Implementation of Parallel Encoder Architecture for Long Polar Codes

Monisha D1, Arul Karthick V J2

1(Department of ECE, Anna University, SNS College of Technology, Coimbatore, India, monishad76@gmail.com)
2(Department of ECE, Assistant Professor, SNS College of Technology, Coimbatore, India, karthick.arul@gmail.com)

Abstract— Polar codes represent an emerging class of error-correcting codes with power to approach the capacity of a discrete memory less channel. The main objective is to perform error correction and detection. The proposed new efficient encoder allows high-throughput encoding with small hardware complexity, it can be systematically applied to the design of any polar code and to any level of parallelism. The delay elements can be reduced by new parallel pipelined architecture. This particular architecture uses folding transformation technique as well as register minimization. Pipelining and parallel processing is used to reduce the power consumption.

Keywords— Polar codes, polar encoder, Very Large Scale Integration (VLSI) optimization

1. INTRODUCTION

Approaching capacity with a practical en/decoding complexity is a central challenge in coding theory. “Turbo-like” code families, such as turbo codes and low-density parity-check (LDPC) codes, have been found to achieve this goal. The key issue is how to practically implement the ideas used in the proof of the channel coding theorem. Coding randomness is introduced by interleavers in turbo codes or by pseudo-random connections between the variable and check nodes in LDPC codes. Among a few manuscripts dealing with hardware implementation, presented a straightforward encoding architecture that processes all the message bits in a fully parallel manner. The fully parallel architecture is intuitive and easy to implement, but it is not suitable for long polar codes due to excessive hardware complexity. For the first time, this brief analyzes the encoding process in the viewpoint of VLSI implementation and proposes a partially parallel architecture. The proposed encoder is highly attractive in implementing a long polar encoder as it can achieve a high throughput with small hardware complexity. Polar code is a linear block error correcting code developed by Erdal Arıkan. It is the first code to provably achieve the channel capacity for symmetric binary-input, discrete, memoryless channels. Polar codes were constructed using a generator matrix created using the Kronecker power of the base matrix. This paper is organized into V sections where section III represents the proposed folding transformation IV section represents register allocation.

2. EXISTING SYSTEM

Arıkan showed that SC decoding can be efficiently implemented by the factor graph of the code which has a structure that of the Fast Fourier transform. Fully parallel architecture for encoding a 16-bit polar code. The fully parallel encoder is designed based on the generator matrix, but implementing such an encoder becomes a significant burden when a long polar code is used to achieve a good error-correcting performance. In practical implementations, the memory size and the number of XOR gates increase as the code length increases. None of the previous works has deeply studied how to encode the polar code efficiently, although various tradeoffs are possible between the latency and the hardware complexity.

3. PROPOSED METHOD

The folding transformation is widely used to save hardware resources by time-multiplexing several operations on a functional unit. A data flow graph (DFG) corresponding to the fully parallel encoding process for 16-bit polar codes is shown in Figure 1 where a node represents the kernel matrix operation F, and wij denotes the jth edge at the ith stage. Note that the DFG of the fully parallel polar encoder is similar to that of the fast Fourier transform except that the polar encoder employs the kernel matrix instead of the butterfly operation. Given the 16-bit DFG, the 4-parallel folded architecture that processes 4 bits at a time can be realized with placing two functional units in each stage since the functional unit computes 2 bits at a time. In the folding transformation,
determining a folding set, which represents the order of operations to be executed in a functional unit, is the most important design factor. To construct efficient folding sets, all operations in the fully parallel encoding are first classified as separate folding sets. Since the input is in a natural order, it is reasonable to alternatively distribute the operations in the consecutive order. Thus each stage consists of two folding sets, each of which contains only odd or even operations to be performed by a separate unit. The folding sets of stage 2 have the same order as those of stage 1, i.e., \{B0, B2, B4, B6\} and \{B1, B3, B5, B7\}, since the four-parallel input sequence of stage 2 is equal to that of stage 1. Furthermore, to determine the folding sets of another stage s, the property that the functional unit processes a pair of inputs whose indices differ by 2s−1 is exploited. In the case of stage 3, two data whose indices differ by 4 are processed together, which implies that the operational distance of the corresponding data is two as the kernel functional unit computes two data at a time. For instance, w2,0 and w2,4 that come from B0 and B2 are used as the inputs to C0. Since both inputs should be valid to be processed in a functional unit, the operations in stage 3 are aligned to the late input data. Cyclic shifting the folding sets right by one, which can be realized by inserting a delay of one time unit, is to enable full utilization of the functional units by overlapping adjacent iterations. As a result, the folding sets of stage 3 are determined to \{C6, C0, C2, C4\} and \{C7, C1, C3, C5\}, where C6 in the current iteration is overlapped with A0 and B0 in the next iteration.

4. LIFETIME ANALYSIS AND REGISTER ALLOCATION

Let us consider the delay elements required in the folded architecture more precisely. When an edge wij from functional unit S to functional unit T has a delay of d, the delay requirements for wij in the F-folded architecture can be calculated as

\[ D(wij) = Fd + t - s \]

where t and s denote the position in the folding set corresponding to T and S, respectively. The delay requirements of the 4-folded architecture, i.e., \( D(wij) \) for \( 1 \leq i \leq 3 \) and \( 0 \leq j \leq 15 \), are. For instance, w2,0 from B0 to C0 demands one delay since \( d = 0 \), \( t = 1 \), and \( s = 0 \). Note that some edges indicated by circles have negative delays. For the folded architecture to be feasible, the delay requirements must be larger than or equal to zero for all the edges. Pipelining or retiming techniques can be applied to the fully parallel DFG in order to ensure that its folded hardware has nonnegative delays. Every edge with a negative delay should be compensated by inserting at least one delay element to make the value of (1) not negative. The two inputs of an operation pass through the same number of delay elements from the starting points. If they are different, additional delay elements are inserted to make the paths have the same delay elements. Four edges with zero delays are specially marked with negative zeros since additional delays are necessary due to the mismatch of the number of delay elements. The delay requirements are recalculated based on units and 48 delay elements in total are enough to implement the 4-parallel 4-folded encoding architecture based on the folding sets. The lifetime analysis is employed to find the minimum number of delay elements required in implementing the folded architecture[13]. The lifetime of every variable is graphically represented in the linear lifetime chart illustrated in Figure 3.

<table>
<thead>
<tr>
<th>( j )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D(w_{ij}) )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( D(w_{ij}) )</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( D(w_{ij}) )</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 2: Delay Requirement calculation for 16bit DFG

Figure 3: Lifetime chart analysis

Since all the edges starting from stage 1 demand no delay elements, only w2j and w3j are present in For instance, w3,0 is alive for two cycles as it is produced at cycle 1 and consumed at cycle 3. The number of variables alive in each cycle is presented at the right side of the chart. The number of live variables at the fourth or later clock cycles takes into account the next iteration overlapped with the current iteration. Consequently, the maximum number of live variables is 12, which means that the folded architecture can be implemented with 12 delay elements instead of 48. Once the minimum number of delay elements has been determined, each variable is allocated to a register. The register allocation is tabulated in Figure 4. In the register allocation table all the 12 registers are shown at the first row, and every row describes how the registers are allocated at the corresponding cycle. With taking into account the 4-parallel processing, variables are carefully allocated to registers in a forward manner. An arrow indicates that a variable stored in a register is migrated to another register, and a circle indicates that the variable is consumed at the cycle. For example, w2,0 and w2,4 are consumed in a functional unit to execute operation C0 that generates w3,0 and w3,4. At the same time, w2,1 and w2,5 are consumed in another functional unit to execute operation C1 that produces w3,1 and w3,5. The migration of the other variables can be traced by following the register allocation table. Finally, the resulting 4-parallel pipelined structure proposed to encode the 16-bit polar code is illustrated in Figure 5 which consists of 8 functional units and 12 delay elements. A pair of two functional units takes in charge of one stage, and the delay elements are to store variables according to the register allocation table. The hardware structures for stages 1 and 2 can be...
The proposed architecture continuously processes four samples per cycle according to the folding sets and the register allocation table. The proposed encoder takes a pair of inputs in a natural order and generates a pair of outputs in a bit-reversed order. As the functional unit in the proposed architecture processes a pair of 2 bits at a time, the proposed architecture maintains the consecutive order at the input side and the bit-reversed order at the output side if a pair of consecutive bits is regarded as a single entity.

5. CONCLUSION

In the proposed architecture, the number of functional units required in the implementation depends on the code length $N$ and the level of parallelism $P$. Since a functional unit can save the hardware by up to 73% compared with that of the fully parallel architecture. Finally, the relationship between the hardware complexity and the throughputs is analyzed to select the most suitable architecture for a given application. Therefore, the proposed architecture provides a practical solution for encoding a long polar code.

REFERENCES


