Lousy Processing Increases Energy Efficiency in Massive MIMO Systems

Sara Gunnarsson*, Micaela Bortas*, Yanxiang Huang†‡, Cheng-Ming Chen†,
Liesbet Van der Perre† and Ove Edfors*
*Department of EIT, Lund University, Lund, Sweden
†Department of ESAT, KU Leuven, Leuven, Belgium
‡imec, Leuven, Belgium
Contact author: sara.gunnarsson@eit.lth.se

***Abstract*—Massive MIMO (MaMIMO) is a key technology for 5G wireless communication, enabling large increase in both spectral and energy efficiency at the same time. Before it can be deployed, it is important to find efficient implementation strategies. Because of the many antennas, an essential part of decreasing complexity, and further improving energy efficiency, is optimization of the digital signal processing (DSP) in the per-antenna functions.**

Assuming an orthogonal frequency-division multiplexing (OFDM) based MaMIMO system, this paper explores coarse quantization in the per-antenna digital transmit filters and inverse fast Fourier transforms (IFFTs) and evaluates it in terms of performance and complexity savings. Results show that DSP complexity can be greatly reduced per-antenna, and therefore significant power savings can be achieved, with limited performance degradation. More specifically, when going towards MaMIMO and therefore increasing the number of antennas from 8 to 64, it is possible to reduce the complexity in each transmit filter by 55%. Also, when using 6 bits to represent the input signal and 6 bits for the filter coefficients, this results in an SNR degradation of less than 0.5 dB compared to floating-point performance. Consequently, we conclude that the overall system energy greatly benefits from lousy per-antenna processing.

***Index Terms*—Massive MIMO, energy efficiency, digital signal processing, low accuracy, quantization**

I. Introduction

Global mobile data traffic is continuously increasing as the use and applications of wireless communication spreads more and more. The number of users and communicating devices follows the same trend. At the same time, energy consumption for networks is increasing faster than the total worldwide electricity use. To be able to meet these challenges, a technology that can provide higher spectral efficiency, at the same time as being energy efficient, is needed.

Massive MIMO (MaMIMO) is one of the most attractive technologies for fulfilling the 5G requirements, since it can provide increased spectral efficiency while still enabling more energy efficient solutions. By using spatial multiplexing in a time-division duplex (TDD) mode, great capacity can be achieved in these systems [1]. Furthermore, the array gain and linear processing results in energy savings in the overall system. It has also been validated that MaMIMO works in a wide variety of situations in real-life testbed experiments, achieving a world record in spectral efficiency [2]. For these reasons, MaMIMO has become a clear candidate when standardizing 5G. What is yet to be progressed before it can be deployed, are efficient implementation solutions.

The basic idea, which MaMIMO is built on, is to use a large number of antennas relative to the number of active terminals. As proposed, the number of antennas will be into the hundreds. Since the per-antenna processing in the base stations is dominating the complexity in the digital signal processing (DSP) part, it is essential to investigate the per-antenna functions when developing energy efficient solutions. We address MaMIMO systems based on OFDM where, more specifically, the complexity in the per-antenna functions is dominated by the transmit filter and IFFT [3].

One way to make MaMIMO more energy efficient is to reduce complexity and resolution in the system. For example this can be done by utilizing error-prone digital hardware [4], or by lowering the accuracy at the end of the digital transmit chain. It has been shown that a MaMIMO system can operate correctly with only 2 or 3 bits with an SNR degradation as small as 1 dB [5].

This paper demonstrates that with MaMIMO array gain, not only the power amplifier (PA) power, but also the digital complexity for each antenna can be reduced. We explore and evaluate performance as well as calculate complexity savings, when complexity in the per-antenna functions is reduced by lowering processing accuracy. The focus is on the final quantization implementations, combined with the power perspective in terms of complexity. We start with quantizing the signal in combination with quantization of either filter coefficients or IFFT twiddle factors. Finally, the effect of this quantization is evaluated in terms of performance and complexity savings per antenna-chain when increasing the number of antennas.

Section II describes the system model and presents the scenario of the simulations. Section III elaborates on the complexity analysis which is later applied in comparisons. Section IV presents the quantization performed in the transmit filter and IFFT, followed by results and evaluation of the performance including numbers for possible complexity savings, when increasing the number of antennas. Finally, Section V presents the conclusions.

978–1–5386–3873–6/17/$31.00 © 2017 IEEE

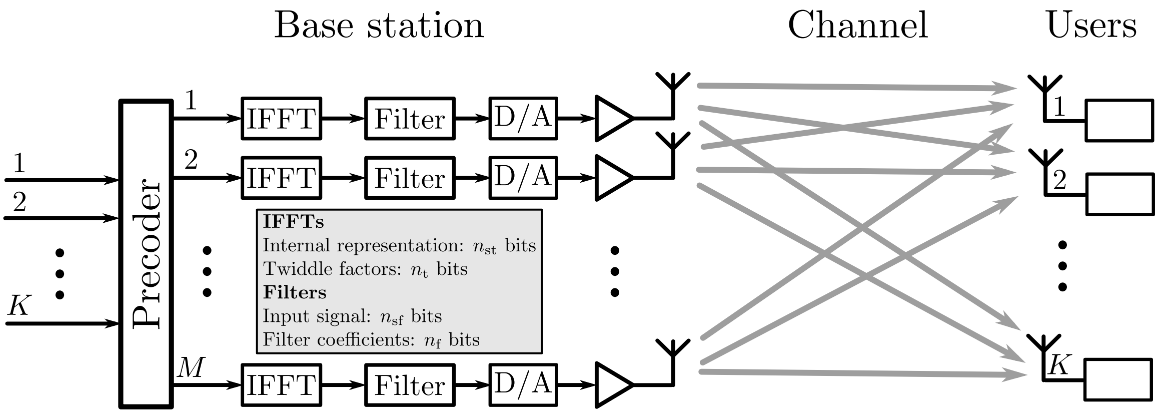
Base station
Channel
Users
Precoder
1
2
K
1
2
M
IFFT
Filter
D/A
IFFTs
Internal representation: n_{st} bits
Twiddle factors: n_{t} bits
Filters
Input signal: n_{sf} bits
Filter coefficients: n_{f} bits
1
2
K

Fig. 1. The MaMIMO downlink system model, with K single-antenna users served by M base station antennas. Transmit chain signals are quantized to n_{st} bits internally in the IFFT and n_{sf} bits on the transmit filter input. Twiddle factors in the IFFT are represented by n_{t} bits and filter coefficients by n_{f} bits.

II. System model

A MaMIMO system consisting of a base station with M antennas and K single-antenna users is considered. Throughout this work, the cases with 8, 32, 64 or 128 antennas and 4 users are usually considered. The different combinations will be stated as $K \times M$. The system is working in TDD mode and perfect knowledge of channel state information (CSI) is assumed. Similar to 3GPP Long-term Evolution (LTE), the bandwidth is 20MHz and 1200 of 2048 subcarriers are used for data transmission, divided into 100 resource blocks [6].

The structure of the system model can be seen in Fig. 1. The focus is on the downlink scenario since it is usually more power consuming because of the higher traffic load. The system starts with signal processing for the K users, including data generation and symbol mapping (not explicitly shown in the figure). The modulation focused on is 64-QAM, in order to explore how a case that normally requires very accurate processing responds to coarse quantization. Channel coding is not included in the simulations unless explicitly specified.

Following the per-user processing is a MaMIMO precoder using zero-forcing (ZF). After that comes the per-antenna processing for the M antennas, which is the main focus of this work. This process includes OFDM modulation, i.e an IFFT, upsampling and filtering. After passing the Rayleigh-fading channel, where each antenna to antenna link is generated independently with a power delay profile (PDP) according to ITU Pedestrian A [7], there is independent data detection for each one of the K users.

III. Complexity analysis

To assess the energy consumption of the digital processing, the arithmetic complexity of the transmit filter and IFFT is used. While the complexity of DSP is often assessed in terms of giga-operations per second (GOPS), a more in-depth study was pursued in order to quantify the effect of coarse quantization. Therefore, the relevant relationships are derived, depending on the word lengths, for the required adders and multipliers needed to implement IFFTs and transmit filters. The resulting complexity, C, in the IFFT and transmit filter can be calculated in terms of number of additions and multiplications being made during the respective operation.

The complexity calculations below are based on the Ladner-Fischer high-speed adder [8], whose complexity is

$$C_{\text{adder}} = n \log_2(n)$$

for a maximum n bit input and the Baugh-Wooley high-speed multiplier [9], whose complexity is

$$C_{\text{multiplier}} = n_1 \cdot n_2$$

for n_1 and n_2 bit inputs.

A. Complexity of IFFT

Following LTE, a 2048-point IFFT is implemented. The total number of butterfly units is $\log_2(2048)$ stages $\cdot \frac{2048}{2}$ butterflies per stage. Each butterfly has 6 adders and 4 multipliers, because of its complex nature. Apart from the arithmetic complexity, the data transfers (memories and registers) also contribute to the complexity of the processing [10]. Therefore, the IFFT complexity is multiplied with an overhead factor of 2 and finally the IFFT complexity per sample is estimated as

$$C_{\text{IFFT}} = 66 \cdot n_{\text{st}} \cdot \log_2(n_{\text{st}}) + 44 \cdot n_{\text{st}} \cdot n_{\text{t}}, \quad (1)$$

where n_{st} is the number of bits used to represent the signal internally and n_{t} is the number of bits used to represent the twiddle factor. In the IFFT implemented in the system model, the same internal representation is used through all the stages.

B. Complexity of transmit filter

The transmit filter used in the simulations has roll-off factor 0.25, filter span 10 and upsampling factor 2. The 21 taps, resulting in 42 adders and 42 multipliers, are used to calculate each output sample. Since processing is in the complex baseband, both real and imaginary parts of the signal needs processing, resulting in the factor of 2. Given this, the filter complexity per sample is estimated as

$$C_{\text{filter}} = 42 \cdot (n_{\text{f}} + n_{\text{sf}}) \cdot \log_2(n_{\text{f}} + n_{\text{sf}}) + 42 \cdot n_{\text{f}} \cdot n_{\text{sf}}, \quad (2)$$

where n_{sf} is the number of bits used for the input signal and n_{f} is the number of bits used for the filter coefficients. The internal representation of the transmit filter is $n_{\text{sf}} + n_{\text{f}} + x$, where x comes from the increased dynamic range when adding contributions from all filter taps. For the most pessimistic case, x would be $\log_2(21) = 5$. In this paper, a more realistic case is considered where x is calculated from the filter coefficients.

IV. Coarse quantization: Exploration and assessment

In order to reduce the complexity in the per-antenna processing, coarse quantization in the IFFT and transmit filter was performed respectively, while assuming full precision for the function not in focus. The complexity of these functions scales with the number of antennas, and therefore, it can result in a significant portion of the overall complexity in the DSP.

By simulating uncoded performance and comparing against a target BER, for various combinations of IFFT and transmit filter quantizations, the quantization combinations that deliver required performance are found. The presented curves show the shortest word lengths meeting, or exceeding, the BER performance requirement. Specifically, when focusing on the corner points, the optimum in terms of complexity per antenna for that specific function, which is needed to achieve the targeted performance, can be found.

The number of bits for internal representation of the signal is used in plots for the IFFT, while the number of bits for representing the input signal is used in plots for the transmit filter. This because the internal representation is the same through all the stages in the IFFT, while the internal representation in the transmit filter will vary depending on the word length of the filter coefficients.

A. Performance analysis of IFFT

For a chosen Additive White Gaussian Noise (AWGN) SNR value at 18 dB and a target BER of 10^{-3}, quantization for the internally represented signal and twiddle factor was performed. The BER 10^{-3}, which is reached at SNR 18 dB for the 4×8 system when using floating point, was chosen as a reasonable BER which with relatively low-complex channel coding can be improved to sufficient performance. The results were assessed based on simulations for four different combinations of $K \times M$ with an uncoded 64-QAM signal.

The graphs in Fig. 2 represent the minimum required bits in the IFFT for different combinations of $K \times M$. What can be seen is that graphs corresponding to the cases with 32, 64 and 128 base station antennas respectively are overlapping, which indicates that an increase of the number of antennas relative the number of users no longer compensates for the loss of accuracy caused by the quantization. The minimum required number of bits for these numbers of antennas are 8 for the twiddle factor and 15 for the internally represented signal. Comparing these three cases to the 4×8 case, it can be seen that processing with lower resolution per antenna is possible when the number of antennas increases.

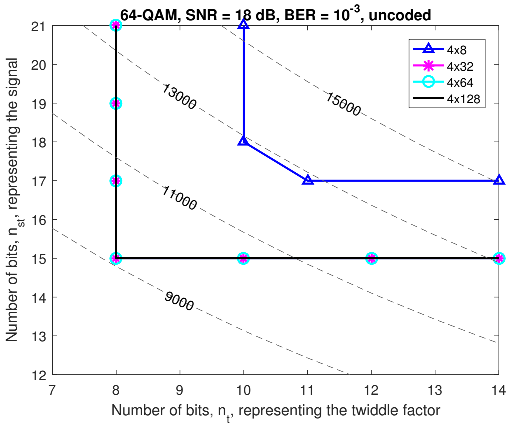

Fig. 2. IFFT quantization with 4 users and between 8 and 128 base station antennas. Number of internal signal representation bits n_{st} and twiddle factor bits n_{t} needed in the IFFT to achieve an uncoded target BER of 10^{-3} at 18 dB SNR using 64-QAM. IFFT complexity contour lines based on Eq. (1) are dashed and grey.

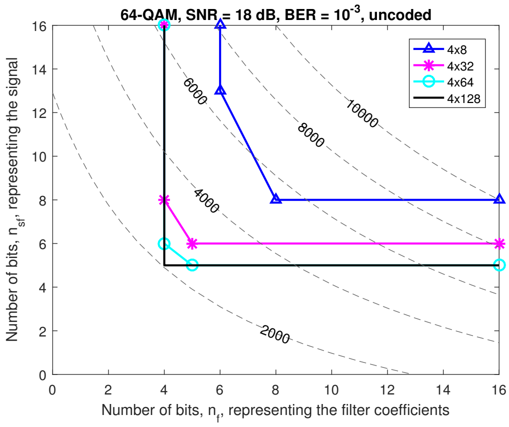

Fig. 3. Transmit filter quantization with 4 users and between 8 and 128 base station antennas. Number of input signal bits n_{sf} and filter coefficient bits n_{f} needed in the transmit filter to achieve an uncoded target BER of 10^{-3} at 18 dB SNR using 64-QAM. Filter complexity contour lines based on Eq. (2) are dashed and grey.

Using calculations from Section III, the IFFT complexity contour lines are included in the graph, in order to improve the comparison between the different options in terms of complexity per antenna-chain. Complexity is, quite naturally, lowest in the bottom left corner of the figure and grows with the number of bits. While there is only one corner point for larger number of antennas, making it clear which option delivers the lowest IFFT complexity per antenna, the contour plot is helpful for the 4×8 case, where it can be seen that the two corner points have roughly the same IFFT complexity per antenna. When increasing the number of base station antennas from 8 to 32 or above, the IFFT complexity can be reduced by 29% in each antenna.

B. Performance analysis of transmit filter

Input signal and filter coefficient quantizations for the transmit filter are investigated using the same BER performance requirement and SNR as for the IFFT. The resulting graphs are shown in Fig. 3. Similar to the IFFT analysis, the complexity contour lines, calculated from the relationship derived in Section III, are included to make it possible to find the least complex quantization combinations fulfilling performance requirements.

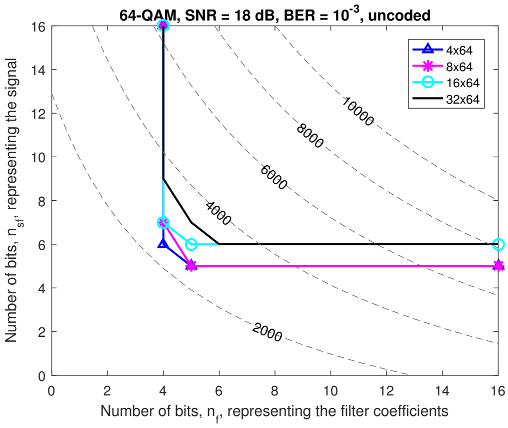

Fig. 4. Filter quantization with 64 base station antennas and between 4 and 32 users. Number of input signal bits n_{sf} and filter coefficient bits n_{f} needed in the transmit filter to achieve an uncoded target BER of 10^{-3} at 18 dB SNR using 64-QAM. Filter complexity contour lines based on Eq. (2) are dashed and grey.

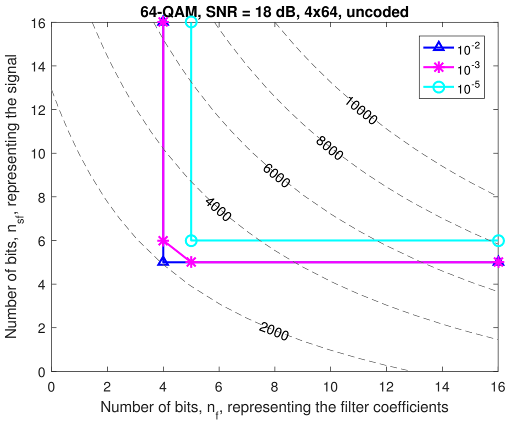

Fig. 5. Filter quantization with 4 users, 64 base station antennas and varying target BER. Number of input signal bits n_{sf} and filter coefficient bits n_{f} needed in the transmit filter to achieve uncoded target BERs of $10^{-2}, 10^{-3}$ and 10^{-5} at 18 dB SNR using 64-QAM. Filter complexity contour lines based on Eq. (2) are dashed and grey.

The cases with 64 and 128 antennas are almost overlapping and for the case with 32 antennas only a few more bits are needed. For the case with 8 antennas the number of needed bits is significantly larger, which also was the case for the IFFT in Fig. 2. Using only 8 antennas is not large enough to benefit from the law of large numbers to the same extent as the other cases. The system load per antenna is high and the number of antennas is too low for coarse quantization effects to effectively average out. A more specific example, when increasing the number of base station antennas from 8 to 64, it is possible to reduce the filter complexity in each transmit filter by 55%.

Comparing results for the transmit filter to the ones for the IFFT, one difference is that the number of required bits in the transmit filter is lower compared to the IFFT. For the case with 64 antennas the corner points with the same complexity are (4, 6) and (5, 5) for filter coefficients and input signal respectively. This gives, for the latter case, an internal representation in the transmit filter of 5+5+3=13 bits at most. Comparing this to the IFFT, where 8 bits for the twiddle factor and 15 bits for the signal (same as the internal representation) was required, the conclusion is that, when increasing the number of antennas, it is possible to push the low accuracy processing in each transmit filter further than in each IFFT.

With the observation that the low accuracy processing in the transmit filter could be pushed more than the IFFT, further assessment was made for the transmit filter. Fig. 4 shows the quantization when the number of antennas is fixed to 64 and the number of users is varying between 4 and 32. With 64 antennas and these various system loads the outcome is similar, although a few more bits are required for higher system loads.

Further investigations included the number of required bits for the input signal and filter coefficients with a fixed combination of 64 antennas and 4 users, but with varying target BER. In Fig. 5, it can be seen that for three different target BERs, 10^{-2}, 10^{-3} and 10^{-5}, there are only small differences. When comparing these curves, it can be concluded that in order to achieve the two better BERs, only one or two more bits are needed, in comparison to the worst BER.

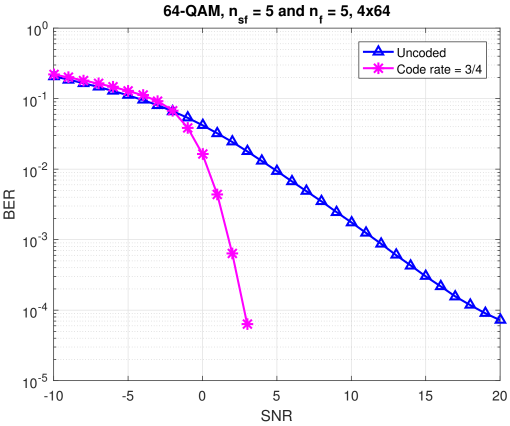

Fig. 6. Uncoded and LDPC coded performance. The corner point simulated with an uncoded 64-QAM signal and with LDPC coding, code rate 3/4, respectively. There are 64 antennas and 4 users and the corner point quantization values are $n_{\mathrm{sf}} = 5$ and $n_{\mathrm{f}} = 5$ bits.

Further on, one corner point from the 4×64 case was chosen for additional evaluation of performance and therefore simulated for a range of SNR values. The chosen corner point was 5 bits representing the input signal and 5 bits representing the filter coefficients. The results of this performance evaluation, as seen in Fig. 6, is starting to show an error floor for higher SNR values. Usually in communication systems, channel coding is applied and the same corner point was therefore simulated with LDPC coding, using block size 672 bits and code rate 3/4. Fig. 6 also shows that when adding LDPC coding the required SNR can be significantly lower while achieving the same BERs, at the cost of higher complexity in the per-user processing, which is scaling with K.

Evaluating the potential performance loss caused by quantization in the transmit filter gives the results shown in Fig. 7, which compares the case using floating-point to different fixed-

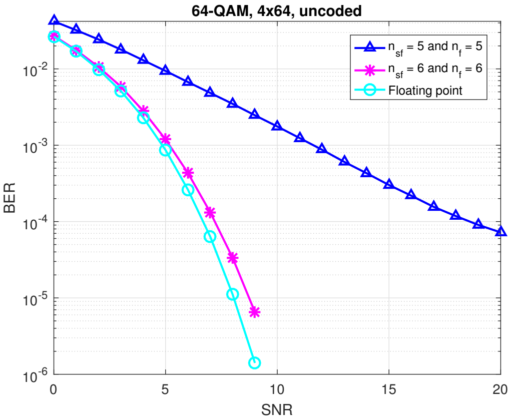

Fig. 7. Comparison between the floating point case and two different quantization combinations, the corner point $n_{\text{sf}} = 5$ and $n_{\text{f}} = 5$ bits and the case with $n_{\text{sf}} = 6$ and $n_{\text{f}} = 6$ bits. There are 64 antennas and 4 users and the signal is uncoded with the constellation 64-QAM.

point combinations. Both the case with the corner point and the case when the number of bits used to represent the input signal and filter coefficients are both increased from 5 to 6 bits, in order to also have a performance closer to the floating-point case to compare with, is visualized. At a BER of 10^{-3} there is an SNR degradation of less than 0.5 dB, for the latter fixed-point combination, compared to when using floating-point. Even for these numbers with quite few bits, great performance is achievable but with the possibility to reduce the complexity per antenna due to the larger total amount of antennas. Also, there are possibilities for even greater complexity savings if optimizing the IFFT and transmit filter internally.

V. Conclusion

This paper focused on simplified DSP for MaMIMO systems, exploiting the large number of antenna signals to reduce the processing accuracy in each antenna and hence, power consumption of the processing. We investigated the performance and calculated complexity savings when coarsely quantizing the IFFTs and transmit filters, in order to decrease the complexity in the per-antenna functions and thereby, the overall DSP complexity. The results show that it is possible to push the transmit filter more than the IFFT, requiring only 6 bits for the input signal and 6 bits for the filter coefficients in order to get a performance close to the floating-point case. For a BER of 10^{-3}, this resulted in an SNR degradation of less than 0.5 dB, despite using a sensitive 64-QAM constellation. A complexity analysis was also made showing that when increasing the number of antennas from 8 to 32 or above, complexity savings of 29% were possible for each IFFT. Similarly, an increment of the number of antennas from 8 to 64, resulted in a possibility to reduce the complexity in each transmit filter by 55%. If optimizing the IFFT and transmit filter internally, even greater complexity savings would be possible. These results show that, when increasing the number of base station antennas, it is possible to reduce the complexity per antenna by lowering the accuracy, having a significant impact on power consumption in MaMIMO systems.

Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 619086 (MAMMOET). We would also like to thank the Erasmus program for the scholarship received while conducting the research and S. Pollin's lab for taking care of us during our stay in Leuven.

References

[1] E. G. Larsson, F. Tufvesson, O. Edfors, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[2] A. Nordrum, "5G researchers set new world record for spectrum efficiency," IEEE Spectrum, [Online]. Available: http://spectrum.ieee.org/tech-talk/telecom/wireless/5g-researchers-achieve-new-spectrum-efficiency-record, May 2016.

[3] C. Desset, B. Debaillie, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, May 2015.

[4] Y. Huang, C. Desset, A. Bourdoux, W. Dehaene, and L. Van der Perre, "Massive MIMO processing at the semiconductors edge: Exploiting the system and circuit margins for power savings," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017.

[5] C. Desset and L. Van der Perre, "Validation of low-accuracy quantization in massive MIMO and constellation EVM analysis," in *2015 European Conference on Networks and Communications (EuCNC)*, Jun. 2015.

[6] E. Dahlman, S. Parkvall, J. Sköld, and P. Beming, *3G Evolution - HSPA and LTE for Mobile Broadband*, 2nd ed. Elsevier Ltd., 2008.

[7] S. Stefania, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From theory to practice*, 2nd ed. John Wiley & Sons Ltd., 2011.

[8] R. E. Lander and M. J. Fischer, "Parallel prefix computation," *Journal of the Association for Computing Machinery*, vol. 27, no. 4, pp. 831–838, Oct. 1980.

[9] C. R. Baugh and B. A. Wooley, "A two's complement parallel array multiplication algorithm," *IEEE Trans. Comput.*, vol. C-22, no. 12, pp. 1045–1047, Dec. 1973.

[10] C. Desset, B. Debaillie, and F. Louagie, "Modeling the hardware power consumption of large scale antenna systems," in *2014 IEEE Online Conference on Green Communications (OnlineGreencomm)*, Nov. 2014.