A 40-nm 13.88-TOPS/W FC-DNN Engine for 16-bit Intelligent Audio Processing Featuring Weight-Sharing and Approximate Computing

Tay-Jyi Lin, Ze Li, Yun-Cheng Chen, Chien-Tung Liu, Tien-Fu Chen*, and Jinn-Shyan Wang

National Chung Cheng University and *National Yang Ming Chiao Tung University, Taiwan

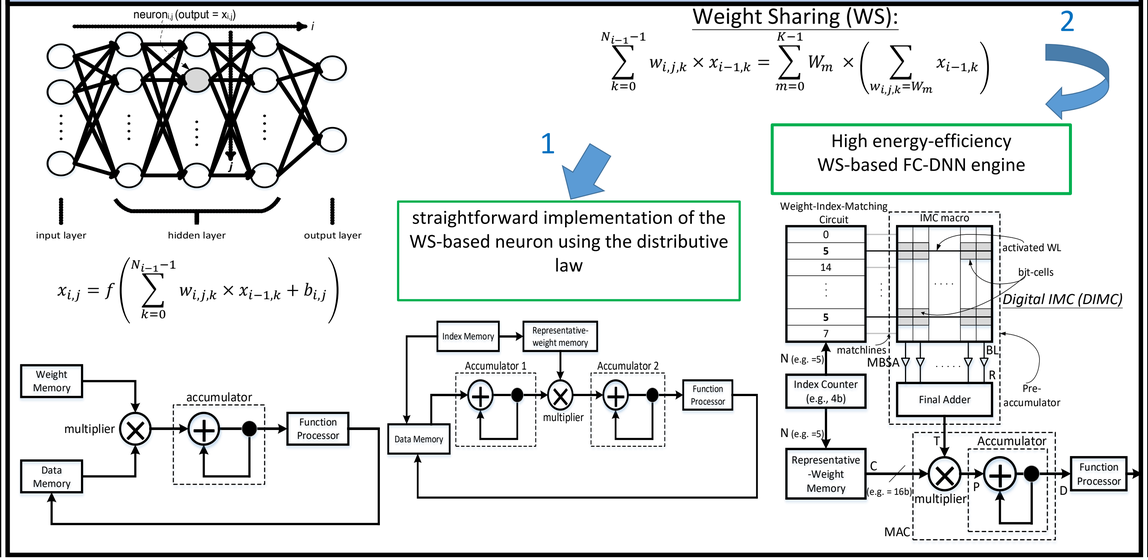
Supported by National Science Council, Taiwan (MOST 111-2221-E-194-047-MY3 & NSTC 113-2640-E-194-001)

國立中正大學 National Chung Cheng University

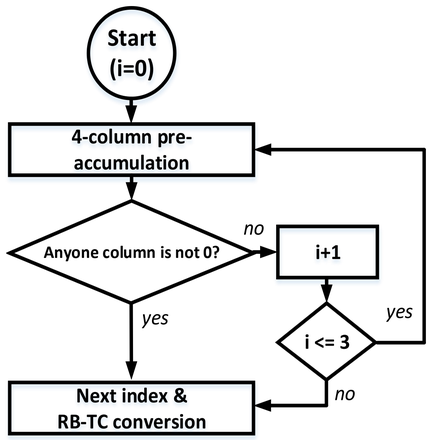
Abstract

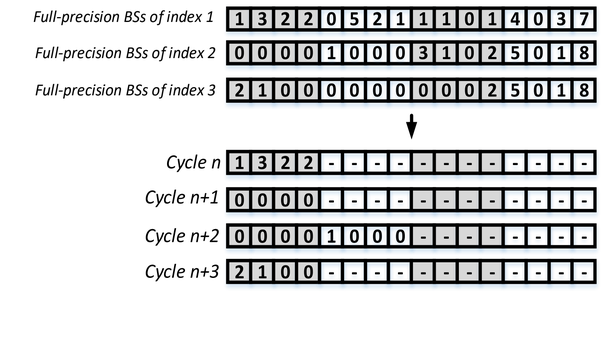
This work presents a novel accelerating engine to efficiently compute the fully connected (FC) deep neural networks (DNN) based on weight sharing (WS). Beyond the highly reduced weights and weight accesses in previous designs, the proposed engine significantly reduces the multiplications by exploiting the distributive law of shared weights. All-digital in-memory computing (IMC) and approximate computing (AC) were applied to accelerate accumulations and improve energy efficiency. We have designed and implemented a test chip in 40nm CMOS for intelligent audio processing of dysarthric voice conversion (i.e., using an FC-DNN composed of neurons with 16-bit inputs, 16-bit weights, and 16-bit activations). When turning AC off and on, the energy efficiency for 16b operation at 0.9V is 7.04 TOPS/W and 13.88 TOPS/W, respectively.

Weight Sharing-based Inner-Product Operation and FC-DNN Engines ①



Approximate Computing (AC) for the All-Digital WS-based FC-DNN Engine ③





For DVC application, the proposed AC provides a 60% reduction in the clock-cycle count to improve the energy efficiency (TOPS/W).

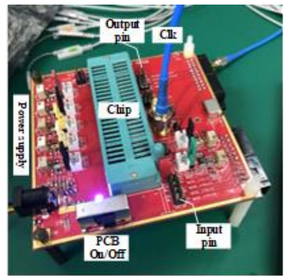
WS-based FC-DNN Engine for ***16b*** Intelligent Audio Processing using Digital-IMC (DIMC) ②

Application example of 16-bit intelligent audio processing:

Real-time Dysarthric Voice Conversion (DVC)

Before conversion

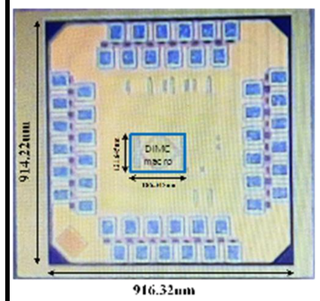
After conversion



New DIMC-based FC-DNN engine for 17X improvement of energy efficiency compared to the work in [Ref]

Ref: T.-J. Lin, C.-Z. Liao, Y.-J. Hu, W.-C. Hsu, Z.-X. Wu, S.-Y. Wang, C.-M. Huang, Y.-H. Lai, C. Yeh, and J.-S. Wang, "A 40nm CMOS SoC for real-time dysarthric voice conversion of stroke patients," in Proc. IEEE Asia and South Pacific Design Automation Conference (ASP-DAC), Jan. 17-20, 2022.

The 40-nm 13.88-TOPS/W FC-DNN Engine Chip ④



	ISSCC'21 [1]	ISSCC'22 [2]	This work	
Chip design	DIMC	DIMC	WS-based FC-DNN w/ DIMC core	
Technology	22nm	5nm	40nm	
Supply voltage (V)	0.8±10%	0.5 ~ 0.9	0.9±10%	
Macro area (mm^2)	0.202 mm^2	0.0133 mm^2	0.023 mm^2 (DIMC core) 0.114 mm^2 (WS FC-DNN)	
Memory types for DIMC	SRAM-DIMC	SRAM-DIMC	SRAM-based DIMC	
Bit cells for DIMC	6T/4T DIMC	12T DIMC	6T/3T DIMC	
Data-memory capacity	128×128 SRAM	n.a.	256×16b data (4Kb SRAM-DIMC)	
Data bits	1 ~ 8	4	16	
Weight bits	4/8/12/16	4	16	
Measurement Conditions	4b/4b @ 0.72V (12.5% tog. rate)	4b/4b @ 0.5V, (10% tog. rate)	16b/16b @ 0.9V (Dysarthric voice conversion)	
TOPS/W	89	253.5	7.04 (AC off)	13.88 (AC on)

[1] "An 89TOPS/W and 16.3TOPS/mm^2 all-digital SRAM-based full-precision compute-in-memory macro in 22nm for machine-learning edge applications"
[2] "A 5-nm 254-TOPS/W 221-TOPS/mm2 fully-digital computing-in-memory macro supporting wide-range dynamic-voltage-frequency scaling and simultaneous MAC and write operations"