IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 6, 2020, accepted December 15, 2020, date of publication December 29, 2020, date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047948

An 8-bit Ring-Amplifier Based Mixed-Signal MAC Circuit With Full Digital Interface and Variable Accumulation Length

JONGHO KIM[1], BEOMKYU SEO[1], YOUNG H. OH[2], JUNG-HOON CHUN[2], (Member, IEEE), JAE W. LEE[3], (Senior Member, IEEE), AND JINTAE KIM[1], (Senior Member, IEEE)

[1]Department of Electrical and Electronics Engineering, Konkuk University, Seoul 05029, South Korea
[2]Semiconductor Systems Engineering, Sungkyunkwan University, Suwon 16419, South Korea
[3]Department of Computer Science and Engineering, Seoul National University, Seoul 08826, South Korea

Corresponding author: Jintae Kim (jintkim@konkuk.ac.kr)

This work was supported in part by the Nanomaterial Technology Development Project of the Ministry of Science and Technology under Grant 2016M3A7B4909668, in part by the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT under Grant 2020M3F3A2A01085756, and in part by the Industrial Innovation Technology Future Semiconductor Project of the Ministry of Industry and Commerce under Grant 10080611.

ABSTRACT An 8-bit switched-capacitor multiply-and-accumulator (MAC) in 65nm CMOS is presented. Based on a cascaded low-power ring-amplifier-based switched-capacitor DACs, the MAC circuit features a programmable accumulation length in MAC computation. Fabricated in 65nm CMOS, the prototype MAC circuit achieves a precision-scaled energy efficiency of 1.32fJ per MAC operation, which is comparable to other state-of-the-art MAC circuits, along with best-in-class linearity. The noise performance has been verified using four real-world convolutional neural networks (CNNs) and 10,000-image data sets with up to 1,000 classes with an accuracy drop of less than 2% compared to the baseline 32-bit floating-point MAC.

INDEX TERMS Switched-capacitor, convolutional neural network, deep learning, ring amplifier, near-memory computing.

I. INTRODUCTION

In-memory or near-memory computing is an increasingly popular design paradigm that aims to minimize the overall power dissipation of the convolutional neural network (CNN) computation for highly energy-constrained machine learning applications. The CNN computation, consisting mainly of numerous multiply-and-accumulate (MAC) operations, is highly memory intensive. As the cost of data movements greatly outweighs that of MAC operation [1], reducing the external memory access is a key to achieving high energy efficiency. Merging the weight-storing on-chip cache memory and the MAC unit, therefore, can greatly reduce the required number of power-hungry DRAM access, leading to overall energy reduction.

While the prevailing MAC unit design is based on digitally-synthesized multipliers and adders, an alternative method utilizing analog-domain signal processing is attracting considerable attention due to its potential to reduce overall power dissipation, particularly for low-resolution MAC computation. For instance, a passive switched-capacitor array followed by an 8-bit ADC realizes an 8-bit MAC function at low power dissipation [3]. While showing excellent power efficiency, the fixed accumulation length due to the passive charge-domain averaging may limit its applicability. This is particularly true for recent light-weight neural networks such as MobileNets where the accumulation length of the convolution varies from 9 to 1024 [4]. A time-domain signal processing circuit is also proposed to implement MAC function. For instance, gated ring oscillator (GRO) based analog MAC in [5] realizes a highly area-efficient solution. However, the inherent nonlinearity of the voltage-to-frequency conversion in VCO may prevent its usage for a general-purpose neural network computation.

Recently, an SRAM based in-memory architecture is being actively explored. A standard 6T SRAM array employing pulse width modulated (PWM) word-line (WL) drivers followed by passive charge sharing 8-bit MAC units and an ADC

The associate editor coordinating the review of this manuscript and approving it for publication was Poki Chen.

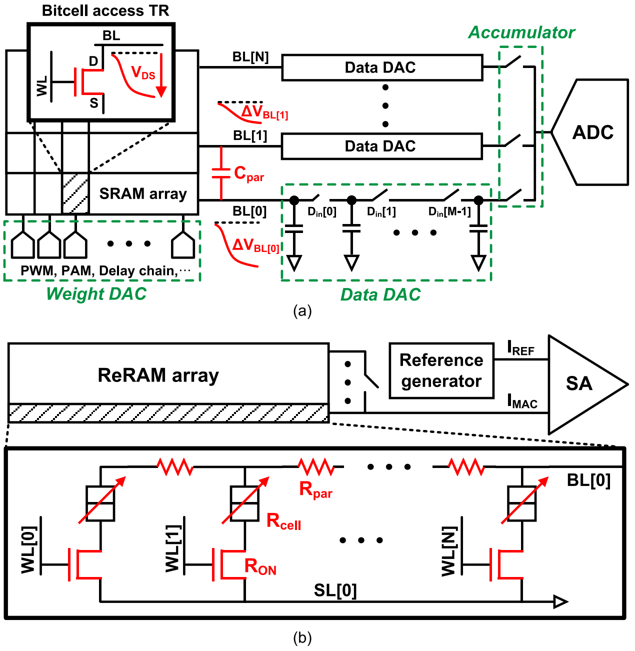
achieves accumulation length of 128 [6, 7], but it suffers from the linearity degradation due to the spatial variation of the threshold voltage of the bit-cells in SRAMs. Some works modify bit-cell design to realize binary or low-resolution MAC computation [8]–[11]. However, such customized bit-cell structures, with inevitable area-per-cell increase, can only be realized at the expense of lower area efficiency when compared to the foundry-provided SRAMs. Also, the state-of-the-art CNN algorithm requires the resolution of at least 8-bit for generality [12].

In this paper, we present a mixed-signal MAC circuit that performs an 8-bit linear MAC computation. Having a fully digital input/output interface, the proposed MAC is natively compatible with standard SRAM macros. In addition, our MAC features a programmable accumulation length in MAC computation as opposed to the fixed accumulation length in passive charge sharing based MACs [3], [6], [7]. The programmable accumulation is made possible by an active integrator circuit in our MAC structure. While the active integrator tends to incur higher power dissipation than the passive one, this work minimizes the overall current consumption by utilizing a low-power ring amplifier in the integrator design, leading to the state-of-the-art precision-scaled energy efficiency when compared to other mixed-signal MACs fabricated in the same process nodes. Section II reviews prevailing mixed-signal MAC circuits in detail. Section III describes the overall MAC architecture. Section IV presents the circuit-level details of the MAC. Section V reports measured results as well as the verification of the MAC computation using a behavioral MAC model based on a software-based inference framework. Section VI concludes the paper.

II. REVIEW OF MIXED-SIGNAL MAC

Previous works on mixed-signal MAC architecture can be classified into two types depending on the physical domain where the MAC operation occurs: voltage and current. Illustrated in Fig. 1-(a), the voltage-based architecture is mainly comprised of a cascade of weight DACs and data DACs, followed by a passive accumulator and an ADC [6]–[8], [11]. The weight DAC is generally combined with SRAM weight storage cache memory with multibit weight inputs applied through pulse-width modulation (PWM) or pulse-amplitude modulation (PAM) word-line (WL) drivers. The access transistor of SRAM bit-cell acts as a constant current source and generates the DAC output on the bit-line (BL). However, the channel length modulation effect due to the BL voltage drop and the capacitive coupling from the adjacent BLs degrade the overall MAC linearity, potentially degrading the inference accuracy in real applications. The data DAC typically uses passive charge sharing to realize the low-power analog multiplier. The accumulation is simply performed by shorting the multiplier outputs, which is equivalent to summation in the charge domain.

As shown in Fig. 1-(b), a current-based MAC is an alternative architecture that is inherently compatible with emerging resistor-based non-volatile memories such as ReRAM or MRAM. The memory array stores the weight and the output of MAC operation is measured in the current domain by a current-domain sense amplifier at the backend. To achieve multi-bit computation, successive approximation via reference generator using a replica array [13], [14] or multi-level cell [15] has been employed. Although having an advantage of being compatible with emerging non-volatile memories, the MAC output precision is constrained by the limited signal margin owing to the finite on-off ratio of cell resistance, mismatch of access transistors, and high parasitic load on the read path. Some of the previous works achieve a peak energy efficiency of nearly hundreds of TOPS/W, showing a great promise when applied to the low power edge device applications. However, many of them resort to the low precision (<4-bit) computation, which limits its practical usage considering that recent mainstream machine learning algorithms often require at least 8-bit of numerical resolution. Also, many previous works require customized memory cells or modification of peripheral circuitry which makes it difficult to use these MACs. Based on these observations, this work proposes a mixed-signal 8–bit resolution MAC structure that is compatible with a fully digital interface and therefore can be seamlessly integrated with standard memory macros.



(a)

(b)

FIGURE 1. Block diagram of (a) voltage-based MAC and (b) current-based MAC circuit.

III. OVERALL MAC STRUCTURE

Fig. 2 displays the proposed multi-bit MAC structure that consists of a cascade of two 8-bit switched-capacitor DACs followed by an 8-bit SAR ADC. The DAC generates an analog-equivalent voltage for a digital input with a scaling factor V_{REF}. Therefore, the front-end input DAC denoted as

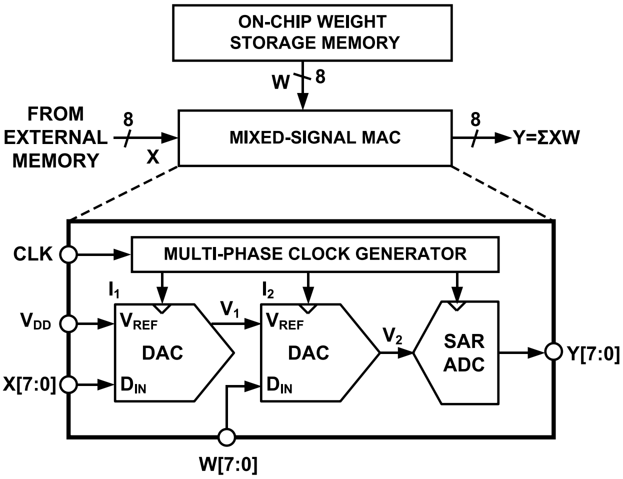


FIGURE 2. Block diagram of the proposed MAC circuit.

I_1 generates a differential analog voltage V_1 corresponding to an 8-bit digital input X[7:0], *i.e.,*

$$V_1 \approx V_{REF} \cdot \frac{X_D}{X_{FS}},$$

where X_D is a decimal representation of the 8-bit input X[7:0], X_{FS} is the full scale of the X_D, and V_{REF} is the reference voltage of the DAC. In this work, we use supply voltage V_{DD}=1.2V as the reference voltage for I_1. The LSB of the 8-bit DAC, which is 4.68mV, is larger than the typical noise level of the supply source such as a low-dropout regulator, so this arrangement does not impose strict noise requirement. Also, we added oversized 0.5nF of on-chip bypass capacitance to reduce the voltage ripple below $500\mu V$ in the reference voltage at the cost of using large silicon area. The following DAC, I_2, takes the output of the first DAC, V_1, as a reference voltage and W[7:0] as an input, generating a differential output voltage V_2, *i.e.,*

$$V_2 \approx V_1 \cdot \frac{W_D}{W_{FS}} = V_{REF} \cdot \frac{W_D \cdot X_D}{W_{FS} \cdot X_{FS}},$$

where W_D and W_{FS} are the decimal representation of the 8-bit input W[7:0] of I_2 and the full scale of the X_D, respectively. Therefore, V_2 is an analog representation of the product of X_D and W_D with an error bounded by the resolution of the DAC.

The DAC, whose detail is shown in Fig. 3, is implemented based on a switched-capacitor integrator. In our structure, adding the accumulation operation is almost effortless. Specifically, the MAC operation over N_{acc} cycles is achieved by resetting the output of the second integrator only once every N_{acc} cycle where N_{acc} is the length of accumulation in the MAC. In other words, the charge in the second integrator is kept in successive multiplication operations during the accumulation. In this structure, the accumulation length of the MAC can be easily adjusted by modifying the period of the reset pulse. Lastly, the output of MAC DAC is converted to digital via an 8-bit SAR A/D converter, generating digital output Y[7:0].

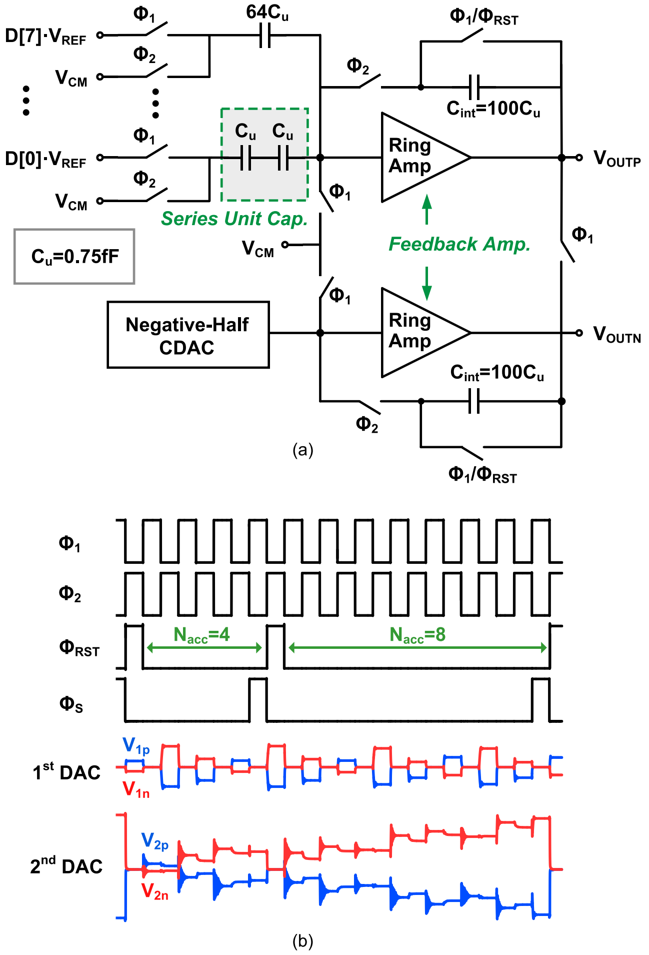


FIGURE 3. (a) Circuit-level detail of the MAC DAC topology and (b) simulated clock and waveforms of the MAC.

From the perspective of reducing overall power consumption, analog-domain accumulation in our structure is advantageous in comparison to implementing the accumulati-on in digital circuits. This is because using the digital accumulation in combination with the analog-domain multiplier requires the A/D conversion to run at every multiplication operation. On the other hand, with analog accumulation, the A/D conversion occurs only once every N_{acc} multiplication operation. With a lower A/D conversion rate, A/D conversion power can be reduced, leading to lower overall power consumption.

IV. CIRCUIT IMPLEMENTATION

Illustrated in Fig. 3-(a), the core of the MAC circuit is the switched-capacitor differential DAC. Two ring amplifiers are used as a differential feedback amplifier for the best possible immunity to the possible common-mode environmental noise such as supply noise. We used a small unit capacitance of C_u=0.75fF, which is smaller than typically allowable unit capacitance for full 8-bit linearity. Specifically, our model-based investigation reveals that the estimated percentage mismatch of the unit capacitor is 0.86%, which is around 2 times of the LSB in an 8-bit accuracy. However, our intended application is tolerant to a few LSBs of errors, which will be verified in the model-based verification in Section V, and we take advantage of such a property to keep C_u small in favor of lower overall power dissipation.

When Φ_1=HIGH, the digital input D[7:0] is sampled as the charge in the binary-scaled capacitor array where the total capacitance is $128C_u$. When Φ_2=HIGH, the charge held in the input capacitor array is transferred to $C_{int}=100C_u$ by the feedback amplifier. From the law of charge conservation, one can show that the differential output voltage is given by

$$V_{OUTP} - V_{VOUTN} \approx \frac{D[7]\cdot 64C_u + D[6]\cdot 32C_u + \cdots + D[0]\cdot 0.5C_u}{100C_u} V_{REF},$$

which is a linearly scaled analog representation of the decimal value of D[7:0] with the scaling factor being a linear function of V_{REF}. Both DACs in Fig. 1 are identical blocks except that the reference voltage of the second DAC is driven by the output of the first DAC. Therefore, the output of the second DAC is a linearly scaled version of the product of X[7:0] and W[7:0]. Note that for the first input DAC, the integration capacitor C_{int} is reset every cycle when Φ_1=HIGH. However, for the second DAC, C_{int} is reset for every N_{acc} cycle when Φ_{RST}=HIGH, thereby realizing the accumulation in the charge domain. Fig.3-(b) shows the clock and signal waveforms of the MAC with an exemplary output voltage of each DAC stage. The simulated waveforms show that the outputs which reset the integration capacitor C_{int}. The waveform in Fig. 3-(b) illustrates two such cases when N_{acc}=4 and N_{acc}=8. Note that the 2^{nd}-stage output exhibits some ringing at the beginning of the integration phase, which is expected behavior for a ring amplifier. The dynamic biasing scheme inherent in a ring amplifier ensures that the feedback is stable after a few ringings during the integration phase.

At the end of each N_{acc} cycle, the backend SAR ADC samples the final value of MAC output voltage when Φ_S=HIGH, which is subsequently digitized by a following A/D converter.

Fig. 4-(a) shows a feedback amplifier circuit. The amplifier is based on the ring amplifier [16], which is essentially three-stage cascaded inverters with dynamic biasing. The sizing and component values are re-optimized for our intended operation. The settling of the ring amplifier is not sensitive to the selection of V_{OS}, which is 20mV in our chip, as long as $V_{OS} \geq |V_{DD} - 2V_T|/2A_2$ is used where A_2 is the gain of the 2^{nd} inverter in Fig. 4-(a) and V_T is the threshold voltage of the transistors [16]. In the design process, we ensured that such a requirement for offset voltage can be maintained over the process and temperature corners. For interested readers, more detailed simulations on the offset voltage requirement for process corners are described in the Appendix. A primary reason to use the ring amplifiers for this application is the process scalability of the inverter-based topology. Since this MAC circuit is intended to be used in conjunction with SRAM, the amplifier topology should be scalable in pace with SRAM scaling. Also, the dynamic biasing scheme in the ring amplifier helps reduce overall power dissipation. Additionally, although the inverter threshold voltage in the ring amplifier is sensitive to the supply voltage and process and temperature variation, the DAC linearity is insensitive to the threshold voltage variation or the current consumption of the inverters. This is because the precision of the DAC output is defined by the ratio of the capacitors as long as the overall gain of the cascaded inverters remains sufficiently high.

During the sampling phase with Φ_1=HIGH, the input signal is sampled on C_S while the C_{CM1} and C_{CM2} sample the common-mode difference between V_{CM} and V_{GS1}, which is used in the common-mode feedback during the amplification phase. At the same time, the capacitor C_{OS1} and C_{OS2} store

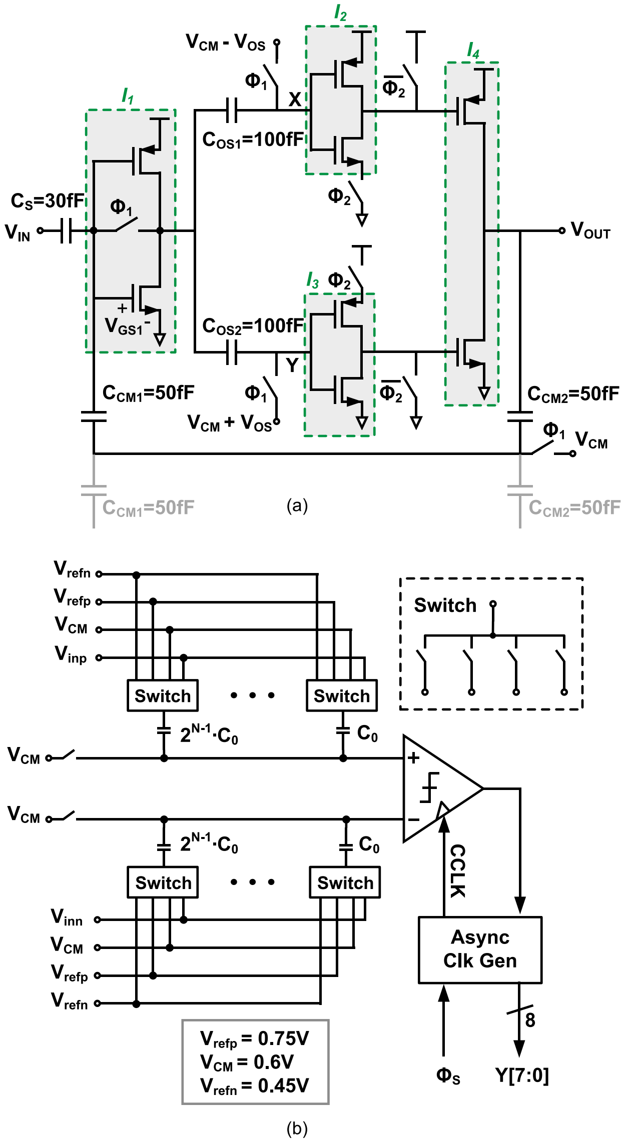


FIGURE 4. Circuit-level detail of (a) the ring amplifier and (b) the back-end SAR ADC.

the offset voltage between V_{GS1} and $V_{CM}\pm V_{OS}$. During the amplification phase with Φ_2=HIGH, the two inverters I_2 and I_3 are activated and amplify the sampled signal initially. The offset voltages between node X and Y stored by two offset capacitors are also amplified such that the amplified offset eventually drives X and Y close to V_{DD} and G_{ND}, respectively. As a result, I_2, I_3, and I_4 in the ring amplifier to draw nearly no current in the steady-state when the output is settled. Note that each ring amplifier is a cascaded three inverters, so the total gain is large enough to ensure sufficiently small gain error for 8-bit accuracy when the output is settled. In addition, it is worth pointing out that the low-frequency noise such as flicker noise or random offset of the inverters is not detrimental as pointed out in [16] because the signal is AC-coupled and the 2^{nd}-stage inverter offset is attenuated by the gain of the 1^{st} stage amplifier.

Fig. 4-(b) shows the structure of the backend 8-bit asynchronous SAR ADC where a binary-scaled capacitor DAC (CDAC) is implemented using unit capacitor C_0=20fF. Unlike the top-plate sample scheme used in [2], a bottom-plate sampling scheme is used for the CDAC, which has the advantage of being immune to the input full-scale range reduction due to the parasitic capacitance of the CDAC. This characteristic is critical because any gain error in A/D conversion will hurt overall MAC operation accuracy.

V. MEASUREMENTS AND VERIFICATIONS

The prototype MAC chip was fabricated in 65nm CMOS process with V_{DD}=1V. Fig. 5 shows a microphotograph of the chip as well as the measurement board. The MAC block occupies an active area of 370μm by 270μm and consumes 101μW when running at 75MHz. V_{REFP}=0.75 [V] and V_{REFN}=0.45[V] for the ADC are externally provided and are chosen to match the ADC input full-scale to the output of the DAC. The performance of the individual ADC and DAC was not characterized since the output of the DAC is only measurable by the internal ADC.

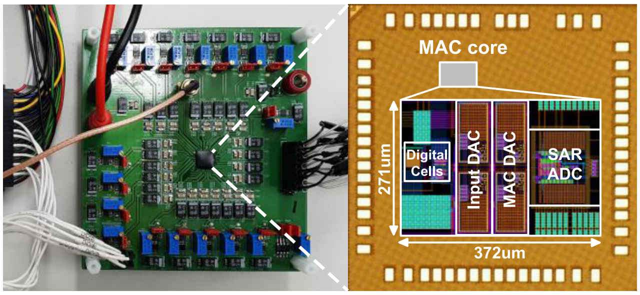


FIGURE 5. Die photograph and test board.

Fig. 6 shows measured multiplier transfer characteristic for 9 different weights while the input code X is swept from −127 to +127. To assess the linearity of the multiplication, attenuation error is removed via least-square fit and the resulting error is also plotted for different W values. The measured nonlinearity error is bounded by ±3LSBs.

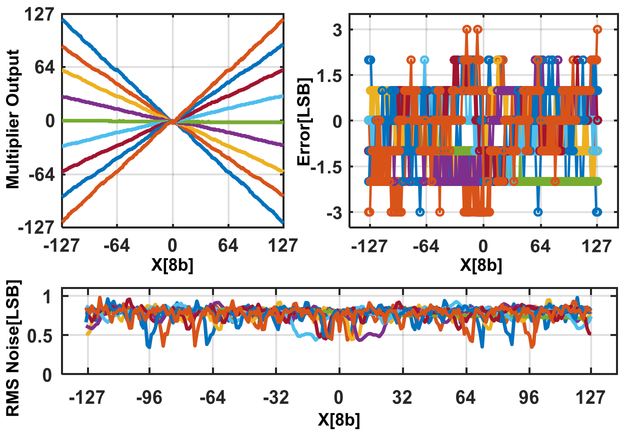


FIGURE 6. Linearity and noise performance.

Unlike digital implementation, analog MAC is affected by the thermal noise. Noise performance is also characterized as well with varying W values, which is also shown in Fig. 6. On average, the RMS noise in the multiplier output is 0.77LSB, which means that the effective resolution of the multiplier is 6.5 effective-bit. As will be verified by our model-based verification flow, our application, the convolutional neural network computation, is very tolerant to random errors, and therefore this level of random error does not incur any noticeable inference accuracy drop.

To demonstrate the active accumulation feature, Fig. 7 shows the measured output when active accumulation is enabled for N_{acc}=4 and N_{acc}=8. As can be seen in the measurement with a fixed weight value, the accumulator outputs versus input data linearly increase with respective slopes for N_{acc}=4 and N_{acc}=8. The error in the final accumulation result is less than 5LSBs. Due to the finite output signal range, the output may saturate when large DC input is accumulated over multiple cycles. However, in our intended applications, the weights of convolutional neural networks are close to zero for the most part, so such saturation is not likely to cause problems.

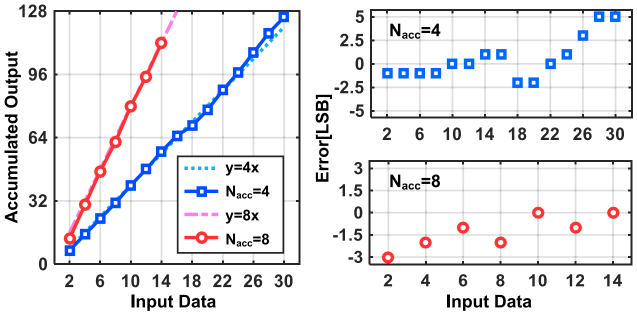
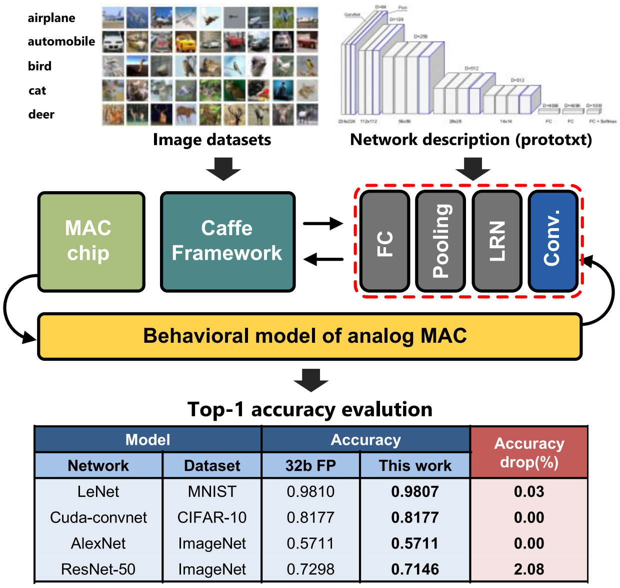


FIGURE 7. Measured accumulation characteristic.

For extensive system-level verification, a behavioral multiplier model is developed as

$$Y = 127 \cdot round\left(\frac{X \cdot W}{127} + 0.77 \cdot randn() - 0.073\right),$$



Model		Accuracy		Accuracy drop(%)
Network	Dataset	32b FP	This work	
LeNet	MNIST	0.9810	**0.9807**	**0.03**
Cuda-convnet	CIFAR-10	0.8177	**0.8177**	**0.00**
AlexNet	ImageNet	0.5711	**0.5711**	**0.00**
ResNet-50	ImageNet	0.7298	**0.7146**	**2.08**

FIGURE 8. **Software-based verification flow.**

where 0.77LSB random noise has been added to match with measured noise performance. This model is then integrated into Caffe [17], a popular deep-learning software framework, to evaluate end-to-end performance with 4 pre-trained CNNs: LeNet, Cuda-convnet, AlexNet, and ResNet-50. The overall verification flow is illustrated in Fig. 8. Note that due to the limited accumulation length in our MAC circuit, longer accumulations are performed in digital domain. Our evaluation reveals the noise performance of the proposed 8-bit MAC DAC is good enough to maintain a comparable classification accuracy to that of the 32-bit digital floating-point (FP) MAC with a worst-case accuracy drop of 2.08%. Table 1 compares the performance of this chip with similar MACs fabricated in similar process nodes. We primarily focus on the comparison using the precision-scaled MAC energy, which is a figure-of-merit for MAC computation used in the latest paper [18]. Table 1 shows that our measured energy efficiency is about 30% better than that of the full digital MACs in the same process node and comparable to those of the mixed-signal MACs with similar resolutions that are fabricated in the same process nodes. In addition, our mixed-signal MAC is the first work that demonstrates the potential of implementing variable accumulation length in analog domain.

TABLE 1. **Performance Comparison.**

	[2]	[5]	[12]	**This Work**
Technology	65nm	65nm	55nm	**65nm**
MAC Architecture	Digital	Switched-capacitor Passive Accumulation	ReRAM based Current Accumulation	**Switched-capacitor Active/Variable Accumulation**
Supporting Variable Kernel Size	Yes	No	N/A	**Yes**
Need Modification in Memory Macro	No	Yes	Yes	**No**
MAC Precision (Feature/Weight /Output)	8b/8b/8b	8b/8b/4b	2b/3b/4b	**8b/8b/8b**
Linearity (Worst-case Error/ RMS Noise)	-	N/A	N/A	**3LSB/ 0.77LSB**
Efficiency [TOPS/W]	1.038	3.125	21.9	**1.46~1.478**
*Precision-scaled MAC Energy[fJ]	1.88	1.25	1.90	**1.32~1.34**
Data set	N/A	MIT-CBCL(N/A)	CIFAR-10(10)	**MNIST(10), CIFAR-10(10), ImageNet(1000)**
Algorithm	AlexNet (CNN)	SVM	N/A (CNN)	**LeNet, Cuda-convnet, AlexNet, ResNet-50 (CNN)**

*Precision-scaled MAC Energy = E_{MAC} / (B_x x B_w x B_y)
(E_{MAC}: MAC energy, B_x: feature precision, B_w: weight precision, B_y: output precision)

VI. CONCLUSION

Our work demonstrates a prototype analog MAC implementation that supports variable accumulation length with true 8-bit linearity. When combined with SRAM-based array implementation, the proposed MAC can fully support a very wide variety of convolution filter sizes in the analog domain. State-of-the-art measured energy efficiency and the inference accuracy verified by the software-based flow prove the viability of the proposed MAC structure. Since the only analog block in our structure is an inverter-like amplifier, the energy efficiency of our MAC is expected to scale well with more aggressively scaled CMOS technology.

APPENDIX

The ring amplifier shown in Fig. 4-(a) exhibits some ringing behaviors. This is expected due to use of a dynamic biasing scheme. To maintain stability, one needs to carefully choose offset voltage V_{OS} that is stored across C_{OS1} and C_{OS2}. The original paper [16] reports that for the stability, the offset voltage has to meet the minimum requirement $V_{OS} \geq |V_{DD} - 2V_T|/2A_2$ where V_T is the threshold voltage and A_2 is the second-stage inverter amplifier gain. One may wonder what will impose the upper bound for V_{OS}. From our experiments, we found that excessively high offset voltage leads to linearity degradation of the entire switched-capacitor DAC using this ring amplifier. We believe that such a linearity degradation results from the reduced output voltage range, and that is again associated with a higher initial offset voltage stored in the AC coupling capacitors. Therefore, one needs to consider both the linearity and the stability in choosing the offset voltage in a ring amplifier.

Fig. 9 shows simulated output of the 2nd-stage DAC to show the settling behavior of the switched-capacitor MAC over three process (fast-fast, slow-slow and nominal-nominal) and three temperature ($0C°$, $27C°$, and $70C°$) corners. Note that we use different offset voltages (displayed in the title of each subfigure) for each process corner, but the settling is stable over all temperature corners for the same V_{OS} value in each process corner. In other words, if we can choose proper offset voltage for a given chip, the operation will be stable over the entire temperature range.

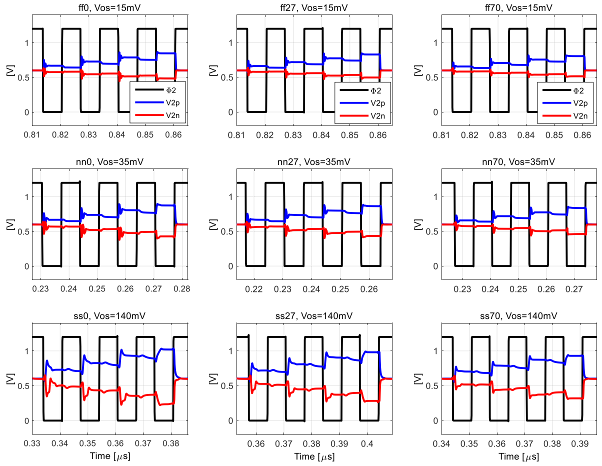


FIGURE 9. The simulated settling response of the switched-capacitor DAC using a ring amplifier over the various process and temperature conditions.

In our prototype chip, the offset voltage is provided externally for the sake of easy chip measurement, but one may need an on-chip process-dependent offset voltage generator for a fully-integrated solution.

ACKNOWLEDGMENT

The authors would like to thank IC Design Education Center (IDEC), KAIST, Daejeon, Korea, for CAD tool support. They also acknowledge Samsung Electronics for the chip fabrication. *(Jongho Kim and Beomkyu Seo contributed equally to this work.)*

REFERENCES

[1] M. Horowitz, "Computing's energy problem (and what we can do about it)," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2014, pp. 10–14.

[2] Z. Yuan, J. Yue, H. Yang, Z. Wang, J. Li, Y. Yang, Q. Guo, X. Li, M.-F. Chang, H. Yang, and Y. Liu, "Sticker: A 0.41-62.1 TOPS/W 8Bit neural network processor with multi-sparsity compatible convolution arrays and online tuning acceleration for fully connected layers," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, Jun. 2018, pp. 33–34.

[3] D. Bankman and B. Murmann, "An 8-bit, 16 input, 3.2 pJ/op switched-capacitor dot product circuit in 28-nm FDSOI CMOS," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Toyama, Japan, Nov. 2016, pp. 21–24.

[4] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[5] K. Yoshioka, Y. Toyama, K. Ban, D. Yashima, S. Maya, A. Sai, and K. Onizuka, "PhaseMAC: A 14 TOPS/W 8bit GRO based phase domain MAC circuit for in-Sensor-Computed deep learning accelerators," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, Jun. 2018, pp. 263–264.

[6] S. K. Gonugondla, M. Kang, and N. Shanbhag, "A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 490–491.

[7] M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, "A multifunctional in-memory inference processor using a standard 6T SRAM array," *IEEE J. Solid-State Circuits*, vol. 53, no. 2, pp. 642–655, Feb. 2018.

[8] A. Biswas and A. P. Chandrakasan, "Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 488–490.

[9] W.-S. Khwa, J.-J. Chen, J.-F. Li, X. Si, E.-Y. Yang, X. Sun, R. Liu, P.-Y. Chen, Q. Li, S. Yu, and M.-F. Chang, "A 65 nm 4 Kb algorithm-dependent computing-in-memory SRAM unit-macro with 2.3 ns and 55.8TOPS/W fully parallel product-sum operation for binary DNN edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers)*, San Francisco, CA, USA, Feb. 2018, pp. 496–497.

[10] J. Yang, Y. Kong, Z. Wang, Y. Liu, B. Wang, S. Yin, and L. Shi, "Sandwich-RAM: An energy-efficient in-memory BWN architecture with pulse-width modulation," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 394–395.

[11] X. Si, J.-J. Chen, Y.-N. Tu, W.-H. Huang, J.-H. Wang, Y.-C. Chiu, W.-C. Wei, S.-Y. Wu, X. Sun, R. Liu, S. Yu, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, Q. Li, and M.-F. Chang, "A twin-8T SRAM computation-in-memory macro for multiple-bit CNN-based machine learning," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 396–397.

[12] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," 2017, *arXiv:1712.05877*. [Online]. Available: http://arxiv.org/abs/1712.05877

[13] W.-H. Chen, K.-X. Li, W.-Y. Lin, K.-H. Hsu, P.-Y. Li, C.-H. Yang, C.-X. Xue, E.-Y. Yang, Y.-K. Chen, Y.-S. Chang, T.-H. Hsu, Y.-C. King, C.-J. Lin, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 65 nm 1Mb nonvolatile computing-in-memory ReRAM macro with sub-16 ns multiply- and-accumulate for binary DNN AI edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 494–495.

[14] C.-X. Xue *et al.*, "A 1Mb multibit ReRAM computing-in-memory macro with 14.6ns parallel MAC computing time for CNN based AI edge processors," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, CA, USA, Feb. 2019, pp. 388–389.

[15] R. Mochida, K. Kouno, Y. Hayata, M. Nakayama, T. Ono, H. Suwa, R. Yasuhara, K. Katayama, T. Mikawa, and Y. Gohou, "A 4M synapses integrated analog ReRAM based 66.5 TOPS/W neural-network processor with cell current controlled writing and flexible network architecture," in *Proc. IEEE Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2018, pp. 175–176.

[16] B. Hershberg, S. Weaver, K. Sobue, S. Takeuchi, K. Hamashita, and U.-K. Moon, "Ring amplifiers for switched capacitor circuits," *IEEE J. Solid-State Circuits*, vol. 47, no. 12, pp. 2928–2942, Dec. 2012.

[17] Caffe. *Deep Learning Framework*. Accessed: Jan. 10, 2019. [Online]. Available: http://caffe.berkeleyvision.org

[18] J. Su *et al.*, "A 28nm 64Kb inference-training two-way transpose multibit 6T SRAM compute-in-memory macro for AI edge chips," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, CA, USA, Feb. 2020, pp. 240–241.

JONGHO KIM received the B.S. degree in electrical engineering from Konkuk University, Seoul, South Korea, in 2018, where he is currently pursuing the Ph.D. degree. He is currently conducting a research on an in-memory-computing analog convolution engine for image processing neural networks. His research interests include data converter circuits and neural network computation engines.

BEOMKYU SEO received the B.S. and M.S. degrees in electrical engineering from Konkuk University, Seoul, South Korea, in 2017 and 2018, respectively. He is currently conducting a research on a low-power switched-capacitor convolution neural-network engine for image processing neural networks. His research interest includes mixed-signal circuit designs for convolutional neural network computation.

YOUNG H. OH received the B.S. degree in information and communication engineering from SungKyunKwan University (SKKU), Suwon, South Korea, in 2013, where he is currently pursuing the M.S. and Ph.D. degrees with the Department of Electrical and Computer Engineering. His research interests include specialized hardware architectures and system-level optimization for DNN accelerators.

JUNG-HOON CHUN (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1998 and 2000, respectively, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2006. From 2000 to 2001, he was with Samsung Electronics, Hwasung, South Korea, where he developed BiCMOS RF front-end IC for wireless communication. From 2006 to 2008, he was with Rambus Inc., Los Altos, CA, USA, where he was involved in high-speed serial interfaces, such as FlexIO, XDR, and XDR2. He is currently a Professor with Sungkyunkwan University, Suwon, South Korea. His current research interests include high-speed serial link, CMOS imagers, on-chip ESD protection and I/O design, and new memory devices. He served on the Technical Program Committee for the IEEE A–SSCC, from 2009 to 2011 and from 2014 to 2020. He received the Gold Medal from Humantech Thesis Competition, in 1998, the Benhamou SGF Fellowship, from 2003 to 2005, the IEEE CICC the Best Paper Award, in 2008, the IEEE SOI Conference the Best Paper Award, in 2010. He was a co-recipient of the IEEE ISSCC Silkroad Award, in 2020.

JAE W. LEE (Senior Member, IEEE) received the Ph.D. degree in computer science from the Massachusetts Institute of Technology (MIT). He is currently an Associate Professor with the Department of Computer Science and Engineering, Seoul National University. His research interests include computer architecture, VLSI design, parallel programming, and computer security.

JINTAE KIM (Senior Member, IEEE) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 1997, and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, CA, USA, in 2004 and 2008, respectively. He held various industry positions at Barcelona Design, CA, USA; SiTime Corporation, CA, USA; and Agilent Technologies, CA, USA, as a Key Technical Contributor for their high-speed A/D converters and timing IC products. He is currently an Associate Professor with the Department of Electrical and Electronics Engineering, Konkuk University, Seoul, South Korea, where he is focusing on low power mixed-signal IC designs for communication and sensor applications. He has served on the Technical Program Committee for the IEEE A–SSCC, since 2016. He was a recipient of the IEEE Solid-State Circuits Predoctoral Fellowship, in 2007.