Performance Walls in Machine Learning and Neuromorphic Systems

Shantanu Chakrabartty
Department of Electrical and Systems Engineering
Washington University in St. Louis
St. Louis, MO, USA
shantanu@wustl.edu

Gert Cauwenberghs
Department of Bioengineering
University of California San Diego
La Jolla, CA, USA
gert@ucsd.edu

***Abstract*—At the fundamental level, an energy imbalance exists between training and inference in machine learning (ML) systems. While inference involves recall using a fixed or learned set of parameters that can be energy-optimized using compression and sparsification techniques, training involves searching over the entire set of parameters and hence requires repeated memorization, caching, pruning, and annealing. In this paper, we introduce three "performance walls" that determine the training energy efficiency, namely, the *memory-wall*, the *update-wall*, and the *consolidation-wall*. While the emerging compute-in-memory ML architectures can address the memory-wall bottleneck (or energy-dissipated due to repeated memory access) the approach is agnostic to energy-dissipated due to the number and precision required for the training updates (the update-wall) and is agnostic to the energy-dissipated when transferring information between short-term and long-term memories (the consolidation-wall). To overcome these performance walls, we propose a learning-in-memory (LIM) paradigm that prescribes ML system memories with *metaplasticity* and whose thermodynamical properties match the physics and energetics of learning.**

***Index Terms*—Machine Learning, Training, Neuromorphic Systems, Energy Efficiency, Memory, Thermodynamics**

I. Introduction

There exists an imbalance between the energy budget required to train a machine learning (ML) system versus the energy budget required for performing inference using the same ML system. As an example, a recent study reported that the carbon footprint of training a single neural architecture search on a 213 million parameter deep neural network (DNN) could be five times the carbon footprint of a US car over its entire lifetime [1], whereas the same size DNN can perform inference at a significantly lower energy-budget. At the fundamental level, this imbalance arises because inference involves recall using a fixed or learned set of parameters that can be optimized through compression, sparsification, and compute-in-memory (CIM) techniques, whereas learning and training involve searching over a large set of parameters and hence require repeated memorization, caching, and pruning. Thus, energy dissipation is dominated by the energy cost of moving data between slow and vast external memory, and fast and sequential processing in the Turing formalism. We quantify this bottleneck using three *performance-walls*, namely: the *memory-wall*, the *update-wall*, and the *consolidation-wall* as shown in Fig. 1(a). The *memory-wall* [8] arises because of energy-dissipation due to repeated memory access (depicted in Fig. 1(b)) and can be addressed using CIM architectures [4], [5]. However, the CIM approach is agnostic to the energy bottleneck due to the number/precision of parameter updates, referred to as the *update-wall* shown in Fig. 1(c), or the energy bottleneck incurred when transferring data between short-term memories (cache, DRAM) and long-term memories (non-volatile SSD), referred to as the *consolidation-wall* shown in Fig. 1(d).

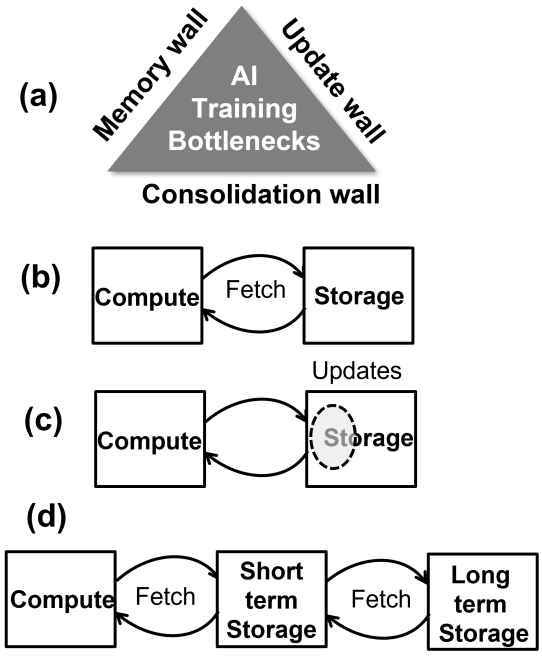

Fig. 1. (a) Three main challenges or performance walls that affect the energy-efficiency of training machine learning systems: (b) *Memory-wall*; (c) *Update-wall*; and (d) *Consolidation-wall*.

Can inspiration from neurobiology provide cues on how to address these performance bottlenecks? While synaptic computations have inspired the CIM paradigm [6], there is growing evidence that biological synapses are inherently a

This work is supported by the National Science Foundation with research grant FET-2208770.

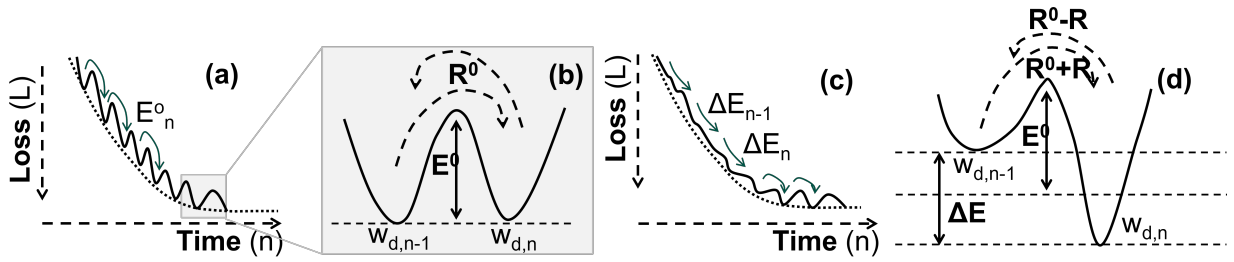

Fig. 2. (a) Learning using conventional memory where states are separated from each other by energy-barriers. (b) A two-state potential barrier between two wells. (c) Adaptive memory profile [7] where the energy-barrier height is modulated. (d) Thermodynamics of memory state transitions during learning.

complex high-dimensional dynamical system in itself [10], [11] as opposed to the simple, static storage unit that is typically assumed in standard neural networks. This neuromorphic viewpoint has been experimentally validated by *metaplasticity* observed in biological synapses [12], [13] where the synaptic plasticity (ease of update) has been observed to vary depending on age and task-specific information accumulated during learning/training. Metaplasticity can be physically emulated in artificial synapses [7] resulting in energy-efficient training and higher memory capacity. In this paper, we argue the benefits of metaplasticity from a thermodynamic and an information-theoretic point of view and analytically show its connection to the *update-wall* and the *consolidation-wall*. Using the analytical expressions we also show how to approach the fundamental energy limits of training by adjusting the plasticity of the memory devices to match the physics of the learning. We refer to this paradigm as *Learning-in-Memory* (LIM).

II. Thermodynamics of Learning and Memory Retention

At the physical level, the memory elements used for storing ML parameters $w_d \in \mathbb{R}, d = 1, .., D$ are static in nature where each of the memory states is separated from each other by a physical energy barrier E^0, as shown in Fig. 2(b). This energy barrier is generally chosen to be high enough to prevent parameter leakage due to thermal fluctuations, especially, during inference when the memory needs to be non-volatile. Therefore, an ML training algorithm that adapts the stored weights in quantized steps ($\ldots$, $w_{d,n-1}, w_{d,n}, w_{d,n+1}$, $\ldots$) so as to minimize some system-level loss-function L consumes energy to overcome the energy-barrier E^0 for each of the parameter/memory updates, despite the overall descending energy profile for L in descending steps ($\ldots$, $\Delta E_{n-1}, \Delta E_n, \Delta E_{n+1}$, $\ldots$) towards convergence. In most memory implementations, the energy incurred per update to traverse this barrier E^0 is irreversibly lost and dissipated as heat. While energy per memory update could be relatively small (for example 13fJ for a single RRAM write [5]), when combined with the total computational requirements for training (which could be greater than 10^{23} for transformer networks), the energy cost could be prohibitively high. As a result, conventional memory elements are unable to exploit the dynamics of the learning process to optimize its energy efficiency, leading to the *update-wall*. In [7] we introduced a LIM paradigm where the energy barrier height E_n^0 separating consecutive memory states during the training process can be adapted, as shown in Fig. 2(c). We showed that by matching the memory retention rates to the process of weight decay used in ML training, energy efficiency could be significantly improved. Here, we derive analytic expressions that will connect the thermodynamics of learning and memory retention with the celebrated Landauer limit.

A. Landauer thermodynamic energy limit

Landauer [2], [3] established physical limits on the energy efficiency of computation based on the thermodynamics of dissipative systems implementing the computation in an irreversible manner, incurring a loss of energy in exchange for negentropy at a rate of $kT\log(2)$ per bit, roughly 3×10^{-21} J or 18 meV at room temperature. However, in practice, states in typical digital systems are separated from each other by energy-barrier greater than $10^6\, kT$ due to memory retention requirements. Also, the Landauer limit is achievable only under adiabatic conditions where the state-transition rate is asymptotically near zero. Below we derive analytic expressions that will show how the memory energy-barrier (or the metaplasticity of a synapse) can be modulated such that training proceeds at a finite rate (information rate) while asymptotically achieving reliable parameter storage (for inference).

B. Energy-barrier profile and energy dissipation

Fig. 2(b) shows a potential-well configuration with two memory states $w_{d,n-1}$ and $w_{d,n}$ separated by an energy-barrier E_n^0. At steady-state equilibrium, reached at convergence in the loss L, the potential well for both states are symmetrical with respect to the barrier-height, such that the state update rate (or equivalently, transition probability) R_n^0 is symmetrical, identical from $w_{d,n-1}$ to $w_{d,n}$ and vice versa, and given by

$$R_n^0 = R_{max} \exp\left(-\frac{E_n^0}{kT}\right). \tag{1}$$

R_{max} is a process and device-specific constant that corresponds to the spontaneous rate in the absence of a barrier. For reliable storage, the transition rate (1) should be $R_n^0 \to 0$ implying a barrier height $E_n^0 \gg kT$ such as $> 10^6 kT$. However, during training, away from equilibrium, an energy

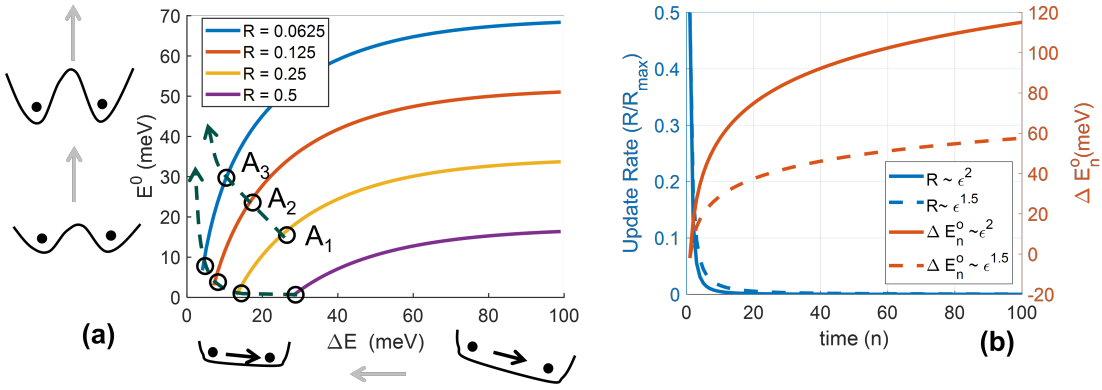

Fig. 3. (a) Energy barrier-height versus gradient energy for different normalized update rates R; and (b) Normalized Update-rate and memory barrier height for a specific learning-rate profile.

differential ΔE_n applies across the potential wells as shown in Fig. 2(d), and the rates across the barriers differ from each other as given by

$$\begin{aligned} R_{n-1\to n} &= R_n^0 + R_n = R_{max}\exp\left(-\frac{E^0_{n-1\to n}}{kT}\right) \\ R_{n-1\leftarrow n} &= R_n^0 - R_n = R_{max}\exp\left(-\frac{E^0_{n-1\leftarrow n}}{kT}\right) \end{aligned} \quad (2)$$

where $E^0_{n-1\to n} = E^0_{n-1\leftarrow n} - \Delta E_n$, and with a net transition rate in the learning update given by the two-sided differential $2R_n$. Hence, the parameter R_n quantifies the *update-wall*. Note that $0 \le R_n \le R_n^0$. From Eqn. (2) one obtains

$$R_n = R_{max}\exp\left(-\frac{E_n^0}{kT}\right)\cdot\frac{\exp\left(-\frac{\Delta E_n}{kT}\right)-1}{\exp\left(-\frac{\Delta E_n}{kT}\right)+1}. \quad (3)$$

Eqn. (3) leads to

$$E_n^0 = kT\log\left[\frac{R_{max}}{R_n}\cdot\frac{\exp\left(-\frac{\Delta E_n}{kT}\right)-1}{\exp\left(-\frac{\Delta E_n}{kT}\right)+1}\right] \quad (4)$$

which specifies how the barrier-height E_n^0 should change with respect to time n for a given information-rate R_n and extrinsic energy ΔE_n.

Eqn. (4) could now be connected to the physics of learning by noting that the extrinsic energy ΔE_n required to support the update-rate R_n is provided by the change in network energy (or loss) function ΔL_n in Fig. 2(c). The change in loss-function can be written in terms of the change in the parameters $\Delta w_{d,n}$ as

$$\Delta L_n = \sum_{d=1}^{D}\left(\frac{\partial L_n}{\partial w_{d,n}}\right)\Delta w_{d,n} \quad (5)$$

where the factor D accounts for the number of training parameters. If the parameter w_d is adapted according to the steepest gradient-descent rule (or equivalently the Lyapunov's criterion) then

$$\Delta w_{d,n} = -\epsilon_n\frac{\partial L_n}{\partial w_{d,n}} \quad (6)$$

where ϵ_n is the learning rate at time instant n. Substituting Eqn. (6) into Eqn. (5) leads to an expression that relates the extrinsic energy applied to each memory element to square magnitude gradient update energy as

$$\Delta E_n = -kT\epsilon_n\left(\frac{\partial L_n}{\partial w_{d,n}}\right)^2$$

where the thermal energy kT has been introduced as a normalization factor. Assuming that the gradient is bounded from above or, without loss of generality, $\left(\frac{\partial L_n}{\partial w_{d,n}}\right)^2 \le 1$, Eqn. (4) leads to a general relationship

$$E_n^0 \ge kT\log\left[\frac{R_{max}}{R_n}\cdot\frac{\exp(\epsilon_n)-1}{\exp(\epsilon_n)+1}\right] \quad (7)$$

that connects the memory energy barrier-height modulation (or synaptic metaplasticity) to two learning parameters: (a) the update-rate R_n quantifying the *update-wall*; and (b) the learning-rate ϵ_n quantifying the *consolidation-wall*. We will further study this relationship in the next section with some simple case studies.

III. Theoretical Results

Fig. 3(a) plots the barrier-energy profile E_n^0 according to Eqn. (4) as a function of the energy gradient ΔE_n for different values of update rates R_n. Note that during the process of learning, computation proceeds at a fixed rate $R_n > 0$ but asymptotically as $n \to \infty$, $R_n \to 0$, $\epsilon_n \to 0$, and $E_n^0 \overset{>}{\to} kT\log[R_{max}/R_n^0]$ which is the equilibrium potential-well configuration in Fig. 2(b). Different learning algorithms will traverse this space along different trajectories, as shown in Fig. 3(a). In the process, the total energy dissipated is given by $E_{diss,n} = E_n^0 + \Delta E_n$ per computational step where the assumption is the external energy ΔE_n that is supplied cannot be recovered. Initially, at $n = 0$ the learning-rate and barrier-height are set such that the update-rate proceeds at a maximum value or $R_0 = R_{max}$. To satisfy the asymptotic conditions $E_n^0 \to kT\log[R_{max}/R_n^0]$ and $\epsilon_n \to 0$ as $n \to \infty$, the update-rate R_n could proceed according to $R_n \approx R_{max}\epsilon_n^{1+\alpha}$ where $0 < \alpha < 1$.

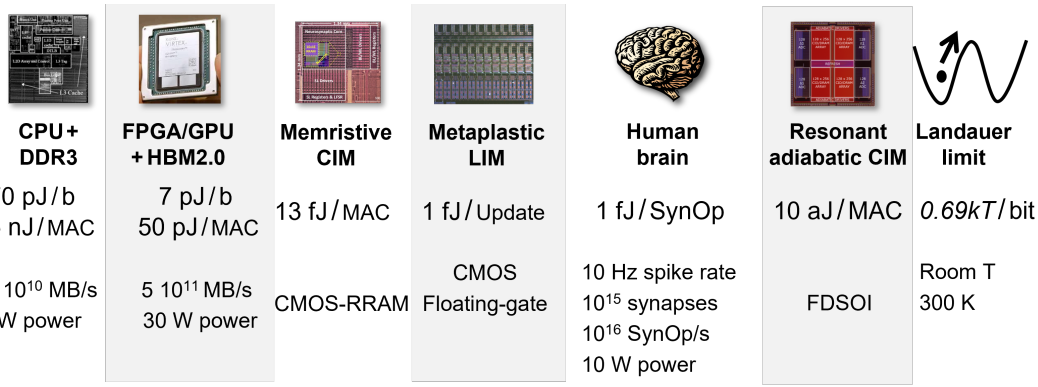

Fig. 4. Energy efficiency spectrum of computation interfacing to memory, relative to the thermodynamic limit.

Fig. 3(b) plots the update-rate profile and barrier-height profile when the learning-rate follow the asymptotics $\epsilon_n = \mathcal{O}(1/n)$. As expected, the barrier-height initially is low and then increases to support the memory retention requirements.

IV. Discussions and Conclusions

The thermodynamic analysis presented in this paper and in particular Eqn. (7) describes how the memory energy barrier-height is connected to two important parameters: (a) the parameter-update rate; and (b) the learning-rate, both which determine two of the three performance walls, namely, the *update-wall* and the *consolidation-wall*. For instance, the *update-wall* is reflected in the profile of the update-rate R_n for each of the parameters and Eqn. (7) shows how a specific update-profile R_n can be achieved by modulating the barrier-profile according to Fig. 3(a), hence, the *learning-in-memory* paradigm. Similarly, the learning-rate ϵ_n determines the *consolidation-wall*. Several adaptive synaptic models have been proposed [10], [11] that show how a specific learning-rate profile can lead to optimal information transfer-rate between short-term and long-term memories. In the LIM paradigm, the memory energy-barrier can be modulated to also control ϵ_n according to Eqn. (7). Energy-barrier modulation supporting the LIM paradigm could be implemented in a variety of physical substrates using emerging memory devices. For instance, recently, we reported a dynamic memory device [7] that could also be used to modulate the memory retention profile and could be an attractive candidate to implement the LIM paradigm. However, note that to approach the fundamental energy limits of training/learning one would need to address all three performance walls. Fig. 4 summarizes some of the computation-to-memory interfaces that could address this. Compute-in-memory (CIM) alternatives where the computation and memory are vertically integrated in massively parallel, distributed architecture (Fig. 4, center) offer substantially greater computational bandwidth and energy efficiency in memristive neuromorphic cognitive computing [5] approaching the nominal energy efficiency of synaptic transmission in the human brain [6]. Resonant adiabatic switching techniques in charge-based CIM [9] further extend the energy efficiency by recycling the energy required to move charge by coupling the capacitive load to an inductive tank at resonance, providing a path towards efficiencies in cognitive computing superior to biology and, in principle, beyond the Landauer limit by overcoming the constraints of irreversible dissipative computing. It is an open question whether the learning-in-memory energy bounded by Eqn. (7) could also be at least partially recovered through principles of adiabatic energy recycling.

References

[1] K. Hao, "Training a Single AI Model Can Emit as Much Carbon as Five Cars in Their Lifetimes," MIT Technology Review, June 6, 2019. https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

[2] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Research & Develpment*, vol. 5, pp 183–91, 1961.

[3] R. Landauer, "Information Is Physical," *Physics Today*, vol. 44, pp. 23–9, 1991. doi:10.1063/1.881299

[4] S. Chakrabartty, G. Cauwenberghs, "A Sub-microwatt Analog VLSI Trainable Pattern Classifier," IEEE J. Solid-State Circuits, vol. 42 (5), 2007.

[5] W. Wan, R. Kubendran, C. Schaefer, S.B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S.P. Wong, and G. Cauwenberghs, "A Compute-in-Memory Chip Based on Resistive Random-Access Memory," *Nature*, vol. 608, pp. 504–512, 2022.

[6] G. Cauwenberghs, "Reverse Engineering the Cognitive Brain," *Proc. Natl. Acad. Sci. (PNAS)*, vol. 110 (39), pp. 15512-15513, 2013.

[7] D. Mehta, M. Rahman, K. Aono and S. Chakrabartty, "An Adaptive Synaptic Array Using Fowler–Nordheim Dynamic Analog Memory," *Nature Communications*, vol. 13, pp. 1670, 2022.

[8] M. Horowitz, "1.1 Computing's Energy Problem (and What We Can Do About It)," in 2014 IEEE Int. Solid-State Circuits Conf. (ISSCC'2014)," Feb. 2014.

[9] Karakiewicz, R., R. Genov, and G. Cauwenberghs, "1.1 TMACS/mW Fine-Grained Stochastic Resonant Charge-Recycling Array Processor," *IEEE Sensors Journal*, vol. 12 (4), pp. 785-792, 2012.

[10] Fusi S, Drew PJ, Abbott L, "Cascade models of synaptically stored memories." *Neuron*, vol. 45 (4), pp. 599–611, 2005.

[11] Kirkpatrick, J. et al. "Overcoming Catastrophic Forgetting in Neural Networks," *Proc. Natl. Acad. Sci.*, vol. 114 (13), pp. 3521-3526, 2017.

[12] Zenke, F., A., Everton J., and Gerstner, W., "Diverse Synaptic Plasticity Mechanisms Orchestrated to Form and Retrieve Memories in Spiking Neural Networks." *Nature Communications*, vol. 6, April 2015, doi:10.1038/ncomms7922.

[13] Yang G, Pan F, Gan WB, "Stably Maintained Dendritic Spines Are Associated with Lifelong Memories," *Nature*, vol. 462 (7275), pp. 920–924, 2009.