A Bit-level Sparsity-aware SAR ADC with Activity-scaling for AIoT Applications

A Bit-level Sparsity-aware SAR ADC with Direct Hybrid Encoding for Signed Expressions for AIoT Applications

Ruicong Chen
raychen@mit.edu
MIT
Cambridge, MA, USA

H.T. Kung
kung@harvard.edu
Harvard
Cambridge, MA, USA

Anantha Chandrakasan
anantha@mit.edu
MIT
Cambridge, MA, USA

Hae-Seung Lee
hslee@mtl.mit.edu
MIT
Cambridge, MA, USA

ABSTRACT

In this work, we propose the first bit-level sparsity-aware SAR ADC with direct hybrid encoding for signed expressions (HESE) for AIoT applications. ADCs are typically a bottleneck in reducing the energy consumption of analog neural networks (ANNs). For a pre-trained Convolutional Neural Network (CNN) inference, a HESE SAR for an ANN can reduce the number of non-zero signed digit terms to be output, and thus enables a reduction in energy along with the term quantization (TQ). The proposed SAR ADC directly produces the HESE signed-digit representation (SDR) using two thresholds per cycle for 2-bit look-ahead (LA). A prototype in 65nm shows that the HESE SAR provides sparsity encoding with a Walden FoM of 15.2fJ/conv.-step at 45MS/s. The core area is 0.072mm^2.

CCS CONCEPTS

• **Computer systems organization → Neural networks**; • **Hardware → Integrated circuits**.

KEYWORDS

analog neural networks (ANNs); SAR ADC; in-memory computing; resistive RAM (RRAM); hybrid encoding for signed expressions (HESE); sigend-digit representation (SDR); bit-level sparsity; artificial intelligence of things (AIoT); convolutional neural networks (CNNs); term quantization (TQ)

ACM Reference Format:
Ruicong Chen, H.T. Kung, Anantha Chandrakasan, and Hae-Seung Lee. 2022. A Bit-level Sparsity-aware SAR ADC with Direct Hybrid Encoding for Signed Expressions for AIoT Applications. In *ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED '22), August 1–3, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3531437.3539700

ISLPED '22, August 1–3, 2022, Boston, MA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9354-6/22/08.
https://doi.org/10.1145/3531437.3539700

1 INTRODUCTION

An internet of things (IoT) system with artificial intellegence (AI) is referred as an AIoT system. The application space of AIoT is huge, ranging from fundamental research to personal daily life. AIoT has great potential in the future. By 2030, ~350 billion AIoT devices are expected to be in operation, reaching $16 trillions, or 14% of total GDP [12]. With the increasing need for edge computing and long battery life, AIoT devices with low standby power and high efficiency for neural network inference are in great demand. The applications can include microphones, industry-monitoring, vital-sign monitoring devices, etc. End devices with speech interfaces can benefit greatly from ultra-low power AIoT devices, such as Voice Activity Detection (VAD) and Keyword Spotting (KWS) [15] systems. Both VAD and KWS systems have to be always-on and highly efficient in inference. AIoT devices are also ubiquitous in machine health monitoring products that minimize downtime with sensor signals [2]. Moreover, the AIoT system with reconfigurable rectenna opens up the opportunities for wireless and battery-less in-body vital-sign monitoring [1].

Conventional AIoT systems, however, still needs to improve their energy-efficiency. For battery constrained AIoT systems, energy-efficient implementations would bring longer battery life and better user experience. For AIoT systems with connected power sources, low power designs are still preferable as they are more environmentally friendly. Some features in conventional AIoT systems can be explored to lower the power consumption and improve the energy-efficiency. For sensing, the input signals usually have low activity. For computing, the data in embedded neural networks are highly sparse. For memory accessing, data reuse can improve energy efficiency associated with the memory wall in the von Neumann architecture.

AI algorithms based on CNNs, coupled with their high computational requirements, have stimulated the development of novel energy-efficient hardware, such as Eyeriss [4]. Analog neural networks (ANNs) with in-memory computing (IMC) using resistive random-access memory (RRAM) [19] are promising architectures to reduce latency and increase energy efficiency for IoT devices [1]. However, interface circuitry including analog-to-digital converters (ADCs) between RRAM and digital components is becoming the bottleneck of the RRAM-based ANNs [14][17].

Previous work applies term quantization to the weights [13] and uses binary encoding [14][5] in ANNs. We propose to apply hybrid

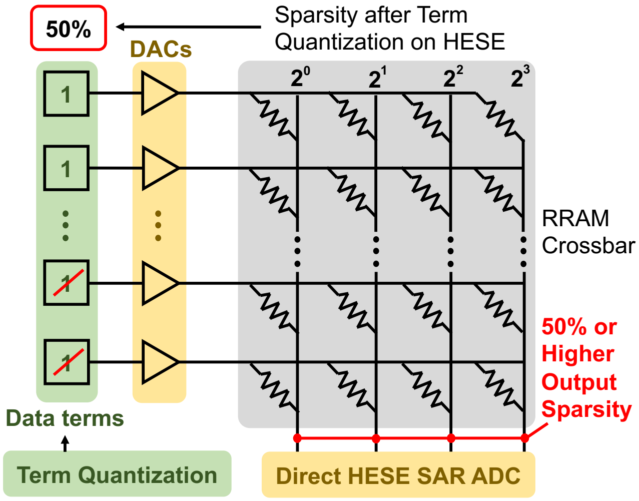


Figure 1: An example crossbar of RRAM with 1-bit RRAM cells and 1-bit input values for computing dot products. Both data and weights are bit-sliced, with each weight term occupying a separate RRAM column. The proposed SAR ADC directly provides the hybrid encoding for signed expressions (HESE) signed-digit representation (SDR) to minimize the number of non-zero terms. Also, term quantization [11] sets low-order power-of-two terms to 0, indicated by red slashes, to satisfy a group term budget. The Figure illustrates a case when 50% of input terms are zeros. In forming the dot product of the input and an RRAM column of weight terms, products on the column to be output for accumulation have 50% or higher sparsity, given that some weight terms may be zeros. The direct HESE SAR ADC introduces extra sparsity other than TQ and reduces the energy of computation.

encoding for signed expressions (HESE) and term quantization (TQ) to the outputs of each layer to further reduce the non-zero terms and increase sparsity (Figure 1). The HESE signed digit representation (SDR) is directly generated during the analog-to-digital conversion. The HESE SDR has both positive and negative terms to reduce the non-zero terms. The TQ prunes out small terms in a group basis.

In this work, we propose the first bit-level sparsity-aware successive approximation register (SAR) ADC which directly produces HESE. The 12-bit resolution can support large ANNs with good accuracy. The proposed HESE ADC has two thresholds for 2-bits look-ahead (LA) and noise averaging (NA) is performed in the last couple of cycles. Section 2 provides an overview of the ANN's system architecture, HESE SDR, term quantization (TQ), and SAR ADC. In Section 3, we describe the schemes and circuit implementation details. Section 4 provides the measurement results. The proposed HESE SAR achieves a FoM of 15.2 fJ/conv.-step at 45MS/s. The core area of the SAR ADC is 0.072mm^2.

The main contributions of the paper are:

- The first direct HESE SAR ADC to provide sparse encoding during the analog-to-digital conversion with 2-bit look-ahead and the noise averaging at the last couple of cycles.
- The ADC achieves a FoM of 15.2fJ/c.-s at 45MS/s in 65nm.
- The use of direct HESE SAR ADC along with term quantization (TQ) to increase the sparsity in analog neural networks (ANNs).

2 BACKGROUND

2.1 Analog Neural Networks for CNNs

Analog Neural Networks (ANNs) typically use RRAM crossbars to both store CNN weights and perform matrix multiplication in-memory in an analog fashion [14][5].

An example crossbar of RRAM with 1-bit RRAM cells and 1-bit input values for computing dot products is shown in Figure 1. Both data and weights are bit-sliced, with each weight term occupying a separate RRAM column. The proposed SAR ADC directly provides the hybrid encoding for signed expressions (HESE) signed-digit representation (SDR) to minimize the number of non-zero terms. Also, term quantization [11] sets low-order power-of-two terms to 0 with red slashes to satisfy a group budget. The direct HESE SAR ADC introduces extra sparsity other than TQ and reduces the energy of computation.

2.2 Hybrid encoding for signed expressions

As shown in Figure 2a, HESE SDR [11][10] is an efficient one-pass encoding scheme to produce minimum-length signed expression representations. Along with the term quantization (TQ) in [13], HESE SDR can further reduce the non-zero terms. This L2R HESE SDR with 2-bit LA is co-designed between the encoding algorithm and the circuit design, which can enable a direct HESE SAR ADC design. The rules for HESE SDR are shown in Figure 2a. Figure 2b shows the illustration of the one-pass HESE SDR encoding. Figure 2c shows the weight distribution of a pre-trained AlexNet and terms distribution. HESE SDR can greatly reduce the non-zero terms and thus increase sparsity.

2.3 Term Quantization

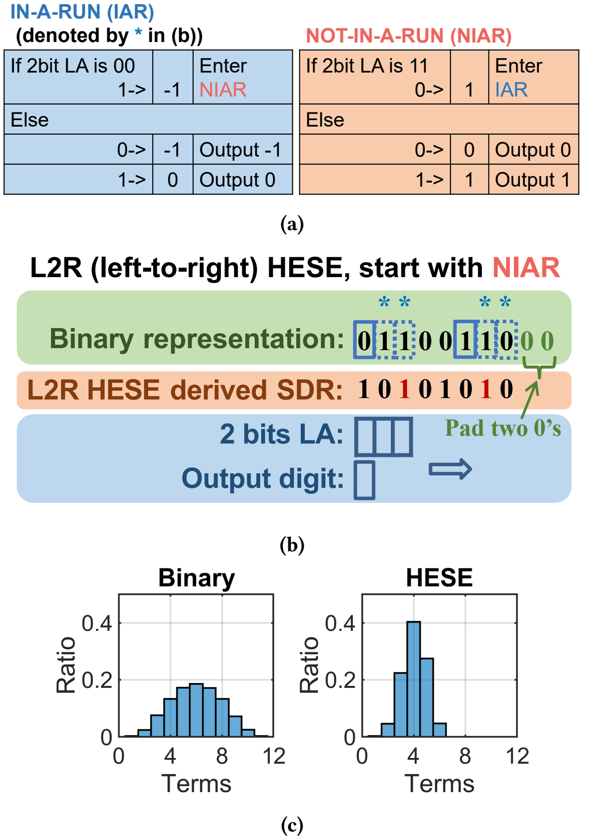
Term quantization (TQ) [11] prunes out small terms in a group of data, as shown in Figure 3. TQ can increase the bit-level sparsity with trivial impact on the classification accuracy. For uniform quantization, all values are truncated uniformly. TQ truncates the data in a group of data, which can keep more information

2.4 Conventional SAR ADC

Due to a large number of ADCs required for an ANN chip, the area and power consumption are two of the most important metrics of the ADCs for ANNs. SAR ADCs are power and area efficient, and various techniques have been reported to increase the sampling rate to 100MS/s or beyond, including pipelined SAR ADCs [9], time-interleaved SAR ADCs [16], and loop unrolled SAR ADCs [7].

The SAR ADC typically consists of the sample-and-hold circuitry, a comparator, a feedback DAC and the SAR logic. In the common capacitor DAC implementation, the sample-and-hold circuitry can re-use the feedback DAC.

There are 3 typical phases for a conventional SAR. The first one is the sampling phase. For bottom-plate sampling scheme, the bottom-plate of the CDAC is connected to the input and the top-plate of



IN-A-RUN (IAR) (denoted by * in (b))

If 2bit LA is 00 1->	-1	Enter NIAR
Else		
0->	-1	Output -1
1->	0	Output 0

NOT-IN-A-RUN (NIAR)

If 2bit LA is 11 0->	1	Enter IAR
Else		
0->	0	Output 0
1->	1	Output 1

(a)

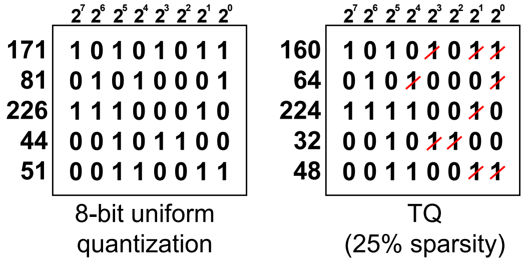
L2R (left-to-right) HESE, start with NIAR
* *
* *
Binary representation: 0 1 1 0 0 1 1 0 0 0
L2R HESE derived SDR: 1 0 1 0 1 0 1 0
Pad two 0's
2 bits LA:
Output digit:

(b)

Binary
HESE
Ratio
0.4
0.2
0
0
4
8
12
Terms

(c)

Figure 2: Hybrid encoding for signed expressions (HESE) SDR is shown in (a). The encoding looks at the current bit and next 2 bits to decide the encoded term. LA stands for look-ahead. Left to right (L2R) HESE SDR finds a minimum-length SDR and red 1 stands for -1, as illustrated in (b). Over 95% of the weights in a pre-trained AlexNet [8] can be represented by only half of the HESE SDR terms due to bit-level sparsity, as shown in (c).



	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
171	1	0	1	0	1	0	1	1
81	0	1	0	1	0	0	0	1
226	1	1	1	0	0	0	1	0
44	0	0	1	0	1	1	0	0
51	0	0	1	1	0	0	1	1

8-bit uniform quantization

	2^7	2^6	2^5	2^4	2^3	2^2	2^1	2^0
160	1	0	1	0	~~1~~	0	~~1~~	~~1~~
64	0	1	0	~~1~~	0	0	0	~~1~~
224	1	1	1	1	0	0	~~1~~	0
32	0	0	1	0	~~1~~	~~1~~	0	0
48	0	0	1	1	0	0	~~1~~	~~1~~

TQ (25% sparsity)

Figure 3: An illustration of term quantization (TQ),which keeps the largest non-zero 10 terms across a group of 5 data.

the CDAC is connected to the virtual ground. After that, the SAR enters the conversion phase in which the ADC conducts the binary search algorithm. The test voltage is always put at the center of the uncertainty range. The SAR logic block connects the corresponding

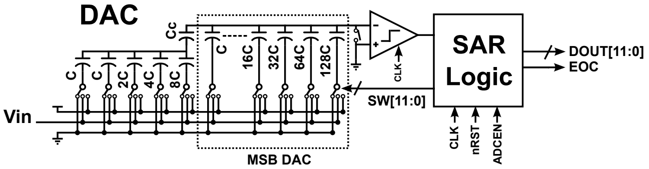


Figure 4: A single-ended 12 bits SAR ADC with 8-4 segmented capacitor array

capacitor to supply voltage at each bit cycle and connects it back to ground if the comparator output is zero. If the comparator output is one, the SAR logic keeps the connection. Similar conversion is done until the LSB comparison. After the conversion phase, the SAR ADC purge all the capacitors by connecting their two plates together.

3 PROPOSED SAR ADC

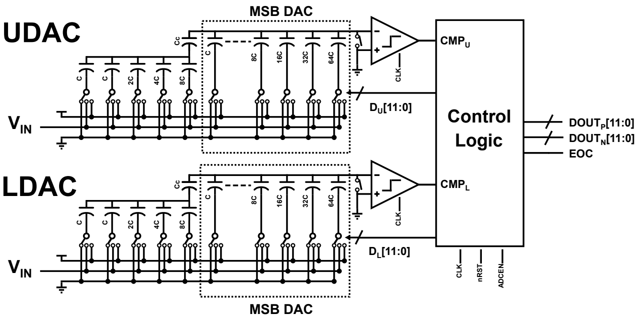


Figure 5: Global architecture of the proposed SAR is shown. Two DACs and comparators are implemented for the 2-bit look-ahead (LA) of hybrid encoding for signed expressions (HESE) SDR. Noise averaging (NA) is used to reduce the capacitor size. The accumulated current is converted to the voltage by sample and hold circuitry.

The global architecture of the proposed SAR ADC is shown in Figure 5. The SAR ADC is split into two half DACs and two half-sized comparators to provide two thresholds. The SAR ADC has two thresholds for each bit-cycling to perform the 2-bit look-ahead (LA). Noise averaging between the two halves is performed in the last couple of cycles to eliminate noise penalty due to half size DACs and comparators. Bottom-plate sampling is used and the sampling switches are bootstrapped to enhance the linearity. The comparators are fully dynamic with no static power. Two foreground calibration schemes are implemented to improve the linearity. One is the bridge capacitor calibration [3]. The other is the 4 largest MSB capacitors calibration [6].

3.1 Conversion Plan of the HESE SAR

Figure 6 shows the flow chart of the conversion plan. The ending states with padded zeros are slightly different and are not drawn for simplicity. N stands for current bit under test.

The HESE SAR switches between the IN-A-RUN (IAR) and the NOT-IN-A-RUN (NIAR) state when the input lies between two

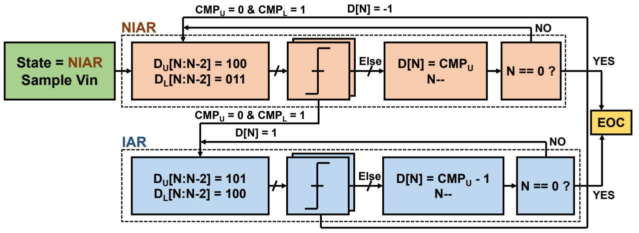


Figure 6: Flowchart of the conversion plan

thresholds. The extra threshold provides the analog 2-bit look-ahead (LA) for direct HESE. V_{IN} stands for the sampled analog input for each conversion. VDAC$_U$ stands for the analog output of upper DAC and VDAC$_L$ stands for the analog output of lower DAC.

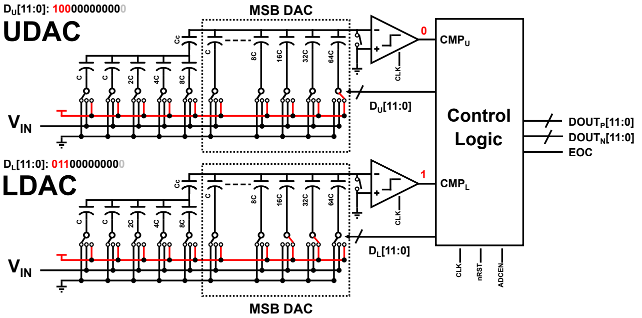


Figure 7: The HESE SAR starts in the NIAR state. D_U[N:N-2] and D_L[N:N-2] are set to 100 and 011, respectively. The SAR can look for 2bit LA of two consecutive 1's with this configuration. When CMP_U is 0 and CMP_L is 1, D[N] is encoded to 1 and the SAR switches to the IAR state.

The SAR starts with the NIAR state. As shown in Figure 7, D_U and D_L are the digital inputs to the upper and lower DAC, respectively. In the NIAR state, D_U[N:N-2] and D_L[N:N-2] are set to 100 and 011, respectively. The SAR can look for 2bit LA of two consecutive 1's with this configuration. When CMP_U is 0 and CMP_L is 1, D[N] is encoded to 1 and the SAR switches to the IAR state. Otherwise, D[N] = CMP_U.

In the IAR state, D_U[N:N-2] and D_L[N:N-2] are set to 101 and 100, respectively. The SAR can look for 2bit LA of two consecutive 0's. When CMP_U is 0 and CMP_L is 1, D[N] is encoded to -1 and the SAR switches to the NIAR state. Otherwise, D[N] = CMP_U - 1.

When N = 1 or 0, the SAR enters the ending states. The ending states are slightly different to handle the padded zeros. If N = 1 or 0 and the SAR is in the NIAR state, the LA is not necessary and 2bit LA cannot be two consecutive 1's because only zeros are padded. If N = 1 and the SAR is in the IAR state, the current bit and next bit of D_U and D_L are set to 11 and 10, respectively, to look for 2bit LA of two consecutive 0's. If N = 0 and the SAR is in the IAR state, the encoded output is -1 regardless of the outputs of the comparators due to the padded two zeros. The NIAR and the IAR switches when CMP_U is 0 and CMP_L is 1. When LA is not necessary, the LDAC and UDAC are connected in parallel to perform noise averaging.

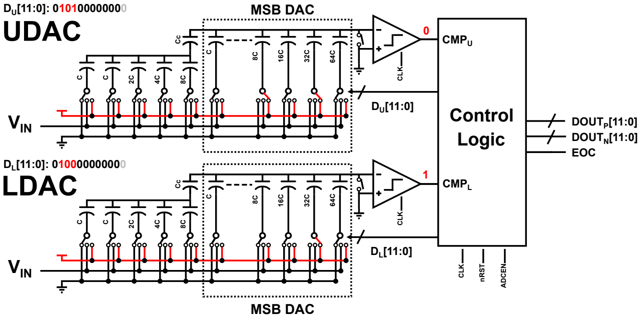
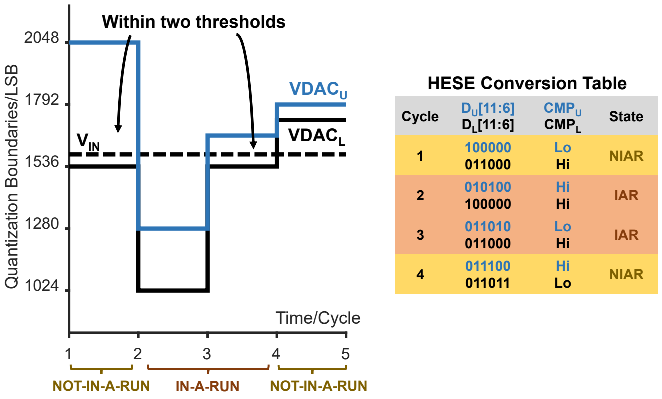


Figure 8: In the IAR state, D_U[N:N-2] and D_L[N:N-2] are set to 101 and 100, respectively. The SAR can look for 2bit LA of two consecutive 0's. When CMP_U is 0 and CMP_L is 1, D[N] is encoded to -1 and the SAR switches to the NIAR state.



Cycle	D_U[11:6] D_L[11:6]	CMP_U CMP_L	State
1	100000 011000	Lo Hi	NIAR
2	010100 100000	Hi Hi	IAR
3	011010 011000	Lo Hi	IAR
4	011100 011011	Hi Lo	NIAR

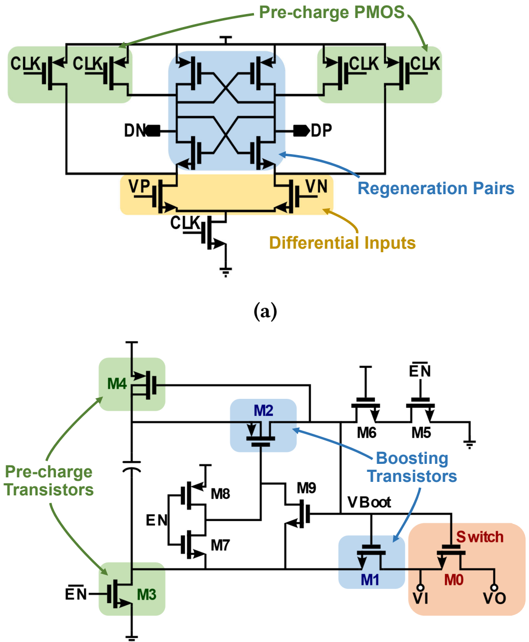
Figure 9: An example conversion waveform of the HESE SAR. VDAC$_U$ stands for the output of upper DAC. V_{IN} stands for the output of lower DAC. CMP_U is the output of the upper comparator. CMP_L is the output of the lower comparator. Comparator output is 1 when $V_{IN} > V_{DAC}$.

Figure 9 shows an example conversion waveform of the HESE SAR. Only 6 MSBs are shown for simplicity. The conversion starts with the NIAR state. In the first cycle, D_U[11:9] is set to 100 and D_L[11:9] is set to 011. If V_{IN} is larger than VDAC$_L$ and smaller than VDAC$_H$, the 2bit LA is 11. The HESE SAR enters the IAR state which would provides negative ones to increase sparsity. In the second cycle, D_U[10:8] is set to 101 and D_L[10:8] is set to 100. V_{IN} is larger than VDAC$_U$ and the HESE SAR stays in the IAR state. In the third cycle, V_{IN} is within two thresholds and the HESE SAR enters the NIAR state.

To reduce the sampled kT/C noise and the comparator noise, both the UDAC and the LDAC are connected in parallel except for the dummy LSB capacitors in the last couple of bit-cycles where LA is not necessary. Since both the UDAC and LDAC have 11-bit resolution, one more bit decision is required to provide an 12-bit result. The LSB capacitors are separately actuated to provide 1 extra bit decision. Compared with conventional SAR ADCs, no power/area/noise penalty is incurred in the proposed direct HESE SAR ADC.

3.2 Comparator and Bootstrapped Switches

Figure 10a shows the dynamic comparator of the SAR ADC. The outputs are connected to two inverters and then to a RS-latch for the digital block. The dynamic architecture only compares the inputs when it's fired. The dynamic comparator only needs one clock for pre-charge and firing. When the CLK is low, the nodes for regenerative circuits are pre-charged to supply voltage and the input pairs are disabled. When the CLK is high, the input pairs are enabled. The voltage difference in the input pairs causes the NMOS of the regenerative circuits to conduct different amounts of current. The regenerative circuits latches the output as the comparison result. There is no static power except for the leakage.



(a)

(b)

Figure 10: Implementation of circuit blocks. The schematic of the dynamic comparator is shown in (a). The bootstrapped switch is shown in (b). DN and DP in (a) are connected to two inverters, respectively. The outputs of the inverters are connected to an RS latch to provide comparator output for the control logic.

Figure 10b shows the bootstrapped switch of the SAR ADC. To avoid the sampling network nonlinearity of the bottom-plate sampling switch, the proposed SAR ADC bootstraps the V_{GS} of the sampling switches by using the charge-pump circuit. A nearly-constant V_{GS} is provided to M0 if the charge-pump capacitor is significantly larger than the parasitic capacitance at the source of M2. M6 and M8 are needed to improve the circuit reliability.

For the bootstrapped switch, M9 and M6 are extra transistors to protect the M2 and M5, respectively. Since the top plate of the capacitor can be larger than the supply voltage, the bulk of M2 and M4 are connected to the top plate of the capacitor.

3.3 DACs

Figure 5 shows the DAC configuration of the proposed SAR ADC. Each DAC is an 11-bit array for 12-bit combined resolution employs the common centroid layout technique to reduce the linear gradient effects. The DAC also uses the equal-edge-ratio layout technique [18] to further suppress the gradient effects. Dummy metal is filled manually in the DAC area for good matching. Bridge capacitor is sized slightly larger than the ideal value for calibration [3].

4 RESULTS

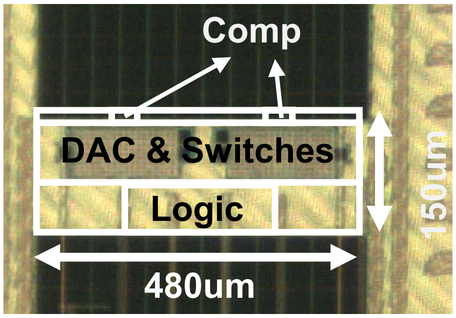


Figure 11: Chip micrograph

A prototype chip is fabricated in low power 65nm technology. The chip micrograph is shown in Figure 11. The prototype demonstrates direct sparse encoding along with the analog-to-digital conversion.

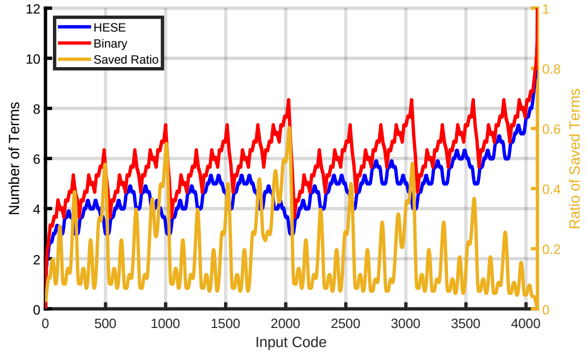


Figure 12: Simulated number of terms of the HESE and the binary encoding for all the 12-bit digital codes. On average, the HESE saves 23% of the terms compared to the binary encoding. Note that the terms refer to the non-zero digits. For HESE, the terms include both positive and negative ones. For binary, the terms include only positive ones since there is no negative ones in the binary encoding.

Figure 12 shows that the HESE SDR minimizes the non-zero terms comparing to the binary encoding. The HESE SDR can save up to 60% of the terms compared to binary encoding.

Figure 13 shows the measured spectrum of the direct HESE SAR. The effective number of bit (ENOB) is 10.6b. Signal bins are

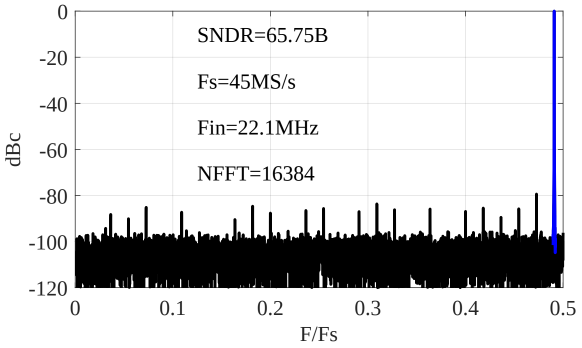
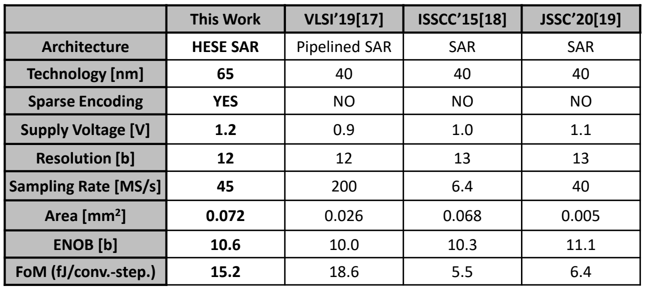


Figure 13: Measured spectrum of the HESE SAR

highlighted. The sampling rate is 45MS/s. 16384 data points are used for FFT calculation. The input frequency is a 22.1MHz sine wave. The positive and negatives outputs are readout separately and reconstructed to binary representations for FFT calculation.



	This Work	VLSI'19[17]	ISSCC'15[18]	JSSC'20[19]
Architecture	**HESE SAR**	Pipelined SAR	SAR	SAR
Technology [nm]	**65**	40	40	40
Sparse Encoding	**YES**	NO	NO	NO
Supply Voltage [V]	**1.2**	0.9	1.0	1.1
Resolution [b]	**12**	12	13	13
Sampling Rate [MS/s]	**45**	200	6.4	40
Area [mm^2]	**0.072**	0.026	0.068	0.005
ENOB [b]	**10.6**	10.0	10.3	11.1
FoM (fJ/conv.-step.)	**15.2**	18.6	5.5	6.4

Figure 14: Summary comparison table

Figure 14 shows the comparison between this work and prior works. The HESE SAR is the first sparsity-aware SAR ADC to demonstrate direct sparsity encoding with competitive energy efficiency, resolution and area. The references only provide binary encodings.

5 CONCLUSION

This work is the first bit-level sparsity-aware ADC in ANNs with direct hybrid encoding for signed expressions (HESE) leveraging algorithm-circuit co-design. ANN with HESE SAR minimizes the non-zero terms and enables a reduction in energy along with the term quantization (TQ). A prototype in 65nm low power technology and achieves Walden FoM of 15.2fJ/conv.-step at 45MS/s. The direct HESE SAR offers a general direction for the ADC design in ANNs leveraging bit-level sparsity. The core area is 0.072mm^2.

ACKNOWLEDGMENTS

This research was supported by DARPA and Navy-ONR under contract N00014-20-1-4005 and MIT Center for Integrated Circuits and Systems.

REFERENCES

[1] Mohamed R. Abdelhamid, Ruicong Chen, Joonhyuk Cho, Anantha P. Chandrakasan, and Fadel Adib. 2020. Self-Reconfigurable Micro-Implants for Cross-Tissue Wireless and Batteryless Connectivity. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking - (MobiCOM)*. Association for Computing Machinery, Article 59, 14 pages. https://doi.org/10.1145/3372224.3419216

[2] J. K. Brown, D. Abdallah, J. Boley, N. Collins, K. Craig, G. Glennon, K. Huang, C. J. Lukas, W. Moore, R. K. Sawyer, Y. Shakhsheer, F. B. Yahya, A. Wang, N. E. Roberts, D. D. Wentzloff, and B. H. Calhoun. 2020. 27.1 A 65nm Energy-Harvesting ULP SoC with 256kB Cortex-M0 Enabling an 89.1µW Continuous Machine Health Monitoring Wireless Self-Powered System. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*. 420–422. https://doi.org/10.1109/ISSCC19947.2020.9063067

[3] Yanfei Chen, Xiaolei Zhu, Hirotaka Tamura, Masaya Kibune, Yasumoto Tomita, Takayuki Hamada, Masato Yoshioka, Kiyoshi Ishikawa, Takeshi Takayama, Junji Ogawa, et al. 2009. Split capacitor DAC mismatch calibration in successive approximation ADC. In *2009 IEEE Custom Integrated Circuits Conference*. IEEE, 279–282.

[4] Yu-Hsin Chen, Tushar Krishna, Joel S Emer, and Vivienne Sze. 2016. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE journal of solid-state circuits* 52, 1 (2016), 127–138.

[5] Teyuh Chou, Wei Tang, Jacob Botimer, and Zhengya Zhang. 2019. Cascade: Connecting rrams to extend analog dataflow in an end-to-end in-memory processing paradigm. In *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*. 114–125.

[6] Ming Ding, Pieter Harpe, Yao-Hong Liu, Benjamin Busze, Kathleen Philips, and Harmke de Groot. 2015. 26.2 A 5.5 fJ/conv-step 6.4 MS/S 13b SAR ADC utilizing a redundancy-facilitated background error-detection-and-correction scheme. In *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*. IEEE, 1–3.

[7] Tao Jiang, Wing Liu, Freeman Y Zhong, Charlie Zhong, and Patrick Y Chiang. 2010. Single-channel, 1.25-GS/s, 6-bit, loop-unrolled asynchronous SAR-ADC in 40nm-CMOS. In *IEEE Custom Integrated Circuits Conference 2010*. IEEE, 1–4.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[9] Lukas Kull, Danny Luu, Christian Menolfi, Matthias Braendli, Pier Andrea Francese, Thomas Morf, Marcel Kossel, Hazar Yueksel, Alessandro Cevrero, Ilter Ozkaya, et al. 2017. 28.5 A 10b 1.5 GS/s pipelined-SAR ADC with background second-stage common-mode regulation and offset calibration in 14nm CMOS FinFET. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 474–475.

[10] HT Kung. 2021. High-order-bit First Conversion for Signed-Digit Representations. In *Annual GOMACTech Conference*. IEEE. http://www.eecs.harvard.edu/~htk/publication/2021-gomactech-kung.pdf

[11] HT Kung, Bradley McDanel, and Sai Qian Zhang. 2020. Term quantization: furthering quantization at run time. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–14.

[12] K. L. Loh. 2020. 1.2 Fertilizing AIoT from Roots to Leaves. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*. 15–21. https://doi.org/10.1109/ISSCC19947.2020.9062950

[13] Bradley McDanel, HT Kung, and Sai Qian Zhang. 2021. Saturation RRAM Leveraging Bit-level Sparsity Resulting from Term Quantization. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5.

[14] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R Stanley Williams, and Vivek Srikumar. 2016. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 14–26.

[15] W. Shan, M. Yang, J. Xu, Y. Lu, S. Zhang, T. Wang, J. Yang, L. Shi, and M. Seok. 2020. 14.1 A 510nW 0.41V Low-Memory Low-Computation Keyword-Spotting Chip Using Serial FFT-Based MFCC and Binarized Depthwise Separable Convolutional Neural Network in 28nm CMOS. In *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*. 230–232. https://doi.org/10.1109/ISSCC19947.2020.9063000

[16] Jeonggoo Song, Kareem Ragab, Xiyuan Tang, and Nan Sun. 2016. A 10-b 800MS/s time-interleaved SAR ADC with fast timing-skew calibration. In *2016 IEEE Asian Solid-State Circuits Conference (A-SSCC)*. IEEE, 73–76.

[17] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* 105, 12 (2017), 2295–2329.

[18] Naveen Verma and Anantha P Chandrakasan. 2007. An ultra low energy 12-bit rate-resolution scalable SAR ADC for wireless sensor nodes. *IEEE Journal of Solid-State Circuits* 42, 6 (2007), 1196–1205.

[19] Peng Yao, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, J Joshua Yang, and He Qian. 2020. Fully hardware-implemented memristor convolutional neural network. *Nature* 577, 7792 (2020), 641–646.