In this paper, we propose and present implementation results of a high-speed turbo decoding algorithm. The latency caused by (de)interleaving and iterative decoding in a conventional maximum a posteriori turbo decoder can be dramatically reduced with the proposed design. The source of the latency reduction is from the combination of the radix-4, center to top, parallel decoding, and early-stop algorithms. This reduced latency enables the use of the turbo decoder as a forward error correction scheme in real-time wireless communication services. The proposed scheme results in a slight degradation in bit error rate performance for large block sizes because the effective interleaver size in a radix-4 implementation is reduced to half, relative to the conventional method. To prove the latency reduction, we implemented the proposed scheme on a field-programmable gate array and compared its decoding speed with that of a conventional decoder. The results show an improvement of at least five fold for a single iteration of turbo decoding.

Keywords: Turbo code, radix-4, center-to-top, parallel decoding, early-stop, FPGA.

I. Introduction

For wireless communications, channel coding is an important tool for improving communications reliability. In satellite communications and 3rd generation mobile communication system, turbo codes have been adopted for high-speed data transmissions. Turbo codes, introduced in [1] and [2], have been shown to perform near the capacity limit in additive white Gaussian noise (AWGN) channels. Such a powerful coding technique offers great promise for improving the reliability of communication over wireless channels. Wireless communication systems are moving today from the conventional narrow-band voice service to the wide-band multimedia service requiring high-speed turbo decoders. Important issues in high-speed applications of turbo decoders are decoding delay and computational complexity. To solve the latency problem, in this paper four decoding algorithms are presented and combined into one decoder architecture. The first algorithm is the radix-4 decoding algorithm, where the previous state at \( t = k - 2 \) goes forward to the current state at \( t = k \), and the reverse state at \( t = k + 2 \) goes backwards from the current one such that the time interval from \( t = k - 2 \) to \( t = k \) is merged into \( t = k \). Therefore, we can decode two source data bits at the same time without any performance degradation while reducing the block size buffered in memory. In order to apply the radix-4 algorithm, we derived the branch metric (BM), forward state metric (FSM), \( \alpha \), backward state metric (BSM), \( \beta \), and log likelihood ratio (LLR), \( \lambda \).

The second algorithm is the center-to-top algorithm. In a conventional scheme, the decoder must wait for finishing the BSM (or FSM) calculations before calculating the LLR. The center-to-top method doesn’t need to wait. The decoder...
calculates the FSM (left to right) and BSM (right to left), simultaneously. When the FSM and BSM reach the same point, then the decoder begins to calculate the LLR.

The third algorithm is the full parallel decoding algorithm. Different from the original turbo decoder consisting of two decoders concatenated in a serial fashion, we propose a parallel decoder structure using the parallel sum, where the two decoders operate in parallel and update each other immediately and simultaneously after each one has completed its decoding. In decoding the estimated data, we use the sum of the LLR outputs of the parallel decoders to reduce the latency to half while maintaining the same performance level.

The fourth algorithm is the hard-decision-aided (HDA) algorithm implementing the early-stop algorithm. It compares each decision generated by the two decoders, and when the two sets of decisions match, it stops decoding the current block and outputs the hard decision bits.

These schemes are desirable in designing a high-speed turbo decoder because they provide significant reductions in decoder memory requirements as well as allowing increased parallelism. To verify the high-speed feature of the decoder, we implemented the proposed combined schemes in a field-programmable gate array (FPGA), Altera FLEX10K, and compared its decoding speed to a conventional design. The computer simulation results of the new high-speed turbo decoder because they provide significant reductions in decoder memory requirements as well as allowing increased parallelism.

**II. High-Speed Decoding Algorithm**

1. **Scheme 1: Radix-4 Algorithm**

   Using the unified approach to state metrics, a $2^{k-1}$-state trellis can be iterated from time index $n-k$ to $n$ by decomposing the trellis into $2^{k-1}$ sub-trellises, each consisting of $k$ iterations of a $2^k$-state trellis. Each $2^k$-state’s sub-trellis can be collapsed into an equivalent one-stage radix-2$^{k}$ trellis by applying $k$ levels of look-ahead to the recursive update. Collapsing the trellis does not affect the decoder performance since there is a one-to-one mapping between the collapsed trellis and radix-2 trellis. An example of the decomposition of a 4-state radix-2 into an equivalent radix-4 trellis using one stage of look-ahead is shown in Fig. 1, where $K=3$, $g_1=(111)$, $g_2=(101)$ with $K$ denoting the constraint length.

   For the radix-4 trellis iterations, time $k$ to $k-2$, the backward state for calculating $\alpha_k^n$, and time $k+2$ to time $k$, the forward state for calculating $\beta_k^n$, can be expressed as

   
   \[
   m_{k-2}(d_{k-1},d_k,m_k) = (d_{k-1} \oplus d_k \oplus m_{0,k}) || (d_k \oplus m_{0,k} \oplus m_{k,1}) \\
   m_{k+2}(d_{k+1},d_{k+2},m_k) = (d_{k+1} \oplus d_{k+2} \oplus m_{1,k}) || (d_{k+2} \oplus m_{1,k} \oplus m_{k,3}),
   \]

   where $m_k$ and $m_0$ denote the uncoded data bit state values at time $k$, and $a||b$ denotes the concatenation of $a$ and $b$. We have that $m_k = [m_{0,k}, m_{1,k}, \ldots, m_{k-1,k}, m_{k+1,k}]$ corresponds to the encoder state with $m_{0,k}$ used for calculating $\alpha_k^n$ and $m_{k+2}$ used for calculating $\beta_k^n$ at time $k$.

   **A. Derivation of $\delta_k^{m,n}, \alpha_k^n$ and $\beta_k^n$**

   We consider an AWGN channel with zero mean and variance $\sigma^2$. In a radix-4 MAP decoder, the branch metric $\delta_k^{m,n}$ for the received signal at time $k$, can be expressed as

   \[
   \delta_k^{m,n} = \Pr(D_k = p, S_k = m_k, R_k) = \Pr(D_k = i_{k-1} || i_k, S_k = m_k, R_k) = \Pr(d_{k-1} = i_{k-1}, S_k = m_k, R_k) = \Pr(x_{k-1} = d_{k-1}, S_{k-1} = m_{k-1}) \cdot \Pr(y_{k-1} = d_{k-1} = i_{k-1}, S_{k-1} = m_{k-1}) \cdot \Pr(x_k = d_k = i_k, S_k = m_k) \cdot \Pr(y_k = d_k = i_k, S_k = m_k) \cdot \delta_k^{i_k} / 2^r
   \]

   \[
   = K_i \exp \left( \frac{2}{\sigma^2} (x_{k-1}d_{k-1} + y_{k-1}Y_{k-1} + x_kd_k + y_kY_k) \right),
   \]

   where $i_k$ is the information bit value at time $k$, and $p = i_{k-1} || i_k$. Therefore, we have $i_k = \{0, 1\}$ and $p = \{00, 01, 10, 11\}; K_i$ is a constant; $x_k$ and $y_k$ are received in-phase and quadrature signals; and $Y_k$ is the coded bit as a function of input bit $d_k$ and encoder
state $m_k$. The forward state metric at time $k$ and state $m$ denoted by $a^m_k$ can be expressed as

$$ a^m_k = \Pr(R_{k-1}^- | D_k = p, S_k = m_k, R_k^N) = \Pr(R_k^N | S_k = m_k) $$

$$ = \sum_{w}^{\varphi} \sum_{p=0}^{k} \Pr(D_{k-2} = p, S_{k-2} = m', R_k^{N-2} | S_k = m_k) $$

$$ = \sum_{w}^{\varphi} \sum_{p=0}^{k} \Pr(D_{k-1} = p, S_{k-1} = m', R_k^{N-4}, R_{k-1} | S_k = m_k) $$

(3)

where $R_k^N = \{R_k, R_{k+1}, ..., R_n\}$, and $b(p, m_k)$ is the previous $k-2$ time stage given input $p$ and state $m_k$, that is,

$$ b(p, m_k) = m_{k-2}(d_{k-2}, d_k, m_k). \tag{4} $$

In a similar way, after the whole block of data is received, we can recursively calculate probability $b^m_k$ from probability $b^m_{k+2}$, that is,

$$ b^m_k = \Pr(R_k^N | S_k = m_k) $$

$$ = \sum_{w}^{\varphi} \sum_{p=0}^{k} \Pr(D_k = p, S_{k+2} = m', R_k^{N+2}, R_k^N | S_k = m_k) $$

$$ = \sum_{w}^{\varphi} \sum_{p=0}^{k} \Pr(D_k = p, S_{k+2} = m', R_k^{N+2}, R_k^N | S_k = m_k) $$

(5)

where $f(p, m_k)$ is the next $k+2$ time stage given input $p$ and state $m_k$. With (1), $f(p, m_k)$ is expressed as

$$ f(p, m_k) = m_{k+2}(d_{k+1}, d_{k+2}, m_k) \tag{6} $$

when taking into account the fact that events after time $k$ are not influenced by observations up to time $k$, that is, $R_k^N$ and $R_{k+2}^N$ are independent. Finally, with (2), (5) and (6), the LLR of the radix-4 algorithm can be written as

$$ \lambda^m_k = \Pr(D_k = p, S_k = m | R_k^N) $$

$$ = \Pr(D_k = p, S_k = m, R_k^N) / \Pr(R_k^N) $$

(7)

In order to decode the two bits, $D_k = d_{k-1} \parallel d_k$, the following equation can be used:

$$ D_k = \max\left\{ \sum_{w}^{\varphi} \lambda^m_k(m), \sum_{w}^{\varphi} \lambda^0_k(m), \sum_{w}^{\varphi} \lambda^1_k(m) \right\}. \tag{8} $$

The overall configuration of the radix-4 Turbo MAP decoder is depicted in Fig. 2. $L_k(x^0_k), L_k(x^1_k), L_k(x^0_{k+1}), L_k(x^1_{k+1})$ and $L_k(d^0_k), L_k(d^1_k), L_k(d^2_k), L_k(d^3_k)$ denote the extrinsic information for correction at next MAP module. It is composed of two constituent decoders, MAP1 and MAP2, concatenated serially with symbol interleavers. Each MAP decoder calculates four LLRs, $\lambda^0_k, \lambda^1_k, \lambda^0_{k+1}$, and $\lambda^1_{k+1}$, where $\lambda^1_k$ corresponds to the LLR value for two input bits '00'.

2. Scheme 2: Center-to-Top Algorithm

Another algorithm is the center-to-top (CT) algorithm. In
conventional algorithms the decoder must wait for finishing the BSM (or FSM) calculations to calculate the LLR. The CT algorithm doesn’t need to wait. The decoder calculates the FSM (left to right) and BSM (right to left), simultaneously. When the FSM and BSM reach the same point, the decoder begins to calculate the LLR. Figure 3 illustrates the operation of the CT algorithm. The procedure of the CT algorithm is as follows.

Step 1: Initialize the forward state metric and backward state metric.

\[
\alpha_i^j(s_i^j(m)) = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{otherwise} \end{cases} (9)
\]

\[
\beta_i^j(s_i^j(m)) = \begin{cases} 1, & \text{if } m = 0 \\ 0, & \text{otherwise} \end{cases}
\]

Step 2: After receiving the whole set of received symbols, forward state metrics (left to right) and backward state metrics (right to left) are calculated simultaneously.

\[
\alpha_i^k(m) = \exp\left(\frac{2}{\sigma^2}(x_i^k + y_i^kY_i^k(m))\right) \sum_{j=0}^{k-1} \alpha_i^j(S_i^j(m)) \#(10)
\]

\[
\beta_i^k(m) = \sum_{j=k}^{N} \beta_i^j(k) \exp\left(\frac{2}{\sigma^2}(x_i^j + y_i^jY_i^j(j,S_j^j(m)))\right) \#(11)
\]

for \( k = 0, \ldots, N/2 - 1 \),

\[
\alpha_i^k(m) = \sum_{j=0}^{k-1} \alpha_i^j(k) \exp\left(\frac{2}{\sigma^2}(x_i^j + y_i^jY_i^j(j,S_j^j(m)))\right) \#
\]

\[
\beta_i^k(m) = \sum_{j=k}^{N} \beta_i^j(k) \exp\left(\frac{2}{\sigma^2}(x_i^j + y_i^jY_i^j(j,S_j^j(m)))\right)
\]

for \( k = N - 1, \ldots, N/2 \).

Step 3: At the middle point, begin to calculate the LLR.

\[
L(d_k)_{right} = \log \frac{\sum_m \alpha_i^k(m)\beta_i^k(m)}{\sum_m \alpha_i^0(m)\beta_i^0(m)} \text{ for } k = N/2, \ldots, N-1 \quad (12)
\]

\[
L(d_k)_{left} = \log \frac{\sum_m \alpha_i^k(m)\beta_i^k(m)}{\sum_m \alpha_i^0(m)\beta_i^0(m)} \text{ for } k = N/2 - 1, \ldots, 0 \quad (13)
\]


The turbo decoding configuration operates in serial mode, that is, “MAP 1” processes data before “MAP 2” starts its operation, and so on. Another possibility, as presented in [4] and shown in Fig. 4 for two codes, is that all decoders operate in parallel at any given time.

4. Scheme 4: Early Stop Algorithm

The hard decision (HAD) algorithm is used as an early-stop algorithm. It compares each decision generated by the two decoders, and when the two sets of decisions match, it stops decoding on the current block and outputs the hard decision bits. Table 1 shows the average number of iterations in an HAD algorithm. At an \( E_b/N_0 \) of 2 dB, it requires about 2.8 iterations relative to 8 for a given performance. This means that the decoding speed is improved or the power consumption (cost) is reduced by 64.6%.

<table>
<thead>
<tr>
<th>( E_b/N_0 )</th>
<th>Average number of iterations</th>
<th>Decoding speed improvement (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.21</td>
<td>22.4</td>
</tr>
<tr>
<td>1.5</td>
<td>4.31</td>
<td>46.1</td>
</tr>
<tr>
<td>2</td>
<td>2.83</td>
<td>64.6</td>
</tr>
</tbody>
</table>
III. Computer Simulation Results

The bit-error rate (BER) performance of the new high-speed turbo decoder architecture combining the four schemes is analyzed in this section. For a comparison purpose, Fig. 5 shows the performance of the new decoder and a conventional one using K=3 turbo codes with generator polynomials \( g_1=(111), g_2=(101) \) as a function of interleaver size \( N \) and as a function of the number of iterations \( I \). In the radix-4 method, the symbol interleaver, takes an information stream of length \( N_s \) composed of 2-bit words and feeds it to a random interleaver. From the figure, it can be verified that the performances of the proposed decoder architecture is very close to the conventional decoder for small block sizes (less than 300 bits). For large interleaver sizes (more than 300 bits), the performance of the new decoder is slightly degraded relative to the conventional one because the randomness of the symbol interleaver is reduced. Therefore, it can be concluded that the proposed decoder reduces the decoding latency while maintaining almost the same performance of a conventional algorithm.

IV. Design of High-Speed Turbo Decoder

A schematic diagram of the high-speed turbo decoder implementation is shown in Fig. 6(a). A detailed block diagram of the radix-4 and CT decoder, MAPI and MAP2, in this figure is shown in Fig. 6(b). In Fig. 6(a), MAP 1 and MAP 2 decoders are operated in parallel, with their outputs being log-likelihood ratios of input bits as shown in Fig. 6(b). For
instance, \( LLR_{00} \) denotes log-likelihood ratios of input bits “00”, and the arrow \( \rightarrow \) denotes the direction of the LLR calculation, that is, left-to-right. To add the extrinsic information to the received symbols exactly in the next iteration, the decoder needs to re-order the blocks. This re-ordering of the blocks is equivalent to arranging the LLRs in a last-in, first-out or first-in, first-out fashion. As shown in Fig. 6(b), the decoder consists of four major units, the radix-4 branch metric unit (R4BMCU), radix-4 backwards state metric unit (R4BSMCU), radix-4 forwards state metric unit (R4FSMCU), and radix-4 for LLR calculator unit (R4LLRCU). The radix-4 algorithm is as described in subsection II.1 and needs four R4BMCUs, two R4FSMUs/ R4BSMUs, and four R4LLRCUs. After receiving the whole set of received symbols, the quantized I and Q samples are fed to the R4BMCU. The branch metrics between the branch codeword “0000” and the received symbols is denoted by \( bm_{00} \). The R4BMCU calculates the branch metrics for four samples of received data. To apply the CT algorithm, the R4BMCU1 calculates the branch metrics in the direction of left-to-right. R4BMCU2 calculates the branch metrics in the direction of right-to-left. Each branch metric is buffered in memory to be used for R4BMCU/R4FSMCU. In calculating the FSM (or BSM) in the interval from 0 to \( N/2 - 1 \) (or \( N-1 \) to \( N/2 \)), we don’t need to calculate the branch metrics but simply refer to the memory of R4BMCU2 (or R4BMCU1). R4FSMCU, R4BSMCU, and R4LLRCU need an E-Table. The argument of the exponential function, that is, the input of an E-Table, was designed to be equal to the address lookup table (LUT). The E-Tables are implemented using an internal ROM (EAB, embedded array block) without using any external ROM.

In order to implement optimally the high-speed decoder, we determined the optimum quantized bits of each block shown in Fig. 6, the demodulator output, in-phase and quadrature signals, R4BMCU output of \( b_{op} \), R4BSMCU output of \( s_{q} \) and log-likelihood ratio \( l_{p} \). By fixed-point computer simulation, the output of the demodulator is quantized to 8 bits. The internal parameters of the turbo decoder were always saturated to 9 bits, and the optimum quantization bits of the turbo decoder derived from the fixed-point simulations are listed in Table 2.

The results of the timing simulation of the high-speed turbo decoder are shown in Table 3. The interleaver size, \( N \), was set to 64 bits due to the limited capacity of the embedded memory of Flex10k100GC503-4. The results from iterative decoding over an AWGN channel at \( E_b/N_0 = 1 \) dB is also presented in Fig. 7. Since the error correcting performance increases with the number of iterations, we concluded that the high-speed decoder implemented in this test-bed of Flex10k100GC503-4, pictured in Fig. 8, is performing well.

In order to compare the decoding speed of a conventional turbo decoder to the proposed one, we implemented a conventional decoder using the same procedures. The required maximum clock of the device for the conventional and proposed decoders was 44 ns and 58 ns, respectively, based on the number of required logics and memories. The total delay latency of the proposed decoder is 27 ns for a single iteration, which is five times faster than that of a conventional algorithm, demonstrating the value of combining various latency improvement schemes into a single decoder design and architecture.

Table 2. The optimum quantized bits of the turbo decoder.

<table>
<thead>
<tr>
<th>Block</th>
<th>Number of optimized quantization bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>( I )</td>
<td>8,8</td>
</tr>
<tr>
<td>( Q )</td>
<td>9</td>
</tr>
<tr>
<td>( s_q )</td>
<td>9</td>
</tr>
<tr>
<td>( l_p )</td>
<td>9</td>
</tr>
</tbody>
</table>

Fig. 6. Schematic diagram of proposed turbo decoder.
V. Conclusion

To extend the application area of turbo decoding to real time services, it is important to reduce the latency in the decoding process of turbo decoders. In this paper, we proposed a new high-speed turbo decoding algorithm and architecture that provides a solution. Two new algorithms are presented, radix-4 and CT algorithms that are combined to two other techniques to reduced latency. The different decoding algorithms are incorporated into a decoder design architecture to show how much the decoding latency can be reduced. Fixed on N=64 and I=1, the results of the implementation in a Flex10k100GC503-4 confirmed that the decoding speed of the proposed decoder is superior to conventional algorithms. The error-correcting performance of the proposed algorithm is the same as that of the conventional one in the case of small block sizes (less than 300 bits) but is slightly degraded for larger interleaver sizes because the randomness of the symbol interleaver is reduced. Given that in the case of voice and data transmission in wireless communications the frame size tends to be very small, this decoder architecture is suitable to high-speed wireless communications.

References


Ji Won Jung received his BS in 1989, MS in 1991, and PhD in 1995 from Sungkyunkwan University, Seoul, Korea, all in electronics engineering. From Nov. 1990 to Feb. 1992, he was with LG Research Center, Korea. From Sept. 1995 to Aug. 1996, he was with Korea Telecom (KT). From Aug. 2001 to July 2002, he was an Invited Researcher at the Communication Research Center Canada supported by NSERC. Since 1996, he joined the Department of Radio Science and Engineering, Korea Maritime University, Busan, Korea. His research interests are channel coding, digital modem, FPGA design technology, and digital broadcasting system.

In Ki Lee received his BS in 2003, MS in 2005, from Korea Maritime University, Busan, Korea. Since July 2005, he joined the Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea, where he is currently a Researcher. His main research interests are channel coding, digital modem, FPGA design technology, and digital broadcasting system.

Duk Gun Choi received his BS in 2004, from Korea Maritime University, Busan, Korea. His research interests are channel coding, digital modem, FPGA design technology, and digital broadcasting system.

Jin Hee Jeong received BS degree in 2005 from Korea Maritime University, Busan, Korea. She is in Korea Maritime Graduate University. Her research interests are channel coding, digital modem, and FPGA design technology.

Ki Man Kim (M'95) received his BS in 1988, MS in 1990, and PhD in 1995 from Yonsei University, Seoul, Korea, all in electronics engineering. From 1995 to 1996, he worked for Yonsei Medical Center as a fellow, where he was involved in developing medical diagnosis equipments. In 1996, he joined the Department of Radio Science and Engineering, Korea Maritime University, Busan, Korea, where he is currently an Associate Professor. His main research interests include speech enhancement, microphone array, synthetic aperture sonar, virtual sound, and hardware implementation by using digital signal processor/CPLD.

Eun-A Choi received the BS degree in mathematics and the MS degree in information communications engineering from the Chonbuk National University (CNU), Chonbuk, Korea, in 1998 and 2000, respectively. She has been with the Digital Broadcasting Research Division, ETRI, Daejeon, Korea, as a Member of Research Staff since her graduation from CNU. In ETRI, she has been involved with Development Projects of Broadband Adaptive Satellite Communications and Broadcasting Convergence System, where she develops design algorithms for variable channel codecs. Her research interests include channel codes, digital communications, satellite communications and satellite broadcasting.

Deock Gil Oh received his BS degree in Electronics Engineering Department from Seoul National University (SNU) in 1980. He received the MS and PhD degrees in electronics engineering from SNU, Korea, in 1984 and 1996, respectively. He joined ETRI in 1982, where he is currently working as a Team Leader in Digital Broadcasting Research Division. His research interests include Wireless Access Technology, Mobile Communication and Broadcasting System and Future Generation Wireless Access System.