# In-Situ Encrypted NAND FeFET Array for Secure Storage and Compute-in-Memory

Zijian Zhao[1*], Yixin Xu[2*], James Read[3], Po-Kai Hsu[3], Yixin Qin[1], Tzu-Jung Huang[1], Suhwan Lim[4], Kijoon Kim[4], Kwangsoo Kim[4], Wanki Kim[4], Daewon Ha[4], Thomas Kämpfe[5], Sumitha George[6], Xiao Gong[7], Suman Datta[3], Shimeng Yu[3], Vijaykrishnan Narayanan[2], and Kai Ni[1]

[1]Rochester Institute of Technology; [2]Pennsylvania State University; [3]Georgia Institute of Technology; [4]Samsung Electronics Co., Ltd; [5]Fraunhofer IPMS; [6]North Dakota State University; [7]National University of Singapore; *equal contribution; (email: zz8118@rit.edu)

2023 International Electron Devices Meeting (IEDM) | 979-8-3503-2767-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEDM45741.2023.10413774

***Abstract*— In this work, we present a lightweight in-situ encryption/decryption technique for high-density NAND memory, aiming to meet the growing need for data privacy and security in storage and computing applications. Using ferroelectric FET (FeFET) as a technology platform for demonstration, we show that: i) using a XOR-based cipher, the encryption/decryption can be simply mapped to in-situ array operations, where the encrypted cipher texts are stored as complementary threshold voltage ($V_{TH}$) states of two consecutive FeFETs in a NAND string and decryption can be simply realized through read operations with key-dependent read gate biases; ii) the proposed technique is scalable to multi-level cells (MLC) by encrypting and decrypting bit-by-bit, thereby significantly increasing the encrypted memory density; iii) a unique advantage of applying XOR-based cipher on NAND array is its capability of supporting high-density and secure compute-in-memory (CiM) (e.g., matrix vector multiplication) with encrypted weights, which is beyond the capability of conventional advanced encryption standard (AES) engine; iv) with integrated NAND FeFET array, we have successfully demonstrated encryption and decryption operations of single-level cell (SLC), MLC, and CiM, showing great promise of the technique.**

## I. Introduction

Large generative artificial intelligence models, notably ChatGPT, take the world by storm and reshape various aspects of society. These powerful models are made possible with their huge number of parameters (Fig.1(b)), which would require a tremendous amount of computing power and storage. In that regard, harnessing high density vertical NAND memory (Fig.1(c)) for compute-in-memory (CiM) (e.g., matrix-vector-multiplication (MVM)) is highly attractive to minimize the required data transfer to save energy and latency (Fig.1(a)). The neural network (NN) weights are stored as the conductance of the NAND memory, either flash or FeFET, and the inputs are sent to the bit lines of the array for in-memory computation in analog domain. Exploiting nonvolatile memories (NVM), these CiM accelerators face a security challenge, i.e., being vulnerable to physical access-based attacks and having the risk of IP stealing. Therefore, encryption is very important for NN accelerators. Conventional encryption of NVMs is mainly based on Advanced Encryption Standard (AES), which incurs significant performance and energy overhead [1] (Fig.1(d)). Also very importantly, AES-based encryption is incompatible with the CiM acceleration, making it less attractive for NN accelerators.

To address this challenge, an XOR-cipher based encryption technique is adopted for the vertical NAND FeFET that only exploits the in-situ memory array operations and is compatible with CiM (Fig.1(e)). The idea is to take two consecutive word line (WL) pages and treat them as one encrypted page (Fig.1(f)). The cipher-text (CT) is stored as the configurations of two cells. Then the decryption is carried out by applying appropriate key-dependent read WL biases on the two pages simultaneously (Fig.1(g)). Only with the correct key, the sensed string current will correspond to that of the plain-text (PT). This design leaves significant design freedom in choosing the granularity of encryption, i.e., a single page, multiple pages, a block, etc., thus being a lightweight yet versatile technique.

Fig.2 shows the working principles of the 2FeFET cell for SLC encryption/decryption. The CT is stored as complementary $V_{TH}$ states of the two FeFETs and the key is mapped into the read gate voltage (i.e., $V_{R1}$ or $V_{R2}$) as shown in Fig.2(a). When the CT is 0/1, the FeFET $F_0$ and $F_1$ are programmed to be the low-$V_{TH}$ (LVT)/high-$V_{TH}$ (HVT) and HVT/LVT state, respectively (Fig.2(b)). The key bit 0/1 corresponds to apply $V_{R1}/V_{R2}$ and $V_{R2}/V_{R1}$ on $F_0$ and $F_1$, respectively (Fig.2(c)). In this way, only when the key and CT mismatches (i.e., XOR result is 1), both FeFETs will be turned ON and enable string current. Therefore, the decryption of the PT can be realized via in-situ memory sensing.

## II. NAND FeFET String Process Integration

The key integration process flow for NAND FeFET is shown in Fig.3(a). The fabrication is performed on a P-type silicon substrate. After phosphorus ion implantation and activation, the isolation oxide in the gate area is removed and the gate area is cleaned for 10 nm $Hf_{0.5}Zr_{0.5}O_2$ gate dielectric deposition through atomic layer deposition (ALD) at 250°C. Source/drain via is opened by reactive-ion etching (RIE) and buffered oxide etchant (BOE). A 90nm thick tungsten (W) layer is sputtered on the wafer for source, drain, and gate metal. The device is finally annealed in the forming gas ($N_2$+$H_2$, 350°C) and $N_2$ (500°C) for ferroelectric crystallization. The SEM and cross-section schematic of a single FeFET and the TEM of the gate stack are shown in Fig.3(b). A memory window (MW) of 0.7 V is obtained by applying a write pulse of ±4 V, 1 $\mu$s (Fig.3(c)). The switching dynamics of a FeFET is shown in Fig.3(d). The retention results of both memory states suggest stable memory states (Fig.3(e)) .

## III. NAND FeFET Array Characterization

The SEM of a NAND string, consisting of three FeFETs, is shown in Fig.4(a). ±4 V, 1 $\mu$s gate pulses are used for program and erase. To read a FeFET in the string, $V_{READ}$=0.9V

Authorized licensed use limited to: UNIVERSITY NOTRE DAME. Downloaded on October 04,2024 at 02:41:18 UTC from IEEE Xplore. Restrictions apply.

is applied to the target device and $V_{PASS}$=1.8V is applied to other devices (Fig.4(b)). The string currents are measured for all eight combinations, where the three FeFETs are each written to the HVT/LVT state (Fig.4(c)).

In a NAND array, an inhibition bias scheme needs to be applied to prevent unwanted programs [2-3]. Fig.4(d) shows an example of the program/inhibition case. FeFETs are first erased to the HVT state. A 4 V 10 $\mu$s pulse is applied on $WL_2$ to program the target cell ($F_{21}$). To inhibit the unselected cell ($F_{22}$), 2 V is applied on the $BL_2$. In this way, $F_{21}/F_{22}$ would end up at the LVT/HVT state, respectively. During the read operation, unselected cells are applied a $V_{PASS}$ and selected cells are applied a $V_{READ}$ (Fig.4(e)). Fig.4(f) shows *I-V* curves of 64 devices, showing a tight $V_{TH}$ distribution. The effectiveness of the inhibition bias scheme is shown in Fig.4(g), where devices remain at the HVT state even with a programming gate pulse.

## IV. Experimental Verification of In-Situ XOR-Based Encryption/Decryption

The SLC encryption/decryption is experimentally verified on an 8×8 encryption cell array. Each cell consists of two FeFETs. An 8×8 checkerboard pattern PT (Fig.5(a)) and a randomly generated key pattern (Fig.5(b)) are used for validation. By performing XOR logic, the CT is created (Fig.5(c)). Then a 16x8 FeFET array is programmed according to the CT (Fig.5(d)). The $V_{TH}$ distribution (Fig.5(e)) is tight, allowing accurate operation. By employing $V_{R1}$=1.7V and $V_{R2}$=0.9V, applying the correct key, the string currents are measured (Fig.5(f)), which can be easily converted to the logic value, which is the PT. Fig.5(g) and Fig.5(h) show the sensed string current for the case of randomly generated key and all-0 key. The accuracy for the random key and the all-0 key is 39% and 50%.

The proposed technique is also applicable for MLC by performing encryption bit-by-bit. Each CT bit is encrypted independently and the final MLC state to be programmed is determined by combining all the MLC bits together (Fig.6(a)). For example, for CT=0X (MSB=0), the $F_0/F_1$ will be the LVT/HVT state, meaning that $F_0/F_1$ will be at $S_3(S_2)/S_0(S_1)$, respectively. Then the CT LSB is also 0, $F_0$ will be $S_3$ and $F_1$ will be $S_0$. The decryption process requires four read voltage values ($V_{R0}$-$V_{R3}$ in Fig.6(b)). The decryption in MLC, in addition to key dependence, also relies on decrypted PT MSB bit. This dependence comes from the fact that sensing a MLC LSB bit requires to apply a read voltage that depends on the MSB. For example, for the case PT=1X and Key=01, the first read voltage applied is $V_{R0}/V_{R2}$ to the $F_0/F_1$ (i.e., determined by the key's MSB only) (Fig.6(c)). After sensing the string current, the PT's MSB is obtained. To determine the read configuration used for sensing PT's LSB, two cases emerge. If PT's MSB=1, $V_{R1}/V_{R2}$ are employed; if PT's MSB=0, $V_{R3}/V_{R0}$ are employed. Fig.6(d) shows the experiment which the two FeFETs in a memory cell are written to four $V_{TH}$ states. Based on these results, the two FeFETs are written to four $V_{TH}$ states ($S_3/S_0$, $S_2/S_1$, $S_1/S_2$, $S_0/S_3$) representing four CTs (i.e., 00, 01, 10, 11). Fig.6(e) shows the parameters used in encryption and sensing string current (decryption). The experimental results are shown in Fig.6(f). All CT and key combinations are covered and plotted in terms of key's MSB and LSB. The experimental results show one-to-one correspondence with the logic truth table, suggesting correct operation.

## I. NAND FeFET based Secure CiM

The proposed encryption scheme enables secure CiM. Fig.7(a) shows the vertical NAND FeFET based CiM with encrypted SLC weights. The input is the BL voltage and the output is the SL current. The weight is encrypted as the $V_{TH}$ states of the FeFETs, similar to SLC (Fig.2). With a correct key to decrypt the weight, the final SL current is the sum of the product between the weight and input BL voltage. Without the correct keys, the cell will yield wrong string current, causing errors in the MVM operation. The SEM and corresponding schematic for the 2×2 FeFET NAND array is shown in Fig.7(b). By using $V_{BL}$ from 0V to 0.07V with a step of 0.01V, a group of linear BL current is obtained (Fig.7(c)). Moreover, by applying all eight BL voltages to both BL1 and BL2, 64 cases are measured (Fig.7(d)). An excellent linear current distribution is achieved. These results are obtained for the case that all correct keys are applied. In the cases that some or all keys are wrong, the string current or computation results are wrong (Fig.7(e)).

Fig.8(a) shows the subarray design for the proposed encrypted CiM architecture including the peripheral circuitry. In the evaluation, 16-bit input vectors and weights are used for the first and last layers while 8-bit inputs and weights for other layers. Based on Resnet-18 trained on ImageNet dataset, the impacts of partial encryption and key guess accuracy of attackers on inference accuracy have been explored with the assumption that attackers use random guess for keys using NeuroSim [4]. Fig.8(b) shows the inference accuracy when encrypting multiple layers. It shows that even if only encrypting the first layer, the inference accuracy sharply drops to ~0%. This indicates our scheme can effectively prevent unauthorized attackers with partial encryption, which means the area overhead introduced by the structure of 2FeFET/cell can be hugely mitigated. Fig.8(c) shows the inference accuracy with single-layer encryption. In addition, the inference accuracy with different key accuracy rate of attackers is also evaluated (Fig.8(d)). The security can be maintained until the attacker can guess the key with >95% accuracy. Fig.8(e) shows a comparison between prior encrypted CiM works [5-6] with this work. Our work demonstrates high energy efficiency and low area cost.

## II. Conclusion

In this study, we present an NAND FeFET based XOR-cipher encryption method by solely leveraging in-situ memory operations, thus introducing minimal overhead. It can be effectively applied to both SLC and MLC memory, thereby significantly boosting encrypted memory density. We validate the viability of our approach through comprehensive integration and characterization of a NAND FeFET array, Furthermore, an exceptional benefit of the XOR-based cipher is its ability to facilitate a secure and high-performance CiM accelerator for neural networks.

**Acknowledgement:** This work was primarily supported by the SUPREME and PRISM, two of the SRC JUMP 2.0 centers and partially supported by the NSF 2312884. **References:** [1] Y. Xu, et al., Arxiv:2306.01863; [2] R. Micheloni, et al., Springer 2010; [3] G. Choe, et al., IEEE TED 2021; [4] S. Yu, et al., IEDM 2019. [5] S. Yu, et al., ICCAD 2020. [6] R. Huang, et al., IEDM 2022.

Authorized licensed use limited to: UNIVERSITY NOTRE DAME. Downloaded on October 04,2024 at 02:41:18 UTC from IEEE Xplore. Restrictions apply.

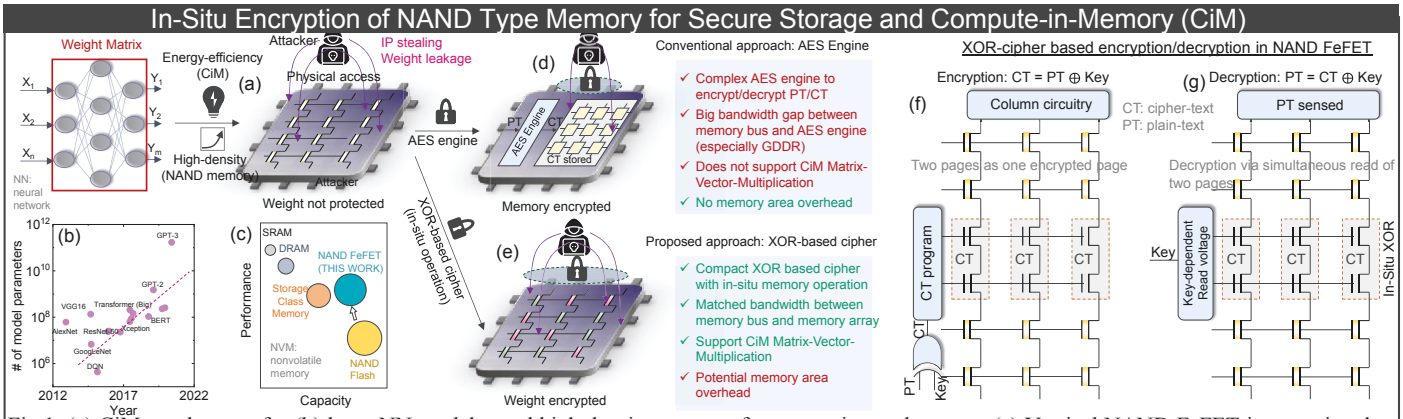



Fig.1. (a) CiM accelerators for (b) large NN models need high density memory for computing and storage. (c) Vertical NAND FeFET is attractive due to its high density and performance. The NVM based CiM accelerators face security challenges of physical access based attacks. (d) Encryption based on AES engine, not compatible with CiM, also has significant performance and power overhead. (e) The XOR-cipher approach supports secure CiM by exploiting in-situ memory operations. (f) Encryption and (g) decryption are mapped to CT programming and key-dependent read gate biases.

## Operation Principles of In-Situ XOR-Based Encryption/Decryption

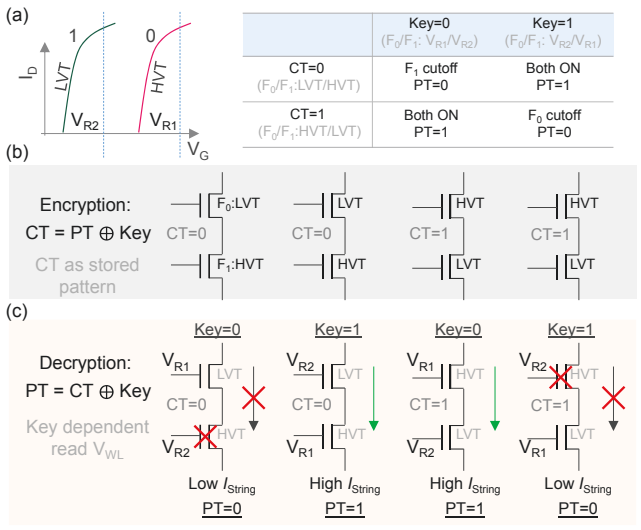



Fig.2. (a) Encryption and decryption scheme. (b) CT, i.e., the PT XOR the key, is mapped as the complementary $V_{TH}$ states of $F_0$ and $F_1$. (c) PT can be decrypted by applying key-dependent read gate voltages, which performs XOR logic between the CT and the key.

## NAND FeFET String Process Integration and Characterization

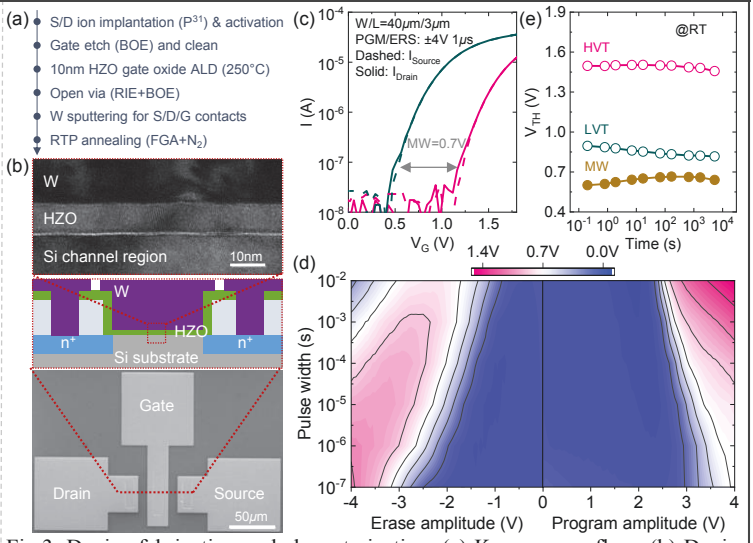



Fig.3. Device fabrication and characterization. (a) Key process flow. (b) Device SEM, cross-section schematic, and gate stack TEM. (c) Measured $I_D$-$V_G$ curves of two memory states. (d) Switching dynamics in a FeFET showing the MW under different amplitudes and pulse widths. (e) FeFET shows decent retention.

## NAND FeFET String and Array Operation and Characterization

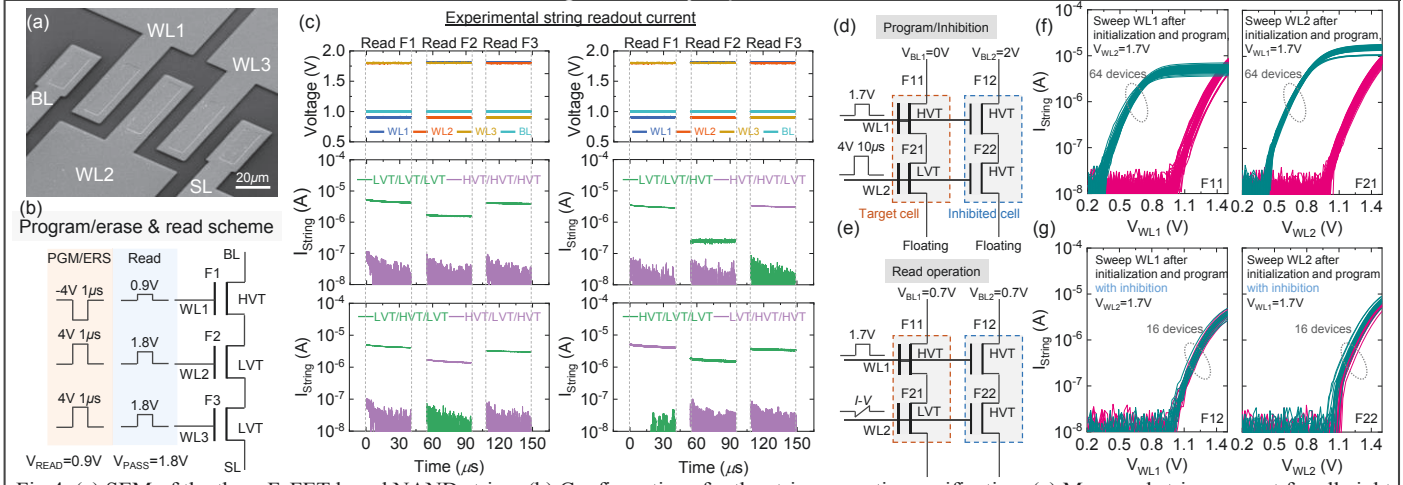



Fig.4. (a) SEM of the three FeFET based NAND string. (b) Configurations for the string operation verification. (c) Measured string current for all eight cases. It shows correct functionality of the FeFET based NAND string. Operation schemes for 2x2 array during (d) program/inhibition and (e) read. SL is kept floating by cutting off the third FeFET (not shown). For inhibited cells, BL is raised to 2V to float the channel to prevent programming. To characterize the memory cell, an *I-V* sweep is performed on one FeFET while a higher pass voltage (e.g., 1.7V here) is applied on the other FeFET. (f) shows the obtained *I-V* curves for the two FeFETs. (g) shows the two inhibited FeFETs remain high-$V_{TH}$ state, showing successful array operation.

Authorized licensed use limited to: UNIVERSITY NOTRE DAME. Downloaded on October 04,2024 at 02:41:18 UTC from IEEE Xplore. Restrictions apply.

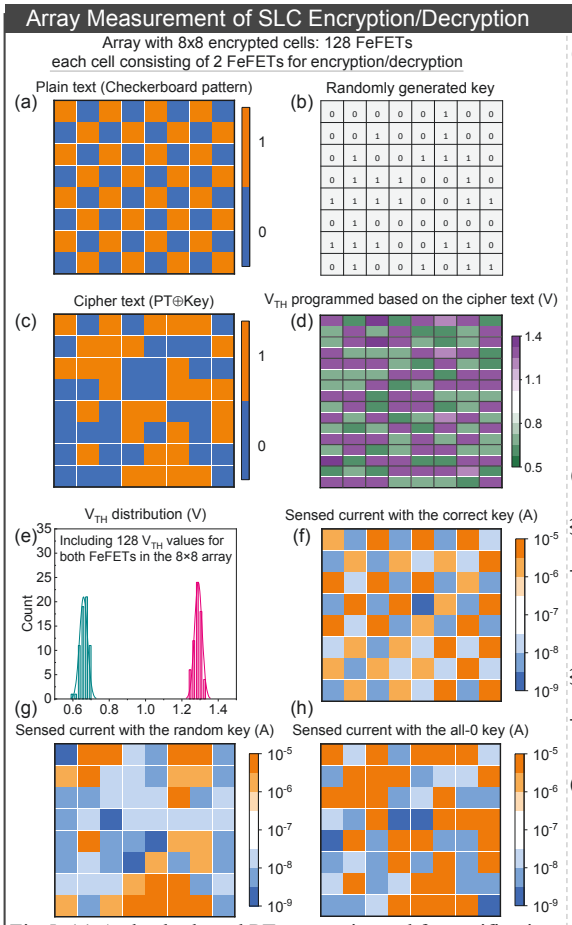



Fig.5. (a) A checkerboard PT pattern is used for verification. (b) A randomly generated key is for encryption. (c) The theoretical CT is calculated and written to an 16×8 SLC array. The corresponding $V_{TH}$ distribution is shown in (d) and (e). (f-h) shows the measured string current read out from the correct key, random key, and all-0 key.

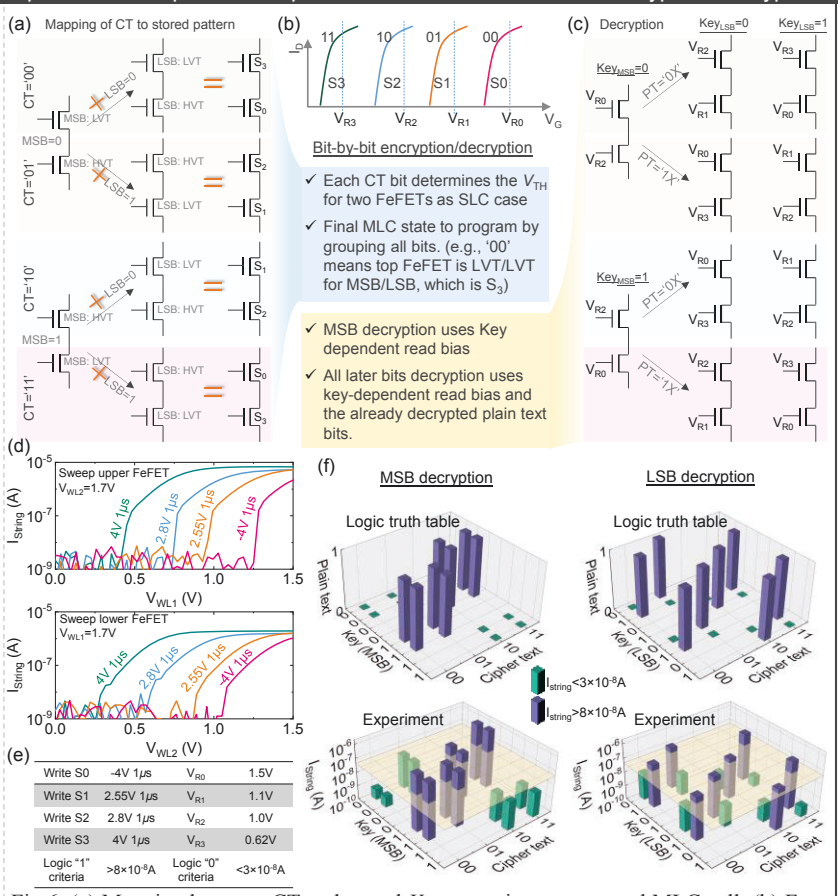
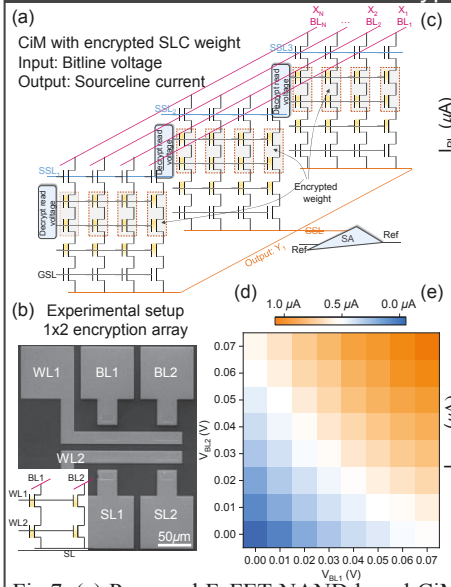



Fig.6. (a) Mapping between CT and stored $V_{TH}$ states in an encrypted MLC cell. (b) Four levels in a FeFET are used for the MLC operation. (c) Decryption depends on key and decrypted PT MSBs. (d) Two FeFETs can be written to four levels by applying different pulse amplitudes. (e) Parameters used in experiment. (f) The MSB and LSB decryption is experimentally verified. The measured results match the logic truth table, thus validating the scheme. Larger sense margin is expected with larger memory window FeFETs.

## Encrypted Compute-in-Memory with NAND FeFET Array



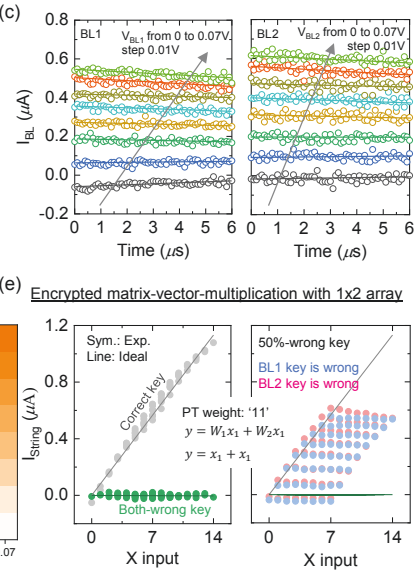
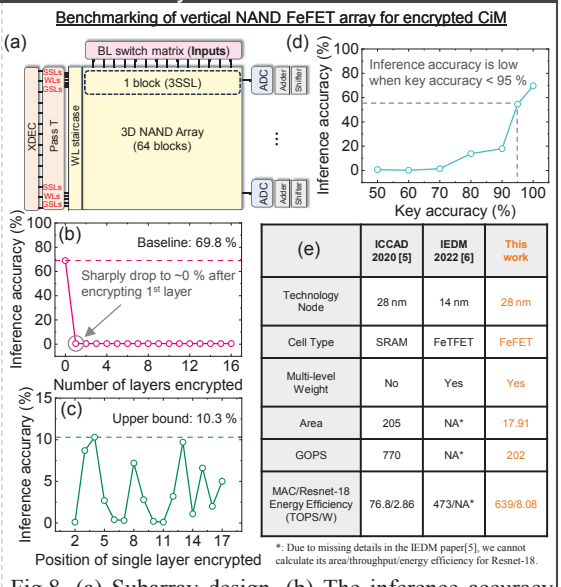




Fig.7. (a) Proposed FeFET NAND based CiM with encrypted SLC weight. (b) SEM and schematic of a 1×2 encryption array for experimental verification. (c) Measured string current for BL1 and BL2 for PT weight of '11'. (d) Total BL current for 3-bits input (8-levels) on both BLs. The BL currents exhibit linear relation. (e) Measured string currents when applying the correct key, all-wrong key, and 50%-wrong key.




| (e) | ICCAD 2020 [5] | IEDM 2022 [6] | This work |
|---|---|---|---|
| Technology Node | 28 nm | 14 nm | 28 nm |
| Cell Type | SRAM | FeTFET | FeFET |
| Multi-level Weight | No | Yes | Yes |
| Area | 205 | NA* | 17.91 |
| GOPS | 770 | NA* | 202 |
| MAC/Resnet-18 Energy Efficiency (TOPS/W) | 76.8/2.86 | 473/NA* | 639/8.08 |

*: Due to missing details in the IEDM paper[5], we cannot calculate its area/throughput/energy efficiency for Resnet-18.

Fig.8. (a) Subarray design. (b) The inference accuracy sharply drops to ~0% for encrypting the first layer only. (c) Inference accuracy varies when encrypting different layers. (d) Unauthorized attacks are prevented. (e) Comparison with SRAM-XOR-based encrypted CiM.

Authorized licensed use limited to: UNIVERSITY NOTRE DAME. Downloaded on October 04,2024 at 02:41:18 UTC from IEEE Xplore. Restrictions apply.