Microelectronics Journal 66 (2017) 76–83

Contents lists available at ScienceDirect

Microelectronics Journal

journal homepage: www.elsevier.com/locate/mejo

Chip implementation of supervised neural network using single-transistor synapses

E.S. Jeng*, S.W. Chou, H.X. Chen, Y.L. Chiang

Department of Electronic Engineering, Chun-Yuan Christian University, Chung Li, Taiwan 32023, Taiwan, ROC

ARTICLE INFO

Keywords:
Single transistor synapse
Neural network chip
Pattern recognition

ABSTRACT

In this work, the newly developed neural chip applied in analog inputs for on-chip training and recognition is presented. We have designed the neural chip using single-transistor synapses which are capable of storing analog weights. The neural chip includes the interface circuit, power switches, analog synaptic array (7 × 4 synapses), and transresistance amplifiers (TR_AMPs) for on-chip training and recognition. Voice signals were acquired using analog signal processing and conditioning circuits for use in verifying the chip's pattern recognition functionality. The experimental results reveal that the synaptic weights of the neural network have adapted with training and have gradually converged to the targets afterwards. Upon system convergence, the recognition rates of the targeted speaker and the three others were evaluated. By using very small amount of synapses, as few as 28 synapses, the system's successful recognition rate for the targeted speaker is 93.5% for 200 tests; whereas, the rate for the other speakers is approximately 6.3% for 600 tests.

1. Introduction

Current computer architecture is based on von Neumann model to process the data and make the decisions. Users run algorithms on the processors to recognition, analytics, and inference. Von Neumann architecture has only one bus which is used for both data transfers and instruction fetches, and therefore data transfers and instruction fetches must be scheduled that they cannot be performed at the same time. With the emergence of big data and artificial intelligence applications, this traditional method faces some of limitations such as energy efficient, scalability, and throughput. Researchers therefore have proposed parallel processing of Artificial Neural Network (ANN) to emulate biological neurons to replace [1–5]. The basic of ANN is shown in Fig. 1(a). Synaptic multiplication is defined as the product of the stored weight and the applied input. The experimental evidence for multiplications with synapses was reported by analyzing motion perceptions in biological nervous systems [6]. When each synaptic multiplication is completed, the output layer sums all the synaptic products from each neuron. The output (a_i) of each neuron in the perceptron is defined as follows,

$$a_i = \sum_{i=1}^{n} P_i \times W_i \tag{1}$$

where P_i and W_i represent the input and weight of each synapse j in the same neuron, respectively.

Programmable FPGAs based ANN systems attain flexible, high speed, and logic densities that rival Von Neumann architecture [1,2]. These solutions have high performance computing platforms to simulate a large number of neurons, but it is highly limited by the throughput. Several studies designed ANN hardware to reduce power consumption and chip area by placing the memories (synapses) and computing units (neurons) in close proximity. The memories and computing units can communicate on-chip routing directly. Digital circuits which use logic devices as synapses and store weight in digital memory have been realized for real time pattern recognition on silicon chip. However, even in the latest study [3], each synapse requires more than 15 transistors. Much effort has been devoted to pursuing a single analog transistor memory as an artificial synapse. Floating gate transistor is a typical artificial synapse which can store and update its weight like a biological synapse [4]. Two-terminal devices such as Resistive RAMs (ReRAMs) have also been proposed as synaptic devices because of their low voltage operation and highly integrated density at the nano-scale [5]. ReRAM is an emerging NVM which has advantages of small bit size, low operating voltage (< 1 V), and scalability below nano-scale. Recently, the novel memory that shows electrically-triggered resistive switching phenomenon has been proposed as plastic device that can be programmed in an analog fashion. In [7], Park et al. has proposed 64-analog levels switching by hybrid pulse scheme. Although ReRAM is potentially used as a synapse, the circuit architecture, power consumption, and learning algorithm are different among

* Corresponding author.
E-mail address: erikjeng@cycu.edu.tw (E.S. Jeng).

http://dx.doi.org/10.1016/j.mejo.2017.06.001
Received 21 November 2016; Received in revised form 3 April 2017; Accepted 2 June 2017

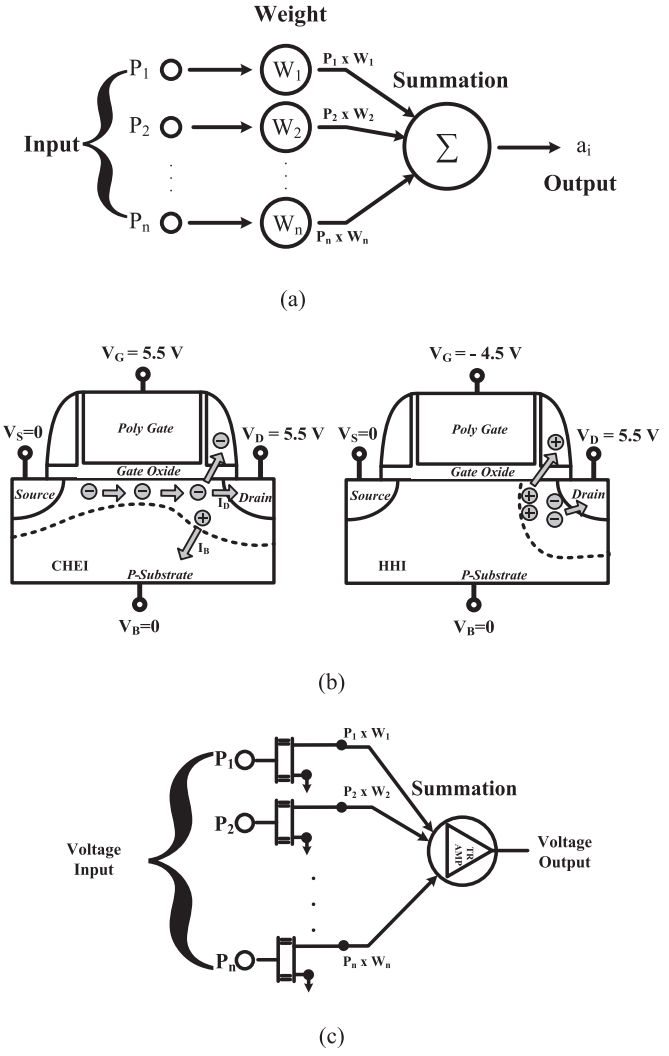

Fig. 1. (a) Multiplication and summation are the basic operations in a single layer artificial neural network. (b) The CHEI and HHI operations in the NOI synaptic MOSFET are depicted. (c) The equivalent circuit of NOI artificial neural network corresponding to Fig. 1(a) is shown.

various research groups. ReRAM devices used in the crossbar-based memristors, spike timing-dependent plasticity (STDP) bilayer resistive switching, two-phase-change-memory synapse, and integrate-and-fire neural circuit have been proposed for neuromorphic hardware applications [8–11]. However, ReRAMs are by far under the development since their material properties and reliability is still immature. On the other hand, the Non-Overlapped Implantation (NOI) device as a single-transistor synapse is the most accessible vehicle for ANN hardware because of the mature CMOS technology. The NOI synapses are fabricated using a typical 0.25 μm CMOS process without the lightly doped drain implantation at the source and drain [12–17]. NOI devices are not only CMOS process compatible, but also have been demonstrated through a single layer artificial neural network in our prior study. NOI devices like other SONOS memories are still facing the reliability challenge when the device technology shrinks to nano-scale [18]. In the previous work [19], NOI synapses have demonstrated the supervised learning capability using digital input patterns in a 4 × 3 NOI synaptic array. The chip is a single layer artificial neural network (ANN). The perceptron is a single layer neural network capable of learning and adapting [20]. The neural input signal is applied to the gate of the NOI synapse. Its weight depends upon the threshold voltage (V_{th}) which is strongly influenced by the trapped charges in the NOI device. The following equation is used to express V_{th} shifting (ΔV_{th}) in a NOI synapse,

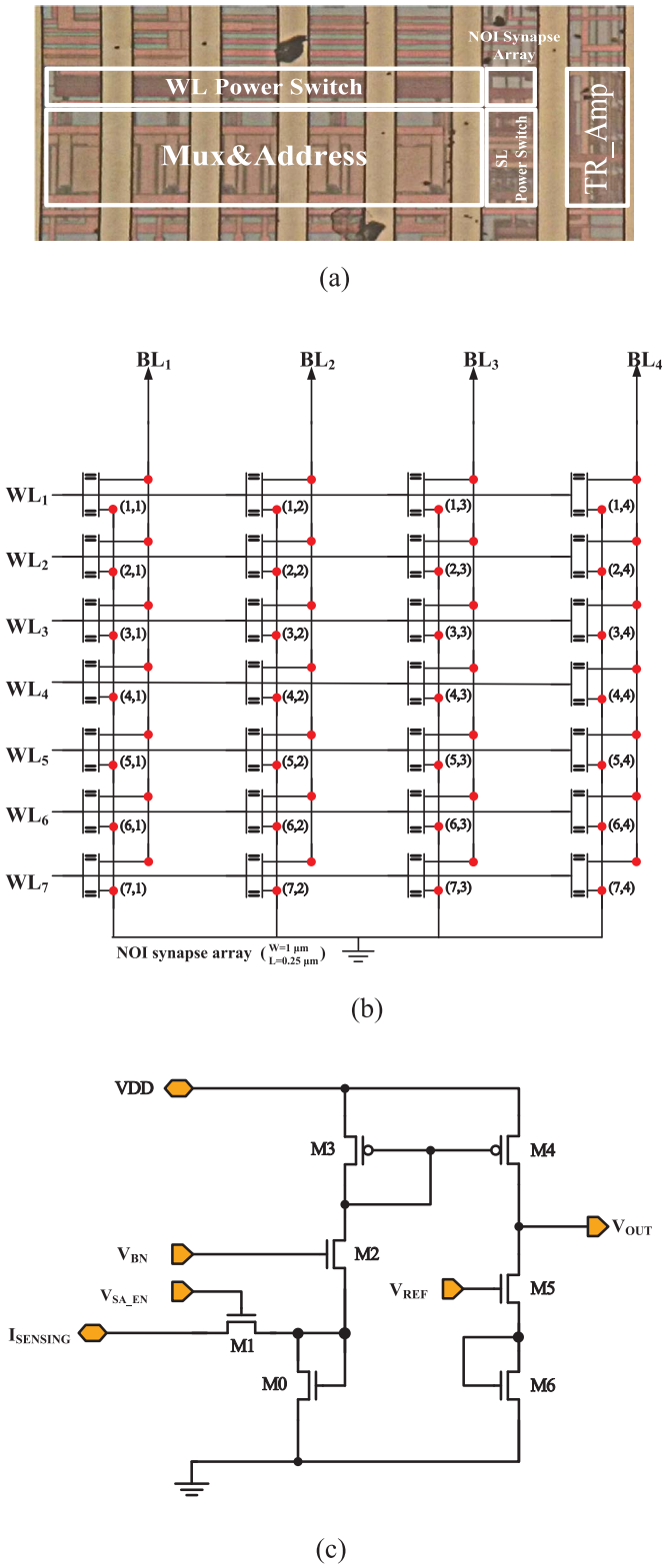

Fig. 2. (a) Micro-photograph of silicon NOI neural network is presented. The schematics of NOI synaptic array and TR_AMP are depicted in (b) and (c), respectively.

$$\Delta V_{th} \cong \frac{Q_{inj}}{C_{fr}} \tag{2}$$

where C_{fr} is the fringing field capacitance which includes the tunneling oxide and the gate spacer capacitance. Q_{inj} is the equivalent quantity of electrons trapped inside the gate spacer.

The NOI synaptic channel current I_{DS} is a function of V_{GS} and V_{th}