FPGA Design of Fixed-Complexity High-Throughput MIMO Detector based on QRDM Algorithm

Xiang Wu*, John S. Thompson
Institute for Digital Communications, University of Edinburgh,
King's Buildings, EH9 3JL, Edinburgh, Scotland
*Corresponding author's Email: x.wu@ed.ac.uk

***Abstract*—This paper presents a field-programmable gate array (FPGA) implementation of an unbiased minimum mean square error (MMSE) metric based QR-decomposition M (QRDM) algorithm for the multiple-input multiple-output (MIMO) systems. Two advanced techniques, namely the merge-sort (MS) based and winner path expansion (WPE) based sorting schemes have been implemented and validated on an FPGA platform for a 4x4 16-QAM MIMO system. The results show that the MS-QRDM is advantageous in the simplified control circuits and leads to less logic resource use, whereas the WPE-QRDM is able to achieve the minimum use of the computational units and results in fewer multipliers. Furthermore, it also shows that both schemes can support up to 1.6Gbps decoding throughput when they are implemented in a fully pipelined parallel architecture.**

Keywords: QRDM, MMSE, FPGA, merge-sort, winner path expansion

I. INTRODUCTION

During the last decade, spatial multiplexing multiple-input multiple-output (MIMO) systems have received considerable attention from both academia and industry as they offer significant increases in the system capacity [1]. However, one of the most challenging tasks for spatially multiplexed systems is to design the hardware efficient detection techniques at the receiver. Therefore, MIMO detectors based on different perspectives and methodologies have been proposed in the literature [2-6]. Among the proposed schemes, an unbiased minimum mean square error (MMSE) metric based QR-decomposition M (QRDM) algorithm that exploits the MMSE metric instead of the maximum likelihood (ML) metric shows superior detection performance and scalability over the alternatives [5]. In the case of the QRDM algorithm, the key technique lies in how to choose the M best candidates during the search procedure in an efficient way so that ultra-high decoding throughput can be obtained. It has been shown that the merge-sort (MS) algorithm is very attractive as it has a considerably reduced complexity compared to other parallel sorting algorithms [7]. Alternatively, a so-called winner path expansion (WPE) technique has been proposed recently in [6] for the K-BEST algorithm, which avoids the need for exhaustive sorting and achieves significant complexity reduction over the bubbling sort method. From two different perspectives, both the MS and WPE approaches have both pros and cons. However, no direct comparison of these two techniques can be found in the literature.

In this work, a field-programmable gate array (FPGA) rapid prototyping methodology has been adopted in order to investigate and identify the feasibility of MS based and WPE based schemes in the context of the QRDM algorithm. Both MS and WPE schemes have been implemented on a FPGA platform and the associated FPGA results show that the MS-QRDM is advantageous due to simplified control circuits and leads to less logic resource use, whereas the WPE-QRDM is able to achieve minimum use of the computational units and results in fewer multipliers. In terms of the throughput per unit power, they are roughly equivalent. Furthermore, the results show that both approaches can support over Gbps decoding throughput when they are implemented in a fully pipelined parallel architecture.

II. MIMO DETECTION

A. System Model

Consider an uncoded MIMO system with N_t transmit antennas and N_r receive antennas ($N_r \geq N_t$), the input-output relationship of this system is given by

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n} \quad (1)$$

where $\mathbf{H}$ is the complex-valued N_rxN_t quasi-static flat Rayleigh fading channel, $\mathbf{r}$ is the N_rx1 dimensional received vector, $\mathbf{s}$ is the N_tx1 dimensional transmitted vector whose elements are chosen independently from a complex-valued constellation alphabet Ω with size P and $\mathbf{n}$ is the N_rx1 dimensional circularly symmetric complex additive white Gaussian noise vector. Additionally, we assume that the channel matrix $\mathbf{H}$ is perfectly known by the detector and that $\mathbf{E}\{\mathbf{s}\mathbf{s}^H\} = \mathbf{I}_{N_t}$ and $\mathbf{E}\{\mathbf{n}\mathbf{n}^H\} = \sigma_n^2 \mathbf{I}_{N_r}$, where $\mathbf{E}\{\cdot\}$ denotes the expectation and $\mathbf{I}_{N_t}$ denotes the identity matrix of size N_t. Also, as in [5], complex vector and matrix notation is adopted here for conciseness, but the simulations are based on the equivalent real-valued model since the suboptimal detection schemes benefit from it [5]. Note that a 4×4 MIMO system ($N_t = N_r = N = 4$) with 16-QAM modulation is considered in this work, unless specified otherwise.

B. QRDM Based on Unbiased MMSE Metric

The main concepts of the unbiased MMSE metric based

CHINACOM 2010, August 25-27, Beijing, China

DOI 10.4108/chinacom.2010.69

QRDM algorithm are briefly revised here for the sake of completeness and the reader is referred to [5] for details. The basic motivation for using the unbiased MMSE metric is the fact that suboptimal detection of non-constant modulus alphabets, such as 16-/64-QAM, can be improved by imposing the constraint of an unbiased estimate. Note that the MMSE metric is ML optimal for constant modulus alphabets [5]. Hence, our QRDM implementation uses the unbiased MMSE metric in the sequel, unless specified otherwise.

In [5], it is shown that the QRDM approach to the detection problem in Eq. (1) is to find a candidate that minimizes the Euclidean distance (ED) metric around the received signal, which can be formulated in a general way as

$$\mu = \left\| \hat{\mathbf{D}}(\hat{\mathbf{r}} - \hat{\mathbf{L}}\mathbf{\Pi}\mathbf{s}) \right\|^2 \qquad (2)$$

The above expression incorporates three metrics, namely the ML, the MMSE, and the unbiased MMSE metrics. The differences between the metrics lie in the diagonal matrix $\hat{\mathbf{D}}$, the unit lower triangular matrix $\hat{\mathbf{L}}$, the permutation matrix $\mathbf{\Pi}$, and the modified received signal $\hat{\mathbf{r}}$. The diagonal matrix $\hat{\mathbf{D}}$ and the unit lower triangular $\hat{\mathbf{L}}$ can be obtained from the symmetrically permuted Cholesky factorization of the MMSE matrix

$$\mathbf{\Pi}(\mathbf{H}^H\mathbf{H} + \sigma_n^2\mathbf{I}_N)^{-1}\mathbf{\Pi}^T = \mathbf{L}\mathbf{D}\mathbf{L}^H \qquad (3)$$

In particular, the unbiased MMSE metric can be rewritten from Eq. (3) as

$$\mu_{\mathrm{UB}} = \left\| \mathbf{D}_{UB}(\mathbf{G}_{UB}\mathbf{r} - (\mathbf{I}_N - \mathbf{F}_{UB})\mathbf{\Pi}\mathbf{s}) \right\|^2 \qquad (4)$$

where $\mathbf{D}_{UB} = \mathbf{D}^{-1/2}(\mathbf{I}_N - \mathbf{D})$ is the transforming matrix to fulfill the requirements for an unbiased estimate. The matrix $\mathbf{F}_{UB} = (\mathbf{I}_N - \mathbf{D})^{-1}(\mathbf{I}_N - \mathbf{L}^{-1})$ is the unbiased MMSE V-BLAST feedback filter and $\mathbf{G}_{UB} = (\mathbf{I}_N - \mathbf{D})^{-1}\mathbf{G}_{MMSE-DFE}$ is the unbiased MMSE V-BLAST feedforward filter. Note that $\mathbf{G}_{MMSE-DFE} = \mathbf{D}\mathbf{L}^H\mathbf{\Pi}\mathbf{H}^H(\sigma_n^2)^{-1}$ is the original MMSE V-BLAST feed forward filter. Note that,

$$\mu = \left\| \hat{\mathbf{D}}(\hat{\mathbf{r}} - \hat{\mathbf{L}}\mathbf{\Pi}\mathbf{s}) \right\|^2 = \sum_{i=1}^{N} \hat{d}_{i,i}^2 \left| \hat{r}_i - s_{b_i} - \sum_{j=1}^{i-1} \hat{l}_{i,j} s_{b_i} \right|^2 \qquad (5)$$

Due to the triangular structure of $\hat{\mathbf{L}}$, the solution to Eq. (5) can be interpreted as a tree search procedure and obtained recursively from the top layer $i = 1$ to the bottom layer $i = N$ using

$$\mu_i = \hat{d}_{i,i}^2 \left| z_i - s_{b_i} \right|^2 + \sum_{j=1}^{i-1} \hat{d}_{j,j}^2 \left| z_j - s_{b_j} \right|^2 \qquad (6)$$

where $z_i = \hat{r}_i - \sum_{j=1}^{i-1} \hat{l}_{i,j} s_{b_j}$.

In Eq. (6), the first term can be seen as the partial Euclidean distance (PED) contribution from the i th level and the second term as an accumulated Euclidean distance (AED) up to level $j = i - 1$. More specifically, the QRDM algorithm starts by enumerating the admissible values of s_{b_1} first. Once the value of s_{b_1} is known, this information will be carried forward to the lower layer and the search will then involve only one undetermined symbol s_{b_2} since s_{b_1} has been found. The search algorithm continues until it reaches to the bottom layer and the symbol sequence $\{s_{b_1}, \ldots, s_{b_N}\}$ with the smallest ED value is chosen to be the final solution [1]. Note that, in the tree search analogy, the symbol sequence $\{s_{b_1}, \ldots, s_{b_k}\}$ corresponds to a node at the kth level, and any full symbol sequence $\{s_{b_1}, \ldots, s_{b_N}\}$ is referred to as a leaf node and the associated nodes up to the root node form a path.

Assuming a real-valued implementation is used, each node will have $\sqrt{P}$ possible children and will result in $M \cdot \sqrt{P}$ children in total during the search traversal. Consequently, a novel sorting/selection scheme should be used to pick the M best candidates from those children efficiently and effectively, which will be discussed in detail in following subsections.

C. Merge-Sort Scheme

The merge-sort (MS) scheme has been investigated in [5, 8], which is able to operate in a fully parallel/pipelined fashion. The most attractive advantage of the MS scheme is that it has a considerably reduced complexity compared with conventional bubble sort or its odd-even sorting counterparts [7]. This is essentially based on the fact that all the child nodes from one parent node can be enumerated in the Schnorr-Euchner (SE) order with ascending PED metrics. Hence, the inputs to the merge-sort unit are M independent sorted lists, each containing $\sqrt{P}$ values. Note that from a hardware implementation point of view, a lookup table (LUT) approach can be adopted to derive the SE enumeration list. As an illustration example, Fig. 1. shows a typical merge-sort architecture for QRDM with $M = 4$, denoted as QRD-4. Two stages of merge-sort units have been deployed in order to pick out the minimum 4 out of 16 values. The detailed merge-sort unit is shown on the right side and it mainly consists of parallel comparators and decision logic. Note that a typical QRDM hardware architecture would process all 4 nodes in parallel using a 4-way parallel structure and the so-called parallelism factor (denoted as λ_{PF}) thus equals 4.

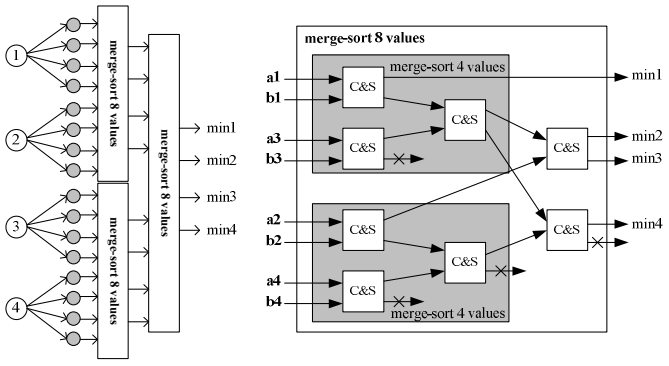

Fig. 1. Merge-sort architecture for QRD-4

For a particular MIMO detector, the throughput implementation can be computed as

$$Q = \frac{N \cdot \log_2 P \cdot f_{clock}}{C} (Mbps) \qquad (7)$$

where f_{clock} is the clock frequency of the design in MHz, which is determined by the critical path delay in the hardware, and C is average number of clock cycles required to detect a MIMO symbol. In the case of the QRDM, C is a constant and equals to one if it is implemented in a fully-parallel structure, i.e. $\lambda_{\mathrm{PF}} = M$.

D. Winner Path Expansion Scheme

The winner path expansion (WPE) scheme has been proposed recently in [6], which aims to find the required M best candidates by extending the minimum number of nodes. Thus, in contrast to the MS scheme, the WPE scheme finds the M best candidates in sequence. More specifically, instead of expanding **all** the child nodes from the parent nodes, only the nearest nodes are enumerated from each parent node. Following that, all these first children are feed into a MIN search unit where the one with the lowest PED is selected as the first candidate. Then that child is replaced by its next best sibling using the SE-enumeration. The above process is repeated $M-1$ times so that all the M best candidates are obtained. As an illustrative example, Fig. 2. shows a typical WPE architecture for the QRD-4 case. The detailed MIN search unit is shown on the right side, and it mainly consists of parallel comparators and decision logic. Clearly, for a parallel/pipelined structure, four MIN search units have to be employed. This leads to higher logic resource use compared with the MS counterpart as shown in Fig. 1. However, the advantage of the WPE scheme over the MS scheme is its minimum node extension, resulting in fewer computation units, which will be discussed in Section III.

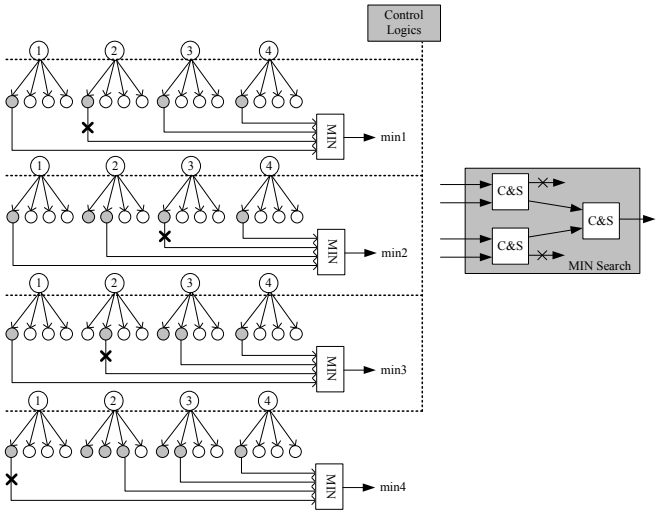

Fig. 2. Winner path expansion architecture for QRDM with M=4

III. RESULTS AND DISCUSSIONS

In order to investigate the feasibility and compare the resource use of different approaches, a so-called FPGA based rapid prototyping methodology is adopted in this work, which mainly relies on Xilinx's DSP System Generator [4]. Both MS and WPE schemes have been implemented for a 4×4 MIMO system with 16-QAM modulation. The FPGA platform from Alpha Data Ltd consists of an ADM-XRC-5T2 board with a Xilinx Virtex-5 SX240T FPGA device.

In Fig. 3, the BER performance of a number of detection schemes is shown in a 4×4 16-QAM MIMO system. It should be noted that both the MS and WPE approaches have the same BER performance, as they only differ in the implementation sorting scheme, so only one result is presented in the figure. From Fig. 3, as expected, the FPGA and MATLAB results of the QRD-4 agreed very well with only a negligible difference due to the quantization process, which verifies that the fixed-point FPGA implementation is functionally correct. Moreover, it can be observed that the QRD-4 with the unbiased MMSE metric (μ_{UB}) outperforms the conventional ML metric with only a small performance degradation compared to the SD benchmark algorithm.

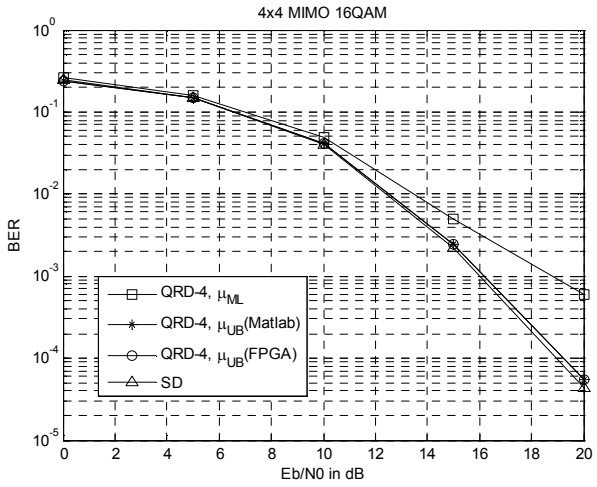

Fig. 3. BER performance of various detection schemes in a 4×4 MIMO system with 16QAM modulation.

The resource use of the FPGA implementation for the both MS and WPE approaches are summarized in Table I. In order to establish a fair comparison, the MS scheme is implemented by the choice of $\lambda_{\mathrm{PF}} = 4$ so that it provides the constant throughput of 1600Mbps at a clock frequency of 100MHz. Note that the original MS scheme has been implemented with $\lambda_{\mathrm{PF}} = 1$ and thus achieves 400Mbps accordingly [5]. Initially, from Table I, we can see that the WPE scheme uses slightly more FPGA logical resources, such as slices, flip-flops (FFs) and look-up-tables (LUTs), than its MS counterpart. This is due to the fact that, although the MIN search unit in the WPE scheme is more resource efficient than merge-sort unit in the MS scheme, 4 of them have to be employed to make algorithm operate in a parallel/pipelined

fashion. Moreover, by introducing the winner path expansion strategy, the WPE approach requires extra control logic resources to identify the nodes that need to be extended subsequently. However, on the other hand, the WPE approach saves noticeably on the computational resource, i.e. DSP48E and Block RAMs. This can be attributed to the fact that the WPE approach requires the minimum number of expanded nodes, which in turn involves fewer computation units to obtain the ED metrics. Also, it is worth noting that, when the parallelism factor λ_{PF} increases from 1 to 4, the computational resources increase considerably whereas the logical resources grow slightly. This is due to the fact that the logical resources can be reused in the parallel structure, while the computational units have to be dedicated to the metric computation tasks.

TABLE I. FPGA RESOURCE USE OF THE MS AND WPE ALGORITHMS.

Xilinx V5SX240T	**MS based QRD-4** ($\lambda_{PF}=1$)	**MS based QRD-4** ($\lambda_{PF}=4$)	**WPE based QRD-4**
Slice (37440)	7644 (20%)	12041 (32%)	14317(38%)
FFs (149760)	21923 (14%)	34691 (23%)	41639(27%)
LUTs (149760)	13759 (9%)	28456 (19%)	30832 (20%)
DSP48E (1056)	96 (9%)	384 (36%)	276 (26%)
Block RAMs (516)	31 (6%)	63 (12%)	45 (8%)
Clock (MHz)	100	100	100
Throughput (Mbps)	400	1600	1600

In order to investigate the power efficiency of each particular design, Xilinx *XPower Analyzer* tool has been used to estimate the approximate power consumption for different algorithms. Note that the total power consumption comprises both quiescent and dynamic power. Quiescent power is the power consumed within an FPGA when it is powered up with no clocks operating, while dynamic power is the additional power consumed through the operation of the device caused by signals toggling and capacitive loads charging and discharging.

TABLE II. POWER CONSUMPTION COMPARISON IN TERMS OF THROUGHPUT PER UNIT POWER.

Algorithms	**MS based QRD-4 (pf = 1)**	**MS based QRD-4 (pf = 4)**	**WPE based QRD-4**
Throughput (T) (Mbps)	400	1600	1600
Quiescent Power (W)	3.39	3.60	3.61
Dynamic Power (W)	0.99	1.48	1.49
T / Total Power (Mbps/W)	91.27	314.55	313.64

From Table II, in terms of the throughput of each approach per unit power, we observed that similar performances are obtained by MS based and WPE based schemes as they are designed in different ways with advantages in either logic or computational resources as discussed previously. Also, it is shown that more power efficiency can be obtained by using a highly parallel structure with either approach.

IV. CONCLUSIONS

Efficient sorting schemes in the QRDM algorithm are crucial to achieve an ultra-high decoding throughput and low power consumption. In this paper, two important approaches to fulfill the task of finding the M best candidates have been investigated. It has been shown that, due to the fact that they arise from different perspectives, MS based and WPE based schemes show their superiorities in different aspects. Moreover, the FPGA prototyping result reveal that both MS and WPE schemes are able to support over Gbps decoding throughput when they are implemented a fully pipelined parallel architecture, making them particularly desirable for realizing high data-rate MIMO systems.

ACKNOWLEDGMENT

This work was supported by the UK Government's Engineering and Physical Sciences Research Council (EPSRC), ISLAY Project: Adaptive Hardware Systems with Novel Algorithmic Design and Guaranteed Resource Bounds. Grant number EP/F03072X/1.

REFERENCES

[1] B. M. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," IEEE Trans. Commun., vol. 51, no. 3, pp. 389-399, Mar. 2003.

[2] A. Burg, M. Borgmann, M. Wenk, M. Zellweger, W. Fichtner, and H. B¨olcskei, "VLSI implementation of MIMO detection using the sphere decoding algorithm," IEEE J. Solid-State Circuits, vol. 40, no. 7, pp. 1566–1577, July 2005.

[3] Wenk, M., Zellweger, M., Burg, A., Felber, N., and Fichtner,W., "K-best MIMO detection VLSI architectures achieving up to 424 Mbps throughput," in IEEE International Symposium on Circuits and Systems, pp. 1151-1154, May 2006.

[4] L. G. Barbero and J. S. Thompson, "Fixing the complexity of the sphere decoder for MIMO Detection," IEEE Trans. Commun., vol. 7, no. 6, pp. 2131-2142, June 2008.

[5] M. Joham, L. G. Barbero, T. Lang, W. Utschick, J. Thompson, T. Ratnarajah, "FPGA Implementation of MMSE metric based efficient near-ML detection," Int. ITG Workshop on Smart Antennas, pp. 139-146, Darmstadt, Germany, Feb. 2008.

[6] Sudip Mondal, Ahmed M. Eltawil, and Khaled N. Salama, "Architectural Optimizations for Low-Power K-Best MIMO Decoders," IEEE Trans. Veh. Tech., vol. 58, no. 7, pp. 3145 – 3153, Sept. 2009.

[7] S. G. Akl, Parallel Sorting Algorithms, Academic Press, Inc., 1985.

[8] N.M. Madani and W.R. Davis, "High-throughput low-complexity MIMO detector based on K-best algorithm," ACM Great Lakes Symposium on VLSI 2009, pp. 451-456, Boston, USA, May 2009.