D20-008 **DESIGN GROUP**

作品名稱

基於神經網路且支援多碼長度之可重構極化碼解碼器

Reconfigurable Neural Network-Assisted Polar Decoder with Multi-Code Length Support

隊伍名稱

安迪我老闆 Andy Is My Boss

隊長

鄧傑方 臺灣大學電子工程學研究所

隊員

陳鈞翔 臺灣大學電子工程學研究所

作品摘要

下世代通訊系統之技術標準有三大應用方向：增強型移動頻寬（Enhanced Mobile Broadband，eMBB），超高可靠性與超低延遲通訊（Ultra-Reliable Low-Latency Communications，URLLC），以及巨量物聯網通訊（Massive Machine-Type Communications，mMTC），如圖一所示。因此在5G通訊系統中，為了達到 eMBB 的要求，就必須要提供相較於 4G 更高可靠、更高吞吐量的傳輸能力；而對於 URLLC 及 mMTC 則分別需要滿足超低延遲及超高能源效率的需求。因此這些多樣化的需求對於通訊系統中的解碼器設計將會是一大挑戰。

自從 2009 年由土耳其 Arikan 教授提出，極化碼旋即受到了極大的關注。極化碼與眾多編碼相比具有高增益、低複雜度等好處，因此已於 2016 年被 3GPP 通訊標準會議採用，作為 5G eMBB 控制頻道的編碼。因此，近年來極化碼解碼器大多注重追求高吞吐量及高可靠度的傳輸能力。然而對於 URLLC 及 mMTC，極化碼也是潛在的編碼方式，為了使極化碼能擴展到下世代通訊系統中更多樣化的應用情境，除了保有高吞吐量及高可靠度之外，在設計上還需達到低延遲、高能源效率，以及支援多碼率、多碼長這三大目標。

雖然過去常用之置信傳播解碼（Belief Propagation，BP）演算法具有高吞吐量之優勢，能符合 5G eMBB 情景所需，但是此演算法所需之疊代次數也造成延遲較長之問題，同時對於吞吐量仍然有進步的空間。為了解決上述的問題，並滿足多樣化的需求，本作品主要創意與貢獻如下：

1. 利用近年來非常火紅的深度學習技術，來訓練 BP 演算法連線上之縮放權重，有效地降低八倍的疊代次數，進而降低延遲，同時提升吞吐量。
2. 提出兩種數值上優化的方法，能減少數據存儲和運算的所需位元數，使得處理單元（Processing Element，PE）的面積和功耗大幅降低 73% 和 67%。因此對於 mMTC 之應用，能達到高面積效率及高能源效率的需求。
3. 提出可重構化的硬體架構，能支援多碼率、多碼長之需求，一共支援 N = 32, 64, 128, 256 四種不同的模式，提高解碼器之功能性及硬體使用效率。同時可以與許多增強機制完美結合，在高吞吐量與高可靠性間進行調整，以滿足 5G New Radio 中的各種應用場景。

本設計作品以 40 奈米製程作為設計實施例，核心面積是 0.18mm^2。量測結果顯示，我們的可重構設計可以支持多碼長解碼，能源效率為 7.8 至 13.6 pJ/b。與最先進的單模 BP 解碼器設計相比，我們的設計能降低 2.3 倍延遲，同時提升 2.3 倍的吞吐量和 10.0 倍的能效。此可重構解碼器，非常適合用於滿足 5G NR 中 eMBB、URLLC 和 mMTC 各種應用場景的需求。同時，我們的設計也是第一個展示了神經網絡輔助數位訊號處理引擎在未來通信系統中的巨大潛力！

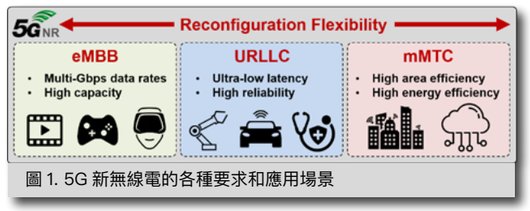

圖 1. 5G 新無線電的各種要求和應用場景

指導教授

吳安宇 臺灣大學電子工程學研究所

- 美國馬里蘭大學電機博士，現為臺灣大學電機工程學系特聘教授。曾任 AT&T Bell Lab 之 Member of Technical Staff (MTS)、工業技術研究院系統晶片中心副主任、臺灣大學電子工程學研究所所長。於 2015 年晉升為 IEEE Fellow，並獲 108 年度中國工程師學會「傑出工程教授獎」以及獲頒美國馬里蘭大學電機系 2019 ECE Distinguished Alumni Award。2020 年起擔任 IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS) 之總主編。
- **研究領域：**VLSI/CAD、通訊積體電路、信號處理

Abstract

Polar codes have become more and more important since they were officially selected as the channel coding in 5G standard. However, there are only a couple of fabricated ASICs featuring polar decoder and most of them target improved throughput rate and error-correction performance. Therefore, the requirements of another two application scenarios of ultra-reliable low-latency communications (URLLC) and massive machine-type communications (mMTC), such as low latency and energy efficiency, are neglected.

In this work, we present an ultra-low latency and energy efficiency 7.8-13.6 pJ/b polar decoder engine fabricated in 40nm CMOS technology. With the proposed recurrent neural network-assisted belief propagation (RNN-BP) decoding algorithm, we can improve the convergence rate by 8 times with reasonable hardware and memory overhead. Then, a reconfigurable architecture is proposed to support multiple code lengths without significantly increased hardware complexity by taking advantage of BP's regular structure. It contributes to 2-8× improved hardware utilization rate and provides a perfect combination with many other enhanced mechanisms. Moreover, two optimization techniques for the design of processing element (PE) are proposed to jointly reduce area and power by 73% and 67%.

From the measurement results, our reconfigurable RNN-BP polar decoder chip has 2.3×, 2.3×, and 10.0× enhancement over prior designs in terms of latency, throughput rate, and energy efficiency, respectively, which make our design perfectly suitable to meet various application scenarios in 5G new radio (NR).

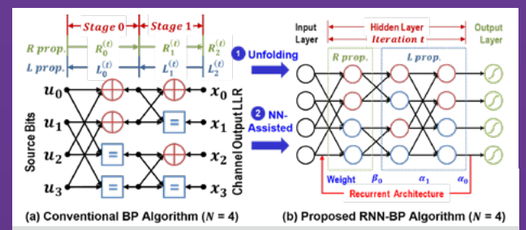

Fig. 2 (a) Conventional belief propagation (BP) algorithm, and (b) the proposed recurrent neural network-assisted belief propagation (RNN-BP) algorithm

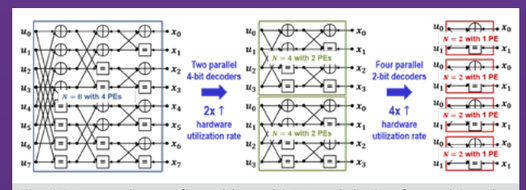

Fig. 3 Proposed reconfigurable architectural design for multi-code length support

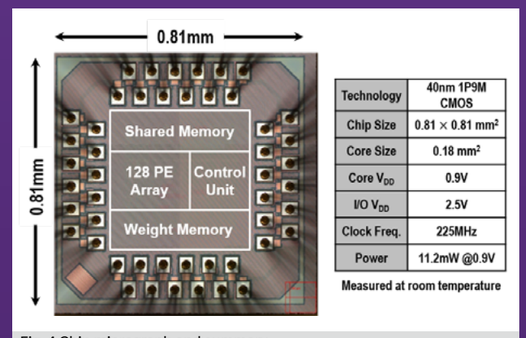

Fig. 4 Chip micrograph and summary