FPGA-based Tabu Search for Detection in Large-Scale MIMO Systems


Published in:
Proceedings of 2014 IEEE Workshop on Signal Processing Systems (SiPS)

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2014 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
FPGA-based Tabu Search for Detection in Large-Scale MIMO Systems

Yun Wu, John McAllister
Queen’s University Belfast
UK
Email: {yun.wu, jp.mcallister}@qub.ac.uk

Abstract—The increasing scale of Multiple-Input Multiple-Output (MIMO) topologies employed in forthcoming wireless communications standards presents a substantial implementation challenge to designers of embedded baseband signal processing architectures for MIMO transceivers. Specifically, the increased scale of such systems has a substantial impact on the performance/cost balance of detection algorithms for these systems. Whilst in small-scale systems Sphere Decoding (SD) algorithms offer the best quasi-ML performance/cost balance, in larger systems heuristic detectors, such Tabu-Search (TS) detectors are superior. This paper addresses a dearth of research in architectures for TS-based MIMO detection, presenting the first realisations in particular is well recorded, no comparable effort that of simple linear detectors. It appears that TS-based detectors have an important role to play in the large-scale MIMO receivers of tomorrow.

Whilst research on SD realisations in general and FSD realisations in particular is well recorded, no comparable effort has been invested in TS-type detectors and it is not clear that the computational efficiency gains observed extend to gains in physical cost and performance. This paper resolves this issue, making three contributions:

1) FPGA-based TS MIMO detectors are presented for 4 × 4 and 10 × 10 MIMO systems. To the best of the authors’ knowledge, these are the largest single-chip detectors on record.

2) The largest recorded single-chip FSD detector on record is presented: a detector for 10 × 10 4-QAM MIMO.

3) It is shown that, despite exhibiting detection efficiency several orders of magnitude better than FSD, TS-based detection lags FSD in terms of both cost and performance, specifically exhibiting reductions in absolute throughput and throughput per unit resource of 47% and 82% respectively.

The remainder of the paper is organized as follows. Section II introduces related background on TS for MIMO detection, before Section III analyzes the performance and cost relative of TS to FSD. Section IV extends this comparative analysis to incorporate FPGA architecture performance and cost.

I. INTRODUCTION

The continually increasing demands for higher data rates in modern wireless communication systems has led to the increased adoption of MIMO systems, where multiple antennas are employed at both the transmitter and receiver ends of the communication channel [1], in wireless standards such as IEEE 802.11n [2] and 802.16e. MIMO schemes achieve superior channel capacity, throughput and diversity over single antennas at the cost of increased baseband signal processing complexity in transceivers. This places huge demand on embedded processing resources in terms of throughput, latency and resource utilization, particularly as the number of antennas increases [3].

The baseband signal processing complexity of MIMO systems is particularly apparent in the embedded detection problem. To achieve close to Maximum Likelihood (quasi-ML) performance, the detection complexity scales in a super-linear fashion. A substantial body of work has addressed the design of embedded detectors which employ Sphere Decoding (SD) algorithms [4], [5], [6] for this operation, motivated directly by the ability of SD algorithms to enable quasi-ML detection performance for relatively small MIMO topologies of 2, 4 or 8 antennas.

However, as the size of MIMO topologies employed increases, the dynamics of performance and cost of different detection algorithms changes. Specifically, [3] compares how the complexity and performance of linear detectors, SDs and heuristic detectors based on Tabu Search (TS), varies with the scale of MIMO system. One major conclusion which may be drawn is that as the number of antennas increases from a few (for instance, 2 or 4) to many, for instance 40, TS becomes substantially more computationally efficient than the Fixed Complexity Sphere Decoder (FSD) [7], a leading highly efficient SD algorithm and even exhibits detection efficiency beyond that of simple linear detectors. It appears that TS-based detectors have an important role to play in the large-scale MIMO receivers of tomorrow.

II. BACKGROUND

MIMO communication systems adopt multiple antennas at both transmit and receive terminals, as shown in Fig. 1. An $M$-element transmit antenna array transmitting a signal $s \in \mathbb{C}^M$ via a multi-path fading channel $\mathbf{H} \in \mathbb{C}^{N \times M}$, results in a received symbol vector $r \in \mathbb{C}^N$. Assuming Additive White Gaussian Noise (AWGN), $w \in \mathbb{C}^N$ at the receiver, $r$ can be modelled as

$$y = \mathbf{H} \cdot s + w. \quad (1)$$

Detection algorithms form an estimate $\hat{s}$ of the transmitted symbol vector $s$ given the received symbol vector $y$ and knowledge of the channel in the form of the channel matrix $\mathbf{H}$. Detection algorithms typically belong to one of three
categories - linear detectors such as Zero Forcing (ZF) or Minimum Mean Square Error (MMSE), non-linear detectors such as Sphere Decoders (SD), and heuristic approaches. The use and realisation of linear and SD algorithms is very well studied, see e.g. [1], [7], [6], [5], [4]. The use of heuristic approaches is not so well developed.

Heuristic detection methods include Genetic Algorithms (GAs) [8], Ant Colony Algorithms (ACAs) [9], [10], Simulated Annealing (SA) [11] and TS [11]. Among these, TS is only approach which has proven capable of quasi-ML performance [12], [13], [14], [15]. Figure 2 shows the tree structure of a TS detection algorithm. TS iteratively updates a global Tabu List (TL) record f stemming from \( \hat{y} \) an equalized version of \( y \); equalization may adopt various strategies, including ZF, MMSE or OSC - in this study we restrict this to ZF. During each iteration \( i \), \( f^{(i)} \) is updated via four steps: initialization, candidate enumeration, tabu move and TL (TL) updating. The iterations of the TS algorithm repeat until a specified stopping criteria is met.

1) Initialization: An initial solution vector, \( x^{(0)} = \hat{s}_0 \), is derived from \( \hat{y}_{ZF} \), the ZF equalized version of \( y \). A simplified ML cost calculation [13] is performed by defining \( y_{MF} = H^H \cdot \hat{y}_{ZF} \) and \( \tilde{H} = H^H \cdot H \); the initial vector, \( f^{(0)} \), for ML cost calculation is given by [13]

\[
f^{(0)} = \tilde{H} \cdot x^{(0)} - \hat{y}_{ZF}.
\]

In addition, TL is initialized while the iteration number and the Tabu Period are set to \( I \) and \( P \) relatively.

2) Candidate Enumeration: The candidate symbol vectors at each iteration are derived from \( \hat{s}_{ZF} \) by selecting the \( K \) nearest symbols in the modulation constellation set for each antenna, selecting \( K \cdot M \) symbol vectors [13]. Letting \( z^{(i)}(m, k) \) represents one of the enumerated candidates at \( i^{th} \) iteration, where \( m \) denotes the antenna index and \( k \) denotes the selected neighbor symbol index, then, only the \( m^{th} \) symbol in \( s_{ZF} \) is changed to its \( k^{th} \) neighbor in modulation constellation set. For each iteration, the starting solution is updated to \( s \) from last iteration.

The detection metric for TS measures the ML cost increment [13]

\[
t(m, c) = 2 \cdot R(e_m^* \cdot f_m^*) + |e_m^*| \cdot \tilde{H}(m, m)
\]

where \( e = z^{(i)}(m, k) - x^i \) is the differential vector between current starting symbol vector and the enumeration candidate, \( (\cdot)^* \) denotes the conjugate function and \( R(\cdot) \) denotes the real part of a complex value.

3) Tabu Move: As shown in Figure 2, \( K \cdot M \) candidate symbol vectors are judged and the starting symbol vector for the next iteration is selected by

\[
x^{(i+1)} = \arg \min_{j \in [1, K \cdot M]} e^{(i)} \notin TL (t_j).
\]

If (4) failed to find a suitable candidate, then the starting symbol vector remain the same as last iteration, hence \( x^{(i+1)} = x^{(i)} \).

4) Tabu List Updating: The TL is an integer matrix of size \( (M \cdot M_J) \times K \) [13]. Letting \( c \) denote the symbol index in modulation constellation set \( \Omega \) at \( m^{th} \) antenna for \( x^{(i+1)}(m, k) \), the position of \( x^{(i+1)}(m, k) \) in TL is at row \((m - 1) \cdot M_j + 1 \) and column \( k \). Hence, the update of TL is performed after each tabu move by assigning the element in TL for \( x^{(i+1)}(m, k) \) to \( P + 1 \) if a successful tabu move is found and decrementing all non-zero elements in TL by 1 [13]. \( f^{(i+1)} \) is given by

\[
f^{(i+1)} = f^{(i)} + e_{\bar{m}} \cdot \tilde{H}_m
\]

where \( \tilde{H}_m \) denotes the \( m^{th} \) column of \( \tilde{H} \).

These iterative steps are performed until the iteration threshold \( I \) is met. TS is promising for large scale MIMO due to the complexity against detection performance, nevertheless, no TS detection architecture for such systems are recorded to date.

Whilst TS-based detectors have been recorded for some time, there has been little record of their realisation. However, [3] points to a prominent future for these approaches as the size of antenna topologies increases. Specifically, it shows that when the number of antennas employed increases to around 40, the detection efficiency of TS-type algorithms is beyond that of linear detectors such as MMSE or SD algorithms. Specifically it shows that, whilst maintaining the same computational complexity, the Bit Error Rate (BER) achieved by TS is an order of magnitude greater than that enabled by MMSE, two orders of magnitude beyond that achieved by FSD and a factor of up to 3 greater than that enabled by MMSE with successive interference cancellation. However, these metrics do not translate directly to implementation cost and performance and as such the implementation efficiency of these detectors
is, as yet, unverified. In Sections III and IV we address this shortfall by comparing the detection performance and real-time performance and cost of TS and FSD-based detectors.

III. TS DETECTION - BEHAVIOUR AND PERFORMANCE

TS detection performs iterative candidate searching until reaching a final solution under the stopping criteria, which enables a trade-off between the detection performance and computational complexity by varying the number of iterations and neighbourhood size via the modulation scheme employed. By dividing the TS detection into three stages, Fig. 3 shows the dataflow diagram of Fig. 2 for three iterations and two neighboring candidates enumerations.

Fig. 3: TS Dataflow for \( I = 3, K = 2 \)

Enumeration produces \( K \) candidate symbol vectors for different antennas, \( NB \) performs the ML cost related calculations of candidate symbol vectors and the TabuMin checks the TL for each candidate and selects the candidate for next iteration. The dashed line box indicates the iterative part of TS detections which is flexible for more iterations and can be repeated for a given number. By changing the number of iterations and neighboring candidates, various detection performance metrics can be achieved.

The variation in computational complexity of FSD and TS are described in Fig. 4. This measures arithmetic complexity as the number of antennas varies between 4 and 40, and when the number of TS iterations \( I \) takes the values 16 or 300 (TS_{16} and TS_{300} respectively). The complexity of each TS stage is also illustrated in Fig. 1, where \( I \) is the number of iterations, \( N_b \) the number of enumerated neighbours, \( M_t \) the number of transmit antennas and \( M \) the constellation size. A number of trends are apparent: as anticipated by the linear increases in complexity with the number of TS iterations, TS_{300} generally exhibits complexity two orders of magnitude higher than TS_{16}. Further, the relative complexities of TS and FSD also varies with the number of antennas - initially FSD is the least complex, but its superlinear complexity growth with number of antennas means that it is more complex than TS_{16} and TS_{300} for more than 10 and 25 antennas respectively.

![Fig. 4: Complexity Comparison of TS and FSD](image)

**TABLE I: Fixed Iteration TS Detection Complexity**

<table>
<thead>
<tr>
<th>Enum</th>
<th>( \pm (x I \cdot M_t) )</th>
<th>( \times (x I \cdot M_t) )</th>
<th>Compare ( (x I \cdot M_t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>( 2(N_b - 1) + 5(N_b &gt; 1) )</td>
<td>( 6 + 5(N_b &gt; 1) )</td>
<td>( 6 + 3(N_b &gt; 1) )</td>
</tr>
<tr>
<td>Tabu Min</td>
<td>( 8N_b )</td>
<td>( 10N_b )</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>( 4 )</td>
<td>2N_b + ( \frac{2}{3} )</td>
<td></td>
</tr>
</tbody>
</table>

Given that the case where \( M = 10 \) antennas denotes the first point where FSD is more computationally demanding than TS_{16}, 4 \times 4 and 10 \times 10 MIMO topologies employing 4-QAM are chosen as comparison points. The BER performance of FSD, TS_{16}, TS_{300} and a linear MMSE detector are outlined in Fig. 5. As Figure 5a shows TS_{300} detection has close-to-FSD performance while TS_{300} detection has 3 dB lower SNR gain than FSD but over 10 dB larger SNR gain than MMSE for 4 \times 4 MIMO. Despite the much higher complexity of the TS schemes, FSD generally performs much better. A similar pattern is repeated in Fig. 5b, with the increased complexity of FSD translating to much superior BER performance. In particular it is notable that the performance of FSD is beyond that of TS_{16}, despite its relatively higher complexity.

Hence, although the TS detection does not achieve ML performance, as the antenna number grows the computational complexity of TS detection outperforms FSD with considerable BER performance. The next section describes the realization of TS detection on FPGA. Section IV studies the impact of these relative computational complexity variations of performance and cost for FPGA implementations.

IV. FPGA-BASED TABU SEARCH

**A. The FPE - an ASIP for FPGA Signal Processing**

The Xilinx-based FPE processor described in [4] and illustrated in Fig. 6 is used as the foundation for the proposed detection architectures. The FPE exploits a RISC load-store architecture with a seven-stage pipeline and enables very high performance with low cost relative to other FPGA soft processors by emphasising three key architecture design principles:
1) **Lean Architecture:** The FPE contains only components critical to ensuring software programmability: a Program Counter (PC) and Program Memory (PM), Instruction Decoder (ID), Register File (RF), Branch Detection (BD), Data Memory (DM) and an Arithmetic Logic Unit (ALU) based around the on-chip DSP48E slice. These are augmented only with components enabling high performance processing on FPGA at low resource cost: Immediate Memory (IMM) for ROM storage of constant data, an off-FPE Communication Module (COMM) and custom coprocessors if required; these features support an instruction set described in Table II.

2) **Scalability:** The FPE is designed to be combined in very high numbers. FPEs can be combined into SIMD FPE Processing Units (FPUs) to exploit data parallelism with the added benefit of amortizing the cost of several components across multiple FPEs (see Fig. 6b); furthermore, large-scale MIMD networks of FPUs communicating via FIFO queues of any size may be created for the application at hand.

3) **Configurability:** The FPE is highly configurable - almost every characteristic may be tuned to the application at hand; frequently components can be omitted if they are not used. Extending to enable configuration of the number of FPU ways, this configurability ensures that there is no barrier to achieving the absolute lowest cost architecture required for a specific application. The configurable characteristics of the FPU are described in Table III.

---

**TABLE II: FPE Instruction Set**

<table>
<thead>
<tr>
<th>Instruction</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTRL</td>
<td>branch if equal/greater/less</td>
</tr>
<tr>
<td></td>
<td>jump</td>
</tr>
<tr>
<td>DC</td>
<td>load/push data from/to channel</td>
</tr>
<tr>
<td></td>
<td>load data from/clear channels</td>
</tr>
<tr>
<td></td>
<td>no operation</td>
</tr>
<tr>
<td>ALU</td>
<td>multiply/add/subtract</td>
</tr>
<tr>
<td></td>
<td>multiply-add/subtract (&amp; forward)</td>
</tr>
<tr>
<td>MEM</td>
<td>load/store data from/to memory</td>
</tr>
<tr>
<td></td>
<td>load/store data from/to IMM</td>
</tr>
</tbody>
</table>

This combination of features is unique in FPGA-based processors and can enable uniquely powerful architectures; when realised on Xilinx Virtex 5 VLX110T FPGA, the computational capability and cost of six FPE configurations - 16 bit Real (16R), 32 bit Complex (32C) and 32 bit Real...
TABLE III: FPE Configuration Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>n_ways</td>
<td>FPU Width</td>
<td>1 - unlimited</td>
</tr>
<tr>
<td>data_ws</td>
<td>Data wordsize (bits)</td>
<td>16/32</td>
</tr>
<tr>
<td>data_type</td>
<td>Data type</td>
<td>Real/complex</td>
</tr>
<tr>
<td>alu_ndsp</td>
<td>No. DSP48E slices</td>
<td>1 - 4</td>
</tr>
<tr>
<td>pm_depth, pm_width</td>
<td>PM Capacity/width</td>
<td>Unlimited</td>
</tr>
<tr>
<td>imm_depth</td>
<td>IMM Capacity/width</td>
<td>Unlimited</td>
</tr>
<tr>
<td>dm_depth, rf_depth</td>
<td>DM/ RF Capacity</td>
<td>Unlimited</td>
</tr>
<tr>
<td>no_tx, no_rx</td>
<td>No. Tx/Rx ports</td>
<td>≤ 1024</td>
</tr>
</tbody>
</table>

(32R) variants - are as described in Table IV\(^1\). To the best of the authors’ knowledge, both the absolute performance and resource metrics quoted in Table IV are the leading metrics amongst recorded soft processors.

TABLE IV: FPE Arithmetic Performance

<table>
<thead>
<tr>
<th>Config</th>
<th>Resource</th>
<th>Latency (Cycles)</th>
<th>Clock (MHz)</th>
<th>Throughput (MMACs/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LUTs</td>
<td>DSP48Es</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 R</td>
<td>90</td>
<td>1</td>
<td>4</td>
<td>483</td>
</tr>
<tr>
<td>16 C</td>
<td>132</td>
<td>1</td>
<td>7</td>
<td>476</td>
</tr>
<tr>
<td></td>
<td>172</td>
<td>2</td>
<td>5</td>
<td>453</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>4</td>
<td>5</td>
<td>474</td>
</tr>
<tr>
<td>32 R</td>
<td>185</td>
<td>2</td>
<td>6</td>
<td>431</td>
</tr>
<tr>
<td></td>
<td>182</td>
<td>3</td>
<td>7</td>
<td>431</td>
</tr>
</tbody>
</table>

B. FPE-based Tabu-Search

The FPGA TS architecture for the TS\(_{16} 4 \times 4\) 4-QAM detector operating on 16-subcarrier OFDM shown in Fig. 7. As this shows, 32 16-way FPs and 92 banks of 16 FIFOs employed in a pipelined chain architecture\(^2\). The FPU behaviours are interleaved - odd-numbered FPs process the candidate search stage schedule illustrated in Fig. 8, including Enum and NVB, while even-numbered SIMDs perform candidate selection via Tabu min, with updated TL. Notice that a TL of size M \(\times M\) \(\times M\) is transferred between each SIMD for candidate selection.

Table ?? reports the performance and cost of the 4 \(\times\) 4 and 10 \(\times\) 10 TS architectures and the 10 \(\times\) 10 FSD architecture. The synthesis result of TS detection for 4 \(\times\) 4 MIMO with 4-QAM modulation on 16 OFDM subcarriers and both TS and FSD detection for 10 \(\times\) 10 on 8 OFDM subcarriers based on Xilinx Virtex-6 FPGA. The 10 \(\times\) 10 architectures reported in Table V are the largest recorded single-chip MIMO detectors on record.

Note the variation in performance and cost across the TS detectors and between the TS and FSD detectors. In particular, it is worth noting that the 10 \(\times\) 10 TS detector has higher cost than FSD of a similar scale, despite having lower computational complexity. Indeed, LUT and DSP48E1 cost have increased by factors of 2.9 and 1.3 respectively, whilst throughput has decreased by 47%. Hence throughput per unit LUT and DSP resource have decreased by 82.5% and 62.5% respectively. These reductions in absolute cost and performance and efficiency are compounded by reductions in detection capability, since FSD 10 \(\times\) 10 offers consistently superior BER performance relative to both TS\(_{16}\) and TS\(_{300}\).

It is probable that the cause of these relatively poor results is the maintenance of a single centralized TL and the iterative nature of TS. This has two major effects - the requirement to communicate this list between processing nodes has a high resource cost. Furthermore the iterative nature of the TS algorithm constrains throughput, since each stage may only be performed on completion of the last.

Accordingly it is apparent that, despite promising high performance detection at low computational complexity, in reality TS-based detection for large-scale MIMO systems has some way to go before it can be considered a practical reality. The lack of research into the embedded realisation of these algorithms, coupled with inherent throughput and data management constraints imposed by the iterative nature of TS and the maintenance of a centralized TL are significant bottlenecks which contrain the performance and increase the cost of current realisations.

V. Conclusion

As the scale of MIMO systems increases the balance of efficiency amongst families of detection algorithms changes. In particular, the work in [3] has shown that the efficiency of heuristic detectors, specifically TS detectors is beyond that of quasi-ML SD algorithms and even simple linear detectors offering, in some cases a 66% reduction in BER for the same computational complexity over MMSE-SIC. The gains relative to SD algorithms are even more spectacular. However, this paper shows that this gain in efficiency does not relate in a straightforward manner to a gain in implementation cost. Specifically, it has shown that, in the case of a 10 \(\times\) 10 MIMO system employing 16 OFDM subcarriers, implementation cost increases dramatically relative to even FSD - indeed when deployed on Xilinx Virtex®-6 FPGA, LUT and DSP48E1 cost increase by factors of 3.0 and 1.33 respectively, whilst performance reduces by 47.8%. It is apparent that the maintenance of a single global Tabu List has serious throughput implications. If the potential of TS algorithms as detectors in large scale MIMO systems is to be realised, substantial research is required to overcome this limitation.
Fig. 7: Architecture of TS $I = 16$, $K = 3$ in 4 x 4 MIMO OFDM System, 4-QAM

Fig. 8: Candidate Search Schedule

ACKNOWLEDGMENT

This work was sponsored by the UK Engineering and Physical Sciences Research Council (EPSRC) under contract number EP/H051155/1. The authors are grateful to Dr. Peng Wang for his assistance in deriving the FPGA multi-SIMD architectures.

REFERENCES


