Solid-State Electronics 79 (2013) 227–232

Contents lists available at SciVerse ScienceDirect

Solid-State Electronics

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

Addition of HfO_2 interface layer for improved synaptic performance of phase change memory (PCM) devices

M. Suri [a,*], O. Bichler [b], Q. Hubert [a], L. Perniola [a], V. Sousa [a], C. Jahan [a], D. Vuillaume [c], C. Gamrat [b], B. DeSalvo [a]

[a] CEA-LETI-MINATEC, Grenoble, France
[b] CEA-LIST, LCE, Gif-sur-Yvette, France
[c] CNRS-IEMN, Lille, France

ARTICLE INFO

Article history:
Received 17 April 2012
Received in revised form 29 August 2012
Accepted 5 September 2012
Available online 16 October 2012

The review of this paper was arranged by Prof. S. Cristoloveanu

Keywords:
Phase change memory
Bio-inspired computing
Spike time dependent plasticity
Visual pattern extraction
Interface engineering

ABSTRACT

In this work, we will focus on the use of phase change memory (PCM) to emulate synaptic behavior in emerging neuromorphic system-architectures. In particular, we will show that the performance and energy-efficiency of large scale neuromorphic systems can be improved by engineering individual PCM devices used as synapses. This is obtained by adding a thin HfO_2 interface layer to standard GST PCM devices, allowing for the lowering of the Set/Reset currents and the increase of the number of intermediate resistance states (or synaptic weights) in the synaptic potentiation characteristics. The experimentally obtained potentiation characteristics of such PCM devices are used to simulate a 2-layer ultra-dense spiking neural network (SNN) and to perform a complex visual pattern extraction from a test case based on real world data (i.e. cars passing on a 6-lane freeway). The total power dissipated in the learning mode, for the pattern extraction experiment is estimated to be as low as 60 μW. Average detection rate of cars is found to be greater than 90%.

1. Introduction

Interest in the field of biologically inspired computational hardware has increased exponentially over the last few years. Researchers at several universities and corporations like IBM [1], Stanford [11], Intel [2], Qualcomm [3], Samsung [4] and HP [5] are exploring neuromorphic hardware due to its promising advantages such as low-power, high tolerance towards defects and variability, and efficient handling of large scale non-linear computations. The main building blocks for any neuromorphic system should include three ingredients; the neuron, the synapse and a plausible learning rule. In terms of basic functionality the neuron acts as a source and integrator of electrical spikes (or action potentials). The synapse is the communication channel connecting two neurons, as shown in Fig. 1. The conductance or strength of the synapse is plastic and can either increase (potentiation) or decrease (depression). The learning-rule is the actual algorithm or protocol which determines when and how the synaptic strength should change. The mammalian cerebral cortex contains an enormously large number of neurons (10^{10}) and synapses (10^{14}–10^{15}) [6]. The idea of achieving such a high synaptic density using pure CMOS synapse circuits is not practical in terms of on-chip silicon area consumption, due to the large number of transistors required (>10) per synapse [7]. Thus hybrid neuromorphic architectures containing CMOS circuits emulating the neuron co-integrated with nanoscale devices emulating the synapse have been proposed [8]. According to neurophysiologic models [9], in order to emulate synaptic behavior a device should have a conductance that can be modulated (i.e. gradual increase or decrease its conductance in response to neuron spikes – named "long term potentiation, LTP" – or "long term depression, LTD", respectively) and it should be non-volatile. Both the criterion of conductance modulation and non-volatility can be satisfied by resistive memory devices. More recently, several types of unipolar and bipolar resistive memory technologies [24] such as phase change memory (PCM) [10,11], conductive-bridge (CBRAM) or programmable metallization cell (PMC) [12,19], and oxide-resistive (OXRAM) memory [13], have been demonstrated as suitable candidates for emulating synaptic behavior. Note that due to the large number of synapses in the neural networks, the learning, computing, and storage become a statistical process inherently tolerant to variability. Indeed, this means that in the case of neuromorphic applications, the noise-margin or the classification of intermediate resistance levels of the resistive memory is far more relaxed compared to conventional multi-bit storage. The firing of a neuron is not dependent on the resistance of any single synapse

* Corresponding author.
E-mail address: manansuri2002@gmail.com (M. Suri).

0038-1101/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.sse.2012.09.006

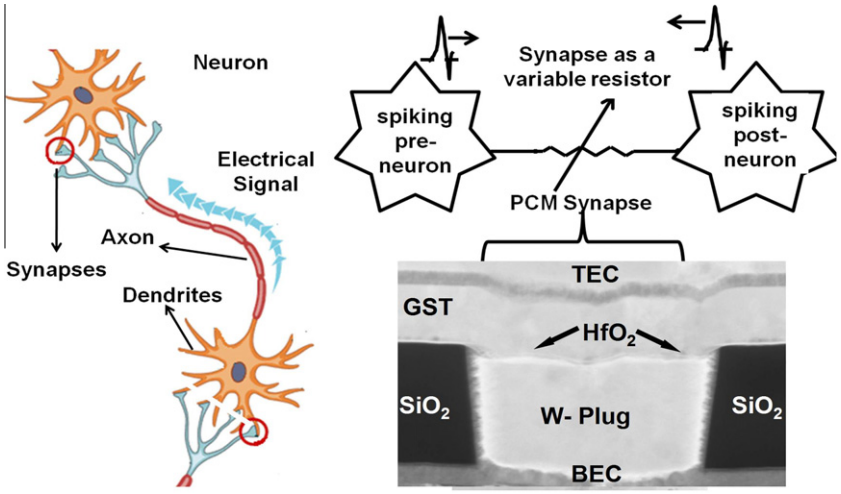

Fig. 1. Illustration showing synapse connecting two neurons and equivalent PCM based circuit. Inset shows a TEM image of GST PCM device with HfO_2 interfacial layer used in this work.

rather it depends upon the resistance of all the synapses connected to it. The synapse acts like a stochastic, programmable, analog non-volatile resistor. Thus the need for emulating synaptic behavior in bio-inspired hardware using nanoscale CMOS compatible devices, and the inherent tolerance of neural computing towards variability, opens up a new and potentially disruptive application for resistive memory technologies. In previous work, we presented a 2-layer ultra-dense spiking neural network (SNN) based on PCM synapses [10]. The system was able to perform complex visual pattern extraction. PCM technology was chosen because of its inherent advantages compared to other resistive memory technologies, such as maturity, scaling capability, high endurance, and good reliability [14]. The synaptic power consumption of such ultra-dense neuromorphic systems was pointed out as one of the main challenge of this architecture. In this paper, we build on our previous work and show that the engineering of individual PCM devices by addition of a thin HfO_2 interfacial layer to standard GST PCM devices, lowers the Set/Reset currents and increases the number of intermediate resistance states (or synaptic weights) in the synaptic potentiation characteristics, thus significantly enhancing the energy-efficiency of the entire neuromorphic system.

2. Our approach

2.1. The 2-PCM synapse

To emulate synaptic behavior (i.e. gradual synaptic potentiation and depression) in PCM devices we formulated the so-called "*2-PCM synapse*" concept (Fig. 2) [10]. In this approach, we use two PCM devices to implement a single synapse and connect them in a complementary configuration to the post-synaptic output neuron. One of the PCM device implements synaptic potentiation (LTP-device), while the other implements synaptic depression (LTD-device). Both devices are initialized to a high resistive amorphous state. When synaptic potentiation is required, the LTP device is crystallized, while when synaptic depression is required, the LTD device is crystallized. Important benefits of the 2-PCM synapse approach are the following. First, exploiting mostly the crystallization phenomenon of chalcogenide devices, it allows defining a programming methodology for the two PCM devices which uses identical neuron spikes (or pulses). In fact, generation of non-identical pulses is to be avoided due to the increased complexity of CMOS neuron circuits, added parasitic effects, such as capacitive line charging, and excessive power dissipation. Second, the "*2-PCM synapse*" is very low power, because the majority of the synaptic events are achieved by crystallization, which is a less energy consuming process for PCM compared to amorphization. Another inherent advantage of this approach is the decrease of the impact of resistance drift on the stored synaptic information. Since we potentiate and depress the synapses by crystallization, the majority of the synaptic information is stored or programmed in low resistance states of the PCM devices. Crystalline or low resistance states are more stable and immune to the resistance drift phenomena compared to high resistance PCM states [15]. All the advantages discussed above define the motivation behind our choice of programming the PCM devices from the reset-to-set front (LTP). However, it is important to note that when the neural network undergoes learning, as time passes the PCM devices become more and more crystallized and finally saturate to a minimum resistance value. In order to enable continuous learning of the network, we defined a refresh-sequence, explained in detail in [23]. In this refresh-sequence, the saturated PCM devices are amorphized and the effective weights of the corresponding synapses are re-programmed. In terms of reducing power-dissipation at the system level, it is thus important to reduce the number of times the refresh-sequence is initiated. Initiation of the refresh-sequence can be decreased if the number of intermediate resistance states in the LTP characteristics of the PCM devices is increased. In other words, if the PCM device takes much longer to attain its minimum resistance value or it crystallizes slowly, it would reduce the number of refresh-sequences required. In this paper, we propose an original solution to address this issue.

2.2. Interface layer PCM

Lance-type PCM devices with a 300 nm diameter cylindrical tungsten (W) heater-plug and 100 nm-thick $Ge_2Sb_2Te_5$ (GST) layer were fabricated (see inset of Figs. 1 and 3). The GST layer is deposited by sputtering of a GST target and is amorphous initially. After deposition it is crystallized by annealing at 200 °C for 15 min. A 2 nm thick HfO_2 layer was deposited between the heater plug and the GST layer by atomic layer deposition (ALD). We observed that adding a thin layer of HfO_2 to the GST devices increases the number of points in the LTP curve and thus reduces the total power dissipation by decreasing the number of refresh-sequences