ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/230727604

Visual Pattern Extraction Using Energy-Efficient "2-PCM Synapse" Neuromorphic Architecture

Article *in* IEEE Transactions on Electron Devices · August 2012

DOI: 10.1109/TED.2012.2197951

CITATIONS
181

READS
1,261

5 authors, including:

Olivier Bichler
Atomic Energy and Alternative Energies Commission
81 PUBLICATIONS **3,533** CITATIONS
SEE PROFILE

Dr. Manan Suri
Indian Institute of Technology Delhi
110 PUBLICATIONS **1,716** CITATIONS
SEE PROFILE

Damien Querlioz
French National Centre for Scientific Research
241 PUBLICATIONS **8,392** CITATIONS
SEE PROFILE

Christian Gamrat
Atomic Energy and Alternative Energies Commission
96 PUBLICATIONS **4,114** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

NeuRAM3 - NEUral computing aRchitectures in Advanced Monolithic 3D-VLSI nano-technologies View project

BAMBI (Bottom-up Approaches to Machines dedicated to Bayesian Inference) View project

All content following this page was uploaded by Damien Querlioz on 16 February 2014.

The user has requested enhancement of the downloaded file.

Visual Pattern Extraction Using Energy-Efficient "2-PCM Synapse" Neuromorphic Architecture

Olivier Bichler, Manan Suri, Damien Querlioz, *Member, IEEE*, Dominique Vuillaume, Barbara DeSalvo, and Christian Gamrat

***Abstract*—We introduce a novel energy-efficient methodology *"2-PCM Synapse"* to use phase-change memory (PCM) as synapses in large-scale neuromorphic systems. Our spiking neural network architecture exploits the gradual crystallization behavior of PCM devices for emulating both synaptic potentiation and synaptic depression. Unlike earlier attempts to implement a biological-like spike-timing-dependent plasticity learning rule with PCM, we use a simplified rule where long-term potentiation and long-term depression can both be produced with a single invariant crystallizing pulse. Our architecture is simulated on a special purpose event-based simulator, using a behavioral model for the PCM devices validated with electrical characterization. The system, comprising about 2 million synapses, directly learns from event-based dynamic vision sensors. When tested with real-life data, it is able to extract complex and overlapping temporally correlated features such as car trajectories on a freeway. Complete trajectories can be learned with a detection rate above 90%. The synaptic programming power consumption of the system during learning is estimated and could be as low as 100 nW for scaled down PCM technology. Robustness to device variability is also evidenced.**

***Index Terms*—Neuromorphic system, phase-change materials, spike-timing-dependent plasticity, spiking neural network, 2-PCM synapse.**

I. Introduction

THE POSSIBILITY of using resistive memory devices as synapses in bio-inspired hardware has received growing interest. Phase-change memory (PCM) devices, in particular, have been proposed to emulate biologically inspired features of synaptic functionality that are essential for realizing neuromorphic hardware [1], [2]. Among the different types of emulated synaptic features, spike-timing-dependent plasticity (STDP) has gained a lot of significance recently. STDP is widely believed to be a foundation of learning mechanisms inside the brain [3]. The basic mechanism of STDP involves reinforcing in average the strength of the synapses that were activated a short time before the activation of the postsynaptic neuron. On the contrary, the synapses that are activated immediately after tend to be weakened. STDP also finds its applications in the fields of computational neuroscience and machine learning.

When the number of neurons and synapses in a neuromorphic system featuring STDP grows large, its implementation on classical computer architecture quickly becomes a severe demonstration of the von Neumann bottleneck [4]. This is a major reason motivating the research for new neuromorphic memory architectures that could allow *in situ*, instantaneous, and fully parallel synaptic-weight updates. From the viewpoint of spiking neural networks (SNNs), the main benefit that PCM-based synapses offer is the ease of conductance modulation by the application of simple electrical pulses. From a technological perspective, PCM is a good candidate for neuromorphic applications because of CMOS compatibility, high scalability, strong endurance, and good retention characteristics.

However, defining practical implementation strategies and useful applications of large-scale cointegrated hybrid neuromorphic systems (CMOS neurons with resistive memory synapses) remains an unsolved challenge. Demonstrating a useful STDP application and developing an implementation approach with an emphasis on low synaptic power consumption constitute the goals of this paper.

In this paper, we introduce a novel low-power architecture *"2-PCM Synapse"* for emulation of synaptic functions in large-scale neural networks. Using this architecture, we designed a fully connected feedforward SNN and implemented a simplified form of the biological STDP learning rule. Finally, we show a real-world application of extracting complex patterns from recorded video data. This paper is organized as follows. In Section II, the state of the art of PCM-based STDP implementations and learning is reviewed. In Section III, our proposed architecture is described. Simulation results are discussed in Section IV, including detailed analysis on learning performances, robustness, power consumption, and scalability.

II. State of the Art

This section briefly describes the state of the art of PCM STDP implementations, with an emphasis on their potential capabilities and constraints. In this regard, the rationale behind our proposed STDP-like learning scheme is then detailed.

Manuscript received December 19, 2011; revised February 29, 2012 and April 13, 2012; accepted April 23, 2012. Date of publication May 17, 2012; date of current version July 19, 2012. The review of this paper was arranged by Editor Y.-H. Shih.

O. Bichler and C. Gamrat are with the Embedded Computing Laboratory, CEA-LIST, 91191 Gif-sur-Yvette Cedex, France (e-mail: olivier.bichler@cea.fr; christian.gamrat@cea.fr).

M. Suri and B. DeSalvo are with the CEA-LETI, MINATEC, 38054 Grenoble Cedex 9, France (e-mail: manan.suri@cea.fr; barbara.desalvo@cea.fr).

D. Querlioz is with the Institut d'Electronique Fondamentale, Université Paris-Sud, CNRS, 91405 Orsay, France (e-mail: damien.querlioz@u-psud.fr).

D. Vuillaume is with the Institute of Electronics, Microelectronics and Nanotechnology, CNRS, 59652 Villeneuve d'Ascq, France (e-mail: dominique.vuillaume@iemn.univ-lille1.fr).

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

Digital Object Identifier 10.1109/TED.2012.2197951

0018-9383/$31.00 © 2012 IEEE

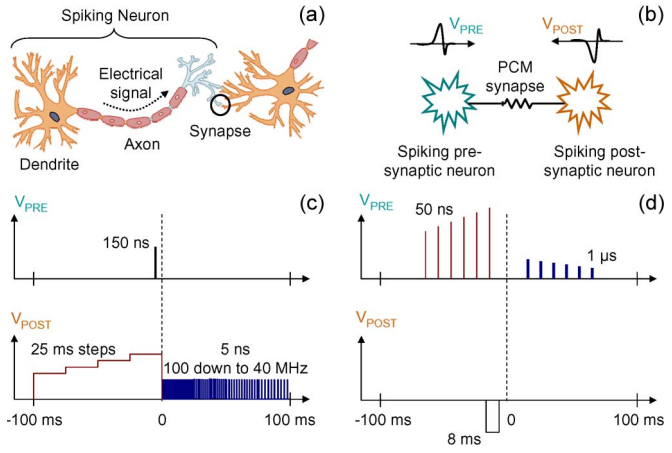


Fig. 1. (a) Biological neuron and synapse. (b) Equivalent circuit for a synaptic connection. (c)–(d) Examples of pre- and postsynaptic pulses for implementing STDP as proposed in (c) [5] and (d) [6]. The indicated timing is the width of the pulses. In (c), the presynaptic pulse is not applied to the second terminal of the PCM but to its associated transistor (not shown).

A. STDP Implementation Proposals

Several STDP implementation proposals with memristive devices have been published over the last couple of years, most of them focusing on bipolar resistive switching devices [7]–[10]. Research on STDP learning with PCM is also performed, which aims to develop neuromorphic architectures that scale to biological levels. PCM is indeed considered to be one of the most promising embedded memory technologies for large-scale cointegration with CMOS [11].

In the current implementations of STDP with PCM [5], [6], [12], [13], synaptic potentiation, inducing long-term potentiation (LTP), is achieved through crystallization and synaptic depression, inducing long-term depression (LTD), through amorphization of the chalcogenide material. Examples of pre- and postsynaptic pulses for two different learning methods are presented in Fig. 1. In the first implementation [see Fig. 1(c)] [5], current is allowed to flow through the PCM only for the duration of the presynaptic pulse. Depending on its relative timing to the postsynaptic pulse, a RESET pulse or a series of SET pulses are applied to the device, with a ΔT-dependent amplitude when $\Delta T < 0$ or a ΔT-dependent frequency when $\Delta T > 0$, respectively (with $\Delta T = t_{\text{post}} - t_{\text{pre}}$). In Fig. 1(d), the basic idea is similar [6], with the postsynaptic pulse acting as the gating function instead of the presynaptic pulse. Pulse amplitude is used instead of frequency to reproduce the biological LTP behavior.

The amorphous region inside the phase change layer can be crystallized by applying SET pulses, thus increasing device conductance. It was shown that the magnitude of the relative increase in conductance can be controlled by the pulse amplitude (as in [6]) and by the equivalent pulse width (as in [5], where the magnitude change is controlled by the number of SET pulses). Amorphization, on the other hand, involves creating a molten region inside the phase change layer and abruptly quenching it, which is a more power-hungry process and is not progressive with identical pulses [14]. The LTD induced by the different pulse voltages is therefore absolute and not relative in value. This means that consecutive LTD (with similar ΔT) would not yield a significant change in the synaptic conductance after the first one. This effect is therefore not apparent in an STDP characterization similar to the one shown in Fig. 2 if, for a given ΔT, only the first conductance change step ΔW is reported (a better STDP characterization may be the recording of a large number of successive measurements with random ΔT, as suggested in [15]).

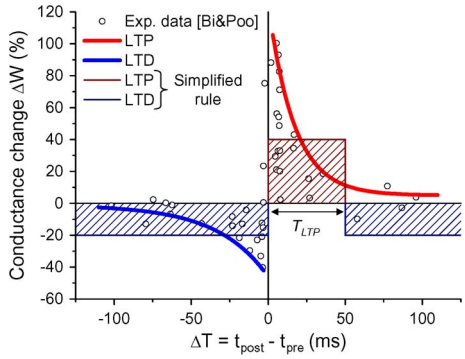


Fig. 2. Biological STDP (from [16]) and simplified STDP used in the proposed PCM implementation. In the simplified rule, a synapse receiving a postsynaptic spike with no presynaptic spike in the LTP window undergoes an LTD regardless of the existence of a presynaptic spike.

B. Simplified Learning Rule

As we have seen, reproducing the complex and ΔT-dependent biological measurements of the STDP learning rule with PCM is not straightforward. The programming pulses need to be tuned both in duration and in amplitude depending on the exact timing difference between the pre- and postsynaptic spike events. The benefit of this biomimetic approach (where we try to match Biology as closely as possible) is that it does not require making assumptions on how the synapses will be exploited for learning in the final system. Indeed, if the biological low-level synaptic update rule is replicated with reasonable accuracy in the electronic system, there is a significant chance that any higher level learning or computation occurring in biological neural networks will be reproducible as well.

There are, however, benefits in designing a specific STDP rule targeted toward our applications. The exact shape of the STDP learning rule may not be required to capture the correct computational behavior. To go further, the measured STDP curve *in vitro* might not reflect the actual behavior of the neurons *in vivo* [17]. Finally, there is not a single STDP rule: a broad family of synaptic update characteristics in function of the pre–post synaptic time difference were recorded [18].

In this paper, we propose a novel approach and aim to solve complex practical problems with a simplified learning rule that is easy and efficient to implement with PCM. In our STDP rule, all the synapses of a neuron are equally depressed upon receiving a postsynaptic spike, except for the synapses that were activated with a presynaptic spike a short time before, which are strongly potentiated [19]. Contrary to a biological synapse, the magnitude of the potentiation or depression is independent on the relative timing between the presynaptic spike and the postsynaptic spike, as shown in Fig. 2.

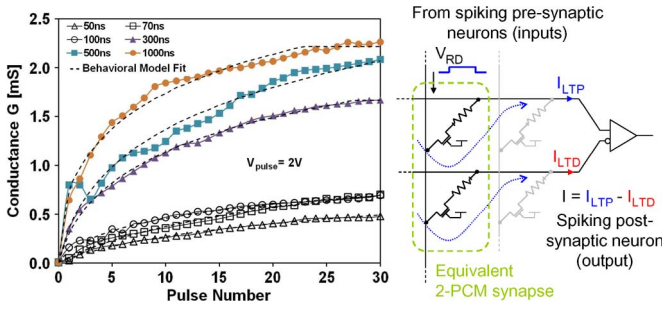


Fig. 3. (Left) Experimental LTP characteristics of $Ge_2Sb_2Te_5$ (GST) PCM devices. For each curve, first, a reset pulse (7 V, 100 ns) is applied followed by 30 consecutive identical potentiating pulses (2 V). Dotted lines correspond to the behavioral model fit used in our simulations. (Right) 2-PCM synapse principle.

III. Proposed Synapse and Architecture

Here, we describe the working principle of our new STDP implementation with PCM. First, the 2-PCM synapse is described. Then, the three modes of operation of the proposed architecture are detailed: read, write, and refresh.

A. 2-PCM Synapse

One of the main limitations of using a single PCM device as a synapse is the implementation of LTD, which is not progressive with amorphization by using invariant or identical pulses [1], [14]. The current required for amorphization is also typically 5–10 times higher than for crystallization, even for state-of-the-art devices (see Fig. 13 at the end of this paper).

To overcome these issues, we propose to implement both LTP and LTD using crystallization, with two PCM devices constituting one synapse, as shown in Fig. 3. The two devices have an opposite contribution to the neuron's integration. When the synapse needs to be potentiated, the so-called LTP device undergoes a partial crystallization, increasing the equivalent weight of the synapse. Similarly, when the synapse must be depressed, the LTD device is crystallized. As the LTD device has a negative contribution to the neuron's integration, the equivalent weight of the synapse is reduced. Furthermore, because gradual crystallization is achieved with successive identical voltage pulses, the pulse generation is greatly simplified.

With this scheme, the conductance of both PCM devices keeps increasing upon undergoing LTP and LTD during the learning, regardless of the effective weight of the equivalent synapse. If no additional action is taken, both devices would eventually reach their maximum conductance, and the equivalent weight of the synapse would always converge to zero. This issue is accounted for by periodically reducing the conductance of the devices while maintaining the equivalent weight of the synapse, with the refresh protocol described in Section III-D. The implications for the learning of real-life stimuli and the overhead in terms of complexity and power consumption are discussed in Section IV.

B. Read Operations

The read operation described here is the normal operation of the network between two output neuron activations. When an input neuron receives an incoming event, it generates a small voltage pulse that is propagated to all its output neurons through its synapses. The resulting current flowing to each output neuron is the difference between the current from LTP devices and from LTD devices, as shown in Fig. 4. The read pulse amplitude and duration can be minimal, as long as it allows reasonably accurate reading of the low-field resistance of the PCM. The output neurons are of type leaky integrate and fire (LIF) [20], [21]. When the integrated current reaches the neuron's threshold, the network enters a write mode operation to update the synaptic weights through STDP, as described in the following section.

Each time an input neuron is activated, it enters or re-enters an LTP internal state for the duration of the STDP LTP window, i.e., T_{LTP}, as shown in Fig. 2 (actual value used in simulations is given in Table I).

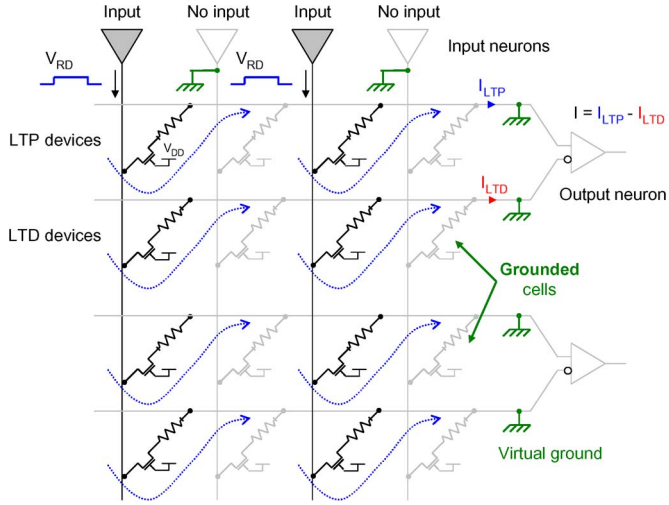


Fig. 4. Read operations. Current from both LTP and LTD PCM devices is integrated in the output neuron, with a positive and negative contribution, respectively.

TABLE I
Neuron Parameters for the Learning. A Different Set of Parameters Is Used Depending on the PCM Materials. See [19] for a Detailed Explanation of the Parameters

Parameter	GST		GeTe	
	1st Layer	2nd Layer	1st Layer	2nd Layer
I_{thres}	2.49 S	0.00437 S	2.50 S	0.00431 S
T_{LTP}	7.59 ms	7.12 ms	11.5 ms	12.9 ms
T_{refrac}	554 ms	410 ms	524 ms	393 ms
$T_{inhibit}$	15.7 ms	56.5 ms	11.8 ms	70.9 ms
τ_{leak}	100 ms	821 ms	115 ms	714 ms
N	30		10	
LTP/LTD	2.0		2.0	

C. Write Operations

When an output neuron fires, it transmits a postspike signal to every input neuron, signaling write operations. In write operations, input neurons generate an LTP pulse of amplitude V_{WR} such that $V_{\mathrm{WR}} < V_{\mathrm{SET}} < 2V_{\mathrm{WR}}$, only if they are in the LTP state. The output firing neuron generates a negative feedback

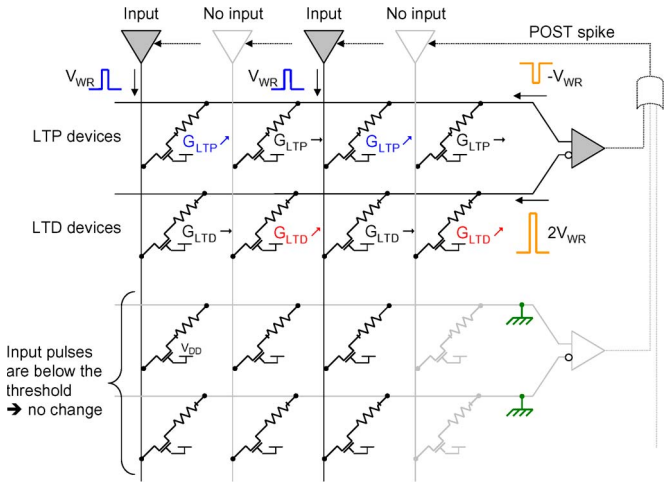


Fig. 5. Write operations based on the simple STDP rule. For a specific PCM, G ↗ denotes an increase in conductance (thus, partial crystallization of the device), whereas G → denotes no change in conductance.

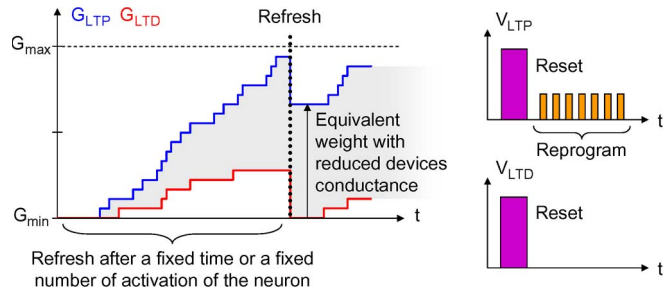


Fig. 6. Refresh principle: The two devices forming a synapse are reset, and the one that had the higher conductance is reprogrammed such that the equivalent weight of the synapse stays unchanged.

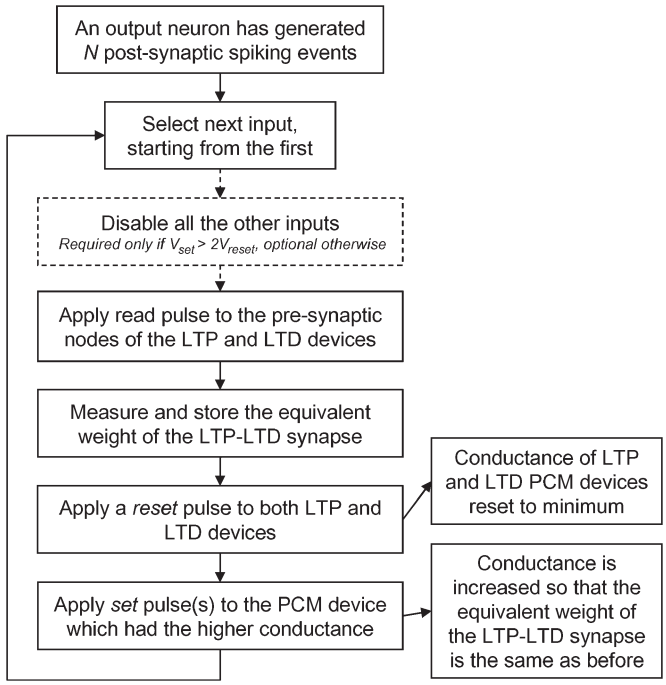


Fig. 7. Refresh operations diagram.

pulse $-V_{\mathrm{WR}}$ for the LTP devices and a positive feedback pulse $2V_{\mathrm{WR}}$ for the LTD devices. When an LTP pulse interacts with the feedback pulses, the effective voltage across the LTP device is $2V_{\mathrm{WR}} > V_{\mathrm{SET}}$, and the voltage across the LTD device is $V_{\mathrm{WR}} < V_{\mathrm{SET}}$. The conductance of the LTP device is then increased. If there is no LTP pulse for a given input, it means that pre–post spike timing difference is not within the LTP window, and thus, the conductance of the LTD device must be increased according to our simplified STDP rule. This is indeed this case, as the voltage across the LTP device is $-V_{\mathrm{WR}} > -V_{\mathrm{SET}}$, and the voltage across the LTD device is $2V_{\mathrm{WR}} > V_{\mathrm{SET}}$ (see Fig. 5).

Selector devices are not required for the write operations, as the input LTP pulse amplitude is below the SET threshold of the PCM devices; hence, the synaptic weights of the other output neurons is not affected. The LTP pulses may, however, significantly alter the integration value of other output neurons. This is a nonissue in the proposed architecture as lateral inhibition is implemented: When a neuron fires, integration of the others is disabled for an inhibit refractory period T_{inhibit}.

D. Refresh Operations

Because the conductance of the PCM devices gradually increases during the learning, a refresh mechanism is introduced to reduce the conductance of LTP and LTD devices while keeping the weight of the equivalent synapse unchanged. The principle of the refresh operation is shown in Fig. 6. When one of the two devices reaches its maximum conductance, they are both programmed to RESET, and the one that had the higher conductance undergoes a series of SET pulses until the equivalent weight is reached again. Because one of the devices stays at minimum conductance, this mechanism enables continued evolution of the weights.

Knowing the average number of conductance steps N achievable for a given PCM technology, with a given SET pulse duration and amplitude, a refresh operation is necessary after N potentiations or N depressions of the synapse (whichever comes first, assuming that one of the devices is initially at minimum conductance). Therefore, output neurons can initiate a refresh operation on their synapses after a fixed number of activations, which would be N in the worst case. Although such a simple mechanism would certainly involve a substantial amount of unnecessary resets, as few synapses would undergo N potentiations or N depressions in a row, it does not require permanent monitoring of the state of the LTP and LTD devices. N can be high (value approaching 100 is shown in [5]), thus reducing the time/energy overhead cost to a minimum. Simulations show that even $N = 10$ incurs only a marginal cost for the system on a real-life learning experiment with almost 2 000 000 synapses (see next section).

Refresh operations are described in the diagram in Fig. 7. The synapses are read, reset, and reprogrammed sequentially. The other neurons are disabled during the process. To strongly amorphize the PCM, a RESET pulse of amplitude V_{RESET} has to be applied across the device, as shown in Fig. 8. If $V_{\mathrm{RESET}} < 2V_{\mathrm{SET}}$, a voltage of V_{RESET} across the PCM can be obtained with the interaction of two pulses of amplitude V_{ER}

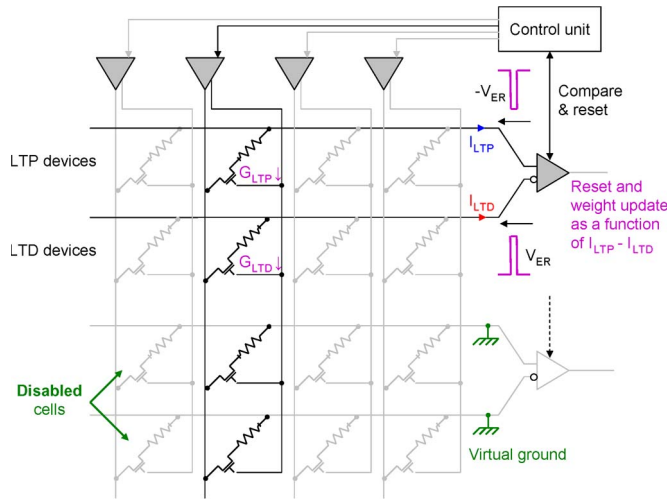


Fig. 8. Refresh operations: RESET pulses generation to reinitialize the LTP and LTD device conductance to the minimum when $V_{\mathrm{RESET}} > 2V_{\mathrm{SET}}$.

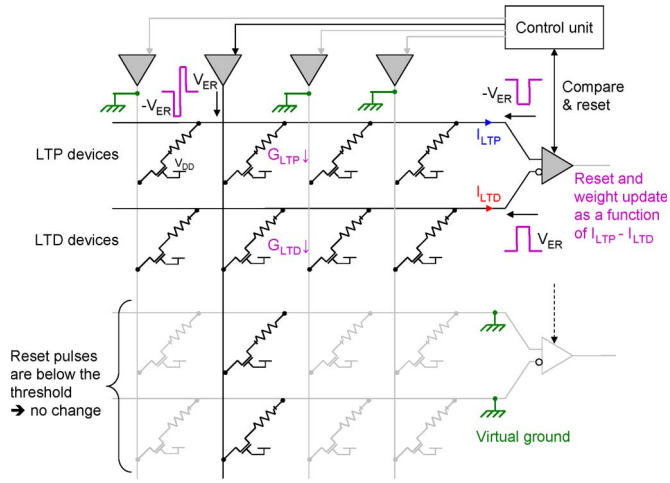


Fig. 9. Refresh operations without selectors: RESET pulses generation to reinitialize the LTP and LTD device conductance to the minimum when $V_{\mathrm{RESET}} < 2V_{\mathrm{SET}}$. The LTD (respectively LTP) device is reset when the negative part $-V_{\mathrm{ER}}$ (respectively positive part V_{ER}) of the erase pulse coming from the input neuron overlaps with the postsynaptic erase pulse.

such that $V_{\mathrm{ER}} < V_{\mathrm{SET}} < 2V_{\mathrm{ER}}$, as shown in Fig. 9. In this case, the voltage across the other synapses in the crossbar is always below the SET threshold, and their conductance is not affected.

Therefore, if the condition $V_{\mathrm{RESET}} < 2V_{\mathrm{SET}}$ is true, no selector device is required for refresh operations. It is noteworthy that this condition is usually verified for scaled down PCMs [22]. As neither the read nor the write operations actually require one, selector devices could be completely eliminated. This would theoretically allow the highest possible PCM integration density in a crossbar and free the underlying CMOS layer for neuron integration.

If $V_{\mathrm{RESET}} > 2V_{\mathrm{SET}}$, the V_{RESET} voltage cannot be obtained with two pulses of amplitude below V_{SET}. Selectors are thus required to disable the other PCM devices and prevent their conductance to be altered, when the 2-PCM synapse being refreshed is reset, as shown in Fig. 8 (disabled PCMs are grayed).

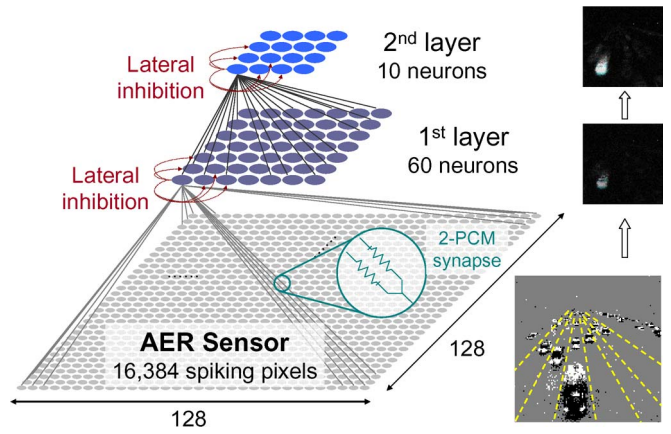


Fig. 10. Network topology used in simulation. It is fully connected, and each pixel of the 128 × 128 pixel AER dynamic vision sensor is connected to every neuron of the first layer through two synapses, receiving positive and negative change in illumination events, respectively.

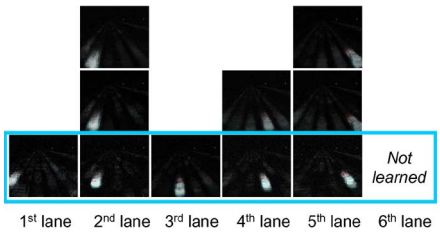
IV. Simulation Results

Here, we present the results of a large-scale learning simulation of our STDP implementation with 2-PCM synapses. We used a special purpose C++ event-based simulator (Xnet) that we developed to simulate large-scale SNNs based on memristive devices for the synapses [19], [23]. The asymmetric behavioral model from [24] is used for the fitting of the LTP characteristic of GST (shown in Fig. 3) and GeTe PCM. The refresh operations using the protocol described earlier are simulated as well.

A. Network Topology and Learning Stimuli

Fig. 10 shows the topological view of the simulated two-layer feedforward SNN [19]. It is a fully connected network, with 60 neurons in the first layer and 10 neurons in the second one. The bottommost layer represents a 128 × 128 pixel address event representation (AER) dynamic vision sensor [25]. A pixel generates an event each time the relative change of its illumination intensity reaches a positive or a negative threshold. Therefore, depending on the sign of the intensity change, events can be of either type ON or type OFF, corresponding to an increase or a decrease in pixel illumination, respectively. There are two synapses per pixel, one for each event type. The total number of synapses in this system is thus $2 \times 128 \times 128 \times 60 + 60 \times 10 = 1\,966\,680$ and thus 3 933 360 PCM devices (2 PCM/synapse). The neurons are standard LIF neurons with a refractory period T_{refrac}. When a neuron spikes, it also disables all the other neurons during a period T_{inhibit}, during which no incoming spike is integrated.

The stimulus used in the following simulations was recorded from the TMPDIFF128 DVS sensor. It represents cars passing under a bridge over the 210 freeway in Pasadena. The sequence is 78.5 s in duration, containing a total of 5.2M events, with an average event rate of 66.1k events per second. The bottom-right picture in Fig. 10 shows a rendering of the sequence, where the traffic lanes have been marked. With STDP, the neural network is able to extract any repetitive pattern from the stimuli



Car Detection %		
Lane	PCM	
	GST	GeTe
1st	N.A.	88
2nd	100	91
3rd	89	96
4th	89	97
5th	96	100
6th	N.A.	N.A.

Fig. 11. (Left) Car detection for the six different lanes by the second layer ten output neurons. Each block shows the sensitive map of one output neuron (here with GeTe PCM). Up to down, the detection rate increases. Lanes 2 and 5 are learn by three neurons each, whereas lane 6 is not learn by any neuron. (Right) Detection rate of five best neurons for each lane.

in an unsupervised way [19]. In the present case, the learned repetitive patterns are the car trajectories, which can be used to detect/count cars on each lane.

Neuron parameters for the two layers are given in Table I, for the simulated networks with GST-PCM and GeTe-PCM. I_{thres} is the neuron's threshold, which is expressed in siemens (S) to make it independent of the read pulse's voltage and duration. τ_{leak} is the neuron's leak time constant. The neuron parameters are adapted to the stimuli dynamic, corresponding to the average spiking activity generated by the cars at the bottom of the retina (where activity is maximal due to the perspective). More information on the meaning and the optimization of these parameters for the learning can be found in [19]. N is the number of activations of an output neuron required to initiate refresh operations. LTP/LTD is the relative strength (or equivalent weight change) of LTP compared with LTD. The LTP/LTD ratio of 2 used in our simulations ensures that repetitively potentiated synapses converge to their maximum equivalent weight quickly enough for the neuron to become selective to a traffic lane. This can be implemented by adding a current gain of 2 on the LTP input of the neurons.

B. Learning Performances and Robustness

Fig. 11 shows the learning results for the AER data set. The neurons are fully selective to single-lane trajectories after only eight presentations of the sequence, corresponding to approximatively 10 min of real-time traffic. STDP learning and lateral inhibition can be disabled altogether for continuous car detection afterward. Output neurons in the second layer are able to detect cars in four traffic lanes out of six for systems based on GST-PCM synapses and five out of six lanes for systems based on GeTe-PCM synapses, respectively. The sixth lane is never learned because it is at the very right of the retina, and cars activate less pixels over their trajectory than those on other lanes. Over the learned lanes, the average detection rate is above 92%, with no false positive (i.e., neurons fire only once per car, and they never fire for cars passing on a different lane that the one they learned). Some learning statistics are given in Table II: The synaptic-weight update frequency (or postsynaptic frequency) is of the order of 0.1 Hz, and the average presynaptic frequency is around 2 Hz. The average frequencies are similar for the two layers.

TABLE II
LEARNING STATISTICS, OVER THE WHOLE LEARNING DURATION ($8 \times 85 = 680$ s). THE SET PULSE NUMBER INCLUDES BOTH THE WRITE PULSES FOR THE LEARNING AND THE ADDITIONAL PULSES TO REPROGRAM THE EQUIVALENT SYNAPTIC WEIGHT DURING REFRESH OPERATIONS (SEE FIG. 6)

	/device	/device (max)	/device/s	Overall
	GST (2 V/300 ns LTP pulses)			
Read pulses	1265	160,488	1.9	4,975,830,080
SET pulses	106	430	0.16	416,334,080
RESET pulses	4.2	7	0.0062	16,585,048
	GeTe (1.5 V/100 ns LTP pulses)			
Read pulses	1265	160,488	1.86	4,975,848,000
SET pulses	190	740	0.28	748,120,539
RESET pulses	20	37	0.030	79,971,200

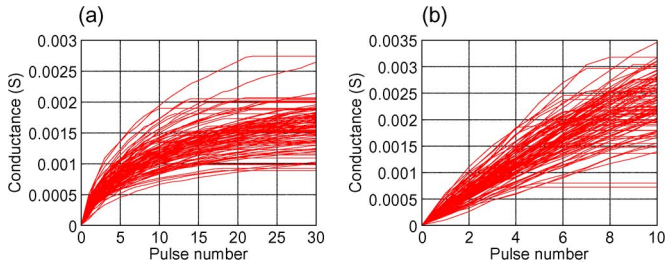


Fig. 12. Simulated variability for (a) GST and (b) GeTe PCM. The plots show the increase in conductance as a function of the number of SET pulses, for 100 different sets of parameters, obtained by applying 20% dispersion (standard deviation of the mean value) from values extracted from fitting.

The frequency of potentiating pulses per device was about 25 times higher than the frequency of RESET pulses for the GST-PCM-based system and about 10 times higher for the GeTe-PCM-based system. This is consistent with the fact that refresh operations were initiated after 30 activations for a given output neuron with GST and only 10 activations for GeTe. As mentioned earlier, this result suggests that the efficiency of the system can be further increased by choosing the right phase-change material with the optimum conductance window to maximize the number of conductance levels reachable with a series of identical SET pulses. The proposed approach of storing maximum synaptic information by crystallization of PCM, with a write frequency of about 2 Hz per PCM device, also reduces the impact of the PCM resistance drift [11], widely ascribed to be a limitation for multilevel PCM programming.

To evaluate the robustness to variability of our neuromorphic system, the simulated neural network included a pessimistic 20% dispersion (meaning that the standard deviation of every parameter is 20% of their mean value) for all the parameters of the PCM asymmetric behavioral model from [24] (G_{min}, G_{max}, α, and β). A sample of PCM characteristics with consecutive identical SET pulses is shown in Fig. 12, for 100 different sets of parameters obtained by adding 20% dispersion from the values extracted from the fitting. In our simulations, the parameters of a PCM device are changed each time it is reset. The 20% dispersion can therefore be seen as including both device-to-device variability and variations between consecutive SET sequences on the same device.

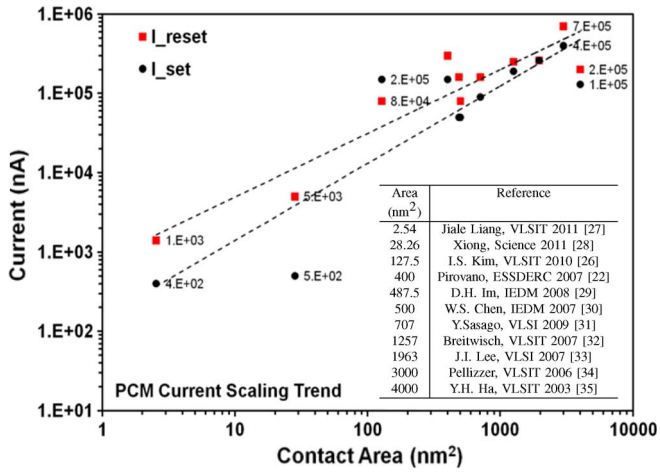


Fig. 13. Scaling trend of RESET and SET current for different PCM technologies (values are extracted from literature). Interpolations are also shown (scaling exponents: ~1.2 for SET and ~0.9 for RESET).

C. Power Consumption and Scalability

Using the learning statistics from Table II, we made a rough estimate of the power consumed for the programming of the GST-PCM devices, i.e.,

$$E_{\text{total}} = E_{\text{SET}} \cdot N_{\text{SET pulses}} + E_{\text{RESET}} \cdot N_{\text{RESET pulses}} \quad (1)$$

$$\text{with} \quad E_{\text{SET}} \approx V_{\text{SET}} \cdot I_{\text{SET}} \cdot t_{\text{SET}} \quad (2)$$

$$\text{with} \quad E_{\text{RESET}} \approx V_{\text{RESET}} \cdot I_{\text{RESET}} \cdot t_{\text{RESET}}. \quad (3)$$

With the SET and RESET voltages and currents measured on our GST devices and $t_{\text{SET}} = 30$ ns, $t_{\text{RESET}} = 50$ ns, $E_{\text{SET}} \approx 121$ pJ, and $E_{\text{RESET}} \approx 1552$ pJ. Using these values, the estimated synaptic power consumption for learning is 112 μW. We did not include the read energy in the calculation as it proved to be negligible. Indeed, in the worst case, the total read energy would be as follows:

$$E_{\text{total read}} = E_{\text{read}_{\max}} \cdot N_{\text{read pulses}} \quad (4)$$

$$\text{with} \quad E_{\text{read}_{\max}} \approx V_{\text{read}}^2 \cdot G_{\max} \cdot t_{\text{read}}. \quad (5)$$

With $V_{\text{read}} = 0.1$ V and $t_{\text{read}} = 10$ ns, we estimated $E_{\text{read}_{\max}} \approx 0.17$ pJ and $E_{\text{total read}} \approx 0.8$ μW. This calculation does not include the CMOS circuitry for the neurons and also neglects the extra energy required for capacitive line charging in the crossbar, which can be significant in modern technology.

Fig. 13 shows that, on average, the current required for RESET and SET scales almost linearly with the PCM area. Table III shows estimations of the synaptic power consumption with several published devices. With extremely scaled PCM technologies, a power consumption as low as 100 nW seems achievable for the ~2 million synapses with continuous STDP learning. If learning would only occur for limited amounts of time, the energy consumption could be orders of magnitude lower.

TABLE III
ENERGY STATISTICS AND SYNAPTIC POWER FOR THE TEST CASE DESCRIBED IN TABLE II, BY USING VOLTAGE AND CURRENT VALUES EXTRACTED FROM LITERATURE

PCM Technology	E_{RESET} (pJ)	E_{SET} (pJ)	Power (μW)
This paper (GST-PCM)	1552	121	112
Jiale Liang, VLSIT 2011 [27]	1.2	0.045	0.056
Xiong, Science 2011 [28]	0.1	0.03	0.02
Pirovano, ESSDERC 2007 [22]	24	4.9	3.6
D.H. Im, IEDM 2008 [29]	5.6	0.9	0.68

With an average SET/RESET frequency per device of the order of 1 Hz, continuous learning for over three years would require an endurance of 10^8 cycles, which is easily achievable with PCM [26]. Performance degradation would be progressive as devices fail due to the high level of fault tolerance of this kind of neural networks [23].

V. CONCLUSION

In this paper, we have demonstrated that PCM devices can be used to devise large-scale synapse-like arrays for neuromorphic systems. A two-layer SNN with about 2 million synapses and 4 million PCM devices has been simulated, showing a complex visual pattern extraction with an average detection rate of 92% and a synaptic power consumption of 112 μW during learning. As mentioned earlier, the extrapolated power consumption for the most recent state-of-the-art devices, if used for the same test case, could be as low as 100 nW. The robustness and scalability of the system were also evaluated.

The low spiking frequency in this type of neural network is remarkable, considering the complex detection task involved, and is a good indicator of the scalability and potentially high efficiency of the association of dynamic vision sensors and SNN compared with the classical synchronous frame-by-frame motion analysis. To go further, we may exploit the novel network topology introduced in [36], with spatially localized neurons, providing similar performances on the car trajectories learning problem with only a tenth of the synapses compared with a fully connected network used in this paper. Requiring only about 130 000 synapses (and thus 260 000 PCM devices), this topology brings the prospect of a practical realization even closer.

Finally, we would like to point out that our proposed STDP implementation is easily adaptable for bipolar memristive devices, provided that gradual increase or decrease in conductance is achievable with successive identical programming pulses. This is the case of several memristive devices, which conductance change characteristic with respect to the number of pulses was modeled with the same behavioral model used in this paper for PCM gradual crystallization [24].

REFERENCES

[1] M. Suri, O. Bichler, D. Querlioz, O. Cueto, L. Perniola, V. Sousa, D. Vuillaume, C. Gamrat, and B. DeSalvo, "Phase change memory as synapse for ultra-dense neuromorphic systems: Application to complex visual pattern extraction," in *Proc. IEDM*, 2011, pp. 4.4.1–4.4.4.

[2] D. Kuzum, R. Jeyasingh, and H.-S. Wong, "Energy efficient programming of nanoelectronic synaptic devices for large-scale implementation of associative and temporal sequence learning," in *Proc. IEDM*, 2011, pp. 30.3.1–30.3.4.

[3] Y. Dan and M.-M. Poo, "Spike timing-dependent plasticity of neural circuits," *Neuron*, vol. 44, no. 1, pp. 23–30, Sep. 2004.

[4] X. Jin, A. Rast, F. Galluppi, S. Davies, and S. Furber, "Implementing spike-timing-dependent plasticity on SpiNNaker neuromorphic hardware," in *Proc. IJCNN*, 2010, pp. 1–8.

[5] M. J. Breitwisch, R. W. Cheek, C. H. Lam, D. Modha, and B. Rajendran, "System for electronic learning synapse with spike-timing dependent plasticity using phase change memory," U.S. Patent 20 100 299 297, Nov. 25, 2010.

[6] D. Kuzum, R. G. D. Jeyasingh, B. Lee, and H.-S. P. Wong, "Nanoelectronic programmable synapses based on phase change materials for brain-inspired computing," *Nano Lett.*, vol. 12, no. 5, pp. 2179–2186, 2012.

[7] G. Snider, "Spike-timing-dependent learning in memristive nanodevices," in *Proc. IEEE Int. Symp. NANOARCH*, 2008, pp. 85–92.

[8] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, 2010.

[9] S. Yu and H.-S. Wong, "Modeling the switching dynamics of programmable-metallization-cell (PMC) memory and its application as synapse device for a neuromorphic computation system," in *Proc. IEDM*, 2010, pp. 22.1.1–22.1.4.

[10] C. Zamarreño-Ramos, L. A. Camuñas-Mesa, J. A. Perez-Carrasco, T. Masquelier, T. Serrano-Gotarredona, and B. Linares-Barranco, "On spike-timing-dependent-plasticity, memristive devices, and building a self-learning visual cortex," *Frontiers Neurosci.*, vol. 5, no. 26, 2011.

[11] A. Fantini, V. Sousa, L. Perniola, E. Gourvest, J. Bastien, S. Maitrejean, S. Braga, N. Pashkov, A. Bastard, B. Hyot, A. Roule, A. Persico, H. Feldis, C. Jahan, J. Nodin, D. Blachier, A. Toffoli, G. Reimbold, F. Fillot, F. Pierre, R. Annunziata, D. Benshael, P. Mazoyer, C. Vallée, T. Billon, J. Hazart, B. De Salvo, and F. Boulanger, "N-doped GeTe as performance booster for embedded phase-change memories," in *Proc. IEDM*, 2010, pp. 29.1.1–29.1.4.

[12] D. Modha and R. S. Shenoy, "Electronic learning synapse with spike-timing dependent plasticity using unipolar memory-switching elements," U.S. Patent 20 100 299 296, Nov. 25, 2010.

[13] B. L. Jackson, D. S. Modha, and B. Rajendran, "Producing spike-timing dependent plasticity in an ultra-dense synapse cross-bar array," U.S. Patent 20 110 153 533, Jun. 23, 2011.

[14] M. Suri, V. Sousa, L. Perniola, D. Vuillaume, and B. DeSalvo, "Phase change memory for synaptic plasticity application in neuromorphic systems," in *Proc. IJCNN*, 2011, pp. 619–624.

[15] F. Alibart, S. Pleutin, O. Bichler, C. Gamrat, T. Serrano-Gotarredona, B. Linares-Barranco, and D. Vuillaume, "A memristive nanoparticle/organic hybrid synapstor for neuro-inspired computing," *Adv. Funct. Mater.*, vol. 22, no. 3, pp. 609–616, 2012.

[16] G.-Q. Bi and M.-M. Poo, "Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type," *J. Neurosci.*, vol. 18, no. 24, pp. 10 464–10 472, Dec. 1998.

[17] J. Lisman and N. Spruston, "Questions about STDP as a general model of synaptic plasticity," *Frontiers Synaptic Neurosci.*, vol. 2, no. 140, Oct. 2010.

[18] G. M. Wittenberg and S. S.-H. Wang, "Malleability of spike-timing-dependent plasticity at the CA3–CA1 synapse," *J. Neurosci.*, vol. 26, no. 24, pp. 6610–6617, Jun. 2006.

[19] O. Bichler, D. Querlioz, S. Thorpe, J. Bourgoin, and C. Gamrat, "Unsupervised features extraction from asynchronous silicon retina through spike-timing-dependent plasticity," in *Proc. IJCNN*, 2011, pp. 859–866.

[20] P. Livi and G. Indiveri, "A current-mode conductance-based silicon neuron for address-event neuromorphic systems," in *Proc. ISCAS*, 2009, pp. 2898–2901.

[21] A. Joubert, B. Belhadj, and R. Heliot, "A robust and compact 65 nm LIF analog neuron for computational purposes," in *Proc. IEEE 9th Int. NEWCAS*, 2011, pp. 9–12.

[22] A. Pirovano, F. Pellizzer, I. Tortorelli, R. Harrigan, M. Magistretti, P. Petruzza, E. Varesi, D. Erbetta, T. Marangon, F. Bedeschi, R. Fackenthal, G. Atwood, and R. Bez, "Self-aligned μtrench phase-change memory cell architecture for 90 nm technology and beyond," in *Proc. ESSDERC*, 2007, pp. 222–225.

[23] D. Querlioz, O. Bichler, and C. Gamrat, "Simulation of a memristor-based spiking neural network immune to device variations," in *Proc. IJCNN*, 2011, pp. 1775–1781.

[24] D. Querlioz, P. Dollfus, O. Bichler, and C. Gamrat, "Learning with memristive devices: How should we model their behavior?," in *Proc. IEEE/ACM Int. Symp. NANOARCH*, 2011, pp. 150–156.

[25] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 $\times$ 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[26] I. Kim, S. Cho, D. Im, E. Cho, D. Kim, G. Oh, D. Ahn, S. Park, S. Nam, J. Moon, and C. Chung, "High performance PRAM cell scalable to sub-20 nm technology with below 4F2 cell size, extendable to DRAM applications," in *Proc. Symp. VLSIT*, 2010, pp. 203–204.

[27] J. Liang, R. Jeyasingh, H.-Y. Chen, and H.-S. Wong, "A 1.4 μA reset current phase change memory cell with integrated carbon nanotube electrodes for cross-point memory application," in *Proc. Symp. VLSIT*, 2011, pp. 100–101.

[28] F. Xiong, A. D. Liao, D. Estrada, and E. Pop, "Low-power switching of phase-change materials with carbon nanotube electrodes," *Science*, vol. 332, no. 6029, pp. 568–570, Apr. 2011.

[29] D. Im, J. Lee, S. Cho, H. An, D. Kim, I. Kim, H. Park, D. Ahn, H. Horii, S. Park, U.-I. Chung, and J. Moon, "A unified 7.5 nm dash-type confined cell for high performance PRAM device," in *Proc. IEDM*, 2008, pp. 1–4.

[30] W. Chen, C. Lee, D. Chao, Y. Chen, F. Chen, C. Chen, R. Yen, M. Chen, W. Wang, T. Hsiao, J. Yeh, S. Chiou, M. Liu, T. Wang, L. Chein, C. Huang, N. Shih, L. Tu, D. Huang, T. Yu, M. Kao, and M.-J. Tsai, "A novel cross-spacer phase change memory with ultra-small lithography independent contact area," in *Proc. IEDM*, 2007, pp. 319–322.

[31] Y. Sasago, M. Kinoshita, T. Morikawa, K. Kurotsuchi, S. Hanzawa, T. Mine, A. Shima, Y. Fujisaki, H. Kume, H. Moriya, N. Takaura, and K. Torii, "Cross-point phase change memory with $4F^2$ cell size driven by low-contact-resistivity poly-Si diode," in *Proc. Symp. VLSI Technol.*, 2009, pp. 24–25.

[32] M. Breitwisch, T. Nirschl, C. Chen, Y. Zhu, M. Lee, M. Lamorey, G. Burr, E. Joseph, A. Schrott, J. Philipp, R. Cheek, T. Happ, S. Chen, S. Zaidr, P. Flaitz, J. Bruley, R. Dasaka, B. Rajendran, S. Rossnage, M. Yang, Y. Chen, R. Bergmann, H. Lung, and C. Lam, "Novel lithography-independent pore phase change memory," in *Proc. Symp. VLSI Technol.*, 2007, pp. 100–101.

[33] J. Lee, H. Park, S. Cho, Y. Park, B. Bae, J. Park, J. Park, H. An, J. Bae, D. Ahn, Y. Kim, H. Horii, S. Song, J. Shin, S. Park, H. Kim, U.-I. Chung, J. Moon, and B. Ryu, "Highly scalable phase change memory with CVD $G_eS_bT_e$ for sub 50 nm generation," in *Proc. Symp. VLSI Technol.*, 2007, pp. 102–103.

[34] F. Pellizzer, A. Benvenuti, B. Gleixner, Y. Kim, B. Johnson, M. Magistretti, T. Marangon, A. Pirovano, R. Bez, and G. Atwood, "A 90 nm phase change memory technology for stand-alone non-volatile memory applications," in *VLSI Symp. Tech. Dig.*, 2006, pp. 122–123.

[35] Y. Ha, J. Yi, H. Horii, J. Park, S. Joo, S. Park, U.-I. Chung, and J. Moon, "An edge contact type cell for phase change RAM featuring very low power consumption," in *VLSI Symp. Tech. Dig.*, 2003, pp. 175–176.

[36] O. Bichler, D. Querlioz, S. Thorpe, J. Bourgoin, and C. Gamrat, "Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity," *Neural Netw.*, 2012, in press.

Olivier Bichler received the M.S. degree in embedded systems from the Ecole Normale Supérieure de Cachan, Cachan, France, in 2009. He is currently working toward the Ph.D. degree in the Embedded Computing Laboratory, CEA-LIST, Gif-sur-Yvette, France.

Manan Suri received the M.Eng. degree in electrical and computer engineering from Cornell University, Ithaca, NY, in 2010. He is currently working toward the Ph.D. degree in the Advanced Memory Technology Laboratory, CEA-LETI, Grenoble, France.

Damien Querlioz (S'06–M'08) received the Ph.D. degree from the Université Paris-Sud, Orsay, France, in 2008.

He is a CNRS Research Scientist with the Université Paris-Sud. He develops new concepts in nanoelectronics relying on bioinspiration.

Dominique Vuillaume received the Habilitation diploma in solid-state physics from the University of Lille, Lille, in 1992.

He is a Research Director with the Centre National de la Recherche Scientifique, Paris, France. He is working on molecular electronics.

Barbara DeSalvo received the Ph.D. degree in microelectronics from the Polytechnic Institute of Grenoble, Grenoble, France.

Since 1999, she has been a Scientist with the CEA-LETI, Gif-sur-Yvette, France.

Christian Gamrat received the M.S. degree in information processing from École Nationale Supérieure d'Électronique et de Radioélectricité, Grenoble, in 1993.

He is a Senior Expert with the CEA-LIST, Gif-sur-Yvette, France, where he is a Chief Scientist.

View publication stats