizes, and the expectation in $\mathbb{E}[Q(\cdot)]$ is taken over the variability in the DL SNR. Note that in the fixed TTI size case $T_{\text{TTI}}$ is a fixed value.

When the BS has hybrid BF with $K > 1$ streams or digital BF, the BS can transmit the UL grant in a fraction of the bandwidth while transmitting DL data to other UEs in the remainder of the bandwidth as shown in Fig. 3b. To evaluate the overhead in this case, suppose that the UL grant is sent $T_g$ seconds over a bandwidth of $W_g$. If the DL power is allocated uniformly across the total bandwidth $W_{\text{tot}}$, then the grant will be received at UE $i$ with an SNR of $\gamma_i^\text{DL}$. Hence, the minimum transmission time for the grant with digital beamforming or multi-stream hybrid beamforming will be

$$T_{\text{ULG, dig}} = Q \left( \frac{b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \gamma_i^\text{DL}} \right),$$

and the overhead will be

$$\alpha_{\text{ULG, dig}} = \frac{W_{\text{g}} p_{\text{UL}}}{W_{\text{tot}} \mathbb{E}[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ Q \left( \frac{b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \gamma_i^\text{DL}} \right) \right].$$

(21)

Now, in principle, $W_g$ can be adjusted so that there is no rounding error in the $Q(\cdot)$ function so that $Q(x) = x$. In this case, (21) simplifies to

$$\alpha_{\text{ULG, dig}} = \frac{p_{\text{UL}} b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \mathbb{E}[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ \frac{1}{\gamma_i^\text{DL}} \right].$$

(22)

which is the same expression as in the analog case (20), but without the quantization. In fact, we remark that the additional overhead incurred by the analog beamforming architecture is due to the difference between $T_{\text{ULG, ana}}$ and its quantized value.

DL grants: Since the DL grant and corresponding DL data are transmitted to the same UE, they can be multiplexed together as shown in Fig. 3c. This multiplexed transmission can be performed for all the beamforming architectures. Thus following the analysis provided for the UL grant, we can express the overhead due to the DL grant as

$$\alpha_{\text{DLG}} = \frac{(1 - p_{\text{UL}}) b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \mathbb{E}[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ \frac{1}{\gamma_i^\text{DL}} \right].$$

(23)

DL and UL ACKs: The DL ACK is sent by the BS in response to UL data received from the UE. Its time frequency allocation has the same constraints as the UL grant. Applying a similar derivation we obtain that with analog beamforming the DL ACK overhead is given by

$$\alpha_{\text{DLACK, ana}} = \frac{p_{\text{UL}} b_g \gamma_{bg}}{G_{BS} G_{UE} W_{\text{tot}} \mathbb{E}[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ \frac{1}{\gamma_i^\text{DL}} \right].$$

(24)

where $b_{\text{ACK}}$ is the number of bits per ACK and $\gamma_{\text{ACK}, \text{DL}}$ is its $E_b/N_0$ requirement. Note that the number of bits may be greater than one due to separate ACKs in different spatial streams as in 3GPP or in the case when subunits within the MAC PDU are ACK-ed individually. Similarly for the case of hybrid beamforming with $K > 1$ digital streams or fully digital beamforming, the overhead is

$$\alpha_{\text{DLACK, dig}} = \frac{p_{\text{UL}} b_g \gamma_{bg} b_{\text{ACK}}}{G_{BS} G_{UE} W_{\text{tot}} \mathbb{E}[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ \frac{1}{\gamma_i^\text{DL}} \right].$$

(25)

The DL ACK is sent by the UE in response to DL data and has similar time-frequency options as the UL ACK, but different overhead. In the analog BF case, it must be sent by itself and the overhead is

$$\alpha_{\text{ULACK, ana}} = \frac{1 - p_{\text{UL}}}{\mathbb{E}[T_{\text{TTI}}]} \cdot \mathbb{E} \left[ \frac{b_g \gamma_{bg} b_{\text{ACK}}}{G_{BS} G_{UE} W_{\text{tot}} \gamma_i^\text{DL}} \right].$$

(26)

In hybrid BF with multiple digital streams or in digital BF, the UL ACK can be multiplexed with UL data from other UEs. In this case, the UE can transmit all its power on the ACK, and the ACK will only be bandwidth limited. Thus, the allocation will only be limited by the spectral efficiency of the UL ACK. Suppose that the UL ACK is transmitted at a spectral efficiency $\rho_{\text{ACK}}$. Then, the UL ACK will require $b_{\text{ACK}}/\rho_{\text{ACK}}$. 


Table I: Parameters used for the system evaluation.

| Parameter | Description                                                                 | Value used |
|-----------|------------------------------------------------------------------------------|------------|
| $\gamma_{\text{th,M}}$ | SNR per bit ($E_b/N_0$) for a given control message M | 6 dB       |
| $G_{\text{BS}}$ | Directional beamforming gain for the base station | 18 dB      |
| $G_{\text{UE}}^{\text{mm}}$ | Max beamforming gain for the BS assuming $N_{\text{UE}}^{\text{BS}} = 64$. | 0 dB       |
| $G_{\text{UE}}$ | Max beamforming gain for the user equipment. Assumes $N_{\text{UE}}^{\text{BS}} = 16$. | 12 dB      |
| $T_{\text{TTI}}$ | Transmission time interval | 125 $\mu$s |
| $t_{\text{sym}}$ | Duration of one OFDM symbol | 4.16 $\mu$s |
| $T_{\text{per,SR}}$ | Period in which all UEs can transmit a SR at least once | 500 $\mu$s |
| $W_{\text{tot}}$ | Total system bandwidth | 1000 MHz |
| $\rho_{\text{ACK}}$ | Spectral efficiency for ACK transmission | $\frac{8}{5}$ bps/Hz |
| $p_{\text{UL}}$ | Fraction of uplink packets | 0.5        |
| $b_{\text{SR}}$ | Size of a scheduling request | 18, 26, 42 bits |
| $b_{\text{UL}}$ | Size of a UL/DL grant | 80, 100 bits |
| $b_{\text{ACK}}$ | Size of a HARQ acknowledgement | 5 bits     |

which will be negligible in the bandwidths of interest for mmWave systems.

V. Evaluation for Realistic Design Scenarios

A. System Assumptions

a) System Parameters: We leverage the above utilization and control overhead analysis to evaluate the different frame structure and signaling options under realistic design scenarios. The parameters are shown in Table I. Following the capacity analysis in [6], we assume $N_{\text{BS}}^{\text{UE}} = 64$ antennas at the BS and $N_{\text{UE}}^{\text{BS}} = 16$ at the UE – reasonable dimensional arrays for mmWave systems. The BS serves $N_{\text{UE}}$ users. We vary $N_{\text{UE}}$ for one BS to capture the effect it has on the control overhead of the system using the parameters in Table I.

b) SNR Distribution: The transmit powers in the DL and UL directions are taken to be 30 dBm and 20 dBm respectively. The noise figures are 7 dB and 4 dB in DL and UL, respectively, consistent with capacity analysis in [6].

We used a path loss model in [6] derived from actual measurements in 28 GHz in New York City [2]. The distribution of the SNRs $\gamma_{\text{DL}}$ and $\gamma_{\text{UL}}$ are then generated from this model under the assumption that the mobiles are uniformly located in the cell radius of 100 m from the base station. The resulting distribution is shown in Fig. 4 along with the 5% and median lines. For the uplink the 5% and median SNRs are -39 dB and -16 dB respectively. For the downlink, the respective values are -32 dB and -9 dB. The 5% values are used as the target SNRs for the minimum SNRs $\gamma_{\text{DL}}^{\text{min}}$ and $\gamma_{\text{UL}}^{\text{min}}$.

c) OFDM Symbol Period and Guard Times: The OFDM symbol period and TTIs are critical in determining the overhead for transmitting small control and data messages – a key aspect of the frame structure design. Very short OFDM symbols and TTIs allow time to be divided into small intervals enabling small data packets to be transmitted with minimal padding. On the other hand, each OFDM symbol must contain a cyclic prefix (CP) whose size is determined by the channel delay spread and synchronization errors. Thus, reducing the OFDM symbol period increases the overhead for a fixed cyclic prefix size.

In the work, we follow the suggested OFDM parameters in [13]. That work proposes, for a 1 GHz bandwidth, a sampling rate of 1106 MHz with a subcarrier spacing of 270 KHz employing a 4096 point FFT. The OFDM symbol duration hence is 3.70 $\mu$s. For small cells (<1 Km in radius), a CP duration of 0.463 $\mu$s is used, giving a total symbol period of 4.16 $\mu$s. This CP duration is sufficient for delay spreads as measured in [40], [41]. The work [13] also uses a fixed TTI with 30 OFDM symbols corresponding to $T_{\text{TTI}}$ of 125 $\mu$s. We use this value for the fixed TTI.

Other OFDM-based mmWave designs have used similar values. For example, in [42], a subcarrier spacing $\Delta f = 180$ KHz is proposed with $N = 512$, giving a sampling frequency of 92.16 MHz and symbol period of 5.56 $\mu$s. Moreover in [43], using 512 point FFT, the authors propose a slot length of 17 symbols where the first (long) symbol is 5.9 $\mu$s and the next 16 short symbols have length 5.88 $\mu$s. The slot length is thus 100 $\mu$s. Additionally the authors also discuss the design with 16 long symbols and with higher values of $N$.

The selection of the other parameters in Table I are justified in the Appendix.

B. Evaluation Results

1) Control Overhead: Following the analysis in Section IV-C, we compare the overheads due to the physical layer con-
TABLE II: Control message overheads for the various design alternatives with $N_{UE} = 8$.

| Control Message | Message Type | Analog TDMA | Hybrid (K=2) TDMA | Digital TDMA |
|-----------------|--------------|-------------|------------------|-------------|
| Scheduling Request | Trigger      | 0.069  | 0.0750 | 0.0333 | 0.0083 |
|                  | Short        | 0.069  | 0.1083 | 0.0333 | 0.0083 |
|                  | Long         | 0.069  | 0.1667 | 0.0333 | 0.0083 |
| Uplink Grant     |              | 0.0167 | N.A    | 0.000177 | 0.000184 |
| Downlink Grant   |              | 0.000177 | N.A | 0.000177 | 0.000177 |
| HARQ Ack         | DL           | 0.0167 | N.A    | 0.00009 | 0.00009 |
|                  | UL           | 0.0167 | N.A    | 0.00016 | 0.00016 |
| Total            |              | 0.1170 | 0.0339 | 0.0089 |

Fig. 5: Control Overhead versus the number of users for analog, hybrid (K = 2) and digital beamforming architectures.

trol signals for analog, hybrid and fully digital beamforming architectures. As an example, the overhead due to the various control signals when a BS serves 8 users is listed in Table II. We note that for analog beamforming the overhead is around 12% while with a $K = 2$ hybrid architecture the overhead dips to 3%. We notice for both these cases that the scheduling request (SR) dominates overhead. For the low power fully digital architecture it is less than 1%. Thus, as discussed in Section III-C the overhead is considerably reduced when hybrid or fully digital beamforming is used.

Fig. 5 plots the overhead as a function of the number of users served by the base station. The linear increase of the overhead with the number of users is attributed to the increase in UL SRs. In Fig. 5 the switch to FDMA based transmission for SRs accounts for the plateau reached by the curves for analog and hybrid beamforming. Thus we can say that when the number of users in high, analog or hybrid beamforming based systems should employ FDMA for UL SR transmission. We note that for analog beamforming based architectures, even for a smaller number of connected users, the overhead is considerably higher than for digital or hybrid architectures. Moreover, for fully digital architecture, the overhead is constant even when the number of connected users grows.

It should be noted here that this gain in overhead comes at the price of increased hardware complexity and power consumption for the hybrid and the fully digital architectures.

In order to limit the power consumption we use low resolution ADCs for the fully digital scheme, which account for the adjustment in effective SNR as given by (3).

2) Utilization:

Utilization with TCP ACKs: Next we compare the utilization of allocated radio resources for the fixed and flexible TTI designs based on the analysis in Section IV-B. Fig. 6 captures the effect of the maximum TTI ($TTI_{max}$) on the utilization for the full buffer TCP model. We see that the fixed TTI scheme gives a constant utilization of around 53% regardless of the TTI size. This implies that there is a dramatic wastage in bandwidth using a fixed TTI. This loss occurs since the TCP ACK packets are much smaller than the TCP data packets. In the fixed TTI mode, the TCP ACK packets are sent in the same TTI frame size as the TCP data packets, and hence the TTIs for the TCP ACKs are essentially wasted.

In contrast, with the flexible TTI scheme, increasing $TTI_{max}$ enables more TCP data packets to be transmitted per TTI on the forward link, which results in more ACKs in the reverse link. But as ACKs are much smaller than the data, these can be transmitted over one or two symbols. Thus for the flexible TTI scheme, the utilization is comparable with the fixed design when $TTI_{max}$ is small but rapidly improves as $TTI_{max}$ increases.

As a second test for utilization, Based on (13) and (14), we compute the resource utilization in the case when the MAC PDUs have large size but arrive at a slow rate. The number of PDUs available for a user in each time interval is modeled as a Poisson random variable with mean 1 packet/s. The sizes of the PDUs (in bytes) are truncated log-normal random variables, between 0.5 MB to 5 MB, with mean PDU size of 2 MB and standard deviation 0.722 MB. Each delivered packet is acknowledged with an ACK by the receiver. Fig. 6 plots the utilization versus $TTI_{max}$ for $N_{UE} = 32$. For the flexible TTI based scheme the trend is similar to the TCP full buffer case. Conversely, with increasing $TTI_{max}$ the utilization for the fixed TTI scheme degrades, as for larger values of $TTI_{max}$, even large data packets drawn from this distribution will not be able to fully utilize the allocated resources.

We next analyzed the utilization with small packet sizes and high average arrival rate. We used the model in [44] with an arrival rate of 5 packets/s, where the packet sizes to be truncated log-normal random variables with mean PDU size 10710 B and standard deviation of 25032 B. The packet sizes generated are between 100 B and 2 MB. Fig. 6 shows the variation of the utilization with $TTI_{max}$ for this scenario. We observe that for fixed TTI based design the utilization decreases rapidly from 20% to 1% as $TTI_{max}$ increases. For flexible TTI based design, the utilization remains somewhat constant around 22%. This is because in this case, most of the time only one symbol is allocated for data transmission in both directions, and thus a constant amount of allocated radio resource is being utilized on average.

Thus we see that a flexible TTI-based design offers significantly improved utilization in comparison to a fixed TTI design in networks where data packets are short and bursty.
Fig. 6: Utilization vs. maximum TTI with symbol length 4.16 $\mu$s for fixed and flexible TTI based frame design with (a) full buffer data, (b) large packets (0.5 MB to 5 MB) arriving at a rate 1 per second (c) packets between 100 B and 2 MB arriving at a rate of 3 per second. ($N_{UE} = 32$)

Fig. 7: The effect of small RRC packets on the data rate of the users for fixed and flexible TTI based designs ($N_{UE} = 8$).

Aggregating packets over multiple arrivals for the same UE may mitigate this problem and allow us to use the fixed design. However, we must note that this approach will severely affect the latency of the system and will not be acceptable for packets with strict deadlines. Moreover, aggregating small packets to saturate an entire subframe might require a wait time longer than that acceptable by most applications.

Utilization with small control messages: Another important source of small MAC PDUs arises for radio resource control (RRC) control messages. RRC messages will be used for a variety of control signaling including interference coordination, measurement reports, resource allocation, etc. 5G mmWave systems will likely need a greater rate of control signaling to handle the more rapid fluctuations in the channel. To analyze this situation, we consider the case when the BS is serving $N_{UE} = 8$ users. Each user is transmitting full buffer traffic in the UL or the DL. At the same time, the BS sends small control messages to each UE at a fixed rate. Similar to current RRC messages such as [45], we assume that each control message is 2000 bits. Fig. 7 plots the change in the data rate, with the increase RRC layer control messages. It can be seen that for the fixed TTI based scheme the data rate falls rapidly with the increase in the rate of RRC messages. However, for the flexible TTI scheme, instead the data rate decreases very slowly (if at all).

This experiment also gives some insight on the effect that short machine type communications (MTC) [21] will have on networks. MTCs are characterized by short data packets in the order of 100 bytes with high priority. A key goal for 5th generation systems is to incorporate a large number of MTC devices. From our experiment we can draw the inference that if the fixed TTI based design is used, the data rates of human-to-human links will degrade considerably. Whereas systems where the flexible TTI scheme is employed can transmit machine type data packets with a negligible effect on user experience.

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have discussed in details the MIMO architectures and the frame design choices available for the design of mmWave cellular systems. Additionally, we have outlined the different options for the physical layer control channels, modeled along the lines of LTE, for the mmWave radio link, and suggested enhancements based on the frame design.

In order to quantify the performance of the mmWave MAC layer we have analyzed the overhead due to the control messages and the utilization of the available network resources. We have analyzed the fixed TTI based design against the flexible TTI based design, in terms of how well they utilize allocated radio resources, and found that the flexible frame structure outperforms the fixed structure in all the traffic scenarios discussed, especially for the bursty traffic with small packet sizes.

For the physical control channel, we have considered the analog, hybrid and fully digital beamforming architecture
based designs. We have shown that the control overhead is negligible with fully digital architectures even when implemented using low resolution ADCs. Thus we can state that for the equal power consumption levels a fully digital design with low resolution ADCs gives a considerably lower control overhead than analog or hybrid beamforming based systems.

A more complete MAC layer analysis will require a study of the effect that the transmission and processing of the control messages have on the data in terms of latency and throughput which is an interesting direction for future work.

APPENDIX

Simulation Parameter Selection Details

In this section we delineate the logic behind the selection of the system parameters used for the simulations and the results. Moreover, we also illustrate some of the considerations that should be made in selecting these values for a practical system.

a) Antenna pattern: We assume a set of two dimensional antenna arrays at both the BS and the UE. On the BS side, the array is comprised of 8 × 8 elements and on the UE side we have 4 × 4 elements. The spacing of the elements is set at λ/2, where λ is the wavelength. These antenna patterns were considered in [6] and shown to offer excellent system capacity for small cell urban deployments. In addition, a 4 × 4 array operating in the 28 GHz band, for instance, will have a size of roughly 1.5 cm × 1.5 cm.

The maximum gain that can be achieved by beamforming with an Nt element antenna array, as pointed out in [6], is given, in dB, as 10 log10 Nt. Thus for the 8 × 8 elements array at the base station the maximum beamforming gain is 10 log10 64 = 18 dB. For the UE the maximum beamforming gain is 12 dB.

b) Spectral Efficiency: The spectral efficiency, ρ, for a given channel is given in [6] as,

\[ \rho = \min \left( \alpha \log_2 \left( 1 + 10^{0.1(SNR-\Delta)} \right), \rho_{\text{max}} \right), \]  

(28)

where α is the bandwidth utilization factor, Δ is the loss factor (in dB) and ρ_{max} is the maximum spectral efficiency. From [6], we get the values Δ = 3 dB and ρ_{max} = 4.8 bps/Hz. The value of α is taken as 0.83, the same as that of LTE as reported in [46].

The spectral efficiency for UL ACK (ρ_{ACK}) is the minimum spectral efficiency required to transmit 100 ACKs over one symbol. Thus

\[ \rho_{\text{ACK}} = \left( \frac{100 b_{\text{ACK}}}{T_{\text{sym}} W_{\text{tot}}} \right) \approx \frac{1}{8} \]

is considered for the calculation of control overhead.

c) Control Message Size: The LTE scheduling request is a trigger that notifies the BS that the user has data to transmit, and carries no further information. For the design with fixed TTI we will use the same scheme for the SR, and to prevent errors and mis-detection a 16-bit CRC is used with the SR. Thus, the size of a SR becomes 18 bits, with 2 bits set as priority bits. In order to provide the scheduler with a more complete information about the UE buffer for the flexible TTI based design we propose that the SR should resemble the BSR.

In our analysis we consider the 8-bit short BSR and the 24-bit full BSR. This with the CRC and the priority accounts for the SR to be either 26 or 42 bits long.

For simplicity, we consider the downlink and the uplink grant to be of the same size. In our analysis we assume that the grants will be 80 bits long for the fixed TTI case and 100 bits long for the flexible TTI based design. The values assumed are nearly double of those used in LTE as we are using higher order MIMO antennas and also have a much higher bandwidth. Moreover, some additional bits are required for the flexible TTI based design to specify the symbols (within a frame or a subframe) which are used for each of the transmissions. The size specified include an attached 16-bit CRC like that of LTE downlink control informations (DCIs).

Considering maximum spectral efficiency (ρ_{max}), for transmission over a 1 GHz bandwidth (W_{tot}) for a slot of period (T) 125 µs, the maximum number of bits that can be transmitted is equal to the number of available degrees of freedom

\[ \rho_{max} W_{tot} T = 600,000. \]

This implies that a maximum of 600,000 bits can be transmitted over this time slot. This accounts for 50 TCP packets, each of 1500 bytes. Sending one HARQ acknowledgement every ten 1500 bytes of data transmitted will lead to a maximum of 5 ACK packets in the control channel, each ACK being a 1 bit message.

REFERENCES

[1] F. Khan and Z. Pi, “An introduction to millimeter-wave mobile broadband systems,” IEEE Commun. Mag., vol. 49, no. 6, pp. 101 – 107, Jun. 2011.
[2] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, “Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!!” IEEE Access, vol. 1, pp. 335–349, May 2013.
[3] S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter-wave cellular wireless networks: Potentials and challenges,” Proc. IEEE, vol. 102, no. 3, pp. 366–385, Mar. 2014.
[4] J. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. Soong, and J. Zhang, “What will 5G be?” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1065–1082, June 2014.
[5] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratrasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. MacCartney, S. Sun, and S. Nie, “Millimeter wave enhanced local area systems: A high data rate approach for future wireless networks,” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1152–1163, June 2014.
[6] M. Akdeniz, Y. Liu, M. Samimi, S. Sun, S. Rangan, T. Rappaport, and E. Erkip, “Millimeter-wave channel modeling and cellular capacity evaluation,” IEEE J. Sel. Areas Commun., vol. 32, no. 6, pp. 1164–1179, June 2014.
[7] T. Bai and R. Heath, “Coverage and rate analysis for millimeter-wave cellular networks,” IEEE Trans. Wireless Commun., vol. 14, no. 2, pp. 1100–1114, Feb. 2015.
[8] J. Hansry, J. Edstam, B.-E. Olsson, and C. Larsson, “Non-line-of-sight microwave backhaul for small cells,” Ericsson Review, Feb. 2013.
[9] T. S. Rappaport, J. M. Murdock, and F. Gutierrez, “State of the art in 60-GHz integrated circuits and systems for wireless communications,” Proc. IEEE, vol. 99, no. 8, pp. 1390 – 1436, Aug. 2011.
[10] T. Baykas, C.-S. Sum, Z. Lan, J. Wang, M. A. Rahman, H. Harada, and S. Kato, “IEEE 802.15.3c: The first IEEE wireless standard for data rates over 1 Gb/s,” IEEE Comm. Mag., Jul. 2011.
[11] R. C. Daniels and R. W. Heath Jr, “60 GHz wireless communications: emerging requirements and design recommendations,” IEEE Veh. Technol. Mag., vol. 2, no. 3, pp. 41–50, Sept. 2007.
[12] R. Daniels, J. Murdock, T. S. Rappaport, and R. Heath, “60 GHz Wireless: Up Close and Personal,” IEEE Microwave, vol. 11, no. 7, pp. 44 – 50, Dec. 2010.
F. Khan, Z. Pi, and S. Rajagopal, “Millimeter-wave mobile broadband with large scale spatial processing for 5G mobile communication,” in Proc. IEEE 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Oct. 2012, pp. 1517–1523.

Z. Pi and F. Khan, “A millimeter-wave massive MIMO system for next generation mobile broadband,” in Proc. IEEE 46th Asilomar Conference on Signals, Systems and Computers (ASILOMAR), Nov. 2012, pp. 693–698.

Z. Pi, Y. Li, and F. Khan, “Methods and apparatus to transmit and receive synchronization signal and system information in a wireless communication system,” Dec. 30 2014, US Patent 8,923,197.

P. Kela, M. Costa, J. Salmi, J. Turkkka, T. Hiltunen, and M. Hronec, “A novel radio frame structure for 5G dense outdoor radio access networks,” in Proc. IEEE 81st Vehicular Technology Conference (VTC Spring), May 2015, pp. 1–6.

E. Lhetkangas, K. Pajukoski, J. Vihrl, G. Berardinelli, M. Lauridsen, E. Tirola, and P. Mogensen, “Achieving low latency and energy consumption by 5G TDD mode optimization,” in Proc. IEEE International Conference on Communications Workshops (ICC), June 2014, pp. 1–6.

E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, 3G Evolution: HSPA and LTE for Mobile Broadband. Oxford, UK: Academic Press, 2007.

F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5G,” IEEE Commun. Mag., vol. 52, no. 2, pp. 74–80, Feb. 2014.

A. Osseiran, F. Boccardi, V. Braun, P. Kekki, M. Martin, E. Tiirola, and P. Mogensen, “3G Evolution: LTE for Mobile Broadband,” Oxford, UK: Academic Press, 2007.

F. Boccardi, R. W. Heath Jr, A. Lozano, T. L. Marzetta, and P. Popovski, “Five disruptive technology directions for 5G,” IEEE Commun. Mag., vol. 52, no. 2, pp. 74–80, Feb. 2014.

A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, A. Ghosh et al., “Scenarios for 5G mobile and wireless communications: the vision of the METIS project,” IEEE Commun. Mag., vol. 52, no. 5, pp. 26–35, May 2014.

H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, A. Ghosh et al., “Machine-type communications: current status and future perspectives toward 5G systems,” IEEE Commun. Mag., vol. 53, no. 9, pp. 10–17, Sept. 2015.

G. Fettweis and S. Alamouti, “5G: Personal mobile internet beyond what cellular did to telephony,” IEEE Commun. Mag., vol. 52, no. 2, pp. 140–145, Feb. 2014.

“LTE-capable transport: A quality user experience demands an end-to-end approach,” Nokia Siemens Networks white paper, 2011, available online at http://www.nokiasiemennetworks.com.

T. Levanen, J. Pirskanen, and M. Valkama, “Radio interface design for ultra-low latency millimeter-wave communications in 5G era,” in Proc. IEEE Globecom Workshops (Gc Wkshps), Dec. 2014, pp. 1420–1426.

F. Khan and Z. Pi, “Millimeter-wave Mobile Broadband (MMB): Unleashing 3-300 GHz Spectrum,” in Proc. IEEE Sarnoff Symposium, Mar. 2011.

K.-J. Koh and G. M. Rebeiz, “0.13-μm CMOS phase shifters for X-, Ku- and K-band phased arrays,” IEEE J. Solid-State Circuits, vol. 42, no. 11, pp. 2535–2546, Nov. 2007.

—, “A Millimeter-Wave (40–45 GHz) 16-Element Phased-Array Transmitter in 0.18-μm SiGe BiCMOS Technology,” IEEE J. Solid-State Circuits, vol. 44, no. 5, pp. 1498–1509, May 2009.

X. Guan, H. Hashemi, and A. Hajimiri, “A fully integrated 24-GHz eight-element phased-array receiver in silicon,” IEEE J. Solid-State Circuits, vol. 39, no. 12, pp. 2311–2320, Dec. 2004.

A. Alkhateeb, O. E. Ayach, G. Leus, and J. Robert W. Heath, “Hybrid precoding for millimeter wave cellular systems with partial channel knowledge,” in Proc. Information Theory and Applications Workshop (ITA), Feb. 2013.