Phase change memory as synapse for ultra-dense neuromorphic systems: application to complex visual pattern extraction

Manan Suri, Olivier Bichler, Damien Querlioz, Olga Cueto, Luca Perniola, Véronique Sousa, Dominique Vuillaume, Christian Gamrat, Barbara Desalvo

Phase Change Memory as Synapse for Ultra-Dense Neuromorphic Systems: Application to Complex Visual Pattern Extraction

Manan Suri[1], Olivier Bichler[2], Damien Querlioz[2], Olga Cueto[1], Luca Perniola[1], Veronique Sousa[1], Dominique Vuillaume[3], Christian Gamrat[2], Barbara DeSalvo[1]
[1]CEA-LETI-MINATEC, Grenoble, France [2]CEA-LIST, LCE, Gif-sur-Yvette, France [3]CNRS-IEMN, Lille, France

Abstract We demonstrate a unique energy efficient methodology to use Phase Change Memory (PCM) as synapse in ultra-dense large scale neuromorphic systems. PCM devices with different chalcogenide materials were characterized to demonstrate synaptic behavior. Multi-physical simulations were used to interpret the results. We propose special circuit architecture ("*the 2-PCM synapse*"), read, write, and reset programming schemes suitable for the use of PCM in neural networks. A versatile behavioral model of PCM which can be used for simulating large scale neural systems is introduced. First demonstration of complex visual pattern extraction from real world data using PCM synapses in a 2-layer spiking neural network (SNN) is shown. System power analysis for different scaled PCM technologies is also provided.

Introduction: The development of biologically inspired neuromorphic circuits, in order to achieve low power, highly parallel, and fault-tolerant systems, has gained a lot of interest over the last few years [1]. Emergence of new resistive memories, suitable for synapse realization, owing to their simple integration scheme and analog memory capabilities, suggests that the fabrication of neuromorphic chips with a synapse density close to that of the human brain could be achievable. Nevertheless, up to now, substantial work has consisted in demonstrating biologically plausible learning rules, such as spike-time dependent plasticity (STDP), just on individual memristive devices [2-6]. Still there lies a considerable gap between the fields of semiconductor devices, computational neuroscience and system-design before a fully functional large scale hardware neuromorphic system will be realized [7].We choose PCM technology because of the advantages it offers compared to the other resistive memory technologies, such as maturity, scaling capability, high endurance, and good reliability [8].

Experiments: Fig.1 shows the biological synapse and the basic concept of emulating it with PCM. An electronic device emulating the biological synapse should be able to gradually increase (long term potentiation, LTP) or decrease (long term depression, LTD) its conductance in response to neuron spikes [9]. To demonstrate these features, lance-type PCM test devices, with a 100nm-thick phase change layer and 300nm-diameter tungsten plug, were fabricated and characterized. Two different chalcogenide materials were integrated: nucleation-dominated GST and growth-dominated GeTe. The R-I characteristics of GST PCM devices are shown in Fig.2. Fig.3 and Fig.4 show LTP-like conductance variation of PCM devices with GST and GeTe phase change layers, respectively. Nucleation dominated behavior leads to more gradual conductance change in GST, compared to GeTe which shows abrupt conductance change due to growth dominated behavior. The saturation of the conductance programming window in GeTe occurs in less than 1/3rd the total number of potentiating pulses required for GST. Fig.5 shows that a gradual decrease in conductance (or LTD) cannot be obtained with the application of identical pulses to PCM. To this aim, the amplitude of the consecutive pulses should increase progressively (Fig.2). Nevertheless, implementing such pulses with varying amplitudes can lead to practical problems, such as capacitive line charging and high power dissipation in large scale neuromorphic systems involving crossbar arrays. We propose an original solution to such problems in the following sections.

Physical simulations: A 2D axi-symmetrical simulator developed in MATLAB and C++ was used to study the LTP, LTD behavior of the PCM devices. Phase change crystallization physics of the devices were modeled using Eqs.1-2 and a level set numerical method [10]. The first few points in the LTP curves (Fig.3, 4) are crucial in determining the number of intermediate conductance states within a given programming window. In Figs.6, 7, we simulate the sensitivity of conductance variation with respect to changes in nucleation and growth rates. The maximum value of conductance is attained in fewer pulses if the growth or nucleation rate is higher. The shape of the bulk amorphous region created after the initial reset pulse depends upon the values of growth and nucleation rates. A high growth rate (GR=10) leads to strong crystal growth from the amorphous-crystalline interface during the falling edge of the reset pulse, thus distorting the shape of the amorphous mushroom. A low growth rate (GR=0.1) leads to much symmetric mushroom shape of the amorphous region after the reset pulse. After the application of the 1st pulse, conductance is more sensitive to changes in the nucleation rate compared to growth.

The 2-PCM synapse: To overcome the problem of abrupt LTD (Fig.5). We propose a new energy efficient synapse circuit consisting of two identical PCM devices (*"The 2-PCM synapse"*) with a realistic and simplified programming pulse scheme (Fig.9). One PCM device (*LTP*) has a positive current contribution, while the other PCM device (LTD) contributes negatively towards the output CMOS neuron current. In the *"2-PCM synapse"*, crystallizing the *LTP PCM* device produces synaptic LTP-like effect, while crystallizing the *LTD PCM* device produces synaptic LTD-like effect. Since gradual crystallization can be obtained by applying simple identical pulses (Fig.3, 4) our approach simplifies the pulse schemes for implementing learning in large scale neuromorphic systems. Moreover, unlike previously demonstrated PCM synapse [6] our "*2-PCM synapse*" is highly energy efficient, as majority synaptic conductance changes (or weight update) events are based on crystallization and not on amorphization of the PCM devices. As shown in Fig.2 and Fig.17, the set (or potentiating) events would consume considerably less amount of energy compared to reset (or depressing) events in PCM technology.

Neural Network & PCM Behavioral Model: To demonstrate complex pattern extraction from real world data a two layer feed-forward spiking neural network (Fig.9) was simulated using a special purpose C++ event-based simulator. The neuron model used is the standard leaky integrate and fire (LIF) [11]. The PCM synapses in the

network were modeled using Eq.3 with the parameters extracted from the fitting of LTP curves in Fig.3, 4. Address-Event Representation (AER) data (Fig.10) recorded with 128x128 pixel silicon retina [13] was used as the input for our PCM neural network. Each of the 128 x 128 pixels in Fig.9 is connected through two synapses to every neuron in layer 1. Likewise each neuron in layer 1 is connected to every neuron in layer 2 with a single synapse, leading to a total of 1,966,680 synapses and thus 3,933,360 PCM devices (2 PCM/synapse). The goal of the neural network is to detect cars passing in different lanes on a freeway in an unsupervised way.

Learning Rule & Programming Scheme: We used a modified form of the biological STDP learning rule (Fig.11). The modified STDP enhances simplicity in programming the synapses. The learning is competitive and lateral inhibition is implemented [12]. When read with a voltage pulse, the two PCM devices have an opposite contribution to the neuron integration (Fig.12). When the integration of an output neuron reaches the firing threshold, the neuron transmits an event to the next layer and a feedback pulse on each of its two input lines corresponding to LTP and LTD devices (Fig.13). A write mode signal is propagated to all the input neurons such that if the input neuron was activated during the LTP time window (Fig.11), a write pulse is emitted and its interaction with the feedback pulse increases the conductivity of the LTP device, resulting in the potentiation of the equivalent synapse. The feedback pulse alone only increases the conductivity of the LTD device, thus depressing the equivalent synapse. A systematic reset scheme (Fig.14), allows reducing the conductivity of the PCM devices by retaining the weight of the synapse (Fig.15).

Results & Discussion: Fig.17 shows the learning results for the AER dataset. To demonstrate the variability robustness of our neuromorphic system we implemented 20% dispersion on standard deviation of all the parameters in Eq.3. Output neurons in the 2nd layer are able to detect cars in 4 traffic lanes out of 6 for systems based on GST-PCM synapses, and 5 out of 6 lanes for systems based on GeTe-PCM synapses, respectively. The frequency of potentiating pulses per device was about 25 times higher than the frequency of reset pulses for GST-PCM based system, and about 10 times higher for GeTe-PCM based system. Indeed, 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. Finally, note that the proposed approach of storing maximum synaptic information by crystallization of PCM, with a write frequency of about 2Hz per PCM device, reduces the impact of the PCM resistance-drift [8], widely ascribed to be a limitation for multilevel PCM programming.

Conclusions: We have developed a unique methodology to use PCM devices as energy-efficient synapses in large scale neuromorphic systems. Advanced electrical characterization, behavioral modeling and circuit level simulations allow us to show a spiking neural network with about 4 million synapses, capable of complex visual pattern extraction with an average detection rate of 92%, and a system power consumption of 112μW for learning.

References

[1] G.S Snider, IEEE/ACM NANOARCH 2008.
[2] S.H.Jo et al., Nano Letters 2010, 10, 1297.
[3] S.Yu et al., IEDM 2010.
[4] A.S.Oblea et al., IJCNN 2010.
[5] S.Yu et al., IEEE Trans. on El. Dev., June '11.
[6] D.Kuzum et al, Nano Letters, June '11.
[7] C.Zamarreno et al., Frontiers in Neuroscience,Vol.5,Art 26, March '11.
[8] A.Fantini et al., Techn. Dig. of IEDM 2010.
[9] X.L.Zhang, Nature Precedings, 21st March, 2008.
[10] A.Gliere et al., SISPAD 2011.
[11] G.Indiveri et.al., Frontiers in Neuroscience,Vol.5, Art 73, May 2011.
[12] O.Bichler et al., IEEE IJCNN 2011.
[13] P. Lichtsteiner et al., IEEE J. Solid-State Circuits, Vol. 43, 2008.
[14] http://sourceforge.net/apps/trac/jaer/wiki/AER%20data
[15] G.Q. Bi et al., J. Neurosci. 1998, 18, 10464-10472.

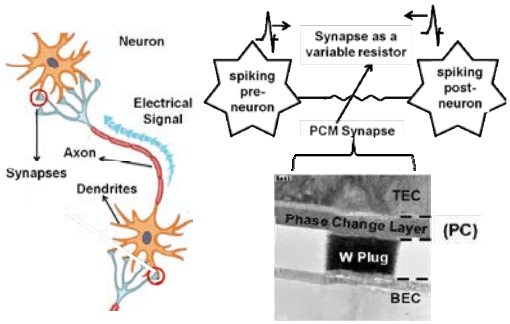


Fig.1 Illustration of biological synapse and concept of using PCM as synapse in neural circuits.

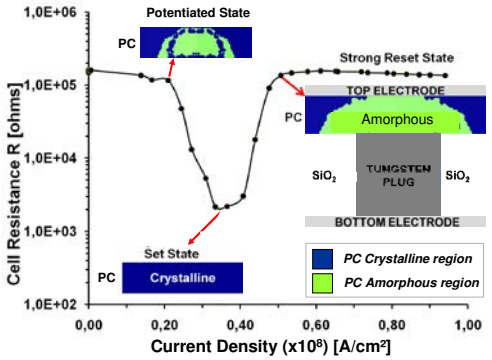


Fig.2 R-I characteristic of GST PCM cell. Potentiated, Set and Reset states with corresponding morphology of the Phase Change layer are shown.

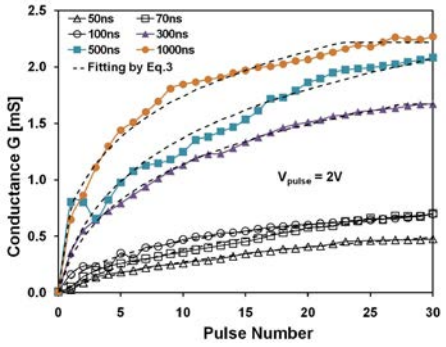


Fig.3 Experimental Long Term Potentiation (LTP) characteristics of GST PCM. For each curve, first a reset pulse (7V, 100ns) is applied followed by 30 consecutive identical potentiating pulses.

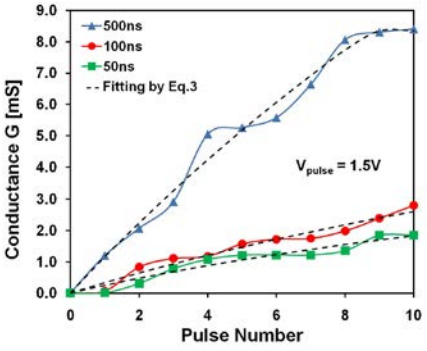


Fig.4 Experimental LTP characteristics of GeTe PCM. For each curve, first a reset pulse (7V, 100ns) is applied followed by 10 consecutive identical potentiating pulses.

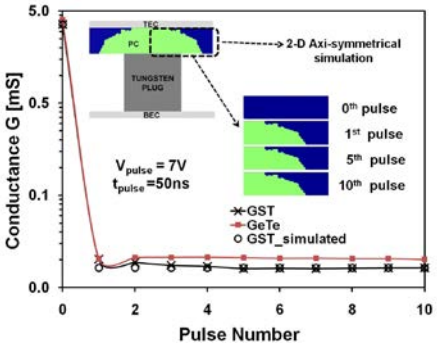


Fig.5 Experimental LTD characteristics of GST and GeTe PCM. Simulations of the GST PC layer morphology with consecutive identical reset pulses shown.

Equations for Physical Simulations

(1) Growth Velocity "V"

$$V = \gamma d\left[1-\exp\left(\frac{-\Delta G_v}{RT}.\frac{M}{\rho}\right)\right]$$

(2) Overall Nucleation Rate "I"

$$I = N_a \gamma O_n Z \exp\left(-\frac{\Delta G^*}{kT}\right)$$

N_a: Nucleation sites/m^3, γ: atomic vibration freq.
ΔG: Free Energy, Z: Zeldovitch parameter
O_n: Number of atoms at critical nucleus surface
M: molar mass d: Inter-atomic distance
ρ: Volumic mass, ΔGv: Diff. in Gibbs free energy of the amorph. and crystalline phase
α, β: fitting parameters

Behavioral Model Equation

(3) Conductance "G" Change:

$$\frac{dG}{dt} = \alpha \exp\left(-\beta \frac{G - G_{min}}{G_{max} - G_{min}}\right)$$

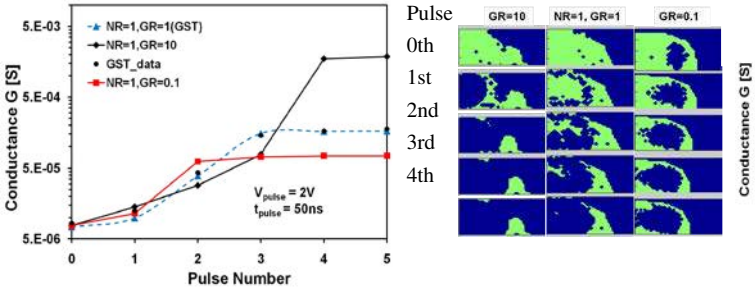


Fig.6 Simulated LTP curves while fixing the nucleation rate (NR) and varying the growth rate GR compared to GST (reference: GR=1, NR=1). Corresponding simulations of PC layer morphology is shown (0th pulse: reset; 1st-5th: potentiating).

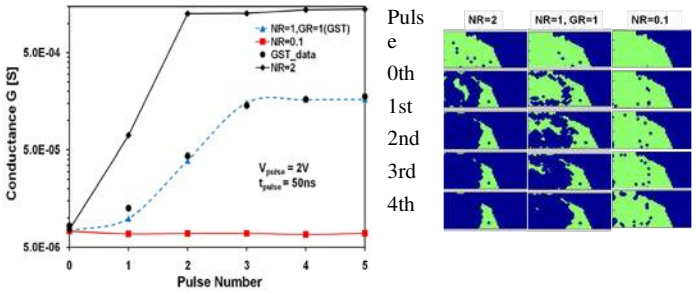


Fig.7 Simulated LTP curves while fixing the growth rate (GR=1) and varying the nucleation rate (NR) compared to GST (reference: NR=1, GR=1).

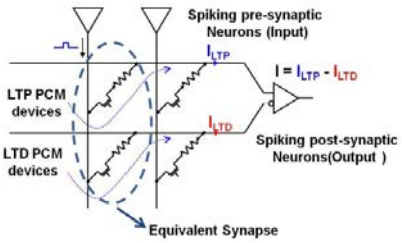


Fig.8 Circuit schematic of the "2-PCM synapse".

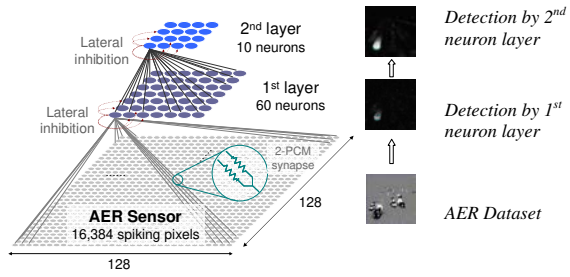


Fig.9 Topological overview of the spiking neural network

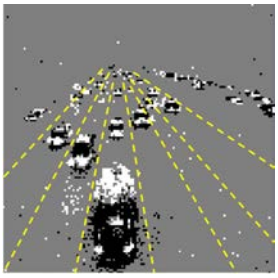
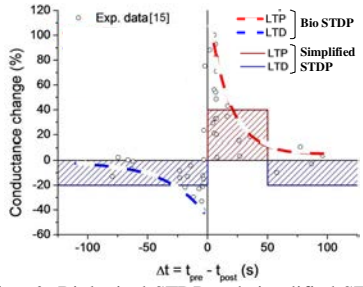


Fig.10 AER dataset for pattern recognition. Represents 6 car lanes [14].



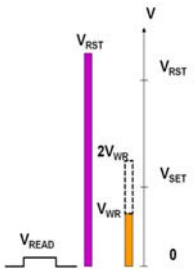


Fig.11 *Left:* Biological STDP and simplified STDP. In the simplified rule, a synapse receiving a post-synaptic spike with no pre-synaptic spike in the LTP window undergoes a LTD regardless of the existence of a pre-synaptic spike. *Right:* Write, reset and read pulses used in Figs. 12, 13, 14.

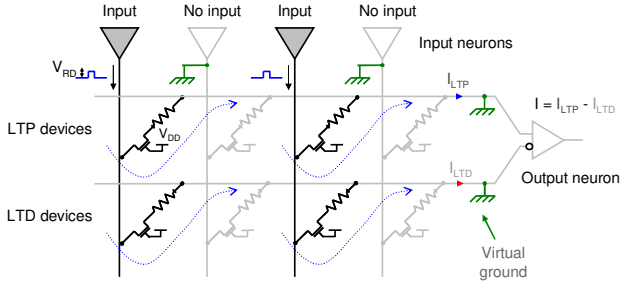


Fig.12 Read circuit scheme. Current from both LTP and LTD PCM devices is integrated in the output neuron, with a positive and negative contribution, respectively.

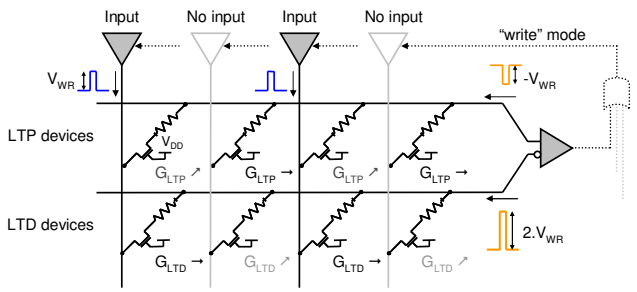


Fig.13 Write circuit scheme based on the simple STDP rule. For a specific PCM, G ↗ denotes an increase while G → denotes no change in conductance.

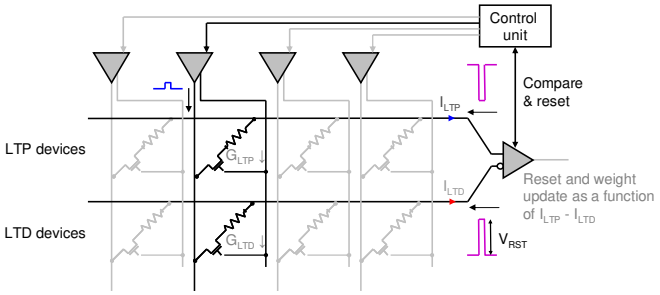


Fig.14 Reset circuit scheme. For a specific PCM resistor, G ↓ denotes a decrease in conductance.

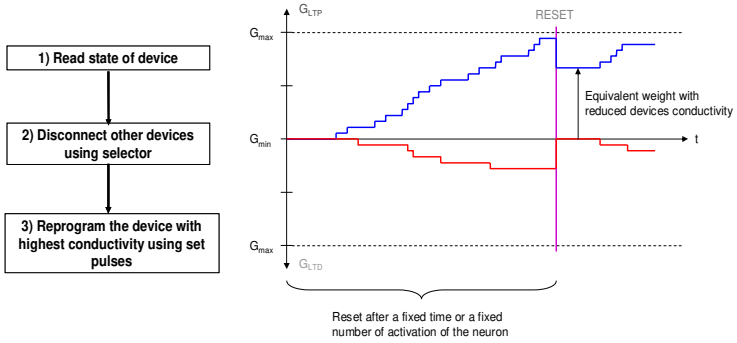
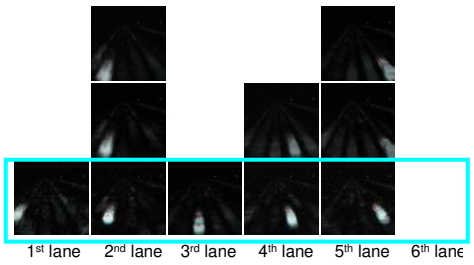


Fig.15 *Left:* Reset methodology. *Right:* Conductance G variation of PCM devices before and after a reset sequence.



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.16 *Left*: Cars detection for the 6 different lanes by the 2nd layer 10 output neurons. Each block shows the sensitive map of one output neuron. Up to down the detection rate increases. Lane 2 and 5 are learnt by 3 neurons each, while lane 6 is not learnt by any neuron. *Right:* Detection % of 5 best neurons for each lane.

Table 1 Pulse statistics for the learning test case shown in **Fig 16**.

Learning Statistics for the PCM-based Neuromorphic Circuit		
Total Synapses: 1,966,680 [= 1st Layer (2*128*128*60) 2nd (60*10)]		
Total PCM cells: 3,933,360 [= 2* Total Synapses (LTP and LTD)]		
Total learning duration = 680 s (same for both GST/GeTe PCM)		
Quantity	Values for	
	GST-PCM	GeTe-PCM
Total read pulses	4,975,830,080	4,975,848,000
Total set pulses	416,334,080	748,120,539
Total reset pulses	16,585,048	79,971,200
Read pulses frequency per PCM [/sec]	1.9	1.86
Set pulses frequency per PCM [/sec]	0.16	0.28
Reset pulses frequency per PCM (/sec]	0.0062	0.030

Table 2 Energy statistics for the test case described in Table 1, by using voltage (V)and current (I) values extracted from literature and the energy equations.

PCM Technology	Ereset (pJ)	Eset (pJ)	System power (µW)
This paper (GST-PCM)	1552	121	112
Jiale,VLSI-2011	1.2	0.045	0.056
F.Xiong, Science-2011	0.1	0.03	0.02
Pirovano, ESSDERC-07	24	4.9	3.6
D.H.Im, IEDM-2008	5.6	0.9	0.68

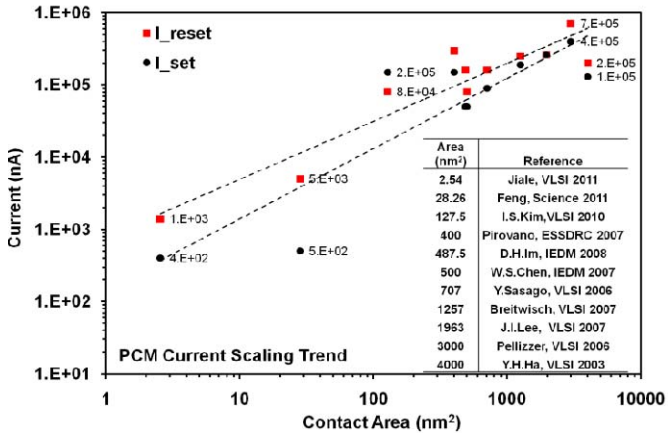


Fig.17 Scaling trend of Reset and Set current for different PCM technologies. Set current values are extracted from literature. Interpolations are also shown

Energy Calculation Equations

Eset ~ Vset * Iset * t_{pulse} (t_{pulse}=30ns);

Ereset ~ Vreset * Ireset * t_{pulse} (t_{pulse} =50ns).

Etotal =(Eset * total set pulses) + (E reset * total reset pulses)

System power = Etotal / total duration of learning