Design Guidelines for Oxide Semiconductor Gain Cell Memory on a Logic Platform

Shuhan Liu, *Student Member, IEEE*, Koustav Jana, *Student Member, IEEE*, Kasidit Toprasertpong, *Member, IEEE*, Jian Chen, *Member, IEEE*, Zheng Liang, Qi Jiang, Sumaiya Wahid, *Graduate Student Member, IEEE*, Shengjun Qin, Wei-Chen Chen, Eric Pop, *Fellow, IEEE*, and H.-S. Philip Wong, *Fellow, IEEE*

***Abstract*—We offer design guidelines with a top–down and bottom–up design approach for oxide semiconductor (OS) transistors, optimized for gain cell memory on a logic platform. With high-density, high-bandwidth on-chip gain cell memory, deep neural network (DNN) accelerator execution times can be shortened by 51–66%, by minimizing access to off-chip dynamic random access memory (DRAM). To balance retention time with memory bandwidth (top–down), atomic layer deposition (ALD) indium tin oxide (ITO) transistors are chosen (bottom–up). The experimentally optimized device exhibits low off-state current (2 × 10^{-18} A/μm at $V_{GS} = -0.5$ V), good on-state current (26.8 μA/μm for power supply <2 V), low subthreshold swing (SS) (70 mV/dec), and good mobility (27 cm^{2}V^{-1}s^{-1}). Using this optimized device, a gain cell memory macro with 64 rows (*WL*) × 256 columns (*BL*) is simulated at the 28 nm node operating at $V_{DD} = 0.9$ V. The simulation results show that hybrid OS-Si gain cell memory achieves 0.98× frequency and 3× density of static random access memory (SRAM), and the OS-OS gain cell memory is projected to operate at 0.5× frequency with *N* times 1.15× density of SRAM with *N*-layer of 3-D stacking.**

***Index Terms*—3-D integration, atomic layer deposition (ALD), gain cell, indium tin oxide (ITO), on-chip memory, oxide semiconductor (OS).**

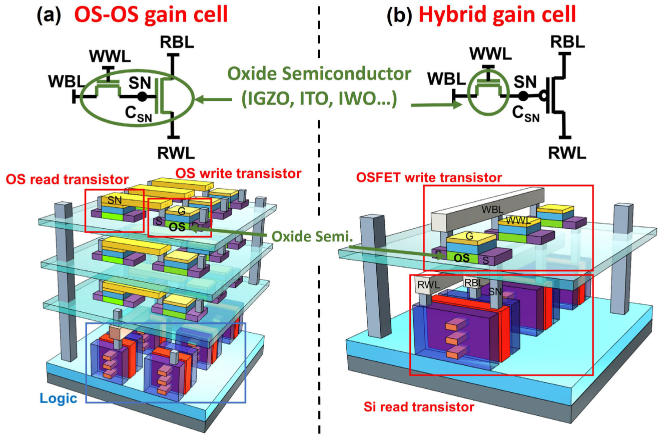

Fig. 1. Monolithic integration with logic of (a) OS–OS gain cell memory and (b) hybrid gain cell memory (OS-Si) to achieve high memory density.

I. Introduction

THE energy and delay consumed by the off-chip memory [dynamic random access memory (DRAM)] and memory-to-logic-chip data movement has become the bottleneck (known as "memory wall") for modern computing systems, especially for neural network accelerators and abundant-data computing [1]. Providing larger on-chip memory capacity with high bandwidth is a solution. Oxide semiconductor field-effect transistor (OSFET)-based gain cell memory [2], [3], [4], [5] on a logic platform provides higher density than static random access memory (SRAM) due to 3-D stacking (Fig. 1) and is an attractive complement to off-chip DRAM. An OSFET with extremely low leakage [6], [7] is used as the write transistor for long retention time, while the read transistor can adopt either OSFET for multiple-layer stacking [OS-OS gain cell in Fig. 1(a)] or Si FET for higher read speed [hybrid gain cell in Fig. 1(b)].

Designing OSFETs for gain cell memory requires more than simply choosing materials and device designs with the lowest off-state leakage current for the longest retention time. This article examines the complex interplay between memory macro level performance, such as bandwidth, memory availability, refresh period, and logic chip voltage compatibility, and device/materials targets, such as channel material, composition, thickness, and stability. To establish the interactions and balance the design tradeoff, we adopt a top–down and bottom–up design approach. The top–down design begins with memory macro simulation, from which we obtain the device specifications with the help of device modeling.

Manuscript received 4 January 2024; revised 27 February 2024; accepted 28 February 2024. Date of publication 17 April 2024; date of current version 24 April 2024. This work was supported in part by Semiconductor Research Corporation (SRC) Center for Processing with Intelligent Storage and Memory (PRISM Center), in part by SRC Center for Heterogeneous Integration of Micro Electronic Systems (CHIMES Center), in part by InnoHK Artificial intelligence (AI) Chip Center for Emerging Smart Systems (ACCESS) Center, and in part by the Member Companies of Stanford Non-Volatile Memory Technology Research Initiative (NMTRI) and Stanford SystemX Alliance. This article is an extended version of a paper presented at IEDM 2023. The review of this article was arranged by Editor G. Meneghesso. *(Corresponding authors: Shuhan Liu; H.-S. Philip Wong.)*

Shuhan Liu, Koustav Jana, Kasidit Toprasertpong, Jian Chen, Qi Jiang, Sumaiya Wahid, Shengjun Qin, Wei-Chen Chen, Eric Pop, and H.-S. Philip Wong are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: shliu98@stanford.edu; hspwong@stanford.edu).

Zheng Liang is with the Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720 USA.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2024.3372938.

Digital Object Identifier 10.1109/TED.2024.3372938

The bottom–up design identifies the material and process development targets that meet the memory macro specifications obtained in the top–down phase. Memory array macro simulations at the 28 nm node using GEMTOO [8] indicate diminishing return for memory availability and bandwidth for retention time >10 s. This provides room to trade the ultralow leakage of OSFETs for higher mobility OS materials [e.g., indium tin oxide (ITO) versus IGZO] and higher I_{ON} device designs (e.g., trading off threshold voltage, V_{TH}) for faster access and higher density. The gain cell memory can achieve higher density and comparable bandwidth as SRAM. Timeloop [9] simulations of deep neural network (DNN) accelerators with the high-density gain cell on-chip memory show 51%–66% reduction in execution time.

II. Top–Down: From Memory Macro to Device

A. Gain Cell Memory Macro Design Considerations

Gain cell memory stores information as the charge on the read transistor gate capacitance and thus needs periodical refresh operations to dynamically retain the data. During refresh operations, the memory macro is blocked from write and read random access. Memory availability is thus defined as the percentage of time that the memory is available for random access [Fig. 2(a)]. Bandwidth depends on frequency as well as availability [8] [Fig. 2(a)], due to refresh operations being an internal behavior without external data transfer.

To associate memory macro performance with device specifications, the key is the storage node (SN) degradation curve [Fig. 2(b)], which illustrates how SN voltage degrades versus time after charge is written onto the SN. The refresh point on this curve [Fig. 2(b)] is the time point when the refresh operation is performed. The choice of the refresh point explicitly sets the refresh period and also impacts the frequency, because the read on-current at the refresh point decides the worst case read delay, which is the critical path delay that determines frequency. Any time point before the read current declines to the sense amplifier resolution limit can be a possible refresh point. This time point should be chosen to be long enough to make the impacts of refreshing negligible while maintaining high enough SN voltage to provide sufficient drive current to achieve high-speed read.

For OS–OS gain cell with both write and read transistor OSFETs, the on-current (I_{ON}) and off-current (I_{OFF}) of the OSFET should be co-designed to meet the retention and frequency requirements at the same time. While for hybrid gain cell, the OSFET is only used as a write transistor with a small SN capacitance load. As such, the OSFET on-current can be relaxed (e.g., by suitable choice of the threshold voltage or channel materials) to achieve lower off-current. Meanwhile, to capitalize on high-bandwidth memory access on a logic platform, OSFETs need to be designed to operate at logic supply voltages (<1 V).

B. Gain Cell Design Space Exploration

To explore the design space of a gain cell memory macro, we use GEMTOO [8], a gain cell memory macro simulation tool that includes all the peripheral circuits (decoder, driver, sense amplifier, etc.) and memory architecture, validated with 28 nm node Si gain cell tape-out. We chose 28 nm node CMOS as a design example and conservatively took the worst case array architecture 256 row × 32 column for the simulations. For a given frequency, the bandwidth will saturate with increasing refresh period due to the saturation of memory availability with refresh period [Fig. 2(c) and (d)]. The refresh period that provides 99.9999% memory availability is defined as the saturation point, conservatively taking into consideration the design margin for device nonidealities like variation and degradation. Fig. 2(e) shows that the saturation point is around 10 s, which indicates that ~10 s retention time is long enough to make the impacts of refresh negligible and that longer retention time offers diminishing return. Note that the saturation retention time depends on frequency [Fig. 2(e)] in a roughly inverse proportional relationship. Intuitively, the product of refresh period and frequency is the number of cycles for memory random access that is not blocked by the refresh operation in a refresh period.

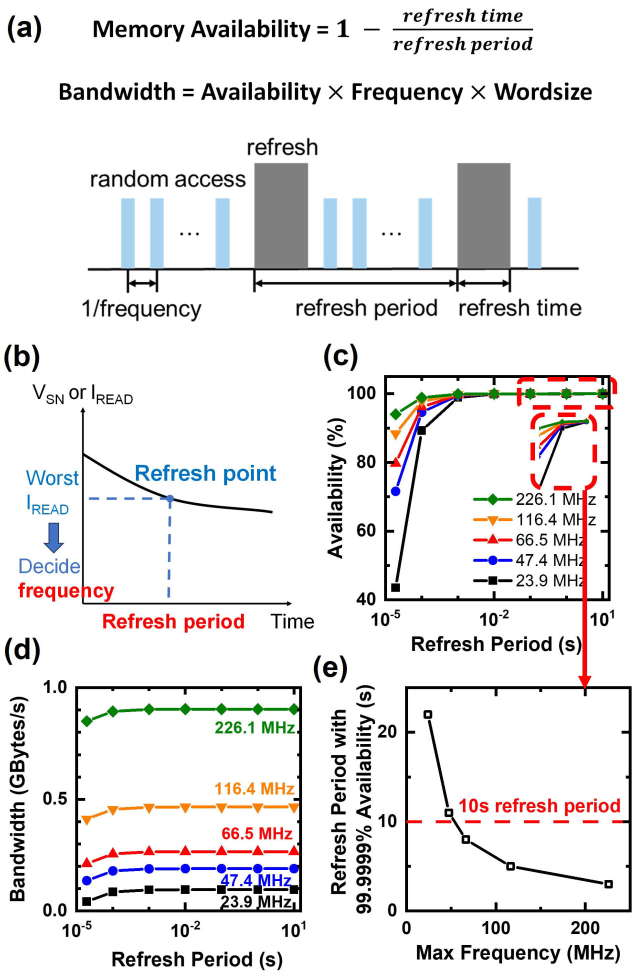

Fig. 2. (a) Definition of memory availability and bandwidth. (b) SN voltage degradation curve on which the refresh point determines refresh period and frequency. (c) Bandwidth saturates with refresh period due to (d) memory availability saturation, based on 28 nm GEMTOO simulation of 256 row × 32 column gain cell memory macro. (e) Conservatively defining 99.9999% availability as the retention saturation point, which is around 10 s for different frequencies. Baseline Si-only gain cell has a maximum frequency of 66.5 MHz.

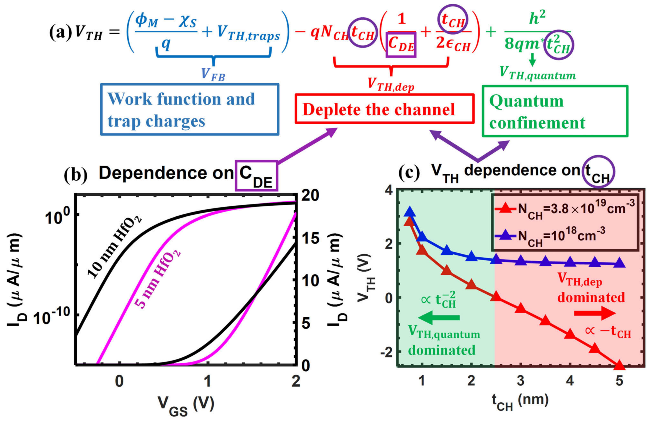

Fig. 3. Design guidelines for OSFET V_{TH} control. Simulations use a coupled Poisson–Schrodinger solver with parameters calibrated based on experimental device with 2 nm ITO and 10 nm HfO_2 (N_{CH} extracted is 3.8 × 10^{19} cm^{-3}). (a) Analytical approximation for OSFET V_{TH}. (b) Impact of gate dielectric thickness and in turn, C_{DE} on V_{TH} (simulation for 10 nm HfO_2 case calibrated to experiment). (c) Impact of channel thickness on V_{TH} for two different N_{CH} values (high and low). $V_{TH,quantum}$ becomes significant only for t_{CH} smaller than around 2.5 nm. For larger t_{CH} with high N_{CH}, $V_{TH,dep}$ dominates.

Based on the memory macro specifications of refresh period and frequency as well as voltage, we can derive the OSFET specifications with the gain cell SN voltage degradation curve. Accordingly, the OSFET off-state current should be ~1 × 10^{-18} A at $V_{DS} = V_{DD} = 1$ V and $V_{GS} = -0.3$ V, assuming $C_{SN} = 0.1$ fF and V_{SN} drops by <0.1 V during the refresh period. The $V_{GS} = -0.3$ V is achieved using negative-level shifter in [10], which adopted a dedicated design with small hardware overhead and limited shift of −0.3 V. OSFET drive current at $V_{DS} = V_{DD} = 1$ V and $V_{GS} = 0.9$ V (after accounting for 0.1 V V_{SN} degradation) as a read transistor should be ~10 μA to achieve the targeted frequency for OS–OS gain cell. For hybrid gain cell, the OSFET on-state current can be relaxed to ~1 μA, because the OSFET only serves as the write transistor that has a small capacitive load (the SN capacitance).

ITO is chosen as the OS channel material, as ITO has sufficiently high mobility [11] to meet the I_{ON} target as well as low I_{OFF} [5]. To achieve the desired $I_{OFF} = 1 \times 10^{-18}$ A at $V_{GS} = -0.3$ V, threshold voltage (V_{TH}) control is critical. As shown in Fig. 3(a), V_{TH} needs to be positive and consists of three main components: V_{FB} (flat-band voltage), $V_{TH,dep}$ (V_{GS} required to deplete the channel), and $V_{TH,quantum}$ (corresponding to conduction band shift caused by quantum confinement). V_{FB} needs to be high enough to ensure positive V_{TH} and is limited by the maximum achievable metal work-function (ϕ_M) corresponding to that of platinum with ϕ_M of 5.65 eV [12]. However, one can push for a higher effective work function by adopting techniques like addition of a dipole layer between the gate metal and dielectric [13]. $V_{TH,dep}$ has a negative contribution and is minimized by increasing the gate dielectric capacitance, C_{DE} [Fig. 3(b)] and most importantly by reducing the OS channel thickness t_{CH} [Fig. 3(c)], especially for a heavily n-doped OS with higher mobility, such as ITO, IWO [14], and In_2O_3 [15]. $V_{TH,dep}$ has a strong negative dependence on t_{CH}, suggesting that ultrathin channels are desired for OSFETs. Ultrathin OS channels can also leverage the positive V_{TH} shift due to quantum confinement [Fig. 3(c)] with a t_{CH}^{-2} dependence. Although both $V_{TH,deep}$ and $V_{TH,quantum}$ lead to a more positive V_{TH} for thinner channels, their relative contributions depend on the effective mass (m^*) and channel doping (N_{CH}) values. We assume $m^* = 0.3$ m_0 for our calculations [16] and neglect any second-order effect, such as the t_{CH} dependence of m^*. Based on this, we observe that the quantum confinement effects become significant only for t_{CH} less than around 2.5 nm, where V_{TH} starts to show a t_{CH}^{-2} dependence for smaller t_{CH}. For larger t_{CH}, $V_{TH,dep}$ dominates with its value strongly dependent on N_{CH} as suggested by the two curves in Fig. 3(c). When N_{CH} is large (3.8 × 10^{19} cm^{-3}), V_{TH} shows a strong ($\propto - t_{CH}$) drop with increasing t_{CH}, whereas the low N_{CH} case (10^{18} cm^{-3}) shows negligible t_{CH} dependence for thicker channels. Thinner channels also help mitigate short-channel effects when channel length is scaled down to 28 nm and below, providing better immunity to V_{TH} and subthreshold swing (SS) roll-off. However, thin t_{CH} is also accompanied by mobility degradation that is mainly attributed to increased surface roughness and thickness fluctuation scattering [17]. This analysis leads to the choice of $t_{CH} = 2$ to 4 nm as the design target and the need for precise control of t_{CH}.

III. Bottom–Up: From Material to Device

A. ALD ITO Material and Process Development

The mobility of ITO transistors with sputtered ITO films degrades significantly for ultrathin sputtered films due to surface roughness scattering [18]. Ultrathin sputtered films also face challenges in achieving good uniformity, especially in the case of large-scale integration. Thus, we target atomic layer deposition (ALD) for ITO thin films [20].

The ALD ITO process uses n-supercycle of m-cycle ALD In_2O_3 and one-cycle of ALD SnO_2 [21], [22]. O_2 plasma is used as the oxygen precursor to avoid H doping during deposition. Tetrakis(dimethylamino)tin(IV) (TDMA-Sn) precursor is chosen for SnO_2 ALD. Indium cyclopentadienyl (CpIn) precursor [21] and trimethylindium (TMIn) precursor [22] are compared for the In_2O_3 ALD process. Based on preliminary X-ray photoelectron spectroscopy (XPS) tests, TMIn precursor temperature of 40 °C gives enough vapor pressure for In_2O_3 film deposition, while CpIn precursor needs 125 °C. With the precursor temperature set to produce the same vapor pressure and precursor pulse time set at the ALD self-saturation point, the ITO films deposited by CpIn and TMIn precursor with various compositions are characterized by XPS. ITO with TMIn precursor shows better incorporation of Sn doping and thus lower indium concentration [23], resulting in transistors with better stability and reliability [24]. Thus, TMIn precursor is used as the standard process in the following device optimization. Growth rates for 19:1 and 9:1 composition are 1.23 and 1.31 Å/cycle, respectively.

B. ALD ITO FET Device Optimization

To optimize the ALD ITO material for transistor and gain cell memory, back-gated ITO FETs were fabricated with the

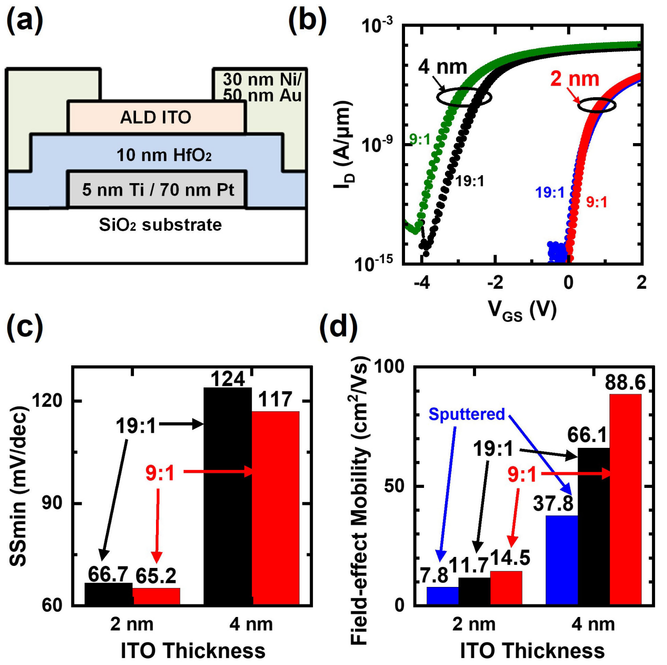

Fig. 4. (a) Device cross section and process flow for back-gated ALD ITO FET. (b) Measured transfer curve and extracted, (c) SS, and (d) field-effect mobility for pristine nonannealed ALD ITO FET with different ALD ITO films by varying the In:Sn composition and thickness. Device $L_G = 2$ μm measured with $V_{DS} = 0.5$ V.

device structure shown in Fig. 4(a), with ITO composition of 19:1 and 9:1, and thickness of 4 and 2 nm deposited by ALD with the substrate temperature at 200 °C. Pt, deposited by *e*-beam evaporation at room temperature, is used as the gate metal for its high work function, and Ni is used as the contact metal for low contact resistance [18], [19]; 10 nm HfO_2 is deposited by ALD at 200 °C as the gate dielectric for lower gate leakage. ITO is patterned and wet etched with diluted HCl, and HfO_2 is dry etched by inductively coupled plasma (ICP). The whole fabrication process is under 200 °C, compatible with back-end of the line (BEOL) integration.

The measured I_D–V_{GS} curve for pristine nonannealed devices in Fig. 4(b) shows that V_{TH} is very negative for ALD ITO FET with 4 nm ITO films, while positive for the 2 nm ones, in agreement with analysis shown in Fig. 3. SS and field-effect mobility are extracted and shown in Fig. 4(c) and (d). ALD ITO FET with 9:1 composition has better SS and field-effect mobility than 19:1 composition. ALD ITO FETs with 4 nm ITO films have higher mobility but worse SS than 2 nm films. Note that ALD ITO films already have better mobility than sputtered ITO films due to smaller film roughness and thus less surface scattering.

Annealing the 4 nm ITO device at 300 °C in O_2 can shift V_{TH} positively, but with annealing time >5 min, V_{TH} still saturates at negative voltage −1 V and −0.5 V for 19:1 and 9:1 composition, respectively [23], which leads us to the final choice of 2 nm ITO films for a gain cell demonstration. To further study the effects of temperature, the 2 nm ITO device is annealed at 300 °C in Ar. After annealing, the device with 9:1 composition shows less shift in V_{TH} and less degradation in SS [23], which indicates better temperature stability than 19:1.

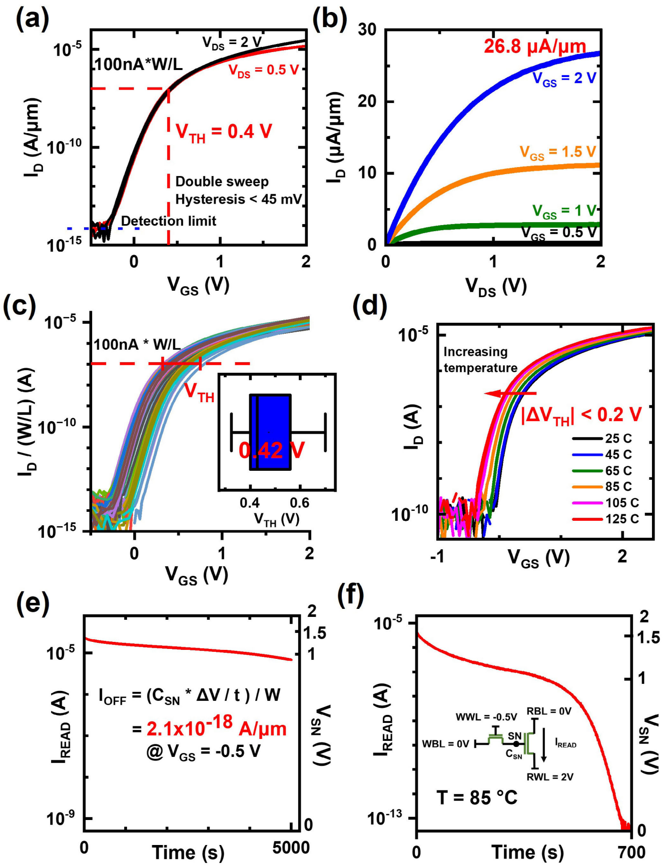

Fig. 5. (a) Transfer and (b) output curves of the $W/L = 4/1$ μm/μm nonannealed ALD ITO FET with 9:1 composition and 2 nm ITO that shows $V_{TH} = 0.4$ V, on-current of 26.8 μA/μm, and small hysteresis <45 mV. (c) Small device variation with median $V_{TH} = 0.42$ V for 34 devices fabricated at different times. (d) V_{TH} shift <0.2 V under 125 °C temperature stability test. OS–OS gain cell measurement setup and measured retention at (e) room temperature and (f) 85 °C.

From the above results and discussion, the ALD ITO process with TMIn precursor, 9:1 composition, and 2 nm thickness yields the optimized materials for the target design of OSFET and gain cell.

IV. Optimized OSFET and Gain Cell

A. Experimental Results of Optimized Device

Fig. 5(a) and (b) shows the transfer and output curves of $L_G = 1$ μm optimized ALD ITO FET without annealing. The device has a positive V_{TH} of 0.4 V, defined as V_{GS} at $I_D = 100$ nA × W/L, as well as small hysteresis <45 mV and small SS of 70 mV/dec. The on-state current is also as high as 26.8 μA/μm at $V_{DS} = 2$ V and V_{GS}–$V_{TH} = 1.6$ V, due to good field-effect mobility of 27 $cm^2V^{-1}s^{-1}$. Device variation is shown in Fig. 5(c), with V_{TH} spread within 0.15 V and median V_{TH} of 0.42 V for a sample size of 34 devices fabricated at different times.

The devices are further characterized with stability tests under temperature and bias stresses. Good stability is shown with low V_{TH} shift <0.2 V under temperature stress up to 125 °C [Fig. 5(d)], low positive bias stress (PBS) shift <0.35 V under 2 V bias stress for 1000 s, and low negative

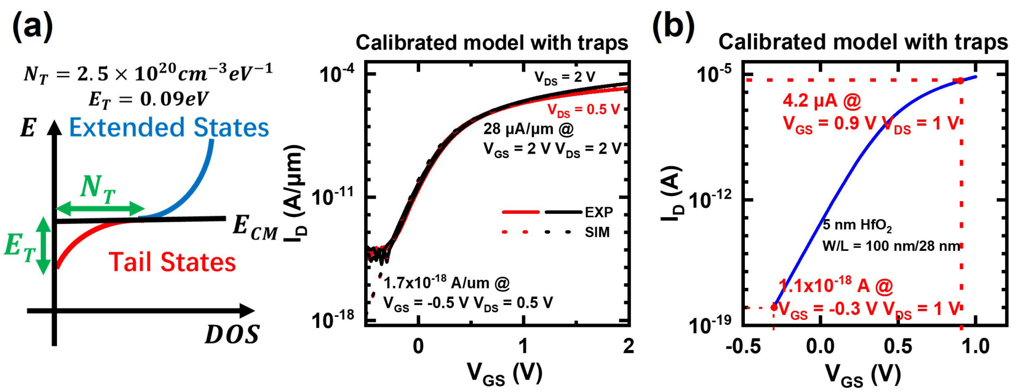

Fig. 6. (a) Calibrated TCAD model with extracted traps from the experimental results of $W/L = 4/1$ μm/μm device and simulated fitting curve. E_{CM}: Conduction band mobility edge. (b) Simulated ITO FET with $L = 28$ nm, based on experimentally calibrated model including traps.

bias stress (NBS) shift <0.1 V under −2 V bias stress for 1000 s [23]. Although OSFET PBS shift is a problem for OS-OS gain cell when storing data "1" with positive gate voltage on the transistor, the hybrid gain cell is much less impacted because Si FET is used as the read transistor. For the OSFET write transistor, ∼ns positive write pulse may induce small stress and shift, which can be recovered during the standby state without any treatment [25].

With the optimized device, OS–OS gain cell was fabricated and measured under standby write wordline (WWL) of −0.5 V. After a 5000 s retention test, the read current dropped from 24 to 7 μA [Fig. 5(e)]. By converting read current to SN voltage with the read transistor I_D–V_{GS} curve, the off-current is extracted to be 2×10^{-18} A/μm. The SN degradation curve at 85 °C was also measured as shown in Fig. 5(f). Initially, the SN voltage degradation rate decreases with time due to smaller voltage difference between SN and WBL. After V_{SN} degrades below V_{TH}, the read current dropped quickly due to exponential dependence on V_{SN} in the subthreshold region. V_{TH} has a positive shift after ∼500 s stress time when measuring long retention of data "1," which is a critical issue for OS–OS gain cell and can be solved by hybrid gain cell instead. After excluding PBS shift, the off-state current at 85 °C is extracted to be 4×10^{-17} A/μm, which is attributed to the negative V_{TH} shift of 0.1 V at 85 °C [Fig. 5(d)] as well as SS degradation at higher temperature.

B. Modeling, Simulation, and Benchmark

Based on the experimental results of the 1 μm long channel ALD ITO FET, we develop and calibrate an OSFET TCAD model in Sentaurus, based on a custom material parameter file for OS. We consider a bottom-gated structure similar to our experimental devices with an n-type OS channel with uniform doping (including the regions under the source/drain contacts). Parameters, such as mobility (μ) and doping (N_{CH}), have been calibrated to match the experimental data. OSFETs operate as junctionless transistors, where the on state corresponds to accumulation mode and off state corresponds to depletion mode, with the model considering this mode of operation for OSFETs. The model also accounts for the trap-like tail states inside the channel due to the amorphous nature of OS, as well as the interface traps. This is done by incorporating acceptor-type traps with exponential density of state (DoS) distribution based on experimental extraction to our model, which fits the experimental data well [Fig. 6(a)]. With the experimentally calibrated model, we simulated a scaled 28 nm short-channel transistor assuming the same ITO material properties and the HfO_2 gate dielectric reduced from 10 to 5 nm. The simulated device in Fig. 6(b) can meet the specifications that we set in Section II. Note that this projection is under the assumption that the ITO material properties (such as μ and N_{CH}) remain unaltered on scaling down channel length. In addition, we assume ideal contacts and neglect any velocity saturation. Also, scaling down the dielectric thickness comes at the risk of increased gate leakage, time-dependent dielectric breakdown (TDDB), stress induced leakage current (SILC), PBS/NBS of the read as well as write transistors leading to

TABLE I
GEMTOO MEMORY MACRO SIMULATION BASED ON SIMULATED 28 nm ALD ITO FET INDICATES THAT OS–OS GC AND HGC HAVE LONG RETENTION AND HIGH FREQUENCY

28nm node, V_{DD} = 0.9V, sub-array size 64 row x 256 col.				
	SRAM[26]	Si GC# [7]	OS GC#	HGC#
Cell size* (μm²)	0.16	0.14	0.14/N	0.06
Refresh Period		19 μs	9 s	9 s
Max Freq. (MHz)	735	242	345	721
Bandwidth (GB/s)		7.6	11	23

Simulated with GEMTOO
* SRAM -- pushed design rule; GC -- logic design rules. For OS gain cell in this work, equivalent cell size depends on 3D stacking number (N) of layers.

TABLE II
USING GC MEMORY WITH HIGH DENSITY, DNN EXECUTION TIME IS REDUCED ∼50% COMPARED WITH SRAM BASELINE FOR DIFFERENT "DNN MODEL:LAYER," SIMULATED USING TIMELOOP, A DNN ACCELERATOR SIMULATOR WITH DATA FLOW OPTIMIZATION FOR ARCHITECTURE DESIGN

DNN with 2X Cache Density/Size Compared with Baseline*				
	ResNet50:12	ResNet50:14	VGG16:14	UNet:2
Execution time	0.43x	0.49x	0.34x	0.42x

*Baseline: 32KB cache per 64 ALU with off-chip DRAM bandwidth of 64bit per cycle potential failure of gain cell operation. Accounting for all the above factors is critical for OSFET development to enable the scaling of gain cell memory.

OS–OS gain cell (GC) and hybrid gain cell (HGC) memory macros based on this 28 nm device are simulated with GEMTOO at the 28 nm node and compared with SRAM (Table I). Hybrid gain cell has 3× density and comparable speed with SRAM. In contrast, the OS–OS gain cell has lower speed (0.5× frequency of SRAM) and can achieve much higher density (N times 1.15× density of SRAM) with N layers provided by 3-D stacking. Both OS–OS and hybrid gain cell memories exhibit long retention of 9 s, much higher than Si gain cell and enough to make the impact of refresh negligible. To evaluate the system performance with our high-density gain cell on-chip memory, we used Timeloop [9], a DNN accelerator simulator with data flow optimization. The gain cell with high density achieves 51%–66% reduction in execution cycle time (Table II) by minimizing off-chip DRAM accesses.

C. Scalability

As observed and discussed in Section II-B, saturation retention is inversely proportional to frequency, assuming that the number of cycles for memory random access in a refresh period remains constant. For more advanced technology nodes, if the clock frequency is scaled up by k, then the required retention time can be scaled down by k. According to the retention equation $t_R = C_{\mathrm{ox}}\ WL \times \Delta V/I_{\mathrm{OFF}}$, because the equivalent oxide thickness (EOT) and V_{DD} almost saturate for very advanced technology nodes [27], the off-state current density I_{OFF}/W of OSFET write transistor should stay constant to satisfy the required retention. Similarly, the on-state current density I_{ON}/W of write OSFET should also remain constant to meet the required frequency. The read transistor scaling depends on bitline interconnect scaling, because $\tau = C_{\mathrm{BL}} \times \Delta V/I_{\mathrm{ON}}$. For a fixed wire aspect ratio, the bitline capacitance is scaled down by k [28]. Thus, the read transistor on-state current I_{ON} should remain the same with scaling to achieve target frequency. In terms of this, hybrid gain cell with Si FET as the read transistor has better scalability. Wire resistance will increase with scaling and with larger array size, and thus, wire RC delay may dominate, which can be addressed by interconnect material and process engineering [28] and memory architecture optimization. The 3× density benefits over SRAM can be scaled down to 28 nm node with planar transistor layout design. For FinFET technology below 28 nm, the array layout needs to be redesigned with process design kits (PDKs) not readily available to academia and subject to further study. In any case, the estimated density of the gain cell will still be higher than SRAM because gain cell memory has fewer transistors per cell and employs 3-D stacking.

For the scalability of OSFET, although a scaled channel length of 8 nm was experimentally demonstrated with small short-channel effect [29], several challenges still need to be addressed within gain cells. With ultrathin OS channel films, the source/drain series resistance may limit I_{ON} for scaled devices. The process optimization of special treatment to channel contact region can be used to reduce the series resistance [30]. The gate leakage will increase for scaled devices with thin gate dielectric and may lead to reduced retention. High-κ gate dielectric [31] with good interface to OS channel material should be studied. Voltage scaling and V_{TH} control is the most challenging with SS limit of 60 mV/dec, especially for very advanced technology node with low operation voltage. More research on physics understanding is needed to improve voltage scaling and V_{TH} control. Higher voltage domain (e.g., I/O transistor) and/or chiplet integration [32] are possible avenues for exploration.

V. Conclusion

This work established design guidelines for gain cell memory on logic platform using a mixed top–down and bottom–up design methodology. With the co-design from memory macro to device and from material to device, we studied gain cell design trade-offs and provided experimental proof-of-concept for OS transistors to be used in high-density gain cell memory on logic platform operating at $V_{\mathrm{DD}} = 0.9$ V. The optimized gain cell achieved good trade-off with high density, high bandwidth, and long retention compared to SRAM. This motivates further research on OS-based gain cell memory, potentially providing relief for the memory wall by providing large-capacity on-chip memory with adequate bandwidth. The same design procedure can be generally used for memory macros optimized for other applications, as well as gain cell memory using other OS materials. Variation control for large memory arrays and integration costs must be further studied because these are key factors for industry adoption.

Acknowledgment

The authors would like to thank Xinyi Wen, Roger Howe (Stanford University), and Vijay Narasimhan (EMD Electronics) for helpful discussion related to the ALD process. They would also like to thank Prof. Alberto Sangiovanni-Vincentelli (University of California, Berkeley) for reviewing the manuscript and collaboration.

References

[1] M. M. S. Aly et al., "The N3XT approach to energy-efficient abundant-data computing," *Proc. IEEE*, vol. 107, no. 1, pp. 19–48, Jan. 2019, doi: 10.1109/JPROC.2018.2882603.

[2] S. Shukuri, T. Kure, and T. Nishida, "A complementary gain cell technology for sub-1 V supply DRAMs," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 1992, pp. 1006–1008, doi: 10.1109/IEDM.1992.307530.

[3] A. Belmonte et al., "Capacitor-less, long-retention (>400s) DRAM cell paving the way towards low-power and high-density monolithic 3D DRAM," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2020, pp. 28.2.1–28.2.4, doi: 10.1109/IEDM13553.2020.9371900.

[4] H. Ye et al., "Double-gate W-doped amorphous indium oxide transistors for monolithic 3D capacitorless gain cell eDRAM," in *IEDM Tech. Dig.*, Dec. 2020, pp. 28.3.1–28.3.4, doi: 10.1109/IEDM13553.2020.9371981.

[5] K. Toprasertpong et al., "Co-designed capacitive coupling-immune sensing scheme for indium-tin-oxide (ITO) 2T gain cell operating at positive voltage below 2 V," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Kyoto, Japan, Jun. 2023, pp. 1–2, doi: 10.23919/vlsitechnologyandcir57934.2023.10185433.

[6] H. Inoue et al., "Nonvolatile memory with extremely low-leakage indium-gallium-zinc-oxide thin-film transistor," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 2258–2265, Sep. 2012, doi: 10.1109/JSSC.2012.2198969.

[7] A. Belmonte et al., "Lowest $\mathrm{I_{OFF}} < 3\times$ 10–21 A/μm in capacitorless DRAM achieved by reactive ion etch of IGZO-TFT," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Kyoto, Japan, Jun. 2023, pp. 1–2, doi: 10.23919/vlsitechnologyandcir57934.2023.10185398.

[8] A. Bonetti, R. Golman, R. Giterman, A. Teman, and A. Burg, "Gain-cell embedded DRAMs: Modeling and design space," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 28, no. 3, pp. 646–659, Mar. 2020, doi: 10.1109/TVLSI.2019.2955933.

[9] A. Parashar et al., "Timeloop: A systematic approach to DNN accelerator evaluation," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, Madison, WI, USA, Mar. 2019, pp. 304–315, doi: 10.1109/ISPASS.2019.00042.

[10] J. Narinx et al., "A 24 kb single-well mixed 3T gain-cell eDRAM with body-bias in 28 nm FD-SOI for refresh-free DSP applications," in *Proc. IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, Macau, China, Dec. 2019, pp. 219–222, doi: 10.1109/A-SSCC47793.2019.9056985.

[11] S. Wahid, A. Daus, A. Kumar, H.-S. P. Wong, and E. Pop, "First demonstration of dual-gated indium tin oxide transistors with record drive current ~2.3 mA/μm at L ≈ 60 nm and VDS = 1 V," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2022, pp. 12.5.1–12.5.4, doi: 10.1109/IEDM45625.2022.10019544.

[12] B. Ofuonye, J. Lee, M. Yan, C. Sun, J.-M. Zuo, and I. Adesida, "Electrical and microstructural properties of thermally annealed Ni/Au and Ni/Pt/Au Schottky contacts on AlGaN/GaN heterostructures," *Semiconductor Sci. Technol.*, vol. 29, no. 9, Sep. 2014, Art. no. 095005, doi: 10.1088/0268-1242/29/9/095005.

[13] R. Bao et al., "Selective enablement of dual dipoles for near bandedge multi-Vt solution in high performance FinFET and nanosheet technologies," in *Proc. IEEE Symp. VLSI Technol.*, Honolulu, HI, USA, Jun. 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265010.

[14] T. Kizu et al., "Low-temperature processable amorphous In-W-O thin-film transistors with high mobility and stability," *Appl. Phys. Lett.*, vol. 104, no. 15, Apr. 2014, Art. no. 152103, doi: 10.1063/1.4871511.

[15] M. Si, Z. Lin, Z. Chen, X. Sun, H. Wang, and P. D. Ye, "Scaled indium oxide transistors fabricated using atomic layer deposition," *Nature Electron.*, vol. 5, no. 3, pp. 164–170, Feb. 2022, doi: 10.1038/s41928-022-00718-w.

[16] S. K. Dargar and V. M. Srivastava, "Design and analysis of IGZO thin film transistor for AMOLED pixel circuit using double-gate tri active layer channel," *Heliyon*, vol. 5, no. 4, Apr. 2019, Art. no. e01452, doi: 10.1016/j.heliyon.2019.e01452.

[17] S. Reggiani, E. Gnani, A. Gnudi, M. Rudan, and G. Baccarani, "Low-field electron mobility model for ultrathin-body SOI and double-gate MOSFETs with extremely small silicon thicknesses," *IEEE Trans. Electron Devices*, vol. 54, no. 9, pp. 2204–2212, Sep. 2007, doi: 10.1109/TED.2007.902899.

[18] M. Si et al., "Indium-tin-oxide transistors with one nanometer thick channel and ferroelectric gating," *ACS Nano*, vol. 14, no. 9, pp. 11542–11547, Sep. 2020, doi: 10.1021/acsnano.0c03978.

[19] S. Wahid et al., "Effect of Top-Gate Dielectric Deposition on the Performance of Indium Tin Oxide Transistors," in *IEEE Electron Device Lett.*, vol. 44, no. 6, pp. 951–954, June 2023, doi: 10.1109/LED.2023.3265316.

[20] Z. Zhang et al., "Atomically thin indium-tin-oxide transistors enabled by atomic layer deposition," *IEEE Trans. Electron Devices*, vol. 69, no. 1, pp. 231–236, Jan. 2022, doi: 10.1109/TED.2021.3129707.

[21] J. W. Elam, D. A. Baker, A. B. F. Martinson, M. J. Pellin, and J. T. Hupp, "Atomic layer deposition of indium tin oxide thin films using nonhalogenated precursors," *J. Phys. Chem. C*, vol. 112, no. 6, pp. 1938–1945, Feb. 2008, doi: 10.1021/jp7097312.

[22] B. Zhao et al., "Atomic layer deposition of indium-tin-oxide as multifunctional coatings on V_2O_5 thin-film model electrode for lithium-ion batteries," *Adv. Mater. Interface*, vol. 7, no. 23, Dec. 2020, Art. no. 2001022, doi: 10.1002/admi.202001022.

[23] S. Liu et al., "Gain cell memory on logic platform—Device guidelines for oxide semiconductor transistor materials development," in *IEDM Tech. Dig.*, San Francisco, CA, USA, 2023, pp. 1–4, doi: 10.1109/IEDM45741.2023.10413726.

[24] Y. Hu et al., "Theoretical and empirical insight into dopant, mobility and defect states in W doped amorphous In_2O_3 for high-performance enhancement mode BEOL transistors," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2022, pp. 8.5.1–8.5.4, doi: 10.1109/IEDM45625.2022.10019366.

[25] E. Fortunato, P. Barquinha, and R. Martins, "Oxide semiconductor thin-film transistors: A review of recent advances," *Adv. Mater.*, vol. 24, no. 22, pp. 2945–2986, Jun. 2012, doi: 10.1002/adma.201103228.

[26] S.-L. Wu et al., "A 0.5-V 28-nm 256-kb mini-array based 6T SRAM with Vtrip-tracking write-assist," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 7, pp. 1791–1802, Jul. 2017, doi: 10.1109/TCSI.2017.2681738.

[27] (2022). *International Roadmap for Devices and Systems 2022 Edition*. [Online]. Available: https://irds.ieee.org/editions/2022/executive-summary

[28] R. Brain, "Interconnect scaling: Challenges and opportunities," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2016, pp. 9.3.1–9.3.4, doi: 10.1109/IEDM.2016.7838381.

[29] Y.-K. Liang et al., "Aggressively scaled atomic layer deposited amorphous $InZnO_x$ thin film transistor exhibiting prominent short channel characteristics (SS= 69 mV/dec.; DIBL = 27.8 mV/V) and high Gm(802 μS/μm at VDS = 2 V)," in *Proc. IEEE Symp. VLSI Technol. Circuits*, Kyoto, Japan, Jun. 2023, pp. 1–2, doi: 10.23919/vlsitechnologyandcir57934.2023.10185343.

[30] S. Subhechha et al., "First demonstration of sub-12 nm Lg gate last IGZO-TFTs with oxygen tunnel architecture for front gate devices," in *Proc. Symp. VLSI Technol.*, Kyoto, Japan, Jun. 2021, pp. 1–2.

[31] B. Wang, W. Huang, L. Chi, M. Al-Hashimi, T. J. Marks, and A. Facchetti, "High-*k* gate dielectrics for emerging flexible and stretchable electronics," *Chem. Rev.*, vol. 118, no. 11, pp. 5690–5754, Jun. 2018, doi: 10.1021/acs.chemrev.8b00045.

[32] J. Wuu et al., "3D V-cache: The implementation of a hybrid-bonded 64 MB stacked cache for a 7 nm × 86–64 CPU," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 65, San Francisco, CA, USA, Feb. 2022, pp. 428–429, doi: 10.1109/ISSCC42614.2022.9731565.