CELLA: FPGA Based Candidate Execution with Low Latency Approach for Soft MIMO Detector

Erulappan Sakthivel1, Kasinathan Pounraj1, Veluchamy Malathi2, Muruganantham Arunraja1, Govindaraj Perumalvignnesh3

1Electrical and Electronics Engineering, SBM College of Engineering, Dindigul, India
2Electrical and Electronics Engineering, Anna University Regional Campus, Madurai, India
3Electrical and Electronics Engineering, The Silicon Harvest, Madurai, India
Email: vlsisakthivel@gmail.com, sbmhodeee@gmail.com, vmeee@autmdu.ac.in, rsearcharunraja@gmail.com, vignesh@siliconharvest.net

Received 16 March 2016; accepted 5 April 2016; published 20 June 2016

Abstract

This paper describes the design and Field Programmable Gate Array (FPGA) based 4 × 4 breadth heuristic Multiple-Input—Multiple-Output (MIMO) decoder using 16 and 64 Quadrature Amplitude Modulation (QAM) schemes. The intention of this work is to observe the performance of Candidate Execution with Low Latency Approach for soft MIMO detector in FPGA (CELLA). The Smart Ordering and Candidate Adding (SOCA), Parallel Candidate Adding (PCA) and Backward Candidate Adding (BCA) give better performance in terms of Bit Error Rate (BER) or chip level service. In order to attain both BER and FPGA level performance in a single system, CELLA is developed in this work. Simulation and experimental results demonstrate the effectiveness of the proposed work under the system 4 × 4 MIMO-OFDM employing 16 QAM and 64 QAM. The proposed experiment is implemented in Xilinx Virtex 5 C5VSX240T. The performance results, in terms of FPGA level 76% slice reduction, 58.76% throughput improvement, 75% power reduction and 87% latency reduction, are achieved. The BER performance is observed and compared with the conventional algorithms. Thus, the proposed work achieves better outcome than the conventional work.

Keywords

Multiple-Input—Multiple-Output (MIMO), Field Programmable Gate Array, Parallel Candidate Adding

http://dx.doi.org/10.4236/cs.2016.78152
1. Introduction

In today’s data-rich world, MIMO has become an energetic element of wireless communication standard for high data rate communications. MIMO is a method for multiplying the capacity using multiple transmit and receiving antennas to make use of multipath propagation. The method of incorporating turbo code in MIMO system is labeled as turbo coded MIMO system. Using the sphere decoder, a simple method is used to detect and decode linear space-time mapping with any channel code and it is called “soft” inputs and outputs [1]. To simplify the exponentially complex search problem in ML decoders for MIMO systems with greater modulation constellations, sphere decoders are accomplished by realizing near-ML performance with reasonable complexity. Moreover, the lattice decoders can be stretched to care soft decision outputs and, hence, be used in an iterative MIMO receiver. The lattice decoding algorithms have two kinds of application strategies. The first one is called Fincke—Pohst strategy and Schnorr—Euchner strategy. To avoid confusion, the lattice decoder using the Fincke—Pohst strategy is called a Sphere Decoder (SD), and the lattice decoder using the Schnorr—Euchner strategy is called SE. A List Sphere Decoding (LSD) algorithm can be used to effectively find the candidate list. The recognized list type of MIMO decoding algorithms is List Sphere Decoding (LSD) algorithm. The two major problems of the List Sphere Decoder (LSD) are variable complexity and the sequential nature of its tree search. To overcome these issues, a Fixed-complexity Sphere Decoder (FSD) is modified as List FSD (LFSD) [2], which performs iterative detection and decodes in turbo Multiple Input Multiple Output (MIMO) systems. This method obtains a list of candidates for calculating likelihood information about the transmitted bits required by the outer decoder [2]. Because of lesser amount of information about expanded candidates, Soft-FSD (SFSD) can be used in turbo-MIMO systems to replace extrinsic soft-information to the outer decoder. It is used to provide the exact performance for the LSD with constant throughput and fixed complexity, by making the algorithm more suitable for hardware implementation [2] [3]. The negative aspects of the soft decoder are variable throughput, noise level and the channel conditions.

In [4] FPGA implementation of a new detection algorithm based on SD is introduced for MIMO systems. This paper algorithm overcomes the main drawback of the SD: its variable throughput, depending on the noise level and the channel environments. The performance results of Barbero et al. show that the algorithm is extremely parallelizable and can be completely pipelined. In [5] PCA scheme is used for the soft-output sphere decoder to identify the partial Maximum A Posteriori Estimation (MAP) estimation based on parent nodes in the place of child nodes. It provides low decoding latency and it is suitable for fully pipelined hardware implementation. In order to achieve optimal error rate performance with minimal power consumption, an adaptive switching detection algorithm is designed to incorporate multiple thresholds within the detector, which is implemented in XilinxVirtex-5 simulation hardware [6].

Michael et al. have introduced a low-cost parallel programmable co-processor, which can achieve high throughput in ASIC/FPGA designs [7]. In [8] a novel MIMO detection algorithm is introduced which is also called Modified Fixed-Complexity Soft-Output (MFCSO) detection. This MFCSO used achieves better performance in terms of bandwidth and hardware implementation cost. The Maximum Likelihood Detector (MLD), QR decomposition with M-algorithm (QRM) and Fixed-complexity Sphere Decoding (FSD) [9] are used to calculate soft information of each coded bit. This approach can retrieve the coded bits and thus can recalculate the soft information with improved error performance under the system with 4 × 4 MIMO-OFDM employing 16 QAM and 64 QAM. Parallel bidirectional scheme [10] obtains a list of candidates to calculate likelihood information in parallel and each stage corresponds to one direction of the path selection process. Due to its parallel nature, it is well suited for hardware implementation, and it achieves better performance with lower complexity than a LFSD. For the data detection stage based on column-norm categorizing fully parallel GPU and multicore are used in [11]. These modules are developed to match the multicore architecture efficiently which has lesser computational time. It supports some of the configurations in the Long Term Evolution (LTE) standard [11].

The rest of this paper is organized as follows. Section 2 addresses the system model. Conventional detectors are discussed in Section 3. Proposed work and its module details are discussed in Section 4. The experimental results are presented in Section 5. Finally, the conclusion is presented in Section 6.

2. System Model

The system model specification is similar to that of Takuma et al. [9]. Here, in Low Density Parity Check (LDPC) coded MIMO system, \( N_t \) represents transmit antennas and \( N_r \) represents receive antennas (\( N_r \geq N_t \)).
These antennas are identically distributed. The binary information (input) is passed through LDPC encoder (by a rate \( \gamma \)) and an interleaver. The final results are mapped into symbols. These symbols are demultiplexed into \( N_t \) sub modules and passed through OFDM section. The vector consisting of the \( N_r \) receives symbols \((y_1, \cdots, y_{N_t})\) is given by,

\[ Y = Hx + z \]

where \( x = (x_1, \cdots, x_{N_t})^T \) is the transmit vector with normalized transmit power \( E[|x|^2] = 1/N_t \); \( z = (z_1, \cdots, z_{N_r})^T \) is the vector of independent and identically distributed; complex Additive White Gaussian Noise (AWGN) with variance of \( \sigma^2 = N_0 \), and \( H \) is the \( N_r \times N_t \) Rayleigh fading channel matrix, where its element \( h_{ij} \) is the complex transfer function from transmitter \( j \) to receiver \( i \) with \( E[|h_{ij}|^2] = 1 \). The receiver section consists of MIMO detector, deinterleaver and LDPC decoder. The calculation of \textit{A Posteriori Probability} (APP) for each of the coded bits \( b_{j,k} \) is taken in the form of Log Likelihood Ratio (LLR). From [9], the LLR equation is refined as per the proposed system.

\[ L(b_{j,k}) = \min_{\omega < \omega_j} \frac{\|y - Hx\|^2}{N_0} - \min_{\omega < \omega_j} \frac{\|y - Hx\|^2}{N_0} \]

The list size upsets the complexity of the system, since the computational load increases linearly with the list size. To avoid this issue, the insertion of the list is removed from [9] and it is represented in Equation (2).

### 3. Conventional Detectors

The conventional algorithm for visited nodes and the list size are listed in Table 1 and Table 2. The maximum level of complexity is MLD, where visited nodes are 69,904 and the list size is 65,536 in 16 QAM.

The above values are reduced by introducing QRM concept. The visited node is 784 and the list size is 16 in 16 QAM whereas in the 64 QAM, the visited node is 5824 and the list size is 30. After that, LFSD, PCA and BCA combination of QRM-BCA and PCA-BCA visited node and list node are listed for both 16 QAM and 64 QAM. On comparing these concepts, the proposed methodology describes lesser consumption of visited nodes in the order of 34 and there is a decrease in the value of list size in the order of 16 in the 16 QAM. On the other hand, if we are considering 64 QAM case, visited nodes and list size do not give a drastic change like the above-described methodologies. This minimizes the size of Look Up Table (LUT). The decrease in the LUT simultaneously reduces the use of nodes and therefore, the slices are partially used in a considerable percentage.

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>VISITED NODES</th>
<th>LIST SIZE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLD [9]</td>
<td>( M(M^N - 1) )</td>
<td>( M^N )</td>
</tr>
<tr>
<td>QRM [9]</td>
<td>( M(M^N - M + 1) )</td>
<td>( M )</td>
</tr>
<tr>
<td>LFSD [9]</td>
<td>( \sum_{j=1}^{N_r} \prod_{i=1}^{N_t} g_{mis} )</td>
<td>( \prod_{j=1}^{N_r} g_{mis} )</td>
</tr>
<tr>
<td>PCA [9]</td>
<td>( N_r M + \frac{1}{2} \frac{\omega}{N_r} (N_r - 1) )</td>
<td>( M + \frac{\omega}{N_r} (N_r - 1) )</td>
</tr>
<tr>
<td>BCA [9]</td>
<td>( N_r M + \frac{1}{2} \frac{\omega}{N_r} (N_r - 1) )</td>
<td>( M + \frac{\omega}{N_r} (N_r - 1) )</td>
</tr>
<tr>
<td>QRM-BCA [9]</td>
<td>( M(M^N - M + 1) + \frac{1}{2} \frac{\omega}{N_r} (N_r - 1) )</td>
<td>( M + \frac{\omega}{N_r} (N_r - 1) )</td>
</tr>
<tr>
<td>PCA-BCA [9]</td>
<td>( N_r M + \frac{\omega}{N_r} (N_r - 1) )</td>
<td>( M + \frac{\omega}{N_r} (N_r - 1) )</td>
</tr>
<tr>
<td>CELLA</td>
<td>( \frac{1}{2} N_r M + \frac{1}{2} \frac{\omega}{N_r} (N_r - 1) )</td>
<td>( M + \frac{1}{\omega} N_r )</td>
</tr>
</tbody>
</table>
Table 2. The number of visited nodes and list size with 16-QAM and 64-QAM.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>16-QAM</th>
<th></th>
<th>64-QAM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Visited nodes</td>
<td>List size</td>
<td>Visited nodes</td>
<td>List size</td>
</tr>
<tr>
<td>MLD [9]</td>
<td>69,904</td>
<td>65,536</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td>QRM [9]</td>
<td>784</td>
<td>16</td>
<td>5824</td>
<td>30</td>
</tr>
<tr>
<td>LFSD [9]</td>
<td>176</td>
<td>64</td>
<td>704</td>
<td>64</td>
</tr>
<tr>
<td>PCA [9]</td>
<td>88</td>
<td>28</td>
<td>292</td>
<td>82</td>
</tr>
<tr>
<td>BCA [9]</td>
<td>88</td>
<td>28</td>
<td>292</td>
<td>82</td>
</tr>
<tr>
<td>QRM-BCA [9]</td>
<td>808</td>
<td>28</td>
<td>5860</td>
<td>48</td>
</tr>
<tr>
<td>PCA-BCA [9]</td>
<td>112</td>
<td>40</td>
<td>328</td>
<td>100</td>
</tr>
<tr>
<td>CELLA</td>
<td>34</td>
<td>16</td>
<td>136</td>
<td>32</td>
</tr>
</tbody>
</table>

4. Proposed Work

SOCA scheme is used to identify the partial MAP using parent nodes [12]. Conventional detectors are complex in nature, because it has maximum number of child nodes to identify the partial MAP. In SOCA, PCA [5] and BCA, the maximum number of child and parent nodes are available. In conventional architecture, the activities of child and parent nodes differ from under working node, but the overall architecture is the same as that of the previous work. In Figure 1, PCA, SOCA and CELLA architecture layouts are indicated. In order to reduce the complexity and for further performance in CELLA, a child node and single PE are represented to identify the partial MAP.

4.1. Candidate Execution with Low Latency Approach (CELLA)

The counterhypotheses of conventional SOCA, PCA and BCA algorithms are added in parallel with other candidates. From the hardware aspects, Look-Up Table (LUT) is introduced for counterhypotheses, and which is used to reduce the latency [5]. The overall resources in SOCA, PCA and BCA slightly differ from one another. Takuma et al. [9] report the visited node and the list size of conventional work, and they are indicated in Table 1 and Table 2. The proposed work is compared with the conventional FPGA family. The major problem of the conventional research (FSD [4], SOCA [12], and PCA [5]) is the performance analysis, which is based on MATLAB Simulink environment model with Xilinx platform.

In order to convert the Simulink to Xilinx, conversion process utilizes more LUT and memory. Due to this complexity, the performance degradation occurs in the conventional architecture in FPGA level. To evade this issue, the conventional architecture is modeled using a VHDL implementation in Xilinx ISE Design Suite 12.1.

The pseudo-code of the CELLA scheme is given in Figure 2 (also refer pseudo-code of the PCA [5] in Figure 1 and pseudo-code of the SOCA [12] in Figure 2). The identification of the partial MAP node is performed in SOCA and PCA. The hardware complexity is slightly reduced in PCA than SOCA. For aggressive low complexity, the proposed concept is introduced. Here, the identification of the partial MAP node takes place within a single PE and MIN Search Unit (MSU). But, PCA and SOCA architectures have same number of extended child nodes with MSUs and parent node. In the proposed system, the identification of partial MAP node is performed by parent node without any additional child node (see Figure 2). This deals with the minimized consumption of the nodes with the above-described elements. It works in a conventional way but with the reduction in the consumption of nodes. This makes the best performance in the usage of nodes and when it is seen microscopically, the motto of the power reduction is achieved at the rate of 141 mw and 40 mw, which correspond to the 64 QAM and 16 QAM, respectively.

4.2. Hardware Architecture

Figure 2 shows the top-level architecture of the SOCA, PCA, and the proposed schemes in 4 × 4 16 QAM and 64 QAM systems. It is assumed that the Processing Elements (PEs) have been implemented to compute the required Euclidean Distance (ED) metrics. This experimental case study of the proposed system is carried out using Xilinx Virtex 5 XC5VSX240T. All 4 × 4 architectures are represented in a pipeline manner so that, it takes
only 4 cycles to obtain the entire list for every transmitted vector after a single latency, while the detector receives a throughput of 400 Mbps at a clock frequency of 100 MHz.

The functional modules of MSU are comparator and decision logic. The MSU is used to identify the local MAP node and the architecture is made so that, each PE processes 4 nodes sequentially. In PCA, 3 MSUs are employed to deal with the counterhypotheses list, which contains \((4 \times 3) + (4 \times 2) + 4 = 24\) nodes. To handle it, 6 numbers of PEs are required, whereas in the proposed scheme, only one MSU and single PE are required to deal with the counterhypotheses list, which contains \(4 \times 1 + 4 \times 1 + 4 = 12\) nodes. As a result of hardware elimination (Removal of 4 PEs and MSU) in the proposed work, the performance improvements in terms of throughput give a range of 970 Mbps at 16 QAM and 830 Mbps at 64 QAM.
The clock frequency obtained here is in the range of 150 MHz for two modulation architectures. They turn out with the power reduction in the range of 40 mw and 141 mw in 16 QAM and 64 QAM, respectively. The above-mentioned features are extended with the previously proposed metrics, and there is an improved standard in terms of reduction in latency, reduction in usage of slices, and improved data throughput. The most important term is the reduction in the usage of power that makes a standard among all the previous mentioned standards. When the throughput of CELLA is compared, the reduction in latency cycles also makes a greater contribution in the throughput for both 16 QAM and 64 QAM architectures. It is detailed in the FPGA level performance, and that is also independent in the amount of noise level.

5. Result and Discussion

The performance results are observed in simulation platform via MATLAB environment and FPGA platform via Xilinx ISE Design Suite 12.1. The Bit Error Rate (BER) results of SOCA, PCA, BCA, PCA + BCA, and CELLA are observed.

5.1. BER Performance

The BER performance has been obtained by computer simulation using 64-subcarrier OFDM per transmit antenna. Block Rayleigh fading is used for the channel model, where the ordering is required only once at the beginning of each received block. The rate of the LDPC code is chosen as $r_c = 1/2$, and the length of the code word is $N = 3072$ bits with maximum number of sum-product iteration $T = 40$, and $L_{\text{max}} = 8$. First, Figure 3 compares the BER performance of the SOCA, BCA, PCA, PCA-BCA, and the CELLA using sibling nodes. It is observed from the case of CELLA work, even if the sibling nodes are selected, there is a performance improvement in the proposed work and it is better than other conventional works. The sibling nodes indicate the difference of throughput $(E_b/N_0)$ of sibling nodes and the difference between the BER with the proposed and also the conventional algorithms.

5.2. FPGA Level Performance

In Table 3, the performance outcomes of the conventional, and the proposed works are compared. The various algorithms are analyzed. In optimized FSD-B algorithm [2], 16 QAM is used to obtain 24,815 slices, 39,800 FFs
and 31,759 LUT. The latency and throughput are 78 and 450 Mbps, respectively. Similarly, these results are obtained for algorithms such as K-Best SD [3], PIPSD [13] and FSD 2 [4], and which are implemented in Xilinx platform. The power values are not observed in these methods. Xiang et al. introduced PCA method [5] to attain better results compared to LFSD and SOCA works. The PCA results in 9471 slices, 22,814 FFs, 24,126 LUT, 174 DSP and 70 BRAM with latency cycles of 109 without power calculation. To overcome these problems and to reduce the area, a new method with various modulation levels (16 QAM and 64 QAM) is introduced in

Figure 3. BER performance using sibling node with the length of code word 3072 rate 1/2 LDPC code with 16-QAM, and 64-QAM in a 4 × 4 MIMO.
Table 3. Performance comparison of conventional and proposed work.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FPGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modulation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLICES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Available)</td>
<td>24,815</td>
<td>8778</td>
<td>NR</td>
<td>12,721</td>
<td>13,577</td>
<td>9570</td>
<td>2215</td>
</tr>
<tr>
<td>FFs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Available)</td>
<td>39,800</td>
<td>6274</td>
<td>NR</td>
<td>15,332</td>
<td>32,934</td>
<td>23,133</td>
<td>4918</td>
</tr>
<tr>
<td>LUT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Available)</td>
<td>31,759</td>
<td>13,417</td>
<td>NR</td>
<td>16,119</td>
<td>34,378</td>
<td>24,523</td>
<td>4982</td>
</tr>
<tr>
<td>DSP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Available)</td>
<td>48</td>
<td>136</td>
<td>NR</td>
<td>348</td>
<td>174</td>
<td>38</td>
<td>151</td>
</tr>
<tr>
<td>BRAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Available)</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td>82</td>
<td>99</td>
<td>70</td>
<td>18</td>
</tr>
<tr>
<td>Latency (cycles)</td>
<td>78</td>
<td>NR</td>
<td>NR</td>
<td>63</td>
<td>133</td>
<td>109</td>
<td>67</td>
</tr>
<tr>
<td>F_clk</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throughput at 20 db (Mbps)</td>
<td>150 MHz</td>
<td>NR</td>
<td>178 MHz</td>
<td>150 MHz</td>
<td>150 MHz</td>
<td>150 MHz</td>
<td></td>
</tr>
<tr>
<td>Power (mw)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CEILLA. They obtain 2215 slices, 4918 FFs, and 4982 LUT with latency of 14 cycles in 16 QAM. Similarly, these results are obtained in 64 QAM. The proposed method achieves the power value of 40 mw, 141 mw in 16 QAM and 64 QAM, respectively. This method is implemented in (FPGA platform) Xilinx Virtex5 XC5VSX240T. The experimental results of the proposed system are implemented in Xilinx Virtex5 XC5VSX240T platform. The performance analysis at FPGA level attains 76% slice reduction, 58.76% throughput improvement at 20 db, 75% power reduction and 87% latency reduction when compared to the conventional work. The BER performance is also computed for the proposed and conventional algorithms. It minimizes the noise level and improves the performance boost at 20 db/s.

6. Conclusion

In this work, a candidate execution with low latency approach has been introduced for a turbo coded MIMO system. The proposed system identifies partial MAP nodes during online process and adds counterhypothesis in parallel within the parent candidate, which indicates the use of child nodes. This is not included here to minimize the ED distance metrics. The experimental results of the proposed system are implemented in Xilinx Virtex5 XC5VSX240T platform. The performance analysis at FPGA level attains 76% slice reduction, 58.76% throughput improvement at 20 db, 75% power reduction and 87% latency reduction when compared to the conventional work. The BER performance is also computed for the proposed and conventional algorithms. It minimizes the noise level and improves the performance boost at 20 db/s. Thus, it proves the execution of the nodes in the confined region compared to other algorithms. Thus, CEILLA achieves the best performance than the conventional work.

References


