Hybrid Beamforming for Massive MIMO: A Survey

Andreas F. Molisch, Vishnu V. Ratnam, Shengqian Han, Zheda Li, Sinh Le Hong Nguyen, Linsheng Li, and Katsuyuki Haneda

Hybrid multiple-antenna transceivers, which combine large-dimensional analog pre/postprocessing with lower-dimensional digital processing, are the most promising approach for reducing the hardware cost and training overhead in massive MIMO systems. The article provides a comprehensive survey of the various incarnations of such structures that have been proposed in the literature.

ABSTRACT

Hybrid multiple-antenna transceivers, which combine large-dimensional analog pre/postprocessing with lower-dimensional digital processing, are the most promising approach for reducing the hardware cost and training overhead in massive MIMO systems. This article provides a comprehensive survey of the various incarnations of such structures that have been proposed in the literature. We provide a taxonomy in terms of the required channel state information, that is, whether the processing adapts to the instantaneous or average (second-order) channel state information; while the former provides somewhat better signal-to-noise and interference ratio, the latter has much lower overhead for CSI acquisition. We furthermore distinguish hardware structures of different complexities. Finally, we point out the special design aspects for operation at millimeter-wave frequencies.

INTRODUCTION

Multiple-input multiple-output (MIMO) technology, that is, the use of multiple antennas at transmitter (TX) and receiver (RX), has been recognized since the seminal works of Winters, Foschini and Gans, and Telatar, as an essential approach to high spectral efficiency (SE). In its form of multi-user MIMO (MU-MIMO), it improves SE in two forms:

- A base station (BS) can communicate simultaneously with multiple user equipments (UEs) on the same time-frequency resources.
- Multiple data streams can be sent between the BS and each UE.

The total number of data streams (summed over all UEs in a cell) is upper limited by the smaller of the number of BS antenna elements, and the sum of the number of all UE antenna elements.

While MU-MIMO has been studied for more than a decade, the seminal work of Marzetta introduced the exciting new concept of "massive MIMO," where the number of antenna elements at the BS reaches dozens or hundreds. Not only does this allow increasing the number of data streams in the cell to very large values, it also simplifies signal processing, creates "channel hardening" such that small-scale fading is essentially eliminated, and reduces the required transmission energy due to the large beamforming gain; see, for example, [1] for a review. Massive MIMO is *beneficial* at centimeter-wave (cmWave) frequencies, but is *essential* in the millimeter-wave (mmWave) bands,[1] since the high free-space path loss at those frequencies necessitates large array gains to obtain sufficient signal-to-noise ratio (SNR), even at moderate distances of about 100 m.

However, the large number of antenna elements in massive MIMO also poses major challenges:

- A large number of radio frequency (RF) chains (one for each antenna element) increases cost and energy consumption.
- Determining the channel state information (CSI) between each transmit and receive antenna uses a considerable amount of spectral resources.

A promising solution to these problems lies in the concept of *hybrid* transceivers, which use a combination of analog beamformers in the RF domain, together with digital beamforming in the baseband, connected to the RF with a smaller number of up/downconversion chains. Hybrid beamforming was first introduced and analyzed in the mid-2000s by one of the authors and collaborators in [2, 3]. It is motivated by the fact that the number of up-downconversion chains is only lower-limited by the number of data streams that are to be transmitted, while the beamforming gain and diversity order is given by the number of antenna elements if suitable RF beamforming is done. While formulated originally for MIMO with arbitrary number of antenna elements (i.e., covering both massive MIMO and small arrays), the approach is of interest in particular to massive MIMO. Interest in hybrid transceivers has therefore been revived over the past three years (especially following the papers of Heath and co-workers, e.g., [4]), where various structures have been proposed in different papers. Thus, the time seems ripe for a review of the state of the art, and a taxonomy of the various transceiver architectures (often simplified to provide computational or chip-architectural advantages) and algorithms. The current article aims to provide this overview, and point out topics that are still open for future research.

This survey covers hybrid beamforming structures using instantaneous or average CSI in the following two sections. A special structure incor-

A version of this article with additional references can be found at arxiv.org/abs/1609.05078.

[1] In a slight abuse of notation, we denote 1–10 GHz as "centimeter waves," and 10–100 GHz as "millimeter waves."

Digital Object Identifier: 10.1109/MCOM.2017.1600400

Andreas F. Molisch, Vishnu V. Ratnam, Shengqian Han, and Zheda Li are with the University of Southern California, Los Angeles; Shengqian Han is also with Beihang University; Sinh Le Hong Nguyen, Linsheng Li, and Katsuyuki Haneda are with the Aalto University School of Electrical Engineering; Linsheng Li is presently with Huawei Helsinki.

0163-6804/17/$25.00 © 2017 IEEE

porating switches between the analog and digital parts is then described. Following that, we clarify constraints at mmWave bands due to propagation conditions and hardware imperfections. A summary and conclusions round up the article.

Hybrid Beamforming Based on Instantaneous CSI

Figure 1 shows block diagrams of three hybrid beamforming structures at the BS, where we assume a downlink transmission from the BS (acting as TX) to the UE (RX). The classification is applicable to both cmWave and mmWave bands. At the TX, a baseband digital precoder $\mathbf{F}_{BB}$ processes N_S data streams to produce N_{RF}^{BS} outputs, which are upconverted to RF and mapped via an analog precoder $\mathbf{F}_{RF}$ to N_{BS} antenna elements for transmission. The structure at the RX is similar: an analog beamformer $\mathbf{W}_{RF}$ combines RF signals from N_{UE} antennas to create N_{RF}^{UE} outputs, which are downconverted to baseband and further combined using a matrix $\mathbf{W}_{BB}$, producing signal $\mathbf{y}$ for detection/decoding.[2] Hence, we use terms "beamformer" and "precoder/combiner" interchangeably hereinafter. For a full-complexity structure, each analog precoder output can be a linear combination of *all* RF signals (Fig. 1, *A*). Complexity reduction at the price of somewhat reduced performance can be achieved when each RF chain can be connected only to a subset of antenna elements, as in Fig. 1, *B*. Different from structures *A* and *B*, where baseband signals are jointly processed by a digital precoder, structure *C* employs the analog beamformer to create multiple "virtual sectors," which enables separated baseband processing, downlink training, and uplink feedback, and therefore reduces signaling overhead and computational complexity [5].

Even assuming full-instantaneous CSI at the TX, it is very difficult to find the analog and digital beamforming matrices that optimize, for example, the net data rates of the UEs [6]. The main difficulties include:

- Analog and digital beamformers at each link end, as well as combiners at the different link ends, are coupled, which makes the objective function of the resulting optimization non-convex.
- Typically, the analog precoder/combiner is realized as a phase-shifter network, which imposes additional constraints on the elements of $\mathbf{W}_{RF}$ and $\mathbf{F}_{RF}$.
- Moreover, with finite-resolution phase shifters, the optimal analog beamformer lies in a discrete finite set, which typically leads to NP-hard integer programming problems.

Two main methodologies are explored to alleviate these challenges and achieve feasible near-optimal solutions.

Approximating the Optimal Beamformer

For single-user MIMO (SU-MIMO), we start with optimum beamforming for the fully digital case with $N_{RF}^{BS} = N_{BS}$ and $N_{RF}^{UE} = N_{UE}$, where the solution is known (dominant left/right singular vectors of a channel matrix $\mathbf{H}$ from singular value decomposition [SVD]). Then one approach (e.g., [2]) is based on eigen decomposition, while another (e.g., [6]) finds an (approximate) optimum hybrid beamformer by minimizing the Euclidean distance to this fully digital one. The objective function of the approximation problem is still non-convex, but much less complex than the original one. For sparse channels (as occur in mmWave bands), minimizing this distance provides a quasi-optimal solution. In non-sparse channels, such as usually occur at cmWave bands, an alternating optimization of analog and digital beamformers can be used. A closed-form solution for each of the alternating optimization steps can be developed for the reduced-complexity structure, while for the full-complexity structure, the non-convex problem can be expanded into a series of convex sub-problems by restricting the phase increment of the analog beamformer within a small vicinity of its preceding iteration.

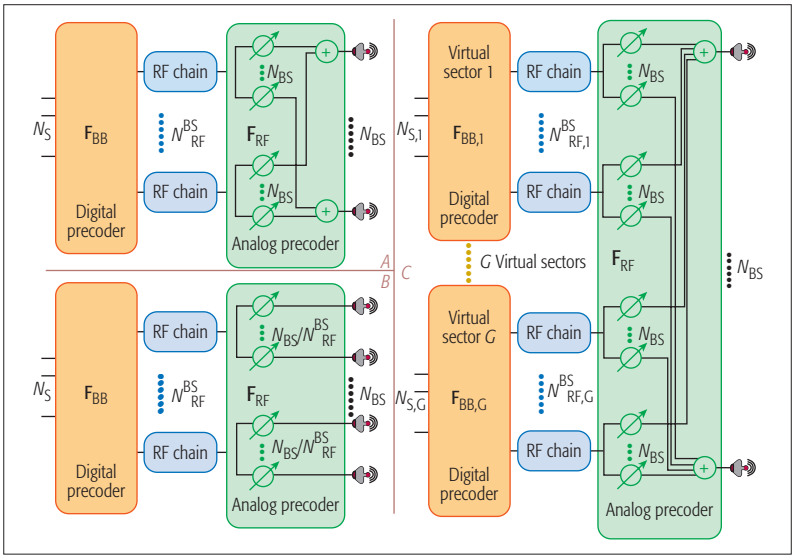

Figure 1. Block diagrams of hybrid beamforming structures at the BS for a downlink transmission, where structures *A*, *B*, and *C* denote the full-complexity, reduced-complexity, and virtual sectorization structures, respectively.

Figure 2 compares the performance of the three structures for downlink transmission of single-cell MU massive MIMO. The full-complexity structure of Fig. 1, *A*, performs the same as the fully digital structure when the number of RF chains is no smaller than the number of users (or streams). Performance loss of structure *B* is rather large for the considered MU case, although it is much smaller for SU-MIMO (not shown here). For structure C, the employed algorithm (JSDM, discussed later in this article) divides the users into four or eight groups, which might lead to a performance floor due to inter-group interference; note that the significantly reduced training overhead of JSDM is not shown here; this is discussed later.

Decoupling the Design of the Analog and Digital Beamformers

One of the main challenges in hybrid beamformer design is the coupling among analog and digital beamformers, and between the beamformers at TX and RX. This motivates decoupling the beamformer designs for reducing the problem complexity. By assuming some transceiver algorithms, optimization of beamforming matrices can be solved sequentially. For example, in order to maximize the net rate for SU-MIMO, one can eliminate the impact of the combiner on the pre-

[2] Obviously, a UE with a single antenna element is a special case.

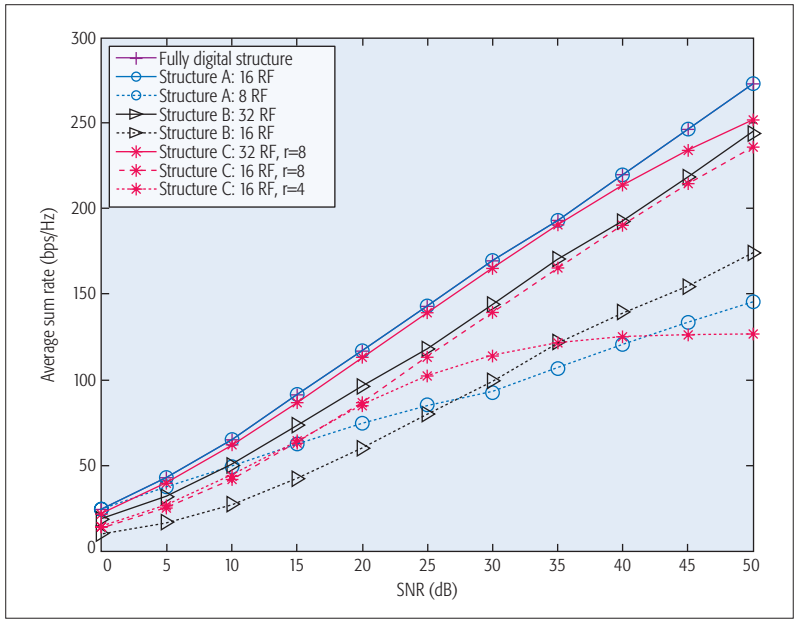

Figure 2. Performance comparison of the three hybrid structures with MU-MIMO; $N_{BS} = 64$, $N_{UE} = 1$, 4 groups of users located in a sector with mean directions [−45°, −15°, 15°, 45°], and each group has 4 users. AoDs of MPCs concentrate around the mean directions of each group with 10° AoD spread. {This analysis assumes ideal hardware and typical channel conditions for cmWaves.)

coder by assuming a fully digital minimum mean square error (MMSE) receiver. Further decoupling of the analog and digital precoder is possible by assuming that the digital precoder is unitary. Subsequently, $\mathbf{F}_{RF}$ is optimized column by column by imposing the phase-only constraint on each antenna. With the known analog precoder, a closed-form expression of the digital precoder can then be obtained.

Alternatively, some simple heuristic decoupling beamforming strategies have been explored. For example, the element-wise normalized conjugate beamformer can be used as the analog precoder, with which the asymptotic signal-to-interference-plus-noise ratio (SINR) of hybrid beamforming is only reduced by a factor of $\pi/4$ compared to fully digital beamforming when letting the number of antenna elements and streams, N_{BS} and N_S, go to infinity while keeping N_{BS}/N_S constant.

Extending to the situation where the UE is also equipped with a hybrid structure for MU-MIMO, one can first construct the RF combiner by selecting the strongest receive beams from the Fourier codebook to maximize the Frobenius norm of the combiner-projected channel. Then the same normalized eigenbeamformer is implemented as the analog precoder on the effective channel. In the baseband, the BS performs block diagonalization (BD) over the projected channel to suppress inter-user interference.

Wideband Hybrid Beamforming

The above discussion focused on narrowband (i.e., single-subcarrier) systems. In wideband orthogonal frequency-division multiplexing (OFDM) systems, however, analog beamformers cannot have different weights across subcarriers; for strongly frequency-selective channels, such beamformers extending over the whole available band adapt to the average channel state.

Frequency-domain scheduling was believed unnecessary for fully digital massive MIMO systems because the sufficiently large number of antennas can harden the channels and provide sufficient spatial degrees of freedom for multiplexing UEs [1]. However, under practical constraints on array size (e.g., according to 3GPP LTE Release 13), frequency-domain scheduling is still necessary for hybrid transceivers [7]. With frequency-domain scheduling, UEs are served on different subcarriers, making the existing narrowband hybrid precoders no longer applicable. Existing works have studied the joint optimization of wideband analog precoder and narrowband digital precoders, aimed at minimizing the BS transmit power or maximizing the sum rate of UEs.

Another important issue in the existing design of hybrid beamforming is control signaling coverage. While narrow analog beams are preferred for user-specific data transmission, wide beams are preferred for broadcasting control signals to all UEs. This problem may be solved, for example, by splitting signaling and data planes so that they are transmitted at different carrier frequencies.

Impact of Phase-Only Constraint and the Number of RF Chains

Hybrid beamforming does not necessarily have inferior performance to fully digital beamforming. Analog beamforming can be implemented by means of phase shifters together with variable gain amplifiers. In this case, analog beamforming can provide the same functionality as digital beamforming, and combine desired multipath components (MPCs) (and suppress interfering MPCs) to the same degree as linear digital processing. Thus, in a narrowband massive MIMO system, with full-instantaneous CSI at the TX, this hybrid beamforming can achieve the same performance as fully digital beamforming if $N_S \le N_{RF}$ [2]. A similar result can be obtained for a wideband system, where the number of RF chains of the hybrid structure should be not smaller than $\min(N_{BS}, N_{S,wb})$ with $N_{S,wb}$ denoting the total number of data streaming over all subcarriers [7].

Since two phase-only entries for the analog precoder are equivalent to a single unconstrained (amplitude and phase) entry, fully digital performance can be achieved with phase-only hybrid structures if $N_{RF}^{BS} \ge 2N_S$ in narrowband systems [2].

Hybrid Beamforming Based on Averaged CSI

Average CSI based Hybrid Beamforming

A major challenge for the beamformers discussed previously is the overhead for acquiring CSI at the BS. Information-theoretic results taking training overhead into account show that for time-division duplexing (TDD) systems, the spatial multiplexing gain (SMG) of massive MIMO downlinks with fully digital structure equals $M(1 - M/T)$, where $M = \min(N_{BS}, K, T/2)$, $K = N_S$ is the number of single-antenna users, and T is the number of channel uses in a coherence time-frequency block [5]. In frequency-division duplexing (FDD) systems, the overhead is even larger, since both downlink training and uplink feedback for each antenna are required. In addition to the coherence time, the

frame structure of systems may provide additional constraints for the pilot repetition frequency and thus the training overhead.

It is evident that for any massive MIMO systems relying on full CSI between all antenna elements of the BS and UEs, the maximal achievable SMG is limited by the size of the coherence block of the channel because N_{BS} and K are generally large. This necessitates the design of transmission strategies with reduced-dimensional CSI to relieve the signaling overhead. Specifically, a number of papers have considered analog beamforming based on slowly varying second order statistics of the CSI at the BS (a two-stage beamformer, with the first analog stage based on the average CSI only, followed by a digital one adapted to instantaneous CSI). The beamforming significantly reduces the dimension of the effective instantaneous CSI for digital beamforming within each coherent fading block by taking advantage of a small angular spread at the BS. Such structures work robustly even with analog beamformers, which cannot usually adapt to varying channels as quickly as digital beamformers.

Hybrid beamformers using average CSI for the analog part were first suggested in [3], which also provided closed-form approximations for the optimum beamformer in SU-MIMO systems. For the MU case, [5] proposed a scheme called "joint spatial division multiplexing" (JSDM), which considered a hybrid-beamforming BS and single-antenna UEs; to further alleviate the downlink training/uplink feedback burden, UEs with similar transmit channel covariance are grouped together, and inter-group interference is suppressed by an analog precoder based on the BD method. Specifically, using the Karhunen-Loeve representation, the N_{BS}-by-1 channel vector can be modeled as $\mathbf{h} = \mathbf{U}\Lambda^{1/2}\mathbf{w}$, where $\mathbf{w} \in \mathbb{C}^{r\times 1} \sim \mathcal{CN}(0, \mathbf{I}_r)$, Λ is an r-by-r diagonal matrix, which aligns eigenvalues of channel covariance $\mathbf{R}$ on its diagonal, $\mathbf{U} \in \mathbb{C}^{M\times r}$ indicates the eigenmatrix of $\mathbf{R}$, and r denotes the rank of the channel covariance. Dividing UEs into G groups and assuming that UEs in the same group g exhibit the same channel covariance $\mathbf{R}_g$ with rank r_g, the JSDM analog precoder is

$$\mathbf{F}_{RF} = [\mathbf{F}_{RF,1}, \ldots, \mathbf{F}_{RF,G}] \text{ with } \mathbf{F}_{RF,g} = \mathbf{E}_g\mathbf{G}_g.$$

By selecting $r_g^* \le r_g$ dominant eigenmodes of $\mathbf{R}_g$, denoted by $\mathbf{U}_g^*$, JSDM builds the eigenmatrix of the dominant interference to the gth group: $\Xi_g = [\mathbf{U}_1^*, \ldots, \mathbf{U}_G^*]$. Then $\mathbf{E}_g$ consists of the null space of Ξ_g, and $\mathbf{G}_g$ consists of dominant eigenvectors of $\mathbf{E}_g^\dagger\mathbf{R}_g\mathbf{E}_g$. This creates multiple "virtual sectors" in which downlink training can be conducted in parallel, and each UE only needs to feed back the intra-group channels, leading to the reduction of both training and feedback overhead by a factor equal to the number of virtual sectors.

In practice, however, to maintain the orthogonality between virtual sectors, JSDM often conservatively groups UEs into only a few groups, because UEs' transmit channel covariances tend to be partially overlapped with each other. This limits the reduction of training and feedback overhead. Once grouping UEs into more virtual sectors violates the orthogonality condition, JSDM is not able to combat the inter-group interference. Eliminating overlapped beams of UEs in different groups is a heuristic approach to solve this problem. In [8], JSDM is generalized to support non-orthogonal virtual sectorization, and a modified MMSE algorithm is proposed to optimize the multi-group digital precoders to maximize the lower bound of the average sum rate.

Two UE grouping methods have been proposed as extensions to JSDM: K-means clustering and fixed quantization. In the large antenna limit, the number of downlink streams served by JSDM can be optimized given the angle of departure (AoD) of MPCs and their spread for each UE group. To reduce the complexity of JSDM, in particular due to SVD, an online iterative algorithm can be used to track the analog precoder under time-varying channels. When considering single-antenna UEs, a Fourier codebook-based analog precoder, and a zero-forcing (ZF) digital precoder, the performance of JSDM can be further improved by jointly optimizing the analog precoder and allocation of RF chains to groups based on second order channel statistics. This principle can be extended to multicell systems, where an outage constraint on the UEs' SINR can be considered.

Decoupling of Analog and Digital Beamformers

Different from the previous section, where both analog and digital beamformers are based on instantaneous CSI, now analog and digital beamformers are based on the average CSI and the instantaneous effective CSI, respectively. Thus, to find the optimal beamformers, one needs to first design the digital beamformer for each snapshot of the channel and then derive the analog beamformer based on their long-term time-average, making their mathematical treatment difficult. Decoupled designs of the analog and digital beamformers therefore make the optimization problem simpler and practically attractive. For SU-MIMO where a UE is equipped with a single RF chain and multiple antennas, the optimal analog combiner is intuitively the strongest eigenmode of the UE-side channel covariance. However, when there are more RF chains at the UE, the strongest eigenmodes are not always the optimal combiners since they may be associated with a single transmit eigenmode of the BS-side channel covariance. For MU-MIMO with multiple RF chains at both link ends, the digital beamformer design needs to consider the UE-level spatial multiplexing and inter-user interference suppression, which will affect the analog beamformer design. In [8], the optimality (in the sense of maximizing the so-called intra-group signal-to-inter-group interference-plus-noise ratio) of decoupling analog and digital beamformers is shown under the Kronecker channel model.

Full-Dimensional MIMO in 3GPP

While the Third Generation Partnership Project (3GPP) standard does not prescribe particular transceiver architectures, hybrid digital-and-analog structures have motivated the design of CSI acquisition protocols in Release 13 of LTE-Advanced Pro in 3GPP, especially the non-precoded and beamformed pilots for full-dimensional (FD) MIMO. The non-precoded beamformer is related to the reduced-complexity structure B in

While the 3GPP standard does not prescribe particular transceiver architectures, hybrid digital-and-analog structures have motivated the design of CSI acquisition protocols in Release 13 of LTE-Advanced Pro in 3GPP, especially the non-precoded and beamformed pilots for FD MIMO.

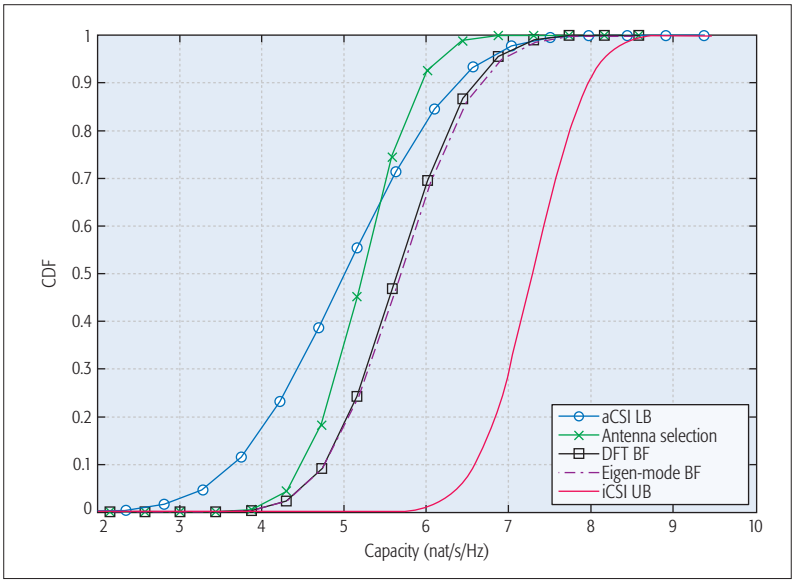

Figure 3. Performance of different analog precoders in a hybrid TX with selection. We consider an SU-MIMO system at cmWaves with ideal hardware conditions, where the RX has full complexity with $N_{UE} = N_{RF}^{UE} = 2$ and the TX has a switched hybrid beamforming structure with $N_{BS} = L = 10$, $N_{RF}^{BS} = 2$. The channels are Rayleigh distributed in amplitude, doubly spatially correlated (both at TX and RX), and follow the Kronecker model of spatial correlation $[R_{BS}]_{ij} = [R_{UE}]_{ij} = 0.5^{|i-j|}$; "aCSI LB" and "iCSI UB" refer to optimal unconstrained precoding with average CSI [3] and with instantaneous CSI, respectively. The RX SNR is 10 dB.

Fig. 1, where a (possibly static) analog precoder is applied to a subset of an antenna array to reduce the training overhead. The beamformed approach may assume the full-complexity structure *A* in Fig. 1, where analog beamformers are used for downlink training signals. The BS transmits multiple analog precoded pilots in different time or frequency resources. Then user feedback indicates the preferred analog beam; given this, the user can further measure and feed back the instantaneous effective channel in a legacy LTE manner. These approaches can, under some circumstances, reduce the overhead in average CSI acquisition, and generally perform well for SU-MIMO but may suffer large performance degradation for MU-MIMO unless the average CSI of all users is fed back. Recently proposed hybrid CSI acquisition schemes in 3GPP combine the above two approaches. First, the BS sends non-precoded pilots to estimate the average CSI at users. Then, based on the analog (non-codebook-based) or digital (codebook-based) feedback of the average CSI from users, the BS determines the analog beamformer and next sends beamformed pilots. These hybrid schemes essentially enable the form of beamforming discussed above in this section, namely, adaptation of the analog beamformer based on long-term statistics, which is then followed by the digital beamformer based on instantaneous effective CSI. Increasing the array size further motivates studies to reduce the training and feedback overhead through, for example, aperiodic training schemes. The JSDM-based structure C in Fig. 1 that separates a cell into multiple "virtual sectors" is one approach to reduce the overhead significantly by simultaneous downlink training and uplink feedback across virtual sectors.

Hybrid Beamforming with Selection

A special class of hybrid systems involves a selection stage that precedes (at the TX) or succeeds (at the RX) the analog processing, called hybrid beamforming with selection hereinafter. The up-converted data streams at the TX pass through the analog precoder $\mathbf{F}_{RF}$, as discussed. However, unlike conventional hybrid beamforming, the number of input ports of the analog block is $L \geq N_{RF}^{BS}$ (and typically, $L = N_{BS}$). A selection matrix $\mathbf{S}$, realized by a network of RF switches, feeds the data streams to the best N_{RF}^{BS} out of the L ports for transmission. The premise for such a design is that, unlike switches, analog components like phase shifters and amplifiers might not be able to adapt to the quick variation of instantaneous channels over time. Therefore, $\mathbf{F}_{RF}$ is either fixed or designed based on average channel statistics as described earlier, and $\mathbf{S}$ picks the best ports for each channel realization, thus making the effective analog processing more channel adaptive. The switching networks are also advantageous over full-complexity analog beamforming in terms of their cost and energy efficiency (EE). Although we focus on the TX for brevity, a switched analog combiner may also be implemented at the RX.

Design of Analog Precoding/Combining Block

The simplest "hybrid beamforming with selection" performs the antenna selection and omits the analog precoding. However, significant beamforming gains can be achieved by introducing analog precoding before the selection, to take advantage of the spatial MPCs. Such an architecture performs signal processing in the beam-space. While $\mathbf{F}_{RF}$ may be designed by discrete Fourier transform (DFT), its performance can be improved by eigenmode beamforming based on the TX correlation matrix [3]. To reduce the CSI feedback overhead for FDD systems, $\mathbf{F}_{RF}$ in the conventional hybrid beamforming can be chosen from a set of a predetermined codebook of matrices. By regarding the codebook entries as realizations from switch positions, this design can be interpreted as hybrid beamforming with selection. The codebook design is discussed, for example, in [9]. The performance of some of these analog precoders is compared in Fig. 3.

Design of Selection Matrix

Since complexity of searching for the best ports in the analog block is exponentially increasing with N_{RF}, many algorithms have been proposed to reduce it. Several greedy algorithms have been proposed to leverage diversity and spatial-multiplexing gains in an MU scenario. Restricted selection architectures allow each RF chain to choose from a subset of the analog ports, thereby reducing both search and hardware complexities. Iterative algorithms with varying search complexities from a linear to sub-exponential order have been proposed. An alternative technique that does not use the instantaneous CSI is called eigen-diversity beamforming [10]. It draws the selection matrix for each channel realization from an optimized probability distribution, thereby leveraging the temporal diversity.

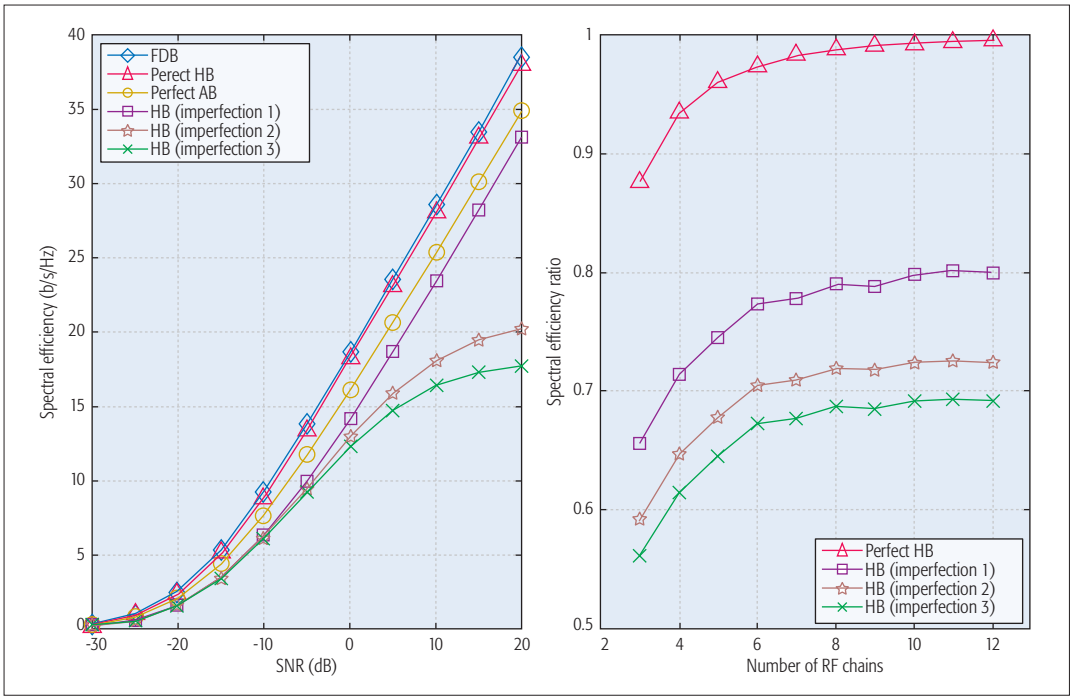

Figure 4. SE comparison of fully digital beamforming (FDB), Hybrid beamforming with perfect RF hardware (Perfect HB), analog-only beamsteering with perfect RF hardware (Perfect AB), and HB with three different cases of RF hardware imperfection: case 1 considers quantization error caused by 6-bit phase shifters; cases 2 and 3 additionally consider residual transceiver impairments at the BS and at both BS and UE, respectively. The spatially sparse precoding [6] is used in the HB. We assume that $N_{BS} = 64$, $N_{UE} = 16$, $N_S = 3$, the radio channel has 3 multipath clusters, and each has 6 rays, as representative of mmWave channels. The residual transceiver impairments at TX and RX are characterized by error-vector magnitude of –20 dB. In the left subfigure, $N_{RF}^{BS} = N_{RF}^{UE} = 6$. In the right subfigure, the SE of the HB with different RF hardware assumptions normalized to the FDB is characterized at SNR = 0 dB and $N_{RF}^{BS} = N_{RF}^{UE}$.

The presence of RF transceiver imperfections degrades SE in various ways. For example, it is harder to accurately generate desired transmit signals when higher beamformer gain is aimed for; non-linear distortion at the RX depends on the instantaneous channel gain and hence the SNR.

Hybrid Beamforming at mmWave

Hybrid beamforming architectures and algorithms in the cmWave band described in the previous sections can in principle be used at mmWave frequencies. In practice, however, propagation channel and RF hardware aspects are significantly different in those bands, and hence novel hybrid beamforming techniques taking into account the practicalities are needed. At mmWave frequencies, the multipath channel experiences higher propagation loss, which needs to be compensated by gain from antenna arrays at the TX, RX, or both. While such arrays have reasonable physical size thanks to short wavelengths, fully digital beamforming solutions become infeasible, and hybrid beamforming becomes harder due to power- and cost-related RF hardware constraints. Moreover, mmWave channels may be sparser, such that fewer spatial degrees of freedom are available. The sparsity can be exploited for optimizing channel estimation and beam training.

Hybrid Beamforming Methods Exploiting Channels' Sparsity

Exploiting the channels' sparsity, the simplest form of hybrid beamforming in SU-MIMO systems focuses array gains to a limited number of multipaths in the RF domain, while multiplexing data streams and allocating powers in baseband. This hybrid architecture is asymptotically optimum in the limit of large antenna arrays [11].

For systems with practical sizes of arrays, which have, for example, 64 to 256 elements for the BS and under 20 elements for the UEs, hybrid beamforming structures are highly desirable. In addition, reduction of the hardware and computational complexity is of great interest. For those purposes, a number of hybrid beamforming methods have been proposed for mmWave SU-MIMO channels that can be categorized into the use of codebooks, spatially sparse precoding, antenna selection, and beam selection.

Use of Codebooks: While having the same principle as the schemes described earlier, the codebook-based beamforming does not directly estimate the large CSI matrix at the RX, but instead performs downlink training using pre-defined beams and then only feeds back the selected beam IDs to the transmitter. To further reduce the complexity of beam search and feedback overhead for large antenna systems, a codebook for full-complexity hybrid architecture can be designed to exploit the sparsity of mmWave channels. Each codeword is constructed based on the Orthogonal Matching Pursuit (OMP) algorithm to minimize the MSE with the pre-defined ideal beam pattern.

Spatially Sparse Precoding: This method finds the approximation of the unconstrained (i.e., fully digital) beamformer as described earlier; at mmWave bands with electrically large arrays and a small number of dominant multipaths, the approximation can be made sufficiently close to the optimal precoder by using a finite number of antenna elements in the array [6]. The multipath sparsity restricts the feasible analog precoders $\mathbf{F}_{RF}$ to a set of array response vectors, and the baseband precoder optimization can be translated into a matrix

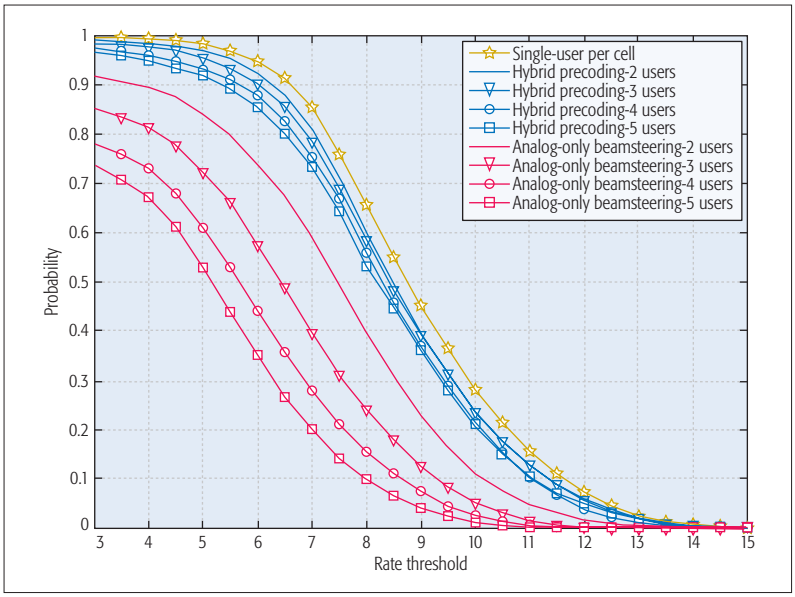

Figure 5. Comparison of achievable rates for hybrid precoding and analog-only beamsteering, from [14]. A single-path model is assumed between the BSs and UEs, and each link is assigned a line-of-sight or non-line-of-sight condition based on a blockage model, that is, the second reference in [14]. Each UE is associated to the BS with the least path loss and the BS randomly selects n = 2, ... ,5 associated UEs to be simultaneously served.

reconstruction with the cardinality constraint on the number of RF chains. The near-optimal solution of $\mathbf{F}_{BB}$ can then be found using sparse approximation techniques (e.g., OMP). The SE comparison of this method with unconstrained fully-digital beamforming and analog-only beamsteering with perfect transmit CSI is shown in Fig. 4.

While the general structure is the same as the one for cmWaves described earlier, in sparse mmWave channels, fast and greedy antenna subset selection [12] performs as robustly as exhaustive antenna search. Hybrid antenna selection can outperform a sparse hybrid combiner with coarsely quantized phase shifters in terms of power consumption when both have the same SE performance. There is still a large gap in SE between the hybrid combiner with switches and a fully digital one with ideal phase shifters.

Beam Selection: Another hybrid beamforming structure is based on continuous aperture phased (CAP)-MIMO transceivers. It uses a lens antenna instead of the phase shifters or switches for RF beamforming, and realizes the beamspace MIMO (B-MIMO) [13] similarly to the spatial DFT with selection discussed previously. An electrically large lens antenna is excited by a feed antenna array beneath the lens. The feed array is called a beam selector since the lens antenna produces high-gain beams that point at different angles depending on the feed antenna. The CAP-MIMO can efficiently utilize the low-dimensional high-gain beamspace of the sparse multipath channel by selecting a couple of feed antennas using a limited number of RF chains, like the spatially sparse precoding.

Hybrid Beamforming in mmWave MU Scenarios

Hybrid beamforming is also a promising solution for mmWave MU-MIMO systems. The hybrid structure at the BS can transmit multiplexed data streams to multiple UEs; each UE can be equipped with an antenna or an antenna array with fully analog beamforming. Figure 5 shows achievable rates of hybrid beamforming in MU multi-cell scenarios [14]. Consider UEs with a single RF chain and many antennas, which distributively select the strongest beam pair to construct analog beamformers. Thanks to the ZF digital precoding at the BS mitigating the inter-user interference, the hybrid structure significantly outperforms the analog beamsteering approach.

The hybrid beamforming based on beam selection and B-MIMO concept can also be extended to MU-MIMO systems with linear baseband precoders. While its effectiveness (compared to its full complexity counterparts) has been demonstrated in mmWave channels, many system and implementation aspects of hybrid beamforming in mmWave MU-MIMO systems, including multi-user scheduling, and 2D and 3D lens array design, are still open for further research.

Impact of Transceiver Imperfections

The presence of RF transceiver imperfections degrades SE in various ways. For example, it is harder to accurately generate desired transmit signals when higher beamformer gain is aimed for; nonlinear distortion at the RX depends on the instantaneous channel gain and hence the SNR. Due to the transceiver imperfections being more pronounced at mmWaves, the SE and SNR of hybrid precoder/combiners no longer scale well with the number of RF chains. Figure 4 compares the SE of spatially sparse hybrid precoding, including RF imperfections, to that from fully digital beamforming based on SVD. The aggregate impact of the transceiver imperfections is modeled as a Gaussian process. The coarsely quantized phase shifters and the transceivers' imperfections significantly degrade the SE. Knowledge of transceiver imperfections at mmWaves is essential for analyzing the scalability of the SE in the large MIMO regime.

Spectral-Energy Efficiency Trade-off

Finally, we discuss a relationship between EE and SE of hybrid beamforming structures at mmWaves based on [15]. The hybrid structure B in Fig. 1 was studied, where the BS uses a sub-array with N_{BS}/N_{RF}^{BS} antennas to serve each user individually. Figure 6 shows the EE-SE trade-off, indicating an optimal number of RF chains achieving the maximal EE for any given SE.

Conclusion

Hybrid beamforming techniques were invented more than 10 years ago, but have seen a dramatic uptick in interest in the past 3 years due to their importance in making massive MIMO systems cost- and energy-efficient. They use a combination of analog and digital beamforming to exploit the fine spatial resolution stemming from a large number of antenna elements, but keep the number of (expensive and energy-hungry) RF up/downconversion chains within reasonable limits. This article categorizes the hybrid beamforming according to:

- Amount of required CSI (instantaneous vs. average) for the analog beamformer part
- Complexity (full complexity, reduced complexity, and switched)

- Carrier frequency range (cmWave vs. mmWave, since both channel characteristics and RF impairments are different for those frequency ranges)

It is clear that there is no single structure/algorithm that provides the "best" trade-off between complexity and performance in all those categories, but rather that there is a need to adapt them to application and channel characteristics in every design.

Acknowledgment

The financial support of the Academy of Finland and the National Science Foundation through the WiFiUS project "Device-to-Device Communica tions at Millimeter-Wave Frequencies" is gratefully acknowledged.

References

[1] E. Larsson *et al.*, "Massive MIMO for Next Generation Wireless Systems," *IEEE Commun. Mag.*, vol. 52, no. 2, Feb. 2014, pp. 186–95.
[2] X. Zhang, A. Molisch, and S.-Y. Kung, "Variable-Phase-Shift-Based RF-Baseband Codesign for MIMO Antenna Selection," *IEEE Trans. Signal Processing*, vol. 53, no. 11, Nov. 2005, pp. 4091–4103.
[3] P. Sudarshan *et al.*, "Channel Statistics-Based RF Pre-Processing with Antenna Selection," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, Dec. 2006, pp. 3501–11.
[4] A. Alkhateeb , "MIMO Precoding and Combining Solutions for Millimeter-Wave Systems," *IEEE Commun. Mag.*, vol. 52, no. 12, Dec. 2014, pp. 122–31.
[5] A. Adhikary *et al.*, "Joint Spatial Division and Multiplexing — The Large-Scale Array Regime," *IEEE Trans. Info. Theory*, vol. 59, no. 10, Oct. 2013, pp. 6441–63.
[6] O. El Ayach *et al.*, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, Mar. 2014, pp. 1499–1513.
[7] L. Kong, S. Han, and C. Yang, "Wideband Hybrid Precoder for Massive MIMO Systems," *Proc. 3rd Global Conf. Signal and Info. Processing*, Orlando, FL, Dec. 2015, pp. 305–09.
[8] Z. Li, S. Han, and A. F. Molisch, "Hybrid Beamforming Design for Millimeter-Wave Multi-User Massive MIMO Downlink," *IEEE ICC '16*, Kuala Lumpur, Malaysia, May 2016.
[9] S. Hur *et al.*, "Millimeter Wave Beamforming for Wireless Backhaul and Access in Small Cell Networks," *IEEE Trans. Commun.*, vol. 61, no. 10, Oct. 2013, pp. 4391–4403.
[10] J. Choi, "Diversity Eigenbeamforming for Coded Signals," *IEEE Trans. Commun.*, vol. 56, no. 6, June 2008, pp. 1013–21.
[11] O. El Ayach *et al.*, "The Capacity Optimality of Beam Steering in Large Millimeter Wave MIMO Systems," *Proc. 13th Wksp. Signal Processing Advances in Wireless Commun.*, Cesme, Turkey, June 2012, pp. 100–04.
[12] R. Mendez-Rial et al., "Channel Estimation and Hybrid Combining for mmWave: Phase Shifters or Switches," *Proc. Info. Theory Appl. Wksp.*, San Diego, CA, Feb. 2015, pp. 90–97.
[13] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for Millimeter-Wave Communications: System Architecture, Modeling, Analysis, and Measurements," *IEEE Trans. Antennas Propag.*, vol. 61, July 2013, pp. 3814–27.
[14] A. Alkhateeb, G. Leus, and R. Heath, "Limited Feedback Hybrid Precoding for Multi-User Millimeter Wave Systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, 2015, pp. 6481–94.
[15] S. Han, C.-L. I, Z. Xu, and C. Rowell, "Large-Scale Antenna Systems with Hybrid Analog and Digital Beamforming for Millimeter Wave 5G," *IEEE Commun. Mag.*, vol. 53, no. 1, Jan. 2015, pp. 186–94.

Biographies

Andreas F. Molisch [F'05] is the Solomon-Golomb — Andrew-and-Erna Viterbi Chair Professor at the University of Southern California (USC), Los Angeles. His research interests include wireless propagation, multi-antenna systems, ultrawideband communication and localization, and wireless video distribution. He has published 4 books, more than 500 journal and conference papers, and more than 80 patents. He is a Fellow of the National Academy of Inventors, AAAS, and IET, and a member of the Austrian Academy of Sciences.

Vishnu V. Ratnam received his B.Tech. degree from the Indian Institute of Technology Kharagpur, where he graduated as the Salutatorian for the class of 2012. He is currently pursuing a Ph.D. degree in electrical engineering at USC. His research interests include the design and analysis of low-complexity transceivers for large antenna and ultra-wideband systems, and resource allocation problems for multi-antenna networks. He is a recipient of the ICUWB 2016 Best Student Paper Award.

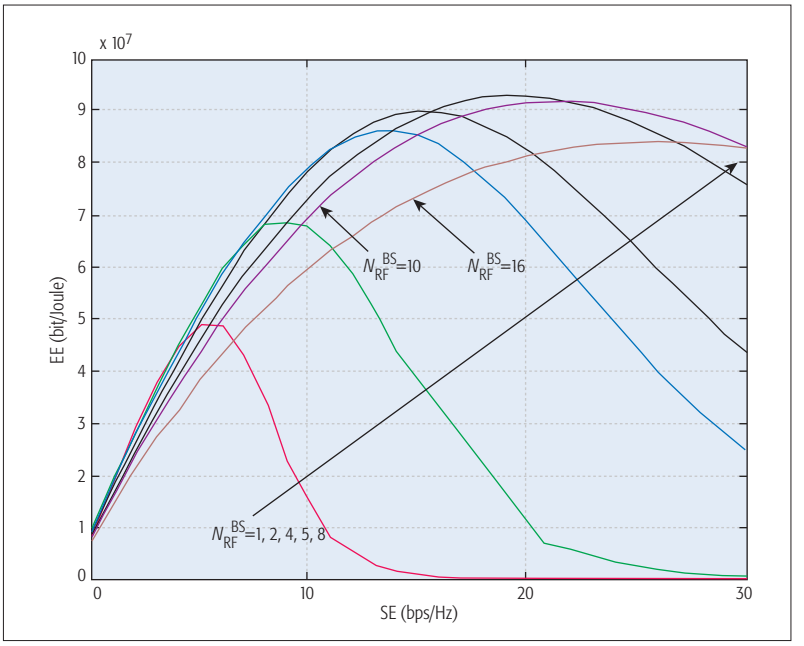

Figure 6. EE-SE relation of mm-wave massive MIMO system [15]. The hybrid transceiver follows Structure B of Fig. 1; N_{BS} = 800, system bandwidth is 200 MHz, noise power spectrum is 10^{-17} dBm/Hz, average channel gain is –100 dB, the efficiency of power amplifier is 0.375, the static power consumptions for each RF chain and each antenna are both 1 Watt, and the other fixed power consumption is 500 Watt.

Shengqian Han [S'05, M'12] received his B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2004 and 2010, respectively. He is currently a lecturer at Beihang University. From 2015 to 2016, he was a visiting scholar at USC. He is an Associate Editor for the *EURASIP Journal on Wireless Communications and Networking*. His research interests include full-duplex communications, multiple-input multiple-output systems, and energy-efficient transmission.

Zheda Li received his B.S. degree in communication engineering from Beijing University of Posts and Telecommunications, China, in 2010 and his M.S. degree (with honors) in electrical engineering from USC in 2012. He is currently pursuing a Ph.D. degree in the WiDeS group of USC. His research interests include massive multiple-input multiple-output systems and millimeter-wave communications.

Sinh Le Hong Nguyen [S'10, M'13] received his Ph.D. degree in electrical engineering from Concordia University, Canada, in 2013. From 2013 to 2014, he was a postdoctoral fellow at McGill University, Canada. He is now with Aalto University, Finland, as a postdoctoral researcher. He has participated in several industrial collaborative and EU-funded projects on 5G technologies. His current research interests include channel modeling and signal processing for mobile communications, compressive sensing, large-scale networks, and the Internet of Things.

Linsheng Li [S'09, M'14] received his Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 2014. From September 2014 to July 2016, he was a research engineer with the Department of Radio Science and Engineering at Aalto University. Since August 2016, he has been an antenna engineer with the Research Center of Huawei, Helsinki. His main research interests include sub-6 GHz and millimeter-wave antennas for 5G and future wireless communication.

Katsuyuki Haneda [S'03, M'07] is an associate professor at Aalto University. He is the recipient of the best paper awards at IEEE VTC-Spring 2013 and EuCAP2013 as a first author, and at a number of international conferences and journals as a co-author. He has been an Editor of *IEEE Transactions on Antennas and Propagation* and *IEEE Wireless Communications*. His research covers wireless and electromagnetic design and modeling for cellular, medical, and emergency applications.