Audio and Image Cross-Modal Intelligence via a 10TOPS/W 22nm SoC with Back-Propagation and Dynamic Power Gating

Zichen Fan, Hyochan An, Qirui Zhang, Boxun Xu, Li Xu, Chien-wei Tseng, Yimai Peng, Ang Cao, Bowen Liu, Changwoo Lee,
Zhehong Wang, Fanghao Liu, Guanru Wang, Shenghao Jiang, Hun-Seok Kim, David Blaauw, Dennis Sylvester
University of Michigan, Ann Arbor, USA. zcfan@umich.edu

2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits) | 978-1-6654-9772-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/VLSITechnologyandCir46769.2022.9830226

Abstract We present an ultra-low-power multimedia signal processor (MMSP) SoC that integrates a versatile deep neural network (DNN) engine with audio and image signal processing accelerators for cross-modal IoT intelligence. The proposed MMSP features 2MB MRAM to store all DNN weights on-chip with an energy-efficient dataflow using an MRAM-cache and dynamic power gating. The SoC achieves up to 3-10 TOPS/W peak energy efficiency and consumes only 0.25-3.84 mW. Being the first to demonstrate CNN, GAN, and back-propagation (BP) on a single accelerator SoC for cross-modal fusion, it outperforms state-of-the-art DNN processors by 1.4 - 4.5× in energy efficiency.

Introduction: DNN-based image and/or audio processing has been widely adopted in intelligent IoT systems. However, the traditional processing flow (Fig. 1 top) that offloads DNN processing to the cloud suffers from bandwidth, energy and privacy problems for resource-constrained IoT applications. To tackle these challenges, the proposed SoC adopts a fully-at-edge processing flow for *audio and image multimodal intelligence* with integrated pre-/post-processing accelerators (JPEG, H.264, FFT, and Mel-spectrum), a GAN/BP-capable neural engine, and 2MB on-chip MRAM for non-volatile weight/data/code storage.

Fig. 2 depicts target application scenarios of the proposed SoC. It can simultaneously receive 12b per pixel VGA images and 8kHz 8b per sample audio signals using dedicated interfaces. The SoC can perform image- or audio-only applications such as face detection or keyword recognition, and also (conditionally) execute image-audio fusion applications such as cross-modal verification and active speaker detection to further increase the detection and recognition credibility. To accomplish this, the neural engine supports different types of DNN operations such as (depth-wise) convolution and fully connected layers (FCLs) for CNN, deconvolution for GAN, back-propagation for BPGAN [3], and on-chip weight retraining (transfer learning of the last layer).

Architecture: Fig. 3 shows the overall SoC architecture. All sub-blocks communicate with each other using an AHB bus. The audio interface performs audio feature extraction using frame buffers, a 256-point FFT unit, a 32/64 channel Mel-filter unit, and a power-to-dB $\log_2$ unit. The image interface features a change detection and on-the-fly JPEG compressed-memory [4] to temporarily store VGA frames in a compressed format. Only the change detected macro-blocks are stored and processed as regions-of-interest (RoIs). The H.264 engine performs image compression [4] on non-rectangular RoIs for compact storage in on-/off-chip memory. The neural engine (NE) is a reconfigurable DNN accelerator that supports various operations including (depth-wise) convolution, deconvolution, FCL, and back-propagation. NE supports DNN weights both in uncompressed (8b/weight) and Huffman-compressed (~2b/weight) formats. The SoC integrates a 2MB MRAM macro to store all weights on-chip, and a 1.5MB multi-bank SRAM activation memory to store all feature maps for backpropagation. For simple applications that only require partial activation memory, unused SRAM macros are dynamically power gated to save leakage power. The main computation unit is an 8×8×8 processing element (PE, each with an 8-bit MAC) array, which enables activation and weight reuse via inter-PE connections. The top 1×8×8 PEs are multi-functional PEs (MPEs) that supports both MAC and max/average pooling operations. Furthermore, ping-pong memory structures for the local memory and row/col buffer enable non-blocking pipelining between data movement and computation to increase PE utilization. The MRAM macro is accompanied with an MRAM cache to enable a dynamic power gating scheme (details in later section).

Power Domain: Fig. 4 shows the power domain design of the SoC. We implement power gating for each SRAM block and MRAM macro to decrease standby leakage power up to 83%. Because the MRAM can retain all data, SRAMs can be power gated when NE is inactive. Most of the computation logic can also be power gated while the always-on block (pads, headers, state registers, and power sequence control logic) consumes only 460nW.

NE Dataflow: Fig. 5 shows the energy-efficient, computation-skipping dataflow scheme implemented in NE. The base dataflow is output stationary, which is used in convolution, strided convolution, and depth-wise convolution. Weights are shared along PE array rows, and input activations are shared across the output channels (OCs). For deconvolution, a zero-skipping dataflow exploits the deterministic and regular pattern of zero padding to increase throughput by 4× by only computing non-zero values in each PE (Fig. 5 left). For back-propagation through a ReLU layer, the zero activation positions are pre-recorded during the forward path to data-gate all computations in the backward path if they (back)propagate to a pre-recorded zero activation position (Fig 5 right).

MRAM-cache Dynamic Power Gating: Although MRAM has the advantage of high density and non-volatility, its active leakage and readout power can be significant for ULP applications. To mitigate this issue, we propose an MRAM-cache architecture and dynamic power gating scheme (Fig. 6). In our scheme, MRAM is power gated until NE executes the load weight (LD_WEIGHT) instruction. During LD_WEIGHT operation, MRAM powers up and weights are read and loaded into the SRAM-based weight cache (WC). MRAM then goes back into either 1) sleep (SLP) mode where the MRAM array is powered off but peripherals remain on, or 2) power-down (PD) mode where the peripherals are also power gated. The optimal selection between SLP and PD depends on the reuse factor of the cached weight, as shown in Fig. 6 (bottom, right), which can be identified during NE programing. The measured MRAM VDIO current waveform shows the whole MRAM-cache sequence. During NN processing, weights are read from the weight cache (while MRAM is in either SLP or PD) for reduced memory readout power compared to MRAM accesses. Based on measured results, the combined weight caching and power gating reduces weight readout power by 95.3% with only slightly increased (4.3%) operation time due to MRAM wake-up latency and cache loading time overhead.

The table in Fig. 6 summarizes the three different power modes: PD, SLP, and stand-by (STB). The analysis concludes that when each cached weight is reused for 353.4 MACs, the PD mode is preferable. Otherwise, SLP mode has an advantage because of the lower overhead in power-up energy (offsetting its higher leakage).

Measurement Results: The SoC was fabricated in 22nm ULL technology and the die photo is shown in Fig. 9. The peak energy efficiency for various NN instructions is shown in Fig. 7 (left). The efficiency for convolution / deconvolution / stride-convolution-backpropagation (CONV / DECONV / S_CONV_BP) is 3.1 / 10 / 10 TOPS/W. The efficiency for DECONV and S_CONV_BP is significantly higher because of the zero-skipping dataflow. The convolution backpropagation can achieve 3.7 TOPS/W with zero-gating dataflow. Depth-wise convolution (DWCONV) and FCL have lower efficiencies due to only 1/8 of the PE array being utilized. Fig. 7 (right) shows the voltage-frequency-efficiency tradeoff for CONV. The SoC achieves the highest energy efficiency at 0.46V (VDD_MAIN) and 1.2MHz, while the total system power is 387uW. Fig. 7 (right) shows that MRAM dynamic power gating enhances the energy efficiency especially at low voltage as it significantly reduces the leakage.

Fig. 8 demonstrates a person-of-interest (PoI) tracking scenario and chip performance for that cross-modal intelligence scenario. This task first performs face detection (FD) on change-detected regions and then face recognition (FR) if faces are detected. The H.264 engine compresses the changed blocks if a PoI is recognized. In the meantime, the audio interface extracts the audio features (AFE), and cross-modal verification (CMV) is performed using both the audio features and face image. If the CMV model confirms the audio is matched with the target face, the audio signal is compressed (AC) by BPGAN. Fig. 8 (bottom) shows the power consumption and latency of each step in this process. Table 1 compares this SoC and other state-of-the-art prior designs in similar application spaces.

References: [1] K.D. Choo *et al.*, ISSCC'19 [2] M. Cho *et al.*, ISSCC'19 [3] B. Liu *et al.*, ICASSP'20 [4] H. An *et al.*, VLSI'20 [5] P. Jokic *et al.*, VLSI'21 [6] J. Giraldo *et al.*, VLSI'19 [7] J. Lee *et al.*, VLSI'19 [8] M. Giordano *et al.*, VLSI'21 [9] S. Kang *et al.*, JSSC'21

Acknowledgement: The authors would like to thank Sony Semiconductor Solutions Corp./Sony Electronics Inc. for supporting this work.

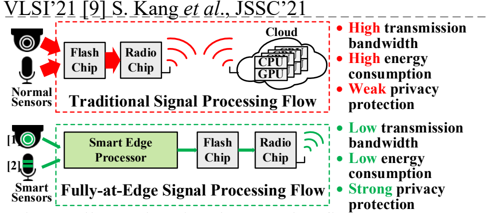
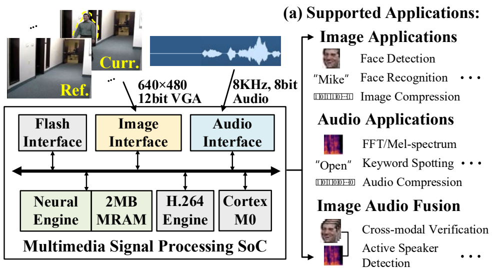

Cloud
Flash Chip
Radio Chip
CPU
GPU
Normal Sensors
Traditional Signal Processing Flow
• High transmission bandwidth
• High energy consumption
• Weak privacy protection
[1]
[2]
Smart Sensors
Smart Edge Processor
Flash Chip
Radio Chip
Fully-at-Edge Signal Processing Flow
• Low transmission bandwidth
• Low energy consumption
• Strong privacy protection

Fig.1 Fully at edge signal processing flow.



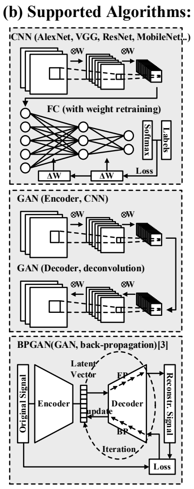

(b) Supported Algorithms:
CNN (AlexNet, VGG, ResNet, MobileNet...)
FC (with weight retraining)
Softmax
Labels
ΔW
ΔW
Loss
GAN (Encoder, CNN)
GAN (Decoder, deconvolution)
BPGAN(GAN, back-propagation)[3]
Original Signal
Encoder
Latent Vector
FP
update
Decoder
BP
Iteration
Reconstr. Signal
Loss

Fig.2 Proposed multimedia signal processing SoC and its supported applications.

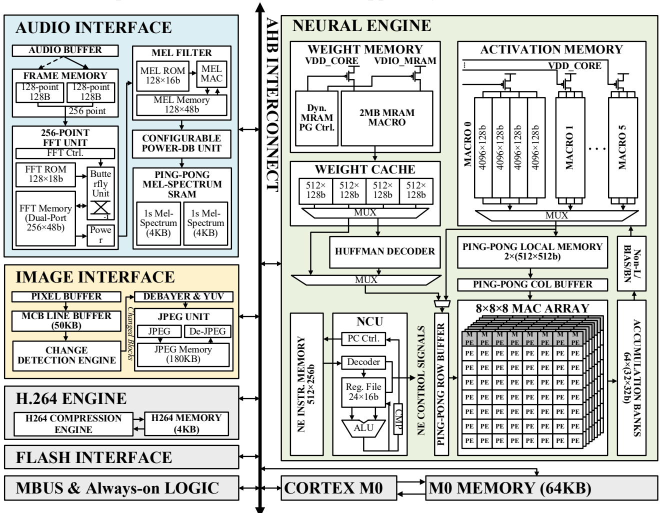


Fig.3 System overall architecture.

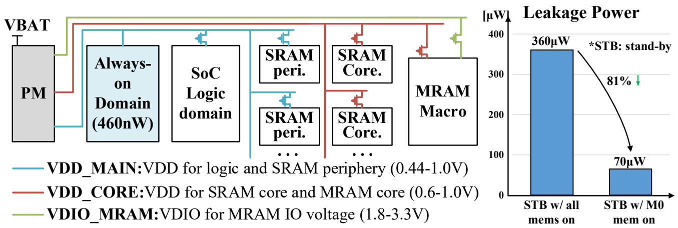


Fig.4 Overall power domain with 460nW always-on domain.

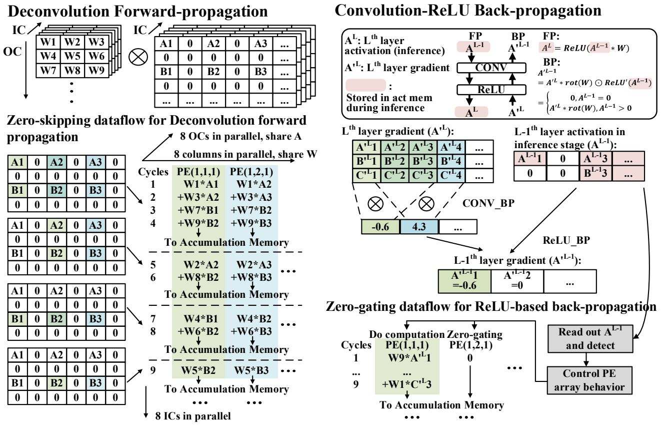
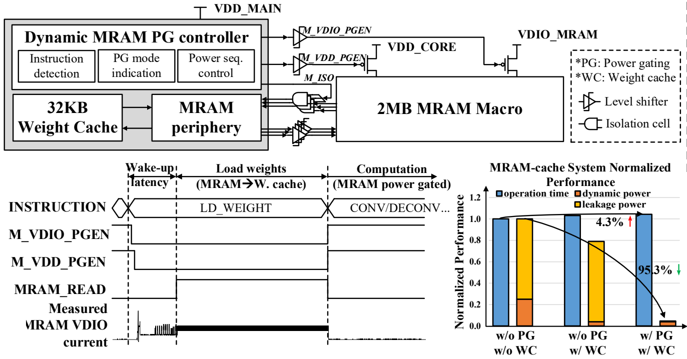

Deconvolution Forward-propagation
IC
OC
W1 W2 W3
W4 W5 W6
W7 W8 W9
A1 A2 A3
B1 B2 B3
Zero-skipping dataflow for Deconvolution forward propagation
8 OCs in parallel, share A
8 columns in parallel, share W
Cycles
PE(1,1,1)
PE(1,2,1)
W1*A1 +W3*A2 +W7*B1 +W9*B2
W1*A2 +W3*A3 +W7*B2 +W9*B3
To Accumulation Memory
W2*A2 +W8*B2
W2*A3 +W8*B3
W4*B1 +W6*B2
W4*B2 +W6*B3
W5*B2
W5*B3
8 ICs in parallel
Convolution-ReLU Back-propagation
A^L: L^{th} layer activation (inference)
A'^L: L^{th} layer gradient
Stored in act mem during inference
FP
BP
CONV
ReLU
FP: $A^L = ReLU(A^{L-1} * W)$
BP: $A'^{L-1} = A'^L * rot(W) \odot ReLU'(A^{L-1})$
$= \begin{cases} 0, & A^{L-1} = 0 \\ A'^L * rot(W), & A^{L-1} > 0 \end{cases}$
L^{th} layer gradient (A'^L):
L-1th layer activation in inference stage (A^{L-1}):
CONV_BP
-0.6
4.3
ReLU_BP
L-1th layer gradient (A'^{L-1}):
$A'^{L-1}1$ =-0.6
$A'^{L-1}2$ =0
Zero-gating dataflow for ReLU-based back-propagation
Do computation
Zero-gating
PE(1,1,1)
PE(1,2,1)
W9*A'^L1
+W1*C'^L3
To Accumulation Memory
Read out A^{L-1} and detect
Control PE array behavior

Fig.5 Efficient dataflow for both forward and backward propagation.



Dynamic MRAM Power Gating Modes Comparison

Power-gating Mode	VDD_MRAM	VDIO_MRAM	Power-up Energy	PG leakage power
Power down mode	off	off	130nJ@4MHz	0.2μW
Sleep mode	off	on	4nJ	22.5μW
Stand-by mode	on	on	0nJ	246μW

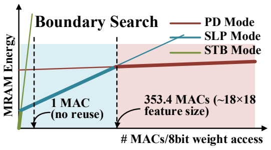


Fig.6 MRAM-cache architecture and dynamic power gating scheme for MRAM.

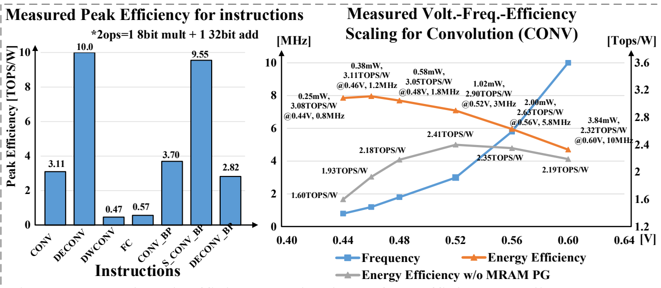
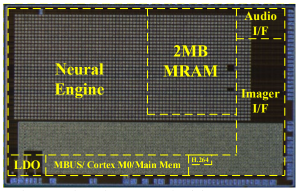
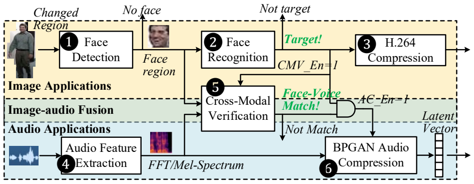


Fig.7 Measured peak efficiency and voltage-freq-efficiency scaling

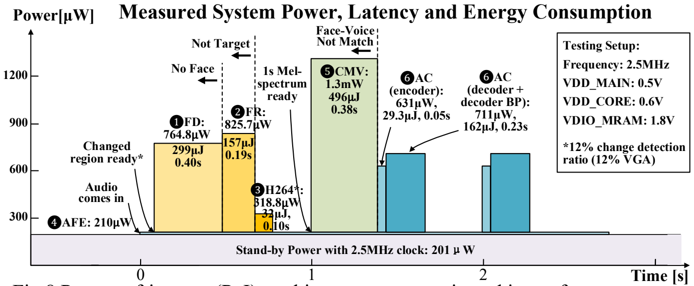


	Specifications
Technology	22nm
Die Area	4.35mm*2.75mm (12mm^2)
Voltage	0.44V-1.0V(Main)/0.6V-1.0V(Core)/1.8V-3.3V(IO)
On-chip Memory	1.98MB SRAM + 2MB MRAM
Max Frequency	70MHz (On-chip CLK gen)
Power Consumption	460nW @ Always-on block
	70μW @ Stand-by Leakage w/ M0 mem on
	0.25mW@0.8MHz, 3.84mW@10MHz
Performance	Sys. Peak Efficiency: 3.1-10TOPS/W

Fig.9 Die photo and chip summary.



Application	# weights [MB]
❶ FD	0.13
❷ FR	0.23
❸ H264	-
❹ AFE	-
❺ CMV	1.20
❻ AC	0.33
Total	1.89



Testing Setup:
Frequency: 2.5MHz
VDD_MAIN: 0.5V
VDD_CORE: 0.6V
VDIO_MRAM: 1.8V

***12% change detection ratio (12% VGA)**

Fig.8 Person-of-interest (PoI) tracking system scenario and its performance.

		This Work	[4]	[5]	[6]	[7]	[8]	[9]
Applications		Face detection, face recognition, audio compression, cross-modal verification	Person detection, face detection, face recognition	Face detection/ face recognition	Keyword spotting, speaker verification	Super resolution	NN inference & training	Face manipulation
Algorithms	CNN/FC	✓	✓	✓	✓(LSTM)	✓	✓	✓
	GAN	✓	-	-	-	✓	-	✓
	BP	✓	-	-	-	-	✓(FC only)	✓
Image processing		Change detection/ JPEG/H264	Change detection /JPEG/H264	-	-	-	-	-
Audio processing		FFT/Mel spectrum	-	-	FFT/MFCC	-	-	-
Technology [nm]		22	40	22	65	65	40	65
Die Area [mm2]		12	27	3.4	2.6	16	29.2	32.4
Non-volatile Memory		2MB MRAM	-	-	-	-	2MB RRAM	-
On-chip SRAM [MB]		1.98	1.13	1.2	0.1	0.56	0.50	0.66
Off-chip Memory access		No	No	-	No	Yes	No	Yes
Precision		INT8	INT8	INT1/16	INT8	INT8	INT8	FP8/FP16
Voltage [V]		0.44-1.0	0.58-0.7	0.65	0.6-1.2	0.75-1.1	1.1	0.7-1.1
Max Frequency [MHz]		70	0.15	180	12.5	200	200	200
Power [mW]		0.25-3.84*	0.17	16.7	0.02	31-211	126	58-647
Sys. Peak Eff. [Tops/W]		3.1-10**	1.5	1.1	0.7	1.9	2.2	1.8

*0.25mW is measured @0.44V, 0.8MHz, 3.84mW is measured @0.6V, 10MHz. **10Tops/W is from deconvolution operation.

Table.1 Comparison table versus state-of-the-art.