SCIENCE CHINA
Information Sciences

October 2023, Vol. 66 200409:1–200409:16
https://doi.org/10.1007/s11432-023-3760-8

• PROGRESS •
Special Topic: Recent Progress of Fundamental Research on Post-Moore Novel Devices

Breaking the energy-efficiency barriers for smart sensing applications with "Sensing with Computing" architectures

Xinghua YANG[1,5*], Zheyu LIU[1,2], Kechao TANG[3], Xunzhao YIN[4], Cheng ZHUO[4], Qi WEI[1] & Fei QIAO[1*]

[1]*Department of Electronic Engineering, Tsinghua University, Beijing* 100084*, China;*
[2]*MakeSens AI Technology (Beijing) Co., Ltd., Beijing* 100084*, China;*
[3]*School of Integrated Circuits, Peking University, Beijing* 100871*, China;*
[4]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou* 310029*, China;*
[5]*College of Science, Beijing Forestry University, Beijing* 100083*, China*

Received 16 February 2023/Revised 16 April 2023/Accepted 5 May 2023/Published online 6 September 2023

Abstract With the developing technologies of artificial intelligence and the Internet of Things, intelligent IoT (iIoT) is prevailing currently. Design and implementation of integrated IoT nodes with continuous perception capability are indispensable to realize various smart terminal devices, which would also be vital to reduce the power consumption, improve the real-time performance, and enhance the security/privacy of the IoT system. In this paper, we present the architecture of "Sensing with Computing" and its chip design for smart sensing applications, which would support multi-modal perception signal processing with multi-dimension extension ability. Specially, we explore the analog/mixed-signal circuit designs and algorithm-hardware co-design methodologies for perception signal processing, and we also study the multi-modal integration of novel sensors and their interface technologies. Additionally, some multi-modal smart sensing systems with "Sensing + Computing in Memory" mixed-signal chips would be fabricated, which would support typical always-on smart sensing tasks.

Keywords low power consumption circuit design, sensing with computing, smart sensors, edge computing, multi-modal sensing

Citation Yang X H, Liu Z Y, Tang K C, et al. Breaking the energy-efficiency barriers for smart sensing applications with "Sensing with Computing" architectures. Sci China Inf Sci, 2023, 66(10): 200409, https://doi.org/10.1007/s11432-023-3760-8

1 Introduction

Recently, with the fast development of artificial intelligence and big data processing, there has been a practical breakthrough in IoT technologies. More and more IoT devices and sensors are deployed "ANYWHERE" and working "ANYTIME", which is profoundly changing the information society. As well, the amount of data produced by IoT devices is also expanding rapidly, and the conventional framework of cloud computing suffers larger energy consumption, longer response time, and even security issues. In view of such a huge amount of data, how to achieve efficient sensor data collection and intelligent processing is a new challenge for IoT technology to further move towards intelligence. With the large-scale deployment of IoT devices, the problem of data bloat is becoming increasingly serious, and the trend of intelligent processing power sinking from the cloud to the terminal is becoming more and more obvious. In view of the natural characteristics of limited resources, limited energy, and limited bandwidth of terminal sensing devices, people have put forward more stringent requirements for the intelligence, security, and continuous working time of sensing terminal devices.

However, to realize this kind of "Sensing with Computing" systems, it would face challenges of sensors/circuits, computing architectures, and design methodologies while deploying intelligent algorithms

* Corresponding author (email: yangxh@bjfu.edu.cn, qiaofei@tsinghua.edu.cn)

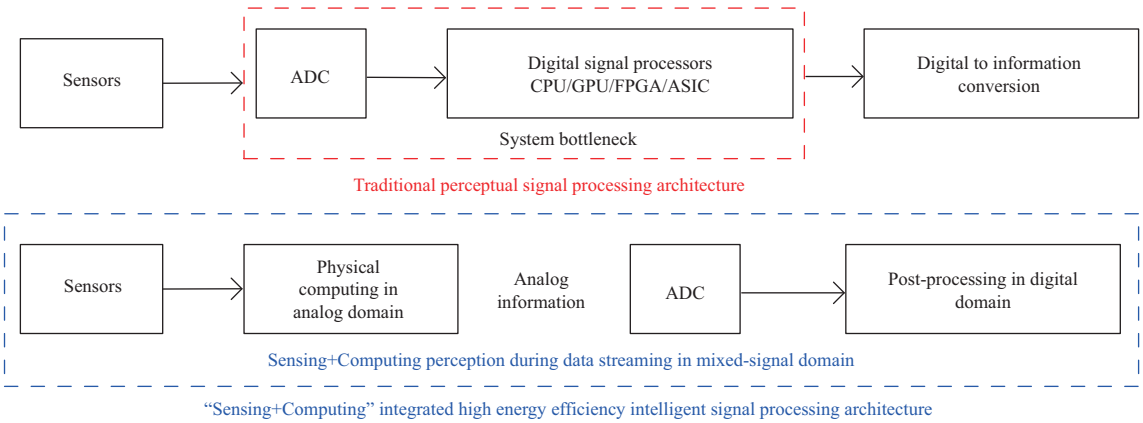

Figure 1 (Color online) Overview of "Sensing with Computing" versus conventional smart sensing configuration.

on the smart sensing terminals. For smart sensing systems, these challenges would include in detail: (1) How to realize multi-modal sensing data collection and expression? Integrated and minimized multi-modal sensors with lower energy consumption would be studied, which support unified data sensing and expression of multiple physical quantities. Then, the sensors would output with interface circuits for information extraction successively. And, (2) how to improve the information extraction efficiency with limited computing and energy resources? Hereon, the boundaries of computing capabilities, as well as energy efficiency and task accuracy, would be carefully explored for the near-sensor computing architectures. Additionally, (3) how to assure the accuracy of information processing in the mixed-signal domain, for the processed sensing data are always in analog and mixed-signal formats? Conventional artificial intelligence algorithms in the digital domain would find their way to be rewritten in the analog/mixed-signal domain, with less accuracy lost, which would also consider the co-optimization of energy, area, latency, and non-ideal hardware characteristics.

This paper summarizes the emerging "Sensing with Computing" architectures and chip-level designs for smart sensing applications, as shown in Figure 1. Section 2 would review the research background of energy-efficient smart sensing systems, with respect to sensors, computing parts, and design methodologies. Next, key issues of the "Sensing with Computing" methods will be explored at both circuit and architecture levels, as well as issues during deployment and optimization, in Section 3. Additionally, in Section 4, some technology progress and methods will be provided to prove the feasibility of the new architectures. Section 5 will show some future directions and challenges. Section 6 concludes the paper with some basic design principles.

2 Background

As the IoT continues to expand and extend in areas such as intelligent manufacturing, autonomous driving, smart homes, and smart cities, the amount of data produced by terminal sensing devices is rapidly expanding. According to a study predicted by IDC, an international data company [1, 2], by 2025, the number of global IoT devices will reach 41.6 billion, and the amount of data produced will be as high as 78.4 ZB. For such a huge amount of data, how to achieve efficient sensing data collection and intelligent processing is a new challenge for IoT technology to further move towards intelligence.

From the current mainstream technology system, the IoT architecture is mainly divided into three layers [3]: the perception node layer, network transport layer, and application layer. Among them, the perception node layer is the source of massive data in the whole IoT system and the core element that promotes the rapid development of the IoT industry. People perceive the physical world in real time through sensor technology and terminal pre-processing equipment within the perception layer, then interconnect and transmit through the network layer, further to the application layer for information processing and knowledge mining, and finally realize accurate cognition, scientific decision-making, and real-time control of the physical world. In recent years, the rapid development of artificial intelligence and machine learning technology has largely driven the progress of sensing technology. However, with the massive deployment of IoT devices, the problem of data inflation is becoming increasingly severe and the trend of intelligent processing power sinking from the cloud to the terminal is becoming more and more

evident [4]. Given the natural characteristics of resource-constrained, energy-constrained, and bandwidth-constrained terminal sensing devices, there are more stringent requirements for the intelligence, security, and continuous working duration of sensing terminal devices. Therefore, continuous sensing systems and related chip design technologies with intelligent processing capability and long working time [5] have become a hot topic in the field of IoT research.

The initial concept of "Sensing with Computing" can be traced back to the 1990s. In 1994, Forchheimer et al. [6] at Linköping University proposed a new paradigm of "Near-Sensor Computing" for image processing. Such chip design will also face various new demands and challenges under the conditions like new sensor forms, diverse intelligent sensing tasks with high arithmetic demands, and advanced IC manufacturing processes. The detailed demands and challenges are listed as follows. (1) Efficient sensing data acquisition. In order to meet the increasingly diverse application needs, a new generation of intelligent sensors is supposed to be equipped with multi-modal information sensing capabilities, for example, similar to the sensing characteristics of human skin, a single receptor that can simultaneously detect a variety of external signals. In addition, due to the expanding sensing edge and the relatively limited space and energy supply at the edge end, the development of miniaturized, low-energy sensing devices will become a key point for large-scale technology applications. (2) Data conversion costs and interface costs. Although the "Sensing with Computing" technology can reduce the amount of data transmission from the sensor to the subsequent processing equipment, the traditional digital signal processing-based solutions still require the sensor to collect a large amount of raw information through the analog-to-digital converter (ADC) to the digital domain, and then post-processing. Therefore, this unnecessary data conversion imposes significant hardware overhead and limits energy efficiency. (3) The problem of access bottleneck. Although the "Sensing with Computing" computing system can reduce the need to access external storage, the computing process still needs to access a large amount of weight data and intermediate results, which means there is an access bottleneck. (4) The limitation of non-ideal circuit factors. Various non-ideal factors such as circuit noise, linearity, and process deviations in mixed-signal computing units lead to limited computational accuracy, and noise and computational errors accumulate along the signal propagation path, which may eventually exert a significant impact on the output quality. As such, these issues and challenges pose difficulties in further improving the energy efficiency and complexity of sensing processing on existing IoT end devices, especially for many limited battery-powered IoT smart nodes and emerging wearable smart sensing devices. As can be seen, the above system-on-a-chip design challenges are not only reflected at the circuit design level, but also in various aspects such as sensor devices, signal processing architecture, application algorithm design, and application requirement definition. With the gradual failure of Moore's law and Dennard Scaling [7, 8], the improvement of energy efficiency of smart sensing ICs is increasingly dependent on cross-level joint design approaches [9].

The "Sensing with Computing" technology can be promoted only after the research of efficient "sensing" technologies, among which new sensor devices and array designs are actively researched. In terms of energy-efficient miniaturized multi-modal sensor design, a new idea is to develop integrated multifunctional sensors that allow a single sensor to detect multiple modes of signals simultaneously, as opposed to the traditional solution of simply assembling a variety of discrete sensor devices. In this way, the complexity, space cost, and energy consumption of sensor devices can be further reduced while realizing multi-modal sensing and fusion technology. In recent years, several reports on integrated multi-modal sensors for applications in soft robotics and wearable devices have been published. Examples are given as follows: Din and his collaborators [10] developed a capacitive sensing-based robot skin that can simultaneously detect shear stress, tensile stress, and normal pressure on a surface. Zhang's research group [11] at Tsinghua University invented an electrical graffiti that can be printed on the surface of the skin to simultaneously detect changes in stress, temperature, and humidity on the body surface. However, research in this area generally is in a preliminary stage, largely on the preparation and characterization of multi-modal sensor devices, while less on arraying, and no complete edge intelligent sensing and computing system has yet been realized.

For computing circuit design, "Sensing with Computing" mainly involves energy-efficient sensing-computing interface circuit design, analog domain feature extraction, and mixed-signal neural network computational circuit design. Circuit design for sensing-computing interfaces typically simplifies or replaces conventional analog-to-digital converter (ADC) at the pixel or pixel array level, thereby reducing the cost of the interface and improving processing efficiency. In 2017, Ceze's research group [12] at the University of Washington proposed an analog-random sequence conversion interface for image recognition directly at the sensor side. In terms of feature extraction circuits, direct feature extraction in the analog

domain is a possible solution to the "Sensing with Computing" technology. In 2016, researchers [13] at KU Leuven proposed a circuit design for directly extracting speech features in the analog domain for voice activity detection (VAD). In the field of visual sensing, a gradient feature extraction circuit with configurable accuracy was proposed by a research group at Stanford University in 2019 [14]. Mixed-signal computing circuits are widely applied in near-sensing neural network processors. In 2019, Stanford and KU Leuven [15] collaborated on a normally opened binary neural network terminal processor using switched-capacitor neurons to achieve highly parallel multiply-accumulate operations in the analog signal domain. In general, analog and mixed-signal circuits are a promising solution for improving the processing efficiency of "Sensing with Computing" circuits because of their better compromise between operational accuracy and circuit power consumption, plus their ability to directly process various analog signals output by sensors. However, existing solutions come with complex circuits and poor scalability, as well as the non-ideal factors of the analog circuit that may lead to a decreased output quality. Therefore, an analog circuit technique with good scalability is urgently needed. Meanwhile, a joint algorithm-circuit design and its optimization are required to counteract the complexity of circuits and the side effect of non-ideal factors on the output quality.

As for architecture, computation, and memory access models need to be considered at a more abstract level than circuit topology design. To shorten the distance between "sensor" and "processor", researchers have explored a new sensing processing architecture to reduce the cost of "Sensing with Computing" data conversion. In 2016, researchers [16] at Rice University proposed the Redeye architecture for mobile visual sensing, which designed an analog signal processing mode that was directly connected to the sensor, and realized the neural network operation by integrating the analog computing circuit at the sensor end, reducing the amount of sensor output data, thereby the interface cost down. In 2019, Tsinghua University [17] proposed a mixed-signal processing near-sensing architecture (processing near sensor architecture, PNSA) and chip design, which, by redesigning the sensor output interface, made an efficient and continuous-time signal processing chip a reality. The above architecture design for sensing computing now confronts challenges of how to optimize the design to offset non-ideal characteristics such as circuit noise, mismatch, and process deviations, and of how to overcome the bottleneck in storage optimization. Breakthroughs in these aspects usually require the design of complex compensating circuits and calibration systems. As for access costs of various sensing processing systems that support neural network tasks, the ShiDianNao architecture proposed by the Institute of Computing, Chinese Academy of Sciences in 2015 is a typical near-sensing computing architecture [18]. It eliminates external access costs by tightly integrating sensors and processing systems. Similarly, there were Eyeriss [19] and Eyerissv2 [20] architectures proposed by MIT researchers in 2016 and 2019, respectively. These "near-data" processing architectures, based on local data organization closer to the computing unit, do not inherently break the access bottleneck. As a promising solution, computing-in-memory or process-in-memory [21, 22] operates in parallel processing in memory to overcome the energy-efficient performance bottleneck of the new computational model in the von Neumann architecture. However, various current research on computing-in-memory architecture [23–29] mainly focuses on the architecture model of a single device, without a deep and joint optimization from perception to storage computing for intelligent perception scenarios. And the numerous data conversion interfaces in the existing computing-in-memory architecture turn out to be a bottleneck in system efficiency instead. To integrate different perception scenarios, circuit and device designs, and computing-in-memory processors for perceptual signal processing neural networks, there still calls for full-stack research from front-end sensor devices, underlying computing devices, and circuits to top-level architectures and algorithms.

3 Key issues of "Sensing with Computing" architectures

3.1 Circuit-level issues

As the circuit-level supporting technology of the "Sensing with Computing" architecture, this issue focuses on the design optimization technology of circuit units in the mixed-signal domain, multi-modal integrated sensor-arrays modeling optimization technology, and interface circuit technology that supports mixed-signal domains. The goal is to use "Sensing with Computing" to directly import the analog signal from the sensors into the memory-in-computing arrays, so as to break through the sensor data conversion and memory access energy efficiency bottlenecks of the sensory integrated chip at the same time.

3.1.1 *Circuits design of mixed-signal domain for "Sensing with Computing"*

Different computing modes and application scenario requirements should be considered to optimize the circuit design for mixed-signal memory-in-computing processing of analog/digital signals transmitted by sensor arrays. We can investigate memory-in-computing circuit units based on CMOS process SRAM/DRAM structure, and conduct comprehensive evaluation and optimization based on device and circuit charge-discharge characteristics. Additionally, it explores neuron array design optimization methods that support the accumulation and inter-layer transmission functions of scalable neural networks.

3.1.2 *Multi-modal integrated sensor devices and array modeling*

Firstly, it is necessary to break through the status quo of a single sensor, which senses a single type of information in the traditional sensor technology field. The design of multi-modal integrated sensor devices and arrays for the joint sensing task of multiple combinations of physical quantities, such as sound, light, pressure, and temperature, through a joint optimization method of materials and device structures, is required to meet the requirements of an edge sensing system for full integration, miniaturization, and low energy consumption. The second is for the structural characteristics and sensing data output of the new multi-modal sensors, there are problems of signal expression, matching interface schemes between sensing and subsequent processing structures, and the fusion processing of multi-modal sensing data in the spatial and temporal domains.

3.1.3 *Interface circuit design for "Sensing with Computing"*

Based on memory-in-computing circuit units, there is a need to solve the problem of multiple interfaces in "Sensing with Computing" architecture. According to the correlation of voltage and current between the output of the sensing array and the input of the memory-in-computing array, it is needed to explore custom design optimization of interface circuits to support digital and analog signal transmission. The other issue is the optimization of the input interface of the scalable memory-in-computing array for a variety of considerations of sensing signal indicators, such as scanning frequency, electrical signal strength, and duration.

3.2 Architecture-level issues

The goal of "Sensing with Computing" is to achieve detection, recognition and classification applications based on neural network algorithms driven directly by sensor output data. It supports processing tasks of different scales, different sensor data sources, and different sensing modalities. In order to solve the above problems, the "Sensing with Computing" architecture needs to have the following characteristics: hierarchical architecture with flexible configuration, multi-dimensional and scalable architecture organization, supporting mixed-signal memory-in-computing and multi-modal processing capability.

3.2.1 *Hierarchical integrated architecture*

This part focuses on the hierarchical and scalable integrated memory-in-computing architecture, including the design of the memory-in-computing array and its scalable signal processing unit architecture. Specifically, we will make research on the low-power storage array and its scalable array optimization method for matrix multiplication in neural networks. Meanwhile, scalable hybrid signal processing unit architecture and its data flow optimization method based on the interconnection between the storage array and neuronal circuits will also be deeply explored.

3.2.2 *Multi-modal "Sensing with Computing" architecture*

"Sensing with Computing" architecture will be deployed that supports the key functions of detection, recognition, and classification directly driven by multi-modal sensing data and oriented to neural network algorithms. The first step is to design and optimize the digital and analog signal interfaces and control modules for multi-mode sensing signals. The second is a hierarchical, scalable "Sensing with Computing" architecture design that supports the fusion of digital and analog signal processing and its algorithm-based requirements. Finally, it is the processing unit configuration and data path optimization strategy for the energy-efficient load metrics of the application scenario.

3.2.3 *Neural network algorithm mapping method for multidimensional scalable "Sensing with Computing" architecture*

Multi-dimensional on-chip mapping methods for neural network algorithms that target deep fusion of digital and analog signals to satisfy the hardware constraints of scalable architectures will be investigated. The first step is to propose the mapping method of a fully connected layer, convolutional layer, and storage-computation integrated architecture of neural network combining sensing array sensory field. The second is the "Sensing with Computing" architecture for the multi-modal sensing task algorithm model, the data organization optimization method, and its hardware configuration scaling strategy.

3.3 Deployment and optimization issues

In order to avoid the problem of errors and noise introduced by analog and mixed signal domain processing, it is important to propose algorithm-circuit hardware-software collaboration methods for the research of circuit error models. This part includes the design of a reliable and stable mixed-signal domain "Sensing with Computing" integrated system for verification and realization of multi-modal intelligent sensing chip for demonstration system. In fact, this technology has played an important role in many practical fields. In industrial automation, "Sensing with Computing" could be used in industrial automation to monitor and control various processes. For example, sensors can detect changes in temperature, pressure, and humidity, and automatically adjust the process to maintain optimal conditions. They can also detect faults in the machinery and notify maintenance personnel. In healthcare, "Sensing with Computing" could be used to monitor patients' vital signs and detect any abnormalities. For example, sensors can monitor a patient's heart rate, blood pressure, and oxygen levels, and alert healthcare professionals if there are any significant changes. In smart transportation, "Sensing with Computing" could be used in transportation to monitor and control traffic flow. It is noteworthy that in these applications, "Sensing with Computing" technology endows sensors with the characteristics of low power consumption, high performance, and intelligence. At the sensor end, we can accomplish most of the intelligent decision-making tasks, greatly reducing communication pressure.

3.3.1 *Algorithm-circuit co-optimization method for non-ideal factors of circuits*

According to the fault tolerance property of the neural network algorithm, the method addresses the influence of non-ideal factors on the circuit, a collaborative hardware-software optimization method for the reliability of computing and the quality of algorithm implementation.

3.3.2 *Optimization of neural network algorithm for "Sensing with Computing" chips*

It is necessary to enhance the robustness and consistency of various sensing tasks when deployed on chip, modeling, and optimization of sensing task algorithms (typical neural network algorithms) considering approximate calculation features in the mixed signal domain. First, the optimization methods of algorithm indicators such as accuracy, network capacity, astringency, training speed, and other parameters for actual sensing tasks, are combined with the approximate computational characteristics of mixed-signal circuits. Then, a behavioral-level algorithm model for quickly simulating and judging the quality of algorithm output and a simulation evaluation framework should be proposed.

3.3.3 *Multi-modal "Sensing with Computing" chip and demonstration system*

In order to verify the interconnection function between sensor devices and near-sensing computing-in-memory arrays, and to validate the functionality, performance, and scalability deployment methods of "Sensing with Computing" architecture, it is necessary to have actual chip implementation and testing, and conduct iterative optimization between sensing arrays and chip samples from small-scale tape out testing. To build a multi-modal "Sensing with Computing" demonstration system for continuous sensing application scenarios, the results of the actual test and evaluation will be applied to the "Sensing with Computing" architecture and design optimized feedback.

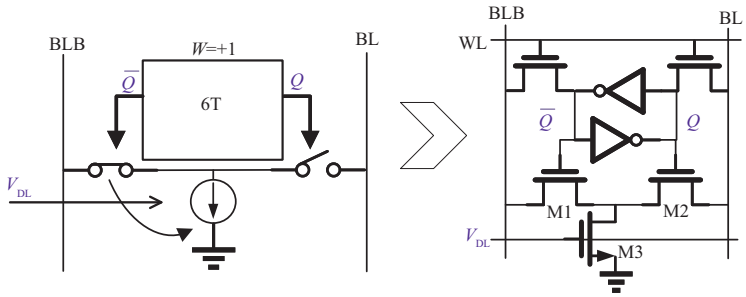

Figure 2 (Color online) Scheme of basic compute-in-memory.

4 Progress and methods

4.1 Near-sensor compute-in-memory circuits design

4.1.1 *Design and optimization of circuit units in the mixed-signal domain*

As representative research in this part, the circuit design of the basic compute-in-memory unit in the mixed-signal domain should be conducted. This unit should be able to store weight value and perform weighting operations on the input signal based on the stored value, thus achieving a fusion of storage and computation. The standard 6T SRAM cell can perform the basic bit storage. In [5], a separate pathway to the 6T SRAM cell is added as shown in Figure 2, where a voltage-controlled current source discharges the bit lines, and the bit data stored in the original SRAM is used to control the direction of current discharge.

At the same time, to accomplish the functions of accumulation and nonlinear operations in neural network algorithms, the design of a mixed-signal domain neuron circuit should be carried out. The basic design principle is to use an analog current integrator circuit to accumulate the partial sums of the outputs of each column of the compute-in-memory array. In the design process, non-ideal factors of the analog circuit, such as the channel length modulation effect, should be avoided as much as possible. Considering to support multiple input sensing data, the neuron circuits should be designed as analog and digital, respectively. The role of the analog neuron is to accumulate the output results of multiple columns in the time domain to obtain the computational results of the hidden layer neurons or output neurons in the neural network; the whole of the digital neurons is to binarize the feature output of the convolutional layer, or to compare the magnitude of multiple columns in a memory channel to obtain the output class in the form of a digital signal.

4.1.2 *Multi-modal integrated sensor devices*

Sensor devices for low energy consumption, high performance sensing, and multi-modal fusion information processing with physical parameters (thermal infrared signal, optical signal, flexible stress, temperature, ion concentration, etc.) should be proposed [30]. In fact, gradient tungsten-doped VO2 films to prepare a thermal infrared signal measurement array as a visual sensor has been proven effective. The incident infrared radiation causes a small temperature increase in the sensing material (WxV1-xO2 film), which in turn is converted into an electrical signal by measuring the change in resistance. Preliminary work shows that, in addition to the sensitivity to temperature, mechanical signals including stress, deformation, and ionic solution concentration can also affect the electrical properties of VO2 materials. Therefore, the VO2 film can achieve multi-modal sensing of temperature, stress, and ion signals. In addition, after the phase transition temperature of the material is lowered to near room temperature by gradient tungsten doping, the sensitivity of the sensor device is improved because the external stimulus signal is more likely to trigger the phase transition of the material [31].

4.1.3 *Interface circuit for near sensor computing-in-memory*

The function of the input interface circuit is to transmit sensing data in the form of analog vectors into a computing-in-memory array. The basic principle is to use the drive line DL in the computing-in-memory array, the interface circuit, and the mirror current tube in the computing-in-memory unit together to form a current mirror structure, which enables the transfer of analog vector signals. As described earlier,

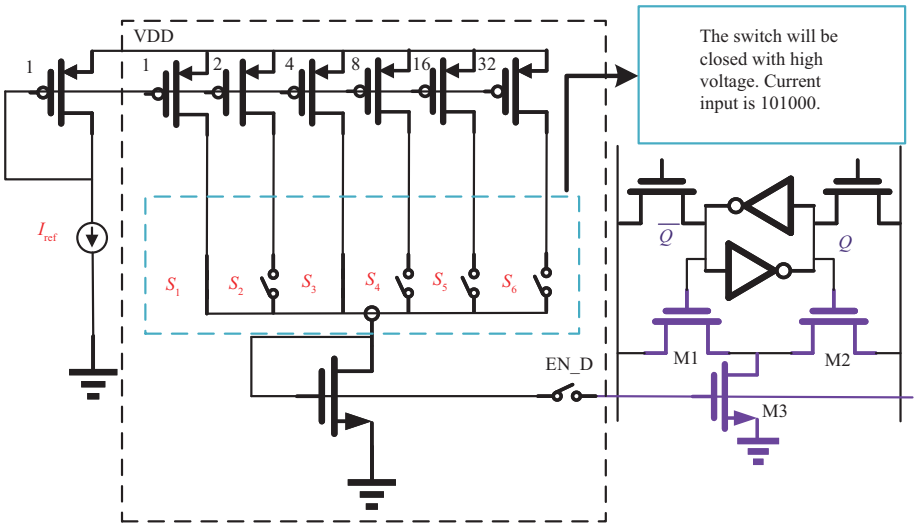

Figure 3 (Color online) Scheme of the digital interface.

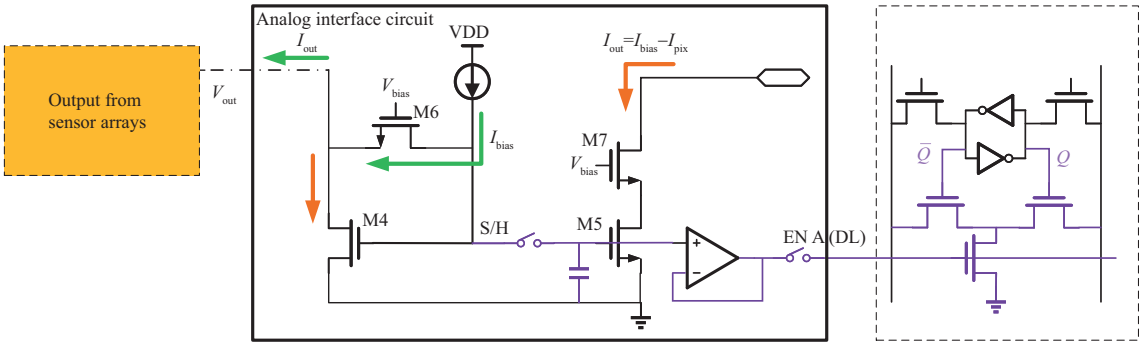

Figure 4 (Color online) Scheme of the analog interface.

the raw signal output from different sensors differs greatly in form, either digital or analog. To meet the processing needs of both signal modes, both types of interface circuits, digital and analog, and implement the multiplexing of DL lines in the circuit structure should be proposed [5, 17, 32]. Figure 3 shows the principle of the digital interface circuit. The digital interface circuit is essentially a small-scale current-mode DAC circuit, which is designed with a digital control switch and a binary proportional weighted mirror current source. Where the current $I_{\rm ref}$ denotes the actual current value corresponding to unit 1, which can be determined by $1:2:4:\cdots:2^d$ proportionally weighted to obtain multiplicative currents, and then these currents are directly superimposed to obtain analog current values with different digital bit equivalent accuracy. Finally, this converted current value is mirrored in the memory unit through the current mirror structure, and the function of converting digital sensing signals into analog currents and reproducing them in the memory unit is realized.

Figure 4 shows the analog interface circuit. The analog interface circuit consists of a current transmitter circuit and an analog buffer. The analog input interface utilizes a currentmirror structure with a negative feedback structure to stabilize the input interface voltage and thus reduce the input impedance. The introduction of negative feedback in the analog interface circuit allows the output of the current-mode pixel to have a stable voltage bias, thus very effectively improving the linearity of the current-mode pixel output current.

4.2 Scalable "Sensing with Computing" architectures

4.2.1 *Hierarchical memory-in-computing architecture*

Designing a memory-in-computing circuit for matrix multiplication operations in neural networks is indispensable, in which the array interface reuses space to build a multi-dimensional scalable architecture with the goal of improving the efficiency of data flow should be explored [5, 33]. This can be completed

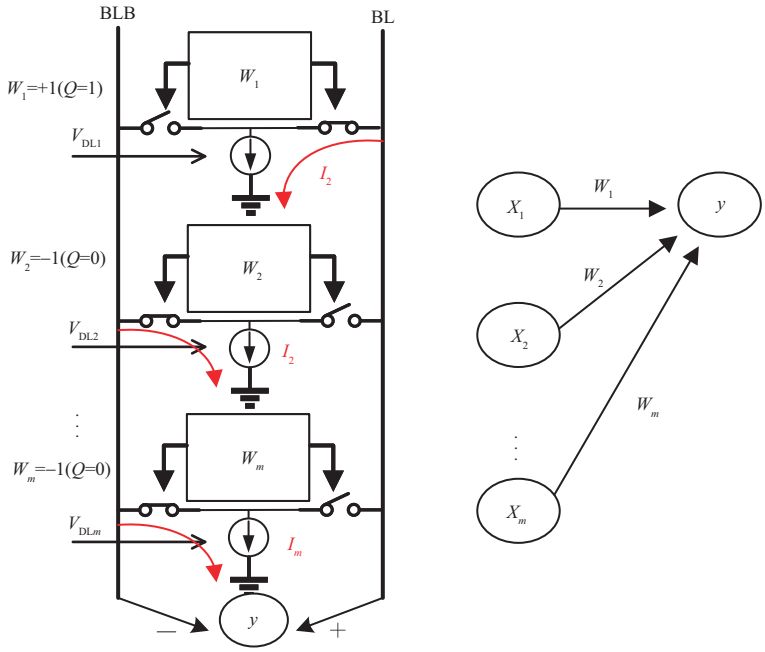

Figure 5 (Color online) Scheme of the vector convolution with computing-in-memory arrays.

by the design of a memory-in-computing circuit based on the previously proposed unit to realize the multiplication and accumulation of multiple input-output nodes. Similar to conventional SRAM arrays, the memory-in-computing circuit is mainly composed of basic memory cells expanded in two dimensions in both horizontal and vertical directions. Specifically, the SRAM read-write word lines and the control voltage signal of the current source are shared horizontally, and the bit lines BL and BLB are shared vertically for cell writing to complete current accumulation, as shown in Figure 5. Besides this conventional SRAM-based circuit, nonvolatile logic-in-memory circuits could also be used [34, 35].

Then, a multi-dimensional and scalable storage and computing architecture could be built with this memory-in-computing array as the core. A possible implementation scheme is shown in Figure 6. It includes the memory-in-computing array, digital/analog interface, and core control module. A 2×2 grid form is used to organize the whole structure, which is divided into the left, right, top, and bottom parts. The left and right parts correspond to different input modes. The arrays on the left side are intended to implement the processing of digital input signals. The arrays on the right side are for the analog input signals. The upper and lower parts correspond to different output modes. The upper part is connected to the analog on-chip network via analog neuron circuitry. The lower part is interconnected to the digital bus through the digital neuron circuit. The left and right sections with the top and bottom sections are connected by horizontal and vertical expansion switches, respectively. This grid-based topology allows fine-grained control in each grid point for parallel processing. Output and input scaling of the neural network layers can be achieved through horizontal and vertical expansion switch configurations. The upper and lower parts use time-division multiplexing to post-process the results of memory-in-computing arrays, thus realizing the expansion of the computing scale. The digital and analog interfaces can obtain sensing data simultaneously to realize the synchronous processing of multi-modal information. In addition, a digital logic circuit implements the core control module of the mixed-signal processing unit, which realizes the configuration of operation parameters, control of external input data, and the operation process.

4.2.2 *Multi-modal "Sensing with Computing" architecture*

Based on the specific requirements of multi-modal sensing algorithms, architecture optimization methods for multi-modal "Sensing with Computing" processors to achieve direct sensing data drive, efficient data flow transfer, and good scalability to support architectures for large-scale data input should be explored.

The "Sensing with Computing" architecture design shown in Figure 7 can be used to achieve multidimensional and scalable multi-modal intelligent sensing. The overall architecture adopts a mixed signal processing model with deep digital-analog fusion. It contains two sets of digital and analog interfaces

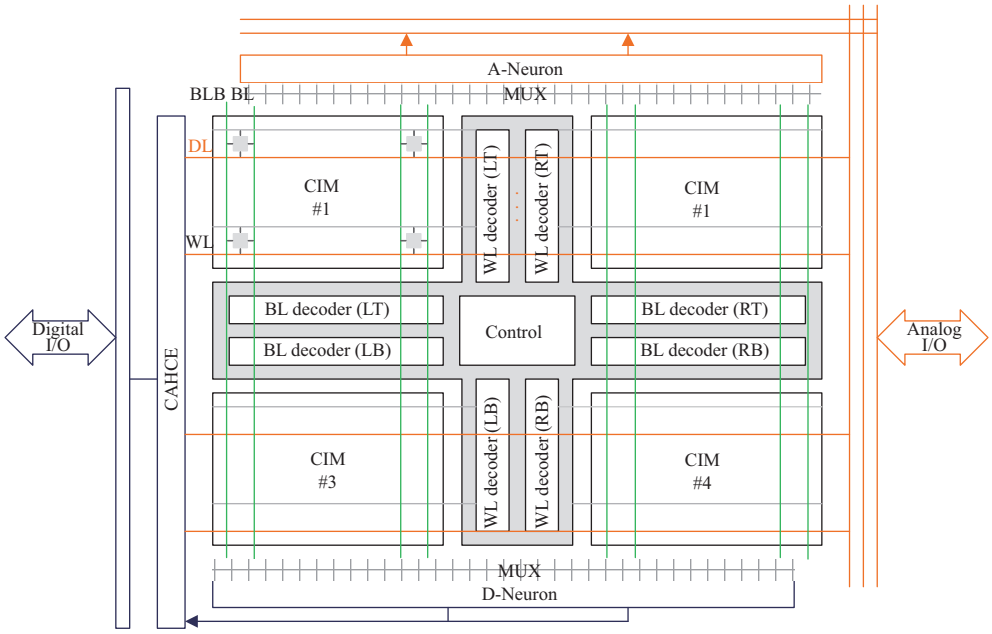

Figure 6 (Color online) Scalable circuit scheme of computing-in-memory arrays.

with different modes and multiple processing units for parallel computation. The dual-mode input and output interfaces are selected for different modalities of sensor data. The core memory-in-computing arrays complete computation. It enables energy-efficient and highly parallel computing while ensuring the scaling of neural network algorithms in terms of the number of layers. The analog sensors are directly interconnected to the memory-in-computing architecture via analog I/O. It implements highly parallel neural network computation driven directly by sensed data. The digital I/O interacts directly with the digital sensors externally. Data is written to the digital buffers and remapped via the data bus. The architecture allows the design and implementation of perceptual computing applications with sensors connected in different modalities and data forms. In addition to this, efficient control instructions are needed to achieve flexible control of the entire top-level processing flow to provide architecture-level support for the implementation of multiple modalities and different scales of sensing tasks. Finite state machines can be used to achieve control of different levels of architecture and circuitry in the chip.

4.2.3 *Neural network algorithm mapping method*

This part will further explore the mapping methods from neural network algorithms to the circuit architecture. It focuses on implementing the mapping of two common types of network layers, fully connected layers (FC) and convolutional layers, to the underlying circuit architecture. It should be acknowledged that although in-sensor computing technology can achieve low-power and high-performance circuit system designs, its generality is weaker compared to traditional technologies. However, from the perspective of applications, most of the application scenarios supported by in-sensor computing technology have strong specific requirements. Therefore, considering the future research and development trends, designers need to conduct comprehensive and in-depth research in various aspects such as technology generality, circuit energy efficiency, and development cost cycle.

For the FC layer, the mapping method shown in Figure 8 maps the FC layer into a single computing-in-memory array. When the dimension of input and output data is larger than the number of channels of the computing-in-memory array, the input and output data will be grouped and the mapping of the FC layer to the memory array is achieved using the time-division multiplexing method. The layer-to-layer data transfer is achieved by simulating an on-chip network with high parallelism and continuous signals. Further, considering that the input layer usually requires a larger input dimension and the shallower hidden layer may require a larger output width, a horizontal and vertical scaling mechanism is introduced for the 2×2 grid expansion of the memory array. It enables the multiplication in single-cycle operations.

In contrast to the fully connected layers, the mapping method of convolutional layers to the underlying

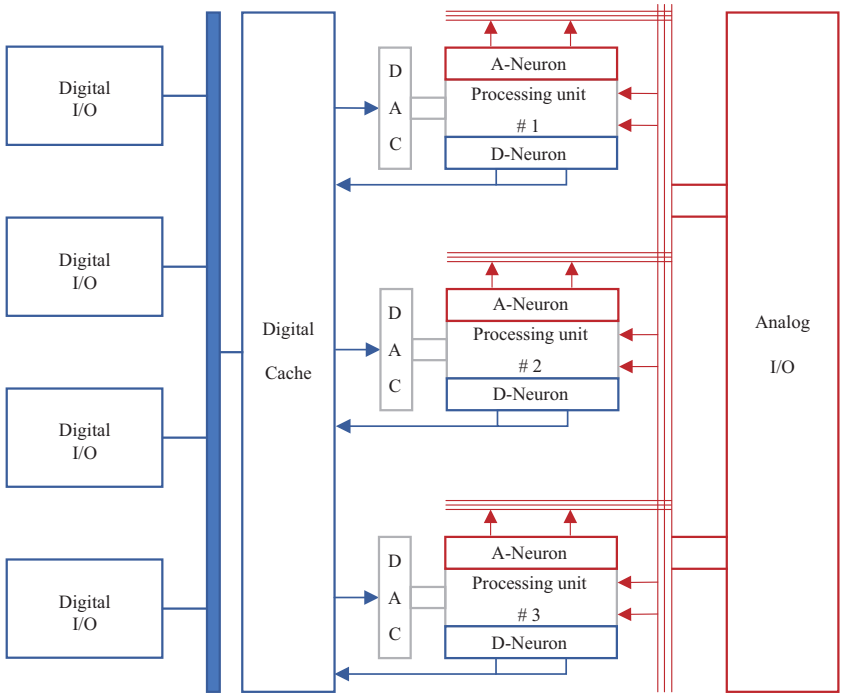

Figure 7 (Color online) Scheme of multi-modal "Sensing with Computing".

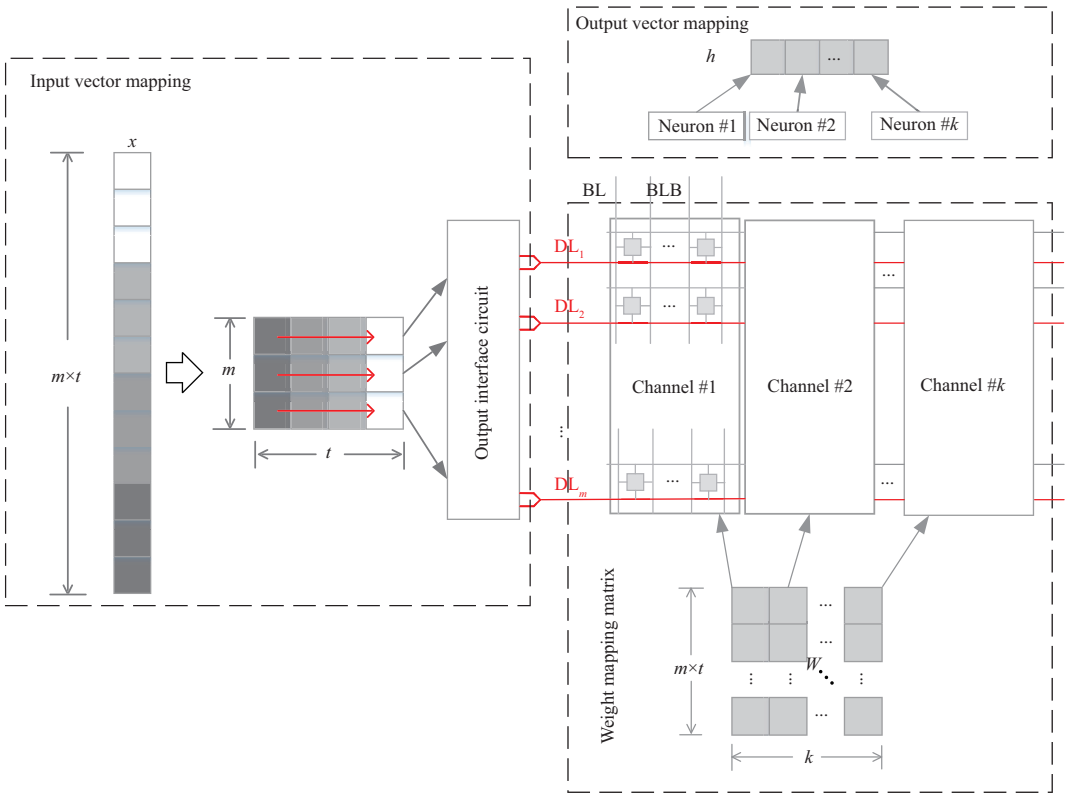

Figure 8 (Color online) Fully connected layer on-chip mapping approach.

architecture should address the issue of array signal unfolding and the correspondence of the output signal position in the feature pattern. This is shown in Figure 9. The mapping method unfolds the input data block while the corresponding weight data in the convolutional kernel are expanded into a column mapped into a storage channel of the arrays. The remaining rows in the channel are used for

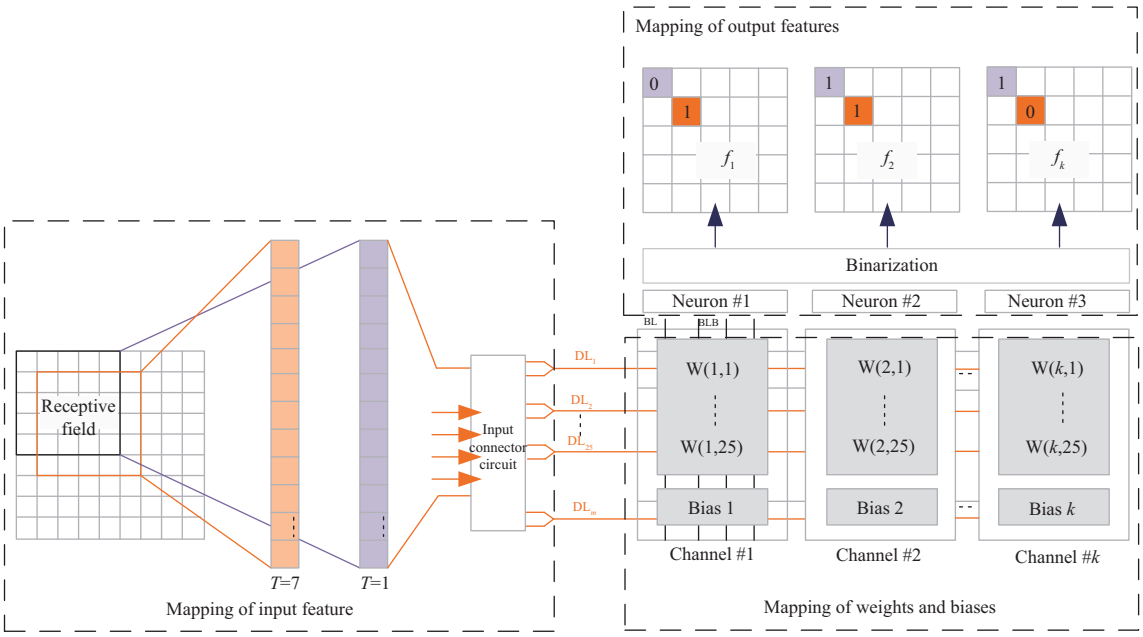

Figure 9 (Color online) On-chip mapping of convolutional layers for "Sensing with Computing" architecture.

bias value. The parallel computation of multiple channels in the memory array corresponds to multiple output channels and multiple convolutional kernels in the convolutional layer, thus computing multiple feature maps of the convolutional layer output in parallel. The data within individual feature maps are computed serially according to the clock. The output results are converted to binary data using a digital neuron circuit for feature map storage in the digital domain.

4.3 Demonstration and optimization for "Sensing with Computing"

4.3.1 *Algorithms-hardware co-design for smart sensing applications*

Based on the above circuit and architecture, this subsection focuses on an algorithm-hardware co-optimization method. The main objective of this method is to model the non-ideal factors including process deviations, device mismatch, and noise so that the optimization could be completed [36]. In general, sensing with computing requires cross-layer collaborative optimization design, mainly because applications that adopt this technology usually have high requirements for power consumption and energy efficiency. Optimization efforts solely focused on circuits or algorithms are not sufficient.

The flow of the optimization method is shown in Figure 10. First, Monte Carlo simulation is carried out in real time during the design of the hardware circuit to construct a behavioral-level model of the computational units. Based on this model, the convolutional neural network algorithm can replace the original exact computational units during training and forward inference. Then, we can get the influence of the non-ideal factor on accuracy of each circuit by robustness analysis. The circuit error magnitude is used as the performance optimization index for the next circuit design step. Finally, the above functional modules are integrated to obtain a complete set of software and hardware analysis methods. The circuit schematic and algorithm model are used as inputs, and the performance analysis results of each algorithm are used as outputs. This optimization method can automate the end-to-end algorithm-hardware analysis.

4.3.2 *Optimization of neural network for "Sensing with Computing" chips*

As the direct introduction of circuit deviations into the network training process may lead to the failure of convergence, a progressive error introduction training method could be used. First, the weights of the network are randomly divided into several groups. Then, the exact computational units are replaced with the behavioral-level model of the circuit units. Finally, multiple rounds of retraining are performed to progressively update the weights. In this way, the pre-set accuracy can be achieved after the non-ideal circuit characteristics are introduced into the set of weights. The above steps are repeated until the accuracy is maintained with the introduction of circuit errors in all the groups [36, 37].

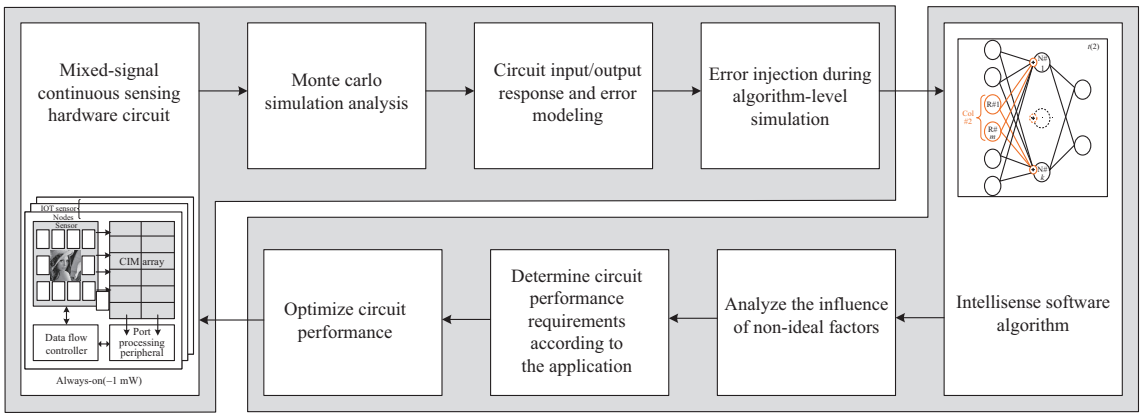

Figure 10 (Color online) Algorithm of the collaborative analysis scheme for circuit hardware and software.

4.3.3 *Multi-modal "Sensing with Computing" chip and demonstration system*

Based on the above optimization method, a complete "Sensing with Computing" chip for multi-modal sensor data could be implemented to build a demonstration system. First, the designers should confirm the sensor scales and algorithm performance index according to the specific sensing task requirements. Then, the design details based on the previous circuit and architecture will be obtained. Finally, the expected performance of the chip will be evaluated and optimized using the optimization method. Tape out after the chip design should meet the application scenario constraints on speed, power consumption, accuracy, and other indicators. After the initial verification is completed, the chip test and demonstration system can be built based on the mature embedded development platform.

In our previous work, presented in [33], we proposed an innovative architecture that combines sensing and computing processing to integrate the multiply-and-accumulation (MAC) operation into the image sensor, enabling low power design for intelligent visual perception. Our circuit design incorporates the feature of correlated double sampling directly into the MAC operation, effectively eliminating power consumption associated with data conversion from photocurrents to pixel values. Instead of sending raw images, our design directly transmits 1st-layer binary activations to binarized neural network (BNN) processors, enabling coarse and simple classification. When an object of interest is detected, the sensor switches to sensor mode, functioning as a conventional rolling-shutter CMOS image sensor (CIS) and sending raw images to a full-precision CNN processor for fine-grained recognition or segmentation. To demonstrate the feasibility of our approach, we fabricated a prototype chip with a 4×4 array, as shown in Figure 11, and the measurement results indicate that our chip design can operate at a high frame rate of 1000 frames per second (fps), achieve impressive energy efficiency of 1.32 tera operations per second per watt (TOP/(s·W)) and save 61% energy than state-of-the-art work [38].

5 Future directions and challenges

In the coming future, more and more sensors will be deployed in our daily lives, which will be of much more intelligence and work days and nights. From the paradigm-shifting of the framework of the information society, conventional cloud computing and big data processing would be updated to some new and diversified aspects, where the edge sides would be of more and more importance. The sensor parts would be of intelligent capabilities and output analyzed results, instead of only collecting and sending raw data to the servers. Tasks would be split, and then deployed on both sensor parts and server parts, according to the system constraints of real-time performance, energy efficiency, and task accuracy. Both sides would be of the same importance to the future information society.

As mentioned in this paper on smart sensing systems, in the coming future, much smarter sensor nodes with smaller sizes and longer working times will lead to urgent challenges for researchers and designers. Those are the following. (1) For the sensor devices, more kinds of sensors, including prevailing image sensors, auditory, and tactile sensors, would be adopted, such as ToF sensors and vibration sensors. The intelligence of all kinds of sensors would be enhanced and the sensor interface would be carefully designed to meet these requirements. We believe that device innovation can also play a significant role in bringing in-sensor computing functionalities to reality, as demonstrated in recent studies [39, 40]. This

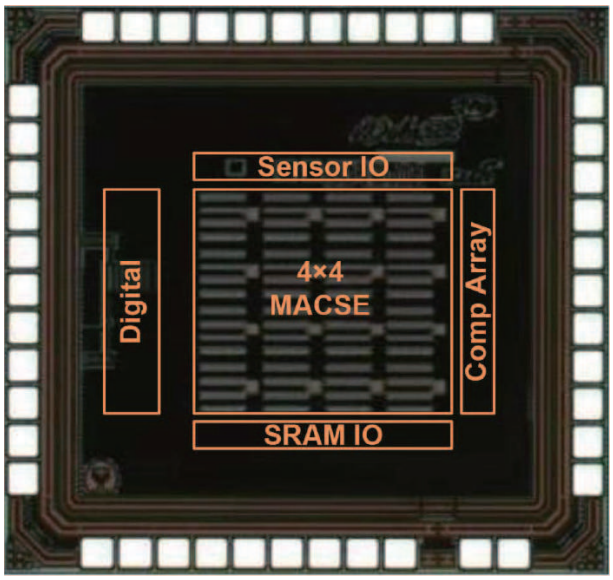

Figure 11 (Color online) Photo of the chip.

can involve the development of novel materials, fabrication techniques, and designs that enable advanced computing capabilities within the sensor itself. Advancements in nanotechnology, microfabrication, and other emerging technologies can lead to the creation of new types of sensors that can process data, perform computations, and make decisions at the sensor level. This can reduce the need for transmitting raw sensor data to external computing resources, which can lead to improved efficiency, reduced power consumption, and lower latency in certain applications. (2) Smart sensing systems would be equipped "ANYWHERE" and working "ANYTIME", which must support more applications and run more complex perception algorithms on the sensor sides. The popular artificial intelligence algorithms, such as DNN, RNN, reinforcement learning, and knowledge graph, would run on sensors and process data directly. Moreover, these kinds of algorithms would be compressed and optimized to be suited to the less computing resource on the sensor sides, to balance the computing efficiency and task accuracy. (3) As for the evolution of the energy efficient computing architectures and chip designs for smart sensors, the "Sensing with Computing" architectures would be optimized with more efficient hardware realizations, such as compute-in-memory and sub-threshold circuits. The different parts of the modules would be packed using some advanced package technologies, such as SiP and chiplet, which are more compatible with the heterogeneous designs. We think that it is possible to do large-scale demonstrations of the in-sensor computing paradigm using Si (silicon) technology tape out, which may require significant resources and expertise in semiconductor design and fabrication. However, with the increasing demand for advanced sensor systems and the continued advancement of semiconductor technology, it is feasible to conduct large-scale demonstrations of the in-sensor computing paradigm using Si technology tape out in the future. Of course, we also see many challenges. One of the significant challenges is the slow development of semiconductor technology in the post-Moore's Law era. The difficulty of development is increasing, and issues such as power consumption and energy efficiency are becoming more severe. Relying solely on clever circuit system designs by designers may not be sustainable. Therefore, we believe that the research and development of new materials and new processes are excellent opportunities for the development of an in-sensor computing paradigm.

Additionally, to enhance the capabilities of smart sensor nodes, integrating energy harvesting modules, wireless communication modules, and security modules, would be the direction for smart sensing systems, which would be helpful to the construction of future "Cloud-Edge" balanced information systems.

6 Summary

This paper provides a review of smart sensing systems with "Sensing with Computing" architectures. "The First Principle" of designing energy-efficient smart sensing systems has been studied, which is "Where there is an interface, there is a bottleneck (we also describe this as '凡是接口, 皆为瓶颈' in

Chinese)", and that is, any interfaces are the performance bottlenecks. Following this principle, the bottlenecks of data converters and memory access have been eliminated in the new "Sensing with Computing" architectures to realize future smart sensing systems. Here, the analog computing technologies have been adopted and carefully designed, for the sensors are the specific devices transferring physical information to analog electrical information (such as analog voltage and current signals). Additionally, the energy-efficient CIM methods also have been used in these architectures to be running artificial intelligent tasks. To sum up, the paper provides details of "Sensing with Computing" methods in circuits and architectures level, as well as algorithm-hardware optimization technologies; then gives some future directions and design challenges of the smart sensing systems.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 92164203).

References

1 Atzori L, Iera A, Morabito G. The Internet of Things: a survey. Comput Networks, 2010, 54: 2787–2805
2 International Data Corporation. IDCs IoT Growth Expection. 2019[2019-06-18]. https://blogs.idc.com/2019/11/04/how-you-contribute-to-todays-growing-datasphere-and-itsenterprise-impact/
3 Sethi P, Sarangi S R. Internet of Things: architectures, protocols, and applications. J Electrical Comput Eng, 2017, 2017: 1–25
4 Shi W, Cao J, Zhang Q, et al. Edge computing: vision and challenges. IEEE Internet Things J, 2016, 3: 637–646
5 Liu Z Y, Ren E X, Qiao F, et al. NS-CIM: a current-mode computation-in-memory architecture enabling near-sensor processing for intelligent IoT vision nodes. IEEE Trans Circuits Syst I, 2020, 67: 2909–2922
6 Forchheimer R, Astrom A. Near-sensor image processing: a new paradigm. IEEE Trans Image Process, 1994, 3: 736–746
7 Waldrop M M. The chips are down for Moore's law. Nature, 2016, 530: 144–147
8 Esmaeilzadeh H, Blem E, Amant R S, et al. Dark silicon and the end of multicore scaling. In: Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA), San Jose, 2011. 365–376
9 Yin X Z. Cross-layer integrated designs for energy efficient computing. Dissertation for Ph.D. Degree. South Bend: University of Notre Dame, 2019
10 Din S, Xu W, Cheng L K, et al. A stretchable multimodal sensor for soft robotic applications. IEEE Sens J, 2017, 17: 5678–5686
11 Wang Q, Ling S J, Liang X P, et al. Self-healable multifunctional electronic tattoos based on silk and graphene. Adv Funct Mater, 2019, 29: 1808695
12 Lee V T, Alaghi A, Hayes J P, et al. Energy-efficient hybrid stochastic-binary neural networks for near-sensor computing. In: Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE), Lausanne, 2017. 13–18
13 Badami K M H, Lauwereins S, Meert W, et al. A 90 nm CMOS, 6 μW power-proportional acoustic sensing frontend for voice activity detection. IEEE J Solid-State Circuits, 2016, 51: 291–302
14 Young C, Omid-Zohoor A, Lajevardi P, et al. A data-compressive 1.5/2.75-bit log-gradient QVGA image sensor with multi-scale readout for always-on object detection. IEEE J Solid-State Circuits, 2019, 54: 2932–2946
15 Bankman D, Yang L, Moons B, et al. An always-on 3.8J/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28 nm CMOS. In: Proceedings of IEEE International Solid-State Circuits Conference, San Francisco, 2018. 222–224
16 LiKamWa R, Hou Y, Gao Y, et al. RedEye: analog ConvNet image sensor architecture for continuous mobile vision. In: Proceedings of ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, 2016. 255–266
17 Chen Z, Liu X, Yang H, et al. Processing near sensor architecture in mixed-signal domain with CMOS image sensor of convolutional-kernel-readout method. IEEE Trans Circuits Syst I, 2020, 67: 389–400
18 Du Z, Fasthuber R, Chen T, et al. ShiDianNao: shifting vision processing closer to the sensor. In: Proceedings of ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA), Portland, 2015. 92–104
19 Chen Y H, Krishna T, Emer J S, et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J Solid-State Circuits, 2017, 52: 127–138
20 Chen Y H, Yang T J, Emer J S, et al. Eyeriss v2: a flexible accelerator for emerging deep neural networks on mobile devices. IEEE J Emerg Sel Top Circuits Syst, 2019, 9: 292–308
21 Ni K, Sharma P, Zhang J, et al. Critical role of interlayer in $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric FET nonvolatile memory performance. IEEE Trans Electron Devices, 2018, 65: 2461–2469
22 Ni K, Jerry M, Smith J A, et al. A circuit compatible accurate compact model for ferroelectric-FETs. In: Proceedings of IEEE Symposium on VLSI Technology, Honolulu, 2018. 131–132
23 Zhang X, Chen X, Han Y. FeMAT: exploring in-memory processing in multifunctional FeFET-based memory array. In: Proceedings of IEEE 37th International Conference on Computer Design (ICCD), Abu Dhabi, 2019. 541–549
24 Ni K, Smith J A, Grisafe B, et al. SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), San Francisco, 2018. 1–4
25 Jain S, Ranjan A, Roy K, et al. Computing in memory with spin-transfer torque magnetic RAM. IEEE Trans VLSI Syst, 2018, 26: 470–483
26 Ni K, Dutta S, Datta S. Ferroelectrics: from memory to computing. In: Proceedings of the 25th Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, 2020. 401–406
27 Chi P, Li S C, Xu C, et al. PRIME: a novel processing-in-memory architecture for neural network computation in ReRAM-based main memory. In: Proceedings of ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), Seoul, 2016. 27–39
28 Reis D, Niemier M, Hu X S. Computing in memory with FeFETs. In: Proceedings of the International Symposium on Low Power Electronics and Design (ISLPED'18), New York, 2018. 24
29 Wang Z, Joshi S, Savel'ev S, et al. Fully memristive neural networks for pattern classification with unsupervised learning. Nat Electron, 2018, 1: 137–145
30 Tang K, Wang X, Dong K, et al. A thermal radiation modulation platform by emissivity engineering with graded metal-insulator transition. Adv Mater, 2020, 32: 1907071
31 Tang K, Dong K, Nicolai C J, et al. Millikelvin-resolved ambient thermography. Sci Adv, 2020, 6: eabd8688

32 Li Q, Liu C L, Dong P Y, et al. NS-FDN: near-sensor processing architecture of feature-configurable distributed network for beyond-real-time always-on keyword spotting. IEEE Trans Circuits Syst I, 2021, 68: 1892–1905
33 Xu H, Li Z R, Lin N C, et al. MACSen: a processing-in-sensor architecture integrating MAC operations into image sensor for ultra-low-power BNN-based intelligent visual perception. IEEE Trans Circuits Syst II, 2021, 68: 627–631
34 Yin X, Niemier M, Hu X S. Design and benchmarking of ferroelectric FET based TCAM. In: Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE), Lausanne, 2017. 1444–1449
35 Yin X, Chen X, Niemier M, et al. Ferroelectric FETs-based nonvolatile logic-in-memory circuits. IEEE Trans VLSI Syst, 2019, 27: 159–172
36 Jia K, Liu Z, Wei Q, et al. Calibrating process variation at system level with in-situ low-precision transfer learning for analog neural network processors. In: Proceedings of the 55th ACM/ESDA/IEEE Design Automation Conference (DAC), San Francisco, 2018. 1–6
37 Liu Z Y, Jia K G, Liu W Q, et al. INA: incremental network approximation algorithm for limited precision deep neural networks. In: Proceedings of International Conference On Computer-Aided Design (ICCAD), Westminster, 2019. 1–7
38 Hsu T-H, Chen Y-K, Wen T-H, et al. A 0.5 V real-time computational CMOS image sensor with programmable kernel for always-on feature extraction. In: Proceedings of IEEE Asian Solid-State Circuits Conference (ASSCC), Macau, 2019. 33–34
39 Zhou F C, Chai Y. Near-sensor and in-sensor computing. Nat Electron, 2020, 3: 664–671
40 Liao F Y, Zhou Z, Kim B J, et al. Bioinspired in-sensor visual adaptation for accurate perception. Nat Electron, 2022, 5: 84–91