1

LP100 — Optimization of 100 Gb/s Short Range Wireless Transceivers under Processing Energy Constraints

Gerd Ascheid, Gaojian Wang, Sebastian Birke

Institute for Communication Technologies and Embedded Systems, RWTH Aachen University

Norbert Wehn, Matthias Herrmann

Division of Microelectronic Systems Design, TU Kaiserslautern

Yanlu Wang, Oner Hanay, Erkan Bayram, Renato Negra

Chair of High Frequency Electronics, RWTH Aachen University

CONTENTS

DOI: 10.26127/BTUOpen-5906

1.1 Introduction

1.1.1 Motivation

There are three important boundaries in the design of a wireless transmission system for data rates beyond 100 Gb/s:

1. Power efficiency
2. Bandwidth efficiency
3. Implementation efficiency

Power efficiency
Transmit power is limited in wireless communication systems for various reasons. With the distant-dependent path loss and the receive and transmit antenna gains, we have a maximum received signal power P. From the well-known equation

$$P = E_S/T_S \Leftrightarrow E_S = T_S P \tag{1.1}$$

it is obvious that an increase in the symbol rate, i.e., a decrease of the symbol duration T_S, decreases the energy per symbol E_S for a given power P. The capacity of the link depends on the signal-to-noise ratio (SNR) given by

$$P/N = \frac{E_S/T_S}{N_0 W} = E_S/N_0 \tag{1.2}$$

with the noise power N, the minimal (two-sided) bandwidth $W = 1/T_S$, and the power spectral density N_0. The achievable data rate is upper-bounded by this capacity. To reach the target rate of 100 Gb/s for acceptable link ranges, the power gain in the antenna arrays must be high enough and power efficiency of the modulation and coding schemes is essential to allow information

bit rates close to capacity.

Bandwidth efficiency
Bandwidth is a limited resource. The available bandwidth depends on the considered carrier frequency range. Below 10 GHz available bandwidth for a link is in the order of 1 GHz. In the 60 GHz carrier frequency range, a few GHz of bandwidth may be used for a link. The largest bandwidths are available in and above the 120 GHz range. Two extreme cases can serve as indicator for the problem size: We consider transmission of 100 Gb/s. Assuming Nyquist rate transmission, this requires 50 GHz bandwidth for a basic 2 bit (1 bit per complex dimension) modulation. To reduce the signal to a bandwidth of 1 GHz, a modulation with 100 bits per symbol (50 bits per complex dimension) would have to be used. These calculations indicate that various methods beyond increased modulation orders, e.g., methods achieving an SNR increase and/or a multiplexing of transmitted streams in the spatial domain, have to be considered.

Implementation efficiency
The continuing increase of transistor density (due to the ongoing technology scaling) will not make the gate count the critical factor in the digital baseband processing. However, a critical limit both for analog and digital integrated circuits is the power consumption, in particular, to avoid overheating. Starting from a limit for the total power consumption of the transceiver, we can immediately derive the energy available to transmit and detect one information bit

$$E_{processing\ per\ bit} = T_{bit} P_{processing} \tag{1.3}$$

Here, T_{bit} is the inverse of the bit rate, and $P_{processing}$ is the total average power required for processing. Note that in 1.1 the transmit power is considered while in 1.3 energy and power for processing are considered (which may include the transmit power). Obviously, for a given maximum power consumption the available energy to process one bit decreases with the bit duration, i.e., for increasing bit rates. As will be demonstrated later, this is one of the most critical issues in the design of a beyond 100 Gb/s wireless transmission system.

With limited battery capacity of portable wireless devices, implementation efficiency is typically considered only after the communication system has been defined. A simple calculation shows, however, that for very high data rates on-chip power consumption becomes a limiting factor. Assuming a limit in the order of 1W for an integrated circuit without forced cooling and a data rate of 100 Gb/s yields 1 W/100 Gb/s = 10 pJ/bit. About 20 G 16-bit-MAC (Multiply-Accumulate) operations are possible in an energy envelope of 200 mJ for a dedicated intellectual property (IP) block implementation in state-of-the-art 65 nm CMOS technology. This yields about 200 mJ/20 GMAC = 10 pJ/MAC, i.e. one MAC-operation per information bit.

It is well known that higher bandwidth and power efficiency requires more elaborate signal processing and, thus, higher processing energy per information bit. This motivated our approach to consider implementation efficiency already in the communication system design. Achieving data rates beyond 100 Gb/s requires a design space exploration considering jointly wireless communication performance and implementation efficiency. Note that implementation efficiency also includes trade-offs between analog and digital signal processing since functions can often be implemented more energy efficient in analog technologies - although sometimes at the cost of a communication performance loss.

1.1.2 Relevant Boundary Conditions

For wireless communication systems there are various constraints. Key constraints are on frequency and bandwidth. Not all frequencies are available for wireless communication, frequency ranges are either blocked for other uses or not suitable due to excessive propagation loss. In the frequency range below 10 GHz bandwidths are strongly regulated and limited. Available bandwidths for a transmission are in the order of 1 GHz at most (e.g. UWB). Due to the progress in RF-IC-design higher frequency ranges become technically and commercially feasible. A first already feasible range is around 30 GHz to 60 GHz, a still more challenging range is from 100 GHz up to 300 GHz. In these frequency ranges much wider bandwidths are available ranging from several GHz per transmission in the medium GHz range (30-60 GHz) up to 10's of GHz in the high GHz range (above 100 GHz).
Therefore, we will consider three scenarios in the subsequent paragraphs:

- Scenario 1: Carrier frequencies below 10 GHz; available bandwidth per transmission in the order of 1 GHz
- Scenario 2: Towards sub-mm and optical frequencies above 120 GHz; very large bandwidths available.
- Scenario 3: Intermediate frequency range, mainly around 60 GHz; several GHz of bandwidth available.

In the following the relevant boundary conditions for these three scenarios will be discussed from three different perspectives, RF front-end, transmission scheme, and general implementation aspects.

Scenario 1
Traditional cellular and wireless local area network (WLAN) systems are operating in the range below 10 GHz. Cost-effective CMOS RF implementations are state-of-the-art. In combination with orthogonal frequency division duplex (OFDM) multi-carrier modulation, high spectral efficiencies despite some implementation challenges can be reached. A main approach besides high modulation orders is the use of multi-antenna systems. In rich scattering environ-

ments spatial multiplexing through multiple MIMO modes can be used. One critical part in MIMO systems with spatial multiplexing is the spatial equalization / detection to separate the transmitted streams. MIMO detection is a computationally complex operation and thus also costly from an energy point of view. One can further make use of the polarization domain, which provide significant efficiency gains in line-of-sight (LOS) environments, however, in non-LOS situations gains are limited.

While spatial multiplexing in favorable scattering conditions and a high modulation order allow higher bandwidth efficiency, the scarce bandwidth availability makes it very challenging to reach the desired throughput of 100 Gb/s and beyond. In addition, high modulation orders come at the cost of a low power efficiency. With the transmit power constraints imposed by regulators this leads to very short ranges for very high bandwidth efficiencies.

Scenarios 2
Recent progress in CMOS (Complementary Metal Oxide Silicon) integrated circuits (IC) technology has made it possible to consider CMOS as an alternative for realization of capable and economical systems that operate at 120 GHz and beyond. Along with the higher resolution, submillimeter-wave transceivers enjoy the advantage that they can easily be integrated jointly with antennas on a chip. Such arrays can be used for beam steering and multiple-input multiple-output (MIMO) transmission schemes.

The vast amounts of unlicensed bandwidth available around 120 GHz and beyond have made this region very attractive and result in relaxed requirements for the spectral efficiency. Due to the huge propagation losses these communication systems will most likely be employed for short-range LOS transmission in indoor scenarios.

Scenario 3
CMOS integrated circuits for this frequency range are becoming commercial state of the art. The amount of available bandwidth is less than above 100 GHz but still much larger than in the range below 10 GHz. Also, the propagation loss is higher than in the low GHz range but lower than the propagation loss above 100 GHz. The number of paths in the channel from reflections or scattering is highly reduced compared to the low GHz range. Yet, communication is not only limited to LOS.

1.1.3 Motivation for the choice of the scenario

A trade-off factor resulting from the choice of the carrier frequency and thus the available bandwidth is the modulation order. The selection of the modulation order and the bandwidth significantly affects the system architecture as well as the complexity on the RF side. Therefore, it also affects the energy efficiency. It is clear that regarding RF and analog/digital conversion a low modulation order is preferable. However, to achieve high data rates, the

modulation order needs to be high enough so the signal fits into the available bandwidth. Therefore, an analysis of the trade-off between bandwidth requirement and modulation order must be performed under consideration of the energy/power consumption.

In the 60 GHz range several GHz of unlicensed bandwidth are available worldwide, e.g., 7 GHz in the USA. Therefore, high data rates above 100 Gb/s can be achieved with moderate modulation orders. As discussed above, this is important for achieving a high energy efficiency. Also, a - though limited - number of strong scatter paths can be expected in addition to LOS. This offers two advantages: communication is also possible when the LOS is blocked and the scatter paths can be used for spatial multiplexing, which is one of the options for higher bandwidth efficiency without excessive reduction of energy efficiency. Above 100 GHz even more bandwidth is available but the propagation is dominated more by LOS than at 60 GHz. Also above 100 GHz RF technology currently is still in a less mature state than 60 GHz RF technology. Therefore, the focus of the subsequently studied system is on 60 GHz carrier frequency, with the perspective to go up to the 100 GHz and beyond range.

1.2 System Concept

1.2.1 Energy Considerations

As shown in Section 1.1.1 the overall energy budget available for transmitting an information bit is in the order of 10 pJ/bit under power density constraints on chip level. Improvement in energy efficiency due to technology progress achieves about one order of magnitude, i.e., the gain of energy efficiency in a 7 nm technology node compared to a 28 nm technology node, is 8.17. Thus, an extrapolation of state-of-the-art wireless techniques even down to 7 nm technology cannot produce the necessary gain in energy efficiency. To achieve high power efficiency, high bandwidth efficiency and ultra-high throughput under such energy constraints, the wireless system architecture must be designed for processing-energy efficient implementation. The consequences of such an processing-energy minimization approach will be discussed subsequently. As important basis for the system architecture considerations, the channel model for wireless transmission at 60 GHz and beyond have to be looked at.

1.2.2 Channel Model

Since many groups have already worked on sounding and characterization of the 60 GHz channel, the the derivation of suitable channel models can be based on published work.

First, the attenuation of reflecting surfaces is much higher at 60 GHz than

in the range below 5 GHz [1, 2]. Thus, multiple reflected rays are not relevant. Secondly, due to high atmospheric absorption at 60 GHz paths much longer than LOS are negligible. Rather the 60 GHz channel will comprise a possible LOS path and a small number of dominating non-LOS paths [3]. In the considered indoor environment most of the reflecting objects ("clusters") will either be close to the access point and to the UE (e.g. objects on the desk) or will be nearby larger objects (walls, boards, etc.). Measurements indicate that the number of reflection clusters to be considered is in the range from 4-8 [4]. The angular spread for each reflection cluster is small leading to a small delay spread within a cluster.

In conclusion, we assume 4-8 dominating clusters with small delay spread of reflections within the cluster but larger delay spread in the range of 3-30 ns between different clusters [4]. Due to the position of the reflecting objects angular separation of clusters may be large at transmitter and small at the receiver (objects near the transmitter) or vice versa (objects near the receiver) or on both sides (walls, boards, etc.).[5]

1.2.3 System Architecture

Based on the channel modeling studies, frequency selectivity will already be relevant for signal bandwidths in the MHz-range. The most common way to enable feasible equalization is the use of orthogonal frequency-division multiplexing (OFDM). In [6, 7, 8], for example, different criteria were proposed to design a RF beamformer for frequency selective channels by employing OFDM. However, OFDM is very costly in terms of power consumption. Due to its high peak-to-average power ratio (PAPR) OFDM requires high resolution AD- and DA-converters and high linearity ranges of the (power) amplifiers. This leads to excessive power consumption (e.g. 3 W for 1 GS/s, 10 bit see www.ti.com) and is, thus, not a suitable approach for getting the highest data rate under a processing power constraint. Another option to combat the frequency selective fading is single carrier modulation with frequency-domain equalization (SC-FDE). In [9] hybrid beamforming for frequency selective MIMO channels using SC-FDE is proposed. Compared with multicarrier OFDM systems, single-carrier systems can achieve a better power efficiency in AD-/DA- conversion and in the RF front-end as well as better synchronization robustness. However, for the targeted data rates above 100 Gb/s bandwidths are wide and, thus, frequency selectivity so severe that equalization becomes extremely complex. This issue can be solved as follows:

Since antennas are small at 60 GHz, multiantenna transceivers can be implemented efficiently. Thus, the channel can be made frequency-flat by RF domain beamforming separating the dominating paths, and subsequent per-beam time-delay compensation (TDC). For paths separable by the transmitter, pretransmission delay compensation is used. For paths separable by the receiver, beamforming at the receiver side is used. This architecture now can be used to trade-off data rate and processing energy. For example, more an-

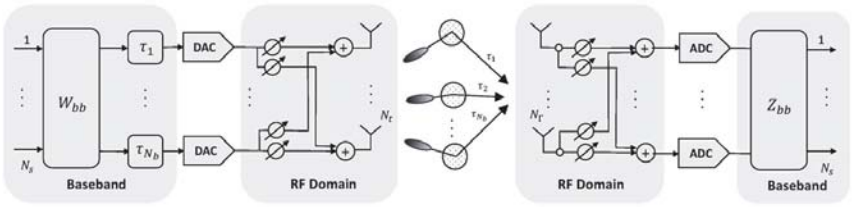

Figure 1.1
System Architecture. From [5] © 2017 IEEE

tennas per beam provide a better separation (less interference) and higher gain but at the cost of more processing power (more RF chains). Different paths may be combined in the receiver to get a larger gain or for spatial multiplexing.

A RF chain for each antenna would result in a very complex, expensive hardware and high power consumption. To minimize power consumption the number of AD-converters must be reduced and processing be shifted to the analog part. Therefore, hybrid beamforming in which precoding and combining is split into analog and digital domains is employed [10], resulting in the system architecture shown in Fig. 1.1. Compared to pure analog beamforming, the digital beamforming layer provides more freedom (i.e. a larger design space) for the pre- and postprocessing design. Hybrid beamforming, thus, enables a trade-off between high data rates and better performance for multistream transmission versus processing power.

1.2.4 Beamforming with Time Delay Compensation

Time delay compensation, if implemented precisely, yields no delay differences between different paths and achieves a flat channel over wide bandwidth. Based on indoor channel measurements, the geometric modeling approach [10] consists of multiple clusters, each with multiple rays. In order to compensate the delay spread both, intra- and intercluster, we introduce beamforming in RF domain and a buffer in baseband. Firstly, the local scatterers in each cluster only introduce very small delay variations and we can further minimize the delay spread by employing RF beamforming. By increasing the number of antennas, a narrower beam can be obtained, which reduces the delay spread. Thus, each cluster can be made frequency flat over a wider bandwidth. Secondly, we compensate the intercluster time delay difference by using a buffer in digital baseband. Obviously, the frequency selectivity of the channel depends on the time resolution of the buffer yielding a trade-off between processing power and performance (both increase with resolution).[11]

The influence of buffer resolution on the BER performance of the above mentioned system with different number of transmit antennas N_t has been analyzed [10]. In Fig. 1.2 the effect of different sample durations of the buffer

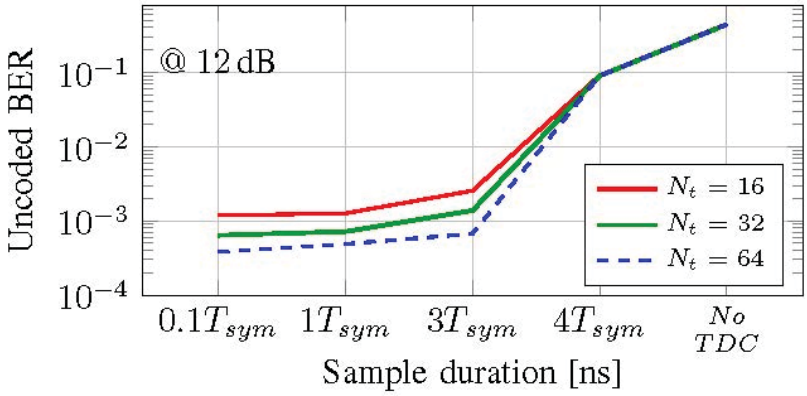

Figure 1.2
Uncoded BER performance with 3-tap equalizer. From [10] © 2015 IEEE

on the BER performance is illustrated. A finite tap-length MMSE equalizer at the receiver side is employed to fix the complexity at the user side for the purpose of a fair comparison. It can be seen in Fig. 1.2 that decreasing the resolution of the buffer has an adverse effect on the BER performance. The BER performance is almost constant for sample durations up to one symbol when using a 3-tap MMSE equalizer. Note that performance without TDC (no TDC) is also shown for reference.

To achieve a flat fading channel, best possible beam separation is necessary. We propose an average signal-to-interference ratio (SIR) constrained design to achieve the best possible beam separation. As is known, the SIR of a particular beam is related to the number of antennas used. In this design algorithm, we optimize the number of antennas to achieve the best possible directivity and at the same time satisfy the SIR constraint, which also minimizes the processing power. The approach will be discussed in subsection 1.4. At this point a tradeoff between performance and complexity arises: beamforming can be done in the analog/RF domain, in the digital domain or in both domains (hybrid beamforming). This tradeoff is exemplified in Fig. 1.3, where we compare a hybrid spatial processing strategy with a fully digital transceiver. With 4 RF chains, the information rate of the optimum scheme is very close to that of a fully digital transceiver in both 8x8 and 16x16 hybrid systems. Moreover, the low complexity transceiver strategy results in an acceptable performance loss.[12]

After obtaining a near-flat fading channel, the optimization of the baseband pre- and postprocessing is formulated based on the capacity maximization criterion under a total transmit power constraint. The joint design of the pre- and postprocessing matrices to maximize the data rate of the proposed system will be discussed in Section 1.4.1.

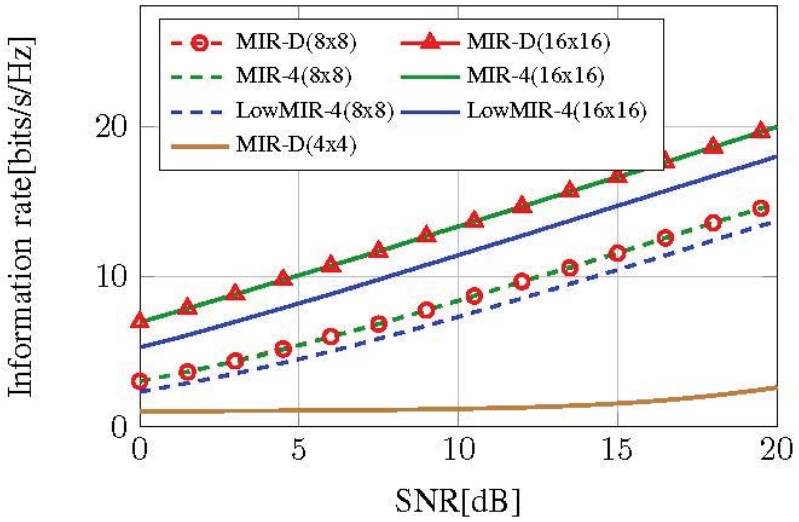

Figure 1.3
Information rate achieved by fully digital, hybrid, and low-complexity hybrid system under a maximum information rate criterion (MIR). From [12] © 2014 VDE

1.3 RF Front-end Architecture

Under stringent power and energy constrains, exploration and evaluation of different RF front-end architectures, which are suitable for high data-rate signal processing, are critical. For a well design exploration for the suitable RF front-end architectures, the factors such as frequency bands, modulation schemes, and power budget have to be taken into consideration.

In the sub 10 GHz bands, the modulation scheme requires a high bandwidth efficiency since the available signal bandwidth is limited. When the frequency bands increases to around 60 GHz and 120 GHz, they provide a wide signal bandwidth to achieve high data-rate, however, their output power is limited. Therefore, in this frequency bands modulation scheme with a high power efficiency should be utilized. Two domains, the bandwidth limited and the power limited regions will be explored for the possible modulation schemes. In the early stage of the design, it is valuable to model RF front-end architecture regarding their behavior and power consumption.

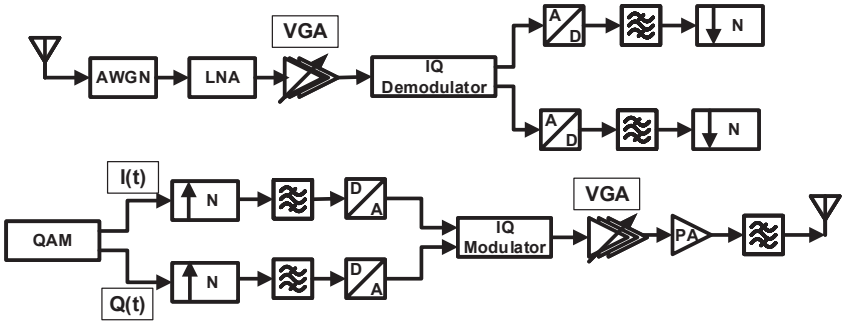

Figure 1.4
Block diagram of the RF front-end architecture.

1.3.1 Power modeling and energy budgeting

Based on previous publications and qualitative arguments, the estimation of power consumption for analog and digital parts are performed in this section. Building blocks are modeled regarding their qualitative behavior and power consumption. Fig. 1.4 presents the block diagram of the transceivers RF front-end architecture. In the communication system, the transceiver works in four modes: active, idle, sleep and transient mode. The total energy consumption is the sum of the energy in the four modes [13]. In this work, only the active mode is under the consideration because it consumes the most power. According to the investigated system, the system energy consumption for different modulation orders and symbol rates is demonstrated in Fig. 1.5. Meanwhile, DAC and PA are the two main components dominating power consumption. Targeting a power budget of 10 pJ/bit for the whole system, a higher order modulation scheme should be considered. That requires high resolution DACs, which could increase the implementation complexity. [14] suggests that a 64-QAM requires at least a 10 bit DAC. Therefore, we consider choosing an RF architecture which does not require a DAC. In such architecture, the power consumption of PA must be taken into account. Usually for the constant envelope RF signal, the efficiency of PA can be maximized. Therefore, from the power limited region, a modulation scheme providing a reduced peak-to-average power ratio (PAPR) is highly predesignated.

1.3.2 Power limited region

1.3.2.1 Power limited system

In a power limited system, constant envelop modulation systems like continuous phase frequency shift keying (CPFSK) are valuable alternatives, which fully exploit the power capability of the available power amplifiers and hence

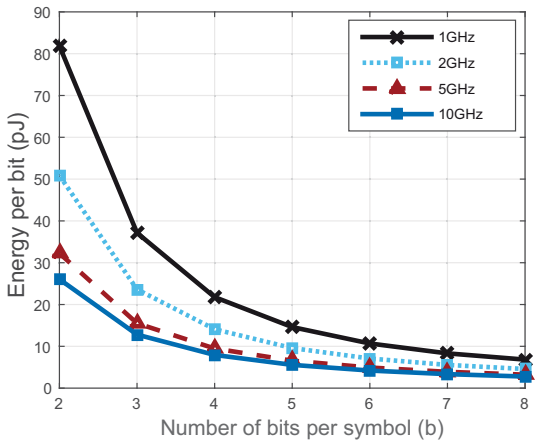

Figure 1.5
Transceiver energy consumption for different symbol rates at 60 GHz. From [15] © 2017 DeGruyter.

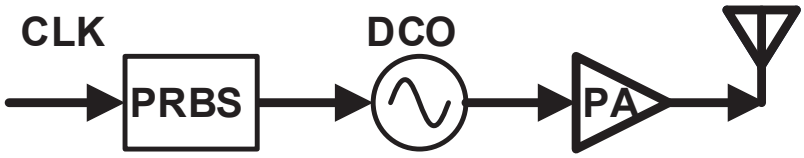

Figure 1.6
Block diagram of the FSK transmitter.

maximize the conversion efficiency. The first advantage of CPFSK modulation is the possibility to use highly nonlinear but high efficiency amplifiers with low power consumption. Another advantage is the elimination of the DACs. Furthermore, the CPFSK modulation scheme has low implementation complexity for both transmitter and receiver architectures.

1.3.2.2 CPFSK transmitter

Fig. 1.6 shows the block diagram of the CPFSK transmitter. The baseband data from the pseudo-random bit sequence (PRBS) generator triggers the digitally-controlled oscillator (DCO) to generate the carrier frequencies and the output signal then will be amplified by a power amplifier (PA). Therefore, the PRBS generator and the DCO are the crucial components in the proposed prototype of the CPFSK transmitter.

PRBS generators are mostly based on linear feedback shift registers

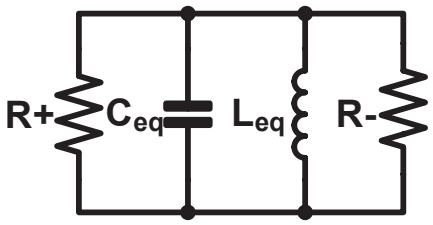

Figure 1.7
Small-signal model of LC-tank oscillator. From [23] © 2019 IEEE

(LFSR). The order (L) of the LFSR determines the maximum sequence length of 2^L-1 under the synchronization of a clock. To achieve a high-data-rate PRBS generator, multiple low-speed sequences are combined together using multiplexers (MUX). The trade-off between power consumption and data rate should be considered. In terms of high speed, all the circuits such as inverters, DFFs and XOR gates are designed in current-mode logic (CML). The CML structure is based on the differential MOS pair which provides an excellent protection to switching noise. Hence, a high supply noise suppression can be achieved. Furthermore, the current-mode approach allows high speed switching [16, 17]. Inductive peaking technique is widely applied to enhance circuit bandwidth[18, 17, 19]. In [20], a 2^7-1, 11.8 Gb/s full-rate PRBS generator with low-power consumption is demonstrated in a standard 65 nm CMOS technology.

Literature research shows recently mmW DCO for modulation have not been explored so far. The DCO in the high data-rate CPFSK transmitter requires a wide output frequency spacing which allows the demodulator to detect the carrier frequencies easily, and also have a better signal-to-noise ratio (SNR). In [21][22], the wide tuning range DCOs, however, consume considerable power and hence they are not suitable for low-power applications. Therefore, an ultralow power consumption, 1-bit DCO for the application in high data-rate CPFSK modulator is targeted. Fig. 1.7 shows the equivalent small-signal model of the oscillator. C_{eq} is the total capacitance of the resonator. L_{eq} and $R+$ are the equivalent inductance and the total resistance of the LC tank, respectively. The negative resistance, $R-$, is provided by the cross-coupled pair which approximately equals to $-2/g_m$, where g_m is the small-signal transconductance of the transistors. $R-$ must be smaller than R_{eq} to guarantee the oscillation start-up condition [22].

In a DCO, the digital input data controls the instantaneous frequency either by directly switching the capacitance in the tank or the inductance. For a 1-bit DCO requiring a wide tuning range, the capacitance is easily controlled by the switch. Therefore, the switch in the design provides the capacitance

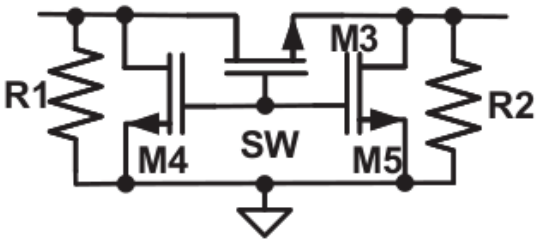

Figure 1.8
Schematic of high speed switch. From [23] © 2019 IEEE

difference during the switch on and off states, and also determines the capacitance difference between the two states to generate the output frequency range. Fig. 1.8 shows the schematic of the switch, which is used to control the equivalent capacitance of the tank. It composed of the transistors, $M3$, $M4$, $M5$ and resistors $R1$ and $R2$. The three transistors turn on and off simultaneously so that the injected channel charge caused by $M3$ can be quickly discharged by $M4$ and $M5$. Two resistors also help to improve the switch speed.

To guarantee the oscillation start-up condition, $R+$ should be more than two times $R-$. Therefore, this design requires very high-Q capacitors and inductors. In [23], it describes the design of the high-Q capacitor and inductor by optimization of the physical layout of passive components. The comparison results show the Q factor of the proposed capacitor is more than 5 times than the one provided from the technology (65 nm CMOS); the Q factor of the proposed inductor also shows the advantageous over the one from the technology library. As a result, a 1-bit DCO with output frequency swing of 3.5 GHz at 71.5 GHz with power consumption of 12 mW.

1.3.2.3 CPFSK receiver

For the CPFSK receiver, to detect the carrier frequencies, there are two methods for the demodulation: frequency-to-amplitude conversion (FAC) and frequency-to-phase conversion (FPC), respectively. Fig. 1.9 shows the basic block diagram of the demodulation based on FPC. It composes of a power divider, a mixer and a phase shifter. The phase shifter is used to convert the frequency information to the phase information which could be detected by the mixer. For the different frequency signals, the phase shifter could cause different phases and moreover, the output of the mixer depends on the phase difference between the two input signals. FPC has been applied at a carrier frequency of 120 GHz and a data-rate of up to 7.4 Gb/s in [24] with a power consumption of 41 mW. In the behavioral FPC model, the error vector magnitude (EVM) is below 10% when the data rate up to 10 Gb/s. Fig. 1.10 describes

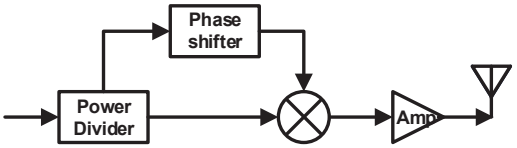

Figure 1.9
Block diagram of frequency-to-phase converter.

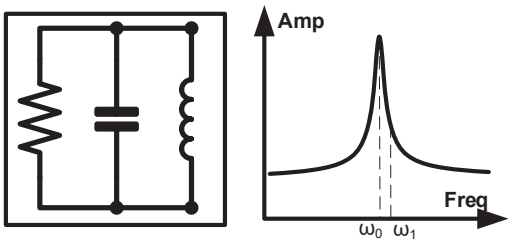

Figure 1.10
Block diagram of frequency-to-amplitude converter.

the basic idea of FAC demodulation method. The design of the *LC* tank aims to make the resonance frequency to be as close to one carrier frequency as possible. The purpose is to maximize the difference of the two amplitudes. The method of injection-locking-oscillator based on *LC* tank is also under the consideration because of the reduced phase noise. This architecture requires a high *Q* factor to maximize the amplitude difference however, it can cause the injection-locking range to be narrow. The trade-off between Q factor and the bandwidth of the *LC* tank should be considered. Moreover, since the CPFSK transmitter is based on the DCO modulation, there is the possibility of variation in the carrier frequencies. The tuning method of the resonance frequency has to be explored.

1.3.3 Bandwidth limited region

The required wireless communication data rates follow an increasing trend. Essentially, the targeted high data-rates are not realizable within the congested frequency spectrum of todays mobile communication frequencies. As well as Shannon's channel capacity theorem implies that increasing data-rate requires higher bandwidth for a limited signal-to-noise ratio. In high bandwidth millimeter-wave applications multicarrier transmit schemes provide advantageous over single carrier schemes regarding channel effects. Nevertheless,

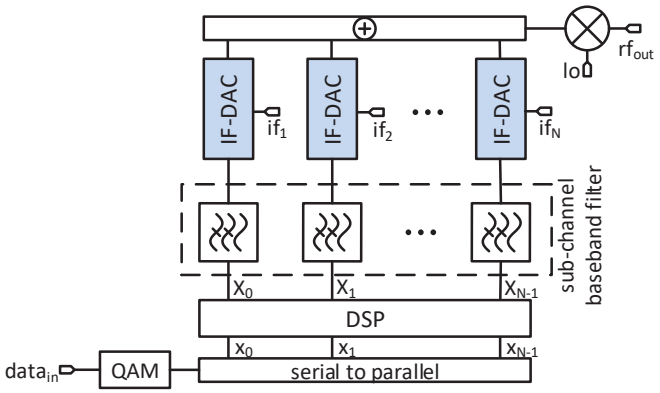

Figure 1.11
Block diagram of the proposed FBMC transmitter. From [25] © 2016 IEEE

the complexity of digital baseband processing restricts the use of transmit schemes such as OFDM since the high bandwidth signal needs to be oversampled, digitally filtered and sampled by the digital-to-analog converter (DAC) at a very high speed.

1.3.3.1 Introduction to FBMC

The proposed topology for the filter-bank multicarrier (FBMC) transmit scheme reduces the digital baseband filtering complexity by keeping the advantageous of multicarrier systems, as shown Fig. 1.11. The baseband data is quadrature amplitude modulated and represented in the time domain. Using a proper digital signal processing such as Fourier Transformation the baseband signal is separated in several subchannels so that the oversampling and filtering for each subchannel is performed for a fraction of the total bandwidth. The digital filters are designed with a certain overlap factor [26]. Nonetheless, because of the filtering of each subchannel the FBMC transmit scheme reduces the out-off-band transmission compared to OFDM [27]. The subchannels are fed to individual IF-DAC based transmitter which is then shifted to the carrier frequency by mixing with an additional LO signal. Hence, the subchannels are equidistantly spaced in the spectrum. Each of the IF-DACs requires an IF signal with a constant frequency offset to the previous one. The switching rate of the IF-DACs can therefore be reduced by factor N at the cost of area which increases by the same amount. The proposed IF based FBMC transmitter aims to operate in the 60 GHz ISM-band with a bandwidth of 2 GHz and a parallelization factor of N = 16 so that the intermediate frequency (IF)

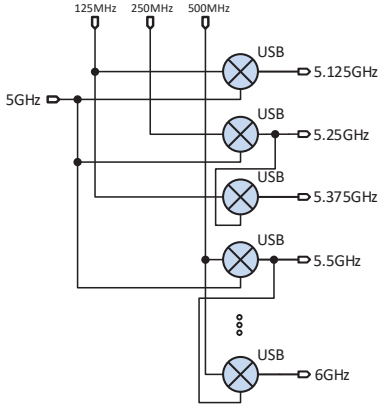

Figure 1.12
Block diagram of a mixer bank generating the upper side band (USB) IF signals. The generation of the lower side band (LSB) IF signals is done by the same mixer bank using LSB mixers. From [25] © 2016 IEEE

offset is 125 MHz. The IF-DACs shift the baseband signal to an IF of 5 GHz which is then upconverted by a millimeter-wave mixer to the 60 GHz range.

1.3.3.2 Frequency Synthesis

The simultaneous synthesis of multiple equidistantly spaced frequencies comes with several challenges. Phase-locked loops have disadvantages for this purpose because their integration is expensive. It requires significant chip area and integration of several voltage-controlled oscillators on a single chip causes pulling effects leading to performance degradation. Furthermore, single high frequency PLL can be combined with fractional frequency dividers. However, this approach delivers a nonconstant frequency spacing between the individual frequency components. The proposed frequency synthesis solution is based on a single sideband mixer. It is capable of generating several IF signals for the proposed FBMC transmitter structure with a uniform frequency spacing of 125 MHz between 4 GHz and 6 GHz. Fig. 1.12 shows the block diagram of the mixer. Fig. 1.13 shows the post-layout simulation results of the synthesized tones.

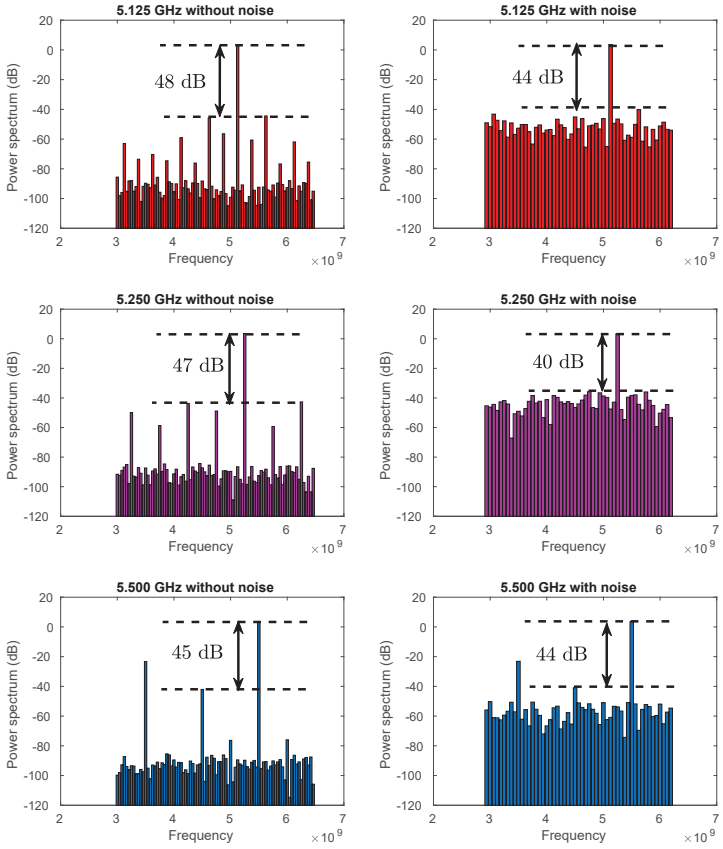

Figure 1.13
Post-layout simulated spectrum of the synthesized equidistant IF tones with and without noise.

1.3.3.3 Radio-frequency Digital-to-analog-converter (RF-DAC) Design

In modern communication systems data is transmitted digitally. This digital data is given to the transmitter with a certain sampling rate. To feature large bandwidths, the sampling rate of the transmitter must be high enough and at least twice as high as the maximum supported bandwidth in the particular case of FBMC twice as the sub-channel bandwidth. The transmitter must hence provide a digital to analog conversion on the one hand and an up-mixing to the transmit band on the other hand. Modern CMOS technologies only propose a very poor analog performance hence realizing an integrated analog mixer is not feasible. The RF-DACs shift the digital to analog conversion towards the output by integrating both the up-mixing and the conversion

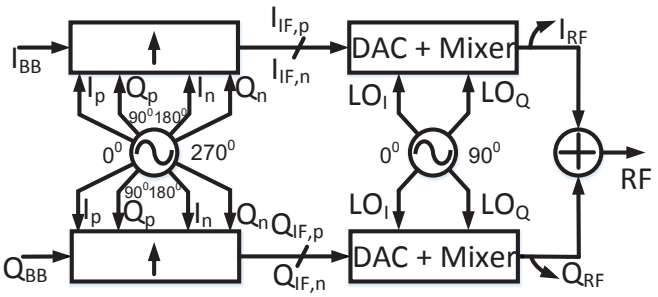

Figure 1.14
Block diagram of the proposed digital IF-DAC structure. From [28] © 2019 IEEE

in one single block. Thus, the RF-DAC is the main building block of the transmitter: It executes both the digital to analog conversion as well as the up-mixing to the transmit band and therefore determines the sampling rate as well as the resolution of the transmitter. This means it mainly determines the performance of the complete transmitter.

In Fig. 1.14 the proposed IF-DAC is shown. Here, baseband I and Q data is upconverted digitally to the IF-domain and converted and upmixed to the analog RF-domain in the final DAC+Mixer block which is based on a RF-DAC structure. Post-layout simulation based constellation diagram is shown in Fig. 1.15. The calculated error vector magnitude (EVM) is less than 6%.

1.3.4 Proposals of RF-architectures

With regard to the power efficiency at high frequency within available bandwidth, constant envelope modulation scheme such as CPFSK provides the advantageous; According to the bandwidth efficiency, at low frequency a higher order modulation scheme should be considered. The proposed FBMC based transmitter realizes it with reduced power consumption.

1.4 Frontend processing

Frequency selectivity is a severe issue especially when using single carrier with a wide bandwidth. Therefore Time Delay Compensation (TDC) is used to compensate the delay spread of several paths [10]. If implemented precisely, no delay between different paths causes a flat channel over a wide range.

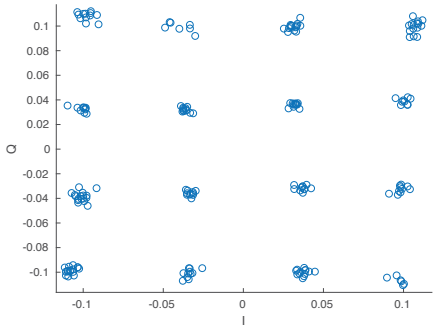

Figure 1.15
Post-layout simulation based constellation diagram of a 16QAM signal.

By employing RF beamforming and increasing number of antennas obtain a smaller beam, which reduces the intra-cluster delay spread. Adjustable buffer in the digital baseband by delaying each stream after baseband processing and before analog conversion compensate the inter-cluster delay spread.

An acceptable beam separation is also mandatory to achieve a flat fading channel. Thus, we propose a signal-to-interference (SIR) constrain, that reduces the interference of beams onto each other. Fig. 1.16 shows the effect of SIR threshold, when increasing the threshold, weak paths are dropped, hence the average number of paths that can be resolved decreases. This reduces the channel rank and leads to lower data rates. On the other hand, increasing the SIR threshold leads to a higher received SNR and a significant reduction of frequency selectivity. Hence the equalization complexity at the user side reduces.

While the threshold is high enough, the frequency selectivity becomes negligible and thus the equalizer unnecessary. A further positive effect of increasing threshold is that the gain in each subchannel is higher and, therefore, higher order modulation can be employed. This results in an increase in the overall data rate. Thus, positive and negative effects occur. The overall effect is such that the data rate is reduced by increasing SIR threshold as can be seen in Fig. 1.17. Thus, the SIR threshold enables a trade-off between the performance of the system, i.e. the data rates, and the complexity of the equalizer.[29]

1.4.1 Joint Pre-/Post-processing

The joint design of the pre- and postprocessing matrices to maximize the data rate of the proposed system was presented in [12]. In addition, we describe the influence of N_{rf} and N_s on the BER performance [30]. Fig. 1.18 depicts BERs

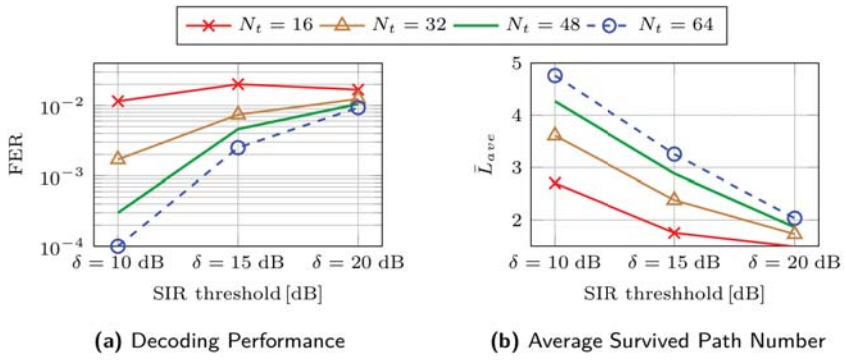

Figure 1.16
Effect of different SIR thresholds on FER and average number of survived paths (N_r: number of receive antennas, N_t: number of transmit antennas, N_s: number of data streams). From [15] © 2017 DeGruyter

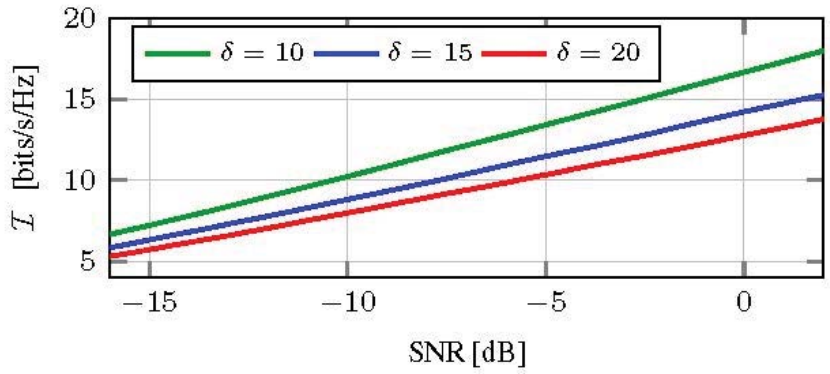

Figure 1.17
Effect of the SIR threshold on the information rate in a 16x16 MIMO system with two data streams, $N_s = 2$. From [29] © 2016 IEEE

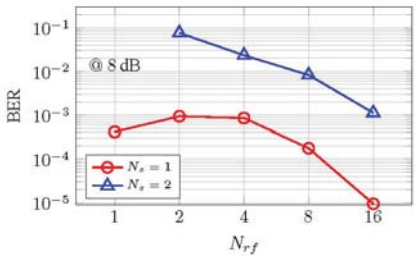

Figure 1.18
Average uncoded BER of a 16x16 MIMO system with various number of RF chains both at transmitter and receiver side. From [30] © 2015 IEEE

obtained by simulation for different number of RF chains with $N_s = 1$. It can be seen that for the case where $N_{rf} = 2$ and $N_{rf} = 4$ the BER performance is worse as compared to the case $N_{rf} = 1$.

This may be attributed to the fact that the increase in diversity gain due to the increase in the number of RF chains does not fully compensate for the beamforming loss due to decrease in the number of antennas in each subarray. However, as the number of RF chains is increased to larger numbers, the increase in diversity gain is much larger as compared to the decrease in the beamforming gain. As a result, the overall gain of the system increases which leads to a better performance. It can be seen from Fig. 1.18 that the BER performance is much better for cases where $N_{rf} = 8$ and $N_{rf} = 162$ when compared to cases with lower values of N_{rf}. This better performance also leads to a higher complexity.

Fig. 1.19 shows the Energy Efficiency (EE) as a function of the number of transmit antennas and the number of receive antennas, where the bandwidth is 1 GHz. Another variable is the amount of transmit SNR. The transmit SNRs are $\rho = 5$ dB and $\rho = 10$ dB. It can be easily understood, that increasing the N_t has a positive influence on the EE. However, with respect to N_r, an optimum point behavior of EE curve is observed. Since not only the data rate but also the circuit operating power consumption is increasing by the increment of the number of receive antennas, the tendency of the EE shows concave shape with respect to the transmit SNR as expected. Therefore, the optimal point for EE appears. Although the absolute value of the EE is increasing along the increment of the transmit SNR, the number of receive antennas to maximize the EE is decreased. The reason is that the rate of increase of data rate is bigger than the power consumption. It is worth noting that the EE curves more sharply for high SNR. Except that, the SE of the system obviously goes on increasing with increasing the N_r as well as N_t.[31]

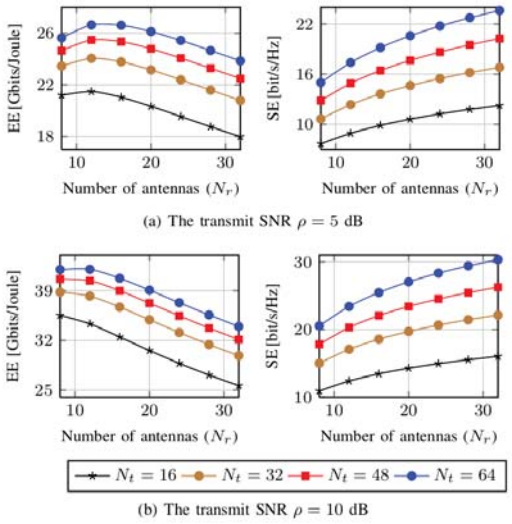

Figure 1.19
Spectral efficiency and energy efficiency for different antenna array sizes. From [31] © 2016 IEEE

1.4.2 Channel Estimation - Beamforming Training

Channel estimation is an important issue for systems with large antenna arrays employing beamforming. The proposed Hierarchical Beamforming Training (HBT)[5] is inspired by the sparsity of a mm-wave channel to lower its complexity. This training process estimates important parameters, such as the path gain, the angle of departure (AoD) and arrival (AoA). We can reconstruct with these the channel matrix modeled with respect to the Extended Saleh-Valenzuela (S-V) model [4].

The training process follows the divide and conquer principle in a greedy manner. During that, we can control the beam width with the number of active antennas and beam direction by the analog frontend. It has to be done before any communication can take place because we would not be able to steer the beams towards the user without any knowledge. We assume that this training process is triggered by the AP and we have an established fall-back channel here as a feedback channel from the user.

From the combinations of AP and User, we test κ^2 Rx-Tx beam pairs and choose the best pair based on the strongest received signal. The selected beam on each side is divided for the next run approaching the final beam resolution. At the last resolution level we derive the first parameter set of path gain, AoA and AoD. This process will be repeated for the next dominant paths in the environment with a small extension. Now the interference by

the previously estimated beam(s) has to be evaluated for the current beam. Before we select the best beam pair out of the training vectors, we have to derive the signal contribution that arises from other estimated beams onto the actual considered one. Although we employ beamforming, the power is steered to a preferred direction but still with some side effects, that depend on the angle difference and the resolution level or even the width of the beam. The feasibility of a VLSI implementation for realizing the signal processing of this training procedure is shown in [32].

1.5 Digital Baseband Signal Processing Beyond 100 Gb/s

This section focuses on the digital baseband signal processing part of the receiver. The receiver structure of a Multiple Input Multiple Output (MIMO) Bit-Interleaved Coded Modulation (BICM) scheme, is shown in Fig. 1.20. The two important building blocks are the MIMO detector and the channel decoder. After the preprocessing stage, an incoming data block is processed in an iterative loop between the MIMO detector and the channel decoder respectively, i.e., the channel decoder feeds back information to the MIMO detector. Since advanced decoding algorithms for Turbo and LDPC codes are decoding a data block iteratively, a MIMO-BICM receiver is a double iterative system. High efficiency w.r.t. spectrum and transmit power, requires high gain in the SNR. Hence, we consider in the following only advanced soft-information-based detection/decoding schemes, i.e. tree-search based MIMO detection and Turbo, Low Density Parity Check (LDPC) and Polar codes.

The MIMO detector and channel decoder are major sources of power consumption and bottlenecks for the transceiver's throughput and latency. Here we are facing two main challenges:

1. achieving high throughput ($\geqslant 100\,\mathrm{Gb/s}$),
2. achieving high implementation energy efficiency in the order of some pJ/bit.

From an architectural point of view, large parallelism, regularity and locality are key for large energy efficiency and high throughput architectures. These implementation characteristics have to be explored on all levels from the algorithmic down to the micro-architecture level. As a consequence, algorithms and implementations have to be jointly explored to find the best trade-off between communications performance and efficient implementation.

Advanced MIMO detection algorithms like sphere decoding are based on a tree traversal in which a dynamic search space reduction is performed at run-time by pruning nodes/branches in the tree. This dynamic control-driven algorithm hampers efficient parallelization. In contrast channel decoding al-

gorithms for advanced codes like Turbo and LDPC codes are data-flow dominated with much less control-flow, but they imply irregularity. Efficient channel coding is grounded on (pseudo-)randomness and occurs in form of an interleaver for Turbo codes and a Tanner graph for LDPC codes. Any regularity and locality in these structures has negative impact on the communications performance. Hence, there is a fundamental discrepancy between information theory and efficient implementations that requires regularity and locality.

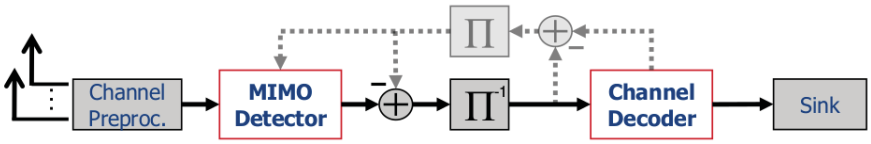

Figure 1.20
Advanced iterative MIMO-BICM receiver

In the following we first present new high throughput channel decoder architectures. In a subsequent chapter we present a high throughput MIMO detector that matches the architectures of the channel decoders resulting in a MIMO-BICM receiver that achieves a coded throughput of 160 Gb/s (information throughput of 100 Gb/s) in a 28 nm Fully Depleted Silicon on Insulator (FD-SOI) technology.

1.5.1 High-throughput Channel Decoder Architectures

Here we focus on the most advanced channel codes, i.e. Turbo, LDPC and Polar codes and present new architectures and its implementation results achieving throughput > 100 Gb/s with high energy efficiency in 28 nm FD-SOI technology.

Let us consider the communications and implementation parameters that impact the overall throughput.

- Let N be the block size and R the rate of a channel code and let I denote the number of iterations that a corresponding iterative decoder requires to decode a code block. $I = 1$ in the case of a non-iterative decoding algorithm.
- Let P denote the degree of achievable parallelism. P is defined as the ratio between the operations that are performed in parallel per clock cycle and the total number of operations necessary to perform one decoding iteration for a complete code block. Note that an operation refers to the critical operation in an algorithm and can be a computation but also a data-transfer. Obviously the maximum value of P is 1, i.e. all operations can be executed in one singe clock cycle.
- Let f be the clock frequency.

With these parameters the throughput (information bits per second) of a Forward Error Correction (FEC) architecture can be roughly estimated by

$$T_{inf} = N \cdot R \cdot \frac{1}{I} \cdot P \cdot f \cdot (1 - \omega), \tag{1.4}$$

ω is a normalized value between 0 and 1 that represents the timing overhead due to e.g. data distribution, routing, memory access conflicts etc.

The different coding schemes and corresponding decoding algorithms have different properties with regard to its efficient hardware implementation. Table 1.1 shows the implementation properties for the different code classes. The achievable parallelism P strongly depends on the properties of the decoding algorithms. Obviously algorithms with inherent parallelism are easier to parallelize on architectural level. The maximum clock frequency f is determined by the critical path in the compute kernels of the corresponding decoding algorithm and is upper limited to 1 GHz due to power and design methodology constraints. The overhead ω increases with N and P and is larger for decoding algorithms that have limited locality and are data-transfer dominated. The impact of ω on the throughput can be considered as an effective reduction of the clock frequency f and/or a decrease in P, if additional clock cycles are mandatory, e.g. due to memory conflicts, that cannot be hidden.

Table 1.1
Implementation Properties of various coding schemes. From [33] © 2018 IEEE

Code	Dec. Algorithms	Parallel vs. Serial	Locality	Compute Kernels	Transfers vs. Compute
Turbo	Maximum a Posteriori	serial/iterative	low (interleaver)	Add-Compare-Select	compute dominated
LDPC	Belief Propagation	parallel/iterative	low (Tanner graph)	Min-Sum/ add	transfer dominated
Polar	Successive Cancelation / List	serial	high	Min-Sum/add/ sorting	balanced

If we are targeting 100 Gb/s throughput with a frequency limit of 1 GHz, the minimum information block size $K = R \cdot N$ is 100 bits. To achieve highest throughput, P has to be maximized and ω minimized that will be discussed in the following sections for the different coding schemes.

1.5.1.1 High Throughput Turbo and LDPC Decoders

A state-of-the-art Turbo decoder consists of two Maximum a Posteriori (MAP) component decoders that are connected through an interleaver/de-interleaver. These component decoders work cooperatively by exchanging extrinsic information in an iterative loop. The respective bit-wise log-likelihood ratios are computed in a forward and a backward recursion on the trellis. Thus, decoding is inherently serial on component decoder (MAP recursions) and on Turbo decoder level (iterative loop).

The kernel operation in the MAP is the Add-Compare-Select (ACS) operation. Depending on the encoder memory depth, several ACS operations have

to be executed to process a single trellis step in the forward and backward recursion respectively. $4 \cdot K$ trellis computations and corresponding interleaving have to be calculated to perform a single Turbo decoding iteration. Hence, P is determined by the number of parallel trellis step computations performed by a decoding architecture. The maximum value of P for one decoding iteration, i.e. $P = 1$, can be achieved by an architecture if 1) the forward/backward recursions of the MAP algorithm are unrolled and pipelined (functional parallelism), yielding the XMAP architecture, or 2) a fully parallel MAP is applied in which every trellis step is spatially parallel executed, yielding the FPMAP architecture. A major disadvantage of the FPMAP architecture is that the interleaver has a strong impact on ω. Furthermore, the FPMAP breaks-up the data dependencies between the Trellis steps of one iteration. This in turn causes a computational overhead since neighboring state metrics need to be estimated in order to mitigate the loss in decoding performance. For a detailed overview and discussion of such highly parallel decoders we refer to [34].

LDPC decoding is based on an iterative message exchange between variable and check nodes on the Tanner graph that is represented by the parity check matrix H. The number of exchanged values in this Belief Propagation (BP) algorithm corresponds to the number of one-entries in the parity-check matrix H and is typically very large. Here, P mainly depends on the number of parallel exchanged messages. The maximum value of P can be achieved if all edges of H are processed in parallel yielding a fully parallel architecture. Since the Belief Propagation is data-transfer dominated (Table 1.1), ω largely increases for increasing N. ω also depends on the structure of H.

Turbo and LDPC decoding are iterative and, thus, sequential since the current iteration has as input the result of the previous iteration. However, the data dependencies due to the iterations can be broken up by unrolling the corresponding iterations and insertion of pipeline stages. In this way, the dependency of the throughput on I diminishes at the cost of additional pipeline memory since at least I blocks are processed in parallel in the decoding pipeline. The parallelism is thus $P = I$ that is the maximum an iterative decoding architecture can achieve.

1.5.1.2 Polar Decoders

Successive Cancelation (SC), Successive Cancelation List (SCL) are the most prominent decoding algorithms for Polar codes. Decoding corresponds to a traversal of the corresponding Polar Factor Tree (PFT) in which the received log-likelihood ratios from the channel are processed by the tree nodes. SC and SCL decoding are depth-first traversals on the PFT and thus exhibit sequential behavior. To achieve a maximum P, the tree traversal can be unrolled and pipelined, alike to the iteration unrolling in Turbo and LDPC decoding architectures. Whenever a node is visited during the tree traversal, a corresponding pipeline stage can be instantiated. In this way, for a block length of N, the maximum number of pipeline stages is $2 \cdot (2N - 2) + 1$ in which

$N \cdot \log N$ operations are performed in parallel and can be reduced by various transformations. For example, if a subtree represents a repetition code or a parity check code, the corresponding subtree can be replaced by a single node. Alike, we can merge rate-0 and rate-1 nodes into their parent nodes or use majority logic decoding in subtrees. These optimizations strongly depend on the position of the frozen bits, i.e. the code structure. Hence, appropriate codes are mandatory.

Table 1.2
Comparison of channel code decoders

Code	Turbo code (4 iter)	LDPC code (10 iter)	Polar code
Blocksize [bit]	384	672	1024
Code rate	1/3	5/8	1/2
Freq. [MHz]	800	400	746
Area [mm^2]	23.6	3.31	2.95
Power [mW]	-	2876	3300
Area efficiency [Gbit/s/mm^2]	4.34	81	259
Energy efficiency [pJ/bit]	**-**	**10.7**	**4.4**
Coded Throughput [1] [Gb/s]	**306**	**268**	**764**

[1] Opposed to information throughput in Equation 1.4.

1.5.1.3 Implementation Results

The above mentioned discussions show that decoder architectures for throughput beyond 100 Gb/s are feasible for all three code classes. The achievable throughput strongly depends on the code class, i.e., the code structure and the decoding algorithm. Maximum throughput can be achieved by heavily pipelined architectures that enable maximum functional parallelism, provide large locality but at the cost of huge number of storage elements and large latency. These storage elements are a major source of the power consumption and imply large challenges on the clock tree. We have shown in [35] that more than half of the power consumption can be consumed by these storage elements only. Hence, optimizing the storage scheme in pipelined decoder architectures is of great importance and has to be performed on various levels: e.g. efficient quantization on algorithmic level, advanced retiming to optimally distribute the pipeline stages between the compute units on architectural level and the use of latch-based design, clock gating etc. on micro architectural level.

We implemented decoders for all three coding schemes using the techniques described above and optimized for highest throughput, i.e. $P = I$ and unrolled iterations for LDPC ($I = 10$) and Turbo ($I = 4$) decoding. Target technology is a 28 nm FD-SOI technology with worst case Process, Voltage and Temperature (PVT) conditions (0.9 V for timing and 1.0 V for power, both 125 °C). Synthesis is performed using Design Compiler, Place & Route

with IC-Compiler, both from Synopsys. Table 1.2 lists the block sizes, code rates, number of iterations, maximum achievable frequency, area, power and efficiency metrics for all three coding schemes. The Turbo decoder is the most complex one. Hence, the maximum block size that can be supported under reasonable area constraints is limited to some 100 bits. LDPC and Polar decoder enable larger block sizes, hence their achievable throughput is higher. The table demonstrates that throughput larger than 100 Gb/s is feasible with an energy budget that is around or below 10 pJ/bit.

Fig. 1.21 shows the corresponding layouts. The different colors of the Turbo decoder represent the 8 MAP engines originating from the 4 unrolled iterations (each iteration requires 2 MAP decoders). The different colors of the LDPC decoder represent the 10 iterations, and each color in the Polar decoder represents a pipeline stage (105 in total).

Figure 1.21
Channel coding beyond 100 Gb/s: **Left:** 306 Gb/s Turbo decoder. The area is 23.61 mm^2. **Middle:** 268 Gb/s LDPC decoder. The area is 3.31 mm^2. **Right:** 764 Gb/s Polar decoder. The area is 2.95 mm^2. From [36] © 2018 IEEE

1.5.2 High-throughput MIMO Receiver

As shown in the previous sections very high throughput channel decoders employ heavily pipelined architectures that require the Log-Likelihood Ratio (LLR) values for a complete code block in each clock cycle. Hence, an architecture for MIMO detection has to match the pipelining scheme of the channel decoders.

Advanced MIMO detectors traverse a tree structure in order to find a shortest path, which is defined over the Euclidean distance from the root to the leaves. Although a tree has a high regularity and large data locality, the detection of the shortest path in this decision tree is difficult to parallelize. Performing an exhaustive search, i.e., computing all paths and then selecting the shortest is infeasible due to the large tree size. Therefore, advanced detection algorithms, like the sphere decoding, use sophisticated pruning techniques to limit the number of visited tree nodes during traversal to a very small subset. However this adaptive pruning at run-time results in a control driven sequential processing of a dynamic set of nodes. This dynamic control driven

algorithm hampers efficient parallelization. Hence, the throughput of sphere decoders is below 1 Gb/s in state-of-the-art silicon technologies [37].

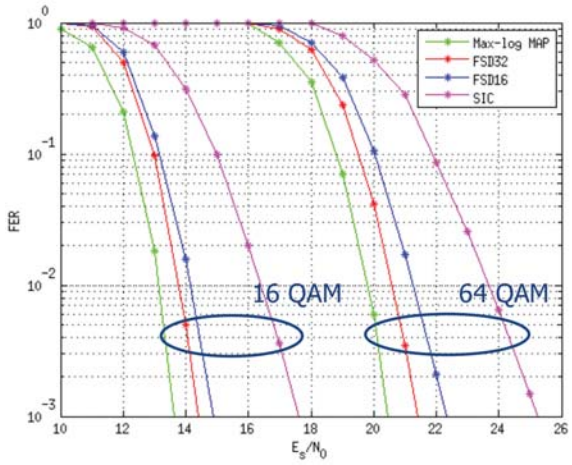

Figure 1.22
Performance comparison of the 4×4 LFSD detector for list sizes 16 and 32 for 16 QAM and 64 QAM after the channel decoder ((672,336) LDPC code, 10 iterations, Min-Sum). For better comparison, performance curves for Max-Log MAP and Successive Interference Cancelation (SIC) detection are shown.

Fig. 1.23 also shows the corresponding layout in a 28 nm FD-SOI technology. The resulting throughput is 6 Gb/s.

Instead of dynamically restricting the search space in the tree, the List Fixed Complexity Sphere Decoder (LFSD) performs an exhaustive search on a constrained tree only. The constrained subtree is defined at design time by fixing the number of nodes per layer and thus the number of paths to a small subset. Hence, the control-flow is completely removed. Analogous to the unrolling of iterative channel decoding algorithms we can completely unroll the tree structure and pipeline the different layers of the tree [38]. Fig. 1.23 shows the resulting architecture for a 4 antenna system. This architectural approach provides the largest throughput, but comes at the cost of a reduced flexibility with respect to the number of antennas and modulation and some degradation of the communications performance (see Fig. 1.22).

Other approaches that have been introduced for maximum throughput are k-best detection [39] and its soft-output extension, the Path Preserving Trellis-Search (PPTS) detector [40]. In [41], the authors present a fully parallel PPTS detector in 65 nm technology that achieves a throughput of 6.4 Gb/s. However, this detector suffers in area efficiency, especially for higher order modulations (16 QAM).

The maximum throughput of MIMO detection is limited by the size of the transmission vector that is processed by the MIMO detector. These vectors

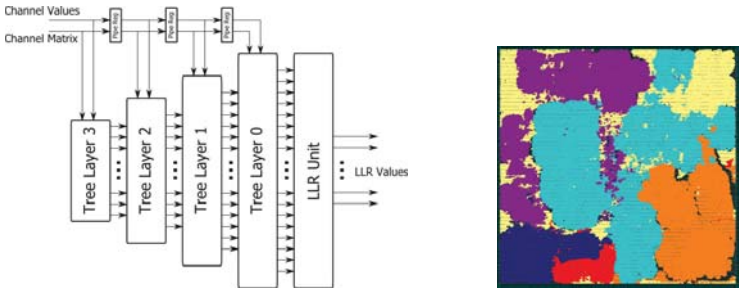

Figure 1.23
Left: High Throughput "unrolled" architecture. **Right:** Layout of MIMO detector. Layers in the detection tree are highlighted with different colors. Area is 0.15 mm^2. From [15] © 2017 DeGruyter

.

typically consist of a small number of bits. Hence, even in a fully pipelined architecture that processes one vector in one clock cycle, the MIMO detector has a much lower throughput compared to a deeply pipelined channel decoder, that operates on a complete code blocks that is composed of many transmission vectors. However, the different transmission vectors in a code block can be considered independent for MIMO detection, they can be detected in parallel on multiple detectors.

Let us consider a 4×4 64 QAM transmission system with a target information throughput of 100 Gb/s and a (672,440) LDPC decoder as described in the previous section. A complete code block for such a decoder is composed of 28 independent transmission vectors, each 24 bits respectively. We can use 28 MIMO detectors in an array structure to match the parallelism of the channel decoder, i.e., the MIMO detector array provides 672 LLR values per clock cycle. Fig. 1.24 shows the layout of the detector array implemented in 28 nm FD-SOI technology. Each color represents a MIMO detection unit. The total area of the detector array is 4.3 mm^2 and the maximum frequency 240 MHz. The MIMO detector array can be connected via a fixed interleaving network to the LDPC decoder forming a complete MIMO-BICM receiver (Fig. 1.24). The receiver achieves a coded throughput of 160 Gb/s and thus the target information throughput of 100 Gb/s at code rate $R = 5/8$. The total area of the MIMO-BICM receiver is approximately 8 mm^2. The energy efficiency is about 25 pJ/bit. Extrapolating the energy efficiency to a 7 nm technology yields less than 10 pJ/bit.

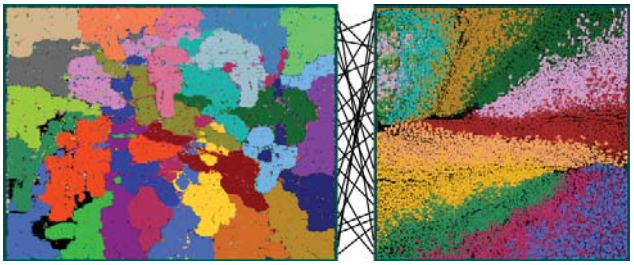

Figure 1.24
MIMO-BICM receiver for throughput of 160 Gb/s: MIMO detector array (left), LDPC decoder (right). From [38]

1.5.3 Improving Energy Efficiency for Pipelined Receiver Architectures

As shown in the previous sections, unrolling and heavy pipelining enables highest throughput beyond 100 Gb/s. However, such architectures largely limit flexibility especially with regard to iterations. The number of iterations, that can be the channel decoder iterations or outer iterations between decoder and detector, has a large impact on latency and energy-efficiency. In less parallel receivers, the inner and outer iterations are limited to a maximum value (which depends on the worst case SNR) and is adapted to the actual SNR by using iteration control techniques. This significantly improves the energy efficiency in higher SNR regions. In unrolled pipelined architecture iteration control is however difficult to implement since the number of iterations (= number of pipeline stages) is fixed at design time. Thus, if the number of pipeline stages is adjusted to the worst case SNR, we waste energy in typical SNR regions, in which less iterations would be sufficient.

A concept to improve the energy efficiency for heavily pipelined architectures is to use different decoder cores for different SNR points. Thus, in high SNR regions we can use low complexity algorithms like Successive Interference Cancelation (SIC) for MIMO detection and an LDPC decoder with smaller number of pipeline stages. Another concept is the so-called "afterburner" approach which combines a simple but hardware-efficient decoder with an advanced post-processor [42] [43]. This post-processor is most of the time in idle state and only activated if the simple decoder fails to find the correct code word. This concept was demonstrated in [43] for an LDPC decoder with a Saturated Min-Sum (SMS) post-processor. The SMS decoder has much higher complexity, but is only activated in 1% of the frames at a target FER of 10^{-3}. With this approach we were able to largely improve the communications performance with only a negligible energy overhead. This concept comes with a certain area overhead that can be efficiently exploited to alleviate the "dark silicon" challenge, which states that large parts of a chip cannot be activated

simultaneously due to thermal design power constraints [44]. We have shown in [45] that the power density can be largely reduced with this architecture.

1.6 Power Estimation

Table 1.3 shows the result and summary of the system power consumption. Note the proceeding to derive those values. At first we assume that we can scale digital VLSI implementation to a up-to-date technology node by using prediction of the International Technology Roadmap for Semiconductors (ITRS). Using this technology scaling we get the Throughput (Gb/s) and Power Consumption (mW). From there we can derive the processing energy below in pJ/bit. We sum this energy processing and get the sum power consumption by multiplying with our goal throughput of 100 Gb/s. Meanwhile we also assume that we can scale the throughput of each component by using multiple components or reducing the architecture and neglect any additional overhead of scaling.

The analog part is based on analytical models, presented earlier in [31] and use the summary from [46] for power figures of phase shifters and results of [47] in particular. From previous analyses we assume here architecture parameter of $N_{rf} = 3$ and $N_r = 8$. The chosen phase shifter implementation from [46] claims to cover a range of more than 10 GHz and provides a suitable gain, because it includes a LNA. Thus, for a fully connected phase array we need $N_{rf} = \times N_r = 24$. Table 1.3 presents both tradeoffs for the RF-Chain on the one hand the modulation order (16- or 64-QAM) and on the other hand the bandwidth or sampling rate respectively (5 GS/s or 10 GS/s). The combination of higher modulation order with 64-QAM and lower sampling rate of 5 GS/s gives depending on the SNR region (15 to 27 dB) a power consumption of about 270 to 375 mW. For similar performance 16-QAM and 10 GS/s results in about 500 to 990 mW. Considering the theoretical throughput of 90 Gb/s and 120 GS/s respectively, leads processing energy of 3 to 4.1 pJ/bit and 4.2 to 8.2 pJ/bit.

Other design approaches and implementation [48] that consider the analog frontend for phased arrays jointly, derive phase shifters with a power consumption of 2 mW each, but shifting part of that complexity into the digital baseband.

Table 1.3
Estimate of System Power Consumption, considering Phase Shifter models. From [46] © 2015 IEEE

	Dig. BB Comb.	LDPC Dec.	MIMO Det.	Sum Digit.	RF Chain	Phase Shift.	Sum Analog
Gbit/s	21	804	12	100	90-120	(100)	
mW	< 1	2151	162	1620	$90 - 990$	1200	< 2200
pJ/bit	< 0.1	2.7	13.5	16.2	$3 - 8.2$	12	< 20.2

1.7 Summary and Conclusions

This chapter has presented a new and fundamentally different approach to the design of a wireless system architecture for data rates beyond 100 Gb/s. As already a rough estimate showed, processing energy per bit is a critical and limiting factor when targeting such high data rates with portable devices, i.e. without forced cooling. Therefore, this design approach directly includes processing energy in the wireless system architecture considerations.

A first consequence is the use of single carrier modulation instead of OFDM. An advantage of OFDM is that each subcarrier is subject to flat fading rather than frequency selective fading, which greatly simplifies equalization. However, it was shown that the channel characteristics and the small antenna size can be used favorably to "flatten" the frequency selective channel and, thus, avoid excessive equalization effort.

A critical component of a high data rate GHz transceiver is the radio frequency part. Therefore, special focus was given to power efficient modulation and digital-to-analog conversion (Section 1.3). An FBMC based transmitter in combination with a RF-DAC was shown to have very advantageous properties for a reduced power consumption transceiver.

Based on published power estimates for RF components and data from the above discussed RF front end research, the next tradeoff is on beamforming, which can be done in the analog domain, in the digital domain and in analog/digital-combination (hybrid beamforming). The key results are the figures comparing beamforming strategies both from a wireless transmission performance view and from a processing energy point of view.

The digital baseband pre- and postprocessing is a key aspect for wireless transmission performance but turned out to be uncritical for the overall energy-per-bit budget.

Finally, MIMO detection and decoding are known to be 'power hungry'. However, it was shown that data rates beyond 100 Gb/s can be achieved even with reasonable energy-per-bit values. Yet, wireless transmission performance is at or above state of the art decoders. The achieved maximum bit rates were unprecedented.

Power consumption estimation always ran in parallel to the performance and design studies. For the purpose of this work, a low-percent precision is not required for the overall system. Rather, differences are in terms of multiples. Also, the estimation method has to be computationally efficient to allow comparison of a large number of alternatives. Therefore, different estimation methods were used. Analog front end component power figures were estimated based on published equations. Digital component power consumption was estimated based on high level power estimation method developed in earlier research projects partly verified by post-layout simulations and on 'real' power numbers from post-layout simulation and on measurement (MIMO detection and LDPC decoding). The overall result show that more than 100 Gb/s are feasible under realistic processing energy constraints but that it is a close call, i.e. the results are close to the limits. Thus, it is a must to consider the processing energy already in the wireless system architecture design phase when data rates beyond 100 Gb/s are targeted with portable devices.