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 188 different block length modes. The BER performance degrades more with shorter block lengths. [5] provides the flexibility to reduce the parallelism by disabling decoding engines. The throughput however reduces when fewer decoding engines are active.

The instantaneous data rate requirement changes with applications from web browsing to HD video streaming. The

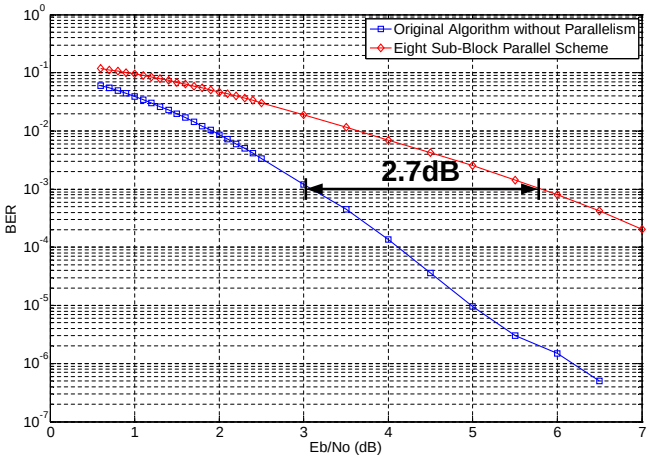


Fig. 1. The BER performance comparison between the algorithm [7] implemented without parallelism and with the eight sub-block parallel decoding scheme [3]–[5] when the block size is 40.

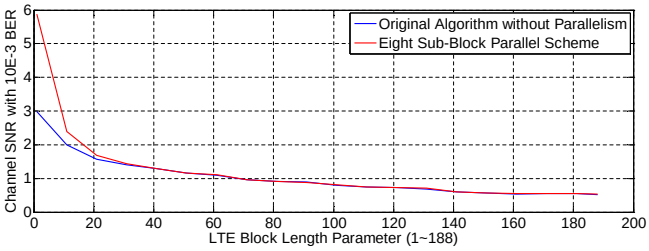


Fig. 2. The channel SNR required by the algorithm without parallelism [7] and the eight sub-block parallel decoding scheme [3]–[5] to achieve 10^{-3} bit error rate in different block length modes.

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

978-1-4244-5759-5/10/$26.00 ©2010 IEEE

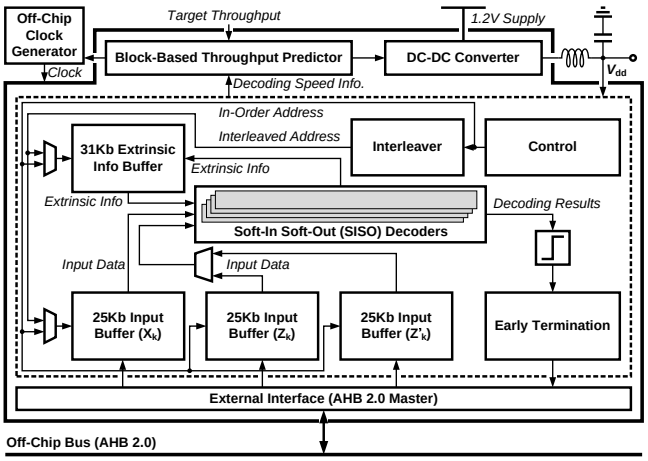


Fig. 3. The system architecture.

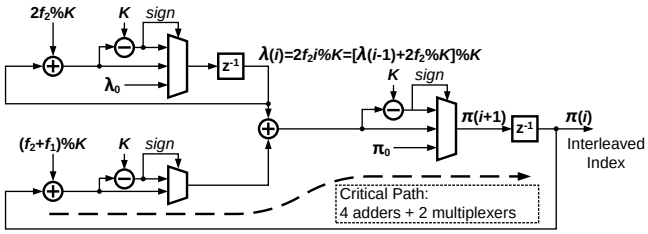


Fig. 4. The interleaver architecture.

from 0.077 to 0.168 nJ/bit/iteration is thus achieved.

The rest of this paper is structured as follows. Section II introduces the system architecture. The developed adaptive sub-block parallel decoding scheme is presented in Sec. III. Section IV describes the design of the DVFS engine and the early-termination scheme. Section V shows the experimental results. Finally, Sec. VI concludes this work.

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 $\pi(i) = (f_1 i + f_2 i^2)\%K$, where K is the block length, and f_1 and f_2 are constants derived from K. We re-express the interleaving function as $\pi(i+1) = [\pi(i) + ((f_1+f_2)\%K) + \lambda(i)]\%K$ 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.

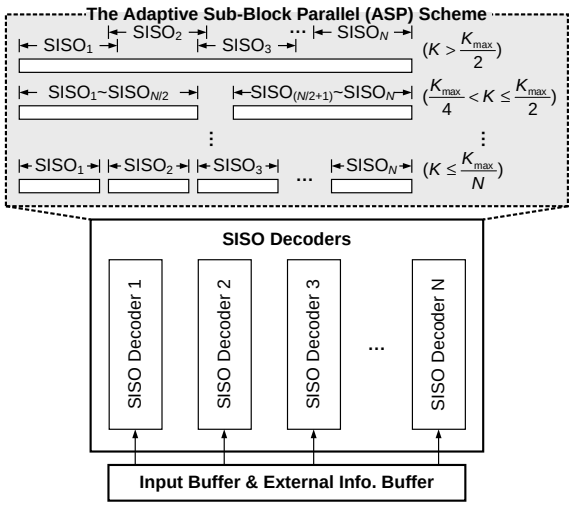


Fig. 5. The adaptive sub-block parallel (ASP) decoding scheme with N SISO decoders. The ASP scheme adjusts the decoding scheme according to the input block length K and the maximum block length defined by LTE K_{max}.

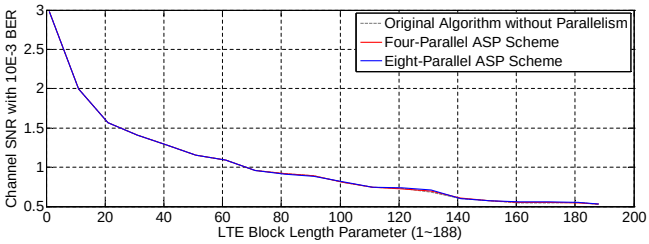


Fig. 6. The channel SNR required by the algorithm [7] implemented without parallelism, with four-parallel ASP scheme and with eight-parallel ASP scheme to achieve 10^{-3} bit error rate in different block length modes.

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_{max}. When the input block length K is less than K_{max}/N, N blocks are buffered on the chip and decoded in parallel. The BER performance drop is eliminated because the blocks are not

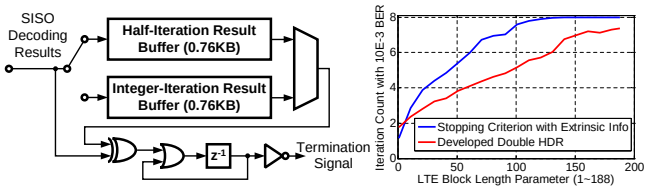


Fig. 7. The developed double hard-decision rule (HDR) early termination scheme and a comparison with the extrinsic info-based stopping criterion adopted in [4].

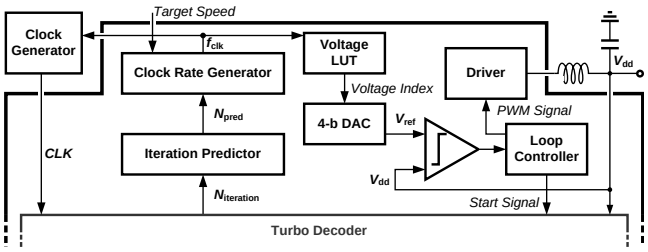


Fig. 8. The developed DVFS scheme including a throughput prediction engine and a DC-DC converter.

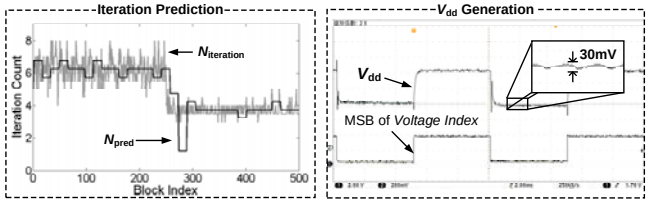


Fig. 9. The results for iteration prediction and the measured step response of the delivered supply voltage.

partitioned into sub-blocks. When $K_{max}/N < K \leq 2K_{max}/N$, $N/2$ 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 $K_{max}/2$, 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 $4\times$ 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. The decoded results are examined twice per turbo iteration by comparing the decoded results with the decoded results obtained one iteration before. A small 1.52 KB buffer is thus required to store previous decoded results. To our knowledge, [4] is the only LTE turbo decoder with an early-termination scheme. The stopping criterion adopted in [4] compares the decoded results and the extrinsic information. Because the extrinsic information changes relatively slowly, it takes longer to detect the convergence. Figure 7 also compares the required iteration count in different block length modes of the double HDR scheme and of the stopping criterion in [4]. Both schemes are tested with 4-parallel ASP scheme and channel SNR values corresponding to 10^{-3} BER. The average iteration count of the double HDR scheme is 5.02 and is 28% lower than the one in [4]. 28% of the energy consumption is thus saved.

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_{pred} of code block n is derived from the accumulated prediction error Err and the average iteration count N_{avg} as follows:

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

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

The target voltage is derived from f_{clk} with a look-up table (LUT). A 4-b charge-redistribution DAC then generates the corresponding reference voltage V_{ref}. A comparator compares V_{ref} with the delivered supply voltage V_{dd}. The loop controller then generates PWM signals in response to the comparator output, and V_{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_{pred} tracking the iteration count and V_{dd} changing with the target voltage index with a voltage ripple of 30 mV.

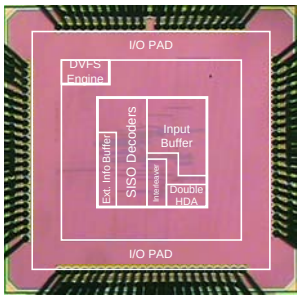
V. Chip Implementation Results

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

TABLE I

COMPARISON TO OTHER 3GPP LTE TURBO DECODERS.

	This Work	ISSCC 2010 [3]	CICC 2009 [4]	VLSI 2009 [5]
CMOS Technology	**65 nm**	0.13 μm	0.13 μm	90 nm
Supported Standard	**3GPP LTE**	3GPP LTE	Wimax/3GPP LTE	3GPP LTE
SISO Decoding Scheme	**4 Adaptive Sub-Block Parallel**	8 Sub-Block Parallel	8 Sub-Block Parallel	1/2/4/8 Sub-Block Parallel
Termination Scheme	**Double HDR Adaptive Termination**	Fixed 5.5 Iteration	Extrinsic Info-Based	Fixed 8 Iteration
Embedded Scalability	**Embedded DVFS**	No	No	No
Supply Voltage	**0.675∼1.2 V**	1.2 V	1.2 V	1.0 V
Throughput	**9.6∼108 Mbps**	390 Mbps	186 Mbps	129 Mbps (8 Parallel)
Active Area	**0.66 mm^2**	3.57 mm^2	10.7 mm^2	2.1 mm^2
Total Power Consumption	**3.7∼90.9 mW**	788.9 mW	N.A.	219 mW
Energy Consumption	**0.077∼0.168 nJ/bit/iteration**	0.37 nJ/bit/iteration	0.61 nJ/bit/iteration	0.21 nJ/bit/iteration
	0.39∼0.85 nJ/bit	2.02 nJ/bit	4.88 nJ/bit	1.70 nJ/bit



Supported Standard	3GPP LTE
Block Size	40~6144 (188 modes)
Technology	65nm CMOS
Core Size	0.66 mm^2
Core V_{dd} Range	0.675 V~1.2 V
Operating Frequency	24 MHz~270 MHz
Throughput	9.6Mbps~108Mbps
Power Consumption	3.7mW~90.9mW

Fig. 10. The die micrograph and the summary of measurement results.

occupies 0.66 mm^2 of area. The DVFS engine delivers a V_{dd} ranging from 0.675 V to 1.2V corresponding to the operating frequency from 24 MHz to 270 MHz. The total power consumption including the DVFS engine and the turbo decoder ranges from 3.7 mW to 90.9 mW while achieving a throughput from 9.6 Mbps to 108 Mbps.

Table I compares the key characteristics with state-of-the-art LTE turbo decoder chips. Other turbo decoders use sub-block parallel decoding scheme which introduces BER performance drop as shown in Fig. 1. [5] could use lower parallelism to reduce BER performance drop. However, this also reduces the throughput. The developed ASP scheme maintains the same throughput without BER performance drop.

[3] fixes the iteration count to be 5.5, and this introduces BER performance drop; [5] fixes the iteration count to be 8, and this introduces redundant computation. The double HDR scheme is developed in this work to adaptively decide the number of iteration. 28% of energy consumption is saved compared with the early-termination scheme adopted in [4].

To further reduce the energy consumption with various throughput requirements, a DVFS engine is developed to lower the clock rate and the supply voltage. The throughput of this work ranges from 9.6 Mbps to 108 Mbps. It satisfies both the 100 Mbps LTE peak data rate [1] and the 60 Mbps estimated system performance [2]. Increasing the parallelism of the ASP scheme can easily increase the maximum throughput to further support the future MIMO configurations, and the BER performance could be still maintained as shown in Fig. 6.

The energy consumption per bit per iteration at 108 Mbps in this work is 0.168 nJ, and it could be reduced to 0.077 nJ due to the DVFS scheme. 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 mm^2 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

[1] *Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and channel coding*, 3GPP TS 36.212 V8.5.0, 2009.

[2] J. J. Sanchez, D. Morales-Jimenez, G. Gomez, and J. T. Enbrambasaguas, "Physical layer performance of long term evolution cellular technology," in *16th IST Mobile and Wireless Communications Summit*, 2007, pp. 1–5.

[3] C. Studer, C. Benkeser, S. Belfanti1, and Q. Huang, "A 390Mb/s 3.57mm^2 3GPP-LTE turbo decoder ASIC in 0.13μm CMOS," in *ISSCC Dig. Tech. Papers*, 2010, pp. 274–275.

[4] J.-H. Kim and I.-C. Park, "A unified parallel radix-4 turbo decoder for mobile WiMAX and 3GPP-LTE," in *IEEE Custom Intergrated Circuits Conference (CICC)*, 2009, pp. 487–490.

[5] C.-C. Wong, Y.-Y. Lee, and H.-C. Chang, "A 188-size 2.1mm^2 reconfigurable turbo decoder chip with parallel architecture for 3GPP LTE system," in *VLSI Symposium Dig. Tech. Papers*, 2009, pp. 288–289.

[6] O. Y. Takeshita, "On maximum contention-free interleavers and permutation polynomials over integer rings," *IEEE Trans. Information Theory*, pp. 1249–1253, Mar 2006.

[7] J. Vogt and A. Finger, "Improving the max-log-MAP turbo decoder," *Electronics Letters*, pp. 1937–1939, Nov 2000.

[8] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Information Theory*, pp. 284–287, 1974.

[9] A. Matache, S. Dolinar, and F. Pollara, "Stopping rules for turbo decoders," *TMO Progress Report 42-142*, Aug 2000.