High Throughput Constant Envelope Pre-coder for Massive MIMO Systems

Prabhu, Hemanth; Rusek, Fredrik; Rodrigues, Joachim; Edfors, Ove

Published in:
2015 IEEE International Symposium on Circuits and Systems (ISCAS)

DOI:
10.1109/ISCAS.2015.7168930

2015

High Throughput Constant Envelope Pre-coder for Massive MIMO Systems

Hemanth Prabhu, Fredrik Rusek, Joachim Neves Rodrigues, and Ove Edfors
Department of Electrical and Information Technology, Lund University, Sweden
{hemanth.prabhu, fredrik.rusek, joachim.rodrigues, ove.edfors}@eit.lth.se

***Abstract*—This study describes a high throughput constant envelope (CE) pre-coder for Massive MIMO systems. A large number of antennas (M), in the order of 100s, serve a relatively small number of users (K) simultaneously. The stringent amplitude constraint (only phase changes) in the CE scheme is motivated by the use of highly power-efficient non-linear RF power amplifiers. We propose a scheme that computes the CE signals to be transmitted based on box-constrained regression (coordinate-descent), with an $\mathcal{O}(2MK)$ complexity per iteration per user symbol. A highly scalable systolic architecture is implemented, where M Processing Elements (PEs) perform the pre-coding for a system with up to $K=16$ users. This systolic architecture results in a very high throughput of 500 Msamples/sec (at 500 MHz clock rate) with a gate count of 14 K per PE in 65 nm technology.**

I. Introduction

Massive Multiple-Input Multiple-Output (MIMO) is a promising technology to meet the ever increasing demands for high throughput and quality-of-service for next generation wireless cellullar systems. In massive MIMO systems, the Base Stations (BSs) are equipped with a very large number of antennas compared to previously considered systems, serving a relatively low number of users simultaneously in the same frequency and time resource. The advantages of large antenna arrays at the BS are well known in literature [1], [2].

Although scaling up MIMO provides impressive theoretical gains, it also leads to many practical implementation challenges. One of the critical challenges is to reduce power consumption, which has become ever-so important considering the environmental impact cellular BS has around the world. Power Amplifiers (PAs) can consume up to 40% of the total base station power [3]. This is mainly due to the high linearity constraints on the PA, which translates into an inefficient operation. Non-linear PAs are highly efficient and employing them requires the Peak-to-Average Ratio (PAR) of the transmitted signal to be low. In [4] (antenna reservation), [5] (convex optimization), different approaches to reduce the PAR for Orthogonal Frequency-Division Multiplexing (OFDM) based Zero-Forcing (ZF) pre-coding massive MIMO schemes are described. Although, these schemes reduce the PAR by 3-10 dB, they still require linear PAs operating at a back-off, since PAR is not 0 dB.

To employ a highly efficient non-linear PA, a very strict constraint on the amplitude (constant) of the transmit signal is enforced, resulting in a 0 dB PAR. This strict amplitude constraint based downlink transmission scheme is known as constant envelope (CE). The CE with its strict constraints, seems sub-optimal compared to other precoding schemes like ZF. However, considering the large degree-of-freedom in massive MIMO, the performance of CE still achieves high sum-rates [6].

Apart from PAs, another evident factor for power consumption in the BS is the signal processing. Massive MIMO, inherently demands more signal processing due to the large number of antennas. The signal processing power consumption associated with obtaining CE transmission has to be lower than the corresponding power savings in the PAs. Furthermore, the downlink signal processing (and data transmission) need to be performed faster than the channel coherence time, hence requiring high throughput and low latency signal processing modules. This requires highly optimized hardware implementation of the CE pre-coder.

In this study we propose a CE pre-coder algorithm based on box-constrained regression (coordinate-descent), which apart from low hardware complexity provides a very good convergence. In case of massive MIMO, where the number of antennas at BS, M is very large (say in 100s), few iterations (less than 3) provides very good performance in-terms of total capacity (or Signal-to-interference-plus-noise ratio (SINR)). Furthermore, the proposed algorithm is hardware friendly, and is implemented as a systolic array. The systolic array is very desirable in hardware because of its simplicity, high throughput and scalability. Each Processing Element (PE) has a gate count of 14 K and can handle 1 user-symbol per clock cycle in a streaming data-flow architecture. These advantages in hardware was one of the main reasons for choosing the algorithm, which is more of bottom-up design flow, which are described in the sections following the system model.

II. System Model

The system model and the pre-coding (downlink) in this section is in line with the corresponding description in [2], where the channel gain between the m-th BS antenna and the k-th user is denoted by h_{km}. The channel matrix to all users is denoted as $\boldsymbol{H} \in \mathbb{C}^{K\times M}$, where h_{km} is the (k,m)-th entry. Let $\boldsymbol{x} = [x_1, x_2, ..., x_M]^{\mathrm{T}}$ denote the transmitted vector from the M BS antennas, which is normalized to satisfy $\mathbb{E}[\boldsymbol{x}^{\mathrm{H}}\boldsymbol{x}] = 1$, and $()^{\mathrm{H}}$ is the Hermitian transpose. The overall symbol vector received by the K autonomous users is

$$\boldsymbol{y} = \sqrt{\frac{P_{\mathrm{T}}}{M}}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{w}\,, \tag{1}$$

where P_{T} is the total transmit power, and $\boldsymbol{w}$ is a $K\times 1$ vector i.i.d complex Gaussian variables with variance σ^2.

A. Pre-coding in massive MIMO

To fully exploit a large antenna array, the user symbols/information at the BS need to be translated or mapped to correct signals in the antennas. Thus, each user receives the information with low (zero) interference from other users. This mapping of weights or pre-coding is expressed as

$$\boldsymbol{x} = \boldsymbol{F}\boldsymbol{s}\,, \tag{2}$$

where $\boldsymbol{s}$ is a $K \times 1$ vector containing the symbols intended for the K users, and $\boldsymbol{F}$ is an $M \times K$ pre-coding matrix which maps the user symbols ($\boldsymbol{s}$) to antenna signals ($\boldsymbol{x}$). A well known linear pre-coding scheme in massive MIMO is ZF, with $\boldsymbol{F}_{\text{ZF}} \propto \boldsymbol{H}^{\text{H}}(\boldsymbol{H}\boldsymbol{H}^{\text{H}})^{-1}$ [7]. This is basically a constrained least-squares solution for an under-determined system, *i.e.* ZF cancels (zeros) all inter-user interference with least transmit energy (min $||\boldsymbol{x}||_2$, subject to $\boldsymbol{s} = \boldsymbol{H}\boldsymbol{x}$).

B. Constant envelope pre-coding

The constant envelope pre-coding is similar to ZF, *i.e.* inter-user interference is suppressed, but with a strict constraint on the amplitude. The optimization problem for the CE pre-coder can be mathematically formulated as

$$\begin{aligned} &\text{maximize} \quad \alpha \\ &\text{subject to} \quad \alpha\boldsymbol{s} = \boldsymbol{H}\boldsymbol{x} \\ &\qquad\qquad |x_m|^2 = 1, \text{for } m = 1, \cdots M, \end{aligned} \tag{3}$$

where α is scaling factor which improves the transmission power-efficiency by utilizing the array gain more effectively.

CE can also be treated as a phase only (fixed amplitude, 0 dB PAR) transmission, with received signal at the k-th user as

$$y_k = \sqrt{\frac{P_{\text{T}}}{M}} \sum_{m=1}^{M} h_{km} e^{j\phi_m} + w_k. \tag{4}$$

The solution to (3) is not trivial, and in this paper we try to simplify it in two steps, first we compute an appropriate value of α and then we minimize the inter-user interference. In the next section we will look into these simplifications and also discuss the performance and complexity of this approach.

III. Proposed CE pre-coder

Increasing α increases the signal strength (*i.e.* increase SINR), but a too high α hinders the ability to cancel interference (hence decrease SINR). Finding the optimal scaling factor α is a non-convex optimization problem, and is also dependent on $\boldsymbol{s}$. Furthermore, frequent changes in α makes it hard for the users (receivers) to keep a track or to predict the scaling factor for detection.

Approximation for α

Due to the aforementioned factors, an approximation of α as a long-term constant (varying slowly over multiple channel estimations) is preferred. A low complexity approximation to find the scaling factor is

$$\alpha_{\text{Tr}} = \sqrt{\frac{\text{Tr}(\boldsymbol{H}\boldsymbol{H}^{\text{H}})}{K}}, \tag{5}$$

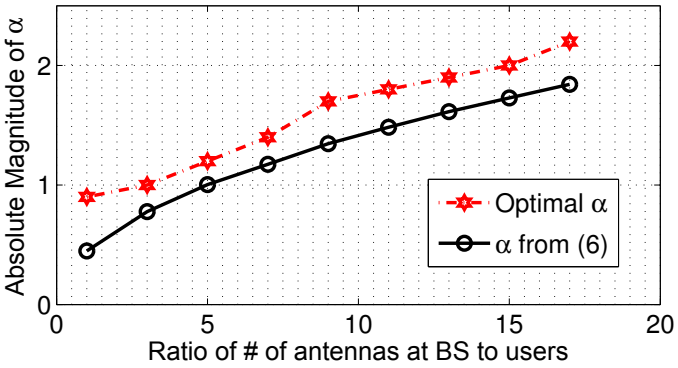

Fig. 1. Optimal and approximated scaling factor (α) for a $K = 10$ massive MIMO system using 4-QAM modulation. The optimal-α is determined by sweeping α over a range of values and solving (3).

where Tr is the trace of a matrix. Fig. 1 shows that the low complexity approximation is close to an optimal α, and the computation is performed as part of the CE pre-coder with very limited hardware overhead (initialization step in Algorithm 1). By computing α from (5), we can simplify (3) as

$$\begin{aligned} &\underset{\boldsymbol{x}}{\text{minimize}} \quad ||\alpha_{\text{Tr}}\,\boldsymbol{s} - \boldsymbol{H}\boldsymbol{x}||_2 \\ &\text{subject to} \quad |x_m|^2 = 1, \text{where } m = 1, \cdots M. \end{aligned} \tag{6}$$

The solution of (6) has multiple local-minima, but fortunately in case of massive MIMO, where $M \gg K$, even the local-minima tends to be close to optimal. To solve the CE pre-coder we use the coordinate-descent algorithm, which is analogous to gradient-descent, barring that the optimization is performed on one coordinate (variable) at a time. This is quite similar to [6], except that $\alpha = 1$ in the latter and also we avoid explicit computation of phases, due to high hardware cost.

After performing appropriate scaling of user symbols, the overall algorithm for solving (6) using coordinate-descent based CE pre-coder is shown in Alg. 1. To reduce the complexity and hardware cost, the previous results and residual vectors are used to avoid straight out matrix-vector computations for

Alg. 1 Proposed CE pre-coder.

▷ Initialization per channel realization	$\mathcal{O}$
for $m = 1 \rightarrow M$ **do**	
$\boldsymbol{a}_m = \frac{\boldsymbol{h}_m}{\|\|h_m\|\|_2^2}$	$4K + 1$
$\alpha = \alpha + \|\|h_m\|\|_2^2$	
end for	
▷ Initialization per user symbol	
$\boldsymbol{r} = \boldsymbol{s}$	
$\boldsymbol{x} = \boldsymbol{0}$	
▷ Main loop (P iterations over M antennas)	
for $l = 1 \rightarrow P$ **do**	
▷ Inner-Update loop	
for $m = 1 \rightarrow M$ **do**	
$update\_x = \boldsymbol{a}_m^H r$	$4K$
$x_{\text{temp}} = x_m + update\_x$	
▷ Truncation	
$x_{\text{trunc}} = \sqrt{\frac{1}{M}}\,\frac{x_{\text{temp}}}{\|x_{\text{temp}}\|}$	5
▷ Update residual vector, $\boldsymbol{r}$	
$\boldsymbol{r} = \boldsymbol{r} - \boldsymbol{h}_m(x_{\text{trunc}} - x_m)$	$4K$
$x_m = x_{\text{trunc}}$	
end for	
end for	

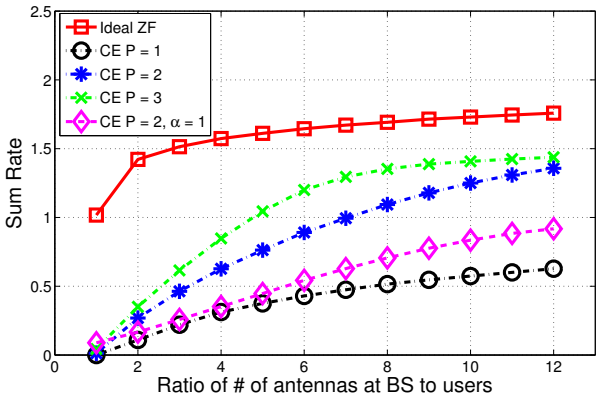

Fig. 2. Per-user ergodic sum-rate for CE pre-coder with respect to ideal ZF pre-coder for an i.i.d channel with $\sigma^2 = 0.01$.

each iterations. This results in an $\mathcal{O}((9K+5)MP)$ real valued multiplications per user symbol $\boldsymbol{s}$.

Performance of CE pre-coder

Fig. 2 shows the per-user sum-rate for the CE pre-coder with different number of iterations (P). When the scaling factor $\alpha = 1$, the CE pre-coding is same as in [6], and has a lower performance compared to our proposed pre-coding. The proposed CE has a lower sum-rate than that of ZF for smaller ratios, due to the inability to suppress Multi-user interference (MUI), and still satisfy the strict amplitude constraint. However, increasing number of antennas provides a large degree-of-freedom (massive MIMO), which improves the performance of CE. In particular for high ratios (>10) only $P = 2$ iterations are needed to achieve performance close to the ZF.

Extension to frequency selective channels

The CE pre-coding in (6), is extended to frequency selective channels by solving a joint optimization problem which spans over a block of transmission or time instances [6]. Another more hardware friendly approach is to look only at the current transmit signals, and mitigate the Inter-Symbol Interference (ISI) due to previous transmitted signals as

$$\boldsymbol{u}_t = \sqrt{\frac{P_T}{M}} \sum_{l=0}^{L-1} \boldsymbol{H}_{(l)} \boldsymbol{x}_{t-l}, \tag{7}$$

where $\boldsymbol{H}_{(l)}$ is the channel at the l-th previous time instance, $\boldsymbol{x}_{t-l}$ is the l-th previous symbol transmitted, and L is the sliding window length or memory of channel. In this study we only consider the hardware for solving (6), since (7) can be easily implemented using a matrix-vector multiplication unit. In the next section the hardware architecture of the CE pre-coder is described, followed by discussion on implementation results.

IV. Hardware Architecture

The CE pre-coder in Alg. 1 is mapped as a very high throughput and pipelined systolic array. This mapping exploits the fact that the optimization is performed on one variable at a time, resulting in vector-dot products, rather than matrix-vector products. The vector-dot product and (update residual) vector-scalar operation is streamlined, resulting in a very efficient processing element.

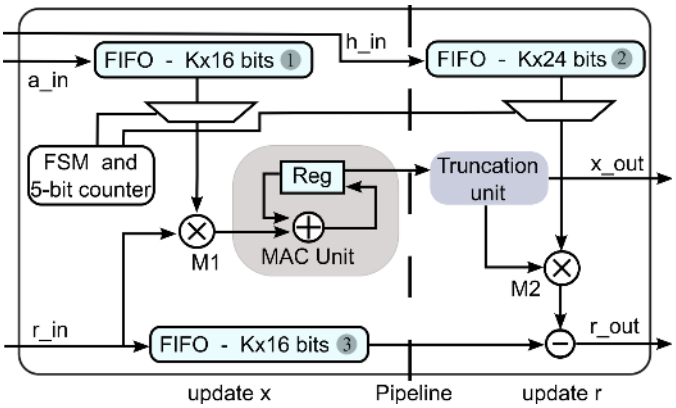

Fig. 3. Processing element for CE pre-coder.

Processing element: In Fig. 3, the PE for the CE pre-coder is shown. The PE basically implements the Inner-update loop (Alg. 1), in a streamlined (serial) architecture. The operations start with a *load-phase*, where the channel column vector ($\boldsymbol{h}_m$) is shifted serially into the First-In-First-Outs (FIFOs)-1,2. After the *load-phase*, the PE enters *wait-phase* and the Multiply-Accumulate (MAC) unit is cleared. A valid data (r_in_valid) triggers the MAC units to perform vector-dot product (*compute-state*) between the serial data (r_in) and normalized column vector (a_in). The serial data is also pushed into the FIFO-3 simultaneously. After the completion of the vector-dot product, the result is truncated as mentioned in Alg.1. The PE is pipelined such that while in the *compute-state* the computation of residual vector $\boldsymbol{r}$ is performed in parallel on the previous data. The utilization of the PE is 100%, with both multipliers M1 and M2 are continuously active, since M1 operates on new data vector and M2 operates simultaneously on the old vector.

Truncation unit: The truncation unit adjusts the amplitude of the result from MAC unit to a fixed value. This is performed by dividing by the absolute value ($|x_{\text{temp}}|$) and scaling appropriately based on the number of antennas. In hardware the truncation unit can be implemented as an independent hardware with one real valued division unit and two real valued multipliers. However, the utilization of this unit will be very low, since it will be used only once every K cycles *i.e.* once after *compute-state*. To avoid this low utilization, the multipliers M1 and M2 are re-used to compute the absolute value and perform the appropriate truncations. For re-use a *trucate-state* is inserted in the state flow after *compute-state*. In this state the multipliers M1 and M2 are multiplexed to perform the absolute value computation and scaling respectively as shown in Fig. 4. Overall this re-use costs 2 additional cycles, but the expensive truncation logic is performed with almost no additional hardware.

Square root and division unit: In this study the user symbols are 8-bits wide (modulation of up to 256-QAM), and this in-turn means that the square-root and division unit is accurately

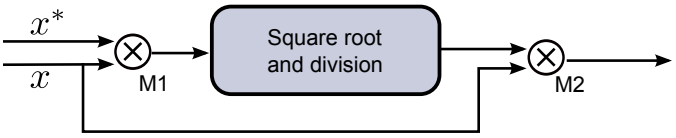

Fig. 4. Performing truncation by re-using multipliers.

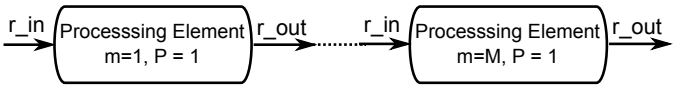

Fig. 5. Uni-directional systolic array of the CE pre-coder with M antennas.

performed without multipliers *e.g.* by a simple Look-Up-Table (LUT). The number of elements in the LUT is optimized by dividing the total range based on statistics and the fact that smaller numbers need to be more accurately represented when performing division. In the next section the aforementioned optimized PE is used to implement a high throughput systolic CE pre-coder, and the results of the latter are discussed.

V. Implementation and Result

The CE pre-coder is implemented by connecting the PE in series, forming a systolic array as shown in Fig. 5, with first PE(m=1) performing computations for the first antenna. The systolic array is very generic and is easily generated based on the requirements, *i.e.* number of antennas (M), and performance (iterations-P).

Hardware cost: The proposed architecture is implemented in RTL, as well as synthesized and routed in 65 nm CMOS. The gate-count break-up of the PE is shown in the pie-chart in Fig. 6. The multipliers (M1 and M2) take the majority of combinational logic, along with the FSM and counters. The non-combinational logic mainly consists of 3 FIFOs, with FIFO-2 being the biggest. The other non-combinational logic mainly consists of pipeline registers required to perform the streaming operations. A major concern with this architecture is the number of registers in the design, which would increase the total power consumption. To combat this a clock-gating scheme is used to reduce the power consumption.

Power reduction and clock-gating: The FIFOs that hold the channel column vector are active only in the *load-phase*, and thus the clocks to FIFOs are gated in the other states. The gating is performed by using a latch-based clock-gating circuit [8]. The power reduction due to clock-gating technique is around 20%, with a negligible area overhead.

Latency and throughput: The latency of the PE depends on the number of users K (Table I), since the operations in the PE are performed serially. However, systolic array provides a high throughput with a maximum clock rate of 500 MHz. The high clock-rate is mainly due to the critical path being isolated inside each PE. Furthermore, the PE can handle 1 user-symbol every cycle in a fully unrolled implementation. As an example, if we consider a massive MIMO system with $K = 10$ (upto

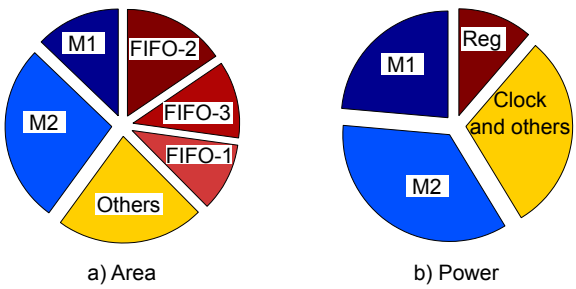

Fig. 6. Area and power breakdown of a processing element with 12-bit internal precision.

TABLE I
Hardware results in 65 nm CMOS technology of CE pre-coder.

	Per Processing Element	For $K = 10$, $M = 100$, $P = 2$
Area [mm^2]$^{\#}$	.030	6.02
Gate Count [10^3]	14.1	-
Max. Clock [MHz]	500	500
Latency [cycles]	$K + 2$	2400
Latency [μsec]	-	4.8
Throughput [G samples/sec]	$\frac{0.5}{K}$	5
Information rate [Gbps]	-	10
Power [mW]*	3.96	792

$^{\#}$ Includes post layout clock tree synthesis
* Power numbers are extracted by performing post-layout simulations with back annotated timing and toggle information

16), $M = 100$, $P = 2$, with 4-QAM (upto 256) modulation as in Table I, the implementation supports total user information data-rate of 10 Gbps with a power consumption of 792 mW.

VI. Conclusion

Constant envelope pre-coding has been proposed for massive MIMO for a while, however, to the authors best knowledge, there has been no studies on the hardware cost and power consumption. In this study we have proposed a CE pre-coder with a scaling factor, which improves the system performance by utilizing the array gain more efficiently. The hardware architecture is systolic, which can be easily scaled and supports very high throughputs. Various procedures such as hardware-reuse, pipelining, clock-gating are employed to optimize hardware. This results in a CE pre-coder with a gate count of 14 K-Gates and power consumption of 3.96 mW per antenna per user per iteration. Taking this into consideration it is clear that a more detailed study is needed, with a total power-consumption analysis including power amplifiers.

Acknowledgement

We thank Lund University, DISTRANT funded by SSF for providing the opportunity to work on this project. Also thank Oskar Andersson for the help in ASIC back-end flow.

References

[1] T. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *Wireless Communications, IEEE Transactions on*, pp. 3590 –3600, Nov. 2010.

[2] F. Rusek, D. Persson, B. Lau, E. Larsson, T. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up MIMO: Opportunities and challenges with very large arrays," *Signal Processing Magazine, IEEE*, vol. 30, pp. 40–60, Jan. 2013.

[3] V. Mancuso and S. Alouf, "Reducing costs and pollution in cellular networks," *Communications Magazine, IEEE*, vol. 49, no. 8, pp. 63–71, Aug 2011.

[4] H. Prabhu, O. Edfors, J. Rodrigues, L. Liu, and F. Rusek, "A low-complex peak-to-average power reduction scheme for OFDM based massive MIMO systems," in *ISCCSP*, May 2014, pp. 114–117.

[5] C. Studer and E. Larsson, "PAR-aware large-scale multi-user MIMO-OFDM downlink," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 2, pp. 303–313, 2013.

[6] S. Mohammed and E. Larsson, "Constant-envelope multi-user precoding for frequency-selective massive MIMO systems," *Wireless Communications Letters, IEEE*, vol. 2, no. 5, pp. 547–550, Oct 2013.

[7] H. Prabhu, J. Rodrigues, O. Edfors, and F. Rusek, "Approximative matrix inverse computations for very-large MIMO and applications to linear pre-coding systems," in *IEEE WCNC*, Apr. 2013.

[8] J. Rabaey, *Low Power Design Essentials*, 1st ed. Springer Publishing Company, 2009.