Hyperdimensional Computing with 3D VRRAM In-Memory Kernels: Device-Architecture Co-Design for Energy-Efficient, Error-Resilient Language Recognition

Haitong Li[1*], Tony F. Wu[1], Abbas Rahimi[2], Kai-Shin Li[3], Miles Rusch[2], Chang-Hsien Lin[3], Juo-Luen Hsu[3], Mohamed M. Sabry[1], S. Burc Eryilmaz[1], Joon Sohn[1], Wen-Cheng Chiu[3], Min-Cheng Chen[3], Tsung-Ta Wu[3], Jia-Min Shieh[3], Wen-Kuan Yeh[3], Jan M. Rabaey[2], Subhasish Mitra[1], and H.-S. Philip Wong[1#]

[1]Stanford University, USA; [2]University of California, Berkeley, USA; [3]National Nano Device Laboratories, Taiwan

Email: [*]haitongl@stanford.edu; [#]hspwong@stanford.edu

Abstract—The ability to learn from few examples, known as one-shot learning, is a hallmark of human cognition. Hyperdimensional (HD) computing is a brain-inspired computational framework capable of one-shot learning, using random binary vectors with high dimensionality. Device-architecture co-design of HD cognitive computing systems using 3D VRRAM/CMOS is presented for language recognition. Multiplication-addition-permutation (MAP), the central operations of HD computing, are experimentally demonstrated on 4-layer 3D VRRAM/FinFET as non-volatile in-memory MAP kernels. Extensive cycle-to-cycle (up to 10^{12} cycles) and wafer-level device-to-device (256 RRAMs) experiments are performed to validate reproducibility and robustness. For 28-nm node, the 3D in-memory architecture reduces total energy consumption by 52.2% with 412 times less area compared with LP digital design (using registers as memory), owing to the energy-efficient VRRAM MAP kernels and dense connectivity. Meanwhile, the system trained with 21 samples texts achieves 90.4% accuracy recognizing 21 European languages on 21,000 test sentences. Hard-error analysis shows the HD architecture is amazingly resilient to RRAM endurance failures, making the use of various types of RRAMs/CBRAMs (1k ~ 10M endurance) feasible.

I. Introduction

Brain-inspired computing aims at energy-efficient emulation of human cognition [1]. Such computing paradigm has advanced rapidly due to progress in emerging synaptic devices [2]-[7] and non-Von Neumann architectures [8]-[13], including hardware implementations of neural networks [1]-[8]. Human cognition features the ability to learn from few examples at a rapid pace, known as one-shot learning [14], whereas modern deep neural networks require large datasets and brute force training of billions of weights iteratively [15]. In contrast, hyperdimensional (HD) computing, a totally different brain-inspired computational framework, is capable of one-shot learning [16]. HD computing requires substantially less number of operations compared with modern deep learning algorithms. It is rooted in the observation that key aspects of human memory, perception and cognition can be explained by the subtle mathematical properties of high-dimensional spaces [16]. In this work, device-architecture co-design of HD language recognition systems using 3D VRRAM/CMOS is presented for the first time (Fig. 1). Extensive cycle-to-cycle (C2C) and device-to-device (D2D) measurements validate the robustness of 3D in-memory MAP kernels. The demonstration of energy-efficient, error-resilient in-memory HD architecture paves the way towards efficient learning machines with hallmarks of human cognition.

II. VRRAM In-Memory MAP Kernels

In the HD computing framework, information (letters, phonemes, DNA sequences, etc.) is represented and distributed in binary vectors with thousands of random '0s and '1's. These HD vectors are manipulated through multiplication-addition-permutation (MAP) kernels to not only classify, but also to bind, associate, and perform other types of cognitive operations in a one-shot manner (Fig. 1). Multiplication (MULT), addition (ADD), and permutation (PERM) are mapped onto 3D vertical RRAM (VRRAM) as in-memory MAP kernels ('1': low resistance states (LRS); '0': high resistance states (HRS)) (Fig. 2). XOR on {0, 1} binary code is equivalent to MULT on {1, -1} bipolar code. Based on the binary code, MULT is therefore performed by programming and evaluating XOR logic along vertical pillars in VRRAMs (Fig. 2(a)). ADD and PERM are performed by current summing (Fig. 2(b)) and in-memory bit transfer (Fig. 2(c)), respectively. To experimentally demonstrate the MAP kernels, 4-layer TiN/Ti/HfO$_x$/TiN 3D VRRAMs integrated with FinFET select transistors were fabricated. Detailed fabrication process was reported in [17]. Typical DC/endurance characteristics from bottom layer-1 (L1) to top layer-4 (L4) are shown in Fig. 3. Random '0's and '1's, the medium for HD computing, are naturally produced within VRRAM utilizing the intrinsic probabilistic switching behaviors (Fig. 4) [18], [19]. D2D statistical distributions of SET probabilities (P_{SET}) are also measured (Fig. 5), which are then incorporated into a variation-aware RRAM compact model on top of cycle-to-cycle variations [20]. P_{SET} can be tuned by programming conditions. Shorter pulses result in tighter D2D spreads around certain P_{SET}, owing to better reproducibility of filament morphology during C2C measurements (Fig. 5). 50% P_{SET} (with +/- 4% D2D variations) is used to produce random '0's and '1's for the following MAP operations.

The experimental implementations of MAP kernels are built upon 'in-memory computing' principle. Voltage division between RRAM cells and linear-region FinFET dynamically changes the pillar voltage (V_P), which leads to SET/RESET/non-switching of upper-layer cells in 3D VRRAM. Thereby, XOR logic kernel can be programmed along the vertical pillar (Fig. 6). During programming, the truth tables of XOR and XNOR logic are memorized thanks to non-volatility. Hence, further logic evaluations are performed by selecting/reading the target pillar with inputs as decoding address. The read-dominant logic evaluations on XOR/XNOR are measured up to 10^{12} cycles without output errors (Fig. 7). Logic evaluation voltage (V_{EVAL}) is 0.1 V. The VRRAM in-memory logic implementation differs from conventional NVM lookup tables [13], in the sense that any other arbitrary functions can be also programmed online in VRRAM (Fig. 8). During each cycle marked with gray background, new functions are programmed online (same principle in Fig. 6). This feature supports variants of MAP kernels within the HD computing framework. Bit error rates (BER) of XOR logic kernel using elevated V_{EVAL} are measured, under room temperature (RT) and 150°C (Fig. 9). The BER data are obtained by performing intensive logic evaluations and monitoring the errors due to disturb on RRAM. Such behaviors are well captured by temperature-dependent compact model [20], under RT, 150°C, and 260°C (solder reflow temperature). Predicted BER under 0.1 V at 260°C is below 10^{-13} (0.0001 ppb), which shows the XOR kernel is extremely robust. Through current summing along vertical pillars, in-memory additions are measured on 4-L VRRAMs that store various 4-bit vectors (Fig. 10). On each pillar, each in-memory addition is repeatedly measured for 10^{11} cycles, and robust additions outputs are obtained without crosstalk or disturb errors among different layers of VRRAMs (Fig. 11). PERM operations are implemented within VRRAM by direct bit transfer among different RRAM cells without needing to first read the content

978-1-5090-3902-9/16/$31.00 ©2016 IEEE

followed by write-back. These cells form a pseudo-series connection, and thus resistance state of one cell can determine another's during pulse programming. Target cell is initialized (RESET) to '0'. After applying a pair of V_{DD}/gnd pulses on the two RRAM cells, bit transfer is completed *in situ* (Fig. 12). The simple in-memory bit transfer does not require extra readout/write-back operations via memory controller/bus. Bit transfer between non-adjacent cells is also feasible for permutations in arbitrary orders (Fig. 13). Since read-dominant XOR and ADD operations are performed after PERM in algorithm, 10^{11} read evaluations (0.1 V) are conducted after each cycle of permutations to emulate system-level behaviors (Fig. 14). Correct and robust permutations are maintained.

Furthermore, wafer-level VRRAM in-memory MAP kernels are experimentally demonstrated and verified to support circuit design and implementation (Fig. 15). D2D measurements across 16 dies and 64 4-L VRRAM pillars (256 RRAM cells in total) are conducted. First, correct XOR outputs (stored in L4 cells) are obtained among the gray-code input combinations (Fig. 15(a)). Second, measured current summing corresponds to vector-wise 4-bit additions, where the output current levels correctly match the desired results (Fig. 15(b)). Last, permutations are performed by bit transfer from L1 (start) to L4 (destination) cells. In Fig. 15(c), the digits around arrows illustrate the pre-stored bits in L1/L4 cells. After permutations, the measured new data in L4 layer (color maps) correctly match the values stored in L1 layer, indicating the success of PERM.

III. In-Memory HD Computing Architecture

Device-architecture-algorithm co-design is leveraged for in-memory HD computing systems recognizing 21 European languages. For training, 21 sample texts (100k~200k words/text) are taken from Wortschatz Corpora [21]. For inference, 21,000 unseen sentences (1,000 sentences/language) are taken from Europarl Parallel Corpus (independent sources) [22]. There are 3 levels of abstraction: algorithm, architecture design, and device operations (Fig. 16). In the algorithm level, 26 letters of the Latin alphabet plus ASCII space are represented by 27 1-kb HD vectors, and trigram (3 consecutive letters) encoding scheme is chosen. In the architecture level, 36-layer 3D VRRAM subarray is designed with vertical partition to carry out different stages of the algorithm pipeline. Individual HD vectors are loaded/stored in horizontal planes, and manipulated by MAP operations either vertically (vector-to-vector) or horizontally (vector-wise). At the device level, ~50% P_{SET} is employed for random projection. In-memory MAP operations are essentially 'regular' R/W memory operations. Therefore, all the standard peripheral R/W analog/digital circuits are included. The language recognition system works as follows: an input text is sampled and projected into HD space by a sliding window of 3 consecutive letters. The letter HD vectors are first permuted (ρ) in the *Letter* layers, and then multiplied (i.e., XORed in {0, 1} system) to compute trigram HD vectors that are temporarily stored in *Trigram* layers:

$$\rho\,(\rho\ letter1) \oplus \rho\ letter2 \oplus letter3. \qquad (1)$$

The generated trigram HD vectors are continually added (accumulated). A final text vector is generated/stored after comparing the sum with the threshold and writing '0's and '1's into the subarray. The 21 trained language vectors are stored in six *LangMap* layers (4 kb/layer), as visualized in Fig. 17. These language maps efficiently encode/capture the causal relations. During inference, 21,000 input sentences go through the same pipeline. Each generated test vector is then XORed with the 21 language vectors and summed in *HamD* layers, yielding Hamming distance (HamD) to identify and select the language (minimum HamD). It's observed that a test vector is very 'far' from the median of 20 unselected language vectors in HD space (Fig. 18), which is the underlying mechanism for robust recognition. Language recognition accuracy is 90.4% considering the D2D variability (50% P_{SET}, +/- 4%) of 3D VRRAM (Fig. 20). The choice of N-gram encoding scheme and HD vector dimension leads to different accuracy performance and circuit size requirement (Fig. 21).

Using 28-nm technology, the 3D in-memory architecture is compared with a low-power (LP) digital design. The LP digital design uses the RTL implementation reported in [23], which is also a non-Von Neumann architecture with distributed registers (no SRAMs) in encoding/search modules. Place and route and post simulations are conducted using the same 28-nm PDK. Running on the same sub-dataset with 1-kb HD vectors, 52.2% energy reduction is obtained by 3D in-memory architecture (Fig. 21). The benefits come from the energy-efficient MAP kernels and the 3D in-memory architecture eliminating long interconnect of the planar design. There was no intentional optimization of peripheral analog circuitry (SA/MUX/PG) for the HD design. Owing to the cost-effective 3D memory-centric structure, total area can be reduced by more than 400 times as compared with planar CMOS design (Fig. 22). When 10-kb HD vectors are chosen for HD computing, under 9-metal/chip area constraint (<10× the total area of all standard cells), the LP digital design fails in routing cleanly. As for the system robustness, for various levels of RRAM endurance considered, the in-memory HD architecture is amazingly error resilient in terms of RRAM endurance failures. This result was obtained by introducing hard stuck-at errors into entire architecture during simulations (Fig. 23). Using various RRAMs/CBRAMs having endurance from 1k ~ 10M (or more) cycles is feasible for HD computing. VRRAMs in this work have 1M cycles endurance (Fig. 3). HD computing also shows superior error resilience over conventional machine learning algorithm [23] (Fig. 24). Higher dimensionality is even more robust. These promising features are attributed to the high-dimensional and holographic representation: every piece of information in a HD vector is 'distributed' equally over all the components, making even the hard errors not "contagious".

Finally, device-architecture co-optimization is performed. Sparsity is introduced into HD computing by initially tuning RRAM P_{SET} while using the same MAP operations. Energy is reduced with more HRS cells involved during MAP operations. The penalty of accuracy drop can be mitigated by computing in higher dimensionality (Fig. 25). The 3D VRRAM design is further scaled up into a larger memory subsystem (Fig. 26). Two mats are activated at the same time for parallelism, and RRAM endurance can be relieved additionally by 16× since data/operations are distributed among multiple subarrays.

IV. Conclusions

Key achievements: (1) probabilistic switching of RRAM is utilized to efficiently generate random '0's and '1's for HD computing; (2) non-volatile VRRAM in-memory MAP kernels are experimentally demonstrated with verified reproducibility and robustness; (3) improved energy and area efficiencies are obtained from the novel 3D in-memory architecture over LP digital design with post-layout simulations; (4) RRAMs/CBRAMs having wide ranges of endurance (1k ~ 10M+) can be used in the error-resilient HD systems.

Acknowledgement

This work is supported in part by NSF ENIGMA Project, STARnet SONIC, Stanford NMTRI and Stanford SystemX Alliance. Device fabrication is performed by NDL facilities and supported by the Ministry of Science and Technology, Taiwan.

References

[1] C. Mead, Proc. IEEE, p.1629, 1990. [2] M. Prezioso *et al.*, Nature, p.61, 2015. [3] G.W. Burr *et al.*, IEDM, p.697, 2014. [4] D. Garbin *et al.*, IEDM, p.661, 2014. [5] M. Suri *et al.*, IEDM, p.235, 2012. [6] S. Park *et al.*, IEDM, p.231, 2012. [7] S.H. Jo et al., Nano Lett., p.1297, 2010. [8] P. Merolla, *et al.*, Science, p.668, 2014. [9] M.-F. Chang *et al.*, ISSCC, p.318, 2015. [10] J.J. Yang *et al.*, Nature Nanotech., p.13, 2013. [11] B. Chen *et al.*, IEDM, p.459, 2015. [12] T. Hasegawa *et al.*, Adv. Mater., p.252, 2012. [13] H. Noguchi *et al.*, IEDM, p.617, 2013. [14] S. Thorpe *et al.*, Nature, p.520, 1996. [15] A. Krizhevsky *et al.*, NIPS, 2012. [16] P. Kanerva, Cog. Comput., p.139, 2009. [17] H. Li *et al.*, VLSI, p.194, 2016. [18] S. Yu *et al.*, Front. Neurosci., p.186, 2013. [19] N. Raghavan *et al.*, IEDM, p.554, 2013. [20] H. Li *et al.*, DATE, p.1425, 2015. [21] U. Quasthoff, *et al.*, LREC, 2006. [22] P. Koehn, MT Summit, 2005. [23] A. Rahimi *et al.*, ISLPED, p.64, 2016.

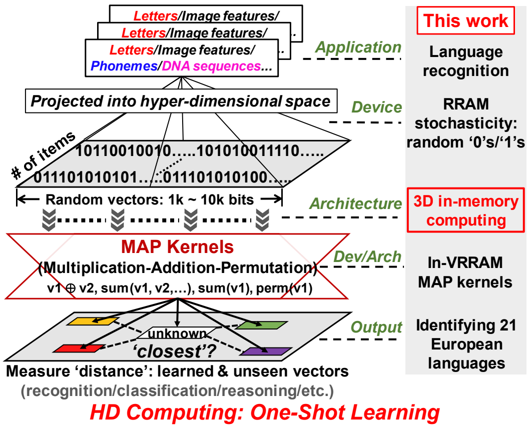

Letters/Image features/
Letters/Image features/
Letters/Image features/
Phonemes/DNA sequences...
Application
Projected into hyper-dimensional space
Device
# of items
10110010010.....101010011110.....
011101010101.....011101010100.....
Random vectors: 1k ~ 10k bits
Architecture
MAP Kernels
(Multiplication-Addition-Permutation)
v1 ⊕ v2, sum(v1, v2,...), sum(v1), perm(v1)
Dev/Arch
Output
unknown
'closest'?
Measure 'distance': learned & unseen vectors
(recognition/classification/reasoning/etc.)
This work
Language recognition
RRAM stochasticity: random '0's/'1's
3D in-memory computing
In-VRRAM MAP kernels
Identifying 21 European languages
HD Computing: One-Shot Learning

Fig. 1 Illustration of hyperdimensional (HD) computing framework and its association with this work.

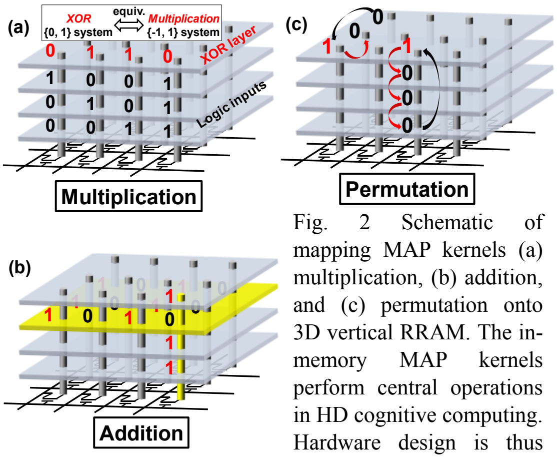


Fig. 2 Schematic of mapping MAP kernels (a) multiplication, (b) addition, and (c) permutation onto 3D vertical RRAM. The in-memory MAP kernels perform central operations in HD cognitive computing. Hardware design is thus highly algorithm-driven.

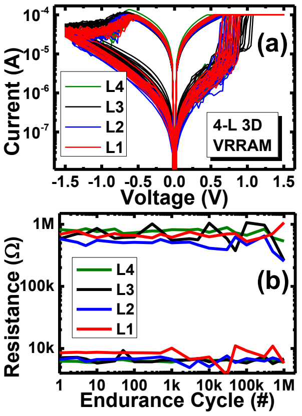


Fig. 3 (a) I-V curves of 4-L 3D VRRAM. (b) Endurance data.

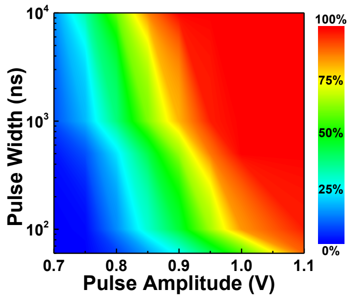


Fig. 4 Measured SET probabilities (P_{SET}) as a function of SET pulse amplitude and pulse width. Each probability value is obtained from 200-cycle strong-RESET/weak-SET operations. Random '0'/'1' bits are produced via P_{SET} ~50%.

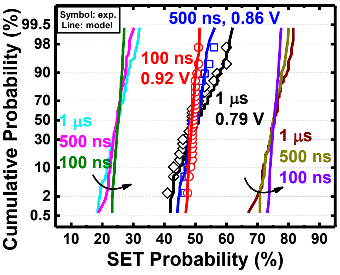


Fig. 5 Measured and modeled device-to-device (D2D) statistical distributions of P_{SET} around {25%, 50%, 75%}. Shorter pulses achieve tighter D2D distributions, as captured by both exp. and model. 25 devices are measured for each spread in the 50% P_{SET} trials.

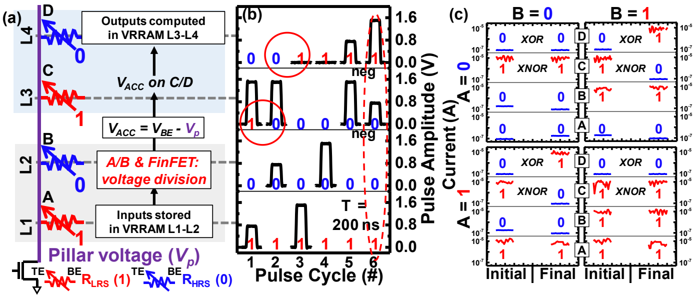


Fig. 6 (a) Schematic of executing Boolean logic on 4-layer 3D VRRAM/FinFET. Voltage division between RRAM and linear-region FinFET changes V_P, triggering different desired switching events on upper-level cells with logic outputs stored *in situ*. (b) Applied pulse train at L1-L4 BE to program XOR/XNOR (input AB = '10'). Digits show the states (LRS:1; HRS:0) of L1-L4 after each pulse cycle. (c) Measured initial/final states of L1-L4 for all combinations. Correct truth tables are measured.

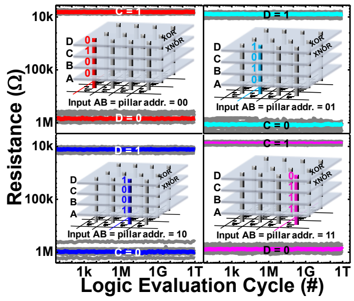


Fig. 7 Measured logic evaluations on XOR/XNOR pillars up to 10^{12} cycles. Four pillars store unique inputs/outputs. Each colored line shows the median of 10 evaluation cycles (gray) after reproducible XOR programming on each pillar.

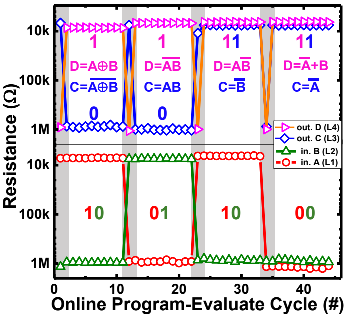


Fig. 8 Measured 4-L VRRAM states during online programming. Each cycle either computes a new function (gray background) or evaluates the new logic. 8 different functions are measured correctly.

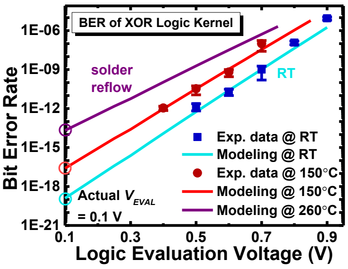


Fig. 9 Measured and modeled bit error rates (BER) of XOR logic kernel for different V_{EVAL} and temperatures. Errors are due to disturb on RRAM during data-intensive logic read evaluations. Predicted BER for 0.1 V at 260°C is below 10^{-13} (0.0001 ppb).

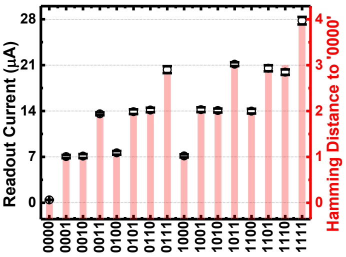


Fig. 10 Measured in-memory addition on various 4-bit vectors (0000~1111) stored in 16 4-L VRRAMs, which are efficiently realized by current summing along vertical pillars. Bit summation along a certain vector calculates the Hamming distance relative to all-zero vectors (measure 'difference').

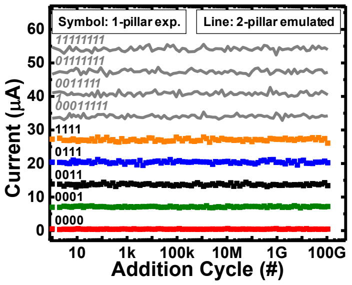


Fig.11 Measured consecutive in-memory addition outputs on 0000~1111 vectors up to 10^{11} cycles without disturb/crosstalk errors. Summation across 8 bits is emulated by combining exp. data from two pillars. Applied read voltage is 0.1 V.

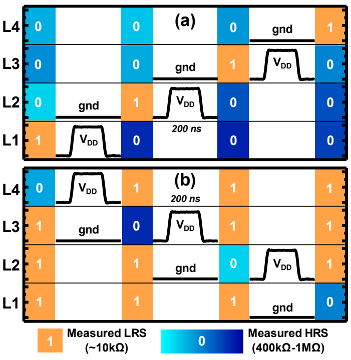


Fig.12 Measured resistance evolution (color-coded scale) of 4-L VRRAM during ordered-permutation of (a) bit '1' and (b) bit '0', in two sequences along the vertical pillars.

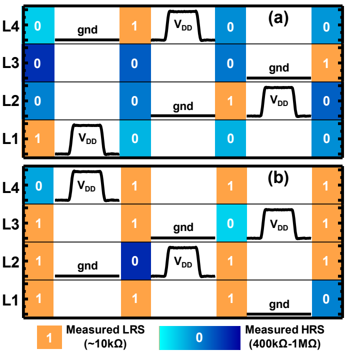


Fig. 13 Measured resistance evolution (color-coded scale) of 4-L VRRAM during arbitrary-permutation of (a) bit '1' and (b) bit '0', for two arbitrary orders along the vertical pillars.

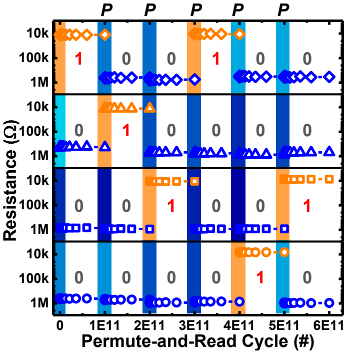


Fig. 14 Measured resistances (color-coded scale) of 4-L VRRAM during arbitrary permutation-and-read cycles. 10^{11} read cycles emulating logic evaluations are performed after each permutation.

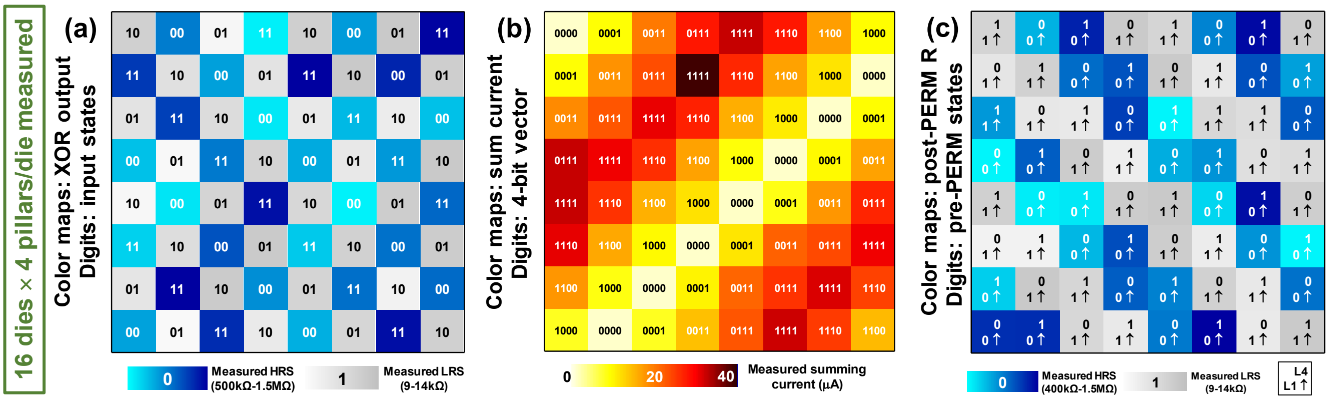


Fig. 15 Wafer-level demonstration and verification of MAP kernels across 16 dies and 64 4-L VRRAM pillars (256 RRAM cells in total). (a) Measured L4 cells' resistances as the correct XOR outputs after programming and evaluating XOR logic. Digits are inputs stored in 1st/2nd layers. (b) Measured sum current of 4-bit vectors stored along VRRAM pillars. Digits are L1-L4 bit states. (c) Measured resistances of L4 cells after permutations (post-PERM R). The digits/arrows illustrate the bit transfer path from L1 to L4. Measured post-PERM data in L4 correctly represent the digits in L1 (arrow's left side).

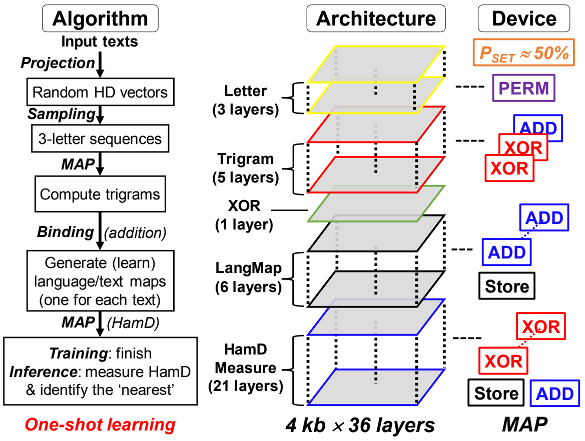


Fig. 16 Algorithm pipeline and 3D in-memory device-architecture co-design for the language recognition system.

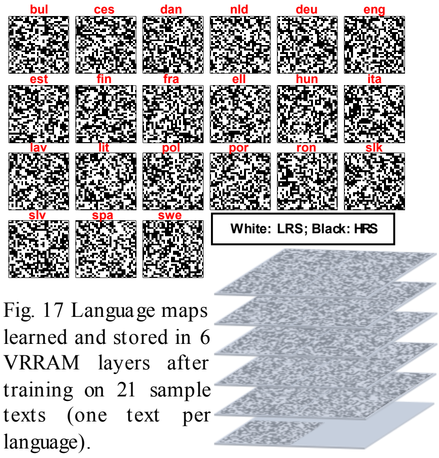


Fig. 17 Language maps learned and stored in 6 VRRAM layers after training on 21 sample texts (one text per language).

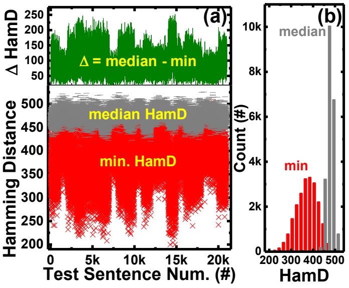


Fig. 18 (a) Hamming distances (HamD) between 21,000 test sentences and 21 learned language maps. Min. distance leads to the identified language. (b) HamD distributions. Medians of HamDs are near 1/2 HD vector dimension.

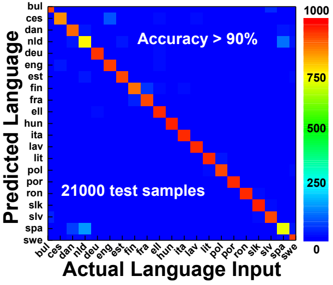


Fig. 19 Confusion matrix of language recognition on 21,000 test sentences. Most samples can be predicted correctly, even though the 21 European languages are somewhat correlated. Recognition accuracy is 90.4% considering the D2D variability of 3D VRRAM.

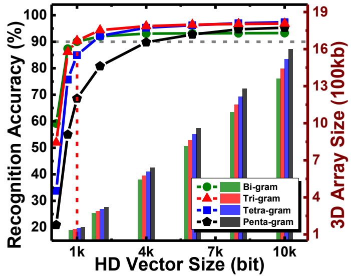


Fig. 20 Recognition accuracy and required 3D VRRAM array size as a function of HD vector dimension and N-gram encoding scheme. Using 1-kb or 2-kb HD vectors with trigram scheme can achieve 90%+ accuracy with modest VRRAM circuit size requirement.

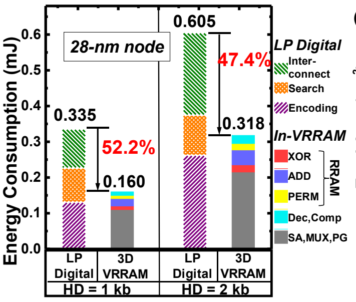


Fig. 21 Energy breakdown for low-power digital design and 3D in-memory architecture for HD systems under 28 nm node. Energy is reduced by 52.2% with in-memory MAP kernels and unoptimized peripherals.

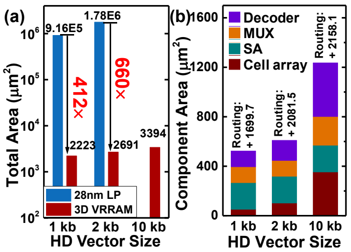


Fig. 22 (a) Total area comparison between LP digital design and in-memory architecture. Digital design for 10-kb vectors is not routable under 9-metal/chip area constraints. (b) Area breakdown for VRRAM circuits. 36-L VRRAM cells have aspect ratio = 30 and trench slope = 89°.

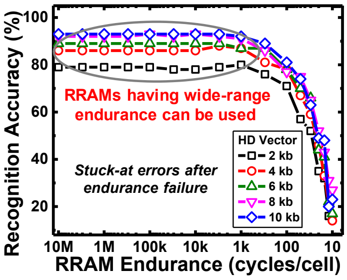


Fig. 23 Recognition accuracy as a function of RRAM endurance and vector size. The simulations assume the memory cells are stuck after endurance failures. Various types of RRAM/CBRAM with a wide range of endurance (1k~10M) can be used in error-resilient HD systems.

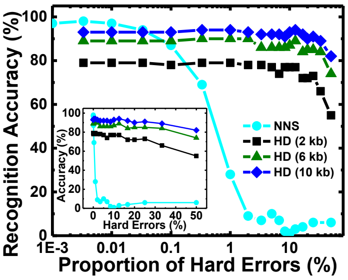


Fig. 24 Error resilience of HD computing and conventional nearest neighbor search (NNS) with hard stuck-at errors in text vectors/histograms during testing. Higher dimensionality in HD computing shows superior error resilience. Inset: linear-scale plot.

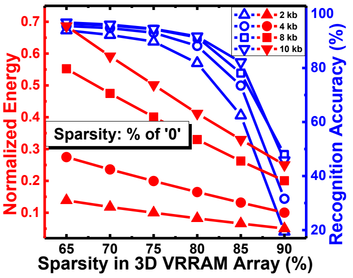
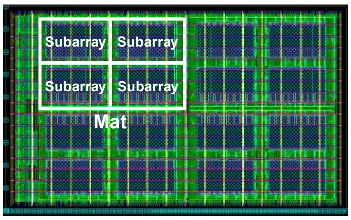


Fig. 25 Energy and accuracy as a function of sparsity (% of '0's) in 3D VRRAM array, initialized by tuning RRAM SET probabilities. Sparse representation is energy efficient, with the penalty of accuracy drop (less drop in higher dimensionality).



Speedup (parallelism)	Endurance relief	Bank energy overhead	Total area
8×	16×	7.7%	31977 μm²

Fig. 26 Layout of one RRAM bank for the 28-nm in-memory HD computing system. The table summarizes the benefits/costs of scaling up VRRAM circuits into a larger memory-centric system for HD computing, compared with the single VRRAM subarray design.