Systolic Like Soft-Detection Architecture for 4x4 64-QAM MIMO System

Pankaj Bhagawat
Dept. of E.C.E
Texas A&M University
College-Station,TX-77840
pankaj-bhagawat@neo.tamu.edu

Rajballav Dash
Dept. of E.C.E
Texas A&M University
College-Station,TX-77840
rajballavdash@neo.tamu.edu

Gwan Choi
Dept. of E.C.E
Texas A&M University
College-Station,TX-77840
gchoi@ece.tamu.edu

***Abstract*— MIMO systems (with multiple transmit and receive antennas) are becoming increasingly popular, and many next-generation systems such as WiMAX, 3-GPP LTE and IEEE802.11n wireless LANs rely on the increased throughput of MIMO systems with up to four antennas at receiver and transmitter. High throughput implementation of the detection unit for MIMO systems is a significant challenge especially for higher order modulation schemes. To achieve superior Bit Error Rate(BER) or Frame Error Rate (FER) performance, the detector has to provide soft values to advanced Forward Error Correction (FEC) schemes like Turbo Codes. This paper presents a systolic soft detector architecture for high dimensional(eg. 4x4, 64-QAM) MIMO systems. A Single detector core achieves, throughput of 215Mbps and power consumption of 23.6mW, whiles using only 33.1K gate equivalent(for l^2 norm). Impressive SNR gains of almost 2dB are observed with respect to the hard detection counterpart over a block fading channel(at an FER of 1%). Additionally, the architecture can be stacked to give linear increase in throughput with linear increase in hardware resources.**

I. Introduction

The scarcity of available radio frequency spectrum combined with the increasing need for higher data rates has led to the use of multiple input-multiple output (MIMO) wireless systems, which offers higher throughput without any overhead in terms of bandwidth or transmitter power as compared to single input single output (SISO) wireless system. Future generation wireless standards such as IEEE802.11n, 3-GPP LTE, Wi-Max etc. all have MIMO as a key enabling technology.

Designing an efficient hardware for soft detection of high-dimensional MIMO systems such as 4x4-64-QAM is hard challenge. A soft detector not only computes binary (or hard) estimates of the transmitted bits, but also provides "reliability"' (or soft decisions) of the binary estimates. The soft decisions from the detector is then fed to FEC schemes such as turbo-decoder or a Viterbi decoder. In all the cases soft values based FEC decoding provides much better BER performance than its hard counterpart[12]. In the past, very few authors have addressed the issues of implementing a soft detector for highly complex systems such as 4x4 with 64-QAM MIMO systems, some of the notable ones are [6,7]. However, none of the design are able to meet the exacting requirements on throughput that future standards place.

The choice of algorithm and architecture has a significant bearing on the final hardware complexity and reconfigurability. Apart from the BER/FER performance, we focus on architectural issues like pipelining,and parallelism. The detection algorithms can be broadly classified into linear, and non-linear. Linear algorithms like zero-forcing(ZF) [11], or Minimum Mean Squared Error(MMSE)[11] are low complexity but incur high penalty in BER/FER performance. Non-linear detectors like Successive Interference Cancellation(SIC)[11] are low complexity too, but provide only modest gain over their linear counterparts. Moreover, neither ZF nor SIC based receivers do well in a wireless channel with limited diversity[11]. Authors in [11,14] provide excellent comparative study of various detectors in different channel conditions. In such channel conditions, it is clear that more sophisticated algorithms(tree search based) need to be considered for practical systems due their superior BER/FER performance.

To get close to the optimum BER/FER performance researchers have proposed many algorithms, that do non-exhaustive tree search, such as List Sphere Decoder (LSD)[12], however, its complexity is still too large for higher order MIMO systems, and is very hard to map onto a parallel, pipelined architecture. Furthermore, LSD converges in a random fashion making it difficult to incorporate in a practical system. On the other hand, algorithms based on Breadth First Search(BFS) such as K-best, provides constant throughput but involves sorting operation which is very expensive. As in the case of LSD, higher order modulation schemes like 64-QAM only makes matter worse. One of the reported implementation of a soft MIMO detector that supports 64-QAM is presented in [13]. Recently [14] presented an algorithm that takes BFS based approach to the problem, this algorithm is called Layered ORthogonal Lattice Detector(LORD). It can be implemented in a highly parallel and pipelined manner, and has a fixed throughput. However, it involves multiple QR decomposition operations which are not only expensive but also require large memory to store the decomposed matrices. Multiple QR decompositions also adversely impact the "latency" of the detection process, which is a very important design parameter.

Contributions: Using Algorithm/Architecture co-design approach a novel soft detection algorithm and a very high speed systolic architecture is developed for 4x4 64-QAM MIMO systems. The detector has fixed throughput(215Mbps) and achieves almost 2dB SNR gain w.r.t the hard decoded counterpart on a block fading channel. Additionally, clock gating has been successfully incorporated to make it energy efficient. Higher throughput can be simply achieved by instantiating multiple cores and have them process OFDM tones concurrently. The architecture achieves very high resource usage.

The paper is organized as follows: Section II describes the basics of the channel model and the sphere detection algorithm. Section III describes the proposed scheme and its architecture in detail. In Section IV we discuss the results. Section V concludes the paper.

978-3-9810801-5-5/DATE09 © 2009 EDAA

II. MIMO Detection

A. Channel Model and Optimal Hard MIMO detection

A generalized MIMO system with M_T transmit and M_R receive antennas can be expressed in terms of matrices as shown in eqn.1[1].

$$\mathbf{y} = \mathbf{Hs} + \mathbf{n} \tag{1}$$

where **y** received vector, **s** transmitted vector (will be referred to as a MIMO symbol in the sequel), **n** is $M_R \times 1$ zero mean complex Gaussian noise vector, and **H** is a $M_R \times M_T$-dimensional complex matrix. Each element in **s** can take η values. In this paper we will assume $M_T = M_R = 4$, unless specified otherwise.

The objective of the MIMO detector is to estimate $\hat{\mathbf{s}}$ of **s** based on the the observation of **y** along with the knowledge of **H**. It has been shown that the optimal or the Maximum Likelihood (*ml*) estimate $\hat{\mathbf{s}}_{ml}$ of **s** is given by eqn.2 [12]:

$$\hat{\mathbf{s}}_{ml} = arg \min_{\mathbf{s} \in \Omega^{M_T}} \|\mathbf{y} - \mathbf{Hs}\|^2 \tag{2}$$

Furthermore, **H** can be triangularized using QR decomposition: $\mathbf{H} = \mathbf{QR}$. where, **R** is an upper triangular matrix, and $\mathbf{Q}^H$ is the Hermitian of a unitary matrix **Q**. Hence, the cost function given by (2) can now be rewritten as [5],

$$\hat{\mathbf{s}} = \|\mathbf{y} - \mathbf{Hs}\|^2 = \|\hat{\mathbf{y}} - \mathbf{Rs}\|^2, and\ \hat{\mathbf{y}} = \mathbf{Q^H y} \tag{3}$$

Eqn.(3) can be further expanded as shown in eqn.(4)-(6).

$$d_i(s^{(i)}) = d_{i+1}(s^{(i+1)}) + |e_i(s^{(i)})|^2 \tag{4}$$

$$|e_i(s^{(i)})|^2 = |c_{i+1}(s^{(i+1)}) - R_{ii}.s_i|^2 \tag{5}$$

$$c_{i+1}(s^{(i+1)}) = \hat{y}_i - \sum_{j=i+1}^{M_T} R_{ij}.s_j \tag{6}$$

The quantity $|e_i(s^{(i)})|^2$ will be called the Incremental Euclidean Distance (IED), and the term $d_i(s^{(i)})$ will be called Partial Euclidean Distance (PED) for $i > 1$, and Euclidean Distance (ED) for $i = 1$. The fact that **R** is upper-triangular ensures that each term on LHS of eqns.(4)-(6) depends only on the current level i, and the history of the path to reach that level(note that in eqn.(6), the index j runs from $i+1$ to M_T). Because the PED's depend only on $s^{(i+1)}$, they can be associated with corresponding nodes in a η-*ary* tree with M_T levels. The computation of the terms $d_1(s^{(1)})$ can then be interpreted as a traversal of the tree from the root($i = M_T$) to the leaf ($i = 1$)corresponding to **s**. The estimate can now be obtained by searching the leaf with the smallest ED and returning the path from the top level to that leaf as $\hat{\mathbf{s}}_{ml}$. The complexity of this tree search can be greatly reduced by noting that IEDs are always positive, and hence if the PED of a node exceeds a predefined threshold (called radius) the subtree rooted at that node can be excluded from further search. This approach is commonly known as sphere decoding.

B. Soft MIMO Detection

The objective of a soft MIMO detector is to output the reliability associated with each hard output bit. This reliability is expressed in terms of the Log-Likelihood Ratio (LLR) of each bit, and is defined as $L(x_{i,j}) = ln\frac{P(x_{i,j}=1|\mathbf{y})}{P(x_{i,j}=0|\mathbf{y})}$, where $x_{i,j}$ is j^{th} bit in label of the i^{th} constituent QAM symbol. This can be approximated [12] as:

$$L(x_{i,j}) \approx \min_{\mathbf{s} \in \mathbf{X}_{i,j}^{(0)}} \{\Gamma(\mathbf{s}, \mathbf{y})\} - \min_{\mathbf{s} \in \mathbf{X}_{i,j}^{(1)}} \{\Gamma(\mathbf{s}, \mathbf{y})\} \tag{7}$$

where $\Gamma(\mathbf{s}, \mathbf{y}) = \|\mathbf{y} - \mathbf{Hs}\|^2$, $\mathbf{X}_{i,j}^{(0)}$ and $\mathbf{X}_{i,j}^{(1)}$ are sets of vector symbols, with j^{th} bit in the label of i^{th} constituent QAM symbol, as 0 and 1 respectively.

In eqn.7 there are two minimization problems (eqn.2 has only one), i.e for each bit $x_{i,j}$ it requires identification of the most likely transmit sequence where $x_{i,j} = 0$ and the most likely one where $x_{i,j} = 1$(first and second term in eqn.7 respectively) along with their respective metrics. The difference between these two metrics gives us the LLR value of $x_{i,j}$(one of the two minima in eqn. 7 is always given by the metric associated with the *ml* bit sequence). Hence, if we let $d_{i,j}^{ml} = \|\mathbf{y} - \mathbf{Hs}_{ml}\|^2$(since $d_{i,j}^{ml}$ is independent of *i,j*, these subscripts will be dropped in future) , and the other minimum in eqn.7 be $d_{i,j}^{\overline{ml}}$, where the counter-hypothesis $x_{i,j}^{\overline{ml}}$ is the complement of the j^{th} bit in the label of i^{th} QAM symbol in $\mathbf{s}_{ml}$, then eqn.7 can be rewritten as:$L(x_{i,j}) \approx d_{i,j}^{\overline{ml}} - d^{ml}$, if $x_{i,j}^{ml}$=1 and $L(x_{i,j}) \approx d^{ml} - d_{i,j}^{\overline{ml}}$, if $x_{i,j}^{ml}$=0. Thus, to compute the LLRs the detector has to compute $\mathbf{s}_{ml}$, d^{ml}, and $d_{i,j}^{\overline{ml}}$ for i=1,2,..M_T and j=1,2,...,$log_2\eta$ (For 4x4 64-QAM system M_T=4, and η=64).

III. Proposed Soft Detection Algorithm and Architecture

Intuitively, the FER/BER performance of the soft MIMO detector will depend on the signs and magnitudes of the LLRs being fed to the FEC decoder. From the earlier discussion it is clear that sign of the LLRs crucially depend on the effectiveness of the detector to get to $\mathbf{s}_{ml}$. Fixed Sphere Decoder(FSD)[3] was proposed as an efficient alternative for providing quasi *ml* hard decoding performance, hence it is a suitable candidate for computing $\mathbf{s}_{ml}$ and d^{ml}.

In FSD all children of the root node are processed, thereon, only their best child are extended. The hard decoded MIMO symbol is the path from root to leaf node that has the minimum ED(*ml* path). Fig.1a shows the reduced tree structure for the hard decoding based on FSD algorithm. To compute the soft values of the associated bits we propose to use not only the *ml* path, but also the "surrounding" paths. As noted earlier that for every bit, one term eqn.7 is always associated with the *ml* path. To compute the other term we search for the paths with opposite bit and pick the one with least ED. If a path with a valid counter hypothesis is not found we simply assign the corresponding LLR, a *clipping* value with appropriate sign. Clipping is also applied to limit the maximum magnitude of the LLR.

A. High Level Architecture and Data Flow

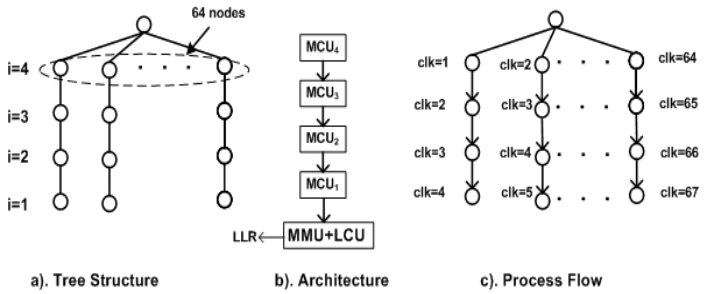

Fig. 1. Tree Structure and High Level Architecture/Process-Flow

Fig.1b shows the high level architecture of the proposed decoder. It consists of an one dimensional systolic array of Metric Computation Units (MCUs). A MCU_i evaluates eqns.(4)-(6) for a particular i. These units feed the Metric Management Unit(MMU),

and the LLR Computation Unit(LCU). We will provide details on these units in subsequent subsections.

Fig.1c shows the process flow of the detection process(assuming one MCU takes one cycle to process), it shows the sequence in which the nodes in the tree are processed. MCU_4 is being utilized for cycles from 1 to 64, MCU_3 from 2 to 65, and so on. Note that even though it takes 67 cycles to process one MIMO symbol, a new MIMO symbol can be fed into the pipeline after(at MCU_4), and hence it effectively takes 64 cycles to process one MIMO symbol.

Thus, the throughput of the architecture is given by: $\theta = \frac{24}{64} freq$, where $freq$ is the operational frequency of the architecture(each QAM symbol constructed using 6bits, and since there are 4 such QAM symbols, total bits in a MIMO symbol is 24)

B. MCU architecture

The MCU computes eqns.(6)-(4) in that order. Fig.2 shows the detailed structure of an MCU at level 1. The MCU in turn is composed of 1) Product Computers (PCs) 2) Adders 3) a Slicer 4)a Norm Computer (NC).

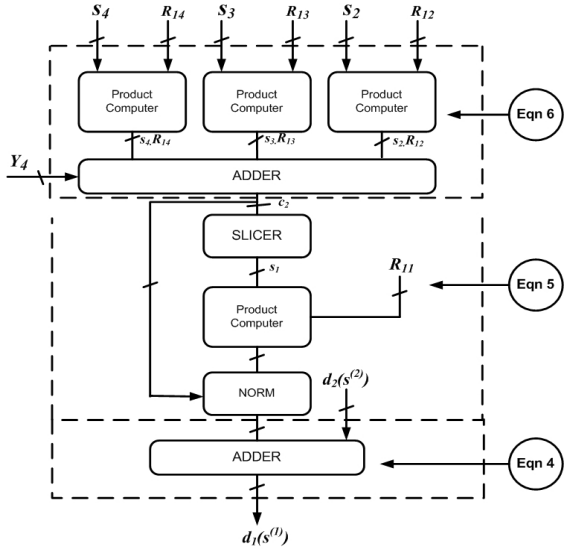

Fig. 2. MCU at Level 1

Product Computer: This unit computes the product of R_{ii} and s_i as required in eqn.5 and eqn.6. This "product" can be implemented simply a by shift and add operation, because the QAM constellation points only take on a finite number of integer values (e.g. in 64-QAM scheme the real and imaginary part of $s_j \epsilon \{-7, -5, -3, -1, 1, 3, 5, 7\}$).

Slicer: This unit picks the "best" child (nodes with least $|e_i|^2$) of a parent node. From eqn.(5) it can be seen that, in order to minimize $|e_i|^2$, we need to compute s_i such that the *distance* between c_{i+1} and $R_{ii}s_i$ is minimized. The slicer block picks the nearest *scaled* QAM symbol($R_{ii}s_i$) to c_{i+1}. This operation involves independently comparing real and imaginary parts of c_{i+1} with appropriate decision thresholds and picking the closest points on each axis. The best child of a parent, which is a complex number, can be constructed using the results of the independent comparisons on real and imaginary axis.

Norm Computer:This unit computes the Euclidean norm or l^2 norm using eqn.5.

C. Metric Management Unit

The MMU keeps track of the appropriate terms required to compute the LLR values using eqn.7. It operates concurrently on the data stored in memory locations $a_{i,j}$, $b_{i,j}$, $c_{i,j}$ and d_a, d_b. Where, $a_{i,j}$ is the $(i,j)^{th}$ bit of the current *ml* hypothesis, and d_a is the metric associated with it. Similarly, $b_{i,j}$ is the $(i,j)^{th}$ bit of the incoming vector symbol, and d_b is the metric associated with it. Finally, $c_{i,j}$=$d^{\overline{a}}_{i,j}$ is the ED of the current best counter-hypothesis of $a_{i,j}$. If $d_a > d_b$, it means the incoming vector is the new *ml* hypothesis. Hence, for all i, j, d_a will become new $c_{i,j}$, if $a_{i,j}$ and $b_{i,j}$ are complements of each other. This would be followed by assignments $a_{i,j}$=$b_{i,j}$, and d_a=d_b. If $d_a < d_b$, it means the incoming vector cannot be a new *ml* hypothesis. It may however, still effect the EDs of counter-hypothesis. Hence, for all i, j, such that $a_{i,j}$ and $b_{i,j}$ are complements of each other it needs to assign $c_{i,j}$=d_b, if $d_b < c_{i,j}$.

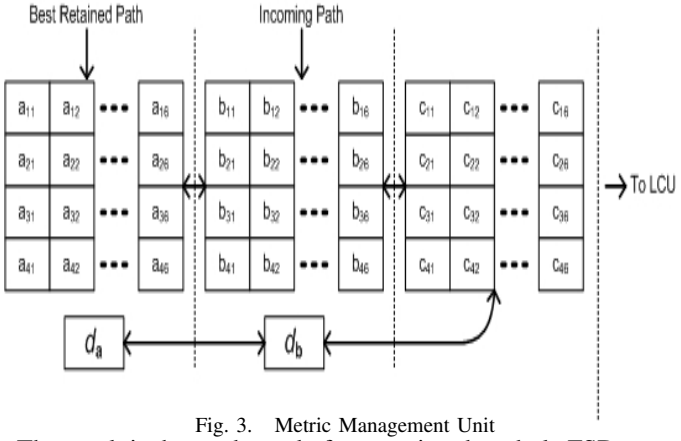

Fig. 3. Metric Management Unit

The result is that at the end of processing the whole FSD tree we have de-mapped $\mathbf{s}_{ml}$ in matrix, $\mathbf{a}$, d^{ml} in d_a, and $d^{\overline{ml}}_{i,j}$ in matrix $\mathbf{c}$. As noted earlier, these are the quantities that are needed to compute the LLR values.

D. Node Pruning to Lower the Energy Consumption

As mentioned in section II.A, sphere decoder reduces the search complexity by updating the radius value whenever a leaf node is reached. We apply same concept to our detector, except that we use (d_a + clip) as radius. This way we can preclude(via clock gating) some MCUs from carrying out computations, thereby reducing energy consumption. Pruning is usually most aggressive

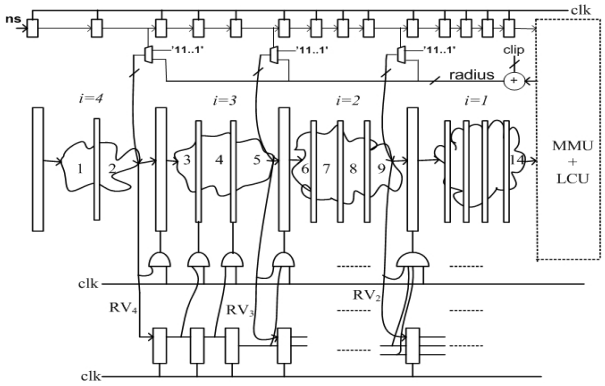

Fig. 4. Clock Gating Details

when the top level nodes in the FSD tree are processed in increasing order of their PEDs. One way to do this efficiently, is by enumerating them as described in [5] or in [4]. However,

approaches in [4-5] are not conducive for pipelining because of the inherent loop that occur in the hardware realization of the procedure. Hence, we propose a suboptimal approach to carry out enumeration. Finding the exact enumeration on real(or imaginary) axis can be implemented using counters generating a zig-zag pattern. In our approach, we first find the zig-zag enumeration pattern the symbols on real axis and imaginary axis(similar to Schnorr-Euchner enumeration). We then fix the real part constant while we pick imaginary part per the enumeration pattern until the column corresponding to the real part is exhausted. We then fix the next real part in the pattern and keep it constant while we traverse its column. We do this until all the 64 points are visited.

In hardware, node pruning can be achieved by clock gating as shown in Fig.4. d_a from MMU is the current best metric, which is added to clip to get the radius. to distinguish between current vector symbol and the next one we use "ns" bit to drive the value of radius to a very large value('111..1'), this is preclude the radius of older vector symbol to interfere. Each combinational "cloud"' consists of MCU_i and a comparator to check for radius violations (RVs). RVs are basically the clock gating signals that propagate along the pipeline as shown. Note that, by introducing clock gating we have introduced a loop in the MCU array. However, this loop can be run at a high speed since it has a two operand (7 and 3 bits each) adder and a 2-to-1 MUX (this delay comes to about 0.8ns based on our synthesis results).

IV. Results and Discussion

We evaluated the algorithm on a block fading channel of 120 information bits encoded by a rate 1/2 convolutional encoder with generator polynomial of [7,5]. Hence 240 coded bits were transmitted over which the fading matrix **H** was constant. During the next block H was generated independently. We counted 100 frame errors to get an estimate of FER.

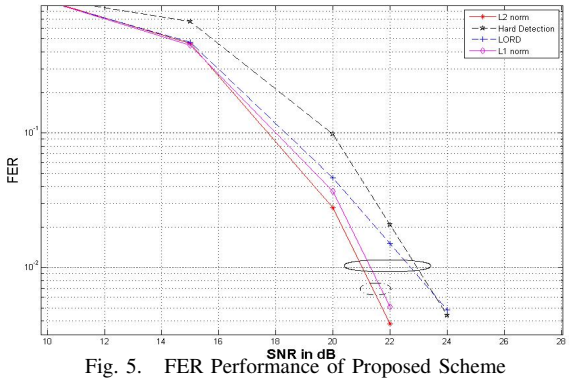

Fig. 5. FER Performance of Proposed Scheme

The FER plot for the proposed detector is shown in fig.5. We see that at FER of 1% the proposed detector gains almost 2dB wrt to the optimal hard detector(for clip=3). It also outperforms LORD(an implementation friendly algorithm [14]) by about 1.6dB. Use of l^1 norm causes the FER to degrade by about 0.4dB.

We can introduce pipelines to achieve high clocking frequency. Let p_i denote the number of pipelines in level i. For l^2 norm we chose p_i=9,9,8,4 for i=1,2,3,4. We chose eleven bit fixed point quantization(internal precision was maintained) for negligible FER degradation. The RTL coding was done using Verilog HDL. Nangate 45nm CMOS standard cell library was used for the design flow. Synopsys Design Compiler was used to synthesize the gate level net-list and to get power, area, and delay estimates. Throughput and synthesis result are summarized in Table.I.

TABLE I
Synthesis Results and Comparisons

	Proposed	[13]	[7]
Gate Equivalent	33.1K	280K	70K
Power Consumption at 20dB(mW)	23.6	94	114
Energy per Bit at 20dB(nJ)	0.11	N.A	0.61
Frequency(MHz)	574.7	270	500
Throughput(Mbps)	215	8.57	187.5
SNR Gain wrt to hard detection	2dB	N.A	N.A
Tech. Library	45nm	130nm	45nm

V. Conclusion

A novel high speed systolic MIMO detector architecture and its ASIC implementation estimate is presented in this paper. By using multiple detectors operating concurrently the throughput scales linearly with linear increase in hardware. This detector is highly suitable for MIMO-OFDM systems which require very high throughputs.

References

[1] W. Wolniansky, et al., *"V-BLAST:An architecture for realizing very high data rates over the rich-scattering wireless channel"*, Proc. IEEE ISSSE 1998, pp.295-300, Sept. 1998.

[2] Z. Guo and P. Nilsson, *"Algorithm and implementation of the K-best sphere decoding for MIMO detection"*, IEEE Journal on Selected Areas in Communications, Volume 24, Issue 3, March 2006, pp 491-503.

[3] L. Barbero and J. Thompson, *"Rapid Prototyping of a Fixed-Throughput Sphere Decoder for MIMO Systems"*, in IEEE International Conference on Communications (ICC '06), Istanbul, Jun. 2006.

[4] Burg, A.,et al., *"VLSI implementation of MIMO detection using the sphere decoding algorithm"*, IEEE Journal Solid State Circuits, vol.40, pp 1566-1577, July 2005.

[5] Bhagawat,P., Ekambavanan,S., Das,S., Choi,G., Khatri.S, *"VLSI Implementation of a Staggered Sphere Decoder Design for MIMO Detection"*, Forty-Fifth Annual Allerton Conference, September 26-28, 2007, University of Illinois at Urbana-Champaign, IL, USA.

[6] Sizhong Chen,Tong Zhang, Goel, M.*" Relaxed tree search MIMO signal detection algorithm design and VLSI implementation"*, Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on, 21-24 May 2006.

[7] Bhagawat,P., Dash,R., Choi, G., *"Dynamically Reconfigurable Soft Output MIMO Detector"*, accepted for publication in XXVI IEEE Conference on Computer Design, ICCD, Oct.2008.

[8] Huang,X., Liang,C., Ma, J., *"System Architecture and Implementation of MIMO Sphere Decoders on FPGA"*, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol 16, No.2, pp. 188-197, Jan.2008.

[9] Shariat-Yazdi, R.Kwasniewski, T., *"Challenges in the Design of Next Generation WLAN Terminals"*, Canadian Conference on Electrical and Computer Engineering(CCECE), pp. 1483-1486,April.2007.

[10] Bhagawat,P., Dash,R., Choi, G., *"Architecture for Reconfigurable MIMO detector and its FPGA Implementation"*, accepted for publication in 15th IEEE International Conference on Electronics, Circuits, and Systems, ICECS 2008.

[11] Michalke,C., Zimmermann,E., Fettweis, G., *"Linear Mimo Receivers vs. Tree Search Detection: A Performance Comparison Overview"*, IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC),pp.1-7, Sept. 2006.

[12] Hochwald,B. M., TenBrink,S., *"Achieving Near-Capacity on a Multiple-Antenna Channel"*, IEEE Trans. on Commun., 51:389399, Mar. 2003.

[13] Chen,S., Zhang,T., Xin, Y., *"Relaxed K-best MIMO Signal Detector Design and VLSI Implementation"*, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 15, issue 3, pp. 328-337, March 2007

[14] Siti, M., Fitz, M.P., *"A Novel Soft-Output Layered Orthogonal Lattice Detector for Multiple Antenna Communications"*, IEEE International Conference Communications, 2006. ICC '06.

[15] Wang, R., Giannakis, G., *"Approaching MIMO channel capacity with reduced-complexity soft sphere decoding,"*, in Proc. of IEEE Wireless Communications and Networking Conf. (WCNC), vol. 3, Mar. 2004, pp.16201625.

[16] J. Stine, et al., *"FreePDK: An Open-Source Variation-Aware Design Kit."*, Proceedings of the 2007 IEEE International Conference on Microelectronic Systems Education, 2007.