(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) **World Intellectual Property Organization**
International Bureau

(43) **International Publication Date**
27 June 2024 (27.06.2024)

WIPO | PCT

(10) **International Publication Number**
WO 2024/137291 A2

(51) **International Patent Classification:**
H10B 12/10 (2023.01)

(21) **International Application Number:**
PCT/US2023/083667

(22) **International Filing Date:**
12 December 2023 (12.12.2023)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**
63/434,627 22 December 2022 (22.12.2022) US

(71) **Applicant: THE BOARD OF TRUSTEES OF THE UNIVERSITY OF ILLINOIS** [US/US]; 352 Henry Administration Building, 506 South Wright Street, Urbana, IL 61801 (US).

(72) **Inventors: CAO, Qing**; 4409 Curtis Meadow Drive, Champaign, IL 61822 (US). **CUI, Jinsong**; 506 East Tomaras Avenue, Savoy, IL 61874 (US).

(74) **Agent: RITTNER, Mindy, N.** et al.; Crowell & Moring LLP, 455 North Cityfront Plaza Drive - Suite 3600, Chicago, IL 60611 (US).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH,

(54) **Title:** SOLID-STATE ELECTROCHEMICAL RANDOM ACCESS MEMORY (ECRAM) AND METHODS OF MAKING AND OPERATING A SOLID-STATE ECRAM

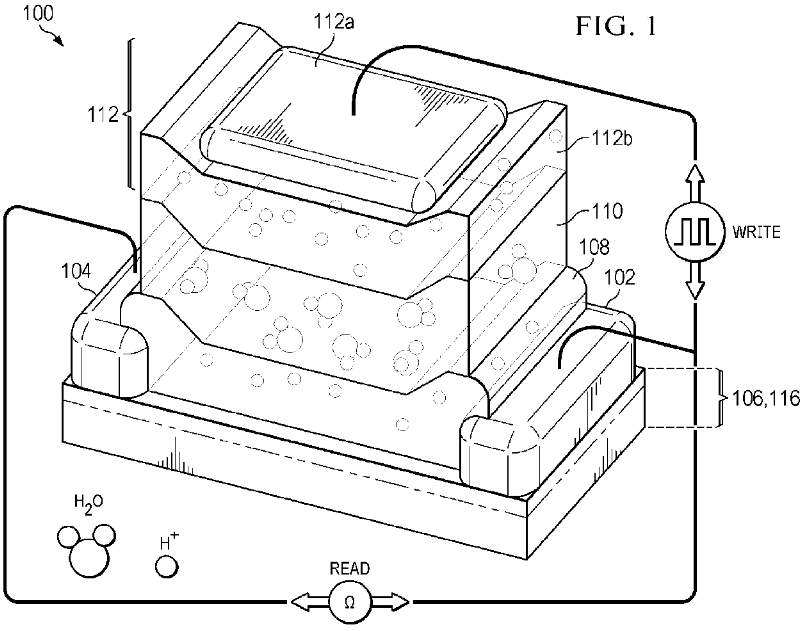


FIG. 1

(57) **Abstract:** An electrochemical random-access memory cell comprises: source and drain electrodes on a substrate; a channel layer comprising an inorganic protonic intercalatable material on the substrate between the source and drain electrodes; a solid-state protonic electrolyte layer on the channel layer; and a gate electrode bilayer comprising the inorganic protonic intercalatable material on the solid-state protonic electrolyte layer. The inorganic protonic intercalatable material may comprise a hydrogenated tungsten oxide, e.g., $H_{x1}WO_3$for the channel layer, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1$ 0.4, and $H_{X2}WO_3$ for the gate electrode bilayer, where x2 represents initial hydrogen concentration of the gate electrode bilayer and

[Continued on next page]

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

$0 < x2 < 0.4$. Upon application of a voltage pulse to the gate electrode bilayer, protons are reversibly inserted into the channel layer from the solid-state protonic electrolyte layer to modulate channel conductance.

SOLID-STATE ELECTROCHEMICAL RANDOM ACCESS MEMORY (ECRAM) AND METHODS OF MAKING AND OPERATING A SOLID-STATE ECRAM

RELATED APPLICATION

[0001] The present patent document claims the benefit of priority under 35 U.S.C. 119(e), to U.S. Provisional Patent Application 63/434,627, which was filed on December 22, 2022, and is hereby incorporated by reference in its entirety.

FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0002] This invention was made with government support under 1950182 and 2139185 awarded by the National Science Foundation. The government has certain rights in the invention.

TECHNICAL FIELD

[0003] The present disclosure is related generally to non-volatile memory and more particularly to solid-state electrochemical random access memory.

BACKGROUND

[0004] Deep learning has made phenomenal progress, giving computers capability to even outperform humans in complicated tasks. However, such improvement comes with the cost of aggressively increased depth and size of the neural-network models utilized, and thus exponentially increasing computational load and energy consumption to train and execute them. In-memory-computing architectures based on crossbar arrays of non-volatile memory are considered for deep-learning accelerators. With their compatibility with the back-end-of-line (BEOL) integration on top of silicon logic circuits and excellent scalability, memristive-crossbar-based accelerators allow on-chip data storage with density and capacity high enough to eliminate the off-chip memory access during data-intensive computations, and can parallelly execute analog matrix operations without incurring data movements. They are therefore projected to achieve much higher computational efficiency with lower chip-area cost than using microprocessors based on the conventional von Neumann architecture.

[0005] Despite their promise of breaking the memory wall and enabling high parallelism, there are hurdles to implementation of memristive arrays. To successfully carry out the training of deep neural networks with memristive arrays, a fundamental objective is to accurately map the weight movement in the training algorithms into the characteristics of the memory hardware, where the device conductance change, representing the weight modulation, may be both symmetric, *i.e.*, independent of the sign of the stimulus, and precise, *i.e.*, with small cycle-to-cycle variability. Otherwise, the training accuracy may be dramatically affected. However, this requirement remains a significant challenge as for most memristive devices, including those based on phase-change materials (PCMs), filament-forming metal oxides, and ferroelectric oxides, their conductance increase (potentiation) and decrease (depression) characteristics are highly asymmetric with large variability as limited by their underlying material and device physical mechanisms. For example, for metal-oxide memristors, they typically have more abrupt potentiation as assisted by the positive feedback between the filament growth and the electric field, and they also demonstrate large variability as the filament-formation process is intrinsically stochastic. Although such non-ideal asymmetric programming can be compensated through adopting more complicated training algorithms, it requires doubling the chip area with penalties on both latency and energy consumption. Some recent work demonstrated memristors with improved symmetry by employing complex switching media, but they still suffer from large variability, and many of them are not compatible with direct BEOL integration.

[0006] Considering these challenges, electrochemical synaptic transistors have become a promising candidate to realize deep-learning accelerators based on the in-memory-computing architecture. As a device specifically designed for neuromorphic computing rather than digital information storage, their channel conductance can be precisely modulated by the electrochemical intercalation reactions controlled with the bias applied on the separate gate terminal, which provides multistate analog programming with the desired high symmetry and low variability. However, the technological promise of the current electrochemical random-access memory (ECRAM) demonstrations is severely limited by their poor compatibility with the silicon complementary metal-oxide-semiconductor (CMOS) technology, which must be monolithically integrated together with the memristive arrays as part of the accelerator to provide peripheral functions. The electrolytes employed in most existing ECRAM devices are either liquid or organic polymers. Despite their good performance, it is very difficult, if not impossible, to incorporate these devices in circuits as scaled memory cells with the long-term stability and reliability required for electronic applications, and they are functional only in a controlled environment.

BRIEF DESCRIPTION OF THE FIGURES

[0007] FIG. 1 is a schematic illustrating an example of the structure and read/write operations of a CMOS-compatible, all-inorganic protonic ECRAM cell.

[0008] FIG. 2 is a cross-sectional scanning transmission electron microscopy (STEM) micrograph showing an exemplary device gate stack, where the scale bar is 50 nm.

[0009] FIG. 3 is a top-down view of a memory cell comprising an ECRAM cell integrated with a silicon MOSFET.

[0010] FIG. 4A is a top-down view of an exemplary fabricated ECRAM-cell array.

[0011] FIG. 4B is a schematic of a bonded chip including the ECRAM-cell array.

[0012] FIGS. 5A-5I are schematics illustrating one example of a process flow to fabricate an ECRAM cell.

[0013] FIG. 6 illustrates an exemplary hydrogen-spillover process to intercalate protons into a WO_3-based channel and gate electrode of an ECRAM cell or cell array.

[0014] FIGS. 7A-7L are schematics illustrating an exemplary process flow to integrate an ECRAM cell with a silicon MOSFET, *e.g.* to form a 1-transistor-1-ECRAM cell.

[0015] FIG. 8 shows results from programming an ECRAM cell (channel length L_{ch}=10 μm and width W=3 μm) with gate-voltage pulses (64 potentiation then 64 depression pulses with amplitude of ±4 V and width of 3 sec), where the dotted line is the mirror image of the potentiation curve; the absolute device channel conductance and the relative modulation with regard to the baseline conductance (G/G0) are used for the left and right y-axis, respectively.

[0016] FIG. 9A is a schematic of a protonic ECRAM employing only a metal (*e.g.*, Cr/Au) as the gate electrode.

[0017] FIG. 9B shows results from symmetric gate-current pulse programming of the (FIG. 9A) device channel conductance under gate-current pulses (100 potentiation then 100 depression with the gate-current pulse amplitude of 100 pA and width of 3 sec).

[0018] FIG. 9C shows results from gate-voltage pulse programming of the same (FIG. 9A) device under gate-voltage pulses (50 potentiation then 50 depression with the gate-voltage pulse amplitude of ±4 V and width of 3 s), the asymmetric response is caused by the non-zero built-in open-circuit potential.

[0019] FIGS. 10 and 11 provide zoom-in views of the programming characteristics showing the readout of discrete conductance states with sufficient noise margin (FIG. 10, source-drain bias for read was 0.1 V) and the gate current during the weight-update (FIG. 11), driven by gate voltage pulses (blue, right axis). Current values measured during read and write operations are represented with triangular and circular symbols, respectively.

[0020] FIG. 12 demonstrates reproducible and highly symmetric programming with 10 μsec and 5 μsec write pulses (amplitude=±4 V).

[0021] FIG. 13 demonstrates reproducible and symmetric switching characteristics of a scaled (L_{ch}=W=150 nm) ECRAM modulated with 10 μsec and 5 μsec write pulses (amplitude=±4 V).

[0022] FIG. 14 demonstrates programming of an ECRAM fabricated on a more hydrogen-rich H_xWO_3 channel with voltage pulses of identical width of 3 sec but different amplitude of ±4 V and ±5 V.

[0023] FIG. 15 shows average ΔG per gate programming pulse (triangles, left axis) and the total charge injected by the gate current (ΔQ, circles, right axis) as a function of the pulse amplitude, with the line representing the linear fitting to the data.

[0024] FIG. 16 shows retention of selected analog states (0.1 V read) under zero gate bias in ambient and the corresponding drift coefficient (ν), where the inset shows retention with the gate floating, and the dotted line represents the fitting to the power decay function of $G(t)=G_{t0}(t/t_0)^{-\nu}$.

[0025] FIG. 17 shows programming of a micron-scale (L_{ch}=10 μm, W=3 μm) ECRAM cell with 300 μsec gate-voltage pulses (±4 V).

[0026] FIG. 18 is a double logarithmic scale plot showing the average ΔG per weight-update step as a function of the pulse width with the same pulse amplitude of 4 V, where the dotted line represents the linear fitting to the data, and the error bars represent the standard deviation.

[0027] FIG. 19 shows measured transient variation of the sense current (solid line, right axis) during the ECRAM read operation performed with a voltage pulse (dotted line, left axis) applied on the drain electrode and the source/gate grounded, where the shading highlights the settling time t_{read}.

[0028] FIG. 20 shows the recovery waveform of the sense current of read post write pulses, where the light shading indicates the write pulses of 300 μs, and the darker shading highlights the settling time $t_{read\text{-}after\text{-}write}$ required to reach steady states.

[0029] FIG. 21 shows a line drawing based on a scanning-electron microscopy (SEM) image of an exemplary nanometer-scale (L_{ch}=150 nm, W=150 nm) ECRAM cell, where the scale bar is 500 nm.

[0030] FIG. 22 shows waveforms of the read voltage (dotted line, left axis) and the sense current (solid line, right axis) of the scaled ECRAM to extract its t_{read} as highlighted by the shading.

[0031] FIG. 23 shows time-resolved source-drain current of the scaled ECRAM in a read performed immediately after a write pulse (light shading), showing a faster settling time (darker shading) to steady state (dashed line).

[0032] FIG. 24 shows an endurance test for 10^8 write-read pulses, showing no device degradation with the intermediate switching cycles plotted after 10^5, 10^6, 10^7, 3×10^7, 5×10^7, 7.5×10^7, and 10^8 pulses.

[0033] FIG. 25A is a STEM image showing the gate stack of an exemplary ECRAM after operation with 100 million read-write cycles, where the scale bar is 100 nm.

[0034] FIG. 25B provides depth profiles from the FIG. 25A gate stack showing the atomic fractions of W, Hf, and O before and after a 10^8 cycle endurance test, as measured by energy dispersive X-ray spectroscopy, where the dashed line serves as a visual guide to mark the interface between WO_3 and HfO_2.

[0035] FIG. 26 shows programming of a ECRAM cell by voltage-pulse trains composed of 64 potentiation followed by 64 depression steps, with the selector transistor turned either on (solid line) or off (dashed, inset).

[0036] FIGS. 27A and 27B show current-voltage characteristics of a silicon MOSFET selector measured before (FIG. 27A) and after (FIG. 27B) fabricating the protonic ECRAM layer on top with 40 nm HfO_2 as the interlayer dielectric, where V_{DS}: source-drain bias and I_{DS}: source-drain current.

[0037] FIG. 28 shows a circuit diagram of an exemplary 3 by 3 ECRAM pseudo-crossbar array.

[0038] FIG. 29 shows parallel row-by-row programing of an ECRAM array, with the normalized conductance modulation ($\Delta G/G$) indicated by grayscale; the voltages applied on the word lines (WLs) and the source lines for the weight update (SLNs) are indicated.

[0039] FIGS. 30A and 30B show achievable accuracy and energy consumption respectively for an ECRAM-CMOS hybrid in-memory-computing accelerator to learn the classification of the MNIST dataset using different write-pulse widths, as benchmarked against the SRAM-based digital accelerator or software.

[0040] FIG. 31 shows simulated accuracy of ECRAM (triangles, operated with 10 µsec gate-voltage pulses) and SRAM-based (circles) accelerators to learn the classification of the CIFAR (Canadian Institute for Advanced Research)-10 dataset with the VGG (Visual Geometry Group)-8 network as a function of the training epochs performed, where results from learning in software (dotted line) are also displayed for comparison.

[0041] FIG. 32 shows an x-ray diffraction (XRD) spectrum of the WO_3 film deposited by sputtering a WO_3 target in an oxygen-rich environment with the wafer substrate temperature held at 300 °C during the process.

[0042] FIG. 33A shows a measured Rutherford backscattering (RBS) spectrum (solid line) and the simulation result (dashed line) of tungsten oxide deposited on silicon substrate, showing an O:W stoichiometry of 3:1.

[0043] FIG. 33B shows O:W stoichiometry determined by RBS for tungsten oxide films deposited by reactive sputtering using either an elemental W (light shading) or WO_3 (darker shading) target as a function of the O_2 to Ar gas-flow ratio.

[0044] FIGS. 34A and 34B show optical transmittance (FIG. 34A) and sheet resistance (FIG. 34B) of 160 nm H_xWO_3 films with different proton concentrations on glass substrates prepared via the H-spillover process, where aluminum films with different thicknesses (5 nm, 10 nm, 20 nm green) on top of stoichiometric WO_3 reacted with HCl-based aqueous buffer solution; depending on the aluminum-film thickness, a different amount of hydrogen molecules was produced for their *in-situ* intercalation into the WO_3 lattice.

[0045] FIG. 35 shows change of the two-terminal conductance of H_xWO_3 resistors (channel length L_{ch}=10 μm and width W=3 μm) as a function of annealing time in ambient at 150 °C, where the H_xWO_3 is formed from 80 nm WO_3 subjected to the hydrogen spillover from the redox reaction between HCl and a 10 nm aluminum film deposited on top.

[0046] FIG. 36 shows width normalized device resistance R as a function of L_{ch} for H_xWO_3 with Pt contact switched between high (180 $M\Omega \cdot sq^{-1}$) and low (150 $k\Omega \cdot sq^{-1}$) sheet resistance corresponding to the operating dynamic range of ECRAM, showing a degradation of the device on/off ratio with the scaling of the L_{ch}.

[0047] FIG. 37A shows a schematic of a protonic ECRAM employing an aqueous electrolyte (0.1 $mol \cdot L^{-1}$ H_2SO_4) and H_xWO_3 channel; in weight update, the source/drain electrodes are grounded, and the voltage pulse is supplied with a gold work electrode immersed in the electrolyte.

[0048] FIG. 37B shows a double logarithmic scale plot showing the average ΔG per weight-update step as a function of the pulse width for protonic ECRAMs employing the aqueous electrolyte (circles, applied gate-voltage amplitude is 2 V) in comparison with that of ECRAMs built on the hydrogenated ZrO_2 solid electrolyte (triangles, applied gate-voltage amplitude is 4 V), where the dashed lines are linear fittings to the data.

[0049] FIGS. 38A-38D illustrate a benchmarking of the performance of deep-learning accelerators based on the protonic ECRAM cells described in this disclosure with those built on other BEOL-compatible non-volatile memory technologies.

DETAILED DESCRIPTION

[0050] In-memory-computing architectures based on memristive crossbar-arrays can enhance the computing efficiency for deep-learning with massive parallelism. To fulfill their potential, the core memory devices are engineered as described in this disclosure with the objectives of providing high-speed and symmetric analog programming with small variability, compatibility with silicon technology, and size reduction into a nanometer-size footprint. In particular, electrochemical synaptic transistors built with CMOS-compatible metal oxides and operating by shuffling protons within a symmetric gate stack are provided to meet all these stringent requirements. Such electrochemical synaptic transistors can be monolithically integrated with silicon transistors to form pseudo-crossbar arrays where parallel, precise, and symmetric programming of the channel conductance can be executed with gate-voltage pulses. High-speed programming with frequency approaching megahertz, endurance above 100 million read-write pulses, and device critical dimensions down to 150×150 nm^2 or smaller, may be realized, as described below.

[0051] A schematic of an exemplary device is illustrated in FIG. 1, with its gate stack revealed in a cross-sectional STEM image (FIG. 2). The electrochemical random-access memory cell 100 includes source and drain electrodes 102,104 on a substrate 106, and a channel layer 108 comprising a protonic intercalatable material (that is, an inorganic material capable of intercalating, or reversibly inserting, protons within its structure) on the substrate 106 between the source and drain electrodes 102,104. The protonic intercalatable material may comprise a hydrogenated tungsten oxide, *e.g.*, $H_{x1}WO_3$, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$. A solid-state protonic electrolyte layer 110 is on the channel layer 108, and a gate electrode bilayer 112 comprising the protonic intercalatable material is on the solid-state protonic electrolyte layer 110. The protonic intercalatable material of the gate electrode bilayer 112 may comprise a hydrogenated tungsten oxide, *e.g.*, $H_{x2}WO_3$, where x2 represents initial hydrogen concentration of the gate electrode bilayer 112 and $0 < x2 \leq 0.4$. Upon application of a voltage pulse to the gate electrode bilayer 112, protons are reversibly inserted into the channel layer 108 from the solid-state protonic electrolyte layer 110 to modulate channel conductance. Notably, the reversible insertion of protons into the channel layer 108 may not involve a phase transformation. The gate electrode bilayer 112 may include a top layer 112a comprising a metal and a bottom layer 112b comprising the $H_{x2}WO_3$, where the bottom layer 112b is in contact with the solid-state protonic electrolyte layer 110. In some examples, $0 < x1 \leq 0.3$ and/or $0 < x2 \leq 0.3$. Also, x1 may be equal to or unequal to x2.

[0052] The solid-state protonic electrolyte layer 110 may comprise a hydrogenated metal oxide, such as zirconium oxide, yttria-stabilized zirconia, cesium oxide, titanium oxide, $La_2Ce_2O_7$, $Ce_{0.9}Gd_{0.1}O_2$, a perovskite metal oxide such as $BaZr_{1-x}In_xO_{3-\delta}$, and/or a Brownmillerite $A_2B_2O_5$-based oxide such as $Ba_2In_2O_5$. The channel layer 108 and the gate electrode bilayer 112 may include amorphous WO_3, which may be stoichiometric or nearly stoichiometric.

[0053] The channel layer 108 may have lateral dimensions in a range from 10 nm by 10 nm to about 1 μm by 1 μm. Typically, the channel layer has a thickness of at least about 5 nm and no more than about 100 nm. The solid-state protonic electrolyte layer 110 may have a thickness of at least about 2 nm and no more than about 30 nm.

[0054] In some examples, the memory cell 100 may include a passivation layer 114 on the gate electrode bilayer 112, where the passivation layer 114 comprises a dielectric material such as hafnium oxide or aluminum oxide. Also or alternatively, the substrate 106 may include a dielectric layer 116 thereon capable of blocking electrons and protons. The substrate 106 may comprise a silicon substrate with a thermal oxide thereon, and the dielectric layer 116 may be on the thermal oxide. The dielectric layer 116 may comprise hafnium oxide. Typically, the source and drain electrodes 102,104 each comprise a metal.

[0055] The electrochemical random-access memory cell 100 may have a base conductance G_0 determined by the initial hydrogen concentration x1 of the channel layer 108. The channel conductance may be modulated to store multiple bits.

[0056] Advantageously, the electrochemical random-access memory cell 100 may exhibit endurance above 100 million read-write operations. In other words, the memory cell 100 may exhibit no degradation in terms of programming symmetry, cycle-to-cycle variability, base conductance, and/or conductance modulation over 100 million read-write operations. Also or alternatively, read and write operations may be decoupled. The electrochemical random-access memory cell 100 may exhibit an operation time as low as 5 ns, symmetric programming characteristics, multi-level conductance with low cycle-to-cycle variability, and/or low energy consumption.

[0057] Referring to FIG. 3, the electrochemical random-access memory cell may be integrated with a silicon metal-oxide-semiconductor field-effect transistor (MOSFET). The electrochemical random-access memory cell may be positioned on the silicon MOSFET with a dielectric layer capable of blocking electrons and protons (*e.g.*, hafnium oxide) in between. An interconnect may electrically connect the gate electrode bilayer to a source electrode of the silicon MOSFET. The electrochemical random-access memory cell may be constructed entirely of inorganic materials. Referring to FIGS. 4A and 4B, a deep-learning accelerator may be constructed from an array of the electrochemical random-access memory cells, where each of the electrochemical random-access memory cells is integrated with a silicon MOSFET.

[0058] Fabrication of the ECRAM is discussed here and also in the Methods section below. Stoichiometric amorphous WO_3, which may be deposited by reactive sputtering from a WO_3 target, is chosen as the ECRAM channel as it gives both low base conductance and long proton retention (See Supplementary Note 1, below). Compared to VO_2, the intercalation of protons into WO_3 does not involve the phase transformation of the crystal structure, ensuring faster ECRAM operations with better device endurance. A hydrogen-spillover process is then utilized to introduce precisely controlled concentration of protons into the WO_3 on wafer scale (See Supplementary Note 2, below), followed by the deposition of an ultrathin film of ZrO_2 by atomic layer deposition (ALD) as the solid-state protonic gate electrolyte. The secondary-ion mass spectrometry (SIMS) depth profile indicates that a substantial amount of protons are diffused into the ZrO_2 from the H-doped WO_3 channel during ALD, which help to passivate the surface and grain boundaries of ZrO_2 with hydroxyl groups, forming pathways for fast protonic conduction. The gate includes a H_xWO_3

and metal bilayer, where the H-rich H_xWO_3 serves as the proton reservoir and helps to minimize the device built-in potential (See Supplementary Note 3).

[0059] Referring now to FIGS. 5A-5I, fabrication of an electrochemical random-access memory cell includes forming source and drain electrodes on a substrate. The substrate may comprise a silicon on insulator substrate. Also or alternatively, the substrate may comprise an array of silicon metal-oxide-semiconductor field-effect transistors (MOSFETs). In some examples, prior to forming the source and drain electrodes, as shown in FIG. 5B, a dielectric layer (*e.g.*, hafnium oxide or silicon dioxide) capable of blocking electrons and protons may be formed on the substrate, as indicated in FIG. 5A. The dielectric layer may be formed by thermal oxidation, physical vapor deposition, atomic layer deposition, and/or chemical vapor deposition.

[0060] A channel layer comprising tungsten oxide is formed on the substrate between the source and drain electrodes, as illustrated in FIGS. 5C-5E, and hydrogen is incorporated into the channel layer, as shown in FIGS. 5F-5G, using the "hydrogen spillover" process illustrated in FIG. 6. More specifically, a reactive metal film, *e.g.*, an aluminum film, may be deposited on the channel layer, and the substrate may be submerged in an aqueous solution comprising an acid, *e.g.*, HCl. As a consequence of the reaction (*e.g.*, $Al+H^+ \rightarrow Al^{3+}+H_2$, as discussed further below), the reactive metal film is etched and hydrogen is incorporated into the tungsten oxide of the channel layer as proton intercalants. In some examples, after submerging the substrate in the aqueous solution comprising the acid, the substrate may be annealed in air, *e.g.*, at a temperature in a range from about 110°C to 180°C, to adjust a concentration of the proton intercalants. As a consequence of hydrogen incorporation into the channel layer, $H_{x1}WO_3$ is formed, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$.

[0061] After incorporating hydrogen into the channel layer, a solid-state protonic electrolyte layer is formed on the channel layer, as shown in FIG. 5H. For example, zirconium oxide or another metal oxide may be deposited on the channel layer by atomic layer deposition using suitable precursors, such as a metalorganic compound (*e.g.*, tetrakis(dimethylamino)zirconium(IV) (TDMAZ)) and water vapor. The water vapor contributes protons to the solid-state protonic electrolyte layer such that a hydrogenated metal oxide is formed. Other suitable metal oxides beside zirconium oxide may include yttria-stabilized zirconia, cesium oxide, titanium oxide, $La_2Ce_2O_7$, $Ce_{0.9}Gd_{0.1}O_2$, perovskite metal oxides such as $BaZr_{1-x}In_xO_{3-\delta}$, and/or Brownmillerite $A_2B_2O_5$-based oxides such as $Ba_2In_2O_5$.

[0062] A first gate electrode layer including tungsten oxide is formed on the solid-state protonic electrolyte layer, and hydrogen is incorporated into the first gate electrode layer. The same process used to introduce hydrogen into the channel layer is employed to introduce hydrogen into the first gate electrode layer. That is, a reactive metal film (*e.g.*, an aluminum film) is deposited on the first gate electrode layer, and the substrate is submerged in an aqueous solution comprising an acid, *e.g.*, HCl, whereby the reactive metal film is etched and hydrogen is incorporated into the tungsten oxide of the first gate electrode layer as proton intercalants. After submerging the substrate in the aqueous solution comprising the acid, the substrate may be annealed in air to adjust the concentration of the proton intercalants. The annealing typically takes place at a temperature in a range from about 110°C to 180°C. As a consequence of hydrogen incorporation into the first gate electrode layer, $H_{x2}WO_3$ is formed, where x2 represents initial hydrogen concentration of the first gate electrode layer and $0 < x2 \leq 0.4$. A second gate electrode layer including a metal is formed on the first gate electrode layer to produce a gate electrode bilayer comprising the first and second gate electrode layers, as shown in FIG. 5I.

[0063] Fabrication of an exemplary silicon MOSFET and integration with an ECRAM is illustrated in FIGS. 7A-7L and discussed by way of an example in the Methods section. The process includes growing a layer of thermal oxide on a silicon-on-insulator (SOI) substrate (FIG. 7A-7B), photolithographically defining patterns of source and drain regions of the transistors into photoresist (FIG. 7C). Diffusion doping of the underlying silicon to form the heavily *n*-doped source/drain contact regions follows. The spin-on-dopant used for diffusion doping and the thermal oxide mask are then removed (FIG. 7D), and the device islands are patterned by photolithography (FIG. 7E). Source-drain contact electrodes are patterned (FIG. 7F), a gate oxide is deposited (FIG. 7G), and a metal gate electrode is patterned (FIG. 7H) to complete the silicon MOSFET fabrication. An insulator such as HfO_2 is deposited to serve as both the interlayer dielectric and the proton-diffusion barrier (FIG. 7I). The ECRAM layer is then fabricated on top using the process described above (FIG. 7J). The vias for interlayer interconnects are exposed by photolithography (FIG. 7K), and a final lithography and lift-off are performed to pattern the interlayer interconnects connecting the source electrodes of the silicon selector transistors to the gate electrodes of the ECRAMs (FIG. 7L).

[0064] A method of operating an ECRAM includes providing, as described above and illustrated in FIG. 1, an ECRAM comprising source and drain electrodes on a substrate, a channel layer comprising $H_{x1}WO_3$ on the substrate between the source and drain electrodes, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$, a solid-state protonic electrolyte layer on the channel layer, and a gate electrode on the solid-state protonic electrolyte layer. A bias (*e.g.*, a voltage pulse) is applied between the gate electrode and the channel layer, as illustrated, thereby inserting protons from the solid-state protonic electrolyte layer into the channel layer and modulating channel conductance. The method may further comprise applying a reverse bias (*i.e.*, a bias of opposite polarity) between the gate electrode and the channel layer, thereby extracting protons from the channel layer into the solid-state protonic electrolyte layer and further modulating the channel conductance. As described above, the gate electrode may comprise a gate electrode bilayer including $H_{x2}WO_3$ on the solid-state protonic electrolyte layer, where x2 represents initial hydrogen concentration of the gate electrode bilayer and $0 < x2 \leq 0.4$. The method may further comprise applying a bias between the source and drain electrode, as illustrated, to determine the channel conductance and conduct a read operation. The gate electrode may be grounded.

[0065] Referring to FIG. 8, the channel conductance G of the device can be programmed through 64 (6-bit) discrete states with a high level of symmetry, small cycle-to-cycle variability, and a large dynamic range above 20 under voltage pulses (amplitude V_G=±4V), in contrast to devices built with an elemental metal gate (FIG. 9A) which can operate symmetrically only under current pulses, as shown in FIGS. 9B and 9C. The zoom-in view of FIG. 8, which is shown in FIG. 10, illustrates three continuous conductance states out of the total 64 and confirms that the CMOS-compatible protonic ECRAM prototype has low read noise with conductance down to nano-siemens regime. The gate current during write is merely around a few pico-amperes, as shown in FIG. 11, corresponding to a low switching energy of about 10 pico-joule per step, which can be further reduced to below femto-joule with the adoption of up to 10^6 times faster write pulses down to microseconds, as can be seen in FIGS. 12 and 13. By adjusting the initial hydrogen concentration in the H_xWO_3 channel through varying parameters in the hydrogen spillover (See Supplementary Note 2), the device base conductance G_0 and the absolute conductance change ΔG per weight-update step can be effectively modulated. FIG. 14 illustrates the programming of a ECRAM built on a more hydrogen-rich H_xWO_3 channel, whose G_0 is increased from 1 to 200 nS with the average ΔG per step simultaneously increased by ~ 10 times, likely assisted by defect-defect interactions and the increase of proton concentration in the ZrO_2 electrolyte. This tunability may be critical for large-scale array implementations, where the memristive resistance can be optimized to minimize the thermal noise, the voltage drop in the interconnect metal lines, and the overhead on the peripheral circuits, and balance the trade-off between the absolute ΔG per step and the on/off ratio. The channel conductance modulation also depends exponentially on the amplitude of the gate-voltage pulses, as higher gate bias increases the gate-current density and thus the total amount of injected protons, as shown in FIG. 15. These ECRAM cells demonstrate good retention. The drift coefficient v, obtained by fitting the change of channel conductance as a function of time t to the power law of $G(t)=G_{t0}(t/t_0)^{-v}$ where G_{t0} is the initial conductance at time t_0 and v is the fitted drift coefficient, is ultra-low, as can be seen in FIG. 16, especially compared to PCMs. It shows that the slow self-discharging of the ECRAM under zero gate bias ensures that these discrete conductance levels are well preserved. States with large deviation from G_0 exhibits slightly larger drift coefficient, likely caused by the larger gradient for proton diffusion. Such low drift coefficient over at least 10^3 sec is more than enough to ensure that accuracy may not be affected when ECRAM-based accelerators, where weights stored on ECRAM cells are continuously updated during training, are used to train multilayer perceptrons and even convolutional neural networks with weight sharing in convolutional layers. The device retention is further improved if measured with zero gate current.

[0066] High-speed programming of protonic ECRAMs has been demonstrated with fast gate-voltage pulses. With G_0 around 1 μS, a large ΔG of 3 μS can be induced by 32 potentiation-depression pulses with width down to 300 μsec to afford an on/off ratio of 3 as can be seen in FIG. 17. FIG. 18 shows the reproducible cycling with 10 μsec and 5 μsec write pulses. Although the device dynamic range becomes smaller, the programming characteristics are still highly symmetric, and the ΔG after each pulse remains larger than 1 nS with low variability and read noise to enable the storage of 8-bit of information within a single ECRAM cell. The effect of the off-state current can be eliminated with the addition of a dummy column in the crossbar architecture. Compared to existing ECRAMs employing oxygen ions as the intercalant, when operated under the 10 μsec-wide write pulses to achieve the same conductance-modulation ratio, the protonic ECRAMs described herein require about 7 times lower number of weight-update steps under a ~40% lower gate-voltage amplitude, resulting from proton's higher ionic diffusivity. There is some upshift of the baseline G_0, indicating that more weight-update operations are required in the depression branch to bring the conductance back to its initial level. However, such asymmetry associated with the slightly different number of the intermediate states between the minimum and maximum conductance states will not degrade the training accuracy as predicted in simulation, and can be compensated by applying voltage pulses with slightly different amplitude or width for the potentiation versus depression operations. Since the conductance change correlates with the amount of charge, *i.e.*, the number of protons, injected or extracted, ΔG, as well as the corresponding energy consumption per weight-update step, scales linearly with the pulse width, as shown in FIG. 18.

[0067] In addition to the time required for the weight update, the ECRAM device speed is also limited by the read transients. For read without write, the ECRAM channel behaves as a resistor capacitively coupled to the device capacitance. The current flowing across the source-drain electrodes is monitored, showing that the settling time t_{read} after the application of the read-voltage pulse was less than 500 nsec as limited by the resistor-capacitor (RC) delay, as shown in FIG. 19. For read post update, a fast recovery current pulse is observed followed by a slower decay, eventually leading toward a stable sense current, as indicated in FIG. 20. This transient behavior is qualitatively similar to that of the ECRAMs employing lithium or oxygen ions as the intercalants, and it is limited by the dielectric relaxation processes and the charge transfer from the center of the channel to contacts on the edge. But the quantitative settling time $t_{\text{read-after-write}}$ for protonic ECRAMs is >100 times faster due to the high ionic diffusivity of protons.

[0068] Both t_{read} and $t_{\text{read-after-write}}$ are expected to diminish with scaling, which reduces both the gate capacitance and the required lateral diffusion distance of injected protons. In CMOS-compatible, all-inorganic protonic ECRAMs with their channel lateral dimensions scaled down to $150 \times 150\ \text{nm}^2$, as shown in FIG. 21, which represent the smallest lateral ECRAMs ever fabricated, their symmetric and linear programming characteristics are well preserved (FIG. 13), showing the excellent device scalability. The absolute conductance values and the dynamic range of modulation are slightly degraded likely due to the impact from the contact resistance, and the larger read noise could be caused by the miniaturization of the device volume together with the increase of the channel surface-to-volume ratio, which reduce the averaging effect and increase the impact from traps at material interfaces. Even though the gate capacitance is significantly reduced with $>10^3$ times smaller device channel area, the t_{read} of the scaled ECRAMs is only marginally improved to 370 nsec, as can be seen in FIG. 22, indicating that the device capacitance is mainly limited by parasitic from, for example, the overlap regions between the source/drain and gate electrodes with the ZrO_2 sandwiched in between. However, the rate of the current stabilization in the read post update indeed becomes much faster, with the apparent $t_{\text{read-after-write}}$ less than 1 µsec, as shown in FIG. 23, revealing the benefit of ECRAM channel-length scaling to mitigate the device read transients. To improve the device latency down to the desired 50–100 nanosecond regime, it is required to further suppress the device read transients and increase the ΔG per step in weight update, which can be accomplished by scaling down the lateral length of the H_xWO_3 channel, optimizing the device structure to minimize the parasitic capacitance, modifying the composition and nanostructures of the ZrO_2-based electrolyte to increase its ionic conductivity, and reducing the gate stack (H_xWO_3 channel and the electrolyte) thickness (Supplementary Note 4).

[0069] In addition to their fast operations and excellent scalability, the CMOS-compatible protonic ECRAMs exhibited excellent endurance when modulated in air with fast voltage pulses. No degradation in terms of programming symmetry, cycle-to-cycle variability, baseline channel conductance, or conductance modulation was observed after 10^8 write-read operations, which correspond to over one million switches over the full on/off range (see FIG. 24). The moderate dynamic range of ECRAM corresponds to limited modulation of the proton concentration in the H_xWO_3 channel, which ensures that the devices will not degrade due to lattice expansion or ion trapping during operation. There may be a slight oxygen loss accompanying the proton insertion and removal in the amorphous H_xWO_3 channel, but it is not significant even after million times of operations driven by the gate pulses, as measured within the accuracy of TEM-energy dispersive X-ray spectroscopy (EDS), as shown in FIG. 25. The variation of the maximum and minimum conductance after 32 potentiation and depression pulses is small with a relative standard deviation around 7%, which indicates a standard deviation normalized to the entire weight range around $7\%/\sqrt{32}=1\%$ for each weight-update cycle. With the 20 nm thick HfO_2 or Al_2O_3 passivation deposited by ALD at 120 °C, the device operation was not affected even when measured in high vacuum (10^{-6} Torr). It is a significant advantage over previous protonic ECRAMs employing polymer/nanoporous electrolytes and/or PdH_x gate, which are only functional in a humidity-controlled, utmost medium vacuum, environment or forming gas. The stable and symmetric programming characteristics were maintained even after annealing at 200 °C, and the device can be successfully operated at 80–100 °C without degrading the data retention or device endurance, which further verifies the reliability of the all-inorganic protonic ECRAM prototype in harsh environments. Similar as other memristive devices, the device conductance and conductance modulation systematically shift with the increase of temperature due to thermal excitation and activation, which needs to be compensated on the software level with the chip temperature continuously monitored by built-in temperature sensors. Finally, when the performance of the all-inorganic protonic ECRAM is benchmarked with other BEOL-compatible analog memory technologies, it shows clear advantages in symmetry, cycle-to-cycle variability, and energy consumption (See Supplementary Note 5).

[0070] All-inorganic protonic ECRAMs were monolithically integrated with silicon MOSFETs (See Methods and FIGS. 7A-7L for the process flow) in the form of 1-transistor-1-ECRAM cells (FIG. 3), where the silicon transistor acted as the selector to solve the write-disturbance problem; see FIG. 26, which shows negligible ECRAM conductance change with the selector transistor turned off. Note that ECRAM arrays can potentially operate without selectors using the half-selection bias scheme, and here the fabrication and characterization of 1-transistor-1-ECRAM cells is used mainly to verify the CMOS compatibility of all-inorganic protonic ECRAMs. With the ALD HfO_2 deposited on top of the completed silicon MOSFETs and the metal interconnects as both the interlayer dielectric and protonic diffusion barrier, the performance of these underlying silicon devices was not affected by the addition of the ECRAM layer on top, as can be seen in FIG. 27. Beyond individual cells, an integrated pseudo-crossbar array was constructed as shown in FIGS. 4A and 4B. In the pseudo-crossbar array fabric (FIG. 28), parallel row-by-row weight updates were successfully implemented, as in the programming patterns shown in FIG. 29. The average weight-update in the selected cells was 210±20%, while those in the unselected rows were not disturbed with an average ΔG of -1±4%.

[0071] After programming, the pseudo-crossbar arrays can be used to parallelly execute the vector-matrix multiplication, which is the core and the most expensive computing operation in many deep-learning and image-processing algorithms. The color transformation is used as an example, where a transformation matrix is used to generate modified red (R), green (G), and blue (B) pixel data for recoloring. Each element of the transformation matrix was first mapped to the ECRAM array by parallel programming. The color of each pixel represented by an 8-bit number for each R/G/B channel was then converted to a voltage input vector delivered to the bit lines of the fabric. The current measured at the source lines for the weight sum was the result of the dot product between the input voltage vector and the analog conductance matrix following the Kirchhoff's law, and it was finally encoded back to generate the modified RGB values. Despite the non-idealities associated with the device variability and conductance drift, the color transformation result was comparable to what performed by the software with a narrow error band.

[0072] With the functionality of ECRAM pseudo-crossbar arrays verified in experiment, the performance of accelerators built on ECRAM arrays integrated with silicon-CMOS peripheral circuits is simulated. Based on the experimentally measured dynamic range, non-linearity, symmetry, read-write voltage and speed, cycle-to-cycle weight-update variation, and device-to-device variations of CMOS-compatible, all-inorganic protonic ECRAMs (see Supplementary Note 6), the modeling, which considers all the device non-idealities, indicates that such in-memory-computing accelerators can achieve similar level of accuracy (86-93%, the small degradation in accuracy for devices operated with faster pulses is caused by their smaller on/off ratio) in learning the classification of the Modified National Institute of Standards and Technology (MNIST) dataset compared to the SRAM-based digital accelerators or the software (94%), as enabled by ECRAMs' good linearity and symmetry, low variability, and their large number of available states (FIG. 30A). Meanwhile, the overall energy consumption, including contributions from the ECRAM array, the selector transistors, and the periphery circuit, is up to two times lower than SRAM-based accelerators (FIG. 30B), and the chip-area cost, with the analog and high-resistance ECRAMs monolithically integrated on top of the silicon circuits, is reduced by over 10 times. The latency is 50 times longer due to the much slower operation of the current ECRAM prototypes compared to SRAM, but it is competitive against accelerators built on PCM and metal-oxide memristors (Supplementary Note 7). ECRAM accelerators can also be used to train the convolutional neural networks, achieving similar level of accuracy (~85%) as benchmarked with their SRAM-based counterparts or software (FIG. 31), but superior to those employing other emerging non-volatile memories.

[0073] To summarize, a CMOS-compatible, all-inorganic protonic ECRAM technology that shows promising highly symmetric switching characteristics with ultralow cycle-to-cycle variability, endurance above 100 million read-write pulses, and good retention with low drift is described in this disclosure. In one example, the approach employs hydrogenated ZrO_2 as protonic electrolyte to ensure CMOS-compatibility, takes advantage of the high diffusivity of protons to realize fast device operations approaching megahertz, and adopts the symmetric gate stack to enable symmetric programming based on identical voltage pulses. The smallest lateral ECRAMs with active device area down to 150×150 nm^2 and the first pseudo-crossbar arrays composed of ECRAM synaptic memory cells integrated on top of silicon selector transistors, capable of performing both parallel analog programming and vector-matrix multiplication, have been demonstrated in experiment, showing the scalability of the ECRAM prototype into both nanometer-size device footprint and circuit-level complexity. These results have established the technological promise of CMOS-compatible, all-inorganic protonic ECRAMs to realize energy- and cost-efficient in-memory-computing accelerators for deep learning, as benchmarked with existing and emerging memory technologies.

[0074] Methods

[0075] *Fabrication of CMOS-compatible all-inorganic protonic ECRAMs.* The process flow is schematically illustrated in FIGS. 5A-5I. 40 nm of HfO_2 was first deposited as a proton-diffusion barrier and etching-stop layer by ALD (Veeco Nanotech ALD System) on a silicon substrate covered with 90 nm thermal oxide, using Tetrakis(dimethylamido)-hafnium and water as precursors at 200 °C. A photolithography step was then performed to define the source-drain electrode patterns into the photoresist (AZ 5214, exposure dose was $200 mJ \cdot cm^{-2}$), and an electron-beam evaporation (Temescal) was used to deposit 2 nm Cr/ 40 nm Pt, followed by lift-off in acetone to form the source-drain contacts. 80 nm of WO_3 was grown by reactive radio-frequency magnetron sputtering (Kurt J. Lesker PVD 75) of WO_3 target (Kurt J. Lesker) by means of a plasma composed of argon as the carrier gas and oxygen as the reactive gas. After pumping down the sputtering-deposition chamber to high vacuum (5×10^{-7} Torr), pure oxygen and argon were introduced with their relative flow rates adjusted to the desired ratio with the help of the mass-flow controllers to a stable chamber pressure of 1.5 millitorr. During deposition, the substrate was held at 300 °C and the radio-frequency forward input power was maintained at 140 W. After WO_3 deposition to the desired thickness of 50–100 nm, another photolithography step was performed to define the device channel footprint into the photoresist. A subsequent SF_6 reactive-ion etching (RIE, Oxford Mixed ICP-RIE system) removed the WO_3 in the unprotected area to stop on the HfO_2 or Pt metal surfaces. The inductively coupled plasma (ICP) power was 1,000 W, the RIE power was 100 W, the flow rate of SF_6 was 20 s.c.c.m., and the chamber pressure was maintained at 5 millitorr. A thin, *i.e.*, 10 nm thick, aluminum film was then blanketly deposited covering the whole substrate by sputtering (AJA Orion-8 Magnetron Sputtering System). The deposited Al film was etched by soaking the substrate in $4 mol \cdot L^{-1}$ HCl aqueous solution, and part of the hydrogen generated was incorporated into the WO_3 channel as proton intercalants through spillover. After further adjusting the proton concentration in H_xWO_3 by annealing at 150 °C in air, the substrate was immediately transferred to an ALD chamber, where 15 nm of ZrO_2 was deposited using Tetrakis-(dimethylamino)-zirconium (Strem Chemicals) and water vapor as precursors at 120 °C. The low ALD deposition temperature and the adoption of water as precursor are critical to maintain the proton content within the ZrO_2 to enhance its ionic conductivity. Photolithography was used again to pattern the gate electrode, which has overlap with the source-drain contacts, into the photoresist. 20 nm of WO_3 was deposited again by the radio-frequency reactive sputtering with the substrate held at room temperature, and then converted to H_xWO_3 by means of the same hydrogen spillover process as described above. Electron-beam evaporation was used to deposit 2 nm Cr/40 nm Au followed by lift-off of the photoresist in acetone to complete the gate electrode stack composed of H_xWO_3/Cr/Au trilayer. Another 20 nm of HfO_2 or Al_2O_3 was deposited by ALD as a passivation layer to improve the device stability and endurance. A final lithography step was performed to expose the probing pads for the source-drain electrodes, where the ZrO_2 and HfO_2/Al_2O_3 dielectrics on top were etched by HF to complete the device fabrication flow.

[0076] *Fabrication of nanoscale ECRAMs.* To fabricate ECRAMs with their device channel length and channel width scaled down to sub-micron regime, electron-beam lithography (EBL), instead of photolithography, was used to define the device source-drain contacts with 150 nm channel length, the device isolation pattern for the metal hard mask with width down to 150 nm, and the composite H_xWO_3/metal gate, using 6% ethyl lactate (EL-6) and polymethyl methacrylate (PMMA 950k A3, Kayaku Advanced Materials) bilayer as the photoresist. The exposure dose was 750 $\mu C \cdot cm^{-2}$ and the developer was methyl isobutyl ketone (MIBK) 1:2 diluted with isopropyl alcohol (IPA). In the device isolation step, 100 nm Au hardmask defined by liftoff was used instead of the photoresist, which was stripped by soaking the substrate in commercial gold etchant (Transene TFA) post the SF_6 RIE.

[0077] *Monolithic integration of all-inorganic protonic ECRAMs with silicon MOSFETs.* The process, which is schematically illustrated in FIGS. 7A-7L, started from fabricating the silicon MOSFETs on a silicon-on-insulator (SOI) substrate (Soitec, 70 nm lightly *p*-doped device-layer silicon on 2 μm buried oxide). A layer of thermal oxide was first grown by dry oxidation at 1050 °C to a thickness of about 90 nm. Photolithography was performed to define patterns of the heavily doped source-drain regions of the transistors into the photoresist (AZ-5214), followed by removing the SiO_2 in the exposed area with buffered oxide etchant. After stripping the photoresist in acetone, a film of phosphorus-containing spin-on-dopant (Filmtronics P509) was blanketly deposited by spin casting at 3,000 rpm for 30 sec followed by a soft bake at 110 °C for 3 min. Annealing at 850 °C for 20 minutes in a three-zone tube furnace (Lindberg) with N_2 ($2\ L{\cdot}min^{-1}$) and O_2 ($1\ L{\cdot}min^{-1}$) flow caused the phosphorus to diffuse from the spin-on-dopant into the underlying silicon to form the heavily *n*-doped source/drain contact regions. After cooling down to room temperature, the wafer was immersed in HF to remove both the spin-on-dopant and the thermal oxide mask, followed by piranha cleaning to remove the residual phosphorus oxide. The device islands were then patterned by photolithography, and the silicon in the exposed area was removed by timed ICP-RIE (20 mtorr, 20 s.c.c.m. CF_4 flow, 300 W ICP power, and 100 W RIE power for 90 sec) stopping on the buried oxide of the SOI wafer. After removing the photoresist by acetone and cleaning the surface by piranha solution, the source-drain contact electrodes were patterned by the third photolithography step, followed by the electron-beam evaporation of 2 nm Cr/40 nm Au and lift-off in acetone. 40 nm of HfO_2 was then deposited as the high-*κ* gate dielectric by ALD, using Tetrakis(dimethylamido)-hafnium and water as precursors at 200 °C. Another photolithography and the lift-off scheme were subsequently performed again to define the metal gate electrodes composed of 2 nm Cr and 40 nm Au to complete the silicon MOSFET fabrication. Afterwards, another 40 nm HfO_2 was deposited by ALD at 200 °C, serving as both the interlayer dielectric and the proton-diffusion barrier. The ECRAM layer was then fabricated on top using the process described above. The vias for interlayer interconnects were exposed by photolithography, with ICP-RIE utilized to etch through the ZrO_2 gate electrolyte of the ECRAM, the HfO_2 interlayer dielectric, and the HfO_2 gate oxide of the silicon MOSFET to stop on the metal contact (5 mtorr, 10 s.c.c.m. CHF_3 flow and 5 s.c.c.m. Ar flow, 100 W ICP power, and 40 W RIE power).

A final lithography and lift-off were performed to pattern the interlayer interconnects connecting the source electrodes of the silicon selector transistors to the gate electrodes of the ECRAMs.

[0078] *Instrumentation.* The cross-sectional high-angle annular dark field (HAADF) STEM images and the associated elemental configurations were obtained using the FEI Talos F200X G2 STEM equipped with four-crystal EDS system (FEI Super-X). All STEM-EDS data were collected for more than 30 min with a 10 μsec dwell time and ~200 pA probe current, at an accelerating voltage of 200 kV. The sample was prepared using the Thermo Scios2 dual-beam focused-ion beam to a thickness around 50 nm under 5 kV. The SIMS elemental depth profile was obtained using the Phi TRIFT III time-of-flight SIMS system with the material removal performed using low-energy Cs^+ ion source. SEM micrographs were acquired using a Hitachi S4800 microscope. The device electrical characterizations were performed either in ambient or under high vacuum at the desired temperature using a manual probe station (LakeShore Cryotronics CRX–6.5K) connected with a semiconductor parameter analyzer (Keysight B1500A) equipped with integrated high-resolution source-measurement units (Keysight B1517A) and waveform generator/fast-measurement unit (Keysight B1530A). The AZ 5214E photoresist was patterned with a Heidelberg MLA150 aligner, and the EL6/PMMA photoresist was patterned with a Elionix ELS-G150 150 keV EBL system. The *Neurosim* simulations were performed on the Illinois Campus Cluster HAL. The X-ray diffraction (XRD) spectrum was recorded using the Bruker D8 Advanced XRD system. The composition of deposited tungsten oxides was determined by Rutherford backscattering spectrometry (RBS) using the NEC Pelletron accelerator equipped with RBS chamber. Optical transmittances of WO_3 and H_xWO_3 were recorded using an Agilent Cary 5000 ultraviolet-visible absorbance spectrophotometer. XPS spectra of WO_3 and H_xWO_3 were acquired on a PHI Versa Probe III instrument with a monochromatic Al Kα (1486.6 eV) source. The pass energy was 55 eV and spectra were referenced to C1*s* peak (adventitious carbon) at 284.8 eV.

[0079] Supplementary Note 1: Optimization of the WO_3 channel for protonic ECRAMs

[0080] Tungsten oxide, which is a prototypical proton intercalation host, is chosen as the channel of an exemplary CMOS-compatible, all-inorganic protonic ECRAM prototype. The deposited tungsten oxide is amorphous, as evident from the absence of any obvious diffraction peak in its XRD spectrum, , as can be seen in FIG. 32. Compared to the crystalline counterpart, amorphous tungsten oxide exhibits faster response time and larger modulation of its optical and electrical properties upon proton intercalation since its open microstructures facilitate the easy insertion and extraction of ions. However, for amorphous oxides, the stoichiometry is variable depending on the film deposition conditions, and the physical properties of tungsten oxide are intimately related with its oxygen content. To optimize the device performance and uniformity, reactive sputtering deposition of tungsten oxide using either W or WO_3 as the target, under different Ar to O_2 gas-flow ratio, is evaluated to find out the process-composition-property relationship. The compositions of the deposited films were determined by RBS as shown in FIG. 33A. For tungsten oxide $WO_{3+\delta}$ deposited using W target, the O/W ratio increases with increasing oxygen concentration in the sputtering gas mixture (FIG. 33B), and the oxygen content can be tuned over a wide range from under-stoichiometric ($\delta<0$) to over-stoichiometric ($\delta>0$). The oxygen deficient tungsten oxide correspondingly exhibits an increase in its electrical conductivity and optical absorption with the increasing oxygen deficiency, because the oxygen vacancies in tungsten oxide act as effective shallow doners to increase the carrier concentration. It may lead to high base conductance of the ECRAM channel, therefore limiting the on/off ratio of the channel conductance modulation upon subsequent proton intercalation. For super-stoichiometric films, these excess oxygen ions are incorporated as interstitials assisted by their high kinetic energy during the sputtering process. Upon proton intercalation, the injected H^+ cations may bind with these excess oxygen ions first to form neutral compound H_2O, causing irreversible changes of the film properties in the initial few modulation cycles, which degrades the device endurance and long-term reliability. Moreover, the formed H_2O can serve as trapping centers for the H^+ intercalation and diffusion ($H_2O+H^+\rightarrow H_3O^+$), lowering the device operation speed. While for the tungsten oxide deposited using the WO_3 target, the nearly stoichiometric WO_3 films were always obtained within the range of O_2/Ar ratio that underwent testing, which not only ensures optimal device performance but also improves the device uniformity and reproducibility. Therefore, the WO_3 target is preferred. Also evaluated were WO_3 films deposited with different substrate temperatures (room temperature *versus* 300 °C). Although the optical and electrical properties of both films can be sufficiently modulated with the proton intercalation, performed using an electrochemical cell composed of 1 $mol \cdot L^{-1}$ H_2SO_4 aqueous electrolyte, a gold counter electrode, and the WO_3 film deposited on top of an indium-tin-oxide (ITO) coated glass substrate acting as the working electrode, the WO_3 deposited at 300 °C demonstrated a much better capability to retain its properties in ambient air after the proton intercalation. The conductance measured across the substrate dropped by less than 30% after exposing the sample to ambient for 10^4 sec, in contrary to the strong self-bleaching observed for those deposited at room temperature. This is likely caused by their difference in film morphology and density. For both films, negligible change of their optical or electrical properties after the intercalation was observed if they were stored in a nitrogen-filled dry box, under vacuum, or in air with an ALD ZrO_2 or HfO_2 passivation layer deposited on top.

[0081] Supplementary Note 2: Wafer-scale hydrogen spillover to control the H-concentration in hydrogen tungsten bronze

[0082] A modified hydrogen spillover process is developed to introduce protons into isolated WO_3 islands on wafer scale, which allows for control of the base conductance of the ECRAM channel and adoption of the conductive H_xWO_3 as part of the gate electrode for ECRAM arrays. The process is schematically illustrated in FIG. 6. A thin layer of reactive metal, *e.g.*, aluminum, is deposited on top of the WO_3 islands. It then reacts with an acidic buffer solution, producing hydrogen molecules through the reaction of $Al+H^+ \rightarrow Al^{3+}+H_2$. Some of these *in-situ* generated hydrogen molecules are chemisorbed on the WO_3 surface and then dissociate into protons and electrons. For a hydrogen atom adsorbed on the oxide surface, it has a repulsive interaction with an oxygen anion, but when it splits into protons and electrons, an attractive interaction develops. As a result, the activation energy of the splitting is low, and the overall energy of the system is reduced, as predicted in DFT simulation.

The electrons reduce the W^{6+} cations to W^{5+} and W^{4+}, and the protons move at the same time to adjacently attached O^{2-} anions and become interstitial intercalants. As assisted by the hydrogen bonding between protons and oxygen ions, the protons can diffuse easily on the surface and into the bulk of WO_3 with a low activation energy, driven by the concentration gradient, to convert WO_3 into H_xWO_3 with uniform degree of intercalation. The amount of intercalated H^+ directly correlates with the mobile carrier concentration determining the tungsten bronze's optical transmittance and electrical resistivity, and therefore needs to be accurately controlled to ensure the ECRAM can operate with optimum baseline conductance G_0 and range of modulation. In the modified hydrogen-spillover process, the degree of proton intercalation can be quantitatively controlled by adjusting the thickness of the Aluminum film deposited on top and thus the amount of the hydrogen molecules produced (see FIGS. 34A and 34B). In addition, the final proton stoichiometry can be further fine-tuned by annealing the film in air at 150 °C, as shown in FIG. 35. Here, the inserted protons on the surface react with the oxygen in air as $4H^+ + 4e^- + O_2 \rightarrow 2H_2O$ to form water molecules, which are removed from the oxide surface with thermal desorption. The created concentration gradient will drive additional protons to diffuse from the bulk to the surface, which is a slower process compared to the surface chemical reaction, to sustain the depletion of H^+ from the tungsten bronze thin film. This process allows for precise adjustment of the base conductance of ECRAMs from nano-siemens to micro-siemens (comparing FIGS. 8, 14 and 17), with good uniformity on wafer scale.

[0083] To verify that the hydrogen spillover indeed introduces protons as intercalants into the WO_3 channel, the XPS spectra of the WO_3 films are measured and compared before and after the process. Before hydrogen spillover, the photoelectron peaks for W $4f_{5/2}$ and W $4f_{7/2}$ are located at 37.9 eV and 35.8 eV, respectively, which correlate with W^{6+}. After hydrogen spillover, components with lower binding energies are identified, which are consistent with the W cations reduced by the electrons from the dissociated H_2 molecules to lower oxidation states of 5+ and 4+. In addition to the W 4*f* spectra, the O 1*s* spectra provides additional evidence. Before hydrogen spillover, the O 1*s* spectrum has a single peak at 530.9 eV, correlating with the oxygen anions in W-O-W bridges of the stoichiometric WO_3.

After spillover, the O 1*s* photoelectron peak becomes much broader, and its deconvolution reveals extra peaks that could be assigned to hydroxyl (-OH) groups incorporated within the lattice at 532.8 eV and oxygens (531.7 eV) for which anion vacancies can be found in their second neighboring due to the breakage of the W-O-W bridges. However, no peaks associated with the oxygen in water molecules, whose binding energy is in the range of 533.3-534.4 eV, can be identified. These results confirm that the hydrogen element exists in the form of protons as lattice intercalant, rather than free water molecules, which is critical for the ECRAM to operate in harsh conditions such as high vacuum and high temperature.

[0084] Supplementary Note 3: Operating principles of the protonic ECRAMs, and the symmetric gate stack to minimize the open-circuit potential for programming with voltage pulses

[0085] The all-inorganic protonic ECRAM may include a H_xWO_3 channel between the source-drain contacts, a zirconia-based solid-state protonic electrolyte, and an electrically conductive gate electrode. In the write operation, a bias applied between the gate electrode and the channel drives protons dispersed in the hydrogenated ZrO_2 electrolyte to first drift/diffuse toward the channel–electrolyte interface, and then insert into the WO_3 lattice through the Faradic reaction $WO_3 + xH^+ + xe^- \leftrightarrow H_xWO_3$. These injected protons partially fill the conduction band of WO_3 to increase its carrier concentration and thus the electrical conductivity. This process can be reversed by the application of the gate bias in the opposite polarity to extract the protons from the channel back into the electrolyte. In the read process, a small bias is applied between the source-drain electrodes to determine the channel conductance with the gate grounded to minimize the ionic current flowing through the gate stack. Compared to two-terminal memristors, by decoupling the read and write operations, ECRAM exhibits much more symmetric programming characteristics, multi-level conductance with low cycle-to-cycle variability, and low energy consumption due to the utilization of the low-energy-barrier ion-insertion process. Since the proton injection and extraction happen over the entire WO_3 interface with the gate electrolyte, the modulation scales with the device dimensions, as can be seen in FIG. 36.

[0086] A significant problem for ECRAMs adopting the pure metal electrode as the gate (FIG. 9A) is the non-zero built-in open-circuit potential of their gate stack. The built-in potential results from the difference between the electrochemical potentials of the channel and the gate, as dictated by the Nernst equation. Therefore, ECRAMs with pure metal gate can only demonstrate the symmetric switching characteristics when programmed by sending current pulses to their gate electrode, as shown in FIG. 9B. The programming inevitably becomes highly asymmetric for voltage-controlled operations, where the depression is much more abrupt compared to the potentiation (FIG. 9C), simplify because the effective bias across the gate stack deviates from the applied gate voltage as distorted by the built-in potential. The implementation of the current-controlled operations on an array scale is undesirable as it requires much more complicated peripheral silicon-CMOS circuits, leading to both increased chip-area cost and higher energy consumption. This limitation is overcome by adding another layer of H_xWO_3 as part of the ECRAM gate electrode. With the symmetric gate stack, the open-circuit potential across the ECRAM gate stack was kept around zero over the capacity range, and thus highly symmetric programming was achieved using voltage-based programming, as shown in FIG. 8. The degradation to the weight-update symmetry caused by the variation of the proton concentration in the channel and the H_xWO_3 gate during ECRAM device operation is very limited as observed in experiment. Moreover, the symmetric gate stack also helps to minimize the self-discharging of the ECRAM device when the gate electrode is grounded during the read operations, leading to a sufficiently long data retention with ultralow drift coefficient to perform accurate trainings of deep-neural networks, as indicated in FIG. 16.

[0087] Supplementary Note 4: Limiting factors of the switching speed of ECRAM

[0088] The transport of protons in the gate electrolyte is currently limiting the switching speed of the CMOS-compatible, all-inorganic ECRAM prototype. By replacing the hydrogenated ZrO_2 with aqueous electrolyte which has much higher ionic conductivity, as shown in FIG. 37A, over 100 times higher conductance change in the H_xWO_3 channel starting from the identical baseline conductance can be induced by applying programming voltage pulses with the same width but lower amplitude of ± 2V, and ΔG per step > 1nS can be achieved under gate-voltage pulse width down to 1 nsec, as can be seen in FIG. FIG. 37B. These results, combining with the recent development of superionic protonic electrolytes based on zirconium hydrogenphosphate, polyoxometalates, metal-organic-framework, and graphene oxide, whose protonic conductivities approach that of liquid water, suggest the promise of achieving the 50–100 nsec switching speed in all-inorganic solid-state protonic ECRAMs. Reduction of the gate electrolyte and the H_xWO_3 channel thickness will also help to improve the device speed by reducing the distance for protons to diffuse across the ECRAM gate stack in each weight-update operation.

[0089] Supplementary Note 5: Performance benchmark of CMOS-compatible, all-inorganic protonic ECRAMs with other BEOL-compatible non-volatile analog memory technologies

[0090] The performance of the CMOS-compatible, all-inorganic protonic ECRAM is benchmarked with other competing BEOL-compatible analog non-volatile memory technologies. The device key performance metrics and characteristic, including the number of conductance states, non-linearity, on/off ratio, weight-update pulse amplitude and width, cycle-to-cycle variation, and endurance are summarized in Supplementary Table 1. Only identical pulses were considered for weight update, as the non-identical pulse programming scheme, which is commonly used to improve the symmetry for PCM and FeFET, is impractical for chip implementation.

[0091] The CMOS-compatible, all-inorganic protonic ECRAM exhibits best symmetry in linearity (Supplementary Fig. 4), which is quantitatively defined as the difference between the non-linearities (NLs) corresponding to the potentiation and depression branches of the device programming characteristics (See Supplementary Note 6 regarding how the NL parameters were extracted) and smallest cycle-to-cycle variability, which collectively enable the accommodation of a large number of conductance states. These attributes are most critical to ensure a high online-training accuracy when they are implemented as the synaptic cores in the deep-learning accelerators. The ECRAM has a lower on/off ratio. However, the effect of the off-state current can be eliminated with the addition of a dummy column in the crossbar architecture. Here, the memory devices in the dummy column remain in the minimum conductance state so that the readout from the dummy column can be used by the peripheral circuit to subtract the off-state weight sum from the output of all other columns. With total chip-area overhead less than 9%, the impact from the low on/off ratio can be minimized, as verified in the simulation with the device-to-device spatial variation in consideration. The programming pulse width for ECRAM is competitive against CBRAM, RRAMs built on AlO_x/HfO_2 and $Pr_{0.7}Ca_{0.3}MnO_3$, but inferior to state-of-the-art RRAMs built on TaO_x/HfO_2, PCM, and FeFET. The adoption of protonic electrolytes with higher ionic conductivity and the further scaling down of the gate stack thickness and channel lateral dimensions can help to reduce this performance gap (See Supplementary Note 4 for more detailed discussions). The device endurance is competitive against other memory technologies. Although the programming voltage for ECRAM is higher, which requires the inclusion of voltage multipliers on chip, the energy consumption for each weight update step is much lower compared to its competitors due to the low ionic gate current. Specifically, FIG. 15 indicates that ~90 picocoulomb of protonic charge is injected into the H_xWO_3 channel under 4 V of gate voltage after 3 sec. Since the injected protonic charge and thus the channel conductance change is proportional with the pulse width (FIG. 18), the charge movement incurred in one weigh-update transaction becomes about 10^4 to 10^6 times smaller with the gate pulse of 300 μsec and 5 μsec, respectively (FIG. 17 and FIG. 12), which gives an energy consumption of about 0.6 to 36 femto-joule, comparable to the energy consumption per synaptic event in human brain. And it can be further improved through adopting electrolyte with higher ionic conductivity, which allows us to induce the same ΔG with smaller voltage amplitude, whose feasibility has been demonstrated in Supplementary Note 4.

[0092] Supplementary Note 6: Deep-learning accelerator simulation

[0093] The NeuroSim 3.0 package is used to emulate the performance of deep-learning accelerators built on the pseudo-crossbar arrays of CMOS-compatible, all-inorganic protonic ECRAMs monolithically integrated with silicon-CMOS peripheral circuits. Their capability of performing the supervised trainings to classify the MNIST dataset of handwritten digits with a simple 2-layer multilayer perceptron (MLP) neural network was benchmarked with the digital synaptic array architecture based on SRAM and analog synaptic arrays built on other competing non-volatile memory technologies, using the *MLP+Neurosim* framework. Their performance in training an

8-layer VGG network for image recognition using the CIFAR-10 dataset was also emulated with the *DNN+Neurosim* framework. The parallel readout and row-by-row weight update as demonstrated in FIG. 29 were utilized. Experimentally measured ECRAM programming characteristics (FIG. 17 and 12), which correlate the conductance (G) change with the number of gate-voltage pulses applied (P), are used to directly extract the simulation parameters including the total number of the available conductance states (P_{max}), the maximum (G_{max}) and minimum (G_{min}) conductance, as well as the amplitude and the width of the weight-update voltage pulses applied. The read pulse width was set to 400 nsec (instead of the 5 nsec pulse width by default in Neurosim) with amplitude of 0.1 V, which accommodates the transient characteristics shown in FIGS. 22 and 23 to ensure that a stable sense current can be eventually obtained. Note that the reduction of the pule width significantly reduces the G_{max}. The nonlinearity parameters (NLs) were extracted by fitting the experimental ECRAM programming data based on the following two equations:

[0094] Weight increase or LTP:

[0095]
$$G(LTP) = B\left(1 - e^{-\frac{P}{A}}\right) + G_{\min}$$

[0096] Weight decrease or LTD:

[0097]
$$G(LTD) = -B\left(1 - e^{-\frac{P-P_{\max}}{A}}\right) + G_{\max}$$

[0098] where the parameter A correlates with the curve nonlinearity (NL) and B is defined as

[0099]
$$B = \frac{G_{\max} - G_{\min}}{1 - e^{-\frac{P_{\max}}{A}}}.$$

[00100] The cycle-to-cycle weight-update variation is a critical device non-ideality and determined as the standard deviation of the difference between the measured and the fitted conductance normalized to the entire conductance modulation range, as 0.9%, 0.7%, and 0.5% for the ECRAMs operating with 300 μsec, 10 μsec, and 5 μsec write pulses, respectively. These numbers are quantitatively consistent with what was determined from the cumulative change of the minimum and maximum conductance values in the ECRAM endurance test (Fig. 2j). Based on these experimental results, 1% standard deviation of cycle-to-cycle variability is used in the simulation.

[00101] The ECRAM device-to-device variations are also explicitly considered in the simulations. An 8×8 ECRAM array was fabricated in experiment and characterized to extract the parameters. The completed ECRAMs exhibited good spatial uniformity with device yield of 92%. More specifically, 59 out of the 64 devices were functional, and the 5 devices failed because of patterning defects or large gate leakage current likely caused by the pin holes in their gate electrolyte resulting from particle defects on the wafer surface during the ALD deposition. All the functional devices demonstrated comparable NLs in weight update as being mapped in Extended Data Fig. 10b, with the standard deviation of the device-to-device NL variation of 10%. Their minimum and maximum conductance after 32 potentiation or depression cycles showed good spatial uniformity, with the relative standard deviations of 13% and 12%, respectively. In contrary to cycle-to-cycle variations, such device-to-device conductance variations will not significantly affect the capability of the ECRAM-based synaptic cores to perform the accurate training of deep neural networks as long as it is less than 30%, as suggested in both simulation and experiment. These device-to-device variation parameters are comparable to those extracted from the 3×3 1-transistor-1-ECRAM pseudo crossbar array, and they are used in the simulations. The uniform baseline conductance G_0 indicates that the proton injection by the hydrogen spillover process is homogeneous on wafer scale, leading to similar degree of proton intercalation into all WO_3 islands.

[00102] Since ECRAMs can be fabricated on top of the silicon transistors in the form of monolithic integration, the chip area of the synaptic cores built with ECRAM pseudo crossbar arrays is determined by the size of the selector transistor arrays plus the peripheral circuits including the decoders, multiplexers, adders, registers, and 8-bit analog-to-digital conversion, which are all based on the 32 nm-node silicon-CMOS technology with device characteristics extracted from foundry’s process design kit. The simulation parameters are summarized in Supplementary Table 2 of the priority application. The classification accuracy, overall energy consumption, and latency of the accelerators were determined after performing 125 training epochs.

The accuracy of the training performed in software was obtained by setting parameters *useHardwareInTrainingFF*, *useHardwareInTrainingWU*, and *useHardwareInTestingFF* all as *false*.

[00103] Supplementary Note 7: Performance benchmark of deep-learning accelerators with their synaptic cores built on ECRAMs and competing analog memory technologies

[00104] The performance of deep-learning accelerators with their synaptic cores built on ECRAM arrays is compared with those employing competing analog memory technologies, in terms of training accuracy, latency, energy consumption, and chip-area cost, based on the Neurosim 3.0 simulation results. With ECRAM's best symmetry, smallest cycle-to-cycle variability, and large number of available conductance states, ECRAM synaptic cores enable the best training accuracy, as can be seen in FIG. 38A. When operated at a programming speed of 5 μsec, the latency for the ECRAM-based accelerators to complete the training is only inferior to those built on the state-of-the-art TaO_x/TiO_2 memristors and GST PCMs that can be switched with weight-update pulses down to 50 nsec, as shown in FIG. 38B. However, because of ECRAM's low channel conductance and low energy consumption in weight updates, the overall energy consumed by the ECRAM accelerators is expected to become only slightly (~30%) larger than their TaO_x/TiO_2-based counterparts, but six times lower than the PCM-based accelerators whose operations involve the energy intensive melting and crystallization processes, as can be seen in FIG. 38C. The low energy consumption of the PCMO-based accelerators is an artifact as they basically could not learn anything with a low accuracy of 33%. Both the latency and the energy consumption are expected to be improved significantly with the further enhancement of ECRAM speed down to 50–100 nsec (See Supplementary Note 4). Although two-terminal memristors could have a smaller footprint due to their simpler device structures, accelerators based on ECRAMs actually have the smallest chip-area cost, as indicated in FIG. 38D, because the low channel conductance and small programming energy consumption of ECRAMs allow silicon transistors with small width to length ratio being used as selectors and in peripheral circuits. The adoption of vertical channels in ECRAM cells could further reduce the chip-area cost and increase the ECRAM-array integration density.

[00105] In summary, an all-solid-state inorganic ECRAM prototype operating based on the reversible insertion of protons into a H_xWO_3 channel from hydrogenated ZrO_2 electrolyte and H_xWO_3 gate has been described. Such devices can simultaneously meet requirements on the programming symmetry and variability, endurance, CMOS-compatibility, device scalability, and reliability for the ideal memristive devices for constructing deep-learning accelerators. They may exhibit highly symmetric and reproducible, *e.g.*, low cycle-to-cycle and device-to-device spatiotemporal variability, programming under gate-voltage pulses with endurance above 100 million read-write operations and energy consumption below femto-joule per transaction. Both WO_3 and ZrO_2 are compatible with the silicon-CMOS technology and associated wafer-scale microfabrication techniques, allowing the scaling of ECRAM dimensions down to 150×150 nm^2, or smaller. With amorphous HfO_2 as both the interlayer dielectric and the proton-diffusion barrier in some examples, these ECRAMs can be fabricated above silicon circuits without affecting the performance of the underlying logic transistors. ECRAM pseudo-crossbar arrays addressed with silicon selector transistors were then realized for the first time to demonstrate successfully parallel operations and monolithic integration. The small radius of the proton intercalant enables fast ionic dynamics for high-speed write-read programming with frequency approaching one megahertz. These CMOS-compatible all-inorganic protonic ECRAMs are functional in harsh conditions such as high vacuum and elevated temperature, and their non-volatile conductance exhibits low drift and long retention. The channel conductance can be tuned over a wide range from nano to micro-siemens, enabling the construction of large-size arrays with optimal power and performance. These results have established this ECRAM prototype as a prominent candidate for the core memory element in the next-generation analog deep-learning accelerators, with their current level of performance and non-idealities already sufficient to support training accuracy comparable to that of digital accelerators based on static random-access memory (SRAM), but under drastically reduced chip-area cost and energy consumption.

[00106] The subject-matter of the disclosure may also relate, among others, to the following aspects:

[00107] A first aspect relates to an electrochemical random-access memory cell comprising: source and drain electrodes on a substrate; a channel layer comprising an inorganic protonic intercalatable material on the substrate between the source and drain electrodes, a solid-state protonic electrolyte layer on the channel layer; and a gate electrode bilayer comprising of the inorganic protonic intercalatable material on the solid-state protonic electrolyte layer, wherein, upon application of a voltage pulse to the gate electrode bilayer, protons are reversibly inserted into the channel layer from the solid-state protonic electrolyte layer to modulate channel conductance.

[00108] A second aspect relates to the electrochemical random-access memory cell of the first aspect, wherein the solid-state protonic electrolyte layer comprises a hydrogenated metal oxide.

[00109] A third aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the hydrogenated metal oxide includes zirconium oxide, yttria-stabilized zirconia, cesium oxide, titanium oxide, $La_2Ce_2O_7$, $Ce_{0.9}Gd_{0.1}O_2$, perovskite metal oxides such as $BaZr_{1-x}In_xO_{3-\delta}$, and/or Brownmillerite $A_2B_2O_5$-based oxides such as $Ba_2In_2O_5$.

[00110] A fourth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the inorganic protonic intercalatable material comprises a hydrogenated tungsten oxide, *e.g.*, the channel layer may comprise $H_{x1}WO_3$, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$, and the gate electrode bilayer may comprise $H_{x2}WO_3$, where x2 represents initial hydrogen concentration of the gate electrode bilayer and $0 < x1 \leq 0.4$.

[00111] A fifth aspect relates to the electrochemical random-access memory cell of the fourth aspect, wherein the hydrogenated tungsten oxide includes amorphous WO_3, the amorphous WO_3 being stoichiometric or nearly stoichiometric.

[00112] A sixth aspect relates to the electrochemical random-access memory cell of the fourth or fifth aspect, wherein $0 < x1 \leq 0.3$ and/or $0 < x2 \leq 0.3$, and/or wherein $x1 \neq x2$.

[00113] A seventh aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the gate electrode bilayer comprises a top layer comprising a metal, and a bottom layer comprising the inorganic protonic intercalatable material, the bottom layer being in contact with the solid-state protonic electrolyte layer.

[00114] An eighth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the reversible insertion of protons into the channel layer does not involve a phase transformation.

[00115] A ninth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the channel layer has lateral dimensions in a range from 10 nm by 10 nm to about 1 µm by 1 µm.

[00116] A tenth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the channel layer has a thickness of at least about 5 nm and no more than about 100 nm.

[00117] An eleventh aspect relates to the electrochemical random-access memory cell of any preceding claim, wherein the solid-state protonic electrolyte layer has a thickness of at least about 2 nm and no more than about 30 nm.

[00118] A twelfth aspect relates to the electrochemical random-access memory cell of any preceding aspect, further comprising a passivation layer on the gate electrode bilayer, the passivation layer comprising a dielectric material.

[00119] A thirteenth aspect relates to the electrochemical random-access memory cell of the twelfth aspect, wherein the dielectric material comprises hafnium oxide or aluminum oxide.

[00120] A fourteenth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the substrate includes a dielectric layer thereon capable of blocking electrons and protons.

[00121] A fifteenth aspect relates to the electrochemical random-access memory cell of the fourteenth aspect, wherein the dielectric layer comprises hafnium oxide.

[00122] A sixteenth aspect relates to the electrochemical random-access memory cell of the fourteenth or fifteenth aspect, wherein the substrate comprises a

silicon substrate with a thermal oxide thereon, and wherein the dielectric layer is on the thermal oxide.

[00123] A seventeenth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the source and drain electrodes comprise a metal.

[00124] An eighteenth aspect relates to the electrochemical random-access memory cell of any preceding aspect having a base conductance G_0 determined by an initial hydrogen concentration x1 of the channel layer.

[00125] A nineteenth aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the channel conductance is modulated to store multiple bits.

[00126] A twentieth aspect relates to the electrochemical random-access memory cell of the nineteenth aspect exhibiting endurance above 100 million read-write operations.

[00127] A twenty-first aspect relates to the electrochemical random-access memory cell of the twentieth aspect, wherein exhibiting endurance above 100 million read-write operations comprises exhibiting no degradation in terms of programming symmetry, cycle-to-cycle variability, base conductance, and/or conductance modulation over 100 million read-write operations.

[00128] A twenty-second aspect relates to the electrochemical random-access memory cell of any preceding aspect exhibiting an operation time as low as 5 ns.

[00129] A twenty-third aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein read and write operations are decoupled.

[00130] A twenty-fourth aspect relates to the electrochemical random-access memory cell of any preceding aspect comprising symmetric programming characteristics, multi-level conductance with low cycle-to-cycle variability, and low energy consumption.

[00131] A twenty-fifth aspect relates to the electrochemical random-access memory cell of any preceding aspect being integrated with a silicon MOSFET.

[00132] A twenty-sixth aspect relates to the electrochemical random-access memory cell of any preceding aspect being positioned on the silicon MOSFET with a dielectric layer in between, the dielectric layer being capable of blocking electrons and protons, and wherein an interconnect electrically connects the gate electrode bilayer to a source electrode of the silicon MOSFET.

[00133] A twenty-seventh aspect relates to the electrochemical random-access memory cell of any preceding aspect, wherein the dielectric layer comprises hafnium oxide.

[00134] A twenty-eighth aspect relates to the electrochemical random-access memory cell of any preceding aspect being constructed entirely of inorganic materials.

[00135] A twenty-ninth aspect relates to a deep-learning accelerator comprising: an array including a plurality of the electrochemical random-access memory cells any preceding claim, each of the electrochemical random-access memory cells being integrated with a silicon MOSFET.

[00136] A thirtieth aspect relates to a method of making an electrochemical random-access memory cell, the method comprising: forming source and drain electrodes on a substrate; forming a channel layer comprising tungsten oxide on the substrate between the source and drain electrodes; incorporating hydrogen into the channel layer; after incorporating hydrogen into the channel layer, forming a solid-state protonic electrolyte layer on the channel layer; forming a first gate electrode layer including tungsten oxide on the solid-state protonic electrolyte layer; incorporating hydrogen into the first gate electrode layer; and after incorporating hydrogen into the first gate electrode layer, forming a second gate electrode layer including a metal on the first gate electrode layer to produce a gate electrode bilayer comprising the first and second gate electrode layers.

[00137] A thirty-first aspect relates to the method of the preceding aspect, wherein the substrate comprises an array of silicon MOSFETs.

[00138] A thirty-second aspect relates to the method of the thirtieth aspect, wherein the substrate comprises a silicon-on-insulator substrate.

[00139] A thirty-third aspect relates to the method of any preceding aspect, further comprising, prior to forming the source and drain electrodes, forming a dielectric layer capable of blocking electrons and protons on the substrate.

[00140] A thirty-fourth aspect relates to the method of the preceding aspect, wherein the dielectric layer is formed by thermal oxidation, physical vapor deposition, atomic layer deposition, and/or chemical vapor deposition.

[00141] A thirty-fifth aspect relates to the method of any preceding aspect, wherein the dielectric layer comprises hafnium oxide or silicon dioxide.

[00142] A thirty-sixth aspect relates to the method of any preceding aspect, wherein incorporating hydrogen into the channel layer comprises: depositing a reactive metal film, *e.g.*, an aluminum film, on the channel layer; and submerging the substrate in an aqueous solution comprising an acid, *e.g.*, HCl, whereby the reactive metal film is etched and hydrogen is incorporated into the tungsten oxide of the channel layer as proton intercalants.

[00143] A thirty-seventh aspect relates to the method of the preceding aspect, further comprising, after submerging the substrate in the aqueous solution comprising the acid, annealing the substrate in air to adjust a concentration of the proton intercalants.

[00144] A thirty-eighth aspect relates to the method of any preceding aspect, wherein the annealing occurs at a temperature in a range from about 110°C to 180°C.

[00145] A thirty-ninth aspect relates to the method of any preceding aspect, wherein $H_{x1}WO_3$ is formed, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$.

[00146] A fortieth aspect relates to the method of any preceding aspect, wherein incorporating hydrogen into the first gate electrode layer comprises: depositing a reactive metal film (*e.g.*, an aluminum film) on the first gate electrode layer; and submerging the substrate in an aqueous solution comprising an acid, *e.g.*, HCl, whereby the reactive metal film is etched and hydrogen is incorporated into the tungsten oxide of the first gate electrode layer as proton intercalants.

[00147] A forty-first aspect relates to the method of the preceding aspect, further comprising, after submerging the substrate in the aqueous solution comprising the acid, annealing the substrate in air to adjust a concentration of the proton intercalants.

[00148] A forty-second aspect relates to the method of the preceding aspect, wherein the annealing occurs at a temperature in a range from about 110°C to 180°C.

[00149] A forty-third aspect relates to the method of any of the fortieth to the forty-second aspects, wherein $H_{x2}WO_3$ is formed, where x2 represents initial hydrogen concentration of the first gate electrode layer and $0 < x2 \leq 0.4$.

[00150] A forty-fourth aspect relates to the method of any preceding aspect, forming the solid-state protonic electrolyte layer on the channel layer comprises depositing zirconium oxide, the hydrogenated metal oxide includes zirconium oxide, yttria-stabilized zirconia, cesium oxide, titanium oxide, $La_2Ce_2O_7$, $Ce_{0.9}Gd_{0.1}O_2$, perovskite metal oxides such as $BaZr_{1-x}In_xO_{3-\delta}$, and/or Brownmillerite $A_2B_2O_5$-based oxides such as $Ba_2In_2O_5$.

[00151] A forty-fifth aspect relates to the method of the preceding aspect, wherein the depositing includes atomic layer deposition using precursors including a metalorganic compound and water vapor, the water vapor contributing protons to the solid-state protonic electrolyte layer.

[00152] A forty-sixth aspect relates to the method of the preceding aspect, wherein the metalorganic compound includes tetrakis(dimethylamino)zirconium(IV) (TDMAZ).

[00153] A forty-seventh aspect relates to method of operating an electrochemical random-access memory cell (ECRAM), the method comprising: providing an ECRAM comprising: source and drain electrodes on a substrate; a channel layer comprising an inorganic protonic intercalatable material on the substrate between the source and drain electrodes, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$; a solid-state protonic electrolyte layer on the channel layer; a gate electrode on the solid-state protonic electrolyte layer; applying a bias between the gate electrode and the channel layer, thereby inserting protons from the solid-state protonic electrolyte layer into the channel layer and modulating channel conductance.

[00154] A forty-eighth aspect relates to the method of the preceding aspect, wherein the inorganic protonic intercalatable material of the channel layer comprises

$H_{x1}WO_3$, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$

[00155] A forty-ninth aspect relates to the method of any preceding aspect, wherein applying the bias comprises applying a voltage pulse to the gate electrode.

[00156] A fiftieth aspect relates to the method of any preceding aspect, further comprising applying a reverse bias (*i.e.*, bias of opposite polarity) between the gate electrode and the channel layer, thereby extracting protons from the channel layer into the solid-state protonic electrolyte layer and further modulating the channel conductance.

[00157] A fifty-first aspect relates to the method of any preceding aspect, wherein the gate electrode comprises a gate electrode bilayer including the inorganic protonic intercalatable material.

[00158] A fifty-second aspect relates to the method of the preceding aspect, wherein the inorganic protonic intercalatable material of the gate electrode bilayer comprises $H_{x2}WO_3$, where x2 represents initial hydrogen concentration of the gate electrode bilayer and $0 < x2 \leq 0.4$.

[00159] A fifty-third aspect relates to the method of any preceding aspect, further comprising applying a bias between the source and drain electrode to determine the channel conductance and conduct a read operation.

[00160] A fifty-fourth aspect relates to the method of the preceding aspect, wherein the gate electrode is grounded.

[00161] Although the present invention has been described in considerable detail with reference to certain embodiments thereof, other embodiments are possible without departing from the present invention. The spirit and scope of the appended claims should not be limited, therefore, to the description of the preferred embodiments contained herein. All embodiments that come within the meaning of the claims, either literally or by equivalence, are intended to be embraced therein.

[00162] Furthermore, the advantages described above are not necessarily the only advantages of the invention, and it is not necessarily expected that all of the described advantages will be achieved with every embodiment of the invention.

CLAIMS

1. An electrochemical random-access memory cell comprising:

source and drain electrodes on a substrate;

a channel layer comprising a protonic intercalatable material on the substrate between the source and drain electrodes;

a solid-state protonic electrolyte layer on the channel layer; and

a gate electrode bilayer comprising the protonic intercalatable material,

wherein, upon application of a voltage pulse to the gate electrode bilayer, protons are reversibly inserted into the channel layer from the solid-state protonic electrolyte layer to modulate channel conductance.

2. The electrochemical random-access memory cell of claim 1, wherein the solid-state protonic electrolyte layer comprises a hydrogenated metal oxide.

3. The electrochemical random-access memory cell of claim 1, wherein the hydrogenated metal oxide includes zirconium oxide, yttria-stabilized zirconia, cesium oxide, titanium oxide, $La_2Ce_2O_7$, $Ce_{0.9}Gd_{0.1}O_2$, perovskite metal oxides such as $BaZr_{1-x}In_xO_{3-\delta}$, and/or Brownmillerite $A_2B_2O_5$-based oxides such as $Ba_2In_2O_5$.

4. The electrochemical random-access memory cell of claim 1, wherein the protonic intercalatable material of the channel layer and the gate electrode bilayer comprises a hydrogenated tungsten oxide.

5. The electrochemical random-access memory cell of claim 4, wherein the hydrogenated tungsten oxide includes amorphous WO_3, the amorphous WO_3 being stoichiometric or nearly stoichiometric.

6. The electrochemical random-access memory cell of claim 4, wherein the protonic intercalatable material of the channel layer comprises $H_{x1}WO_3$, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$, and

wherein the protonic intercalatable material of the gate electrode bilayer comprises $H_{x2}WO_3$, where x2 represents initial hydrogen concentration of the gate electrode bilayer and $0 < x2 \leq 0.4$.

7. The electrochemical random-access memory cell of claim 6, wherein $0 < x1 \leq 0.3$ and/or $0 < x2 \leq 0.3$, and/or

wherein $x1 \neq x2$.

8. The electrochemical random-access memory cell of claim 1, wherein the gate electrode bilayer comprises a top layer comprising a metal, and a bottom layer comprising the protonic intercalatable material, the bottom layer being in contact with the solid-state protonic electrolyte layer.

9. The electrochemical random-access memory cell of claim 1, wherein the reversible insertion of protons into the channel layer does not involve a phase transformation.

10. The electrochemical random-access memory cell of claim 1, wherein the channel layer has lateral dimensions in a range from 10 nm by 10 nm to about 1 μm by 1 μm.

11. The electrochemical random-access memory cell of claim 1, wherein the channel layer has a thickness of at least about 5 nm and no more than about 100 nm.

12. The electrochemical random-access memory cell of claim 1, wherein the solid-state protonic electrolyte layer has a thickness of at least about 2 nm and no more than about 30 nm.

13. The electrochemical random-access memory cell of claim 1, further comprising a passivation layer on the gate electrode bilayer, the passivation layer comprising a dielectric material.

14. The electrochemical random-access memory cell of claim 13, wherein the dielectric material comprises hafnium oxide or aluminum oxide.

15. The electrochemical random-access memory cell of claim 1, wherein the substrate includes a dielectric layer thereon capable of blocking electrons and protons.

16. The electrochemical random-access memory cell of claim 15, wherein the dielectric layer comprises hafnium oxide.

17. The electrochemical random-access memory cell of claim 15, wherein the substrate comprises a silicon substrate with a thermal oxide thereon, and

wherein the dielectric layer is on the thermal oxide.

18. The electrochemical random-access memory cell of claim 1, wherein the source and drain electrodes comprise a metal.

19. The electrochemical random-access memory cell of claim 1 having a base conductance G_0 determined by an initial hydrogen concentration x1 of the channel layer.

20. The electrochemical random-access memory cell of claim 1, wherein the channel conductance is modulated to store multiple bits.

21. The electrochemical random-access memory cell of claim 20 exhibiting endurance above 100 million read-write operations.

22. The electrochemical random-access memory cell of claim 21, wherein exhibiting endurance above 100 million read-write operations comprises exhibiting no degradation in terms of programming symmetry, cycle-to-cycle variability, base conductance, and/or conductance modulation over 100 million read-write operations.

23. The electrochemical random-access memory cell of claim 1, exhibiting an operation time as low as 5 ns.

24. The electrochemical random-access memory cell of claim 1, wherein read and write operations are decoupled.

25. The electrochemical random-access memory cell of claim 1 comprising symmetric programming characteristics, multi-level conductance with low cycle-to-cycle variability, and low energy consumption.

26. The electrochemical random-access memory cell of claim 1 being integrated with a silicon MOSFET.

27. The electrochemical random-access memory cell of claim 26 being positioned on the silicon MOSFET with a dielectric layer in between, the dielectric layer being capable of blocking electrons and protons, and

wherein an interconnect electrically connects the gate electrode bilayer to a source electrode of the silicon MOSFET.

28. The electrochemical random-access memory cell of claim 27, wherein the dielectric layer comprises hafnium oxide.

29. The electrochemical random-access memory cell of claim 1 being constructed entirely of inorganic materials.

30. A deep-learning accelerator comprising:

an array including a plurality of the electrochemical random-access memory cells of claim 1, each of the electrochemical random-access memory cells being integrated with a silicon MOSFET.

31. A method of making an electrochemical random-access memory cell, the method comprising:

forming source and drain electrodes on a substrate;

forming a channel layer comprising tungsten oxide on the substrate between the source and drain electrodes;

incorporating hydrogen into the channel layer;

after incorporating hydrogen into the channel layer, forming a solid-state protonic electrolyte layer on the channel layer;

forming a first gate electrode layer including tungsten oxide on the solid-state protonic electrolyte layer;

incorporating hydrogen into the first gate electrode layer; and after incorporating hydrogen into the first gate electrode layer, forming a second gate electrode layer including a metal on the first gate electrode layer to produce a gate electrode bilayer comprising the first and second gate electrode layers.

32. The method of claim 31, wherein the substrate comprises an array of silicon MOSFETs.

33. The method of claim 31, wherein the substrate comprises a silicon on insulator substrate.

34. The method of claim 31, further comprising, prior to forming the source and drain electrodes, forming a dielectric layer capable of blocking electrons and protons on the substrate.

35. The method of claim 34, wherein the dielectric layer is formed by thermal oxidation, physical vapor deposition, atomic layer deposition, and/or chemical vapor deposition.

36. The method of claim 34, wherein the dielectric layer comprises hafnium oxide or silicon dioxide.

37. The method of claim 31, wherein incorporating hydrogen into the channel layer comprises:

depositing a reactive metal film on the channel layer; and

submerging the substrate in an aqueous solution comprising an acid, whereby the reactive metal film is etched and hydrogen is incorporated into the tungsten oxide of the channel layer as proton intercalants.

38. The method of claim 37, further comprising, after submerging the substrate in the aqueous solution comprising the acid, annealing the substrate in air to adjust a concentration of the proton intercalants.

39. The method of claim 38, wherein the annealing occurs at a temperature in a range from about 110°C to 180°C.

40. The method of claim 37, wherein $H_{x1}WO_3$ is formed, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$.

41. The method of claim 31, wherein incorporating hydrogen into the first gate electrode layer comprises:

depositing a reactive metal film on the first gate electrode layer; and

submerging the substrate in an aqueous solution comprising an acid, whereby the reactive metal film is etched and hydrogen is incorporated into the tungsten oxide of the first gate electrode layer as proton intercalants.

42. The method of claim 41, further comprising, after submerging the substrate in the aqueous solution comprising the acid, annealing the substrate in air to adjust a concentration of the proton intercalants.

43. The method of claim 42, wherein the annealing occurs at a temperature in a range from about 110°C to 180°C.

44. The method of claim 41, wherein $H_{x2}WO_3$ is formed, where x2 represents initial hydrogen concentration of the first gate electrode layer and $0 < x2 \leq 0.4$.

45. The method of claim 31, wherein forming the solid-state protonic electrolyte layer on the channel layer comprises depositing zirconium oxide, the hydrogenated metal oxide includes zirconium oxide, yttria-stabilized zirconia, cesium oxide, titanium oxide, $La_2Ce_2O_7$, $Ce_{0.9}Gd_{0.1}O_2$, perovskite metal oxides such as $BaZr_{1-x}In_xO_{3-\delta}$, and/or Brownmillerite $A_2B_2O_5$-based oxides such as $Ba_2In_2O_5$.

46. The method of claim 45, wherein the depositing includes atomic layer deposition using precursors including a metalorganic compound and water vapor, the water vapor contributing protons to the solid-state protonic electrolyte layer.

47. The method of claim 46, wherein the metalorganic compound includes tetrakis(dimethylamino)zirconium(IV) (TDMAZ).

48. A method of operating an electrochemical random-access memory cell (ECRAM), the method comprising:

providing an ECRAM comprising:

source and drain electrodes on a substrate;

a channel layer comprising a protonic intercalatable material on the substrate between the source and drain electrodes;

a solid-state protonic electrolyte layer on the channel layer;

a gate electrode on the solid-state protonic electrolyte layer;

applying a bias between the gate electrode and the channel layer, thereby inserting protons from the solid-state protonic electrolyte layer into the channel layer and modulating channel conductance.

49. The method of claim 48, wherein the protonic intercalatable material of the channel layer comprises $H_{x1}WO_3$, where x1 represents initial hydrogen concentration of the channel layer and $0 < x1 \leq 0.4$.

50. The method of claim 48, wherein applying the bias comprises applying a voltage pulse to the gate electrode.

51. The method of claim 48, further comprising applying a reverse bias between the gate electrode and the channel layer, thereby extracting protons from the channel layer into the solid-state protonic electrolyte layer and further modulating the channel conductance.

52. The method of claim 48, wherein the gate electrode comprises a gate electrode bilayer including the protonic intercalatable material.

53. The method of claim 52, wherein the protonic intercalatable material of the gate electrode bilayer comprises $H_{x2}WO_3$, where x2 represents initial hydrogen concentration of the gate electrode bilayer and $0 < x2 \leq 0.4$.

54. The method of claim 48, further comprising applying a bias between the source and drain electrode to determine the channel conductance and conduct a read operation.

55. The method of claim 54, wherein the gate electrode is grounded.

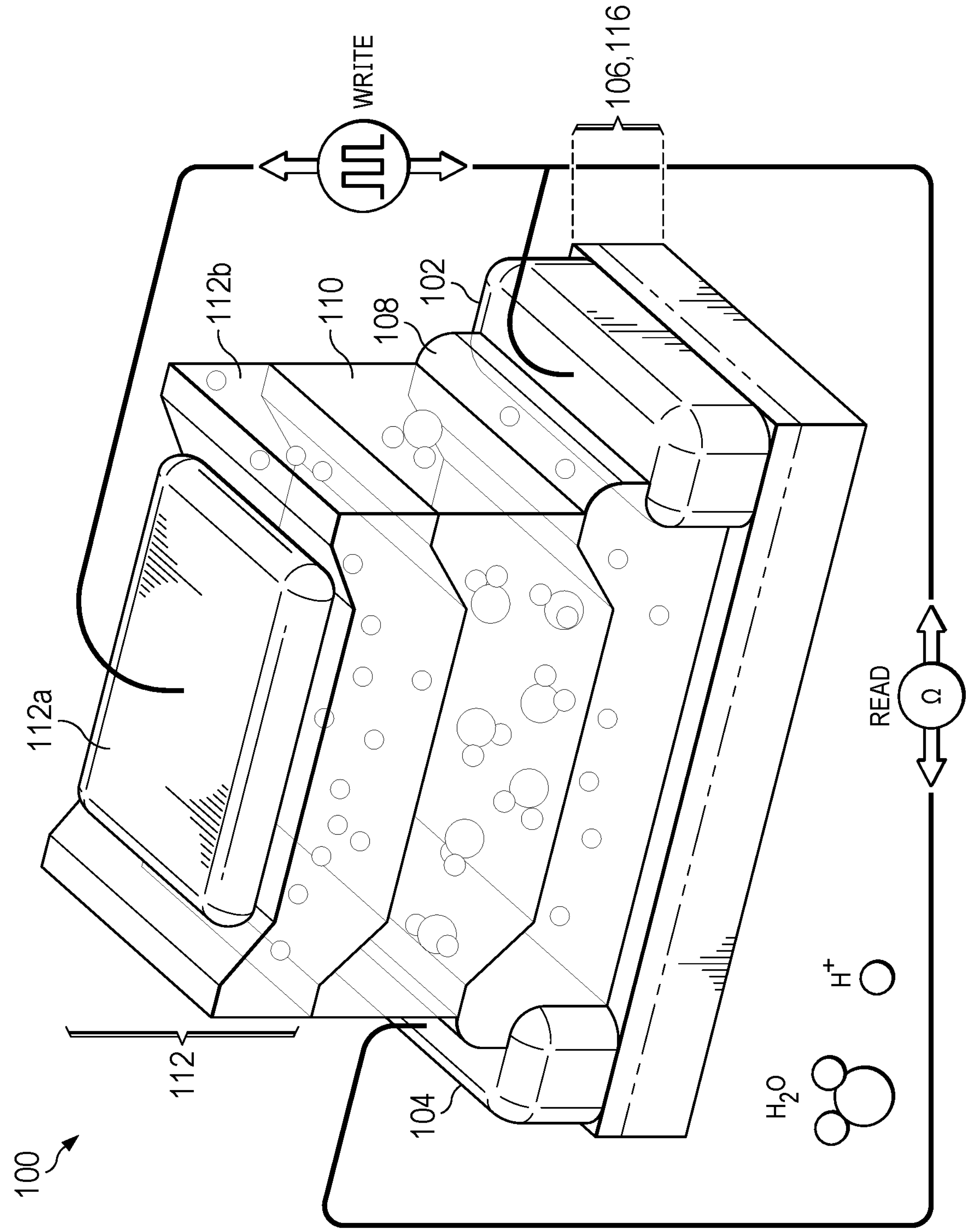


FIG. 1

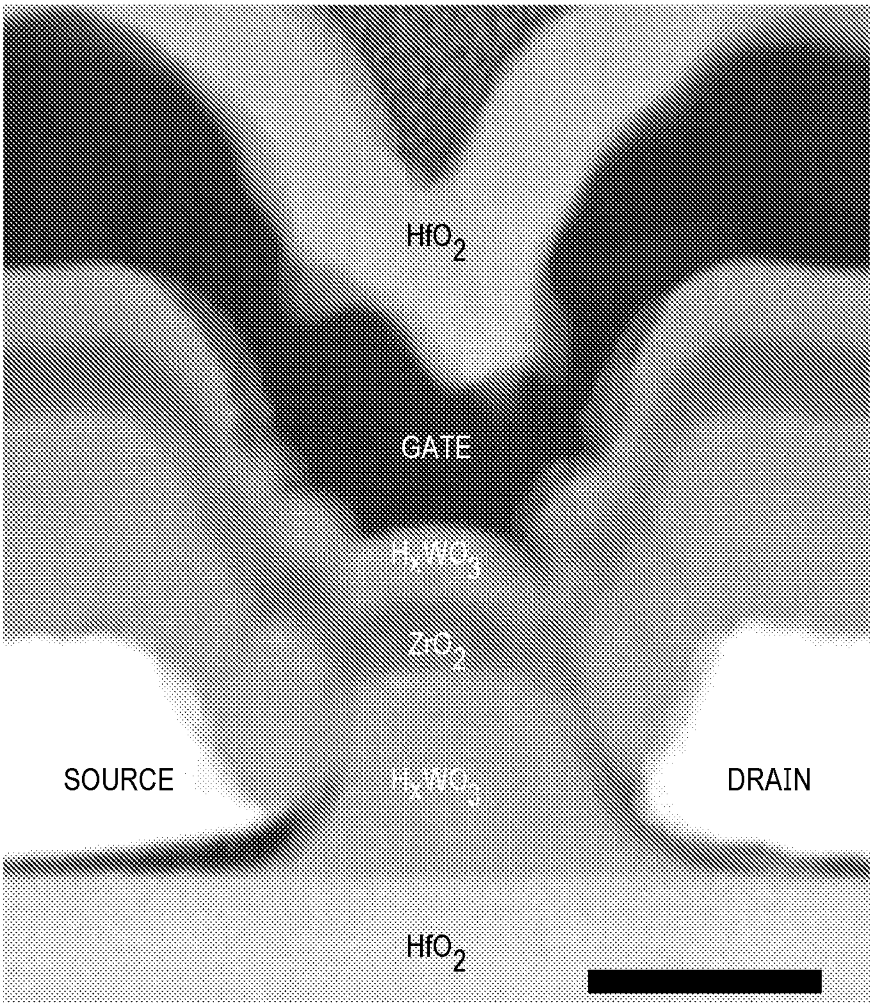


FIG. 2

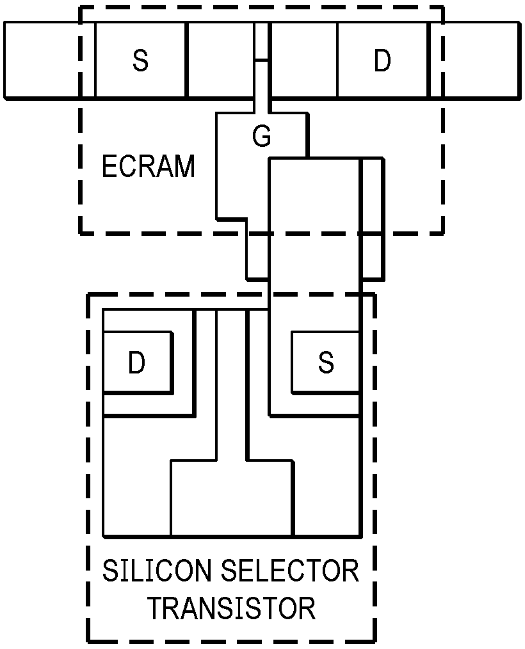

S
D
G
ECRAM
D
S
SILICON SELECTOR TRANSISTOR

FIG. 3

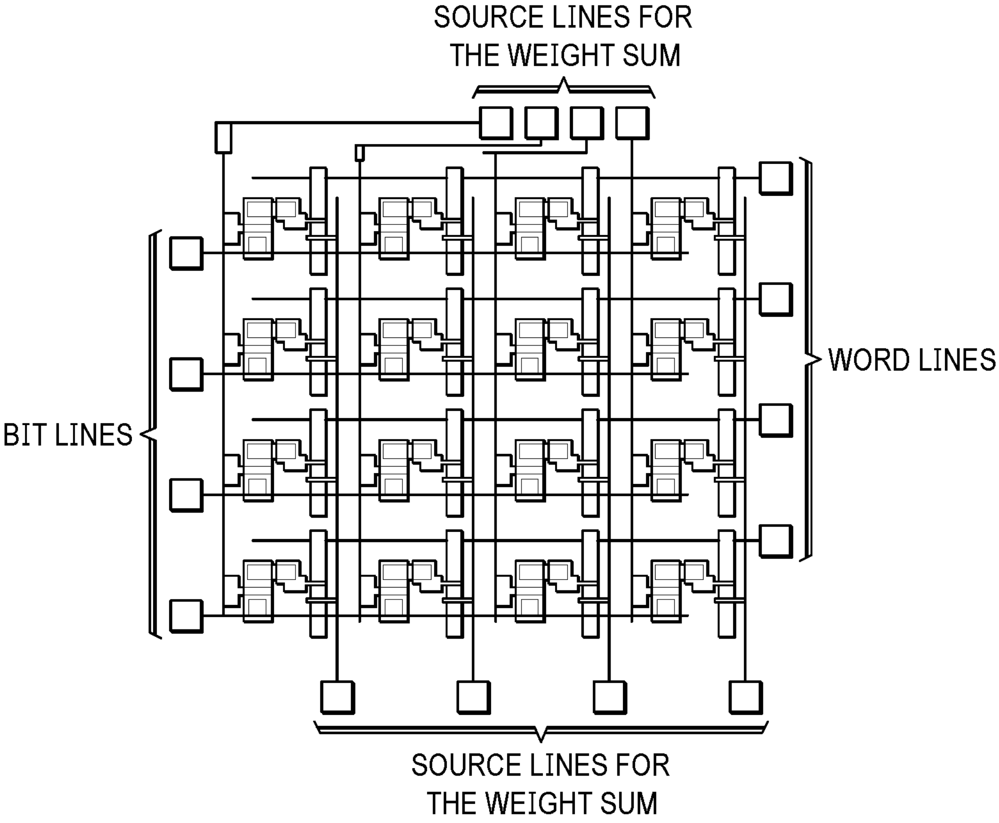

SOURCE LINES FOR THE WEIGHT SUM
WORD LINES
BIT LINES
SOURCE LINES FOR THE WEIGHT SUM

FIG. 4A

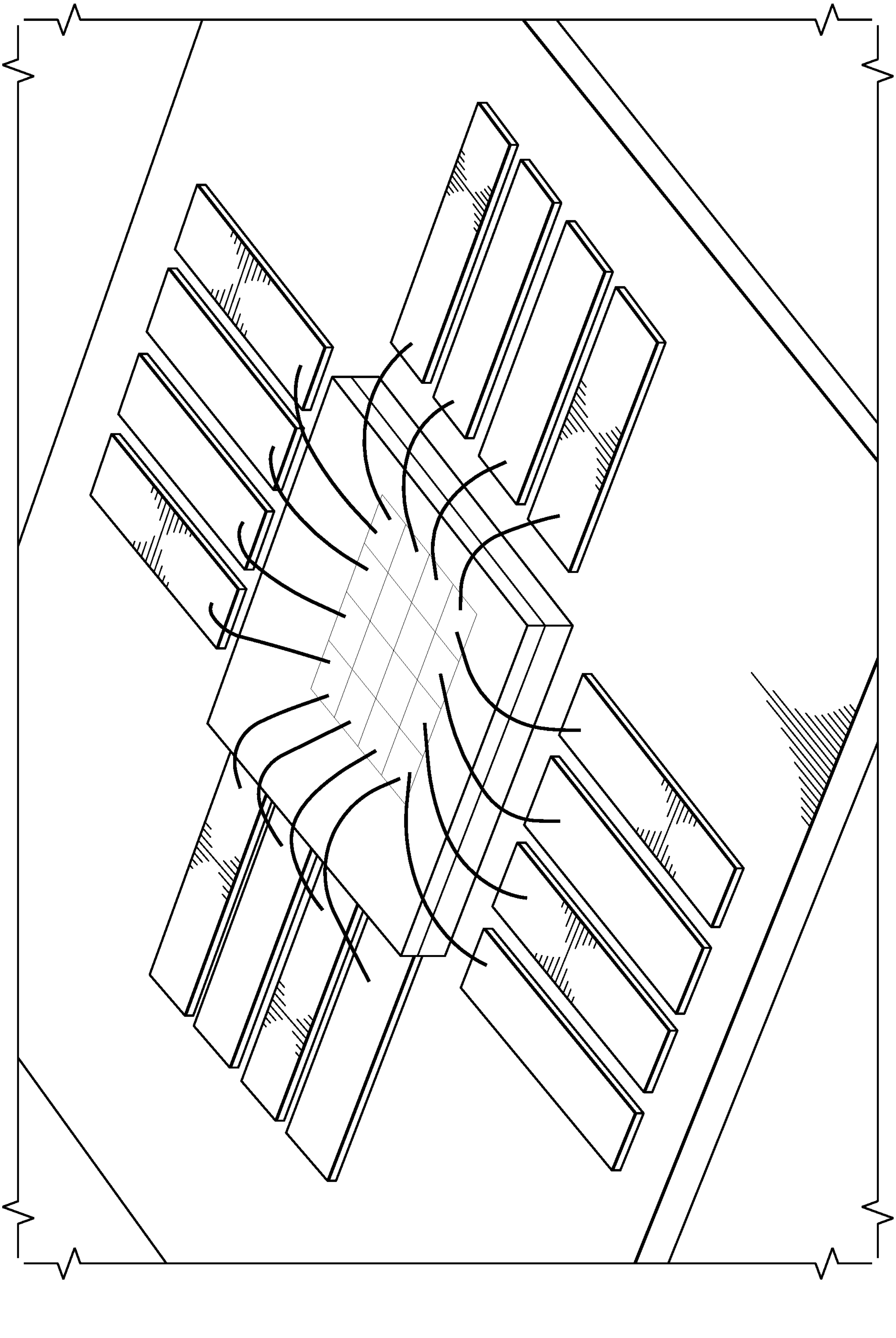


FIG. 4B

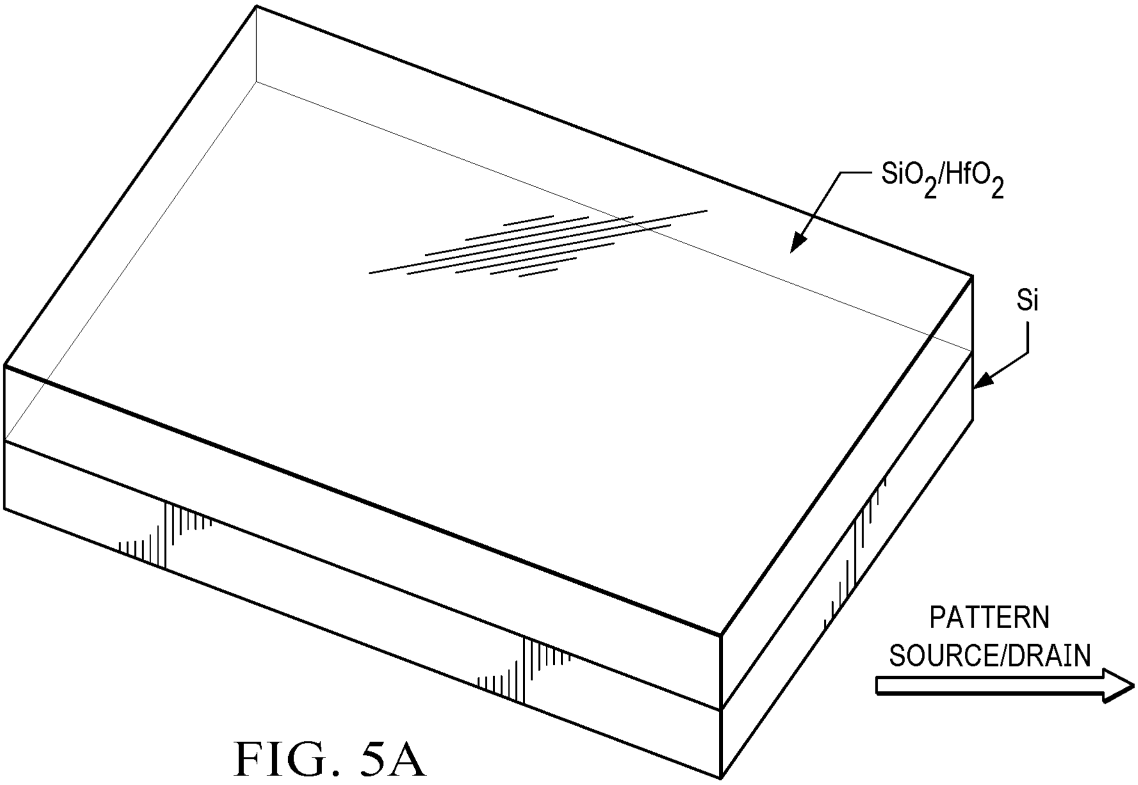


FIG. 5A

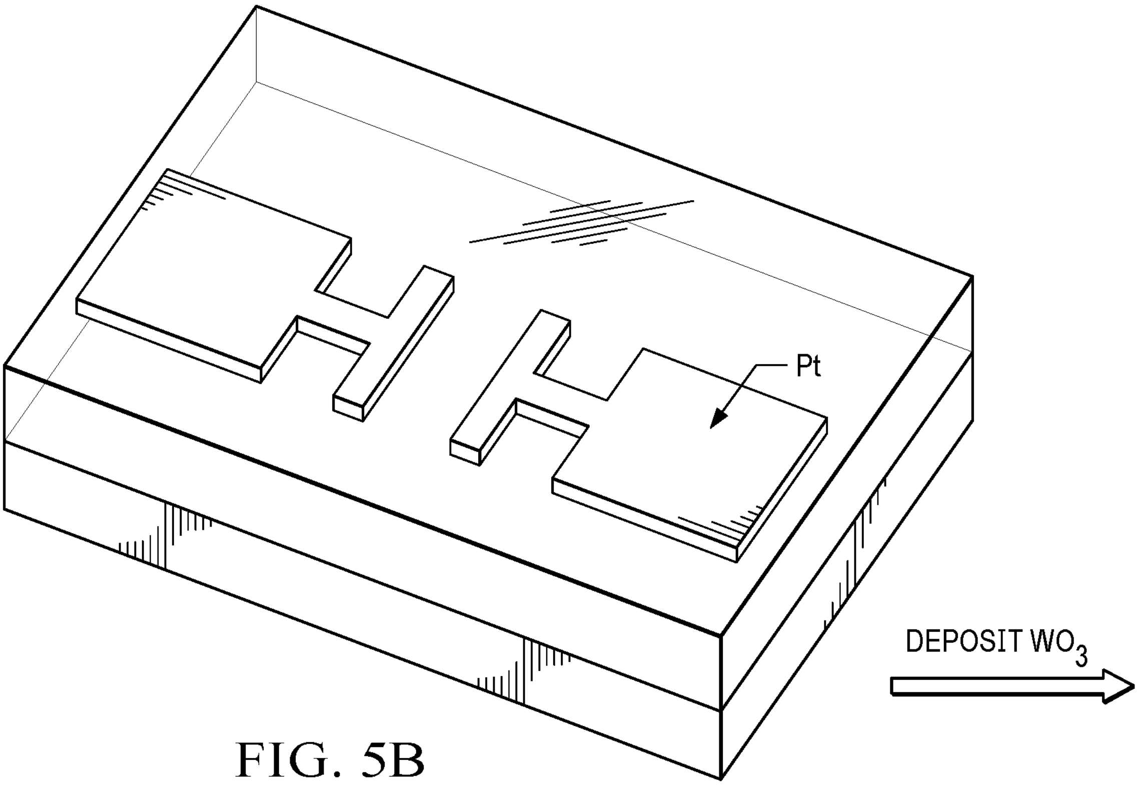


FIG. 5B

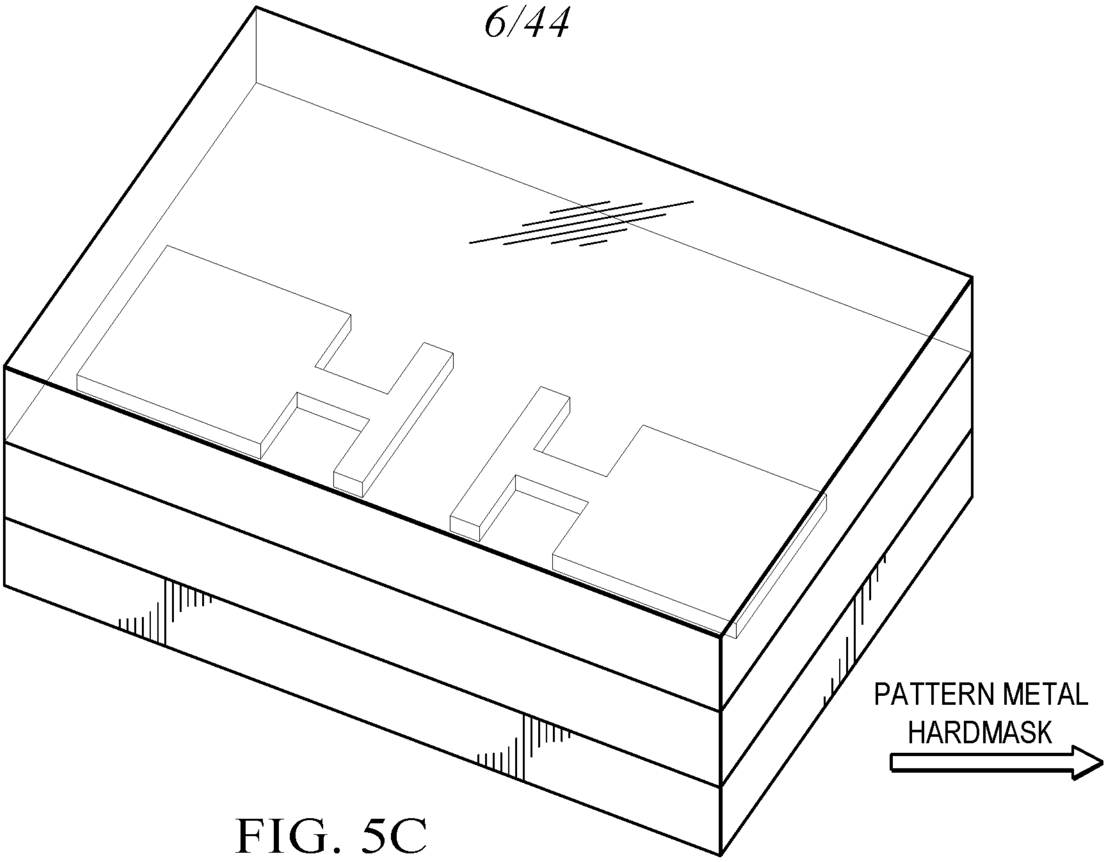


FIG. 5C

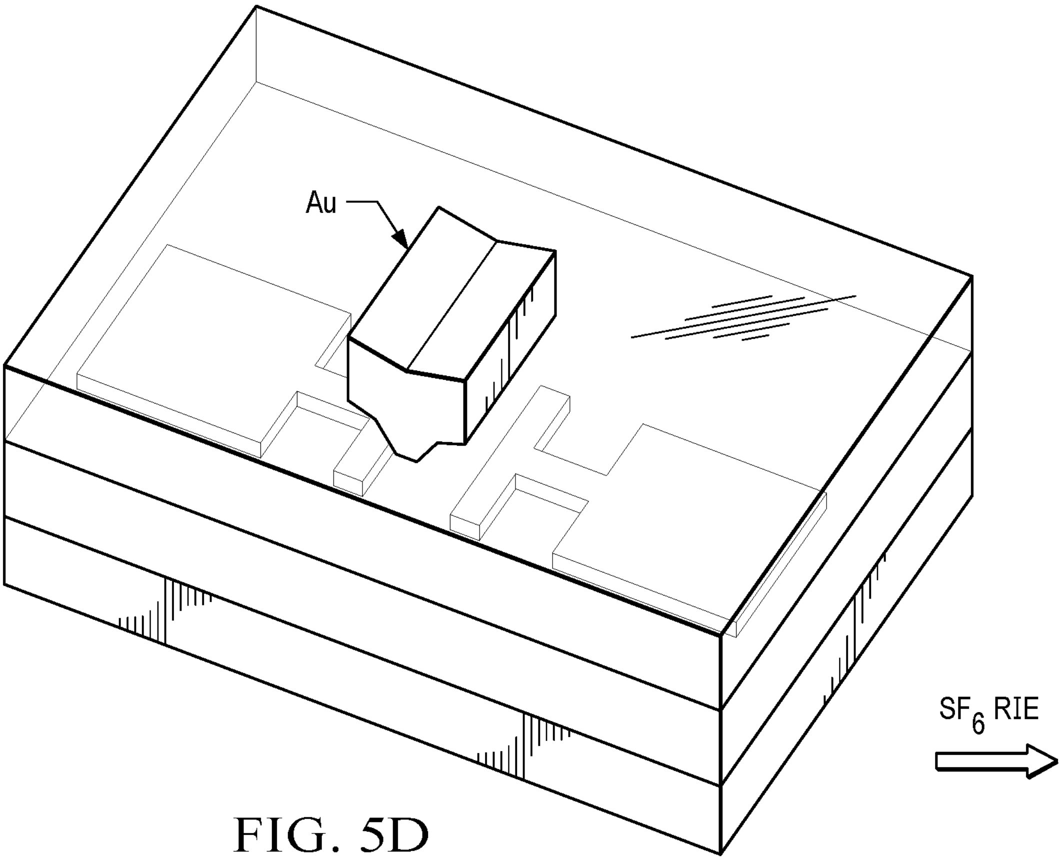


FIG. 5D

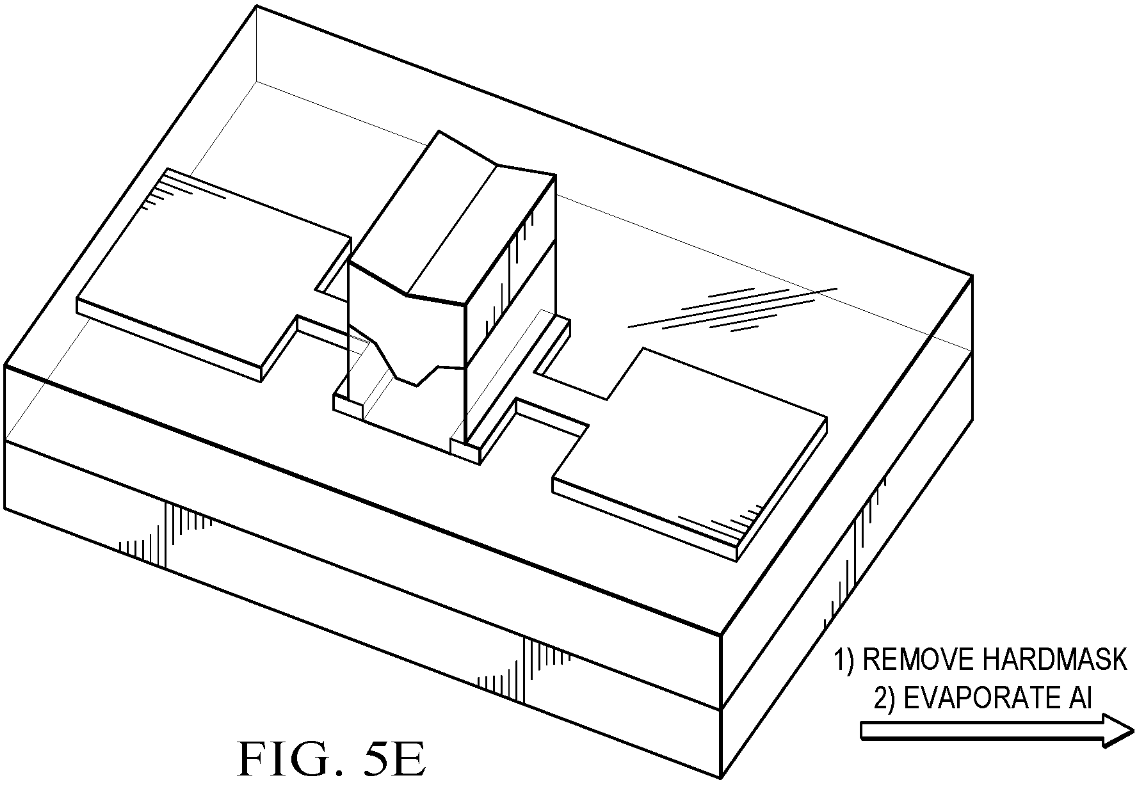


FIG. 5E

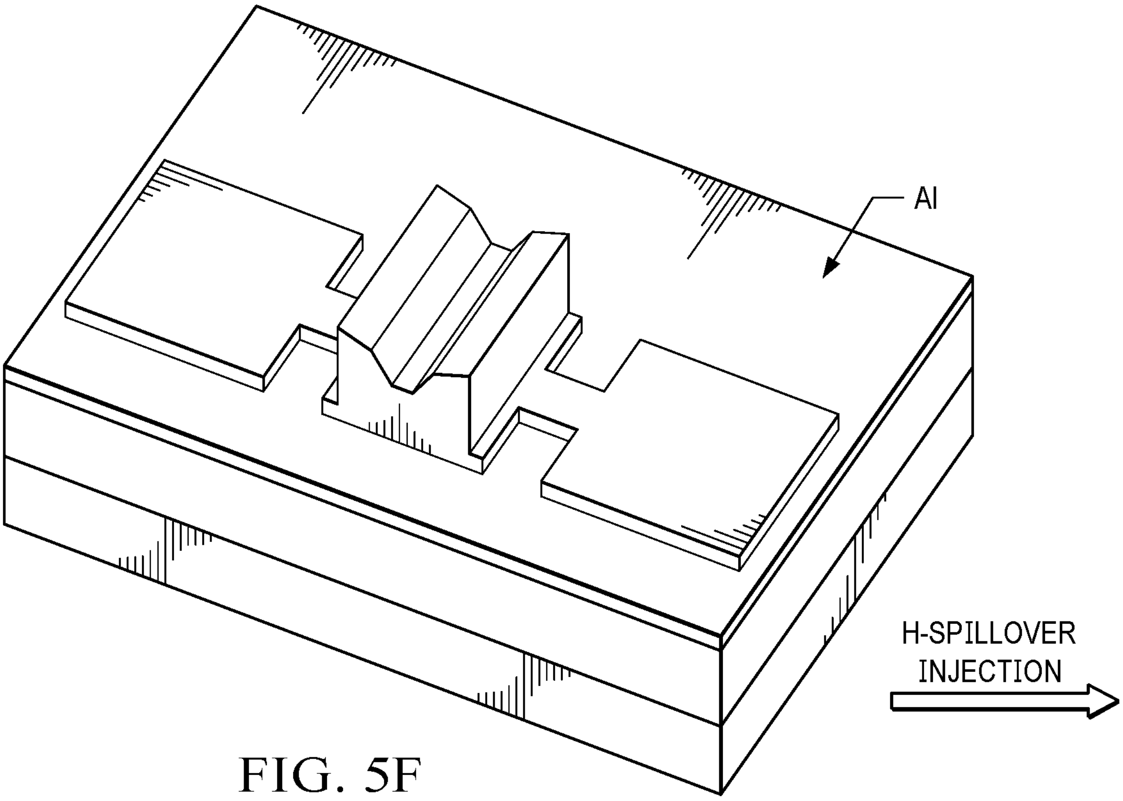


FIG. 5F

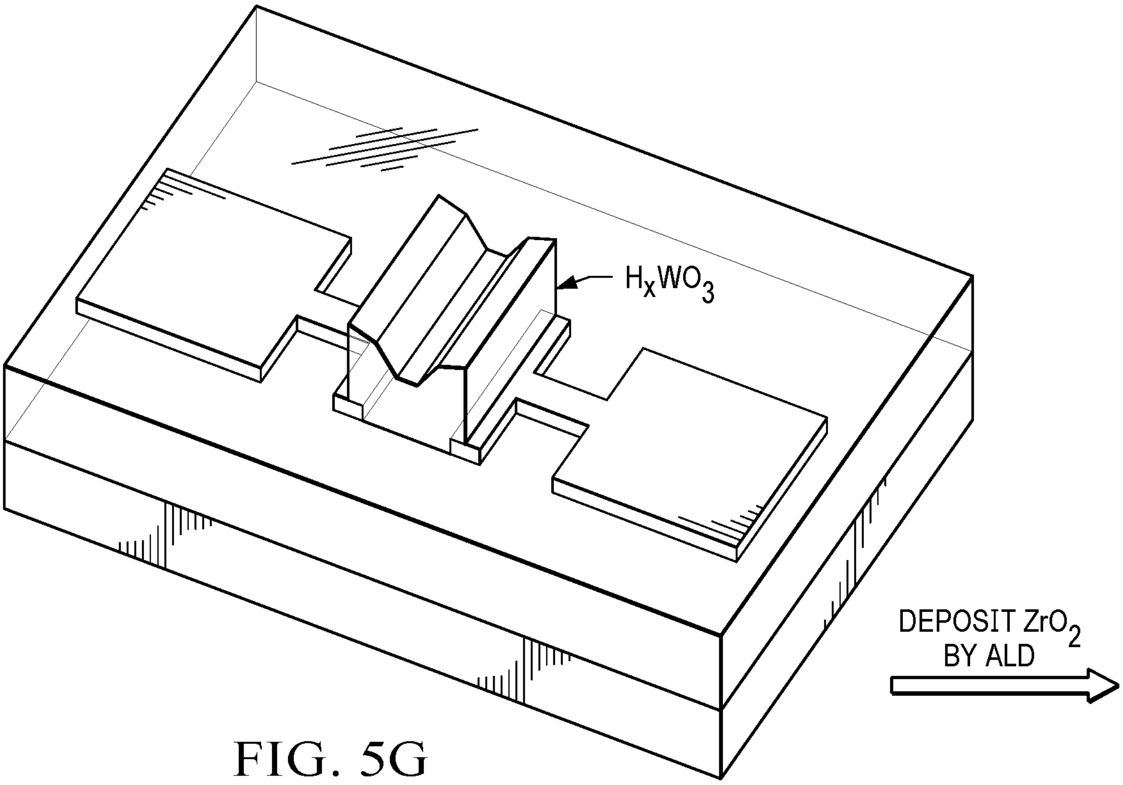


FIG. 5G

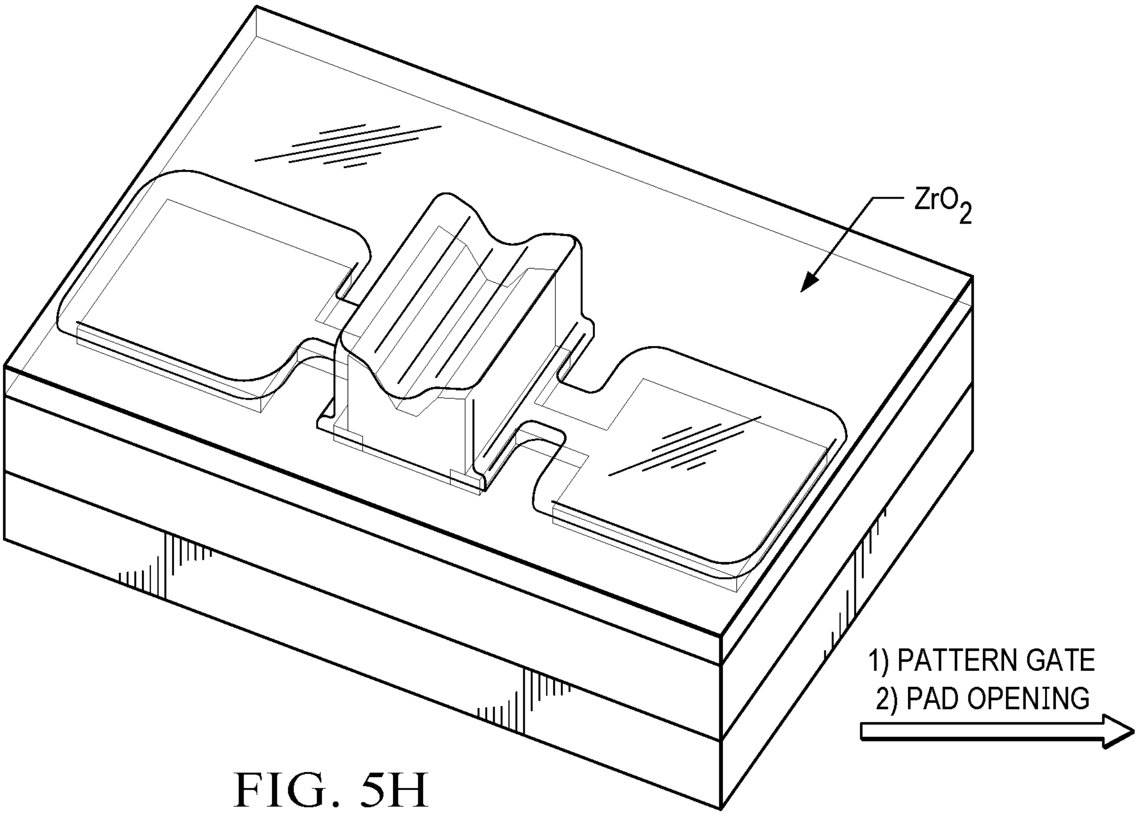


FIG. 5H

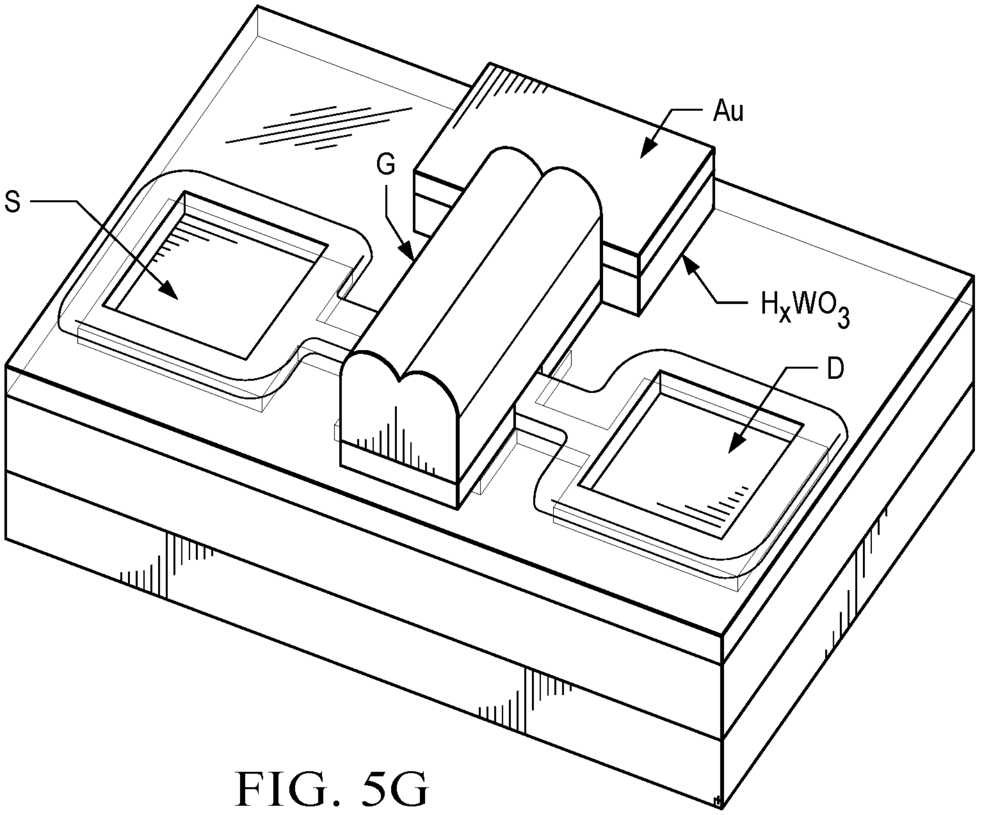

Au
G
S
H_xWO_3
D

FIG. 5G

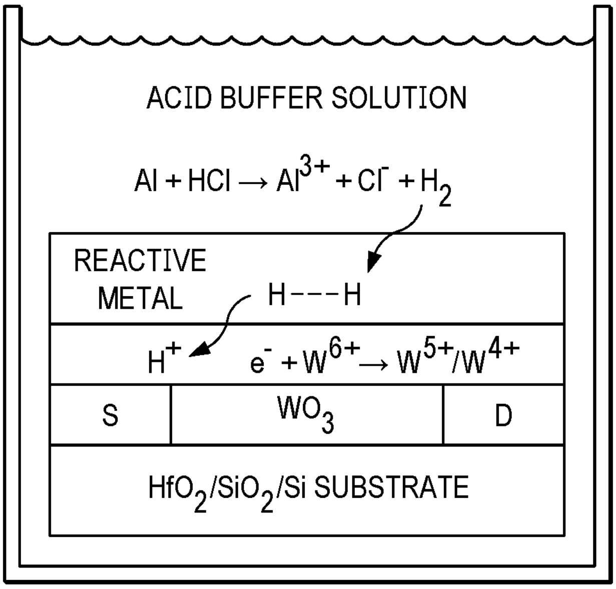

ACID BUFFER SOLUTION
$Al + HCl \rightarrow Al^{3+} + Cl^{-} + H_2$
REACTIVE METAL
H---H
H^+
$e^- + W^{6+} \rightarrow W^{5+}/W^{4+}$
S
WO_3
D
$HfO_2/SiO_2/Si$ SUBSTRATE

FIG. 6

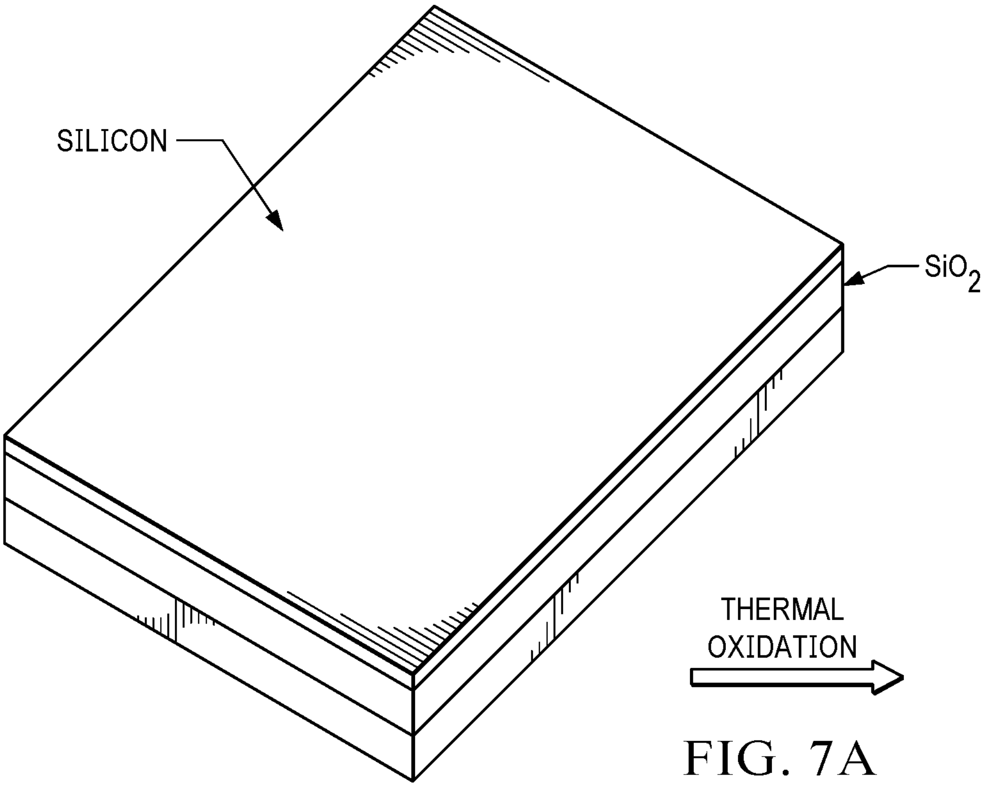


FIG. 7A

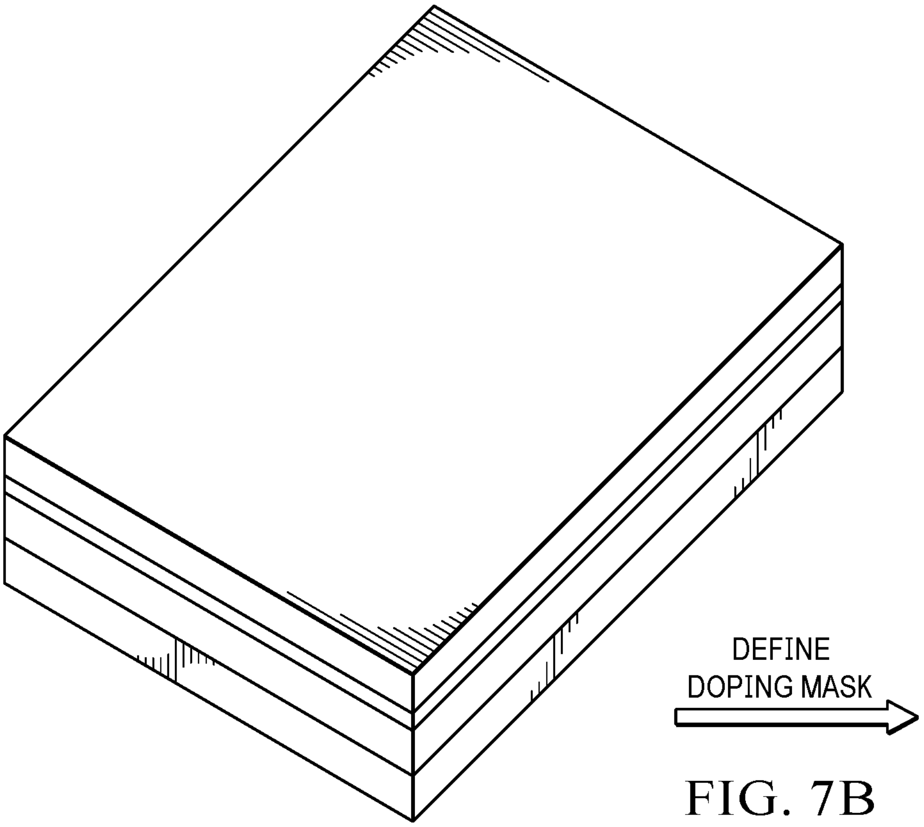


FIG. 7B

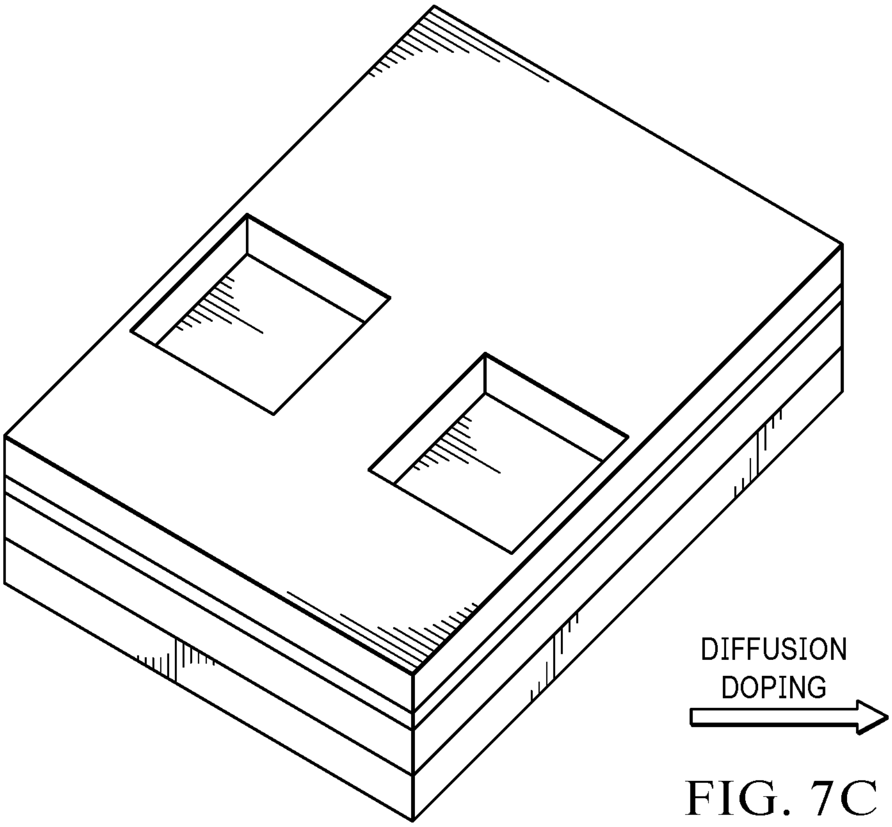

DIFFUSION
DOPING

FIG. 7C

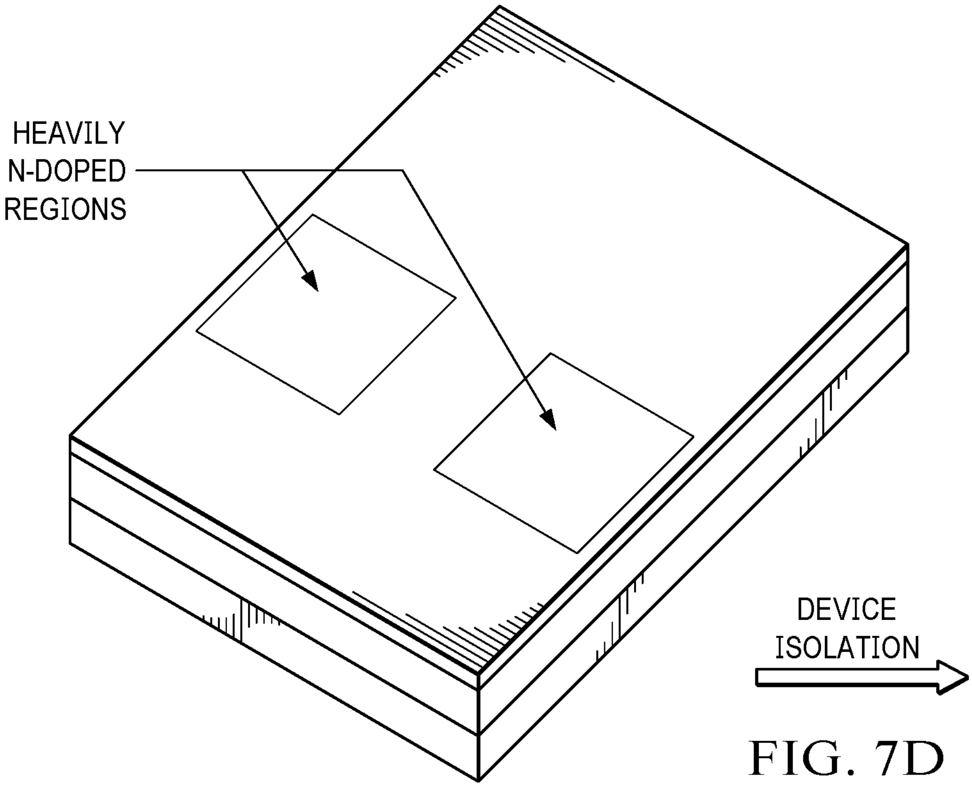

HEAVILY
N-DOPED
REGIONS
DEVICE
ISOLATION

FIG. 7D

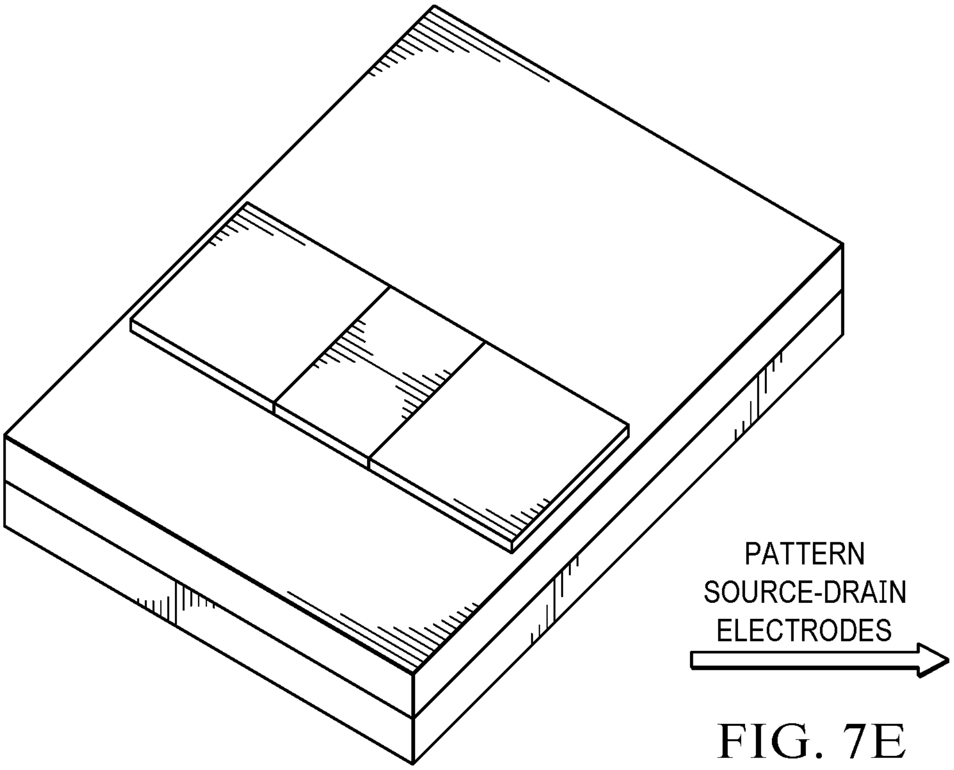

PATTERN
SOURCE-DRAIN
ELECTRODES

FIG. 7E

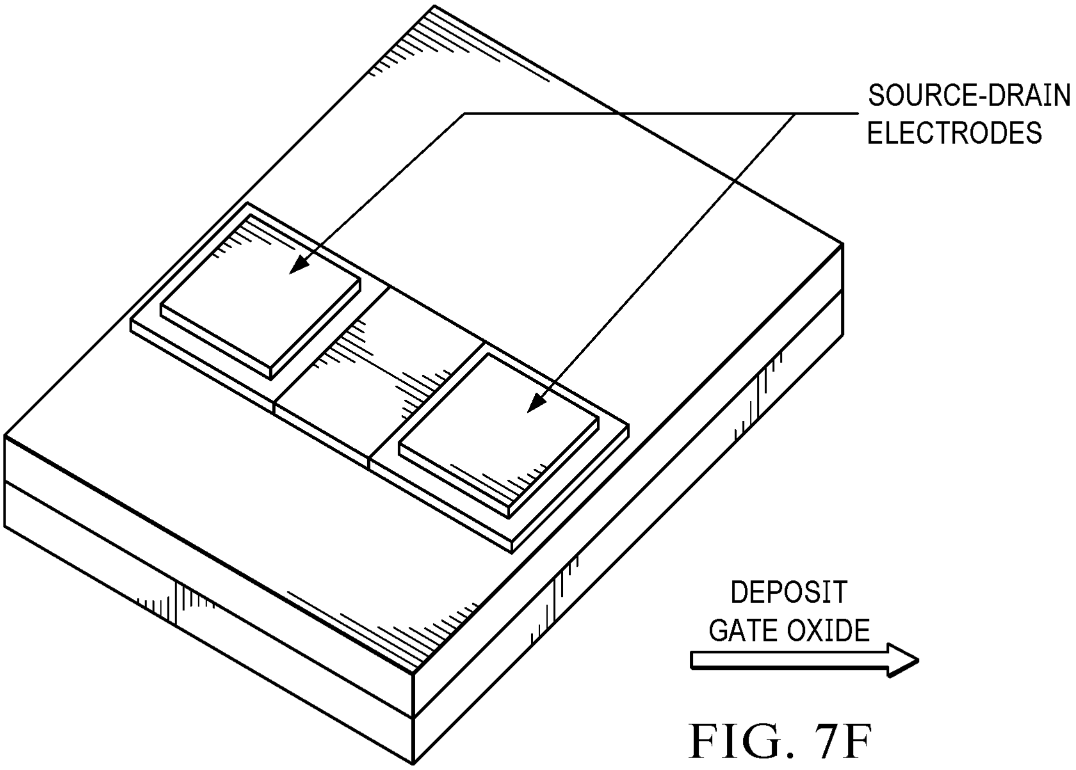

SOURCE-DRAIN
ELECTRODES
DEPOSIT
GATE OXIDE

FIG. 7F

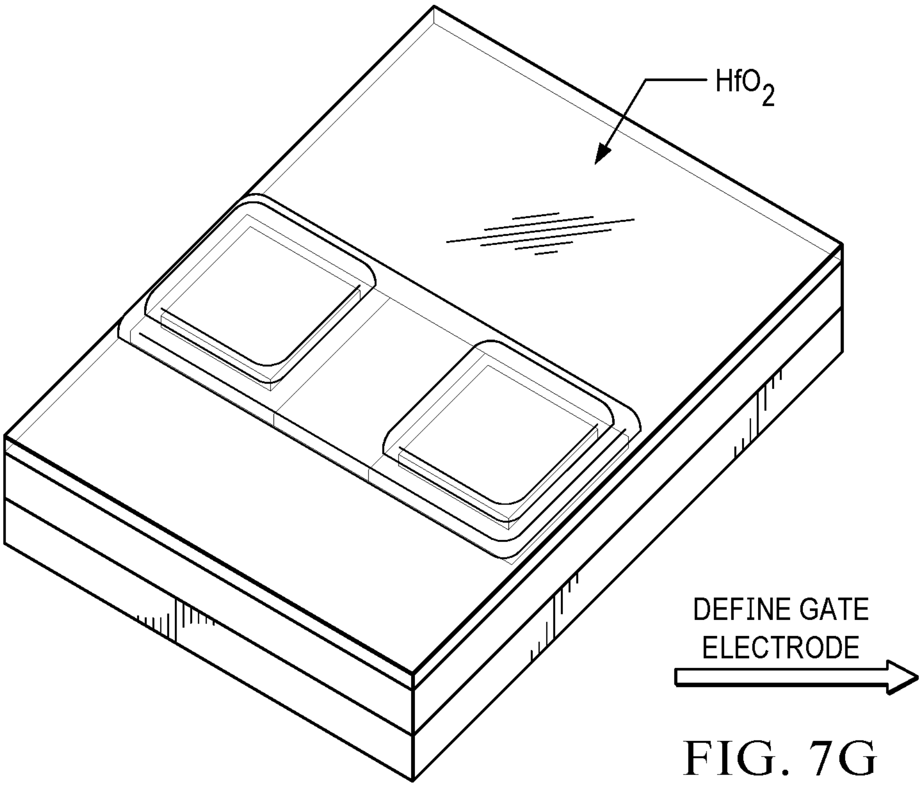

HfO_2
DEFINE GATE ELECTRODE

FIG. 7G

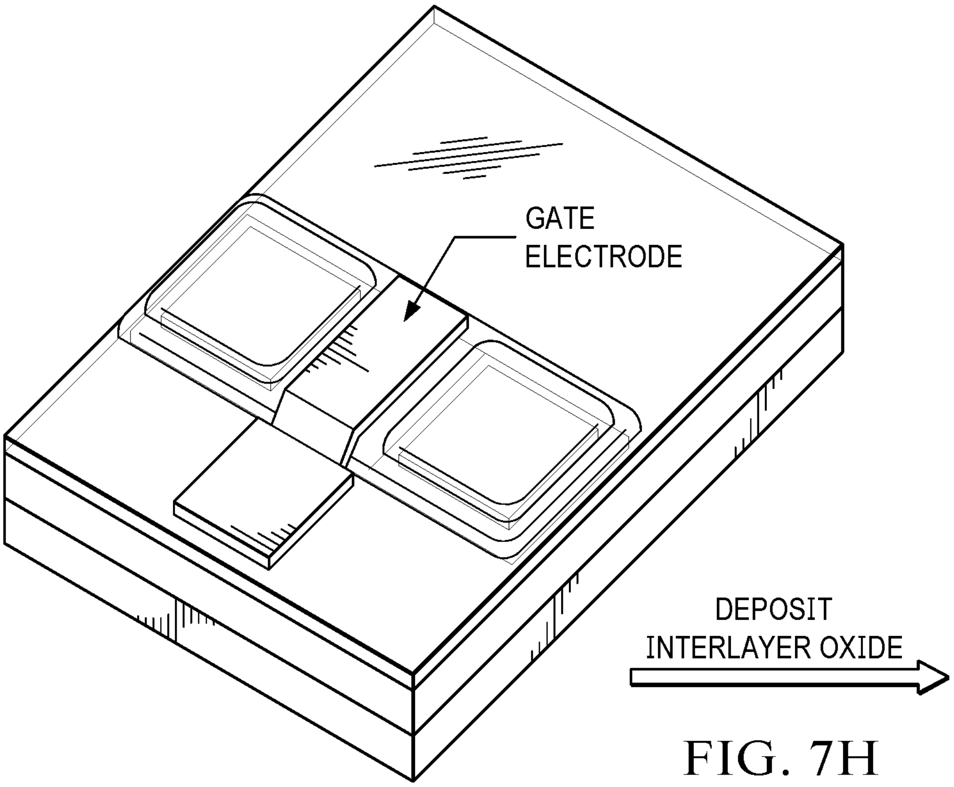

GATE ELECTRODE
DEPOSIT INTERLAYER OXIDE

FIG. 7H

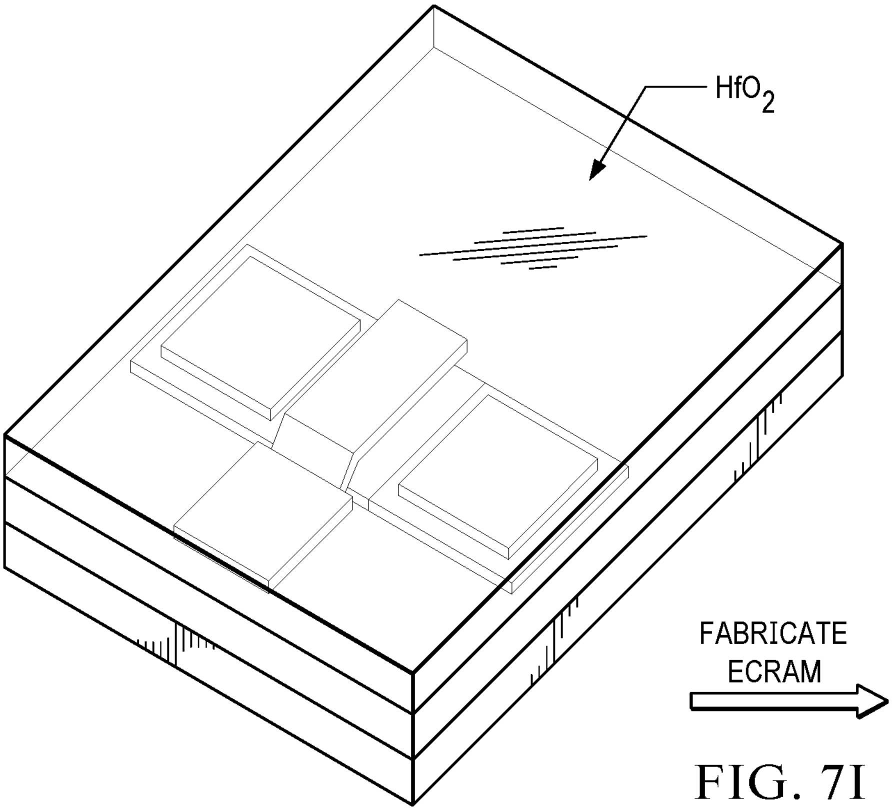


FIG. 7I

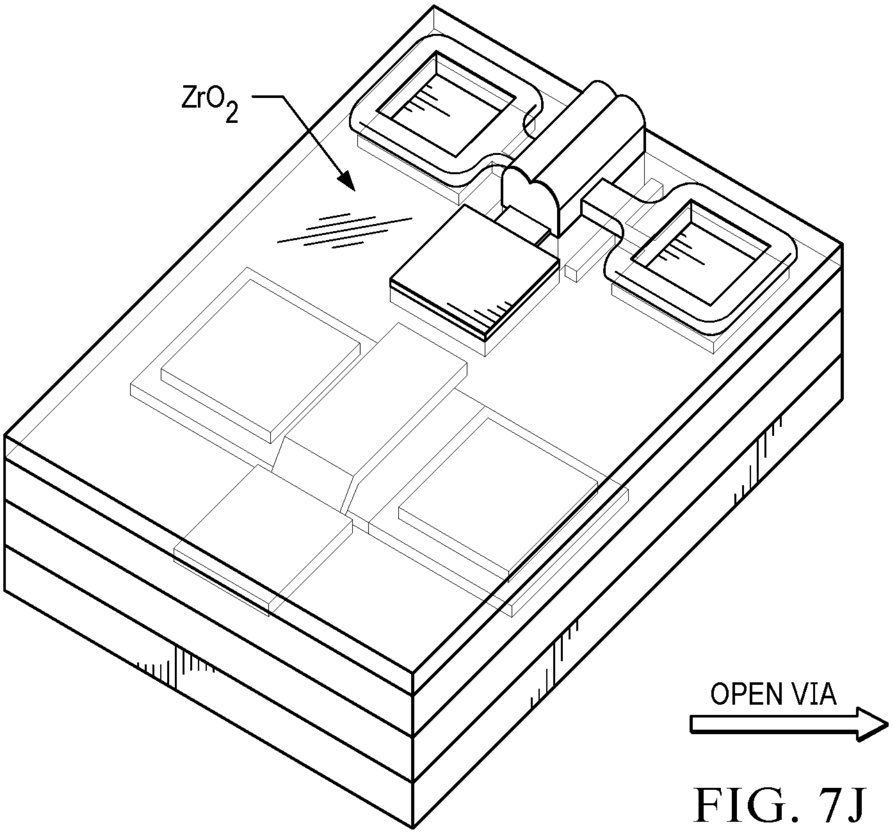


FIG. 7J

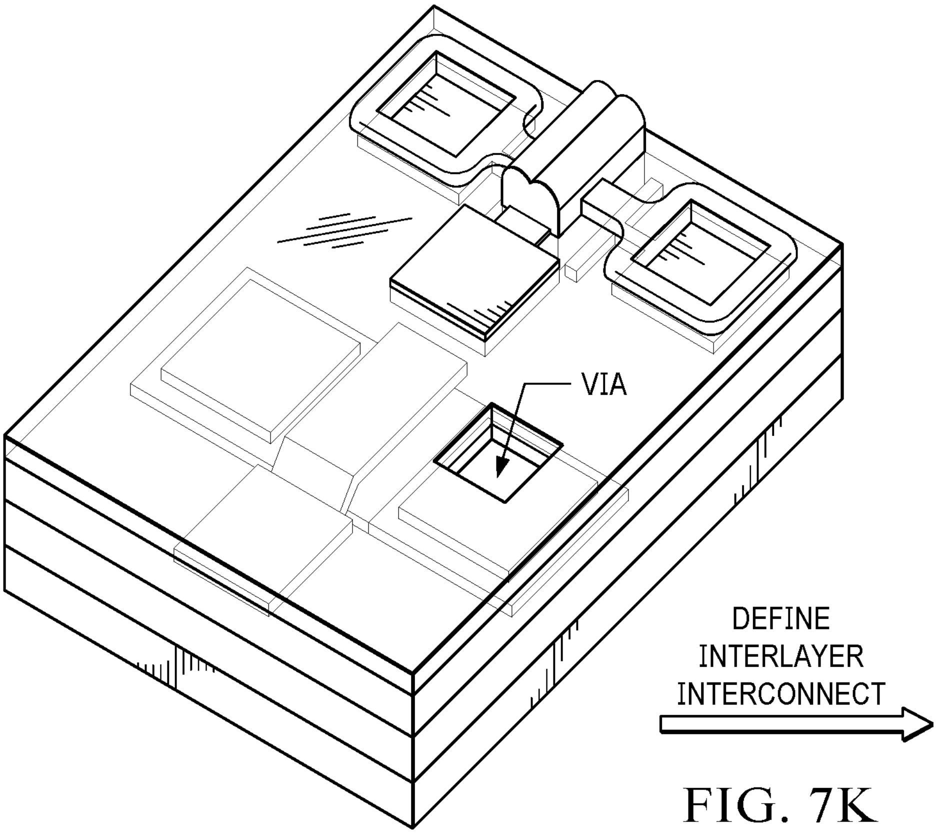


FIG. 7K

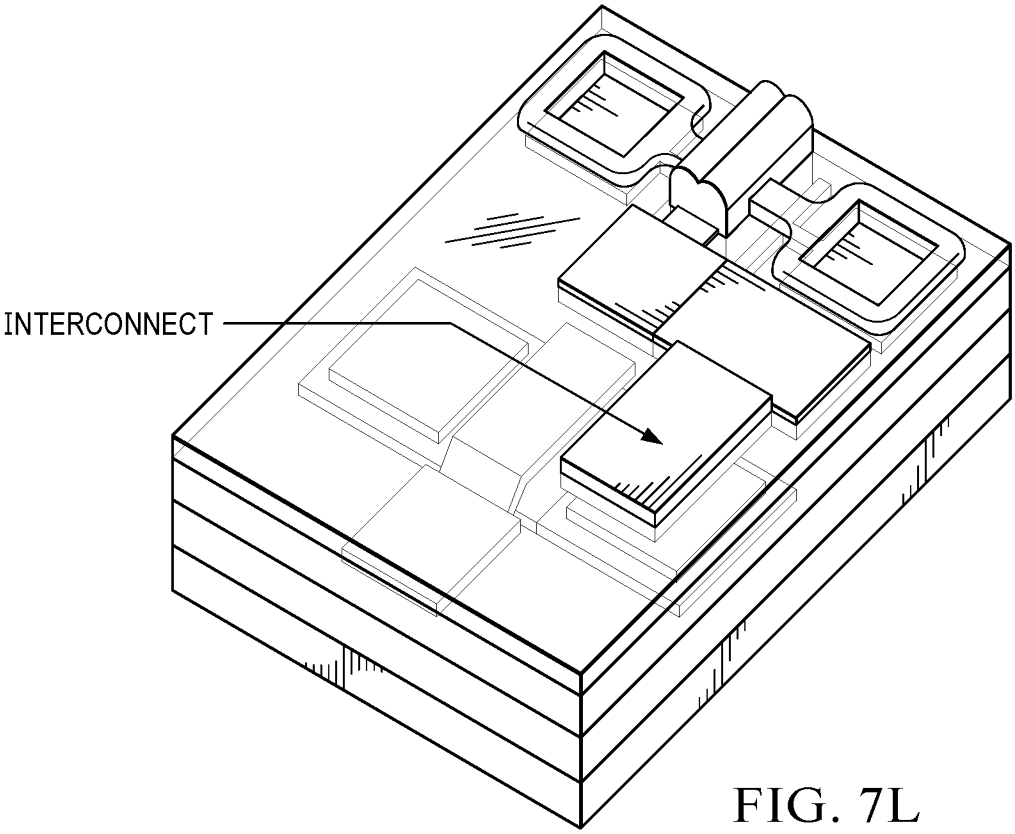
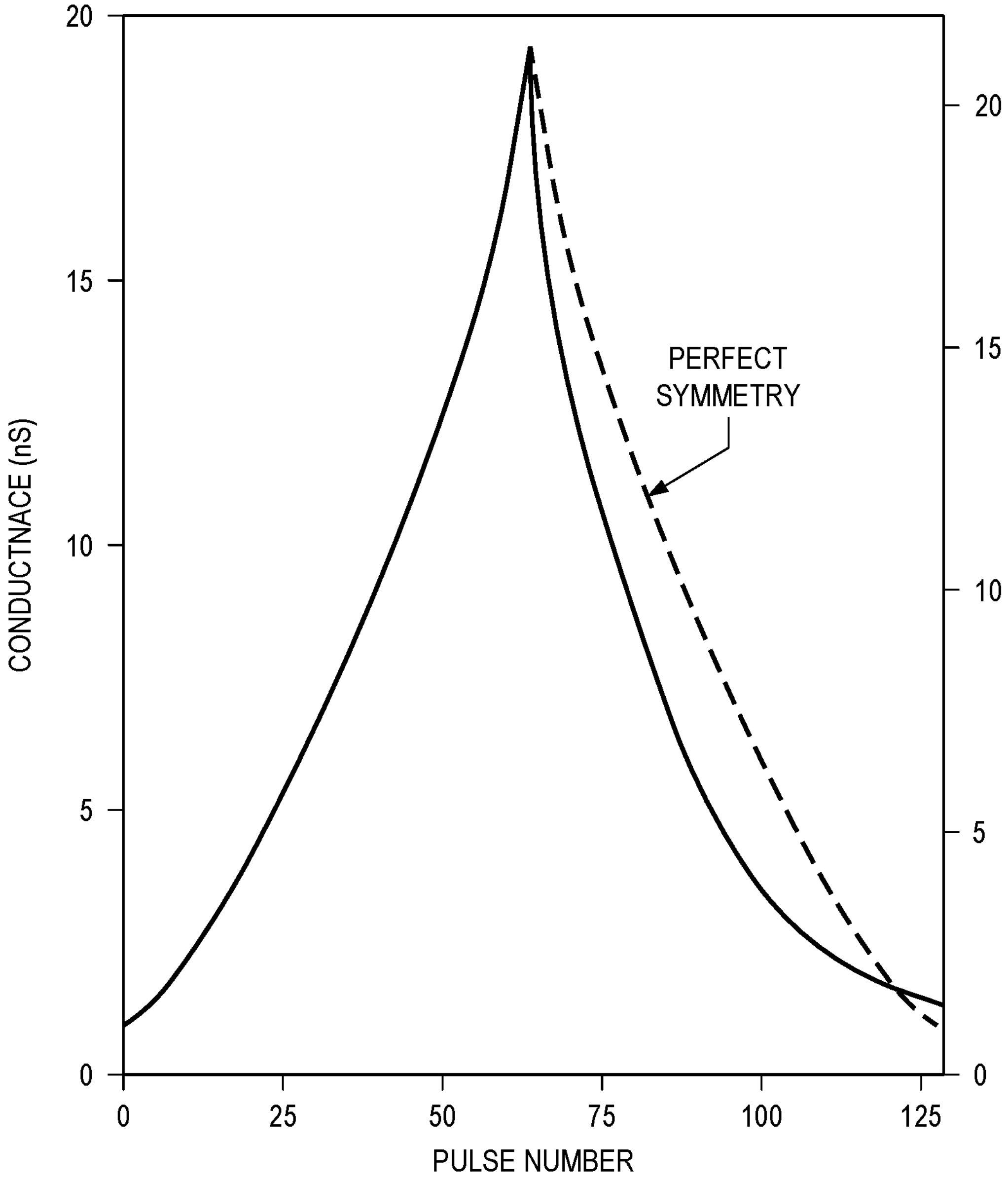


FIG. 7L

FIG. 8


20
15
10
5
0
20
15
10
5
0
0
25
50
75
100
125
CONDUCTNACE (nS)
PULSE NUMBER
PERFECT
SYMMETRY

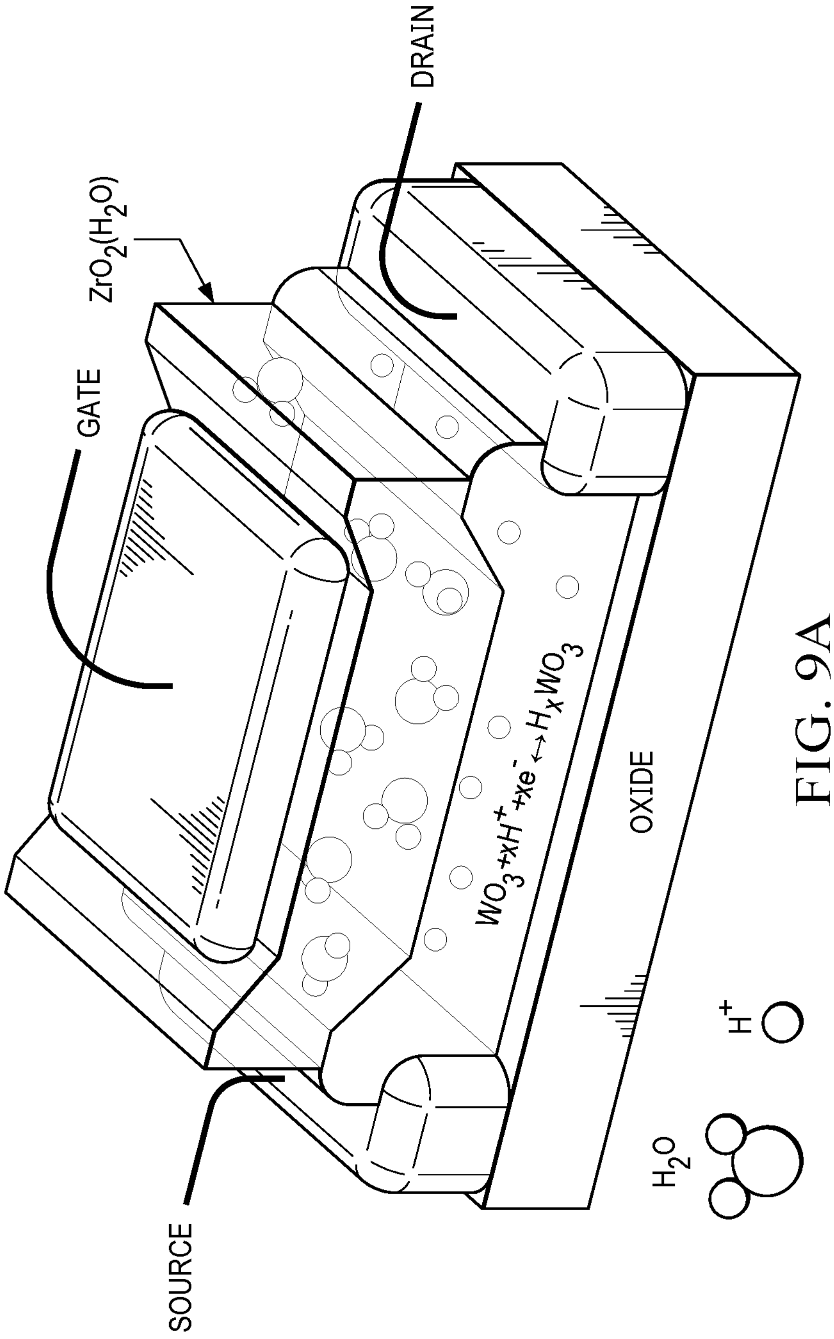

GATE
$ZrO_2(H_2O)$
DRAIN
SOURCE
$WO_3+xH^++xe^- \leftrightarrow H_xWO_3$
OXIDE
H_2O
H^+

FIG. 9A

FIG. 9B

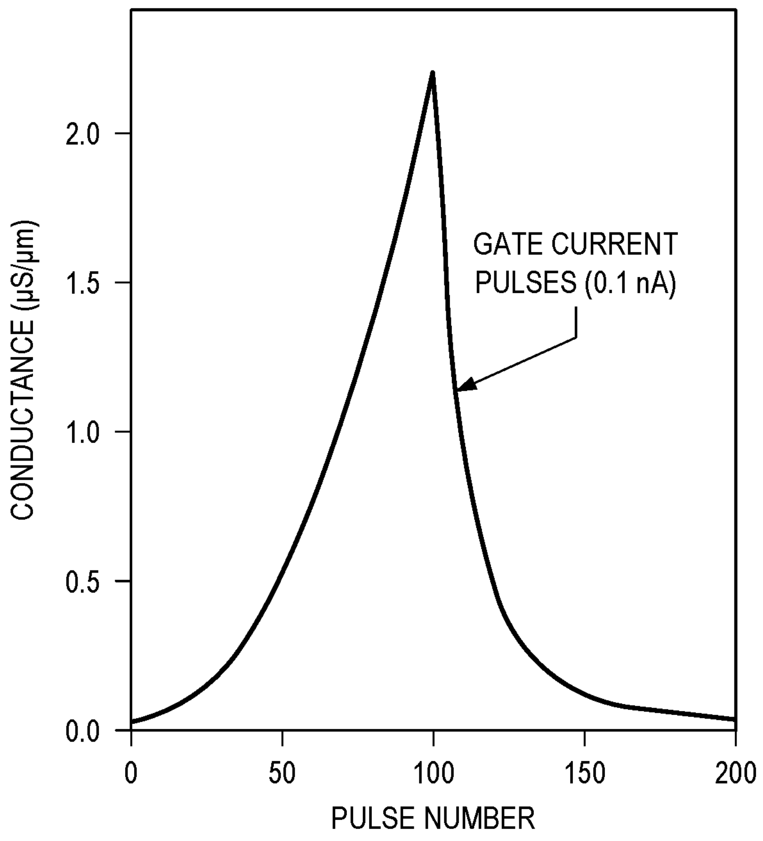

GATE CURRENT
PULSES (0.1 nA)
CONDUCTANCE (µS/µm)
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
PULSE NUMBER

FIG. 9C

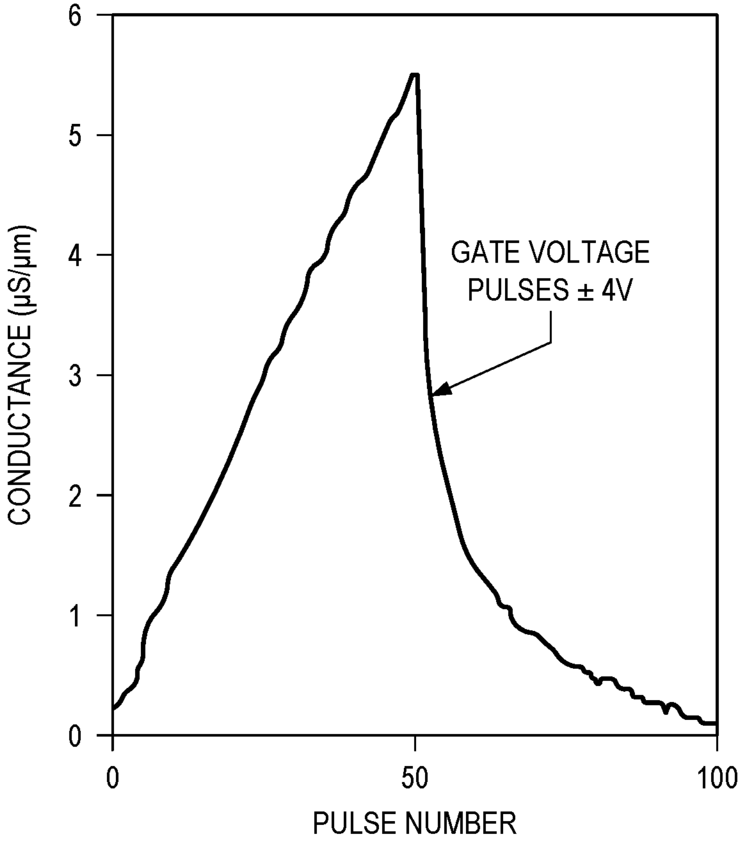

GATE VOLTAGE
PULSES ± 4V
CONDUCTANCE (µS/µm)
6
5
4
3
2
1
0
0
50
100
PULSE NUMBER

FIG. 10

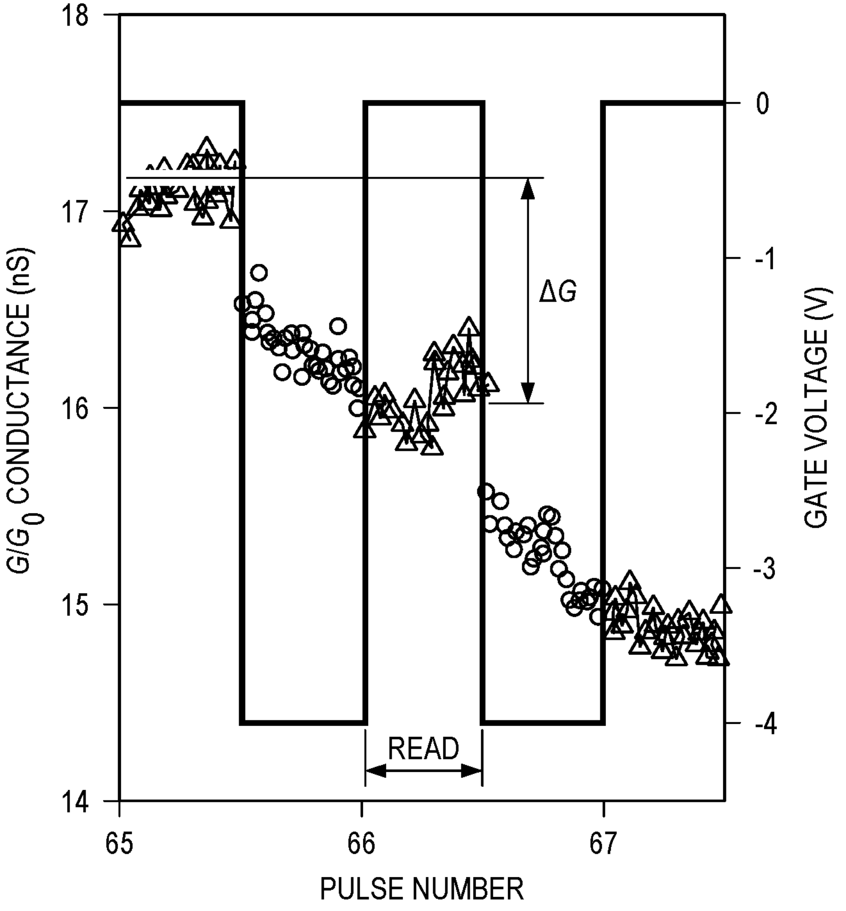

18
17
16
15
14
G/G$_0$ CONDUCTANCE (nS)
0
-1
-2
-3
-4
GATE VOLTAGE (V)
ΔG
READ
65
66
67
PULSE NUMBER

FIG. 11

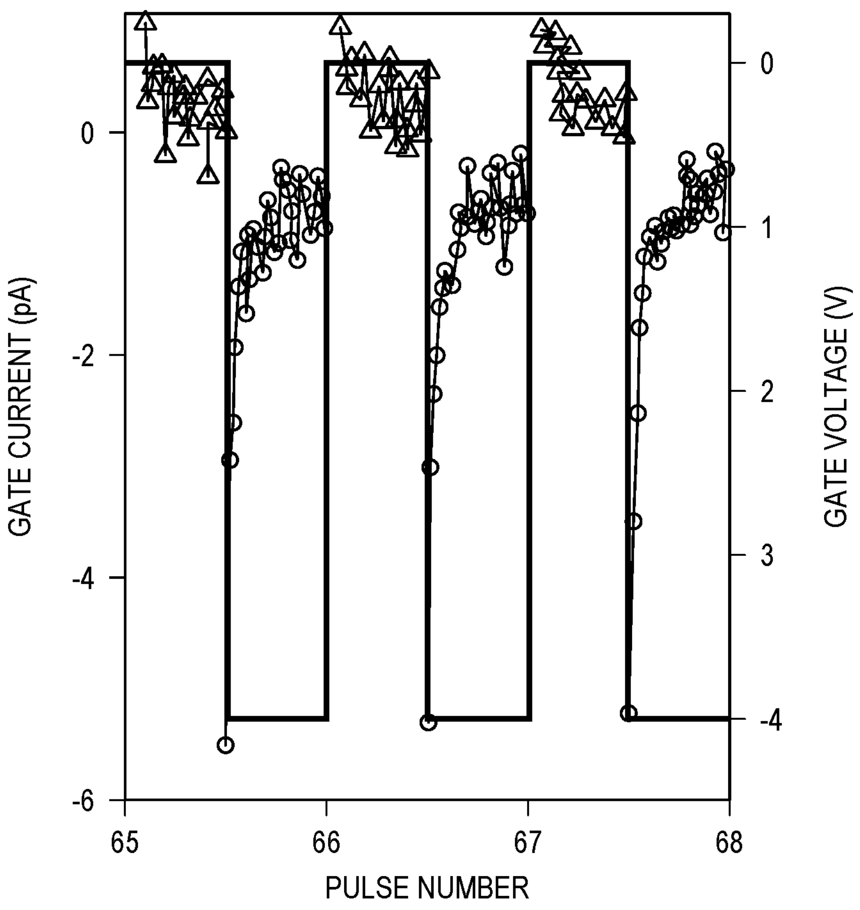

0
-2
-4
-6
GATE CURRENT (pA)
0
1
2
3
-4
GATE VOLTAGE (V)
65
66
67
68
PULSE NUMBER

FIG. 12

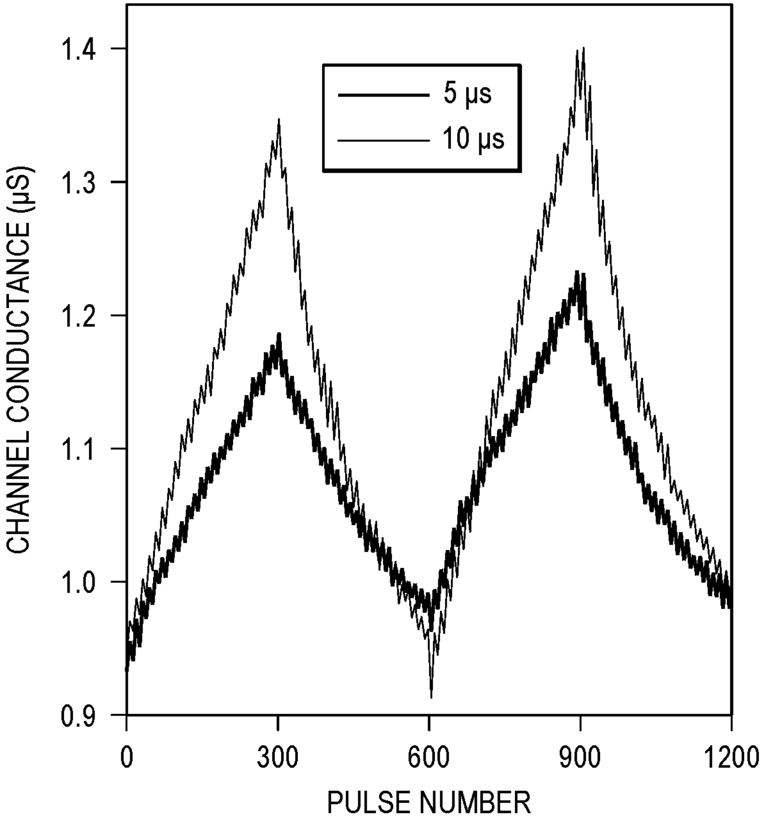

CHANNEL CONDUCTANCE (μS)
1.4
1.3
1.2
1.1
1.0
0.9
0
300
600
900
1200
PULSE NUMBER
5 μs
10 μs

FIG. 13

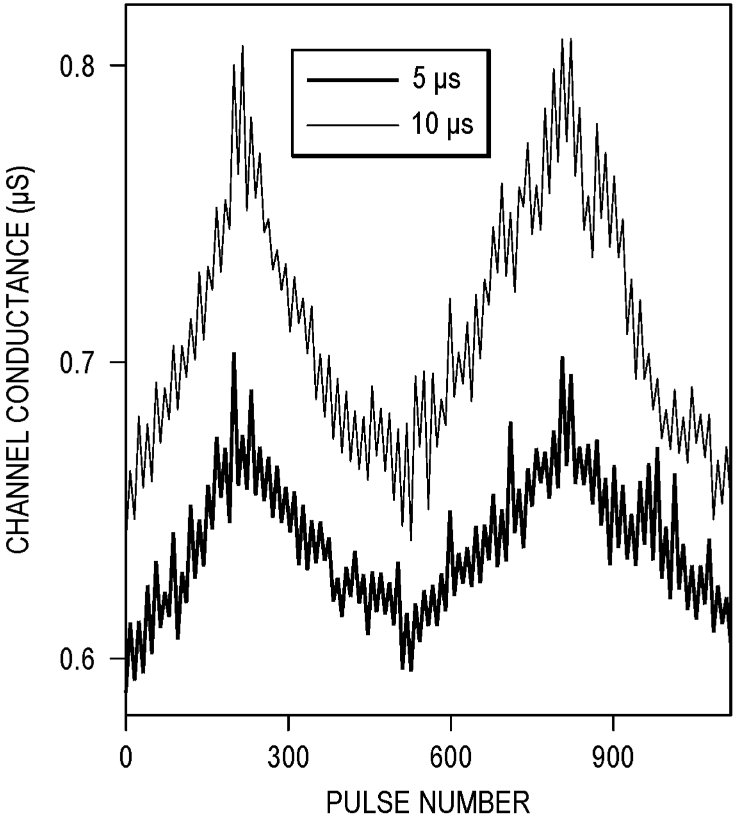

CHANNEL CONDUCTANCE (μS)
0.8
0.7
0.6
0
300
600
900
PULSE NUMBER
5 μs
10 μs

FIG. 14

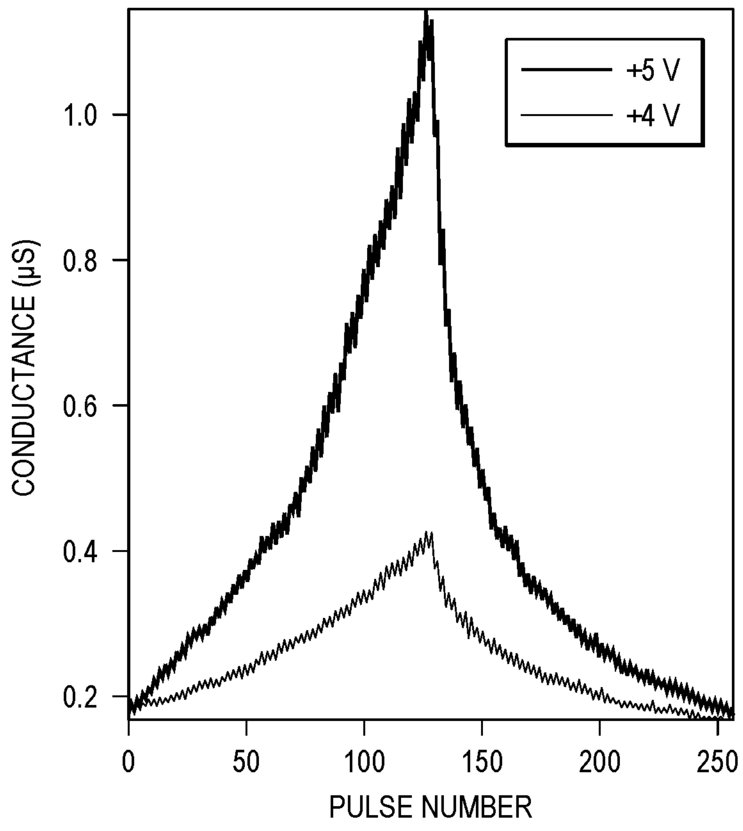

CONDUCTANCE (μS)
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
PULSE NUMBER
+5 V
+4 V

FIG. 15

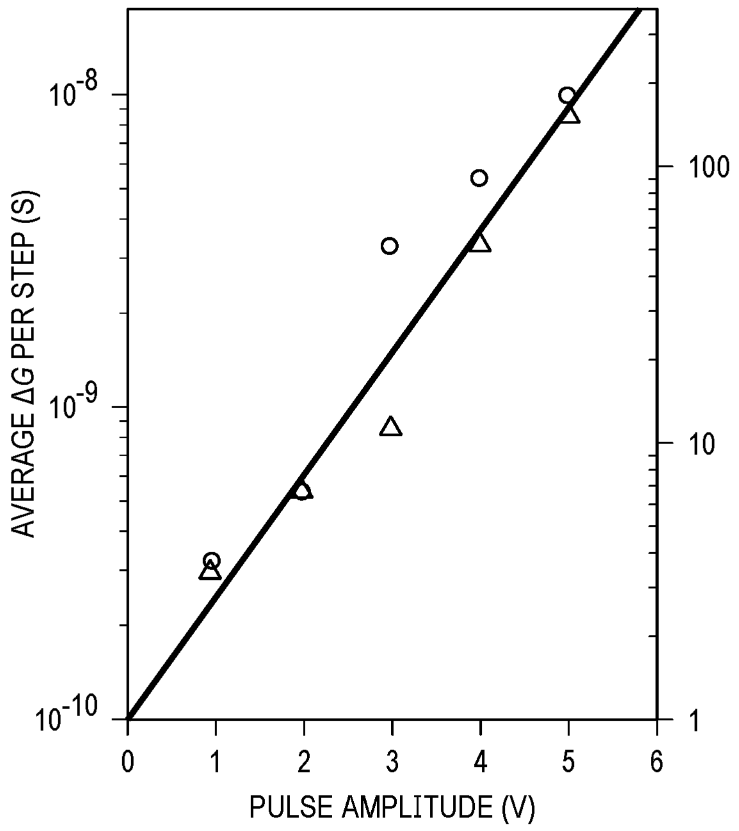

AVERAGE ΔG PER STEP (S)
10^{-8}
10^{-9}
10^{-10}
0
1
2
3
4
5
6
PULSE AMPLITUDE (V)
100
10
1

FIG. 16

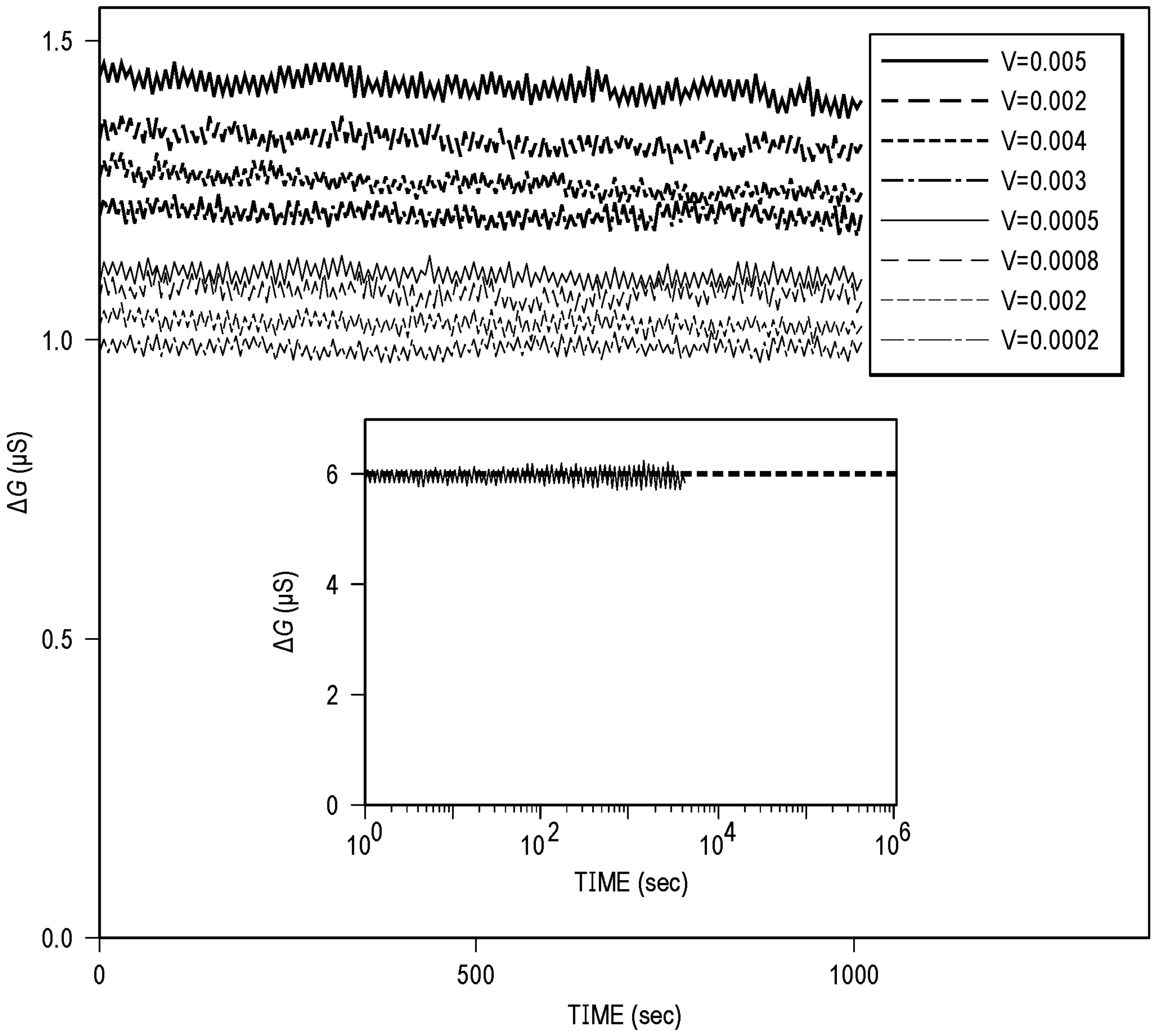

1.5
1.0
0.5
0.0
ΔG (μS)
0
500
1000
TIME (sec)
V=0.005
V=0.002
V=0.004
V=0.003
V=0.0005
V=0.0008
V=0.002
V=0.0002
6
4
2
0
ΔG (μS)
10^0
10^2
10^4
10^6
TIME (sec)

FIG. 17

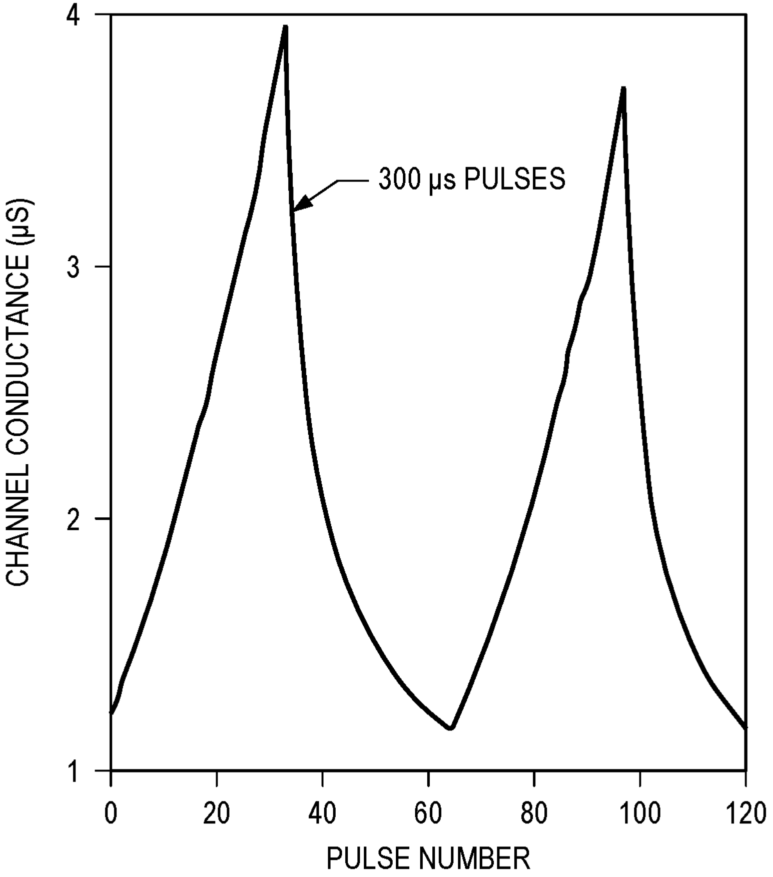

300 μs PULSES
CHANNEL CONDUCTANCE (μS)
4
3
2
1
0
20
40
60
80
100
120
PULSE NUMBER

FIG. 18

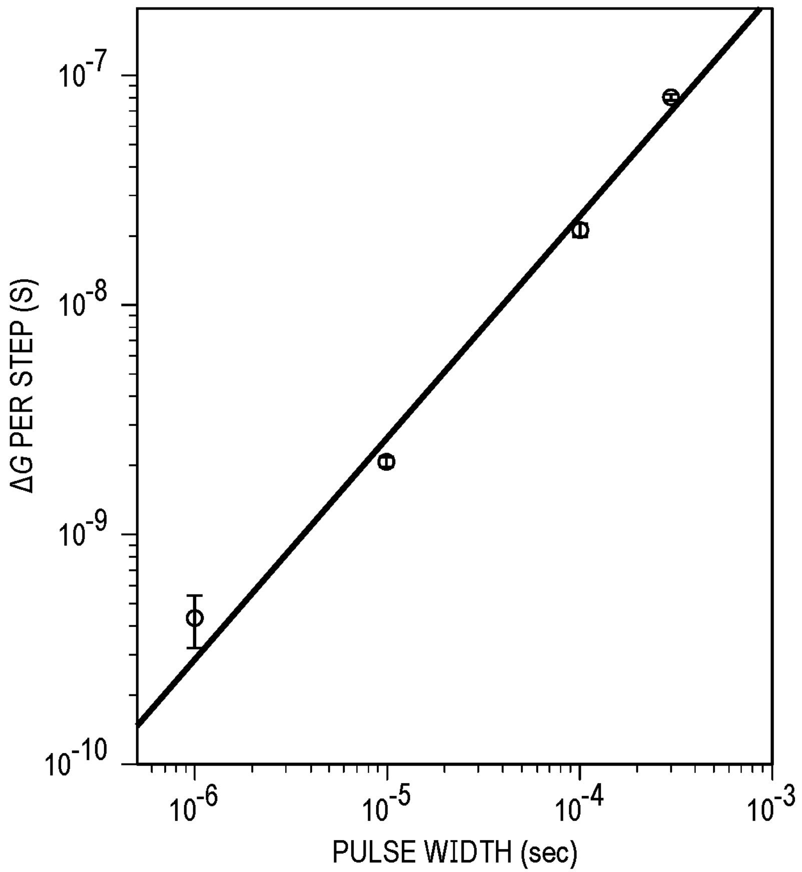

ΔG PER STEP (S)
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
PULSE WIDTH (sec)

FIG. 19

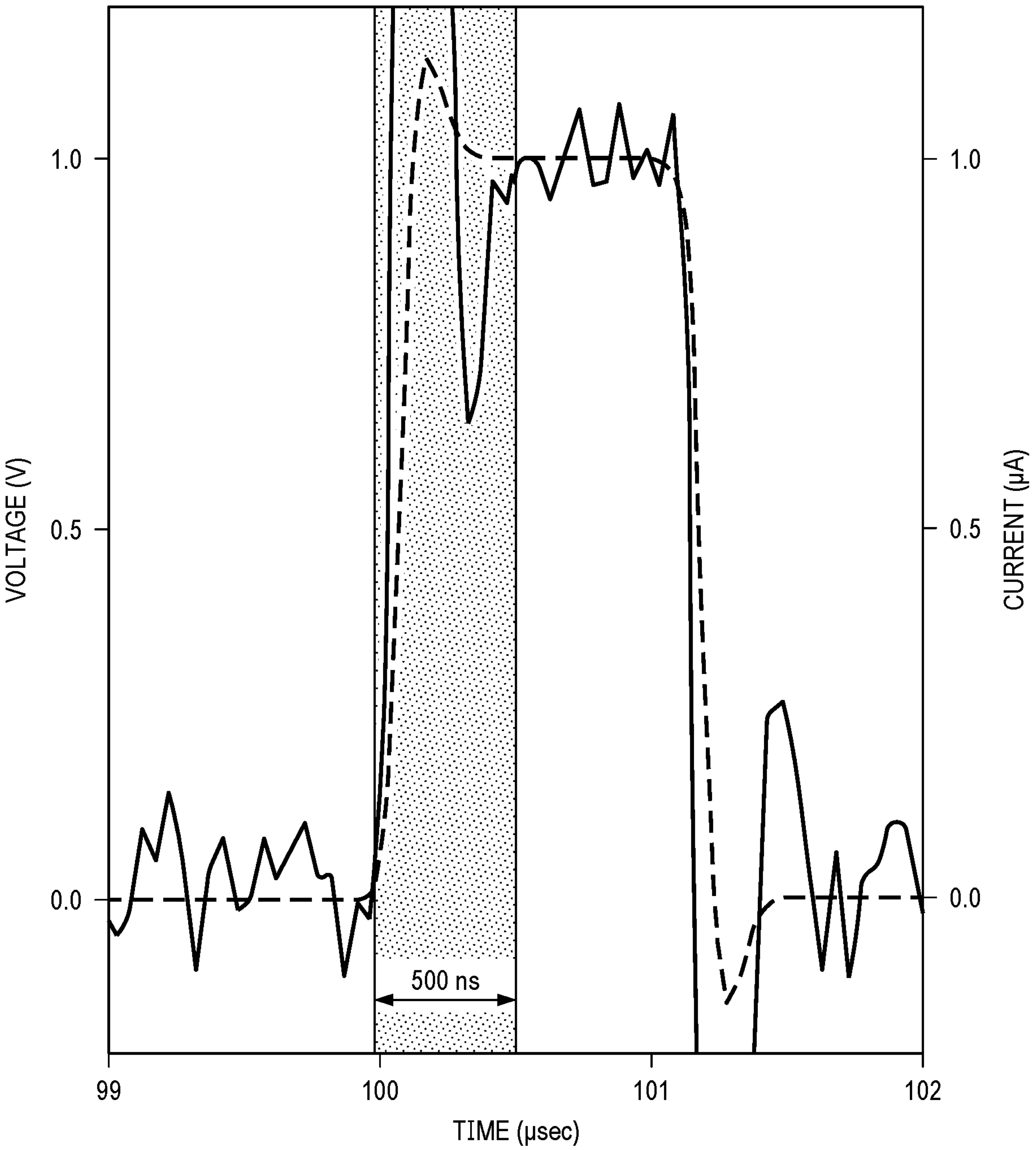
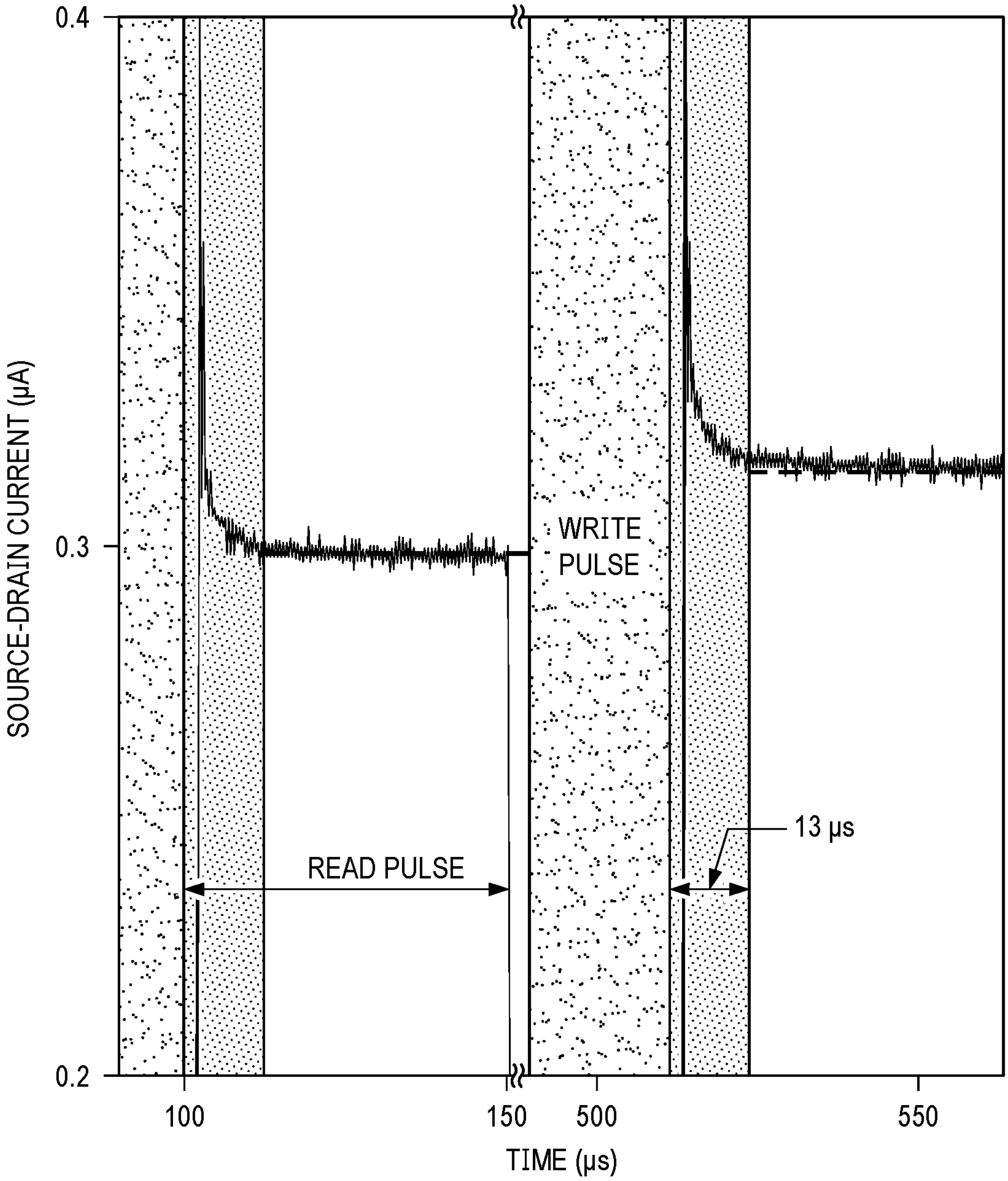

VOLTAGE (V)
1.0
0.5
0.0
CURRENT (μA)
1.0
0.5
0.0
500 ns
99
100
101
102
TIME (μsec)

FIG. 20


0.4
0.3
0.2
SOURCE-DRAIN CURRENT (μA)
WRITE PULSE
READ PULSE
13 μs
100
150
500
550
TIME (μs)

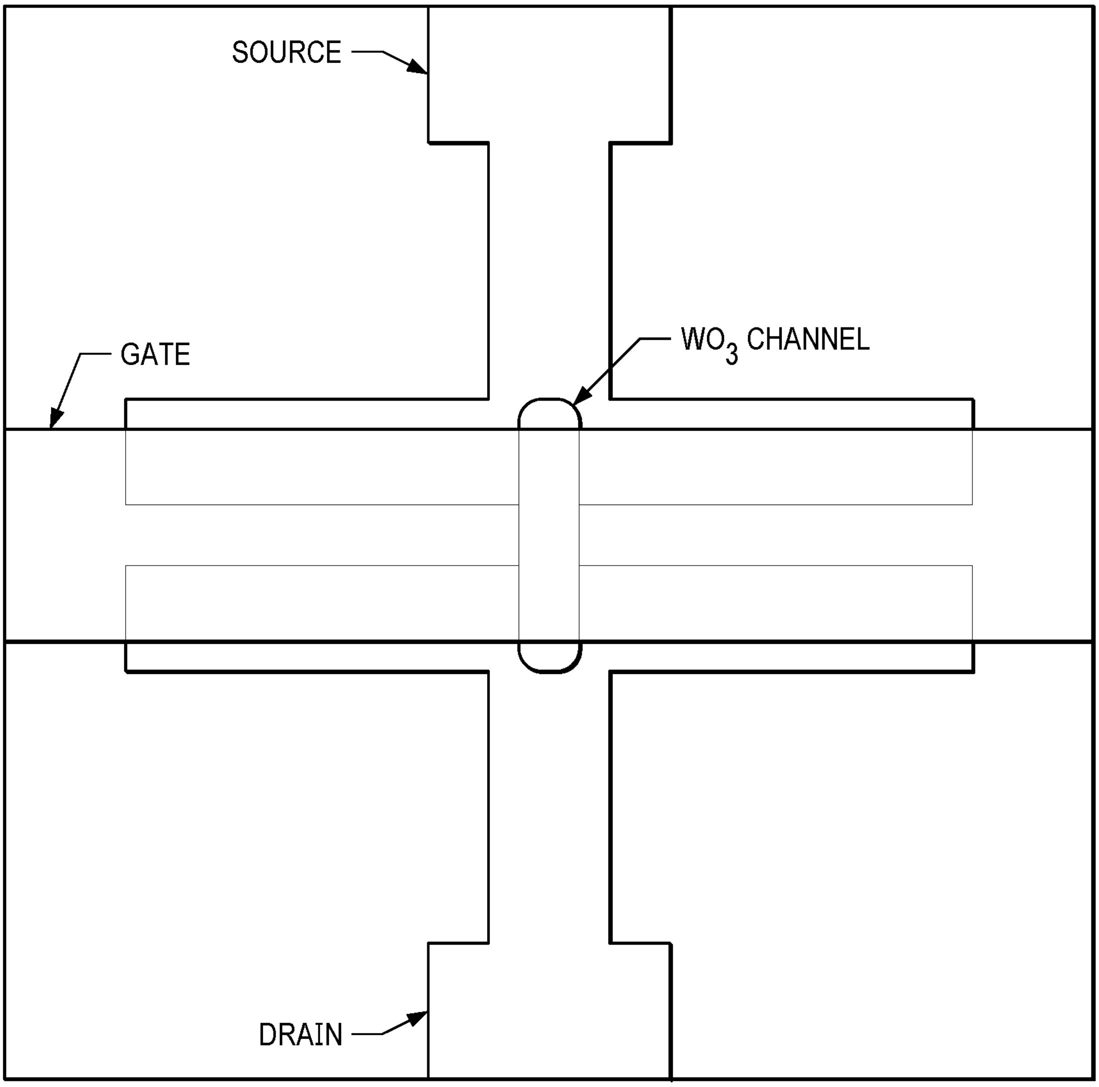


FIG. 21

FIG. 22

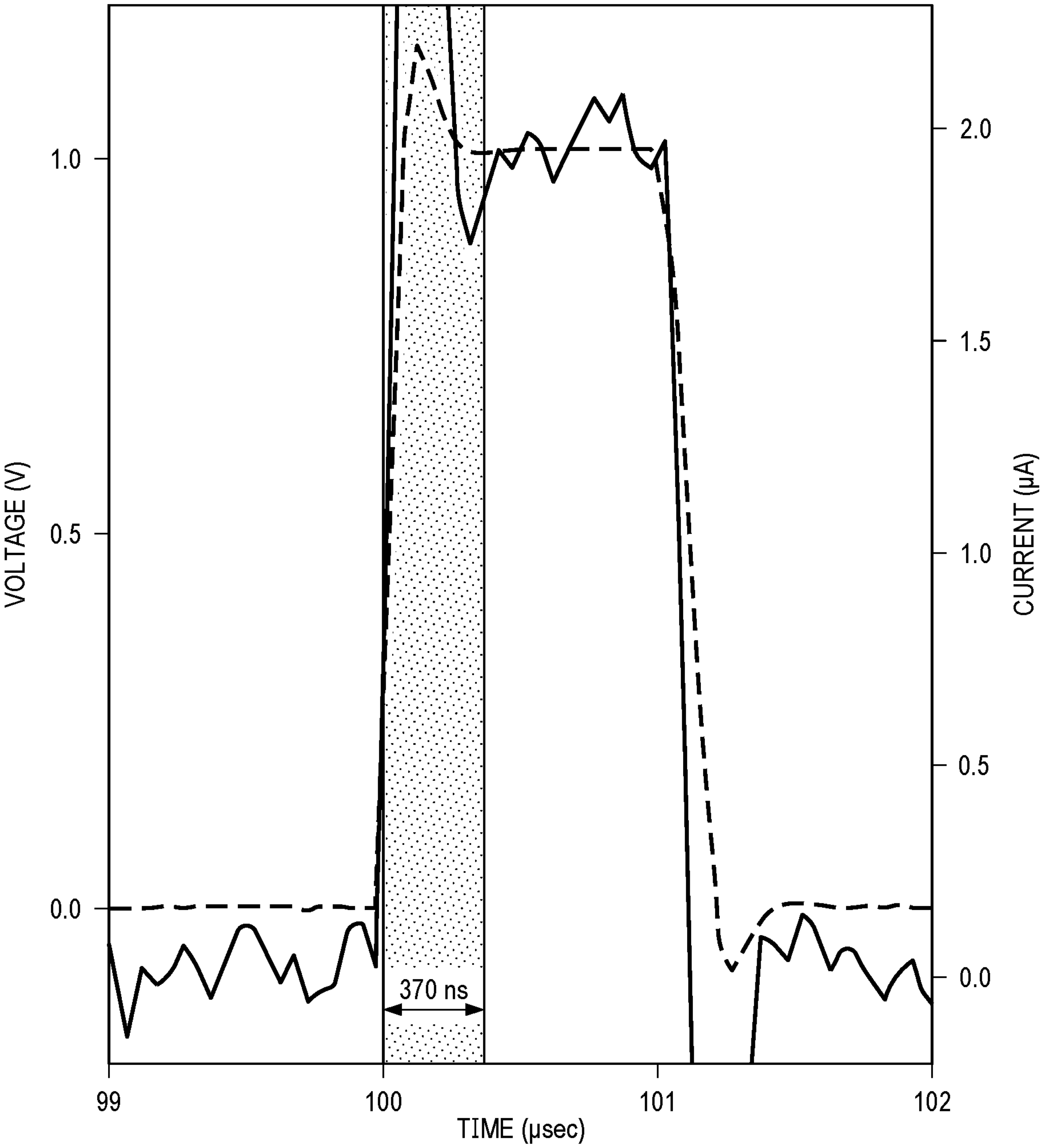

VOLTAGE (V)
1.0
0.5
0.0
CURRENT (µA)
2.0
1.5
1.0
0.5
0.0
370 ns
99
100
101
102
TIME (µsec)

FIG. 23

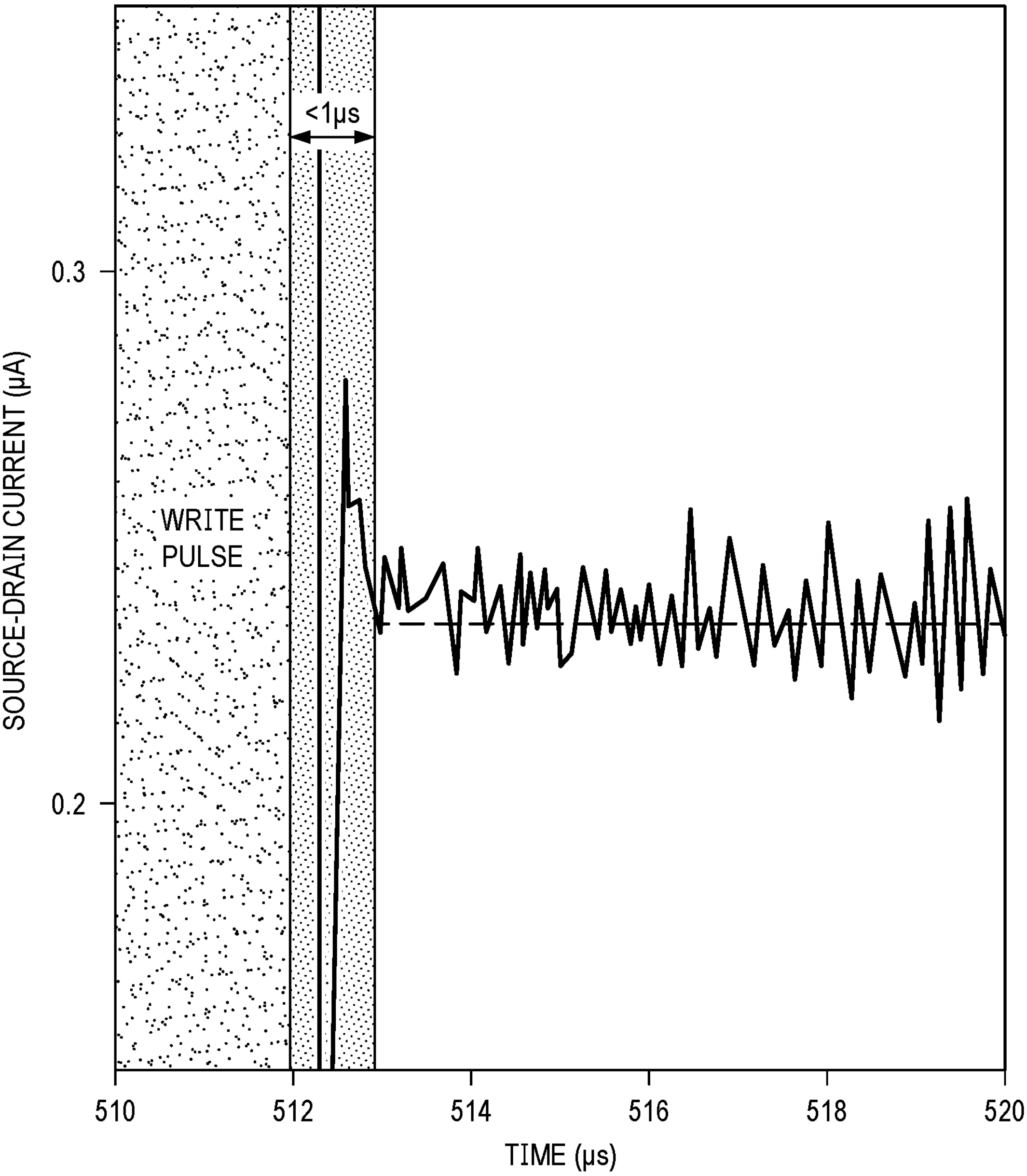
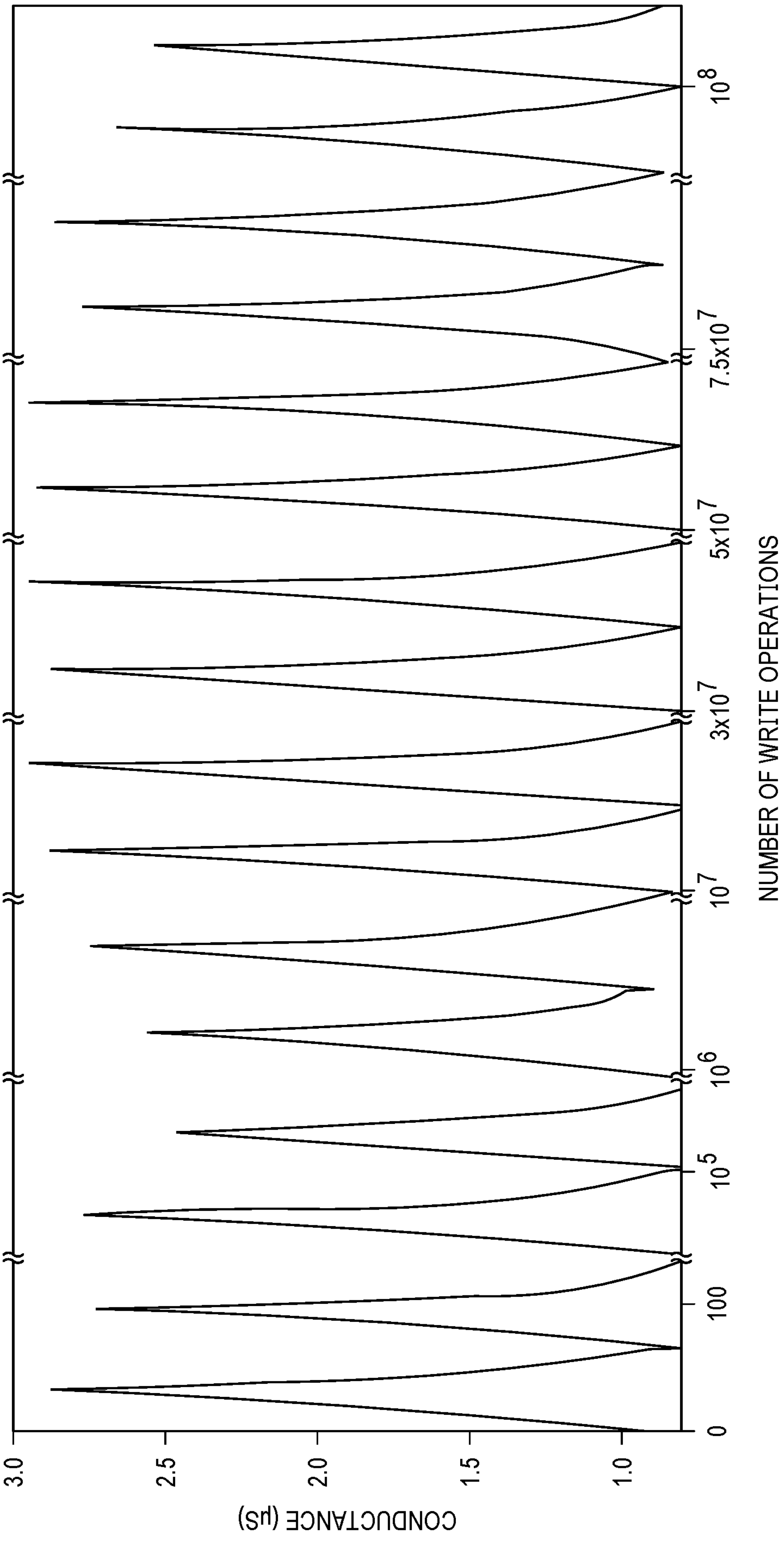

WRITE PULSE
<1μs
SOURCE-DRAIN CURRENT (μA)
0.3
0.2
510
512
514
516
518
520
TIME (μs)

FIG. 24


CONDUCTANCE (μS)
3.0
2.5
2.0
1.5
1.0
0
100
10^5
10^6
10^7
3x10^7
5x10^7
7.5x10^7
10^8
NUMBER OF WRITE OPERATIONS

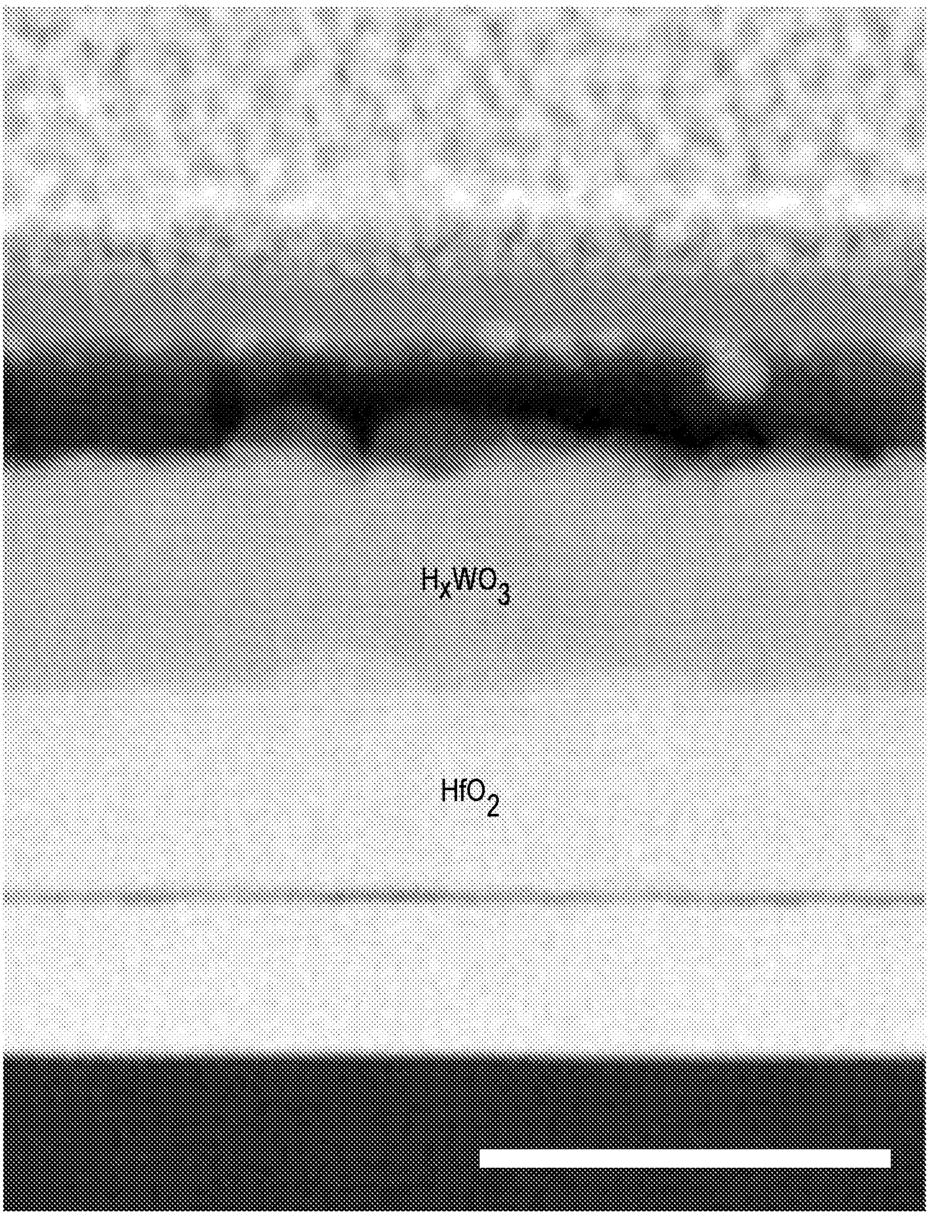


FIG. 25A

FIG. 25B

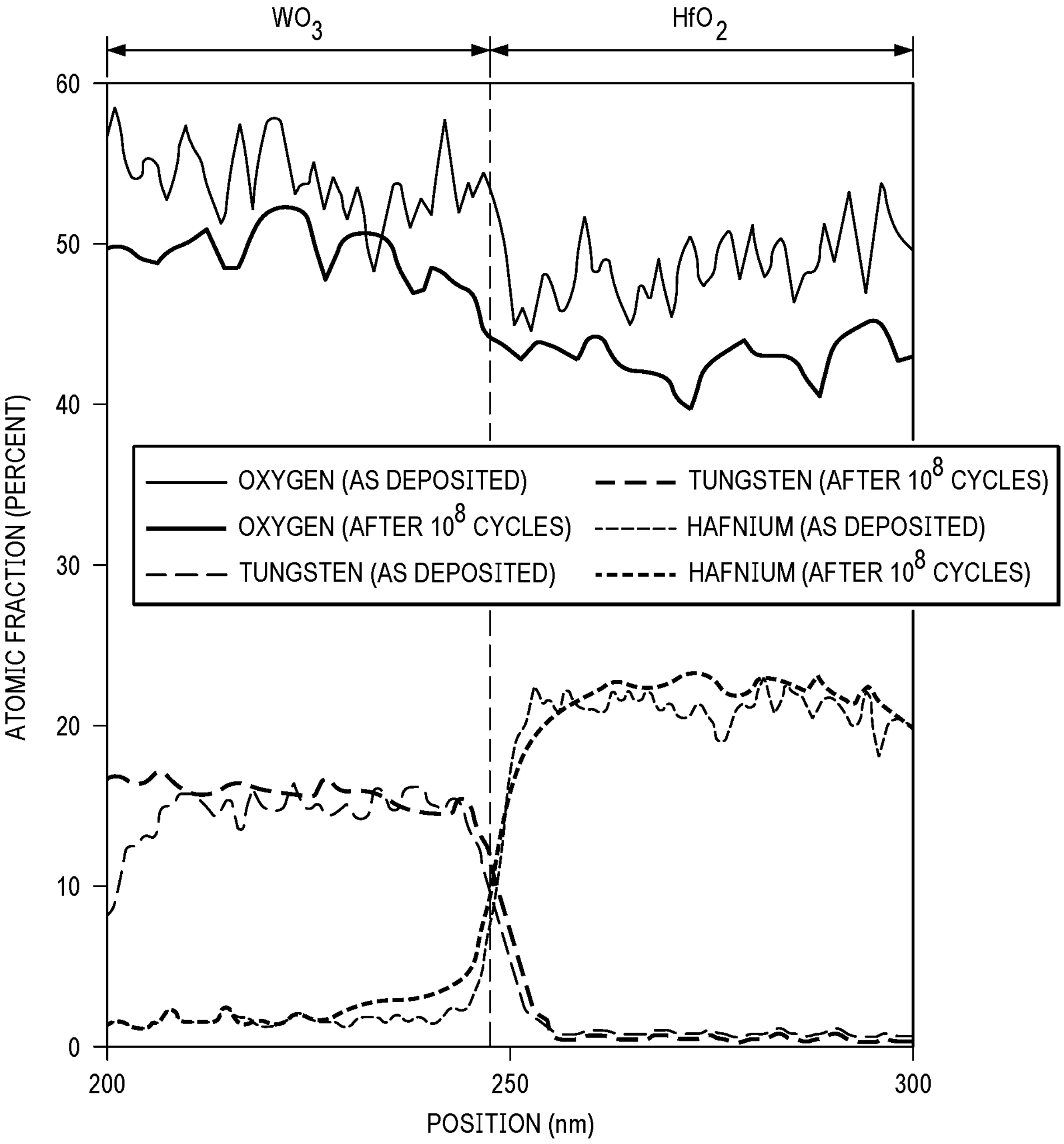

WO_3
HfO_2
ATOMIC FRACTION (PERCENT)
60
50
40
30
20
10
0
200
250
300
POSITION (nm)
OXYGEN (AS DEPOSITED)
OXYGEN (AFTER 10^8 CYCLES)
TUNGSTEN (AS DEPOSITED)
TUNGSTEN (AFTER 10^8 CYCLES)
HAFNIUM (AS DEPOSITED)
HAFNIUM (AFTER 10^8 CYCLES)

FIG. 26

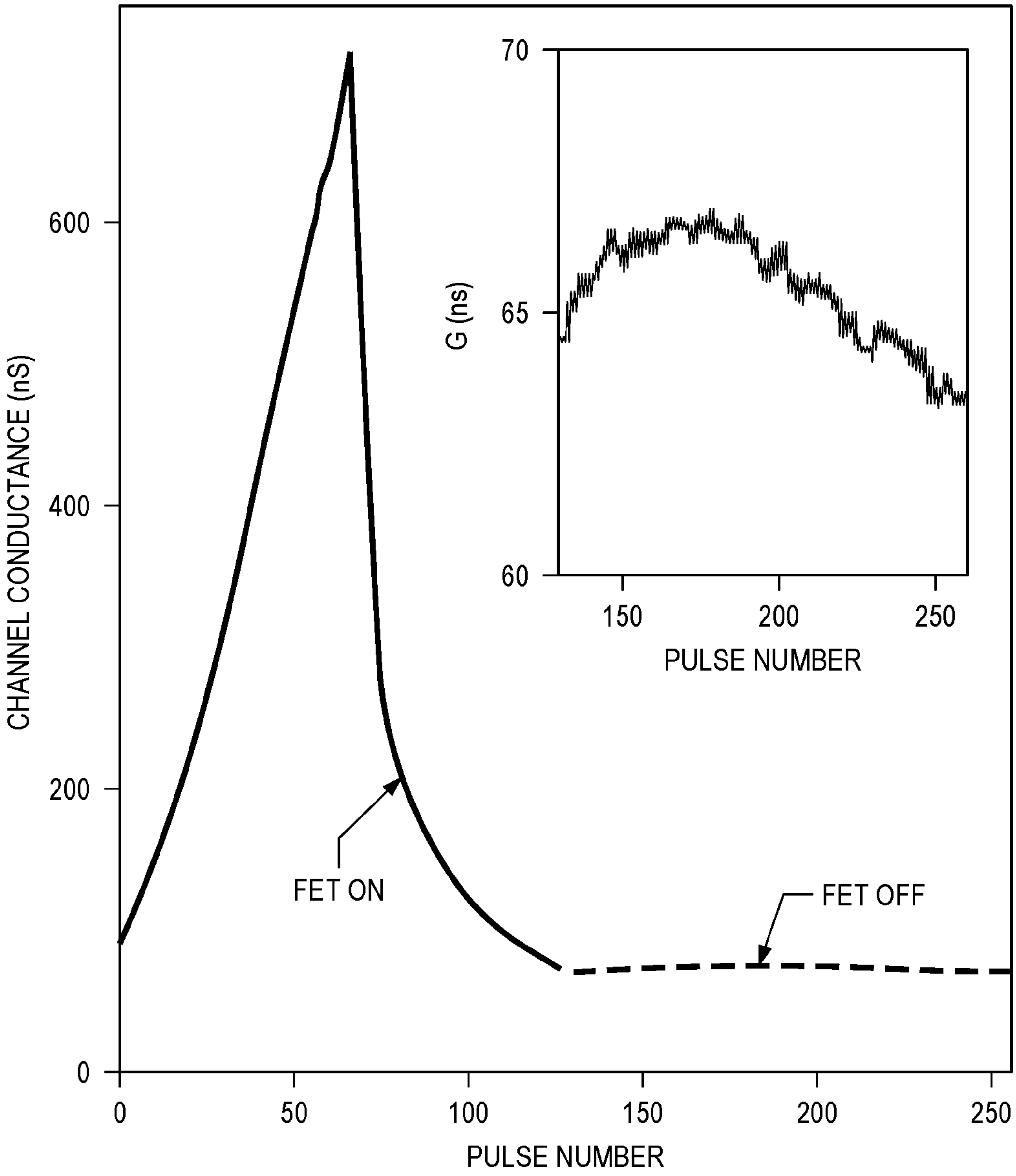

CHANNEL CONDUCTANCE (nS)
600
400
200
0
0
50
100
150
200
250
PULSE NUMBER
FET ON
FET OFF
G (ns)
70
65
60
150
200
250
PULSE NUMBER

FIG. 27A

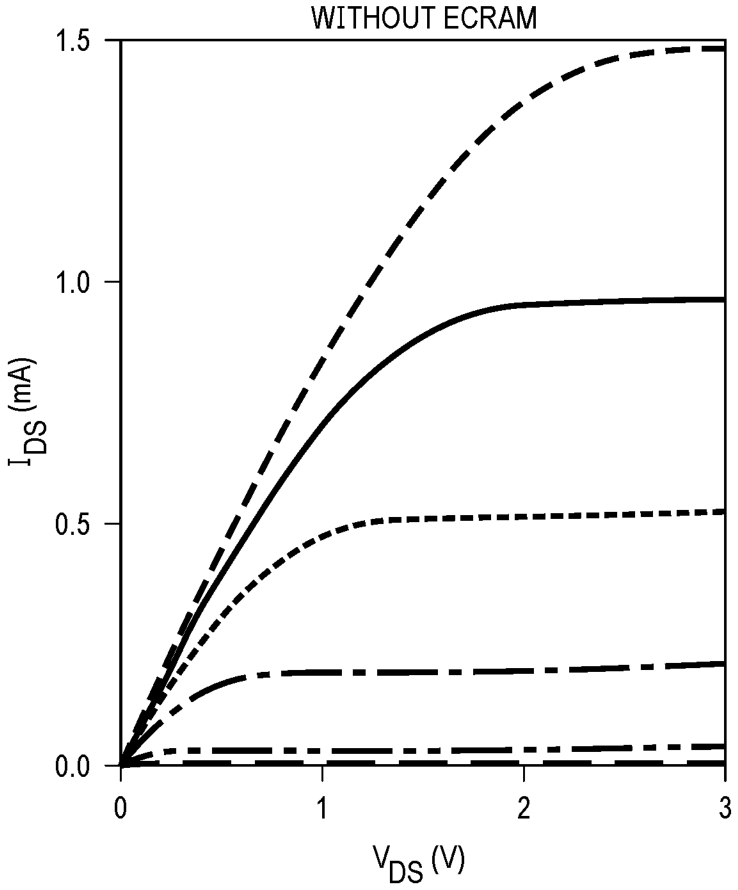
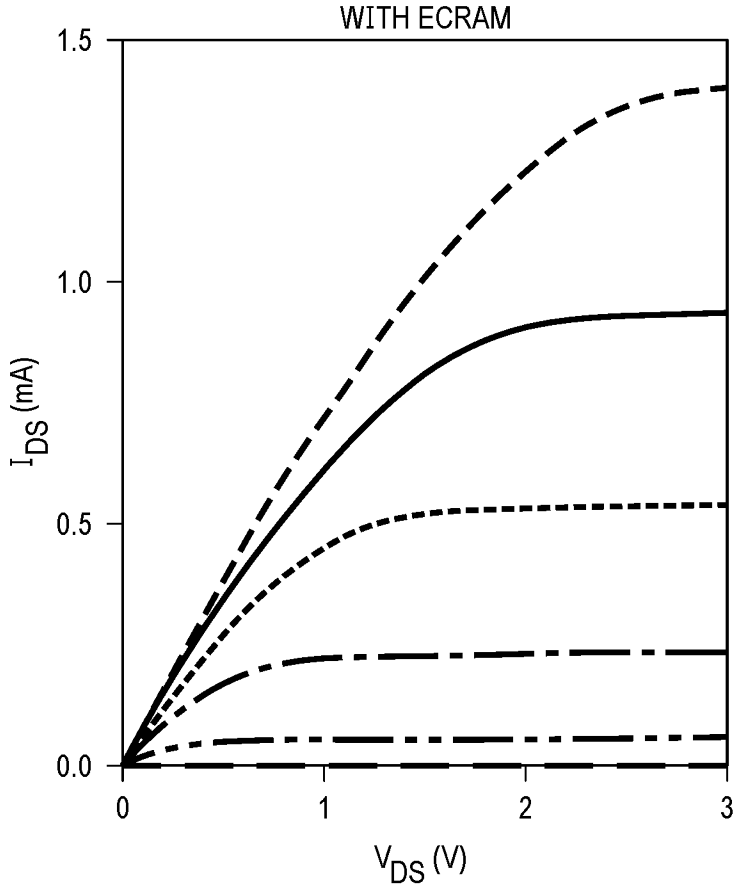


FIG. 27B



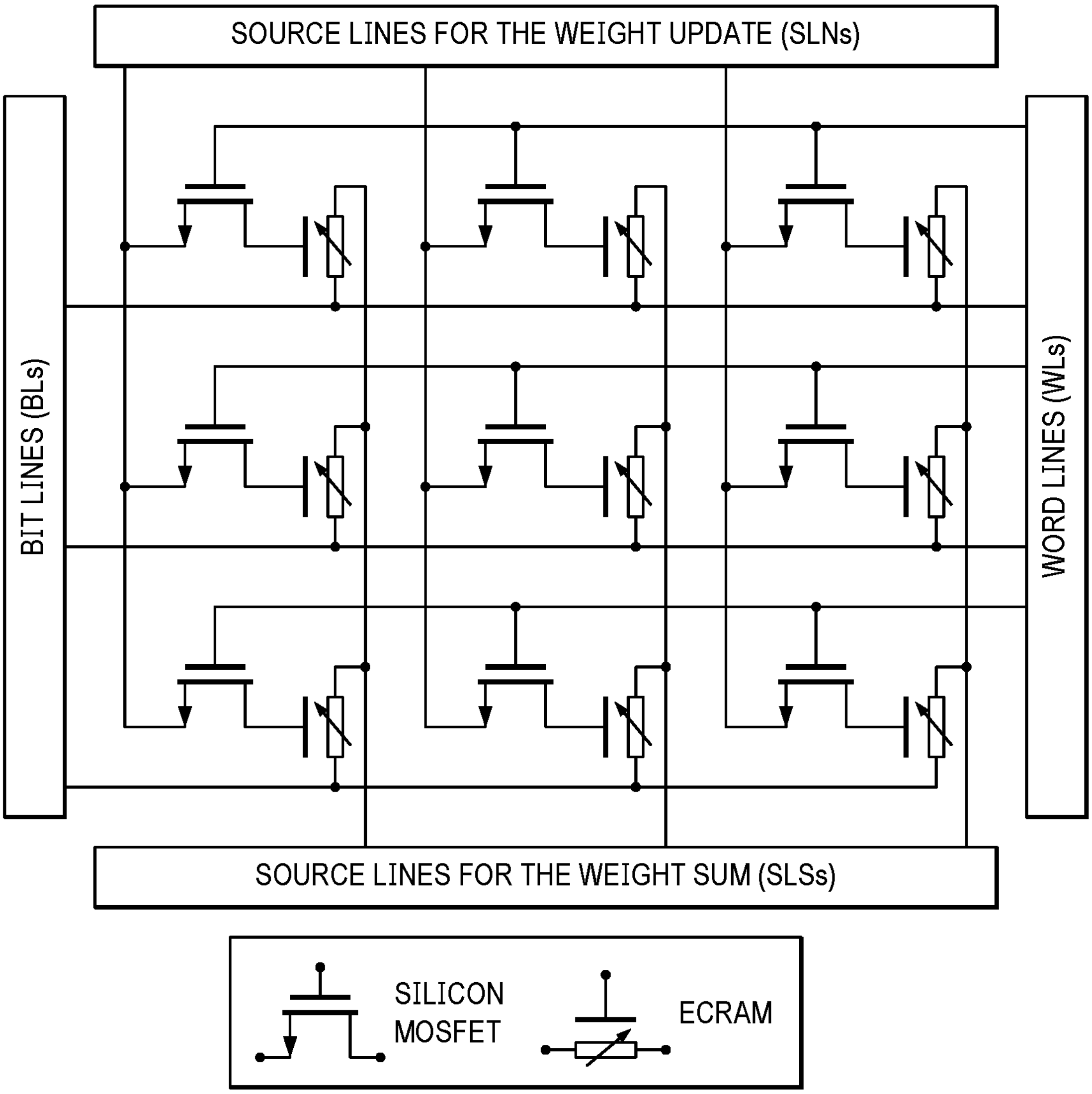

SOURCE LINES FOR THE WEIGHT UPDATE (SLNs)
BIT LINES (BLs)
WORD LINES (WLs)
SOURCE LINES FOR THE WEIGHT SUM (SLSs)
SILICON MOSFET
ECRAM

FIG. 28

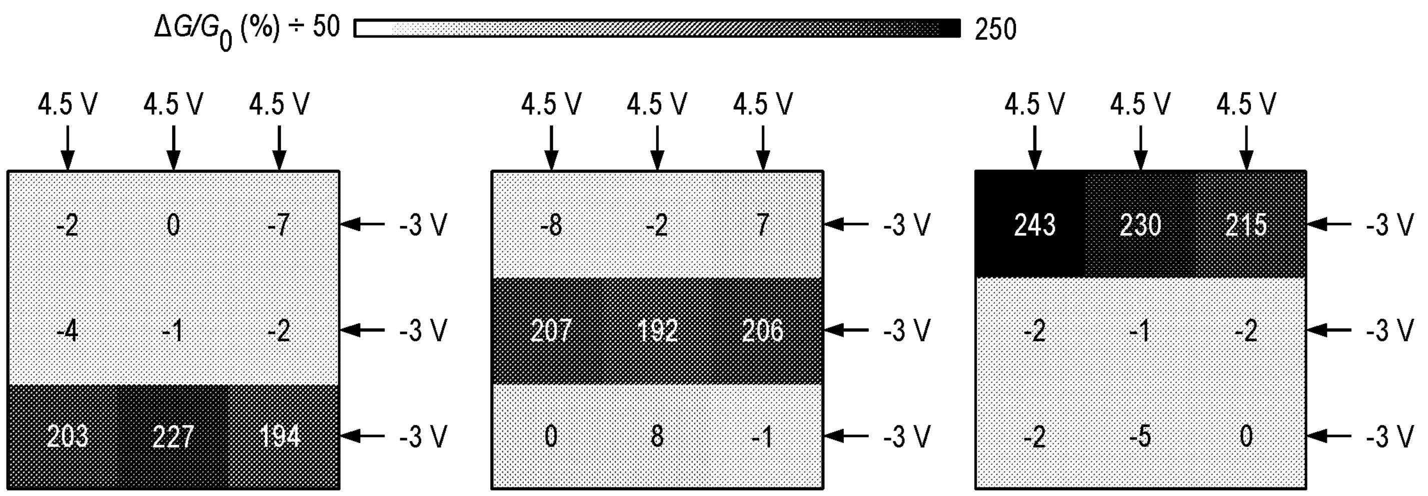

ΔG/G0 (%) ÷ 50
250
4.5 V
4.5 V
4.5 V
-3 V
-3 V
-3 V
-2
0
-7
-4
-1
-2
203
227
194
-8
-2
7
207
192
206
0
8
-1
243
230
215
-2
-1
-2
-2
-5
0

FIG. 29

FIG. 30A

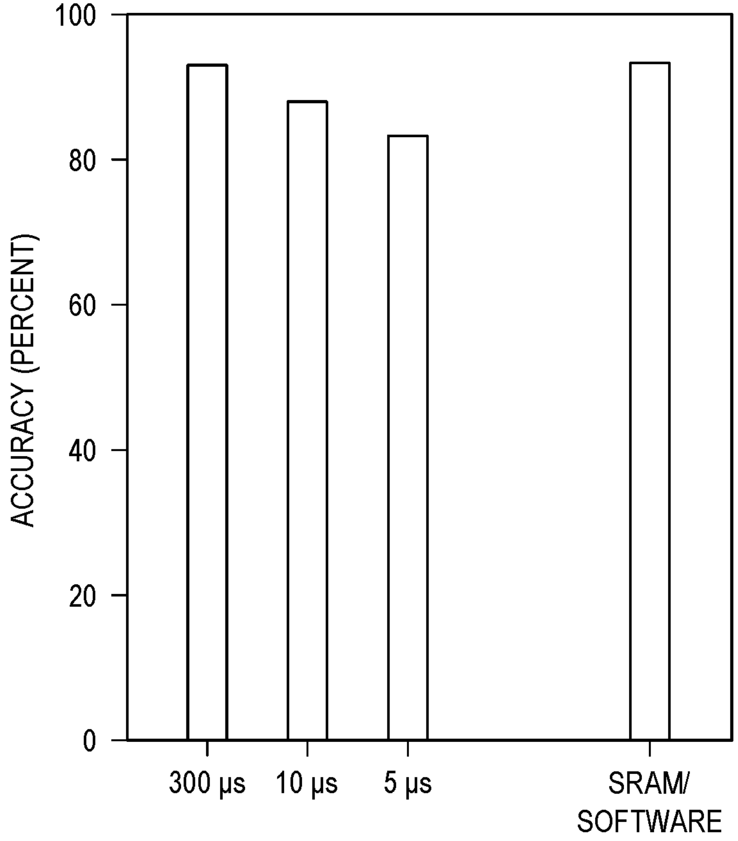


FIG. 30B

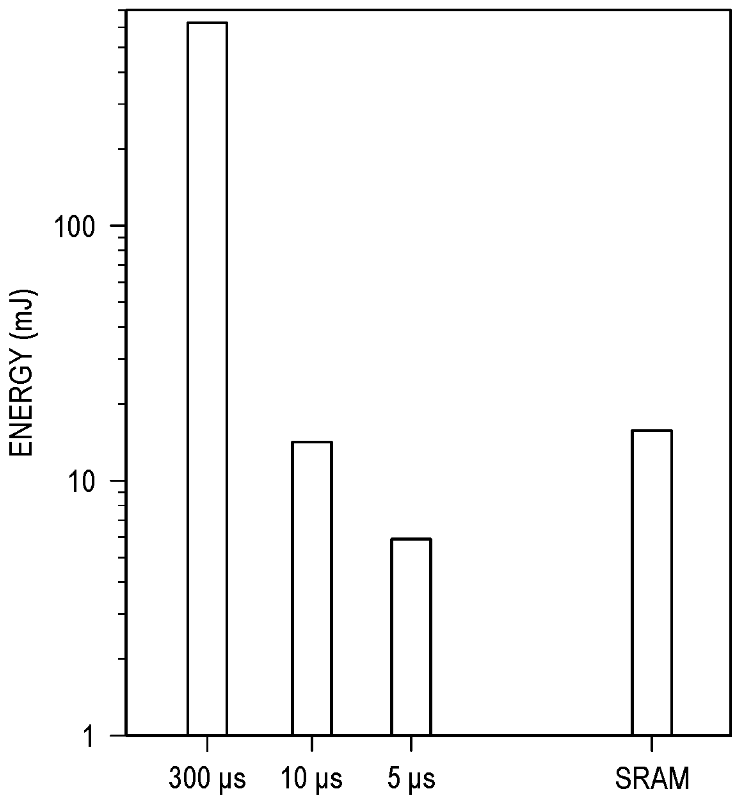


FIG. 31

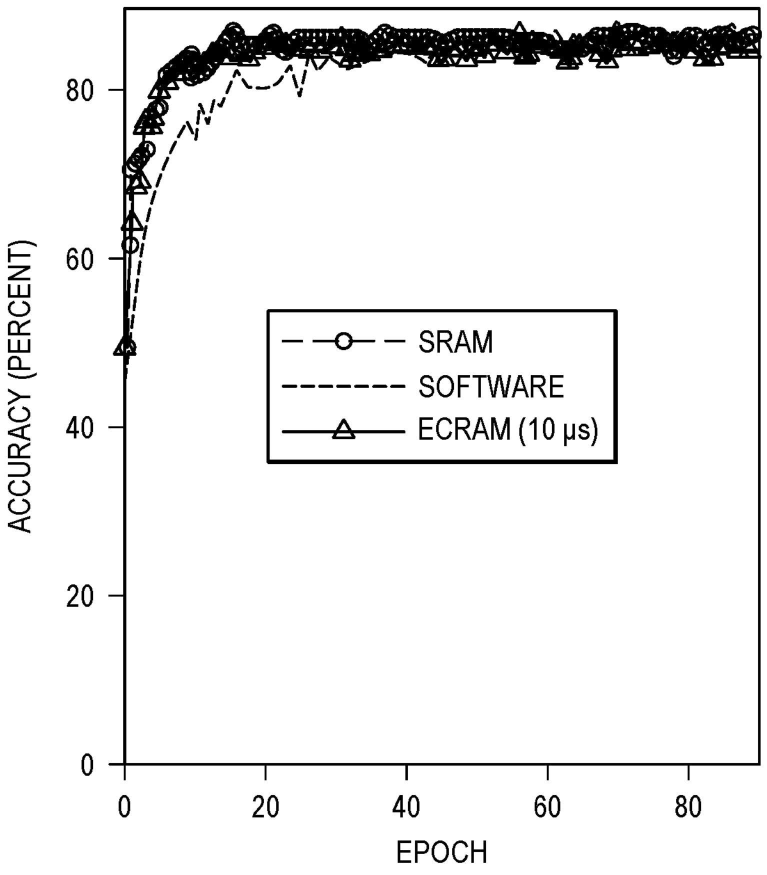

ACCURACY (PERCENT)
0
20
40
60
80
EPOCH
0
20
40
60
80
SRAM
SOFTWARE
ECRAM (10 μs)

FIG. 32

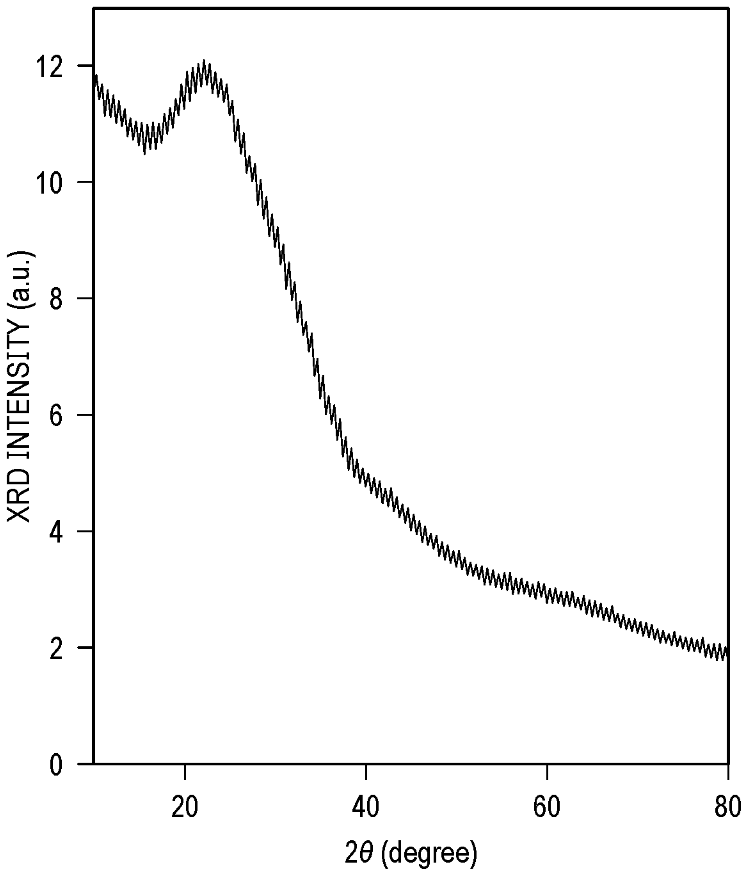

XRD INTENSITY (a.u.)
0
2
4
6
8
10
12
2θ (degree)
20
40
60
80

FIG. 33A

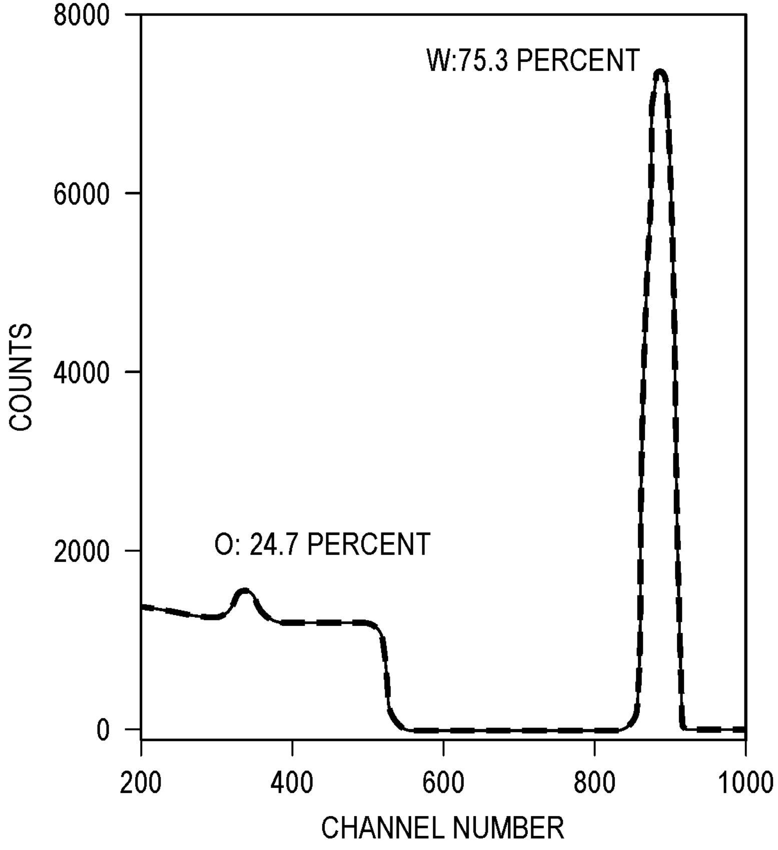


FIG. 33B

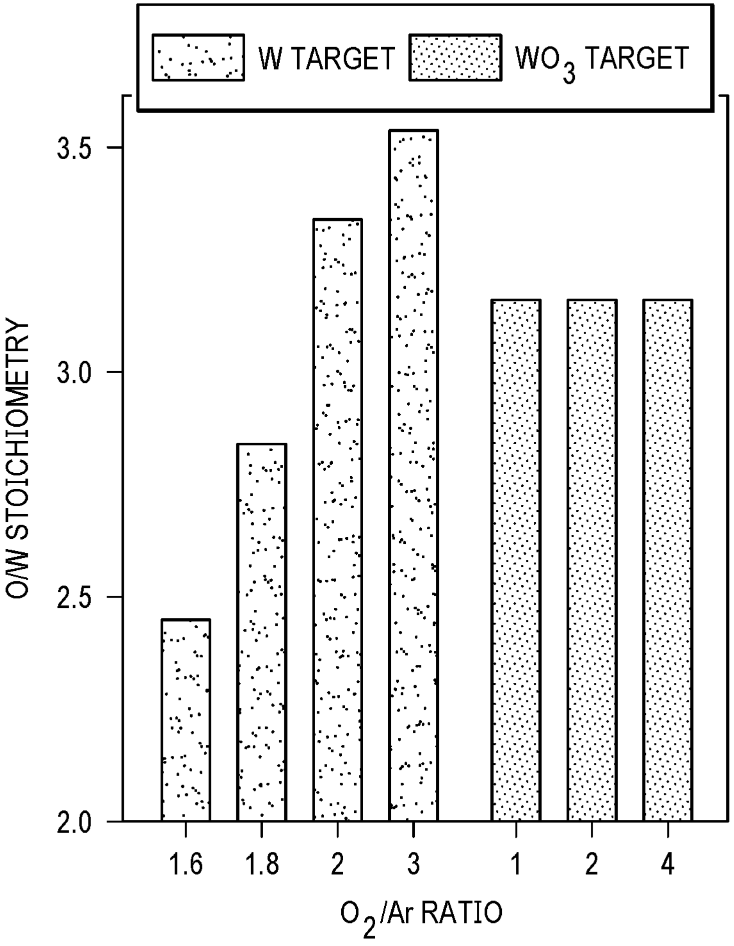


FIG. 34A

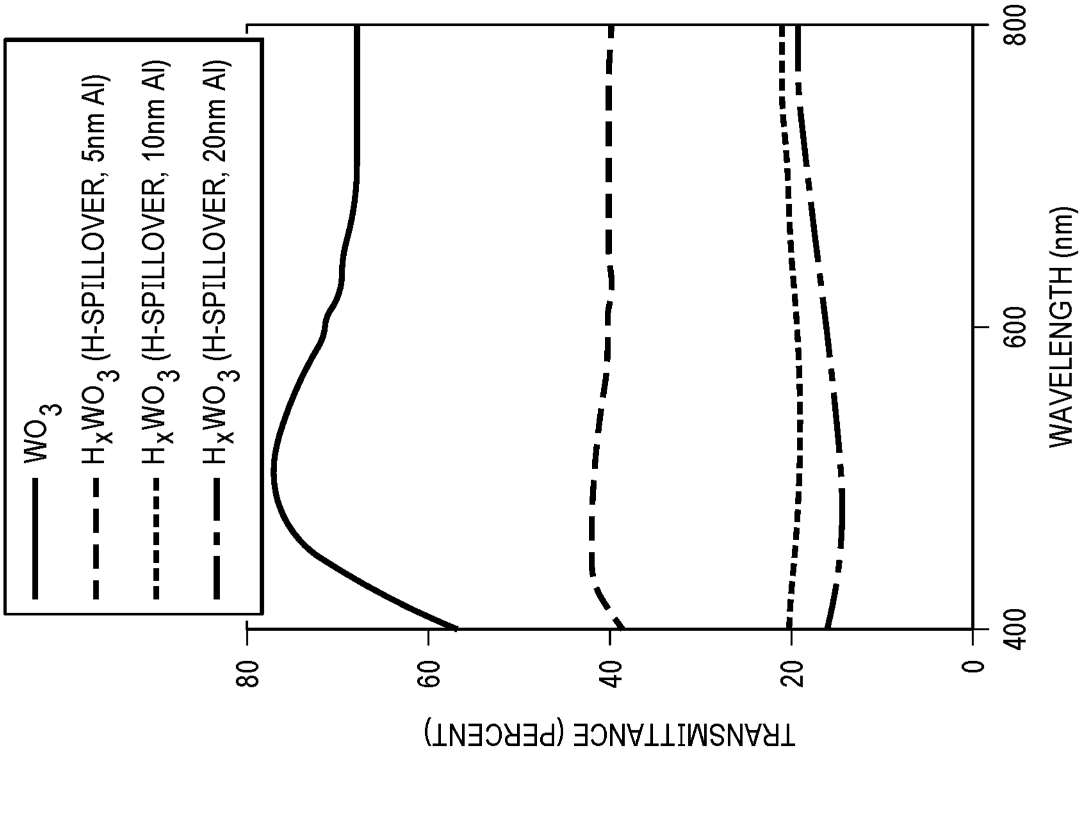
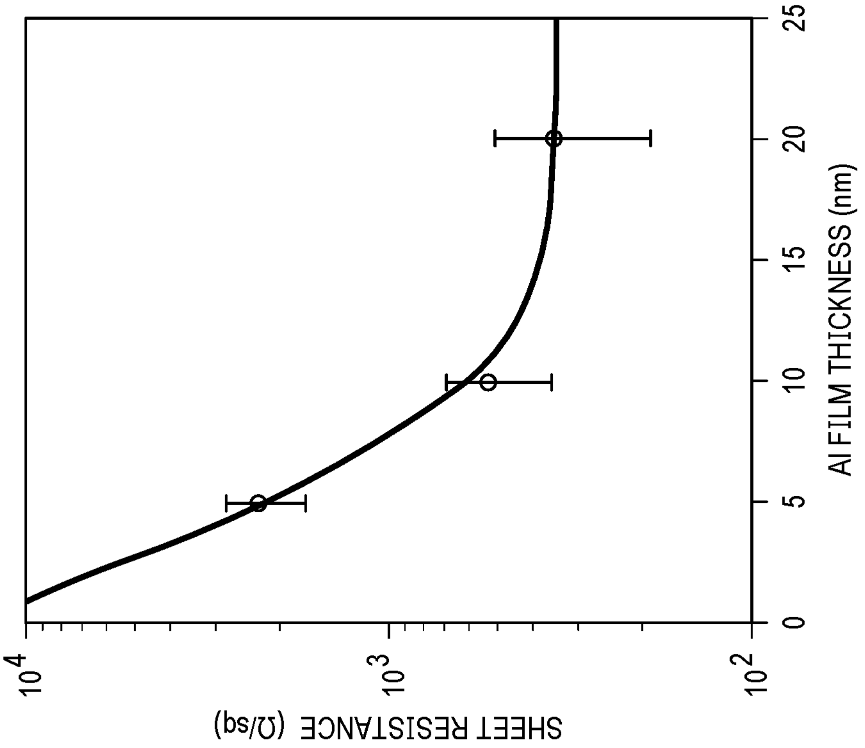


FIG. 34B



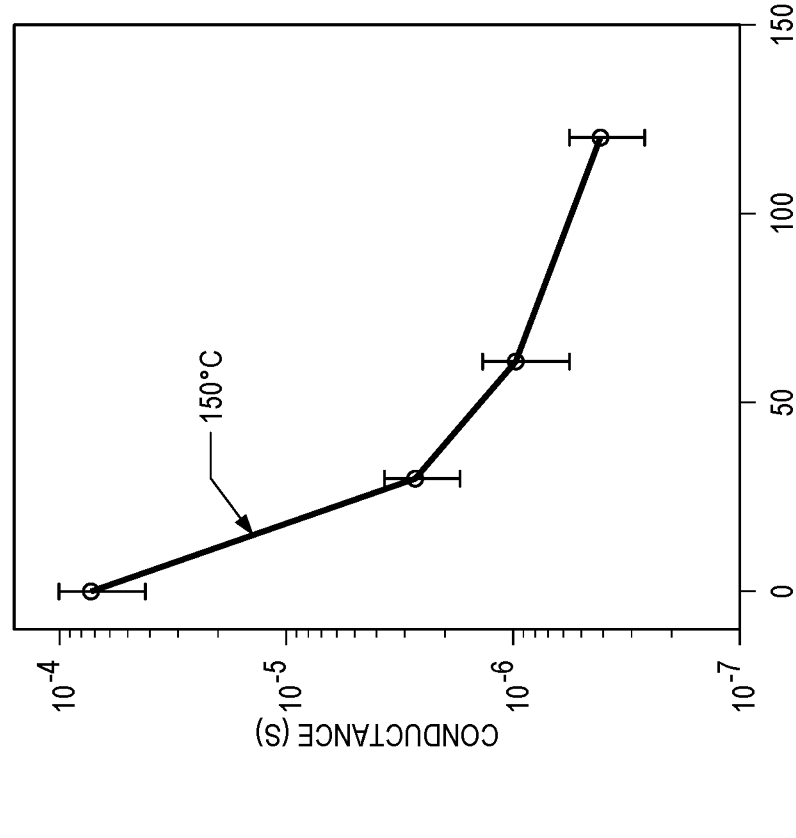

CONDUCTANCE (S)
10^{-4}
10^{-5}
10^{-6}
10^{-7}
150°C
0
50
100
150
TIME (sec)

FIG. 35

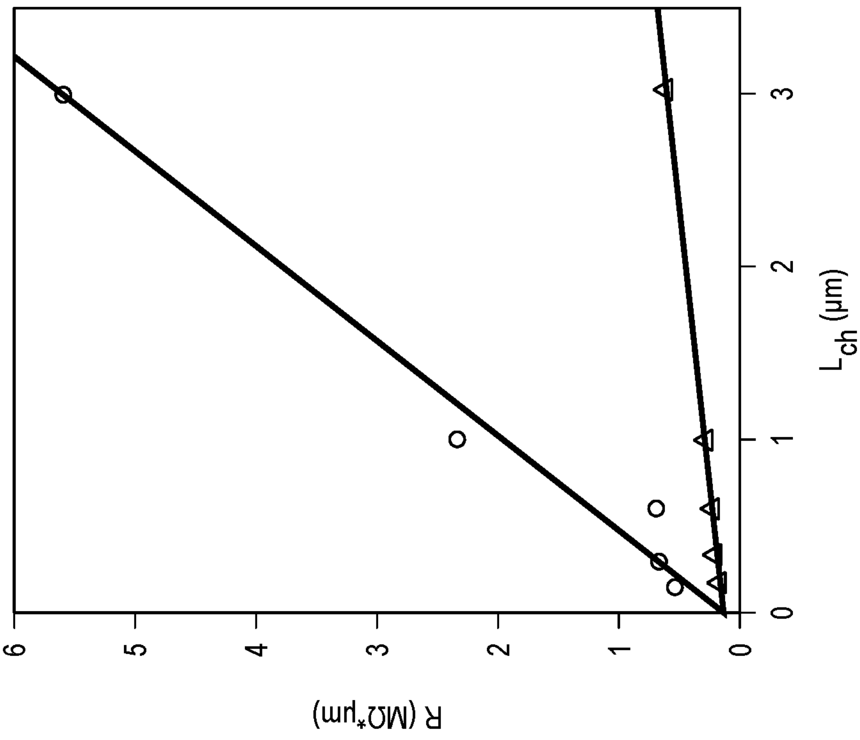

R (MΩ*μm)
6
5
4
3
2
1
0
0
1
2
3
L_{ch} (μm)

FIG. 36

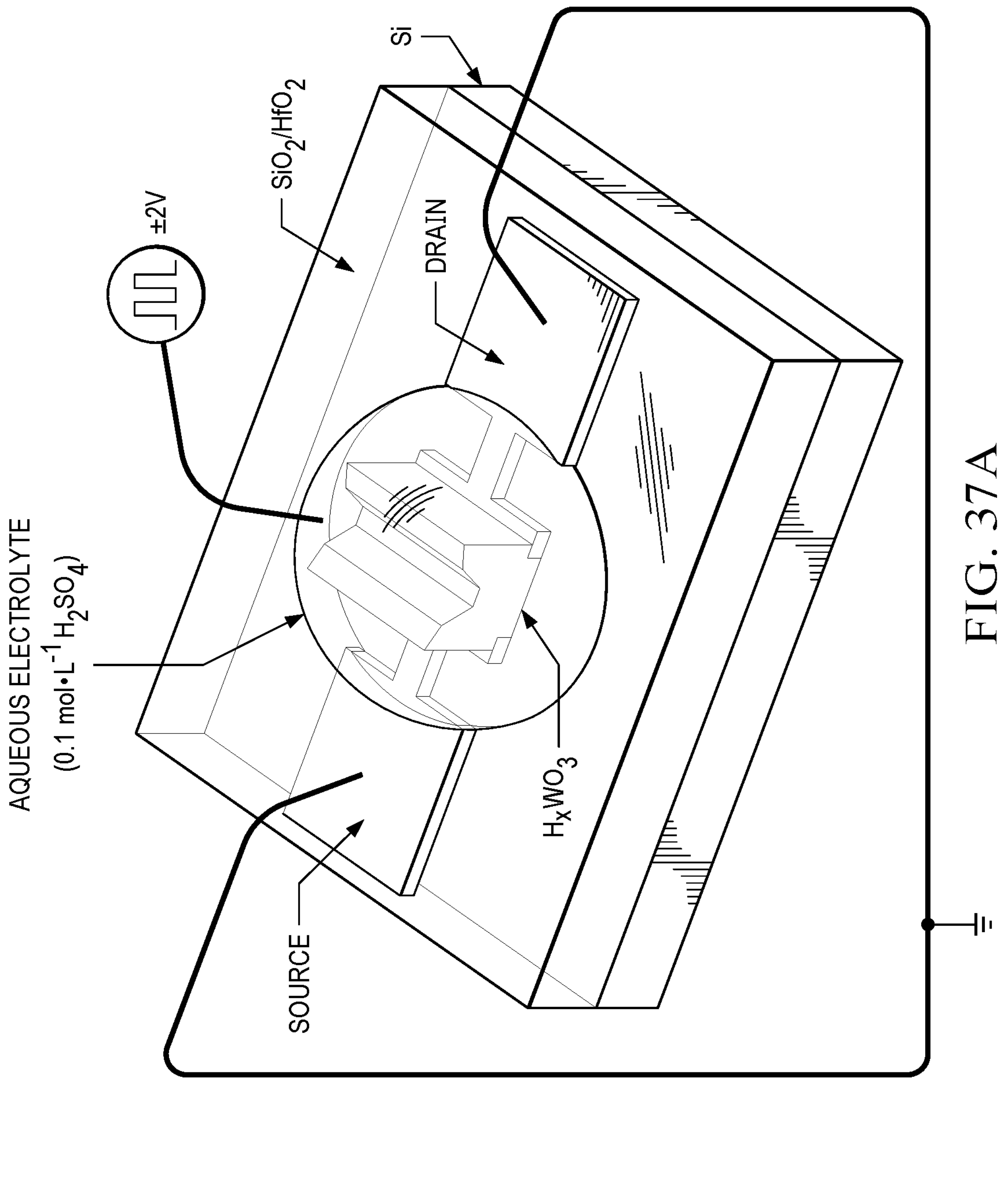

AQUEOUS ELECTROLYTE
(0.1 mol•L^{-1} H_2SO_4)
±2V
SiO_2/HfO_2
Si
DRAIN
SOURCE
H_xWO_3

FIG. 37A

FIG. 37B

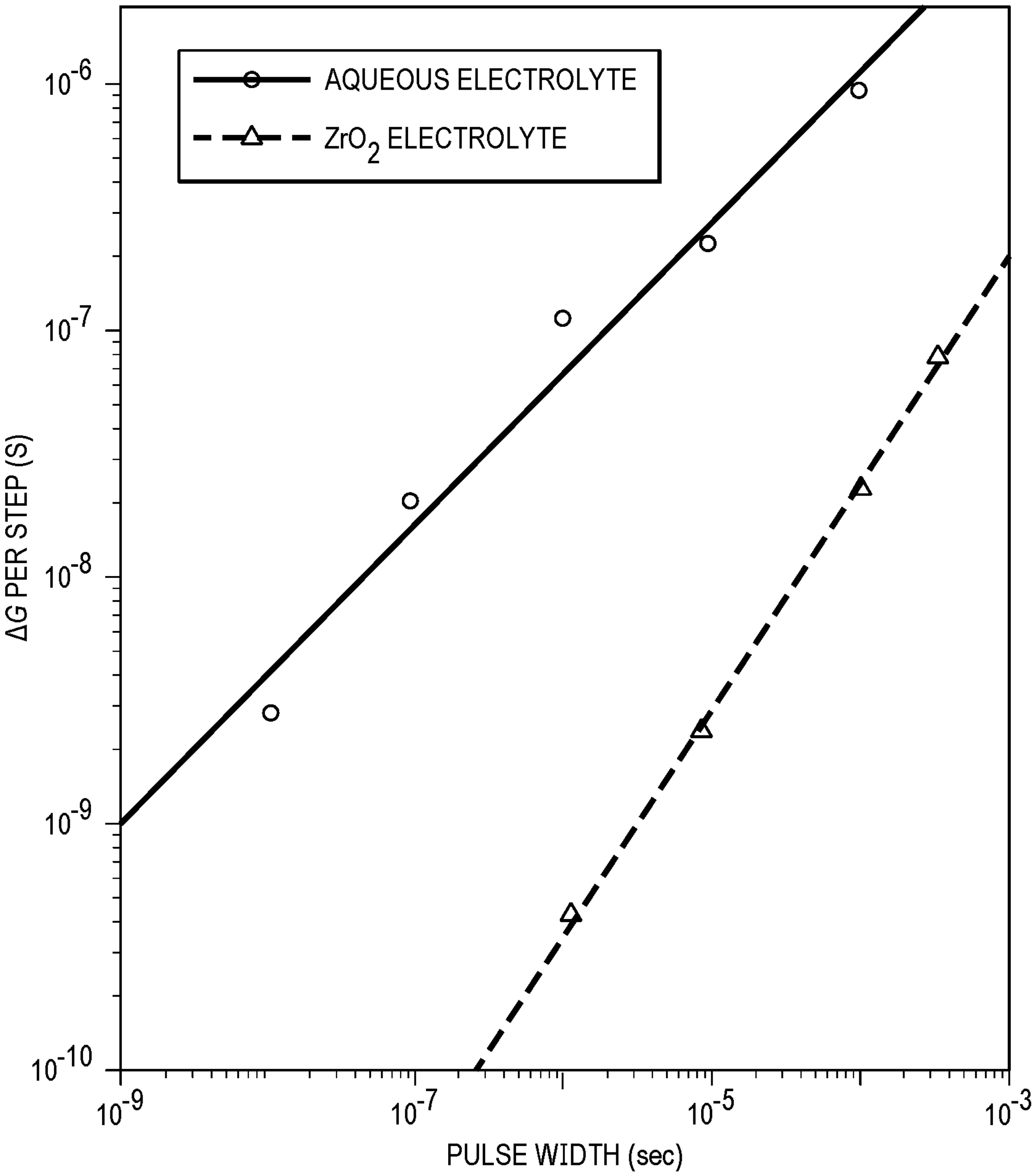

AQUEOUS ELECTROLYTE
ZrO_2 ELECTROLYTE
ΔG PER STEP (S)
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-9}
10^{-7}
10^{-5}
10^{-3}
PULSE WIDTH (sec)

FIG. 38A

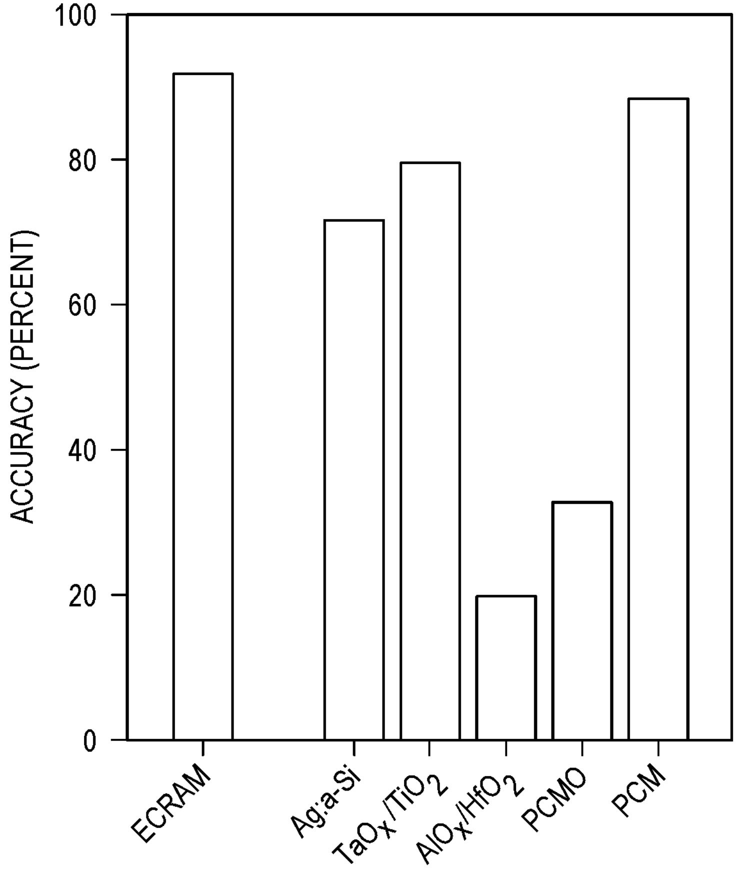


FIG. 38B

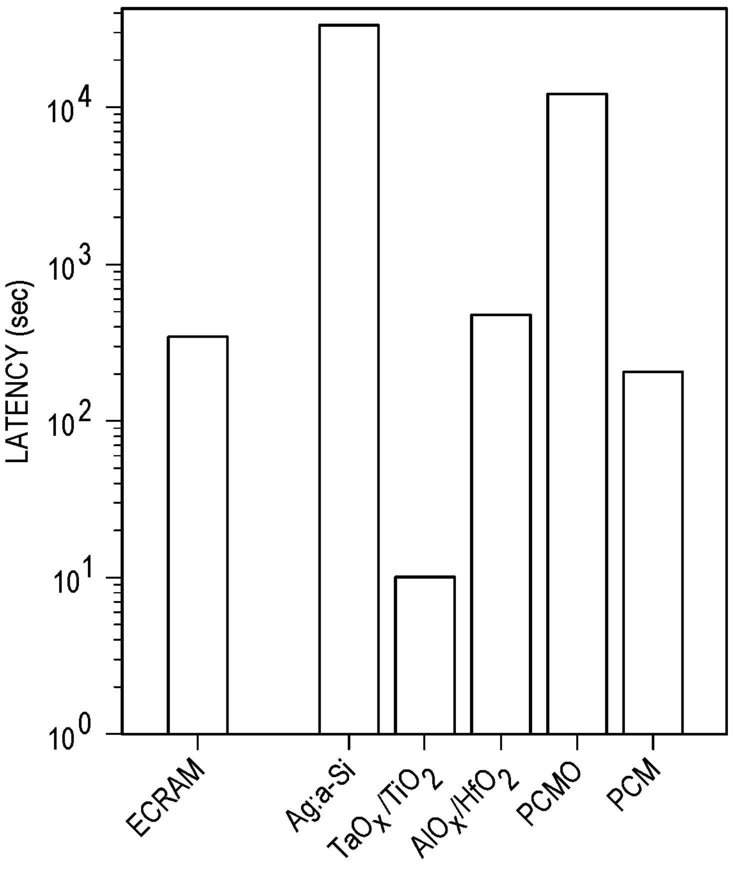


FIG. 38C

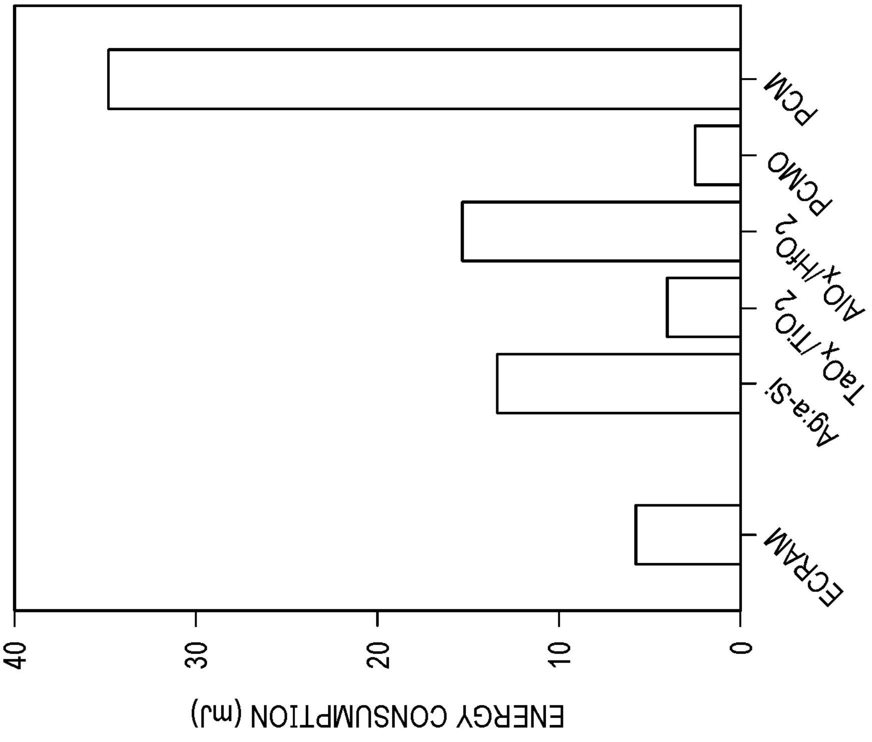
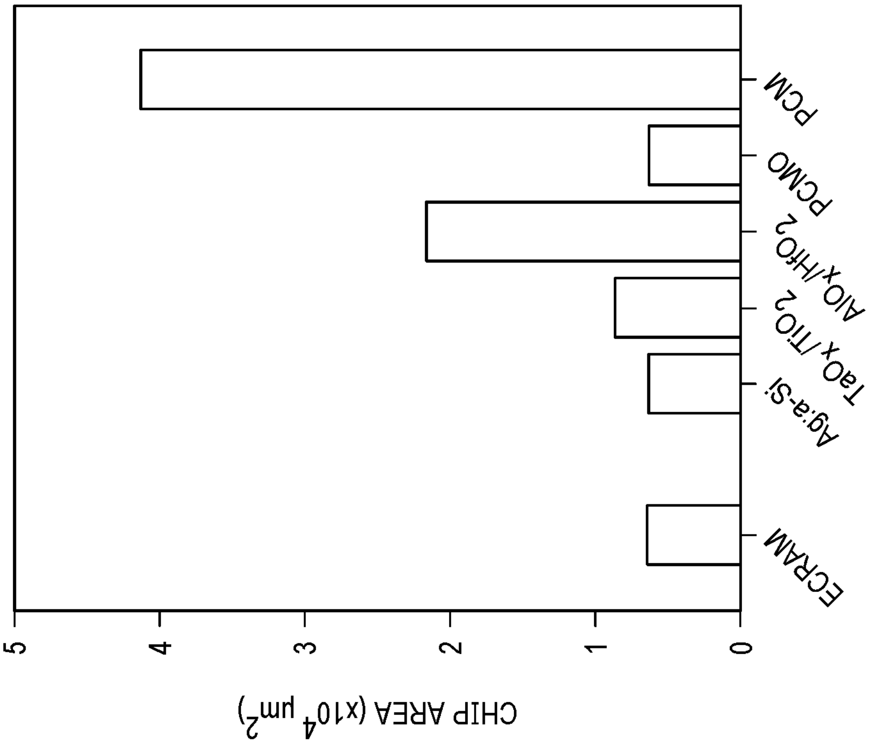


FIG. 38D


CHIP AREA (x10^4 μm^2)
ECRAM
Ag:a-Si
TaO$_x$/TiO$_2$
AlO$_x$/HfO$_2$
PCMO
PCM