Abstract

To meet with the stringent requirements of ultra-low latency communication in 5G, this work presents a polar decoder fabricated in TSMC 40nm CMOS featuring: 1) World’s first neural network-assisted decoder chip with 8x improvement of convergence rate. 2) Fully reconfigurable architecture to support multi-code length operations with a 2-to-8x hardware utilization rate. 3) Optimized fixed-point design of processing element (PE) to reduce 73% area and 67% power consumption.

Introduction

In 2016, polar code was officially adopted by 3GPP as the coding scheme in 5G New Radio due to its capacity-achieving performance [1]. The decoding algorithms are divided into two categories: Successive Cancellation (SC) and Belief Propagation (BP). In general, SC suffers from high latency and limited throughput due to its inherent sequential processing [2-3]. On the other hand, BP can achieve a much higher throughput rate with its parallelized architecture [4]. However, it still suffers from high latency due to the iterative process and unfavorable performance within limited iterations [5].

Fig. 1 depicts the factor graph of a conventional BP [4] and our adopted recurrent neural network-assisted BP (RNN-BP) algorithm [5]. For an \((N, K)\) polar decoder, there are \(n = \log_2 N\) stages and a total of \(N (n+1)\) nodes in the factor graph. In BP, the received channel log-likelihood ratios (LLRs) are iteratively passed from left to right and then right to left through PEs for computing information bits. In [5], the authors unfolded the BP structure of Fig. 1(a) and converted it into a recurrent structure with trainable weights assigned on each connection as shown in Fig. 1(b). After offline training, the weights of the BP network are used as the scaling factors associated with wire connections of the factor graph. At comparable BER performance, this intelligently trained RNN-BP architecture can converge faster within only 5 iterations compared with 40 iterations of conventional BP as shown in Fig. 5. Therefore, it leads to 8x convergence speed while inheriting BP’s advantages of low complexity and high parallelism. However, [5] only shows experimental results without any hardware design. We are the first to implement the RNN-BP chip in 40nm with several architectural innovations.

Reconfigurable Multi-Code Length Support

Fig. 2 shows our polar decoder chip design for multi-code length support. Suppose each code length of BP employs different sets of weights to implement the corresponding RNN-BP, it suffers from severe memory overhead. However, note that the shorter code length is embedded in the longer one due to the regular structure of BP. As a result, the weights of shorter code lengths can be “co-shared” with longer codes. Following this design concept, at the algorithmic level, we propose a new multi-stage training process. The well-trained weights of the shorter code length are set as “frozen” weights and initially assigned to the corresponding positions of longer code lengths. Then, the remaining non-frozen weights can be updated/trained for a longer code length. This weight-sharing mechanism helps to reduce memory overhead by 41%, but with only a slight loss in performance. At the architectural level, for the case of \(N = 8\), the processing engine consists of 4 PEs and can be reconfigured to support \(N = 4\) with 2 PEs and \(N = 2\) with 1 PE, respectively. Hence, two kinds of “turbo” modes: “two parallel 4-bit decoders” and “four parallel 2-bit decoders” can be supported with the reconfigured architecture to increase hardware utilization rate by 2x or 4x. As a result, with a 128-PE design, the RNN-BP can be fully reconfigured to support multi-code length of \(N = 32, 64, 128,\) and 256. The fully running PE array improves system throughput rate by 8-to-2x.

Processing Element Design and Optimization

Fig. 3 shows the overall block diagram of our chip. In this work, we designed a 256-bit polar decoder that consists of 128 PEs. This PE array can compute messages of each stage in one cycle. Two memory banks, “message memory” and “weight memory,” are used to store the computed messages and pre-trained weights, respectively. The control unit is carefully designed to control the memory read/write and data flow of PE array for supporting multi-code length.

Fig. 4 depicts the detailed PE design with two arithmetic-level optimization techniques. 1) Specialized data number representation: The function \(g()\) inside PE consists of one absolute operator and one sign operator. Hence, sign-magnitude representation becomes a perfect choice to realize \(g()\) than 2’s complement. 2) Mechanism of dynamic overflow detection: Because the magnitude of computed messages indicates the likelihood value of being 0 or 1, the occurrence of the overflow will reverse its likelihood value and severely degrade error-correction performance. By adding one saturation operator at outputs of each adder and multiplier, the relative likelihood ratio can be retained for overflow-free operations without affecting the decoding performance, and reduces the required wordlength from 8-bit to 5-bit. These two arithmetic-level optimizations can jointly reduce PE’s area and power consumption by 73% and 67%, respectively.

Measurement Results and Conclusion

Table I summarizes the comparison between the state-of-the-art BP chip and our RNN-BP chip. Fig. 5 shows the chip measurement results. From the comparison chart, we demonstrate that the RNN-BP chip can achieve 8x convergence speed in terms of iteration number at algorithmic level. At room temperature, the chip operates at a maximum frequency of 240MHz with 0.98W supply and consumes 13.8W. Fig. 6 shows the Shmoo measurements and the micrograph and summary of the fabricated chip.

Table II lists the comparisons with state-of-the-art designs. Firstly, our reconfigurable design can support multi-code length and other designs have fixed code length. The measured energy efficiency can be as low as 7.8 to 13.6 pJ/b in supporting multiple code lengths from \(N = 32\) to \(N = 256\), benefiting from 8x convergence rate and aforementioned optimization techniques. After normalization for a fair comparison, the proposed RNN-BP decoder can achieve 3.76 to 6.58 Gb/s for high-performance channel decoding. Operated at a frequency of 225MHz, our multi-code-length design has 2.4x, 2.3x, and 10.0x enhancement in terms of latency, throughput rate, and energy efficiency, respectively, compared with the state-of-the-art single-mode BP design in [4]. On the other hand, compared with the SC-based design in [2], the RNN-BP design also has 8.7x and 8.6x improvement in latency and throughput at comparable energy and area efficiency.

In summary, as the first fabricated NN-assisted decoder, this chip demonstrates ultra-low latency (8x reduction in total iterations), and shows 10.0x improvement in energy efficiency over prior designs. It also demonstrates the great potential of DL-assisted DSP engine in 5G New Radio communication systems.

Acknowledgment

We thank TSRI, Taiwan, for chip fabrication and support.
Table I. Comparison between conventional BP and RNN-BP designs.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Convergence Speed</td>
<td>40 iteration</td>
</tr>
<tr>
<td>Operation per PE</td>
<td>4 adds</td>
</tr>
<tr>
<td>Memory</td>
<td>Shared Memory (1152b)</td>
</tr>
</tbody>
</table>

References