Capacity Optimization of Emerging Memory Systems: A Shannon-Inspired Approach to Device Characterization

Jesse H. Engel[1,2], S. Burc Eryilmaz[2], SangBum Kim[3], Matthew BrightSky[3], Chung Lam[3], Hsiang-Lan Lung[4], Bruno A. Olshausen[1], and H.-S. Philip Wong[2]

[1]Redwood Center for Theoretical Neuroscience, UC Berkeley, Berkeley, CA, 94720, [*]E-mail: jhengel@stanford.edu

[2]Dept. of Electrical Engineering and Center for Integrated Systems, Stanford University, Stanford, CA, 94305

[3]IBM Research, T.J. Watson Research Center, Yorktown Heights, NY, 10598

[4]Macronix International Co., Ltd., Emerging Central Lab, 16 Li-Hsin Road, Hsinchu Science Park, Taiwan

Abstract

Traditional approaches to memory characterize the number of distinct states achievable at a given Raw Bit Error Rate (RBER). Using Phase Change Memory (PCM) as an example analog-valued memory, we demonstrate that measuring the mutual information allows optimal design of read-write circuits to increase data storage capacity by 30%. Further, we show the framework can be used for energy efficient memory design by optimizing simulations of a 1Mb memory array to consume 32% less energy/bit. This work provides an information-theoretic framework to guide the design and characterization of other analog-valued emerging memory such as RRAM and CBRAM.

Introduction

The number of data generating devices connected to the internet is expected to grow exponentially to 20 billion by 2020 [1]. Storing and analyzing the data deluge from this 'internet of things' requires new energy-efficient and high-density memory systems. Multi-level storage has successfully achieved high densities in emerging memories such as Flash, PCM, and RRAM, with systems realizing as many as 16 states/cell [2, 3].

Multi-level cells function by modulating an analog-valued property such as cell resistance or threshold voltage and discretizing the output. The dynamics of setting these properties are stochastic due to atomic kinetics and nanoscale fabrication variations, leading to higher RBER with increasing states/cell. Robust memory storage increasingly requires strong Error Correcting Codes (ECCs) to reach low ($\sim 10^{-15}$) Uncorrected Bit Error Rates (UBER) [4].

The Shannon Capacity (C) of a memory cell is the theoretical maximum bits/device that can be robustly stored and recalled with the help of an ECC [5] (Fig. 1). As modern coding techniques such as Low-Density Parity Check (LDPC) codes can now approach the Shannon Capacity limit [6], and are employed in noisy MLC-Flash memories [7], the Shannon Capacity is a valuable metric to improve performance of the memory cell/controller/ECC system.

From a Shannon perspective, storage and recall in a memory cell can be thought of as an input write voltage pulse distribution, P(V), traveling through a noisy memory channel, P(R|V), and resulting in an output resistance read distribution, P(R) (Fig. 2). Shannon showed that the capacity of such a channel is given by the maximum mutual information over the possible inputs [5]:

$$C = \max_{P(V)} \sum_{V} P(V)P(R|V)\, log_2 \frac{P(R|V)}{P(R)}$$

Since $P(R) = \Sigma_V P(V)P(R|V)$, the capacity of a memory cell is determined entirely by its conditional distribution / transition matrix P(R|V) (Fig. 3).

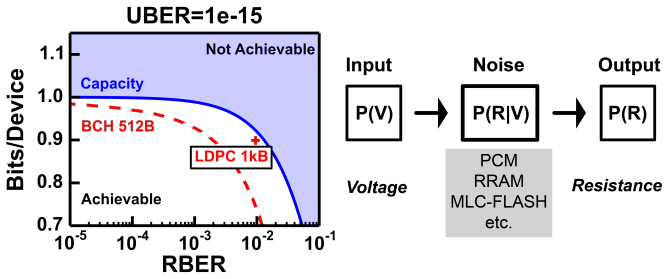

Fig. 1 (left) Capacity sets the frontier between achievable and unachievable rates of reliable storage with error correction. While commonly used Bose Ray-Chaudhuri (BCH) codes perform below the capacity limit, it can be approached by modern LDPC codes with large codewords (ex. 1kB) [4]. **Fig. 2 (right)** Capacity can be measured by viewing information storage as communication through a noisy memory channel. It is determined uniquely by P(R|V).

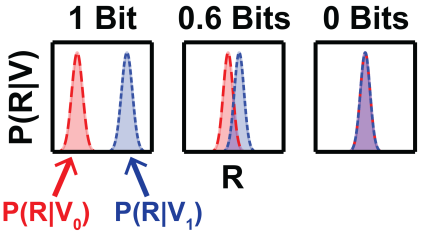

Fig. 3 Overlap of distributions reduces capacity. Capacity is optimized by increasing the number of input states and reducing overlap of output states.

Here, we demonstrate how measuring the full conditional distribution P(R|V), instead of the RBER at a number of distinct states, enables co-optimization of device characteristics and memory controller design to maximize storage capacity (bits/device). Further, we find ECCs that operate without quantization, storing analog signals in analog

devices, intrinsically operate at the highest capacity for a given channel.

Device Characterization

For this study we performed pulsed resistance measurements on 100-device 1T1R PCM arrays (Fig. 4). Device fabrication and characterization of these arrays were previously reported in [8-10]. The cell is reset to high resistance states via short current pulses that heat the film above its melting temperature and quickly quench to form a resistive amorphous cap. Slow Joule heating above the crystallization temperature anneals the amorphous cap and sets the low resistance state. Multiple resistance levels are achieved by switching between different volumes of resistive amorphous and conducting crystalline phases [11].

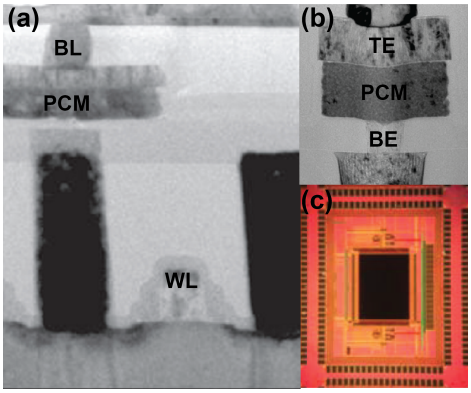

Fig. 4 100-device PCM array (~40nm BE contact diameter) (TEM (a)[8], (b)[9], and (c) optical micrograph) as an example analog-valued memory.

The gate voltage (V_{WL}) on each access transistor controls the current flowing through the cell (Fig. 5). To ensure independent statistics, we apply a pulsing scheme (Fig. 6) where the cell is initialized to a consistent set state, before applying partial reset pulses to the word line (V_{WL}) of varying magnitude and measuring the resulting resistance (R).

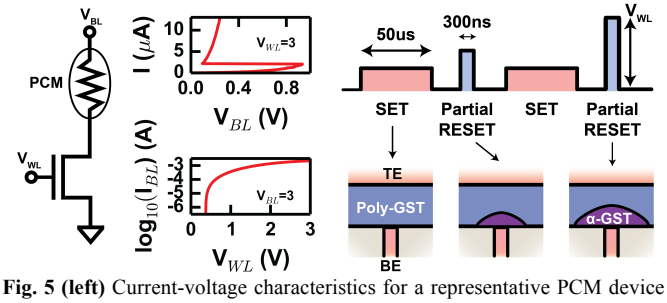

Fig. 5 (left) Current-voltage characteristics for a representative PCM device and access transistor in the array (I_{RESET} ~400μA) **Fig. 6 (right)** Pulsing scheme for partial-RESET operation. Device is SET between each partial-RESET pulse, and R(V_{WL}) is measured.

After collecting 380 trials at each voltage level, we calculate a Gaussian kernel density estimate of the continuous probability density P(R|V_{WL}) (Fig. 7). For simplicity and generality, time-dependent effects of PCM such as resistance drift are not considered here [12]. Similarly, we apply only a single write and read step, even though more robust storage has been demonstrated with read-verify schemes [3]. The results of this investigation can be extended to these modified channel models, and methods for calculating their Shannon capacity have been proposed [13].

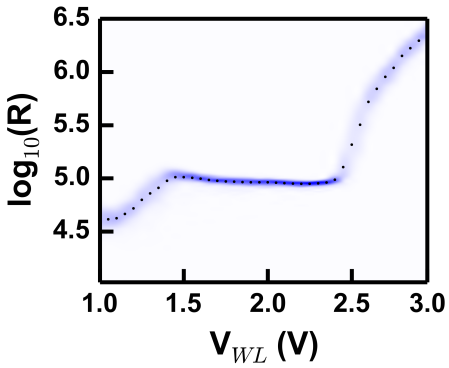

Fig. 7 Kernel density estimation of P(R|V_{WL}) (shaded, darker is higher density), from PCM data over 380 trials. Means of the data at each voltage are represented by dots.

Optimal Discretization

Measuring P(R|V_{WL}) provides a powerful tool to optimize the write pulse and read bin locations of an associated memory controller. We solve for the capacity-achieving input distribution by maximizing the mutual information over P(V) using the Blahut-Arimoto algorithm [14]. While the algorithm only applies to discrete distributions, we can approximate the capacity of analog channels by sufficiently discretizing P(R|V) such that the capacity is not increased by further discretization (ex. >2000 states).

The number of nonzero values in the capacity-achieving input distribution is determined by the balance between using as many input states as possible, and reducing overlap of their outputs (Fig. 8). Interestingly, beyond 12 discrete inputs, no more states are added, as the increased overlap would create a net reduction in mutual information (Fig. 9). The output distributions for these 12 inputs partially overlap, demonstrating higher capacity than achievable with totally distinct states. The unequal input state probabilities are unrealistic for most applications, however, making these probabilities uniform only reduces the channel capacity by approximately 5%.

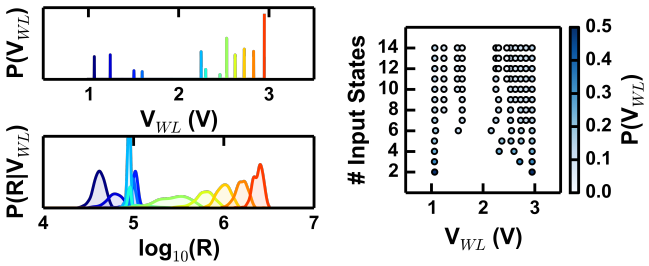

Fig. 8 (left) Optimal 'analog' input distribution $P(V_{WL})$, is actually discrete (12 states). Corresponding output $P(R|V_{WL})$ achieves the highest storage capacity despite moderate overlap of 12 outputs. **Fig. 9 (right)** Optimal input distributions for increasing number of allowed states. Mutual information is maximized by reducing overlap of outputs. Each row of dots is of the same type as the top graph of figure 8, with height represented by color. For 2 inputs, they are as separate as possible. Beyond 12 discrete inputs, no more states are added, as the increased overlap would create a net reduction in mutual information.

We further consider optimal discretization strategies via employing basin hopping search to find discrete inputs/outputs levels with maximum capacity [15]. Comparing discretization schemes between equally spaced, optimally spaced, and analog-valued input/outputs (Fig. 10), capacities are similar between schemes for low and high bit ADC/DACs (Fig. 11-12), as they are limited by the number of input states and intrinsic overlap of P(R|V) respectively. However, at 3-bits, the optimally spaced scheme gains 20% over the equally spaced case, and analog-valued outputs gain 30% over equal spacing. Having more output states than input states creates 'soft information', grayscale belief about the input given the output, a practice currently used to increase capacity of MLC-Flash decoders that perform variations of belief propagation [16].

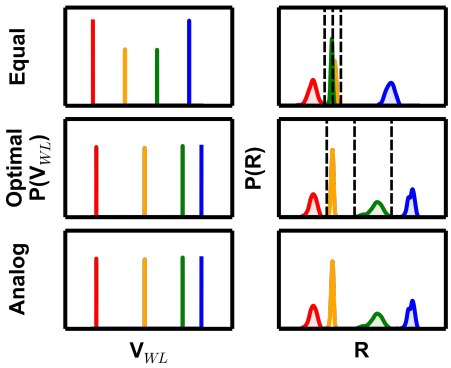

Fig. 10 Quantization schemes for 4 inputs and outputs, drawn from P(R|V) in Fig. 7: equally spaced, optimally spaced, and analog-valued. The equally spaced example has higher probability in the outer states as there is less overlap of the outputs than the inner states.

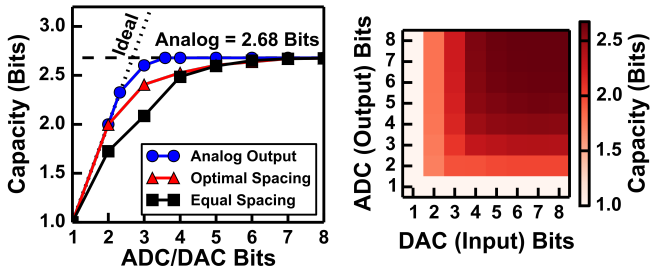

Fig. 11 (left) Higher capacity can be reached for the same number of states if their spacing is optimally chosen. Analog representations, such as in an analog artificial neural network, intrinsically have the same capacity as an infinite bit ADC. **Fig. 12 (right)** The capacity of the PCM device approaches the analog case for increasing numbers of input and output states. Having more output states than input states creates 'soft information', grayscale belief propagation values currently used in MLC-Flash for decoders.

Energy Efficient Storage

We calculate the energy consumption of each pulse from oscilloscope measurements (Fig. 13). The current is transistor limited, indicating that the dynamic resistance of the PCM devices is low and pulses are not strongly affected by parasitic capacitances in the 1T1R structure. Additional power consumed due to line losses are included in simulations of a 1Mb square array with 130nm wide, 1:1 aspect ratio, Cu wires [17] (Fig. 14). We then perform constrained optimization to find the input distributions that maximize capacity per unit energy [14] (Fig. 15). Since larger V_{WL} consume more energy, inputs are constrained to lower voltages in the efficient case (Fig. 16). Although this gives less separable outputs and lower capacities, there is an overall gain in efficiency (nJ/bit). We find an appropriate choice of input pulses can create a 32% reduction in energy/bit for the array.

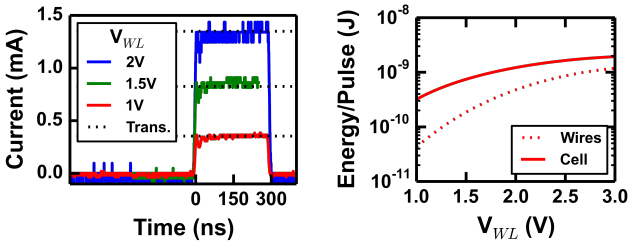

Fig. 13 (left) Current is limited by the access transistor. Current traces measured in oscilloscope match the predicted level (dashed) based on transistor transconductance from Fig. 5. **Fig. 14 (right)** Simulations of square 1Mb array, including I^2R losses in the wires and CV^2 losses from capacitive charging (dashed), and energy dissipated in the memory cell calculated from Fig. 5 (solid).

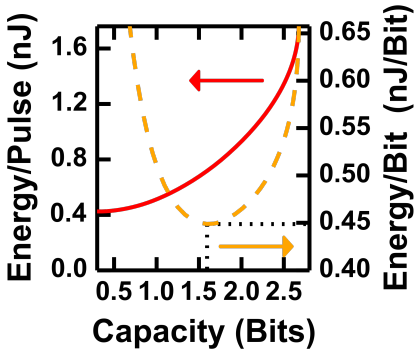

Fig. 15 Simulations of 1Mb array. While capacity (red) decreases with less energy due to less separable outputs, a minima exists for capacity/energy (yellow dashed).

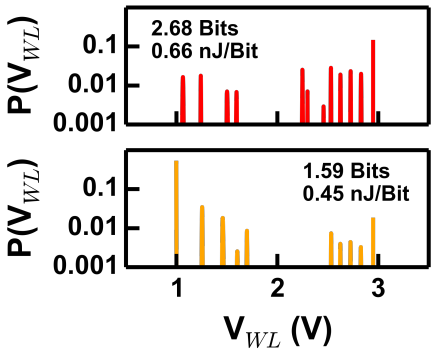

Fig. 16 The ideal input distribution for max capacity (red, 12 states) and max efficiency (nJ/bit) (yellow, 10 states). The efficient case is weighted towards lower V_{WL} that have less current and power per pulse, but also less separable outputs.

Conclusion

We have presented an information theoretic framework for the characterization of analog-valued emerging memory devices. By measuring full device statistics, we have demonstrated the potential for co-optimization of memory device and controller design. Further, the results that analog-valued circuits intrinsically operate at peak capacity are promising for analog artificial neural network designs where emerging memories are proposed as artificial synapses.

Acknowledgements

This work was supported in part by SONIC, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA, and by member companies of the Stanford Non-Volatile Memory Technology Research Initiative (NMTRI).

References

[1] D. Evans, "The Internet of Things How the Next Evolution of the Internet Is Changing Everything," Cisco, http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf, *Internet Business Solutions Group*, , 2011.

[2] C. Trinh et al., "A 5.6MB/s 64Gb 4b/Cell NAND Flash memory in 43nm CMOS," in *2009 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, 2009, pp. 246–247.

[3] T. Nirschl et al, "Write Strategies for 2 and 4-bit Multi-Level Phase-Change Memory," in 2007 *IEEE International Electron Devices Meeting*, 2007, pp. 461–464.

[4] R. Motwani, Z. Kwok, and S. Nelson, "Low Density Parity Check (LDPC) Codes and the Need for Stronger ECC," in *Flash Memory Summit*, Aug. 2011.

[5] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[6] D. J. C. MacKay and R. M. Neal, "Near Shannon limit performance of low density parity check codes," *Electron. Lett.*, vol. 33, no. 6, p. 457, 1997.

[7] S. Tanakamaru, Y. Yanagihara, and K. Takeuchi, "Error-Prediction LDPC and Error-Recovery Schemes for Highly Reliable Solid-State Drives (SSDs)," *IEEE J. Solid-State Circuits*, vol. 48, no. 11, pp. 2920–2933, Nov. 2013.

[8] M. Breitwisch et al., "Novel Lithography-Independent Pore Phase Change Memory," in *2007 IEEE Symposium on VLSI Technology*, 2007, pp. 100–101.

[9] G. F. Close et al., "Device, circuit and system-level analysis of noise in multi-bit phase-change memory," in *2010 International Electron Devices Meeting*, 2010, pp. 29.5.1–29.5.4.

[10] S. B. Eryilmaz et al., "Experimental Demonstration of Array-level Learning with Phase Change Synaptic Devices," *IEEE Int. Electron Devices Meet.*, vol. 25, no. 5, pp. 1–4, Dec. 2013.

[11] H.-S. P. Wong et al., "Phase Change Memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2201–2227, Dec. 2010.

[12] N. Papandreou, H. Pozidis, T. Mittelholzer, G. F. Close, M. Breitwisch, C. Lam, and E. Eleftheriou, "Drift-Tolerant Multilevel Phase-Change Memory," in *2011 3rd IEEE International Memory Workshop (IMW)*, 2011, pp. 1–4

[13] .L. A. Lastras-Montano, M. Franceschini, T. Mittelholzer, and M. Sharma, "Rewritable storage channels," in *2008 International Symposium on Information Theory and Its Applications*, 2008, pp. 1–6.

[14] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 14–20, Jan. 1972.

[15] D. J. Wales and J. P. K. Doye, "Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms," *J. Phys. Chem. A*, vol. 101, no. 28, pp. 5111–5116, Jul. 1997.

[16] G. Dong, N. Xie, and T. Zhang, "On the Use of Soft-Decision Error-Correction Codes in nand Flash Memory," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 58, no. 2, pp. 429–439, Feb. 2011.

[17] J. Liang, S. Yeh, S. S. Wong, and H.-S. P. Wong, "Effect of Wordline/Bitline Scaling on the Performance, Energy Consumption, and Reliability of Cross-Point Memory Array," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 9, no. 1, pp. 1–14. Feb. 2013.

Device data and Python code for analysis are publicly available at https://github.com/rctn/CapacityOptimization.