The Development of Silicon for AI: Different Design Approaches

Kyuho Jason Lee, *Senior Member, IEEE*, Jinmook Lee, *Member, IEEE*, Sungpill Choi, *Member, IEEE*, and Hoi-Jun Yoo, *Fellow, IEEE*

***Abstract*— This paper provides a review of design approaches towards artificial intelligence (AI) System-on-Chip. AI algorithms have progressed over the past decades from perceptron-based neural network (NN) and neuro-fuzzy (NF) system to today's deep neural network (DNN) and neuromorphic computing. Recent DNN hardware accelerators focus on energy-efficient integration of digital circuits to realize real-time DNN operation while neuromorphic processors deploy new memory technologies with analog computation for low power consumption. However, different design approaches can be applied to such processor implementation with their pros and cons. This paper reviews from the early processor designs for NN and NF in both mixed-mode and digital implementations to the recent DNN SoC designs that we have proposed for a decade. The former content deals with NN and NF processors used as a functional building block of a machine vision SoC, while the latter concentrates on integration of the whole DNN function. We also provide a discussion on the approaches, and provide perspective on future research directions.**

***Index Terms*— Mixed-mode SoC, neural network processor, neuro-fuzzy processor, deep learning SoC.**

I. Introduction

PRIOR to the era of Deep Learning (DL) that deals with deep neural network (DNN) architecture, there were various types of machine intelligence algorithm such as multilayer perceptron (MLP), fuzzy inference system (FIS), neuro-fuzzy (NF) system, etc. It took several decades for them to become today's DL that has been rapidly advanced and used for wide range of computer vision applications such as image classification [1], object detection [2]–[5], and autonomous vehicles [6] due to its high accuracy, where variant of DNN is used for different applications. For example, convolutional neural network (CNN) is widely used for image processing and recurrent neural network (RNN) is used for natural language processing. DNNs consist of hundreds of layers that they require huge amount of computations and memory footprints. Thus, most of the applications target at bulky systems with high performance GPUs or servers rather than mobile applications although the needs of mobile DNN are increasing nowadays. More advanced DL researches [7], [8] have given promise of success on mobile applications by providing new user experience or useful functionality with sensory data such as image, video, and voice that can be easily collected on mobile devices. However, the computationally expensive DL algorithms suffer from satisfying processing speed requirements on mobile applications with limited hardware resources and power budget. In spite of algorithmic efforts to lighten DNN architectures [9]– [17], they could not resolve fundamental problem of heavy computational cost. Thus, the importance of hardware accelerator design is gaining more attention worldwide and they are actively investigated to efficiently run DNNs in real time with low power. What lies on the other side of the artificial DNNs are neuromorphic chips that aim at modeling the operations of biological brain to achieve high energy efficiency since human brain is well-known as the most energy-efficient computer. The development of neuromorphic chip also gained lots of attention among researchers. However, there have been many efforts to develop dedicated processors from the early age of neural networks with various hardware design approaches and it is important to analyze the previous approaches to develop advanced System-on-Chips (SoCs).

There are three different hardware design approaches (analog, digital, and mixed-mode implementations) with their own pros and cons. In general, digital neural network (NN) processors [18], [19] have advantages that they can achieve high accuracy, flexibility, and programmability, but they consume huge power and area due to the large amount of data transaction and fast operation speed. On the other hand, analog NN VLSIs enable low-cost parallelism with low-power computation, but their inaccurate circuit parameters induced by noise and low precision degrade accuracy [20], [21]. Several mixed-mode SoCs took the advantages of both analog and digital implementations obtaining low-power consumption within small area, but it suffers from domain conversion overhead cost. In this paper, we provide a review of the design methodologies by introducing various processors with versatile design approaches for NN/FIS/NF/DNN acceleration.

The rest of this paper is organized as follows. In Section II, related works about NN processor design including today's DNN and neuromorphic processors will be explained as previous works. Section III will review several NN/NF processors used as a functional block of an intelligent computer

Manuscript received February 2, 2020; revised April 30, 2020; accepted May 17, 2020. Date of publication June 1, 2020; date of current version December 1, 2020. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) under Grant 2019R1C1C1009857 and in part by the Samsung Electronics. This article was recommended by Associate Editor C. H. Chang. *(Corresponding author: Kyuho Jason Lee.)*

Kyuho Jason Lee is with the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology, Ulsan 44919, South Korea (e-mail: kyuho.jsn.lee@unist.ac.kr).

Jinmook Lee, Sungpill Choi, and Hoi-Jun Yoo are with the School of Electrical Engineering, KAIST, Daejeon 34141, South Korea.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2020.2996625

vision SoC. They are classified into three categories according to their design methodology: analog, digital, and mixed-mode implementation. Each processor is dedicated to different functional blocks of the vision SoC pipeline, e.g. visual attention module, classification, workload prediction, etc. Then, recent DNN processors that incorporate with fully-digital circuit implementation will be explored in Section IV. Section V will discuss neuromorphic processors. Finally, we provide our insights and perspectives on future research directions in Section VI, followed by conclusion in Section VII.

II. Related Works – Neural Network Processors

Many researchers investigated on developing DNN and neuromorphic SoCs recently. Speaking of neuromorphic designs, Lu *et al* [22] implemented clustering algorithm in analog domain with floating-gate non-volatile memory. Zhang *et al* [23] implemented matrix-multiplying ADC that enables multiplications with input samples, which is used for feature extraction in classification algorithm. Kim *et al* [24] proposed sparse coding ASIC to enable training of sparse representation of images for feature detection and recognition using spiking neural network. They also developed a simple object recognition system that is composed of spiking neural network inference module in [25]. Lee *et al* [26] designed energy-efficient matrix multiplier with switched capacitor scheme for classification applications on analog front-end. Ambrogio *et al* [27] used resistive switching memory, so-called RRAM, to emulate the function of spiking neural network, and Zhang *et al* proposed in-memory computation scheme using standard 6-T SRAM array in [28] and [29].

In the viewpoint of DNN processor design, Tsai *et al* proposed a digital DL processor for big-data applications such as data filtering and data estimation kernels [30]. Park *et al* attempted to implement DNN training on silicon with scalable architecture and massively parallel thread-level parallelisms in [31] and [32]. They expanded their work to develop a user experience glass system using embedded DL SoC [33]. It is capable of both simple NN training and inferencing. Lee *et al* [34] developed an SoC for advanced driver assistance system using RNN and FIS accelerators for energy-efficient automotive applications. This work also supported simple on-line learning of RNN using dedicated MAC processing elements (PEs) with SIMD extension cores. Chen *et al* [35] proposed a CNN accelerator utilizing data-reuse pattern and its dedicated hardware architecture, and Sim *et al* [36] also designed a digital CNN processor with multi-range MAC unit and kernel compression scheme to reduce off-chip memory access. Moons *et al* designed a CNN processor with 2-D MAC array and scalable bit precision in [37], and enhanced the energy-efficiency with zero guarding and wide-range voltage-frequency scaling in [38]. Knag *et al* [39] proposed a convolutional restricted Boltzmann machine processor to infer support vector machine classifier with integrated sparse convolution unit. Bang *et al* [40] and Whatmough *et al* [41] introduced efficient accelerators for fully-connected (FC) DNN, with their hardware architectures optimized for matrix multiplication. Desoli *et al* integrated low-power DNN SoC with other hardware components such as DSP or dedicated direct memory access in [42]. More works were exposed to integrate digital circuits for different types of DNNs, from FC-DNNs to CNN and RNN [43]– [48]. The trend of recent design techniques show high-performance DNN SoCs are implemented in digital while ultra-low-power SoCs are mixed-mode implementation.

III. Neural Network / Neuro-Fuzzy Processors as a Functional Building Block of System-on-Chip

There have been many researches to deploy NN and NF algorithms as a functional block of a machine vision system. The details of such building block processors are introduced.

A. Fully-Analog Implementation

There were several attempts to analog circuit implementation of NN [57], [21], [89], [90] and spiking neural network (SNN), which will be introduced in Section V. These fully-analog circuit implementations were proposed for functional blocks used in NN/NF such as synapse or sigmoid activation generation. The reason is that analog circuit lacks of programmability while NN/NF operations requires training and setting many parameters by nature. Most of the analog-based neural networks were assisted by digital circuits for flexible control of the analog arrays [91]–[93]. Thus, full integration of NN/NF could be classified as mixed-mode and digital designs.

B. Mixed-Mode Implementation

Most of mixed-mode processors utilized analog circuits for feedforward operation of NN and NF due to its low power consumption. In addition, analog design saves area since analog multipliers are usually smaller than digital multipliers. Also, current-mode analog circuit employs massively parallel architecture with simple current summation based on Kirchhoff's current law, therefore, no additional adder circuit design is required. Although voltage-mode circuits have some advantages over current-mode circuits, it requires both of multipliers and adders for NN/NF operation. Therefore, current-mode circuits are preferably designed in many cases. On the other hand, digital circuits are used for training and controlling the analog parameters since they provide high programmability and accurate calculations with high bit precision.

Our first NF mixed-mode circuit for object detection [49] was designed as a functional block of the whole object recognition SoC [50], to achieve high performance and low power with small area overhead. The object detection engine detects contour of target objects around the ambiguous object boundaries in the input image. After finding the seed points during region-of-interest (ROI) generation in the entire object recognition pipeline, the NF processor expands the ROI of an object around neighboring pixels of the seed until it finds the object boundary by using three parameters for the classification homogeneity metric: pixel intensities, saliency, and location. In the proposed NF architecture, Gaussian fuzzy membership function and single-layer perceptron are utilized to measure the similarity of neighboring pixels and to classify boundaries, respectively. The processor, depicted in Fig. 1,

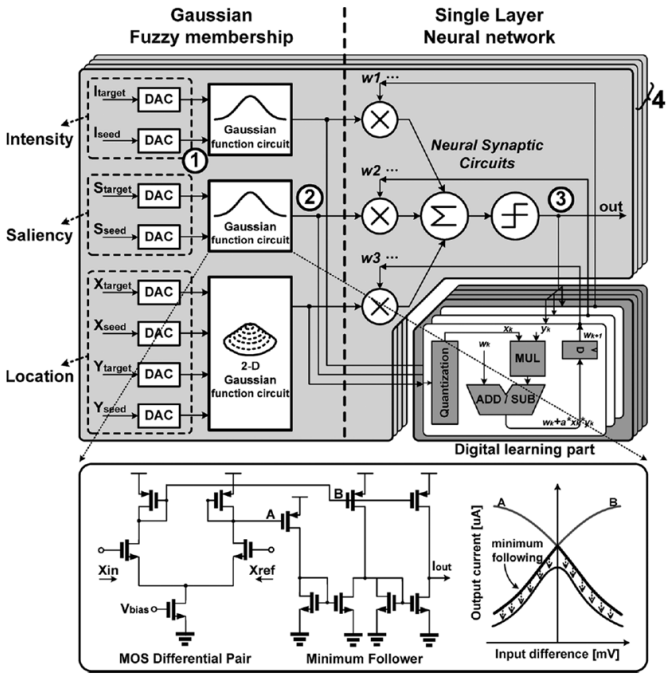


Fig. 1. Hardware architecture of the mixed-mode neuro-fuzzy object detection engine.

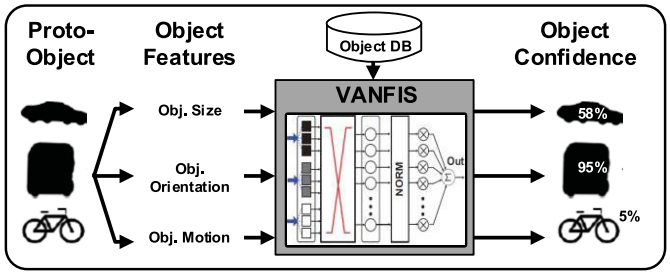


Fig. 2. Example use-case of the VANIFS in object recognition.

consists of a current-mode analog datapath for feedforward neuro-fuzzy operation and a digital processor for training the synaptic weights with Hebbian learning [51]. For weight multiplication with small number of transistors, a binary-weighted current mirror is proposed for current multiplier. The Gaussian function circuits are cascaded to generate 2-D Gaussian functions to measure similarity on pixel location x and y. The object detection engine performs ROI detection for an object within 7 μs at 200 MHz frequency. It is fabricated in 0.13 μm CMOS technology and reduced area and power by 59% and 44% compared with the fully-digital implementation in the same process technology, respectively.

Next, a versatile adaptive neuro-fuzzy inference system (VANFIS) hardware is designed for multiple purposes of classification [52]. The VANFIS was used for object classification and dynamic workload prediction to increase energy efficiency of the object recognition SoC [53] by adapting to different types of input vector. Fig. 2 depicts an example use-case of object classifier. It measures the similarity between proto-objects and fuzzy rules of target objects to estimate the confidence level of the input object. It is also used for efficient hardware control by comparing current status of the SoC

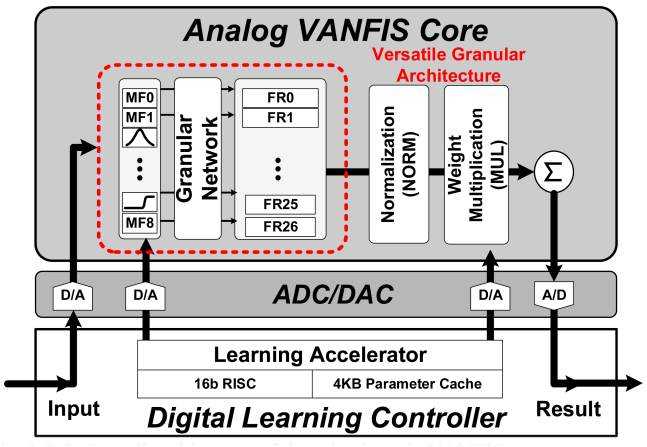


Fig. 3. Overall architecture of the mixed-mode VANFIS core.

with pre-trained workload history to predict the future workload to make the SoC achieve the highest energy efficiency. Fig. 3 shows the mixed-mode VANFIS architecture. The analog VANFIS core performs classification and prediction and the digital learning controller trains weight values. The analog core exploits current-mode circuits for simple current summation and division operations [54]. The granular network selects the configuration of connections between membership function layer and fuzzy rules layer, and provides flexible control with different kinds of input. Since the VANFIS processor uses a hardware-oriented weight perturbation algorithm for learning, it integrated perturbation multiplying-DAC circuits with current rail for signed operation. Implemented with current-mode circuits, the mixed-mode VANFIS processor that is fabricated in 0.13 μm CMOS technology saved area and power by 56% and 85%, respectively, compared with the equivalent digital implementation [53].

Another mixed-mode Intelligent Reconfigurable Integrated System (IRIS) SoC is introduced for multi-purpose application of NN and FIS [55]. The overall architecture is shown in Fig. 4. It is composed of a reconfigurable analog PE cluster that is capable of computing various NN and FIS topologies, a global/local learning accelerator, a MAC unit, memory banks, a RISC controller, and data converters. The analog core contains 32×32 PEs for parallel MAC operation of NN where each PE consists of multiplying-DAC and SRAM cell array, and 32 normalization and nonlinear function circuits for NN activation functions. The reconfigurable processor is capable of accelerating MLP, radial basis function neural network (RBFNN) [56], and RNN by changing signal paths of the analog PE cluster, which takes 6 μs to multiply a 32×32 matrix by a 32-D vector and 560 ns for one inner-product operation. As a result, the mixed-mode design reduced power and area by 71.2% and 54% compared with equivalent digital design; the SoC fabricated in 0.13 μm CMOS technology achieves 1 *mJ/frame* energy efficiency and consumes 57 mW on average for object recognition.

Because RBFNN is widely used as a classifier for its high accuracy [57], a mixed-mode RBFNN classifier deploying current-mode circuits is proposed for low-power yet highly-accurate scene classification in [58]. However, current-mode circuits operating with small current are

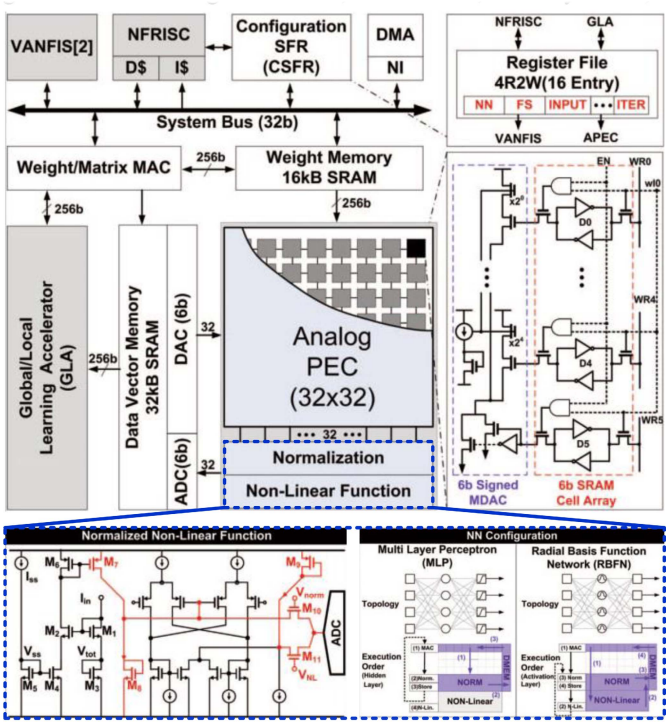


Fig. 4. Overall architecture of the IRIS SoC.

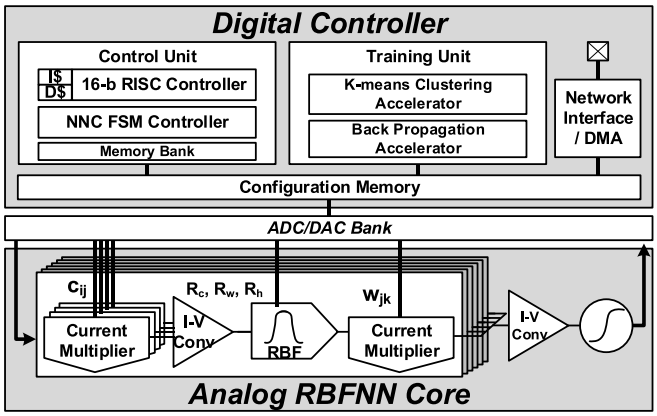


Fig. 5. Hardware architecture of the mixed-mode RBFNN classifier.

vulnerable to environmental noise such as temperature variation and supply voltage variation. Even a small current variation induced by the environmental noise critically degrades the classification accuracy by transforming the shape of pre-trained radial basis functions. The problem becomes worse when it comes to mobile platforms without stable power source such as unmanned aerial vehicles. To make the mixed-mode processor tolerant to such noise, the proposed RBFNN classifier contains temperature and supply voltage variation compensation circuits, which outputs stable current despite the variations. Moreover, the radial basis function circuit provides high programmability that it can also produce sigmoidal functions, therefore can be used for MLP classifier. As the datapath depicted in Fig. 5 shows, classification is performed in the analog core while digital controller is in charge of controlling configuration of analog connections and training of RBFNN parameters. The mixed-mode classifier fabricated in 0.13 μm CMOS technology saves area and

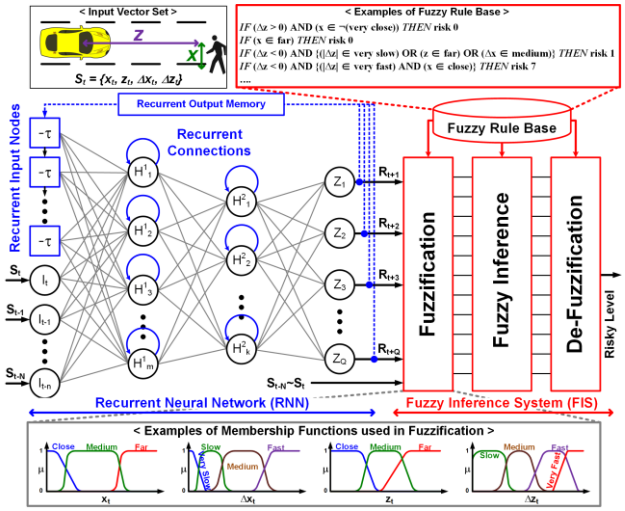


Fig. 6. The RNN-FIS algorithm for deep risk prediction.

power by 84% and 82%, respectively, compared with fully-digital implementation. It classifies global scene by taking HMAX [59] features as input within the brain-inspired object recognition processor [60]. As a result, the whole object recognition processor achieves 84% of visual attention accuracy and 96% of object recognition accuracy with 200 objects.

Another NF processor with RNN-FIS shown in Fig. 6, is proposed for automotive black box in [61]. The RNN part predicts the future motion status $S_t = \{x_t, z_t, \Delta x_t, \Delta z_t\}$ of detected object and the FIS part classifies the object's behavior and returns risky level using membership functions and fuzzy rules. The algorithm alerts drivers to the risky objects that are about to be collided with the vehicle in driving-mode while it triggers surveillance recording when object is getting closer to harm the vehicle in parked-mode to extend life-time of recording with limited capacity of battery. It features dual-mode operations: 1) the processor runs with the whole digital core in driving-mode to ensure high performance computation for real-time operation with deep risk prediction, and 2) mixed-mode circuit operation is activated in parked-mode for ultra-low-power consumption while some of digital cores are turned off (dark colored). Figure 7 shows the mixed-mode RNN-FIS hardware that consists of analog core for feedforward RNN-FIS acceleration and a digital controller for on-line training and control of the analog parameters. The analog core is also implemented with current-mode circuits to save area and power. Moreover, it operates with the maximal current of 8 μA, resulting in 156 μW for one feedforward RNN-FIS operation. To reduce much power, 6 out of 8 neuron units embedding SIMD extension cores as well as the fuzzy accelerator in the digital controller are gated. The SoC is fabricated in 65 *nm* CMOS technology and achieves high performance (502 *GOPS*) in driving-mode and low power (0.984 *mW*) in parked-mode. Thanks to the mixed-mode implementation, the total area and power consumption are reduced by 64% and 39%, respectively, compared with the fully-digital implementation.

Martinez *et al* [82] designed mixed-mode neural network by proposing digitally programmable multipliers for linearization

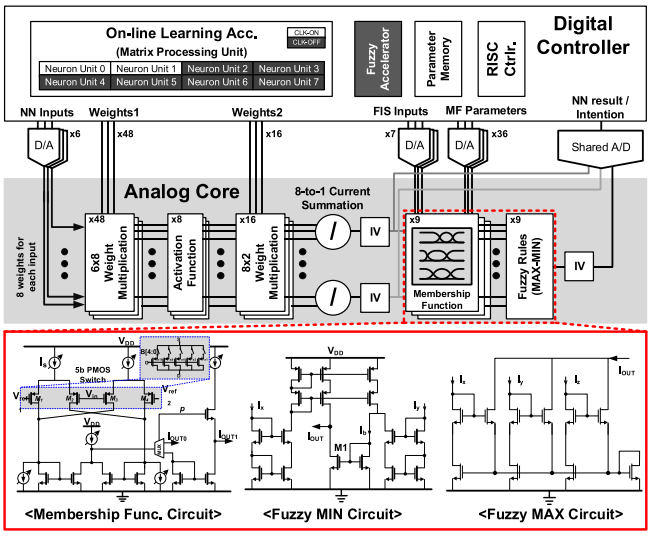


Fig. 7. Hardware architecture of the mixed-mode RNN-FIS deep risk prediction engine.

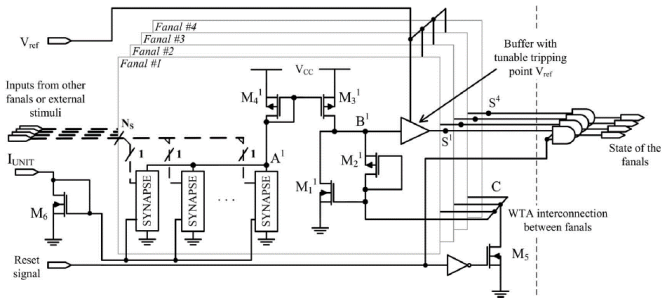


Fig. 8. Hardware architecture of neuron in encoded neural network [86].

of the NTC thermistor. The gain of analog multiplier is controlled by digital switches, but they needed adder circuit since they used voltage-mode circuits. Saffar *et al* [87] proposed multiplying-DAC and analog neurons for XOR net computation. The processor designed with 0.18 μm CMOS technology consumes 0.538 mW with 100 MHz. But they focused on the neuronal circuit design rather than complicate NN applications, containing only 2 hidden neurons which is very primitive.

Recently, Larras *et al* [86] proposed a mixed-signal IC for encoded binary neural network. The chip consists of analog neuron computing nodes for low power and cost design with digital network for communication among the nodes. The synapse and neurons are current mirrors that selects one leading data by winner-takes-all operation, and the resulting signal of each neuron is ANDed (Fig. 8). The neurons are clustered and connected via data bus for communication. The chip is fabricated in 64 nm CMOS technology and achieves 68 fJ ultra-low-energy operation per each synaptic operation. However, signals are input from external FPGA and both NN architecture and the target application, message retrieval, is too simple compared to other processors.

Although analog/mixed-mode design facilitates low-power and energy-efficient designs [88], it is often complicate due to PVT variation as well as the domain conversion overhead in terms of conversion speed, area, and power. High-end ADCs with fast sampling rate [94]–[96] consumes $\sim$10 mW under

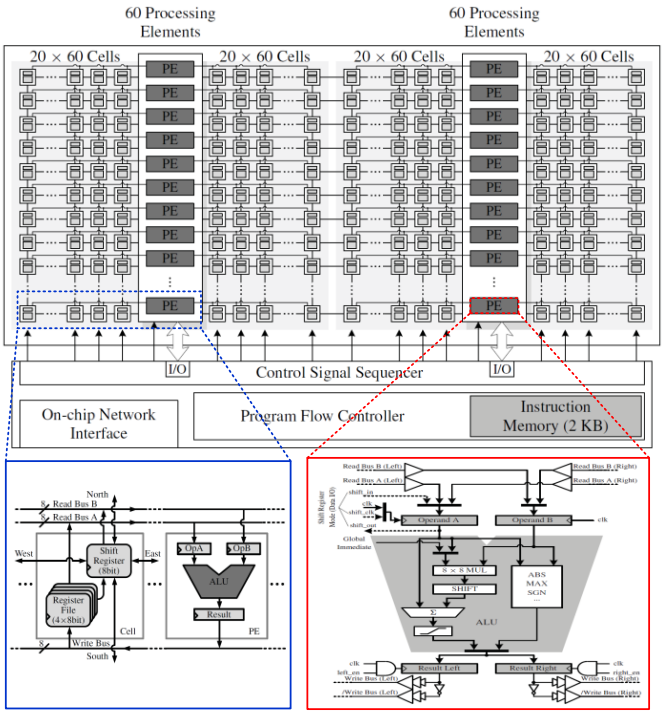


Fig. 9. Hardware architecture of the cellular neural network processor.

$\sim$10,000 μm^2 area. Since digital domain runs with $\sim$250 MHz, ADC could be the bottleneck in operating speed, thus, analog and digital domains must be carefully divided and balanced.

C. Digital Implementation

Although mixed-mode design has advantages in low power and small area implementation, digital processor has advantages in its speed and high precision as the technology gets smaller in addition to the removal of data conversion overhead in image processing.

A cellular neural network processor [62] is designed for visual attention in the brain-inspired object recognition pipeline [63], where visual attention is performed over input image of 80 × 60 pixels. The hardware architecture shown in Fig. 9 combines the flexibility of digital approach with high performance of fully parallel cell topology of analog approach. The 80 × 60 pixel values are organized into four 20 × 60 cell arrays and the 120 PEs are organized into two columns. Each cell consists of register files for data storage and a shift register for inter-cell communication. The PEs execute basic functions of cellular neural network as well as other functions for general purpose image processing. PE arrays are fully pipelined with three stages of read, execute, write; resulting in 42 cycles to execute each instruction on the whole cell array. The processor is implemented in 0.13 μm CMOS technology and occupies 4.5 mm^2. It consumes 84 mW running at 200 MHz. With the help of the processor, incorrect local features are drastically reduced to increase frame rate by 83% and reduce energy/frame by 45% without degradation in recognition accuracy.

The well-known Convolutional NN, CNN, is used for visual attention in the recent object recognition SoC for smart glass system [64]. In the visual attention system, CNN with two sets

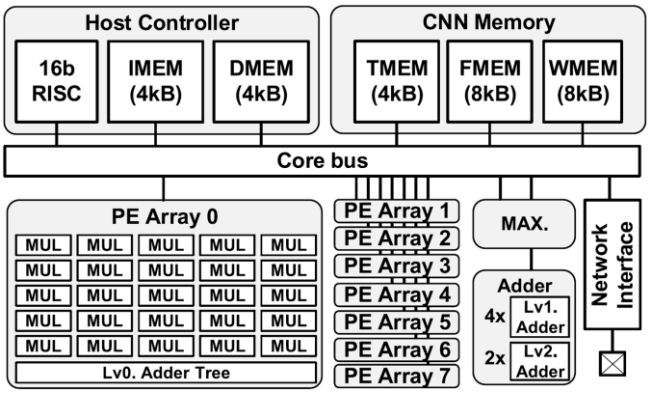


Fig. 10. Overall architecture of the embedded DNN processor.

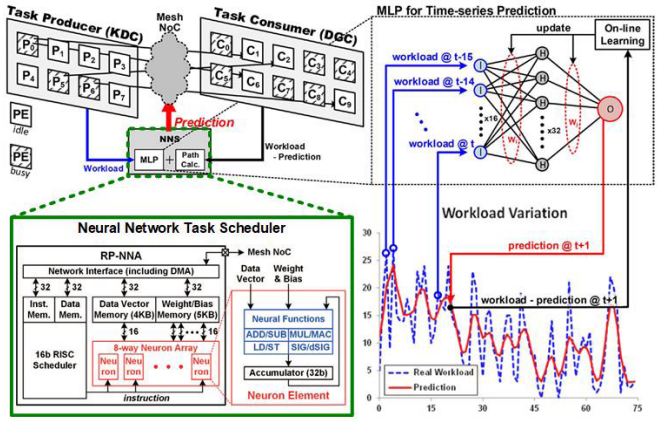


Fig. 11. System architecture of the MLP-based neural network task scheduler for workload prediction.

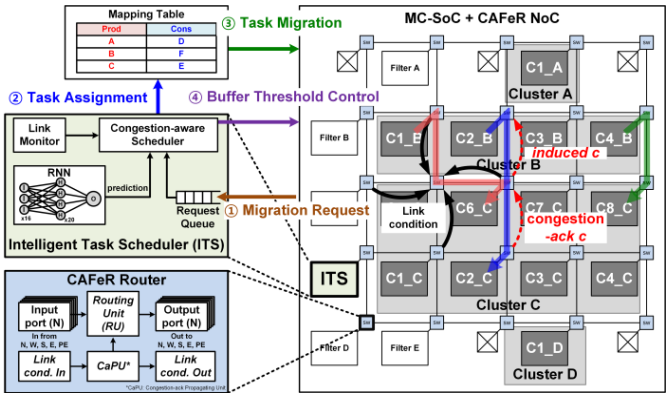


Fig. 12. Overall system architecture with RNN-based task scheduler and energy-efficient network-on-chip architecture for workload prediction.

of convolution and pooling layers extract features from input image tile and MLP at the end classifies ROIs. It reduces total number of ROIs by ∼80% and directly increases energy efficiency of the object recognition processor. The hardware architecture of [65] is shown in Fig. 10. It consists of a RISC controller, 8 PE arrays each of which contains 5×5 multipliers and an adder tree, a max pooling unit, a hierarchical adders, and memory banks. Each PE array is dedicated to acceleration of convolution operations. However, 2-D image convolution and 1-D FC convolution for MLP have different data access pattern. Thus, using 2-D PE arrays for MLP is not efficient. To resolve this issue, hierarchical adder trees are added for dual-mode configuration to accelerate both 2-D and 1-D convolutions. Fabricated in 65 *nm* CMOS technology, the CNN processor consumes 9 *mW* and occupies 1.55 mm^2, achieving 1.9 *nJ/pixel* energy efficiency. Experimental results showed that the CNN processor selects only 27% of image tiles, while degrading only ∼4% of object recognition accuracy.

In another aspect of applying NN as a functional block, neural network task scheduler [66] is designed for workload prediction to enhance energy efficiency by reducing time overhead on core-to-core allocation in a multi-core vision processor [67]. Many multi-core systems today adopt network-on-chip as their communication fabric for fast and parallel data transmission, but data transaction delay caused by network congestion directly degrades energy efficiency because processing cores must run faster to meet the overall system latency. Therefore, we designed a neural network task scheduler that estimates workloads in future frames with MLP and allocates producer-consumer pairs in advance to reduce network congestion as shown in Fig. 11. The neural network task scheduler contains a RISC scheduler and an 8-way neuron array that is capable of reconfigurable precision. The processor is fabricated in 65 *nm* CMOS technology. It consumes 4.9 *mW* while achieving 12.7 *mJ/frame* energy efficiency by reducing 24.4% of network latency on average. To enhance the performance, an advanced version of neural network scheduler using RNN (Fig. 12) with new network-on-chip architecture that is dedicated to SIFT-based object recognition [68], is proposed in [69]. It is also designed as a part of the object recognition SoC [64]. It improved the workload prediction accuracy to 91.4% and system throughput by 50.2%.

IV. Deep Learning / Deep Neural Network System-on-Chips

The processors explained in Section III imply that utilizing NN/NF algorithms as a functional block of SoC brings great improvements if they are used in right place with dedicated hardware architecture. Now, recent SoC implementations for accelerating the entire DL/DNN algorithm will be explored.

TPU by Google [44] targeted for general DNN acceleration, it was not suitable for mobile applications due to its huge power consumption ($>$ 40 *W*) and memory footprint ($>$ 28 *M-Byte*). Then, the DNN processing unit (DNPU) [70] was introduced to accelerate both CNN and RNN in mobile environment with following features to optimize both CNN and RNN on a single chip: 1) heterogeneous multi-core architecture; 2) mixed channel division scheme for CNN computation; and 3) quantization-table (Q-table) for efficient matrix product. Because the most dominant operation in CNN is 2-D convolution, both temporal and spatial locality of operands are very high. It is interpreted as the operational intensity of CNN becomes high and it characterizes CNN "computation bounded". On the other hand, 2-D matrix multiplication dominates in RNN operations and non-reusable weights makes it "memory bounded". Therefore, throughput of RNN is limited by external memory accesses for operands rather than operation itself. The heterogeneous multi-core architecture of DNPU (Fig. 13) aims at fulfilling the two different peculiarities. The

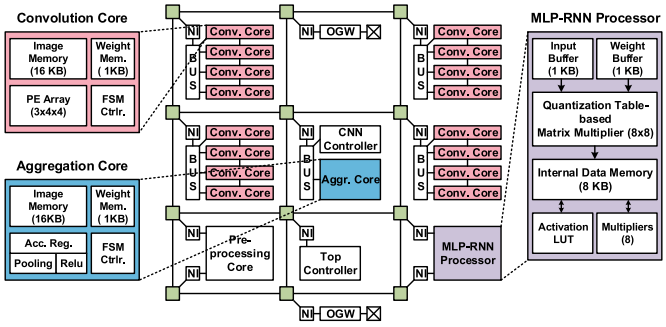


Fig. 13. Overall architecture of the UNPU with homogeneous DNN cores.

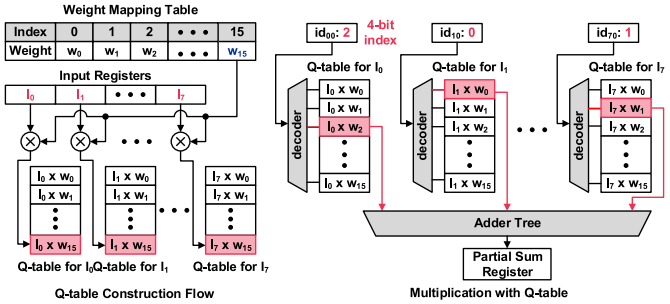


Fig. 14. Bit-serial PE architecture supporting variable weight precision.

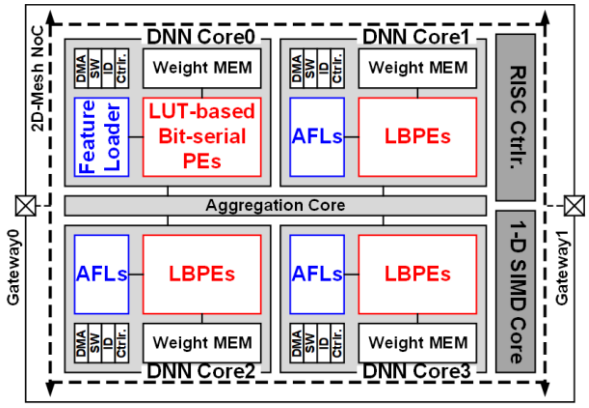


Fig. 15. Overall architecture with distributed memory and chip-to-chip connections of CNNP.

CNN core is dedicated to maximize data reusability with high computational performance over 300 GOPS, and the RNN core was designed to minimize the amount of external memory access with algorithmic support and dedicated PEs. It proposed new computation methodologies to CNN acceleration of which the required memory is too big to fit into the mobile DL accelerator with limited memory capacity. According to [70], the image division method is to divide feature maps into multiple segments along with the direction of image. It is useful when the entire image map cannot fit on a chip. The channel division is to divide feature maps by *channels*. It is useful when the whole weights cannot be stored within the hardware or the channel depth is very large. DNPU utilizes mixed division computation that combines image and channel division methods to enable more efficient processing of lager CNN with limited on-chip memory. For RNN acceleration, the RNN core exploits the redundancy in weight matrix of RNN/FCL [71]. Usually, weight quantization of RNN/FCL involves with tiny accuracy loss but significantly reducing external memory accesses by fetching just weight indices. In addition, partial-sums become also quantized when input activation is multiplied with quantized weights. By using this characteristic, the DNPU equips Q-table based matrix multiplication as described in Fig. 14. The values of Q-table are updated using all of possible partial-products for given input activations. Then, matrix multiplication is performed by accessing table values without using multipliers, reducing the number of multiplication of RNN/FCL by 99%. The SoC is fabricated in 65 *nm* CMOS technology and achieves 8.1 *TOPS/W* at 50 *MHz,* 0.77 *V* where peak power is 279 *mW*.

However, heterogeneous architecture of DNPU limited hardware utilization. For example, if a DL application requires only one of CNN or RNN, the unused part of the DNPU wastes hardware resources. The unified neural processing unit (UNPU) was presented [72] to resolve the limitations of the DNPU by obtaining high hardware utilization on energy-optimal point. It was designed with new microarchitecture with the following features: 1) unified DNN core architecture; and 2) fully-variable weight bit precision. As described in [72], feature map becomes dominant when the weight bit precision is reduced on the accuracy-energy optimal point, so the UNPU shown in Fig. 15 utilizes feature map reuse datapath while DNPU dealt with weight reuse. This unified data-path enabled to combine the separate CNN and RNN cores of the DNPU into a single DNN core, which achieved 1.15× and 13.8× higher peak-performance with the same hardware footprint. In addition, with the help of configurable PEs in DNN core, the UNPU was able to support versatile workload combinations of CNN and RNN. Moreover, the multi-bit precisions of 4-bit, 8-bit, and 16-bit supported by the DNPU were not enough to accelerate DNNs on the optimal bit precision that differs according to different layers and different networks [73], [74]. Many DNN applications require finer bit precision along with the desired accuracy-speed requirement [75]– [77]. The UNPU was designed with dedicated bit-serial PEs that enable fully-variable weight bit-precision as depicted in Fig. 16. For the given feature vector, weights are put into the PEs bit-serially, then multiple PEs make bit-serial partial-products in every cycle. This scheme facilitate the UNPU to support every weight bit precision from 1-bit to 16-bit in trade of sacrificing the latency to support fully-variable bit precision. To relieve the area/power overhead that bit-serial PE involves compared with fixed-point MAC PE, the UNPU proposed LUT to store possible bit-serial partial-products in the PEs. It reduced energy consumption of MAC operation by 23.1% (16-bit), 27.2% (8-bit), 41.0% (4-bit), and 53.6% (1-bit) than the fixed-point MAC PEs.

Contrary to abovementioned ASICs for general purpose DNN acceleration, the CNN processor (CNNP) introduced in [78] targeted user identification application on always-on sensor node, which essentially requires highly accurate face recognition with low-power consumption. To satisfy the requirement, the CNNP proposed two main features for low

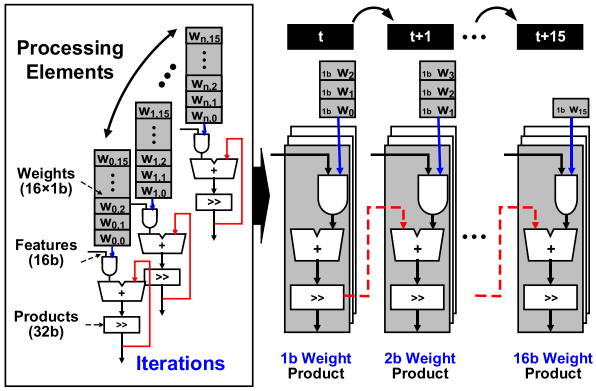


Fig. 16. Separable filter approximation to 1-D convolution.

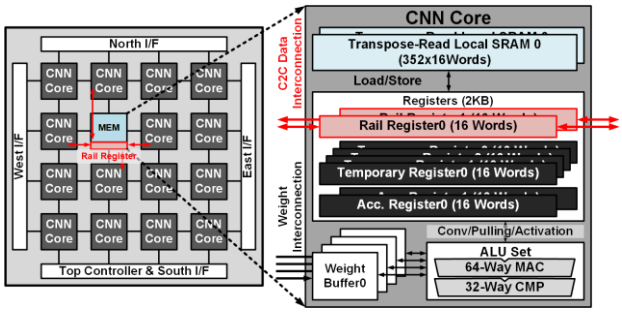


Fig. 17. Overall architecture of the DNPU; It has multi-core processor architecture with distributed memory to optimize both CNN and RNN.

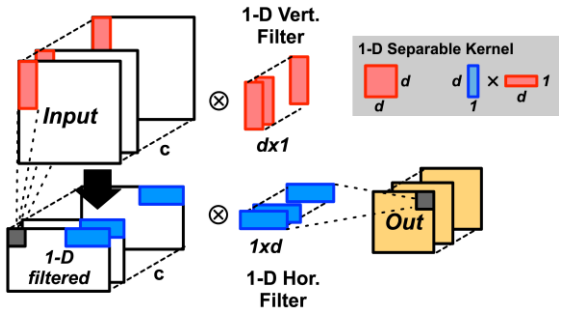


Fig. 18. Quantization-table (Q-table) in RNN core of the DNPU to reduce the amount of external memory access for weights.

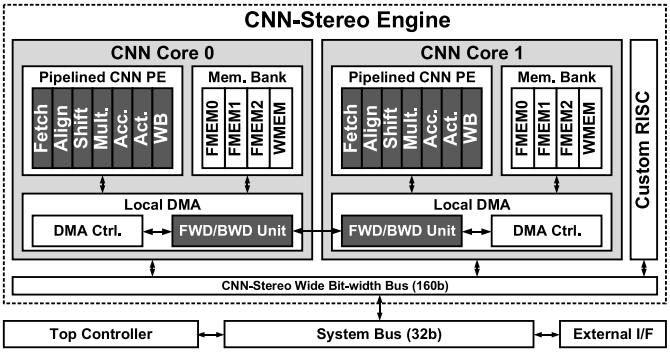


Fig. 19. Overall architecture of CNN-Stereo engine.

power operation: 1) distributed-memory CNN architecture; and 2) separable filter approximation. As shown in Fig. 17, it utilized distributed memory architecttture without separate global routing and memory to lower routing complexity, and brings the external memory into multiple local on-chip memories for lower power CNN operation. To reduce additional power consumption, the CNNP utilized separable filter approximation technique that estimates the 2-D convolution with two cascaded 1-D vertical and horizontal convolutions as depicted in Fig. 18. This approximation reduces the amount of required computation by up to 3× with less than 1% accuracy loss when applied to the real face recognition. The chip embedded full custom transpose-read SRAM for efficient memory access for separable filter approximation, reducing 78% overall energy. The CNNP is fabricated in 65 *nm* CMOS technology and achieved 5.3 *mW*.

However, the CNNP suffers from poor hardware utilization when the size of input feature map is small or when the image size assigned to the core is not multiples of the number of words that is computed at once in the core. It is not sufficient for stereo matching CNN algorithm without pooling layers [81] of which the image size gets smaller in deeper layers, leading to degradation in hardware utilization. Therefore, the CNN-Stereo Engine (CSE) [80] is published to maximize the hardware utilization with two main features: 1) channel-wise parallel 1-D MAC operation; and 2) core-to-core data balancing. The CSE increased parallelism in the channel direction rather than spatial direction. And it was designed to handle 1-D operations to remove the redundancy caused by the 2-D workload assignment method of the CNNP. Each core of the CSE in Fig. 19 performs parallel computation with multiple weight channels for 1-D input feature map. The overall hardware utilization is improved by reducing the number of words computed at once. The CSE hides both the accumulation latency among multiple channels and data fetch latency by pipelining. It achieved 20% higher energy efficiency than the CNNP. Moreover, the FWD/BWD units support exchanging adjacent boundaries of two local feature maps to guarantee the balanced workload allocation (Fig. 20). This operation was hidden behind CNN pipeline and improved the overall performance by 23.9%.

V. Neuromorphic Processors

Neuromorphic processors that aim at ultra-low-power consumption by mimicking neurons are investigated along with the advance of DNN accelerators. The representative processors would be Neurogrid [84], IBM's TrueNorth [99] and Intel's Loihi [100]. Neurogrid is a mixed-mode multichip system of which neuron array is designed in analog while spike TRX and memory are designed in digital with 0.18 μm CMOS. TrueNorth implemented 4,096 neurosynaptic cores of SNN that are connected via scalable routing network, as shown in Fig. 21. Each core consists of 256 digital axons and neurons which are connected by 256×256 crossbar for dendrite design. The large-scale chip (4.3 cm^2) fabricated in 28nm CMOS was applied to multiobject detection with 63 *mW* with 400×240.30 *fps* video. Loihi also is a fully digital architecture implementing 128 cores of 1,024 SNN units and supports sparsity as its performance is evaluated with LASSO

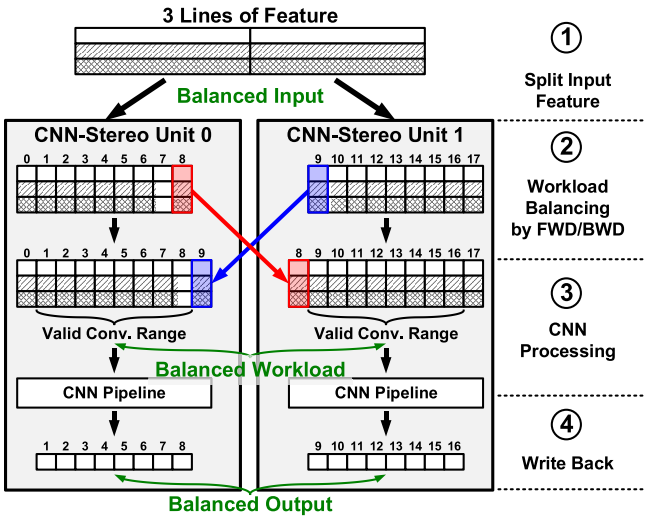


Fig. 20. Core-to-core workload balancing of CSE.

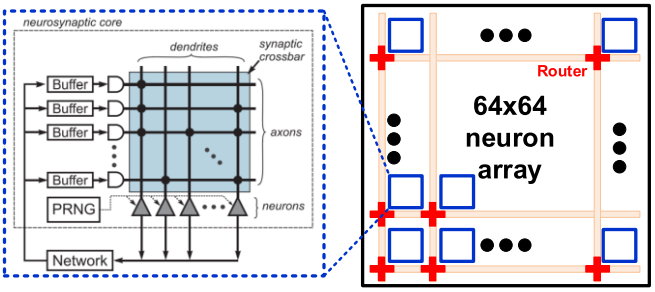


Fig. 21. Overall architecture with neurosynaptic core of TrueNorth [99].

optimization. The chip has a high level of completion that it also contained three x86 cores for message control for the neuromorphic cores. The whole SoC is implemented in 14 *nm* FinFET technology and occupies 60 mm^2. Although their die area were very large, both TrueNorth and Loihi are energy efficiency consuming only ~10 *mW*.

Some recent neuromorphic processors take advantage of analog computing arrays. Miyashita *et al* designed time domain neural network processor in part of Binarized NN with analog MAC circuits in CMOS technology [101]. They employed differential signaling to prevent precision degradation from PVT variation. The chip fabricated in 65 *nm* CMOS technology achieved high energy efficiency of 12.9 *fJ* per synaptic operation. Some neuromorphic processors make use of nonvolatile memory (NVM) such as ReRAM and MRAM in their analog arrays to replace floating-point MAC with parallel analog processing array [83], [85], [102]. The conceptual diagram of analog computing is depicted in Fig. 22. Transconductance of NVM-based crossbar represents corresponding weight. Current flowing through a column becomes summation of product of input vector and weight connected to the column. Bayat *et al* designed MLP processor with ReRAM crossbar arrays for classification. It was capable of limited MLP application with only 10 hidden neurons because number of neurons must fit to the array size. The analog accelerators show significant performance improvement

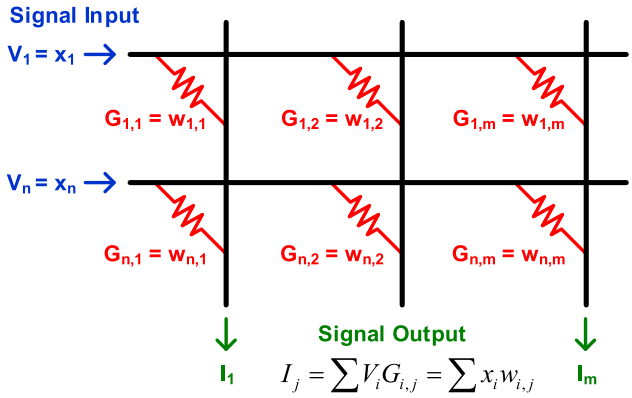


Fig. 22. Conceptual diagram of analog vector MAC array.

than digital processors. According to [102], analog processor achieved 270× energy, 1040× latency, and 1.8× area than digital ReRAM accelerators; 430×, 34×, and 11× compared to SRAM-based digital accelerators, respectively. However, it is important to ensure that implementation details and material properties of the NVM must be aligned with the requirements of NN algorithms [85]. These processors still have problems of their limited functionality. Most of the works were capable of only simple classification tasks with MNIST or CIFAR-10 dataset while DNN/DL SoCs provide higher-level intelligence with much complex dataset.

VI. Discussion and Future Direction

Various circuit design approaches toward NN/NF/DL were explained in this paper. Table I shows the summary table of the mixed-mode NN/NF processors, each of which was the state-of -the-art designs by the time they were proposed [49], [52], [55], [57], [61]. They employed current-mode mixed-mode circuits for area and power reduction compared with the fully-digital implementations. However, the amount of reduction becomes less significant as the technology node goes down to *nm*. The average power reduction using 0.13 μm technology was 70%, while that of 65 *nm* technology was only 40%. This is because the analog circuit cannot be scaled down with the same ratio of technology scaling. Length of MOSFETs must be larger than some degree to reach bias point for analog circuits to operate, and this disturbs circuit designers to have very small circuits in *nm*-scale. But this does not mean that mixed-mode circuits no longer have advantages over digital circuits. It still can take the advantage of having smaller but flexible digital controller as well as utilizing natural parallelism of circuit operations. Also, the analog cores can be reconfigurable with some design techniques to provide versatile functions even though it can be less accurate than digital computations. The analog PE cluster in [55] generates four different types of NN topologies by changing the signal paths. The highly controllable radial-basis-function circuit [57] generates almost 5 different shape of functions including sigmoid function, thus it can also be used for MLP. By having shared datapath to ADC with the digital learning controller, the RNN-FIS processor [61] can handle four different NN/NF topologies. This shows that

TABLE I
SUMMARY TABLE OF MIXED-MODE NN/NF CIRCUITS

Ref.	Type	Application	Reconfigurability	Supported Function	Process	Power [mW]	Area [mm^2]	Power Reduction	Area Reduction
[49]	NF	Object Detection	Low	NF	0.13 μm	2.83	0.163	44.0%	59.0%
[52]	NN	ROI / Workload Prediction	High	NF	0.13 μm	1.20	0.765	85.0%	46.0%
[55]	NN/NF	Object Recognition	Very High	MLP/RNN/RBFNN/NF	0.13 μm	75.0	13.5	71.2%	43.0%
[57]	NN	Classification	Medium	RBFNN/MLP	0.13 μm	2.20	0.140	82.3%	84.0%
[61]	NF	Deep Risk Prediction	High	MLP/RNN/FIS/NF	65 nm	N/A	N/A	39.1%	64.1%
[87]	NN	XOR Operation	None	XOR NN	0.18 μm	0.538	0.0285	N/A	N/A
[86]	NN	Message Decoding	High (FGPA)	Encoded Binary NN	65 nm	0.145	0.0194	N/A	N/A

* Power and Area reductions are the case compared with their equivalent digital implementation using the same process

TABLE II
SUMMARY TABLE OF DIGITAL NN/NF PROCESSORS

Ref.	Type	Application	Process	Power [mW]	Area [mm^2]	Supply Voltage [V]	Frequency [MHz]
[62]	Cellular NN	Visual Attention	0.13 μm	84.0	0.163	1.2	200
[64]	CNN	Visual Attention	65 nm	9.00	0.765	1.2	200
[66]	MLP	Workload Prediction	65 nm	4.90	13.5	1.2	250
[69]	RNN	Workload Prediction	65 nm	4.16	0.140	1.2	200

TABLE III
SUMMARY TABLE OF DNN SYSTEM-ON-CHIPS

Ref.	Type	Application	Process	Peak Power [mW]	Area [mm^2]	Supply Voltage [V]	Frequency [MHz]
[70]	RNN/CNN	General Purpose	65 nm	279	16.0	0.77 ~ 1.0	50 ~ 200
[72]	RNN/CNN	General Purpose	65 nm	297	16.0	0.63 ~ 1.1	5 ~ 200
[78]	CNN	Always-on Face Recognition	65 nm	211	16.0	0.46 ~ 0.8	5 ~ 100
[80]	CNN-Stereo	Depth Estimation	65 nm	21.3	~ 8	0.7 ~ 1.2	10 ~ 100
[35]	CNN	Image Classification	65 nm	278 (conv. only)	16.0	0.82 ~ 1.17	100 ~ 250
[40]	FCL	Always-on DNN	40 nm	N/A	7.10	0.63 ~ 0.9	1.9 ~ 19.3
[48]	CNN/FCL/RNN	General Purpose	40 nm	2,083	122	0.77 ~ 1.1	75 ~ 330

analog circuits can also be programmable assisted by digital controller.

However, analog and digital domains must be carefully divided in mixed-mode design to reduce the domain conversion overhead cost in terms of conversion speed, area, and power. The overhead can be resolved when the processing comes into sensors and computes with analog data before A-D conversion. In [103], pupil edge detection and glint corner detection circuits were deployed into CMOS image sensor, each of which circuit configures pixel array to extract edge in analog voltage and compares charge difference, respectively. Matrix multiplication was integrated into ADC that multiplies analog input with digital signal to reduce additional energy required for A-D conversion in [104]. Kim *et al* [105] designed an image sensor dedicated to stereo matching. Its reconfigurable pixel array computes sparse image rectification to align two different inputs and census transformation is executed in analog domain with simple switch network and comparators. Since those values are computed with analog signal before ADC the digital processor integrated in the sensor excluded additional blocks for rectification and census transformation. All of the works achieved high energy efficiency by putting computations into the sensors, and showed new design paradigm.

Table II shows the summary of digital NN/NF processors as a functional block of the machine vision systems. They reduced the amount of overall computation with visual attention and improved system throughput with workload prediction. It is noticeable that taking NN/NF functions for specific purpose reduces energy efficiency of the whole machine vision SoC, significantly. This implies that the processor does not always have to be dedicated for accelerating end-to-end DNN algorithm but applying small NN/NF system to right places with dedicated hardware architecture for that specific purpose brings performance enhancement.

Although the majority of computing engine for general-purpose DNN acceleration are traditional CPU and GPU, they consume too much power. For example, AlexNet running on Intel Xeon CPU and NVIDIA Titan X consumes 130 *W* and 250 *W* [97]. Even one FC layer of AlexNet consumes lots of power, e.g. 73 *W* with Intel Core-i7 CPU, 159 *W* with Titan X GPU, and 5.1 *W* with Tegra K1 mobile GPU [98]. In contrast, ASIC designs reduce power consumption to the order of *mW*, which makes ASICs suitable for mobile and embedded systems. Table III shows the summary of the DNN SoCs. AlexNet running on [72] consumes only 290 *mW* at 346 *fps* maximum while [48] consumed 2.7 *W* with 76.7 % accuracy.

There are several design challenges in DNN SoC. Although they are designed for low power consumption memory bandwidth becomes crucial problem. The processors [38], [43], [47], [79] tried to deploy data locality for maximizing on-chip data reuse in order to achieve low power yet high throughput. Moons *et al* [38] and Chen *et al* [79] brought with systolic array to reuse data within PEs as many as possible. Shin *et al* [43] and Lee *et al* [47] utilized 2-D MAC array with dedicated logics for data management and scheduler to maximize data reusability. They are on the basis of global buffer to filter the amount of external memory bandwidth that involves with global routing and aggregation path between global buffer and parallel PEs. This approach requires a large memory bandwidth and complex global routing to high parallelism, which is not suitable for low-power consumption. So [78] utilized distributed memory architecture without global routing and memory to lower routing complexity and resolve memory bandwidth bottleneck. Such distributed memory architecture extends to Processing- in-Memory (PIM) architecture (the memory-centric architectures such as in-memory/near-memory processing will be included in PIM in this paper). Kang *et al* [106] employed deep in-memory architecture that computes in the periphery of the memory cell array. This technique minimizes the costs of data access and processing similar to the concept of mixed computing in sensors. Data stored in SRAM bitcell are computed and selected before passing through ADC. The processor achieved 4.9× energy saving and 2.4× throughput improvement compared to a conventional von-Neumann architecture. In [107], results of census transform are stored in stereo-SRAM and they are XORed in the peripheral between bitcell arrays to generate costs before propagated to digital block. Yang *et al* inserted pulse width modulation and shift memory into the bitcell array to efficiently accelerate binary-weight network [108]. Biswas *et al* utilized 1-b multiplication-and-average operation in analog domain right next to the 10T SRAM cell array to accelerate binary-weight network [109] and Yin *et al* proposed SRAM macro that computes ternary-XNOR-and-accumulate operation. These approaches achieve significant improvement in energy efficiency. But the architecture is very primitive that they could only support small neural network with limited dataset. External memory bandwidth is another bottleneck as on-chip data reuse technique advances. 3-D stacking technique with DNN accelerators as reported in [48] will bring immense improvement in external bandwidth overhead.

Unlike CPU and GPU, the DNN SoCs introduced so far are dedicated to specific system and application which induces dependency on DNN architecture and accuracy. For example, target application of [78] was always-on face recognition for user identification where the number of output class is less than ImageNet dataset, or the task is simpler. This fact implies that small CNN architecture is enough to achieve 97.4% of accuracy while only 0.2% degradation is not crucial for the application. Also, the SoCs utilized reduced bit precision to resolve computational complexity that directly harms the accuracy even a small amount. GPU and TPU are preferable than the ASICs introduced in this paper if programmability of DNN algorithm is versatile or the target application requires precise computation for high accuracy such as autonomous driving.

Thanks to the technology scaling, we foresee the future researches on hardware accelerators for high performance DNN applications will be digital to maintain high throughput and precise computation and it will bring integration of higher-level intelligence with complicate DNN architectures with energy-efficient digital implementations. On the other hand, neuromorphic processors will be investigated for ultra-low-power applications with simple intelligence by utilizing energy-efficient analog computations. However, many neuromorphic processors highly rely on the development of new technologies such as ReRAM or MRAM. These technologies are still being researched but its time-to-market until mass production is unpredictable. In the meantime, it is believed that mixed-mode implementations and PIM architectures of neural network processor that uses current CMOS technology is promising for ultra-low-power DNN applications until the neuromorphic devices become in mass production.

VII. Conclusion

This paper provides a review of dedicated processors for neural network, neuro-fuzzy system, deep learning & deep neural network processors in various design approaches of analog, digital, and mixed-mode implementation over decades. We provided both mixed-mode and fully-digital implementations of neural network / neuro-fuzzy processors that are deployed as a functional building block of machine vision systems. We also provided the comparisons upon the different design approaches. Finally, the most recent designs of deep learning processors were introduced. Neuromorphic chips are highly dependent on the development of new technologies but no one can assure how long it will take for the technologies to be in mass production. Before then, it is believed that mixed-mode implementation of neural network processor is promising for ultra-low-power applications while digital implementation will fit for high-performance applications.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[4] S. Xie *et al.*, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2017, pp. 5987–5995.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[6] Z. Chen *et al.*, "End-to-end learning for lane keeping of self-driving cars," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2017, pp. 1856–1860.

[7] K. Ota, M. S. Dao, V. Mezaris, and F. G. B. De Natale "Deep learning for mobile multimedia: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 3, pp. 1–22, Aug. 2017.

[8] Y. Shuochao *et al.*, "DeepSense: A unified deep learning framework for time-series mobile sensing data processing," in *Proc. Int. Conf. World Wide Web*, Jun. 2017, pp. 351–360.

[9] D. Li, X. Wang, and D. Kong, "DeepRebirth: Accelerating deep neural network execution on mobile devices," 2017, *arXiv:1708.04728*. [Online]. Available: http://arxiv.org/abs/1708.04728

[10] A. Mathur *et al.*, "Deepeye: Resource efficient local execution of multiple deep vision models using wearable commodity hardware," in *Proc. Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2017, pp. 68–81.

[11] Q. Cao *et al.*, "MobiRNN: Efficient recurrent neural network execution on mobile GPU," in *Proc. Int. Workshop Deep Learn. Mobile Syst. Appl. (EMDL)*, Jun. 2017, pp. 1–6.

[12] N. D. Lane *et al.*, "DeepX: A software accelerator for low-power deep learning inference on mobile devices," in *Proc. ACM/IEEE Int. Conf. Inf. Process. Sensor Netw.*, Apr. 2016, pp. 1–12.

[13] S. I. Venieris and C.-S. Bouganis, "FpgaConvNet: A toolflow for mapping diverse convolutional neural networks on embedded FPGAs," 2017, *arXiv:1711.08740*. [Online]. Available: http://arxiv.org/abs/1711.08740

[14] J. Wu *et al.*, "Quantized convolutional neural networks for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4820–4828.

[15] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," 2016, *arXiv:1611.06440*. [Online]. Available: http://arxiv.org/abs/1611.06440

[16] A. Tulloch and Y. Jia, "High performance ultra-low-precision convolutions on mobile devices," 2017, *arXiv:1712.02427*. [Online]. Available: http://arxiv.org/abs/1712.02427

[17] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," 2015, *arXiv:1511.06530*. [Online]. Available: http://arxiv.org/abs/1511.06530

[18] P. Ienne, T. Cornu, and G. Kuhn, "Special-purpose digital hardware for neural networks: An architectural survey," *J. VLSI Signal Process. Syst.*, vol. 13, no. 1, pp. 5–25, Aug. 1996.

[19] F. Yang and M. Paindavoine, "Implementation of an RBF neural network on embedded systems: Real-time face tracking and identity verification," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1162–1175, Sep. 2003.

[20] J. Lont and W. Guggenbuhl, "Analog CMOS implementation of a multilayer perceptron with nonlinear synapses," *IEEE Trans. Neural Netw.*, vol. 3, no. 3, pp. 457–465, May 1992.

[21] S. Peng, P. Hasler, and D. Anderson, "An analog programmable multi dimensional radial basis function based classifier," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 10, pp. 2148–2158, Oct. 2007.

[22] J. Lu, S. Young, I. Arel, and J. Holleman, "A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 μm CMOS," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 270–281, Jan. 2015.

[23] J. Zhang, Z. Wang, and N. Verma, "A matrix-multiplying ADC implementing a machine-learning classifier directly with data conversion," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 332–333.

[24] J. K. Kim, P. Knag, T. Chen, and Z. Zhang, "A 6.67mW sparse coding ASIC enabling on-chip learning and inference," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Honolulu, HI, USA, 2014, pp. 1–2.

[25] J. K. Kim, P. Knag, T. Chen, and Z. Zhang, "A 640M pixel/s 3.65mW sparse event-driven neuromorphic object recognition processor with on-chip learning," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Kyoto, Japan, 2015, pp. 50–51.

[26] E. H. Lee and S. S. Wong, "24.2 A 2.5GHz 7.7TOPS/W switched-capacitor matrix multiplier with co-designed local memory in 40nm," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 418–419.

[27] S. Ambrogio *et al.*, "Novel RRAM-enabled 1T1R synapse capable of low-power STDP via burst-mode communication and real-time unsupervised machine learning," in *Proc. IEEE Symp. VLSI Technol. (VLSIT)*, Jun. 2016, pp. 1–2.

[28] J. Zhang, Z. Wang, and N. Verma, "A machine-learning classifier implemented in a standard 6T SRAM array," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Jun. 2016, pp. 1–2.

[29] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6T SRAM array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[30] C.-H. Tsai *et al.*, "A 7.11mJ/Gb/Query data-driven machine learning processor (D2MLP) for big data analysis and applications," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2014, pp. 1–2.

[31] S. Park *et al.*, "An energy-efficient and scalable deep learning/inference processor with tetra-parallel MIMD architecture for big data applications," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 6, pp. 838–848, Dec. 2015.

[32] S. Park, K. Bong, D. Shin, J. Lee, S. Choi, and H.-J. Yoo, "A1.93TOPS/W scalable deep learning/inference processor with tetra-parallel MIMD architecture for big-data applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.

[33] S. Park, S. Choi, J. Lee, M. Kim, J. Park, and H. Yoo, "A 126.1mW real-time natural UI/UX processor with embedded deep-learning core for low-power smart glasses," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 254–255.

[34] K. J. Lee *et al.*, "A 502GOPS and 0.984mW dual-mode ADAS SoC with RNN-FIS engine for intention prediction in automotive black-box system," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 256–257.

[35] Y. Chen, T. Krishna, J. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 262–263.

[36] J. Sim, J. Park, M. Kim, D. Bae, Y. Choi, and L. Kim, "A 1.42TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2016, pp. 264–265.

[37] B. Moons and M. Verhelst, "A 0.3-2.6 TOPS/W precision-scalable processor for real-time large-scale ConvNets," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jun. 2016, pp. 1–2.

[38] B. Moons *et al.*, "14.5 Envision: A 0.26-to-10 TOPS/W subword-parallel dynamicvoltage-accuracy-frequency-scalable convolutional neural network processor in 28 nm FDSOI," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 246–257.

[39] P. Knag, C. Liu, and Z. Zhang, "A 1.40mm2141mW 898GOPS sparse neuromorphic processor in 40nm CMOS," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Jun. 2016, pp. 1–2.

[40] S. Bang *et al.*, "A 288 W programmable deep-learning processor with 270KB on-chip weight storage using non-uniform memory hierarchy for mobile intelligence," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 250–251.

[41] P. N. Whatmough *et al.*, "A 28nm SoC with a 1.2GHz 568nJ/prediction sparse deep neural network engine with >0.1 timing error rate tolerance for IoT applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 242–243.

[42] G. Desoli *et al.*, "A 2.9 TOPS/W deep convolutional neural network SoC in FD-SOI 28 nm for intelligent embedded systems," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 238–239.

[43] D. Shin, J. Lee, J. Lee, and H.-J. Yoo, "DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2017, pp. 240–241.

[44] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," *ACM SIGARCH Comput. Archit. News (ISCA)*, vol. 45, no. 2, pp. 1–12, 2017.

[45] S. Yin *et al.*, "A 1.06-to-5.09 TOPS/W reconfigurable hybrid-neural network processor for deep learning applications," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Jun. 2017, pp. 26–27.

[46] K. Ando *et al.*, "BRein memory: A 13-layer 4.2 K neuron/0.8 M synapse binary/ternary reconfigurable in-memory deep neural network accelerator in 65 nm CMOS," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Jun. 2017, pp. 24–25.

[47] J. Lee *et al.*, "UNPU: A 50.6 TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 218–219.

[48] K. Ueyoshi *et al.*, "QUEST: A 7.49 TOPS multi-purpose log-quantized DNN inference engine stacked on 96MB 3D SRAM using inductive-coupling technology in 40nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 216–217.

[49] M. Kim *et al.*, "A 22.8GOPS 2.83mW neuro-fuzzy object detection engine for fast multi-object recognition," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Jun. 2009, pp. 260–261.

[50] J.-Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H.-J. Yoo, "A 201.4 GOPS 496 mW real-time multi-object recognition processor with bio-inspired neural perception engine," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 32–45, Jan. 2010.

[51] D. O. Hebb, *The Organization of Behavior*. New York, NY, USA: Wiley, 1949.

[52] J. Oh, S. Lee, and H. J. Yoo, "1.2-mW online learning mixed-mode intelligent inference engine for low-power real-time object recognition processor," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 21, no. 5, pp. 921–933, May 2013.

[53] S. Lee, J. Oh, J. Park, J. Kwon, M. Kim, and H.-J. Yoo, "A 345 mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 42–51, Jan. 2011.

[54] F. Vidal-Verdu *et al.*, "A design approach for analog neuro/fuzzy systems in CMOS digital technologies," in *Computers and Electrical Engineering*, vol. 25. Amsterdam, The Netherlands: Elsevier, 1999, pp. 309–337.

[55] J. Oh, J. Park, G. Kim, S. Lee, and H.-J. Yoo, "A 57mW embedded mixed-mode neuro-fuzzy accelerator for intelligent multi-core processor," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 130–131.

[56] J. Mark and L. Orr, "Introduction to radial basis function networks," Tech. Rep., 1996.

[57] Y. Ou and Y. Oyang, "A novel radial basis function network classifier with centers set by hierarchical clustering," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2005, pp. 1383–1388.

[58] K. Lee, J. Park, G. Kim, I. Hong, and H.-J. Yoo, "A multi-modal and tunable radial-basis-function circuit with supply and temperature compensation," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 1608–1611.

[59] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, Nov. 1999.

[60] J. Park *et al.*, "A 646GOPS/W multi-classifier many-core processor with cortex-like architecture for super-resolution recognition," *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2013, pp. 168–169.

[61] K. J. Lee *et al.*, "A 502-GOPS and 0.984-mW dual-mode intelligent ADAS SoC with real-time semiglobal matching and intention prediction for smart automotive black box system," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 139–150, Jan. 2017.

[62] S. Lee *et al.*, "24-GOPS 4.5-mm^2 digital cellular neural network for rapid visual attention in an object-recognition SoC," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 64–73, Jan. 2011.

[63] K. Kim *et al.*, "A 125 GOPS 583 mW network-on-chip based parallel processor with bio-inspired visual attention engine," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 136–147, Jan. 2009.

[64] I. Hong *et al.*, "A 2.71 nJ/pixel gaze-activated object recognition system for low-power mobile smart glasses," *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 45–55, Jan. 2016.

[65] S. Park, I. Hong, J. Park, and H.-J. Yoo, "An energy-efficient embedded deep neural network processor for high speed visual attention in mobile vision recognition SoC," *IEEE J. Solid-State Circuits*, vol. 51, no. 10, pp. 2380–2388, Oct. 2016.

[66] Y. Kim, G. Kim, I. Hong, D. Kim, and H.-J. Yoo, "A 4.9 mW neural network task scheduler for congestion-minimized network-on-chip in multi-core systems," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2014, pp. 213–216.

[67] G. Kim *et al.*, "A 1.22 TOPS and 1.52 mW/MHz augmented reality multicore processor with neural network NoC for HMD applications," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 113–124, Jan. 2015.

[68] D. G. Lowe, "Object recognition from local scale-Invariant features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 1999, pp. 1150–1157.

[69] K. Lee, J. Park, I. Hong, and H.-J. Yoo, "Intelligent task scheduler with high throughput NoC for real-time mobile object recognition SoC," in *Proc. IEEE Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2015, pp. 100–103.

[70] D. Shin, J. Lee, J. Lee, J. Lee, and H.-J. Yoo, "DNPU: An energy-efficient deep-learning processor with heterogeneous multi-core architecture," *IEEE Micro*, vol. 38, no. 5, pp. 85–93, Sep. 2018.

[71] J. Lee, D. Shin, and H. Yoo, "A 21mW low-power recurrent neural network accelerator with quantization tables for embedded deep learning applications," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Nov. 2017, pp. 237–240.

[72] J. Lee *et al.*, "UNPU: An energy-efficient deep neural network accelerator with fully variable weight bit precision," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 173–185, Jan. 2019.

[73] L. Lai, N. Suda, and V. Chandra, "Deep convolutional neural network inference with floating-point weights and fixed-point activations," 2017, *arXiv:1703.03073*. [Online]. Available: http://arxiv.org/abs/1703.03073

[74] P. Judd *et al.*, "Stripes: Bit-serial deep neural network computing," in *Proc. Annu. IEEE/ACM Int. Symp. Microarchitecture*, 2016, pp. 1–12.

[75] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," 2016, *arXiv:1612.01064*. [Online]. Available: http://arxiv.org/abs/1612.01064

[76] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-net: Imagenet classification using binary convolutional neural networks," 2016, *arXiv:1603.05279*. [Online]. Available: http://arxiv.org/abs/1603.05279

[77] Q. He *et al.*, "Effective quantization methods for recurrent neural networks," 2016, *arXiv:1611.10176*. [Online]. Available: http://arxiv.org/abs/1611.10176

[78] K. Bong *et al.*, "A low-power convolutional neural network face recognition processor and a CIS integrated with always-on face detector," *IEEE J. Solid-State Circuits*, vol. 53, no. 1, pp. 115–123, Jan. 2018.

[79] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, Jan. 2017.

[80] S. Choi, J. Lee, K. Lee, and H. Yoo, "A 9.02mW CNN-stereo-based real-time 3D hand-gesture recognition processor for smart mobile devices," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 220–222.

[81] W. Luo *et al.*, "Efficient deep learning for stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2016, pp. 5695–5703.

[82] J. Martinez-Nieto *et al.*, "Integrated mixed mode neural network implementation," in *Proc. Eur. Conf. Circuit Theory Design*, Sep. 2017, pp. 1–4.

[83] F. M. Bayat *et al.*, "Implementation of multilayer perceptron network with highly uniform passive memristive crossbar circuits," *Nature Commun.*, vol. 9, p. 2331, Mar. 2018.

[84] B. V. Benjamin *et al.*, "Neurogrid: Mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[85] W. Haensch *et al.*, "The next generation of deep learning hardware: Analog computing," *Proc. IEEE*, vol. 107, no. 1, pp. 108–122, Jan. 2019.

[86] B. Larras *et al.*, "Ultra-low-energy mixed-signal IC implementing encoded neural networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 11, pp. 1974–1985, Nov. 2016.

[87] F. Saffar *et al.*, "A neural network architecture using high resolution multiplying digital to analog converters," in *Proc. IEEE Int. Midwest Symp. Circuits Syst.*, Aug. 2017, pp. 1454–1457.

[88] S. Moon, K. Shin, and D. Jeon, "Enhancing reliability of analog neural network processors," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 27, no. 6, pp. 1455–1459, Feb. 2019.

[89] S. Lin, R. Huang, and T. Chiueh, "A tunable Gaussian/square function computation circuit for analog neural netowrk," *IEEE Trans. Circuits Syst. II. Analog Digit. Signal Process*, vol. 45, no. 3, pp. 441–446, Mar. 1998.

[90] T. Kettner, C. Heite, and K. Schumacher, "Analog CMOS realization of fuzzy logic membership functions," *IEEE J. Solid-State Circuits*, vol. 28, no. 7, pp. 857–861, Jul. 1993.

[91] M. Milev and M. Hristov, "Analog implementation of ANN with inherent quadratic nonlinearity of the synapses," *IEEE Trans. Neural Netw.*, vol. 14, no. 5, pp. 1187–1200, Sep. 2003.

[92] J. Schemmel, F. Jieres, and K. Meier, "Wafer-scale integration of analog neural netowrks," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 2008, pp. 431–438.

[93] O. Rossetto, C. Jutten, J. Herault, and I. Kreuzer, "Analog VLSI synaptic matrices as building blocks for neural networks," *IEEE Micro*, vol. 9, no. 6, pp. 56–63, Dec. 1989.

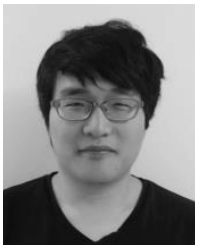
[94] B. Hershberg *et al.*, "A 3.2GS/s 10 ENOB 61mW ringamp ADC in 16nm with background monitoring of distortion," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 58–59.

[95] A. Ramkaj *et al.*, "A 5GS/s 158.6mW 12b passive-saampling 8x-interleaved hybrid ADC with 9.4 ENOB and 160.5dB FOMS in 28nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 62–63.

[96] L. Jie, B. Zheng, and M. P. Flynn, "A 50MHz-bandwidth 70.4dB-SNDR calibration-free time-interleaved 4th-order noise-shaping SAR ADC," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 332–333.
[97] S. Bianco, R. Cadene, L. Celona, and P. Napoletano, "Benchmark analysis of representative deep neural network architectures," *IEEE Access*, vol. 6, pp. 64270–64277, 2018.
[98] S. Han *et al.*, "EIE: Efficient inference engine on compressed deep neural network," in *Proc. ACM/IEEE Annu. Int. Symp. Comp. Archit. (ISCA)*, Jun. 2016, pp. 243–254.
[99] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communitcation network and interface," *Science*, vol. 345, pp. 668–673, Feb. 2014.
[100] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018.
[101] D. Miyashita *et al.*, "A neuromorphic chip optimizec for deep learning and CMOS technology with time-domain analog and digital mixed-signal processing," *IEEE J. Solid-State Circuits*, vol. 52, no. 10, pp. 2679–2689, Oct. 2017.
[102] M. J. Marinella *et al.*, "Multiscale co-design analysis of energy, latency, area, and accuracy of a ReRAM analog neural training accelerator," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 86–101, Mar. 2018.
[103] K. Bong *et al.*, "A 0.5 error 10mW CMOS image sensor-based gaze estimation processor," *IEEE J. Solid-State Circuits*, vol. 51, no. 4, pp. 1032–1040, Apr. 2016.
[104] Z. Wang, J. Zhang, and N. Verma, "Realizing low-energy classification systems by implementing matrix multiplication directly within an ABC," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 6, pp. 825–837, Dec. 2015.
[105] C. Kim *et al.*, "A CMOS image sensor-based stereo matching accelerator with focal-plane sparse rectification and analog census transform," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 12, pp. 2180–2188, Nov. 2016.
[106] M. Kang, "An in-memory VLSI architecture for convolutional neural network," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 3, pp. 494–505, Sep. 2018.
[107] J. Lee *et al.*, "A 31.2pJ/disparity pixel stereo matching processor with stereo SRAM for mobile UI application," in *Proc. IEEE Symp. VLSI Circuits (VLSIC)*, Jun. 2017, pp. 158–159.
[108] J. Yang *et al.*, "Sandwich-RAM: An energy-efficient in-memory BWN architecture with pulse-width modulation," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2019, pp. 394–395.
[109] A. Biswas and A. P. Chandrakasan, "CONV-SRAM: An energy-efficient SRAM with in-memory dot-product computation for low-power convolutional neural networks," *IEEE J. Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.
[110] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-memory computing SRAM macro for binary/ternary deep neural networks," *IEEE J. Solid-State Circuits*, Early Access, Jan. 14, 2020, doi: 10.1109/JSSC.2019.2963616.

Kyuho Jason Lee (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2012, 2014, and 2017, respectively. Before joining UNIST as a Faculty Member, he had worked for Samsung Research America as a Hardware Designer in 2016. From 2017 to 2018, he was a Post-Doctoral Researcher with the Information Engineering and Electronics Research Institute, KAIST, Daejeon, South Korea. He is currently an Assistant Professor with the School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST). His research interests include mixed-mode neuromorphic SoC, deep learning processor, network-on-chip architectures, and intelligent computer vision processor for mobile devices and autonomous vehicles. He has also been serving as a TPC member for the IEEE Asian Solid-State Circuits Conference and the ACM/IEEE Design, Automation and Test in Europe since 2018.

Jinmook Lee (Member, IEEE) received the B.S. degree in electrical engineering from Hanyang University, Seoul, South Korea, in 2014, and the M.S. and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2016, and 2020, respectively. His research interests include energy-efficient deep learning inference/training accelerator design and verification, embedded system development with FPGA programming, and deep learning algorithm for sequence recognition.



Sungpill Choi (Member, IEEE) received B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2013, 2015, and 2019, respectively. He was a Post-Doctoral Researcher with the Information Engineering and Electronics Research Institute, KAIST, from 2019 to 2020. He is currently with Samsung Research, South Korea. His research interests include stereo matching processor, hand gesture recognition interface processor, deep learning processor, and hardware-efficient computer vision algorithms.

Hoi-Jun Yoo (Fellow, IEEE) received the bachelor's degree from the Electronic Department, Seoul National University, Seoul, South Korea, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1985 and 1988, respectively. Since 1998, he has been the Faculty Member with the Department of Electrical Engineering, KAIST. From 2001 to 2005, he was the Director of the Korean System Integration and IP Authoring Research Center, Seoul. In 2007, he founded the System Design Innovation and Application Research Center, KAIST. Since 2010, he has been the General Chair of the Korean Institute of Next Generation Computing, Seoul. He is currently a Full Professor with KAIST. He has authored or coauthored the *DRAM Design* (South Korea: Hongrung, 1996), the *High-Performance DRAM* (South Korea: Sigma, 1999), the *Future Memory: FRAM* (South Korea: Sigma, 2000), the *Networks On Chips* (Morgan Kaufmann, 2006), the *Low-Power NoC for High-Performance SoC Design* (CRC Press, 2008), the *Circuits at the Nanoscale* (CRC Press, 2009), the *Embedded Memories for Nano-Scale VLSIs* (Springer, 2009), the *Mobile 3-D Graphics SoC From Algorithm to Chip* (Wiley, 2010), the *Biomedical CMOS ICs* (Springer, 2011), the *Embedded Systems* (Wiley, 2012), and the *Ultralow-Power Short-Range Radios* (Springer, 2015). He has authored more than 400 articles. His current research interests include artificial intelligence, computer vision system-on-chip (SoC), body area networks, and biomedical devices and circuits. He was a recipient of the Electronic Industrial Association of Korea Award for his contribution to DRAM Technology in 1994, the Hynix Development Award in 1995, the Korea Semiconductor Industry Association Award in 2002, the Best Research of KAIST Award in 2007, the Scientist/Engineer of this month Award from the Ministry of Education, Science and Technology of Korea in 2010, the Best Scholarship Awards of KAIST in 2011, and a co-recipient of the ASP-DAC Design Award 2001, the Outstanding Design Awards from 2005 to 2007, 2010, 2011, and 2014 A-SSCC, and the Student Design Contest Award of 2007, 2008, 2010, and 2011 DAC/ISSCC. He received the Order of Service Merit from the Ministry of Public Administration and Security of Korea in 2011. He has served as a member for the Executive Committee of ISSCC, the Symposium on VLSI, and A-SSCC, and the TPC Chair of the A-SSCC 2008 and ISWC 2010, the IEEE Distinguished Lecturer from 2010 to 2011, the Far East Chair for the ISSCC from 2011 to 2012, the Technology Direction Sub-Committee Chair for the ISSCC in 2013, the TPC Vice Chair for the ISSCC in 2014, and the TPC Chair for the ISSCC in 2015. From 2003 to 2005, he was the full-time Advisor to the Minister of Korea, Ministry of Information and Communication, and National Project Manager of SoC and Computer.