A 0.077 to 0.168 nJ/bit/iteration Scalable 3GPP LTE Turbo Decoder with an Adaptive Sub-Block Parallel Scheme and an Embedded DVFS Engine

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

Citation: Chih-Chi Cheng et al. “A 0.077 to 0.168 nJ/bit/iteration Scalable 3GPP LTE Turbo Decoder with an Adaptive Sub-block Parallel Scheme and an Embedded DVFS Engine.” 2010 IEEE Custom Integrated Circuits Conference (CICC), 2010. 1–4. © Copyright 2012 IEEE

As Published: http://dx.doi.org/10.1109/CICC.2010.5617396

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/72198

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.
A 0.077 to 0.168 nJ/bit/iteration Scalable 3GPP LTE Turbo Decoder with an Adaptive Sub-Block Parallel Scheme and an Embedded DVFS Engine

Chih-Chi Cheng*,†, Yi-Min Tsai†, Liang-Gee Chen† and Anantha P. Chandrakasan*

*Massachusetts Institute of Technology, Cambridge, MA
†National Taiwan University, Taipei, Taiwan

Abstract—3GPP LTE requires a 100 Mbps of peak bandwidth, and the instantaneous throughput demand changes with different applications. Fixed sub-block parallel turbo decoding scheme introduces bit-error rate (BER) performance drop when the block length is short. In this paper, an LTE turbo decoder implemented on a 0.66 mm$^2$ die in a 65 nm CMOS technology is presented. An adaptive sub-block parallel (ASP) decoding scheme that improves the BER performance by up to 2.7 dB while maintaining the same parallelism is developed. A DVFS engine combining with an early-termination scheme is also developed. It generates the supply voltage and the clock rate that lead to the lowest energy consumption given the output bandwidth requirement. The measured energy consumption is 0.077 ̄ 0.168 nJ per bit per iteration and 0.39 ̄ 0.85 nJ per bit.

I. INTRODUCTION

3GPP long-term evolution (LTE) is an emerging 4G wireless technology. LTE channel coding features a 100 Mbps peak data rate and 188 modes with code block length ranging from 40 to 6144 [1]. The overall physical layer throughput is estimated to be 60 Mbps [2].

Sub-block parallel decoding scheme is widely used in LTE turbo decoders to meet the high throughput requirement [3]–[5]. In an $N$ sub-block parallel decoding scheme, one code block is divided into $N$ equal-lengthed sub-blocks, and the sub-blocks are decoded in parallel. Due to the contention-free property of the LTE interleaver [6], memory access collision could be avoided.

However, the sub-block parallel scheme suffers from the bit-error rate (BER) performance degradation. Figure 1 shows the BER performance comparison of the algorithm [7] implemented without parallelism and with eight sub-block parallel decoding scheme [3]–[5] when the block size is 40. Figure 1 shows that an eight sub-block parallel turbo decoder needs the communication channel to be 2.7 dB better to achieve the same bit-error rate. Figure 2 further shows the channel SNR required by the algorithm without parallelism [7] and the eight sub-block parallel decoding scheme [3]–[5] to achieve $10^{-3}$ bit error rate in different block length modes.

Fig. 1. The BER performance comparison between the algorithm [7] implemented without parallelism and with the eight sub-block parallel decoding scheme [3]–[5] when the block size is 40.

Fig. 2. The channel SNR required by the algorithm without parallelism [7] and the eight sub-block parallel decoding scheme [3]–[5] to achieve $10^{-3}$ bit error rate in different block length modes.

required number of decoding iterations also changes with different quality levels of the communication channel. An adaptive decoding scheme that changes the operating point according to the channel quality and the required bit rate could therefore reduce the energy consumption.

In this paper, a 3GPP LTE turbo decoder in 65 nm CMOS with an improved parallel decoding scheme and an embedded dynamic voltage-frequency scaling (DVFS) engine is proposed. With an adaptive sub-block parallel (ASP) decoding scheme, both the throughput and the BER performance could be maintained without area overhead; the developed DVFS engine combining with an early-termination engine could reduce the energy consumption. The energy consumption ranging

978-1-4244-5760-1/10/$26.00 ©2010 IEEE
Early Termination

SISO \( k \)

SISO / g651

SISO3 / g171

\[ \pi \] as defined by LTE. The i-th interleaved memory addresses according to the interleaving order terminates the decoding.

The SISO decoders perform the BCJR decoding. An early termination input buffer and extrinsic info buffer, respectively. The SISO interleaver block. During the decoding process, extrinsic information is generated from the input code block. The SISO decoding processes using BCJR algorithm [8]

\[ \text{Figure 3 shows the system architecture. The blocks in the dashed box handle the turbo decoding operations, and those outside the dashed box belong to the DVFS scheme.} \]

Turbo decoding is an iterative process with several turbo iterations. Each turbo iteration comprises two soft-in, soft-out (SISO) decoding processes using BCJR algorithm [8] with the first one performed on the input code block in the original order and the second one in an order generated by the interleaver block. During the decoding process, extrinsic information is generated and used in succeeding iterations. The input data and extrinsic information data are stored in the input buffer and extrinsic info buffer, respectively. The SISO decoders perform the BCJR decoding. An early termination engine detects the convergence of the decoded results and terminates the decoding.

The interleaver permutes the input code blocks by generating memory addresses according to the interleaving order defined by LTE. The i-th interleaved address \( \pi(i) \) is defined as \( \pi(i) = (f_1i + f_2i^2) \% K \), where \( K \) is the block length, and \( f_1 \) and \( f_2 \) are constants derived from \( K \). We re-express the interleaving function as \( \pi(i+1) = \pi(i) + \left((f_1 + f_2) \% K\right) + \lambda(i) \% K \)

\[ \pi(i) = (f_1i + f_2i^2) \% K \]

\[ \pi(i+1) = \pi(i) + (f_1 + f_2) \% K + \lambda(i) \% K \]

with \( \lambda(i) = (2f_2 \times i) \% K = (2f_2 + \lambda(i - 1)) \% K \). The resulting interleaver architecture is shown in Fig. 4. The critical timing path passes only 4 adders and 2 multiplexers.

On top of the turbo decoding operation, a block-based throughput predictor dynamically predicts the required number of iterations for decoding a code block and then decide the required supply voltage and clock frequency by combining the predicted iteration count and the output bandwidth requirement. A buck DC-DC converter then generates the required supply voltage.

II. THE SYSTEM ARCHITECTURE

Figure 3 shows the system architecture. The blocks in the dashed box handle the turbo decoding operations, and those outside the dashed box belong to the DVFS scheme.

Turbo decoding is an iterative process with several turbo iterations. Each turbo iteration comprises two soft-in, soft-out (SISO) decoding processes using BCJR algorithm [8] with the first one performed on the input code block in the original order and the second one in an order generated by the interleaver block. During the decoding process, extrinsic information is generated and used in succeeding iterations. The input data and extrinsic information data are stored in the input buffer and extrinsic info buffer, respectively. The SISO decoders perform the BCJR decoding. An early termination engine detects the convergence of the decoded results and terminates the decoding.

The interleaver permutes the input code blocks by generating memory addresses according to the interleaving order defined by LTE. The i-th interleaved address \( \pi(i) \) is defined as \( \pi(i) = (f_1i + f_2i^2) \% K \), where \( K \) is the block length, and \( f_1 \) and \( f_2 \) are constants derived from \( K \). We re-express the interleaving function as \( \pi(i+1) = \pi(i) + \left((f_1 + f_2) \% K\right) + \lambda(i) \% K \)

\[ \pi(i) = (f_1i + f_2i^2) \% K \]

\[ \pi(i+1) = \pi(i) + (f_1 + f_2) \% K + \lambda(i) \% K \]

with \( \lambda(i) = (2f_2 \times i) \% K = (2f_2 + \lambda(i - 1)) \% K \). The resulting interleaver architecture is shown in Fig. 4. The critical timing path passes only 4 adders and 2 multiplexers.

On top of the turbo decoding operation, a block-based throughput predictor dynamically predicts the required number of iterations for decoding a code block and then decide the required supply voltage and clock frequency by combining the predicted iteration count and the output bandwidth requirement. A buck DC-DC converter then generates the required supply voltage.

III. THE ADAPTIVE SUB-BLOCK PARALLEL (ASP) DECODING SCHEME

The adaptive sub-block parallel (ASP) scheme adjusts the decoding scheme according to the input block length. The main idea is developed based on two observations in sub-block parallel decoding schemes. Firstly, the BER performance degrades less with longer blocks. Secondly, there is free space in the on-chip memory when decoding short blocks.

Figure 5 shows the ASP scheme with \( N \) parallel SISO decoders. The on-chip storage size is designed to be able to decode blocks with the maximum block length \( K_{\text{max}} \). When the input block length \( K \) is less than \( K_{\text{max}} / N \), \( N \) blocks are buffered on the chip and decoded in parallel. The BER performance drop is eliminated because the blocks are not
partitioned into sub-blocks. When $K_{\text{max}}/N < K \leq 2K_{\text{max}}/N$, $N/2$ blocks are buffered, and each block is decoded by 2 SISO decoders with 2 sub-block parallel decoding scheme. This scheme continues like this. Finally, when the block is longer than $K_{\text{max}}/2$, only one block is buffered on the chip, and $N$ sub-block parallel decoding scheme is employed.

Figure 6 shows the BER performance of ASP scheme with parallelism four and eight. Compared with [7] implemented without parallelism, the $N$-parallel ASP scheme achieves $N \times$ of throughput with negligible BER performance degradation.

In the implemented prototyping chip, four-parallel ASP scheme is adopted. The throughput of 108 Mbps is achieved. The ASP scheme increases the throughput 4× with only 21% area increase, 24% power increase and negligible BER performance drop.

IV. THE DVFS ENGINE AND THE EARLY-TERMINATION SCHEME

In this section, a DVFS engine combining with an early-termination scheme is proposed to reduce the energy consumption given different throughput requirements.

A. The Early-Termination Scheme

Early-termination schemes have been proved to be able to effectively avoid unnecessary turbo decoding iterations by detecting the convergence of the decoded results [9]. Because the required iteration count changes rapidly with time, fixing the iteration count either introduces redundant computation [5] or BER performance drop [3].

Figure 7 shows the developed double hard-decision rule (HDR) early-termination scheme and a comparison with the extrinsic info-based stopping criterion adopted in [4].

B. The DVFS Engine

Figure 8 shows the developed DVFS engine that generates the supply voltage and clock rate according to the speed requirement and the channel quality. An iteration predictor predicts the iteration count and decides if the voltage and clock rate need to be updated. The predicted iteration count $N_{\text{pred}}$ of code block $n$ is derived from the accumulated prediction error $Err$ and the average iteration count $N_{\text{avg}}$ as follows:

$$N_{\text{pred}}[n] = \begin{cases} N_{\text{avg}}[n] + |Err[n]|/32, & \text{if } |Err[n]| \leq 16 \\ N_{\text{avg}}[n-1] + |Err[n]|/8, & \text{otherwise}. \end{cases}$$

The required clock rate $f_{\text{clk}}$ is then derived from $N_{\text{pred}}$ and the target throughput.

The target voltage is derived from $f_{\text{clk}}$ with a look-up table (LUT). A 4-b charge-redistribution DAC then generates the corresponding reference voltage $V_{\text{ref}}$. A comparator compares $V_{\text{ref}}$ with the delivered supply voltage $V_{\text{dd}}$. The loop controller then generates PWM signals in response to the comparator output, and $V_{\text{dd}}$ is obtained by passing the PWM signals to an off-chip L-C filter.

The DVFS energy efficiency is the ratio of the turbo decoder power in the total power, and it ranges from 80% to 87% while delivering 2.9 mW to 75 mW to the turbo decoder. The efficiency of the DC/DC converter is limited by the parasitic resistance of the pads connecting the driver stage and the off-chip inductor, and it could be improved by further optimizing the pad design. The waveform in Fig. 9 shows $N_{\text{pred}}$ tracking the iteration count and $V_{\text{dd}}$ changing with the target voltage index with a voltage ripple of 30 mV.

V. CHIP IMPLEMENTATION RESULTS

The developed 3GPP LTE turbo decoder is implemented in a 65 nm CMOS process. Figure 10 shows the die micrograph and the summary of measurement results. This chip supports all the 188 block types with lengths from 40 to 6144 and
in this work is 0.168 nJ, and it could be reduced to 0.077 nJ.

The developed ASP scheme maintains the same throughput without BER performance drop. [5] could use lower parallelism to reduce BER performance drop. However, this also reduces the throughput. The redundant iterations could be eliminated by the double HDR early-termination scheme, and the energy consumption per bit thus could be relatively even lower. In addition, the developed design occupies the smallest area.

VI. CONCLUSION

A 3GPP LTE turbo decoder is implemented in a 65 nm CMOS technology and occupies 0.66 mm² of area. A throughput of 108 Mbps is achieved without degrading the BER performance by developing an adaptive sub-block parallel (ASP) decoding scheme. It improves the BER performance by up to 2.7 dB compared with 8 sub-block parallel scheme. To reduce the energy consumption for various output bandwidth demands and channel conditions, a DVFS engine and an early-termination scheme are developed. The measured energy consumption is 0.077∼0.168 nJ per bit per iteration and 0.39∼0.85 nJ per bit.

ACKNOWLEDGMENT

The authors thank TSMC for the chip fabrication and National Chip Implementation Center for chip testing facility. This work was supported in part by MediaTek Fellowship.

REFERENCES


TABLE I

<table>
<thead>
<tr>
<th>Supported Standard</th>
<th>3GPP LTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block Size</td>
<td>40~6144 (188 modes)</td>
</tr>
<tr>
<td>Technology</td>
<td>65nm CMOS</td>
</tr>
<tr>
<td>Core Size</td>
<td>0.66 mm²</td>
</tr>
<tr>
<td>Core Vdd Range</td>
<td>0.675 V~1.2 V</td>
</tr>
<tr>
<td>Operating Frequency</td>
<td>24 MHz~270 MHz</td>
</tr>
<tr>
<td>Throughput</td>
<td>9.6~108 Mbps</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>3.7~90.9 mW</td>
</tr>
<tr>
<td>Energy Consumption</td>
<td>0.077~0.168 nJ/bit/iteration</td>
</tr>
</tbody>
</table>

A through-put of 108 Mbps is achieved without degrading the BER performance by developing an adaptive sub-block parallel (ASP) decoding scheme. It improves the BER performance by up to 2.7 dB compared with 8 sub-block parallel scheme. To reduce the energy consumption for various output bandwidth demands and channel conditions, a DVFS engine and an early-termination scheme are developed. The measured energy consumption is 0.077∼0.168 nJ per bit per iteration and 0.39∼0.85 nJ per bit.

ACKNOWLEDGMENT

The authors thank TSMC for the chip fabrication and National Chip Implementation Center for chip testing facility. This work was supported in part by MediaTek Fellowship.