Simulations of Ultralow-Power Nonvolatile Cells for Random-Access Memory

Dominic Lane and Manus Hayne

***Abstract*—Dynamic random-access memory (DRAM), which represents 99% of random-access memory (RAM), is fast and has excellent endurance, but suffers from disadvantages such as short data-retention time (volatility) and loss of data during readout (destructive read). As a consequence, it requires persistent data refreshing, increasing energy consumption, degrading performance, and limiting scaling capacity. It is, therefore, desirable that the next generation of RAM will be nonvolatile RAM (NVRAM), have low power, have high endurance, be fast, and be nondestructively read. Here, we report on a new form of NVRAM: a compound-semiconductor charge-storage memory that exploits quantum phenomena for its operational advantages. Simulations show that the device consumes very little power, with 100 times lower switching energy per unit area than DRAM, but with similar operating speeds. Nonvolatility is achieved due to the extraordinary band offsets of InAs and AlSb, providing a large energy barrier (2.1 eV), which prevents the escape of electrons. Based on the simulation results, an NVRAM architecture is proposed for which extremely low disturb-rates are predicted as a result of the quantum-mechanical resonant-tunneling mechanism used to write and erase.**

***Index Terms*—InAs/AlSb, memory, nonvolatile memory (NVM), nonvolatile RAM (NVRAM), resonant tunneling (RT).**

I. Introduction

PRODUCTION and sales of electronic memories are dominated by dynamic random-access memory (DRAM) and Flash. DRAM is the workhorse of active memory in current electronics. It is fast and cheap to produce and has very high endurance. However, it also has some inconvenient properties, notably volatility and destructive read. As a result, persistent data refreshing is required, negatively affecting the bandwidth, scaling capacity, and energy consumption of the memory [1]. Consequently, the search for alternative memory concepts with all the advantages of DRAM and none of the disadvantages, sometimes called "universal memory," continues. Universal memory cells should be nonvolatile, have low voltage, have low energy, should be nondestructively read, should be cheap, should be fast, and have high endurance, providing a universal solution for all memory requirements. Implementing such a memory as a nonvolatile RAM (NVRAM), for example, would produce a paradigm shift in computing. However, a seemingly insurmountable stumbling block comprises the apparently contradictory requirements of nonvolatility, which necessitates a very robust programmed state, and fast low-voltage (low-energy) write and erase, which implies a state that can be readily changed. This has led to the view that the universal memory concept is not realistic [2].

Here, we report on a novel memory [3] that exploits the quantum properties of a triple-barrier resonant tunneling (RT) structure to allow the contradictory combination of nonvolatility with low-voltage write and erase. Due to the large (2.1 eV) barrier, the intrinsic (thermal excitation) electron storage time of our InAs/AlSb system was predicted [4] to exceed substantially the age of the Universe. Clearly, in real devices, the presence of other loss mechanisms will lower the actual storage time dramatically. Nevertheless, the barrier of 2.1 eV exceeds that of NAND Flash (1.6 eV), so such devices are expected to be nonvolatile, and this has been demonstrated in recent work [9]. Despite this, write and erase require $\leq$2.3 V. The simulation results detailed here are from a specially developed, room-temperature model implemented using a combination of commercial software. The nextnano multi-scattering Büttiker (MSB) software [5], [6] was used to investigate the transport of carriers through the RT structure (write and erase), nextnano++ to model the channel (read), and Simulation Program with Integrated Circuit Emphasis (SPICE) [7] to determine the corresponding overall device and circuit-level properties. The simulation parameters used to model the device physics are provided in Table I and are fixed to experimentally observed constants [6], [8]. The chosen structure of the device is based on very recently reported memory cells operating at low voltages at room temperature [9]. In these devices, the read process used a depletion mode channel that is "normally ON," i.e., is conducting at zero gate bias. However, this inhibits its implementation in a RAM, as devices in the array that are not being addressed cannot be switched OFF. Here, to overcome this obstacle, the thickness of the channel used for the read cycle is reduced to form a quantum well (QW), exploiting quantum confinement to create a channel with a threshold voltage for conductivity to read the device. This structural adaptation produces the "normally OFF" channel that is required for an operational floating gate (FG) RAM. Combining the results of the RT simulations and QW channel (QW_{CH}) simulations into a SPICE program

Manuscript received October 4, 2019; accepted November 27, 2019. Date of publication January 1, 2020; date of current version January 27, 2020. This work was supported by the Joy Welch Educational Charitable Trust. The work of D. Lane was supported by the Engineering and Physical Sciences Research Council (EPSRC) for providing a Scholarship under Grant EP/N509504/1. The review of this article was arranged by Editor T.-H. Kim. *(Corresponding author: Dominic Lane.)*

The authors are with the Department of Physics, Lancaster University, Lancaster LA1 4YB, U.K. (e-mail: d.lane@lancaster.ac.uk; m.hayne@lancaster.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TED.2019.2957037

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

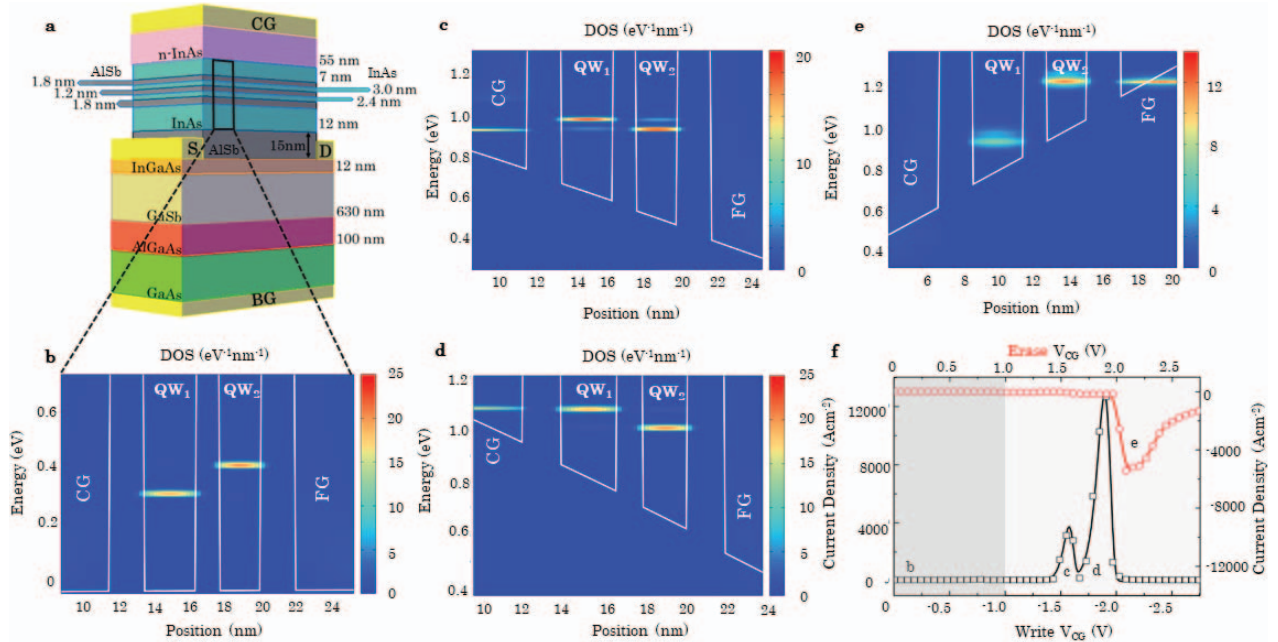

Fig. 1. Simulation results (300 K) for the tunneling region of the device. The model used is strictly 1-D. (a) Schematic of a potential device structure. Device includes CG, back gate (BG), source (S), and drain (D) contacts. (b)–(e) QW energy levels for the structure are shown, where the color scale indicates the electron DOS. No states are shown in the collector, which is interpreted as supplying a current in the software as electrons tunnel through the barriers. All voltages mentioned will be applied to the device terminals, as the 15-nm AlSb blocking barrier has been accounted for in the nextnano++ modeling of the band structure under applied biases. (b) 0-V bias (store). (c) −1.6-V CG bias for the write cycle. (d) −1.9-V CG bias for the write cycle. (e) +2.1-V CG bias for the erase cycle. (f) Current density to CG-channel voltage relation for the write (black) and erase (red) cycles. Labels (b)–(e) correspond to the simulation results in the respective parts of the figure.

TABLE I
nextnano.MSB MATERIAL PARAMETERS

Parameter	InAs	AlSb
Band-edge offset (eV)	1.390	1.385
Band-edge gap (eV)	0.417	2.386
Band-edge α (eVK^{-1})	0.276E-3	0.42E-3
Band-edge β (K)	93	140
Effective mass m_0	0.026	0.14
Static dielectic constant	15.15	12.04
Optic dielectric consant	12.25	10.24
Deformation potential (eV)	-6.66	-8.12
Material density (kgm^{-3})	5.61E3	4.26E3
LO phonon energy (meV)	30	42
LO phonon width (meV)	3	3

Material parameters used for simulation in nextnano software packages. These can be found in the program database and are fixed to experimental values [6, 17]. LO = longitudinal optical.

predicts that this memory can operate as a disturb-free, fully functional RAM at DRAM speeds, but with the additional advantages of nonvolatility and nondestructive read.

II. Device Concept

The construction of the device is illustrated schematically in Fig. 1(a). The memory features a tunneling junction constructed from thin InAs/AlSb layers to form a triple-barrier structure. The key characteristic of the tunneling junction is that it does not allow electrons to pass through it under zero bias, but will under small potentials between the control gate (CG) and the channel (≤2.3 V). Within a small and specific voltage range (∼0.5 V), electrons are rapidly transported through the junction via RT to (or from) the FG. This results in sharp and high current-density peaks that allow the memory to achieve nonvolatility and RAM capabilities. It is important to understand this process and simulate transport through this region to investigate the performance characteristics of the device.

The FG is an electron-confining layer that stores any charge that tunnels through the thin AlSb barriers, which form the tunneling region [see Fig. 1(a)]. It is this charge storage region that defines the state, similar to the FG metal–oxide–semiconductor field-effect transistor (FGMOSFET) cells used in Flash memory [9]. Logic "1" is assigned to the state in which there are no charges inside the FG. When a suitable voltage pulse is applied, charges tunnel quantum mechanically from the CG into the FG, where they are trapped by an AlSb charge-blocking layer. This state is defined as logic "0," achieved by adding charges to the FG (write cycle). Similarly, a voltage pulse of opposite polarity can be used to remove the charges from the FG in order to return to the "1" state (erase cycle) [3], [9].

III. Write and Erase via RT

The triple-barrier construction of the tunneling region forms two QWs within the structure [see Fig. 1(b)], causing electrons to be confined to distinct energy levels [9]. Two QWs are required to produce a sufficiently thick barrier to prevent leakage via conventional tunneling (i.e., not via a resonant state), while simultaneously utilizing thin QWs raises the confined states to produce a well-defined RT peak. Furthermore, the well thicknesses are sufficiently dissimilar to prevent the energy-state alignment between the two wells, which would otherwise reduce the electron-blocking capability of the central barrier. Applying a voltage across the tunneling junction tilts the conduction band such that the energy levels relative to the

energy of the incident electrons (emitter) change. In the case of this structure, the electrons outside the tunneling junction are in a quasi-bound state due to the formation of a triangle-shaped well from the applied voltage [11]. This is shown by the color scale of the density of states (DOS) of the write process displayed in Fig. 1(c) and (d). In these figures, the conduction band is at a gradient due to an applied voltage at the CG of the device. A similar DOS plot is used for the erase process with an opposite polarity voltage, as shown in Fig. 1(e).

Coherent RT allows the energy levels of the well to act as a filter, allowing only electrons with similar energy to transmit. An applied bias lowers the energy level of the well state relative to the energy of the incident electrons from the emitter, which is the quasi-bound state of the electrons at their source, i.e., at the CG for the write cycle, and the FG for the erase cycle. At a specific applied bias, the energy of the incident electrons and the energy level of the well on the other side of the AlSb barrier are the same, resulting in a sharp increase in transmission through the barrier. Once the applied bias is such that the emitter energy exceeds the QW energies, the transmission through the barrier drops sharply [12]. This is demonstrated by the current-density plot of the tunneling junction of the device in Fig. 1(f), where the applied voltage is across the device terminals (i.e., the 15-nm AlSb barrier is accounted for). The results show two sharp current peaks for the tunneling junction under negative CG bias for the write process. The smaller peak at -1.6 V is the characteristic of the emitter and well energy alignment for QW_2 (QW nearest the FG), where the electron wave function of QW_2 is also spatially present in QW_1, the first well of the tunneling junction [see Fig. 1(c)]. This allows tunneling from the CG to the FG via QW_1 and QW_2 in a fast, coherent process. Similarly, the second, larger peak at higher voltage (-1.9 V) is due to alignment of the quasi-bound emitter energy state with the energy of QW_1 [see Fig. 1(d)]. The applied bias for the DOS plots, labeled c and d in Fig. 1(f), corresponds to the peaks in the tunneling current for the write process, demonstrating that the current–voltage relation of the write cycle is a result of coherent RT through the InAs/AlSb triple-barrier structure from the combined QW_1 and QW_2 energy alignments.

The simulation of the tunneling junction was repeated using opposite polarity voltages for the erase cycle. The results are similar to the write cycle, with a current peak corresponding to the FG electron energies aligning with the QW energies in the tunneling junction [see Fig. 1(e)]. However, the peak is shifted to a higher applied bias due to the difference in energy between the two QW states [see Fig. 1(b)], which is a result of the InAs wells QW_1 and QW_2 having different widths (3.0 and 2.4 nm, respectively). A consequence of this is that the erase voltage is higher than the write voltage.

The resulting current peaks indicate that electrons can be transported both into and out of the FG at low voltages ($\leq$2.3 V), and that the current flowing is zero at zero voltage. Thus, the tunneling junction operates effectively for charge-storage memory device applications, since there is no leakage current through the barriers when the applied bias is removed and a large current density when the appropriate write (or erase) bias is applied. The absence of any current density at 0 V and an extremely small $<$1 Acm^{-2} current density up to $\pm$1 V indicates a good data retention as expected from the 2.1-eV barrier height of the InAs/AlSb system.

The simulations of this process allow us to transfer these results into another model (SPICE) to characterize the more performance-based properties of the memory device using the current density relations of Fig. 1(f). An important realization from the current density results is seen directly from the sharpness of the peaks, with a very small current ($<$1 Acm^{-2}) at voltages away from the peaks [see Fig. 1(f)]. This allows the voltages required for the write and erase cycles to be split between the CG and the channel [with drain D and a back gate (BG) grounded], where they combine to perform the desired write or erase cycle. Crucially, applying one of these half-voltages does not change the logic state of the cell. Later, we will show how this enables us to realize an architecture for a RAM.

IV. Read Operation

To read the data stored in a memory chip, we must be able to determine the logical state of the individual devices (bits) within a large array. In Flash memories, device-level readout is achieved using a threshold voltage, defined as the bias on the CG at which the channel transitions from an insulating to a conducting state. As charge is added to the FG of a device, it partially screens the potential applied across the device at the CG. This shifts the threshold voltage to a larger value, with the magnitude of the voltage shift given by

$$\Delta V_T = \frac{Q_{\text{FG}}}{C_{\text{FG}}} \tag{1}$$

where C_{FG} is the capacitance between the CG and FG (calculated from a parallel plate approximation as 1.2 μFcm^{-2} for our devices) and Q_{FG} is the charge stored in the FG [14]. Note that as both Q_{FG} and C_{FG} are directly proportional to the cross-sectional area, it is eliminated from the above equation. This results in a 1-D equation for the threshold voltage shift, justifying the strictly 1-D simulation used here.

The threshold voltage shift creates a system in which there is a different threshold voltage for the memory device when there is no charge present in the FG (1), compared with the device when charge is present in the FG (0). The difference between these two thresholds creates the threshold voltage window (ΔV_T) [15], within which we can apply a reference voltage (V_{REF}) to determine the logic state of the device: the channel will conduct if it is logic 1 (applied voltage is above threshold) and will not if it is logic 0 (applied voltage is below threshold). Here, we propose to use a similar read technique. The threshold voltage in this device is produced by applying a voltage between the CG and the BG. In the simulations presented here, we use a 12-nm-$In_{0.8}Ga_{0.2}As$ channel for the device [see Fig. 1(a)], although other compositions and thicknesses would have a similar effect: 5 nm of InAs or 14 nm of $In_{0.7}Ga_{0.3}As$, for example. This produces threshold voltages, which, in turn, allow the logical state of an individual device to be read within a large array. This modification also reduces the overall strain on the device in comparison with the previous samples [9]: the substantial reduction in the channel layer thickness compensates for the increased lattice mismatch from introducing a small composition of gallium [16].

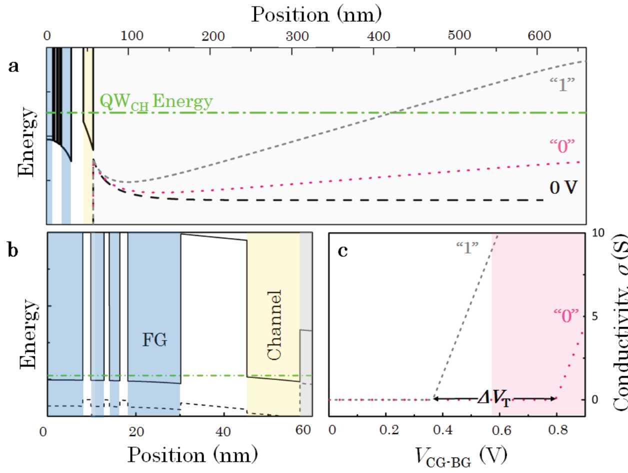

Fig. 2. Read operation of the device. (a) Simulated band diagram (300 K) for the read operation, showing the GaSb valence band relative to the channel QW state (green dashed-dotted line) at 0 V (black dashed line), at V_{REF} for logic 0 (pink dotted line), and at V_{REF} for logic 1 (gray short dashed line). When a portion of the GaSb valence band lies above the QW$_{CH}$ ground-state energy, electrons may flow from the GaSb into the $In_{1-x}Ga_xAs$ channel. (b) Simulated details of the conduction band and valence band for the RT structure, FG barrier, and channel parts of the memory under zero bias. (c) Channel conductivity versus $V_{CG\text{-}BG}$ determined from the simulation results to define logical 1 and 0.

The channel forms a QW (QW$_{CH}$), which raises the minimum energy requirement for electron occupation above the valence band energy of the adjacent GaSb [see Fig. 2(a)]. Consequently, at zero or low bias on the CG, the electrons in the GaSb valence band cannot move into the QW$_{CH}$, resulting in an unoccupied (and, therefore, insulating) channel. Applying a potential (V_{CG-BG}) between the CG and BG raises the GaSb valence band. When a portion of the GaSb valence band exceeds the QW$_{CH}$ ground-state energy, electrons are transferred from the GaSb valence band into the QW$_{CH}$, causing a transition from an insulating state to a conducting state, i.e., there exists a threshold voltage for the transition. This is shown in the simulation results of the read operation of Fig. 2(a) for the reference voltage (V_{REF}), where the QW$_{CH}$ state [Fig. 2(a) and (b) green dashed-dotted line] formed by the $In_{1-x}Ga_xAs$ conduction band is partially below the valence band energy of the GaSb (gray short-dashed line): the channel is occupied and conductive and the device is in logic 1. For a cell in logic 0 with the same reference voltage, the valence band lies underneath the QW$_{CH}$ ground-state energy and the channel remains insulating (pink dotted line).

The density of electrons in the channel, and hence the conductivity, is thus a function of the potential between the CG and the BG. The conductivity of the channel is

$$\sigma = e n_{2D} \mu \tag{2}$$

where e is the charge of an electron and μ is the mobility of the electrons in the $In_{0.8}Ga_{0.2}As$ channel [17]. The electron occupancy of the channel at a given CG–BG voltage is calculated using the 2-D DOS. Thus, the 2-D carrier density

$$n_{2D} = 2\frac{m^*_{CH}}{\pi \hbar^2}\Delta E \tag{3}$$

where m^*_{CH} is the effective mass of the electrons in the channel [17], $\hbar$ is the reduced Planck constant, and ΔE is the energy overlap between the GaSb valence band and the QW$_{CH}$ energy state [18]. Combining (2) and (3) with the simulated energy overlaps (ΔE) for the device [see Fig. 2(a)] allows us to directly obtain a conductivity–voltage relation for reading the device, as depicted in Fig. 2(c).

Similar to Flash technology, adding charge to the FG will partially screen the potential across the device—in this case, the CG–BG potential (V_{CG-BG}). This shifts the entire conductivity–voltage curve to a higher voltage during the write cycle in accordance with (1), represented by the pink dotted line in Fig. 2(c). Likewise, the erase cycle shifts the relation back toward the original state as charge is removed from the FG. The resemblance of the read technique with Flash technologies has no bearing on how the device can perform as an RAM. Indeed, utilizing a similar read technique allows us to assemble the arrays of multiple devices while also enabling single-bit access: it is the triple-barrier RT mechanism that allows this memory to operate as an NVRAM.

V. SPICE Electrical Model

A SPICE program (ltSPICE) was used to combine the write/erase and read simulation results, which were produced using the software packages nextnano.MSB and nextnano++, respectively [7]. There are many examples of SPICE models that have been used to characterize FG memories [13], [19], [20]. However, they are usually focused on modeling a device that has already been fabricated, extracting information for the model from experimental measurements such as capacitive coupling coefficients and tunneling parameters (tunneling parameters can also be modeled [20]). These are then inserted into the simulation to compare directly with experimental data [19], [20]. In this article, where there are no established models or experimentally derived parameters available, the data for the tunneling mechanism are represented by a voltage-controlled current source (VCCS), modeling the current (for a device area, A_{tun}) from a multiple peaked asymmetric-Gaussian fit to the simulated tunneling results of Fig. 1(f). The result is dependent on the voltage applied across the tunneling region. The voltage across the tunneling region comes from two biases during the write and erase processes: the CG voltage and the source (S) voltage. The combined bias across the tunneling region is determined from separate investigations of the band structure gradient (and RT alignments) using a Poisson–Schrodinger solver for an extended nextnano++ simulation of the device with voltages applied from both the CG and S. These provide a relationship between the voltages across the contacts with the voltage seen by the tunneling region of the device. Fig. 1(f) already includes these corrections for a CG voltage only. This gives us a physical model of the tunneling voltages that is likely to be more accurate than the capacitive coupling approximation [20].

Further voltage adjustments are made for the effect of band bending of the highly doped (n+) CG layer, also using nextnano++. We also have to consider the voltage-screening effect due to the presence of charge on the FG, which changes during the write or erase process, so the current supplied by the VCCS changes as its own current output screens the input voltage, i.e., build up, or loss of, charge in the FG during the write and erase pulses, respectively.

The simplest way to model this system is to connect the VCCS that contains all the above information to a capacitor with capacitance C_T, the total capacitance coupled to the FG

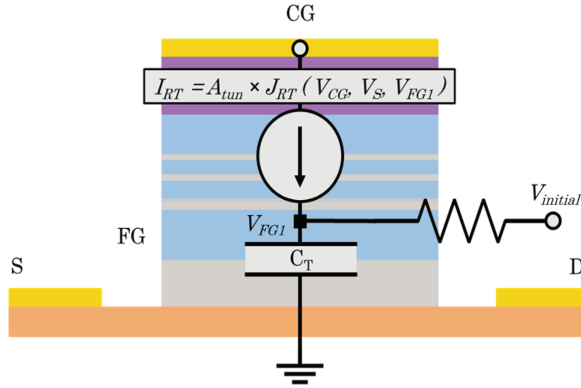

Fig. 3. SPICE simulation of the device using a VCCS containing the RT results of Fig. 1, where the tunneling voltage is given as a function of the CG voltage (V_{CG}), source voltage (V_S), and charge-screening voltage (V_{FG1}). $V_{INITIAL}$ allows us to add an initial screening voltage (used for the erase cycle).

from the tunneling junction and the charge blocking barrier (calculated from a parallel-plate approximation as 2 μFcm^{-2}; see Fig. 3). When a voltage pulse is applied, it is converted into the voltage across the tunneling junction, from which the VCCS responds according to the RT simulation results of Fig. 1 to release a current, continuously adapted to consider the changing charge on the FG. The electrons released in the write process are stored on the FG capacitor, and a voltage V_{FG1} is created (see Fig. 3)

$$V_{FG1} = \frac{Q_{FG}}{C_T}. \quad (4)$$

This result then feeds back into the VCCS as a voltage-screening effect. Similarly, this setup can be used to simulate the charges leaving the FG (erase), where an initial voltage, $V_{INITIAL}$, defines the previously written state for the device. Combining (1) and (4) with the capacitances for the device, approximated as parallel-plate capacitors using the layer thicknesses and dielectric constants of the materials, allows us to obtain an equation for the threshold voltage shift of the channel as a function of V_{FG1}, that is

$$\Delta V_T = \frac{C_T}{C_{FG}} V_{FG1}. \quad (5)$$

The result is that we can track the threshold shift for any given voltage pulse in a transient simulation to determine the change in the conductivity relation of the channel discussed in Section IV [see Fig. 2(c)].

VI. Memory Architectures

The similarities between the device reported here and Flash memory cells readily allow compatibility with Flash architectures, i.e., it could be implemented in a NAND-type architecture, with devices connected in series in large strings. This will allow for a low-power high-endurance alternative to NAND Flash. However, large-scale use would require 3-D implementation and consequent increase in areal bit density to compete with the transition from planar to 3-D NAND Flash. An alternative is use in niche applications, where reliable data retention, high speed, and low energy are preferred to the high-bit density of FGMOSFET-based Flash memory.

More interesting is the implementation in an architecture suitable for active memory (RAM). The most important feature of an active memory is that it allows fast access to individual bits (devices) at the command of the user [21]. For our devices, this can be realized by implementing a NOR-type architecture (see Fig. 4). Note that we introduce a new device symbol in Fig. 4, similar to the well-known FGMOSFET device symbol but combined with an RT diode symbol to specify the write/erase mechanism. Due to the nature of RT, the current peaks for the write and erase processes are very sharp [see Fig. 1(f)]. This allows for the use of half-voltages, where half of the required voltage for writing or erasing data is applied to the CG and the other half to the channel. When only a single half-voltage is applied to any device, the state of the device remains intact. This feature can be used to target individual devices in an array by selecting half-voltages on the desired wordline and bitline, which we designate as CG and S, respectively. These combine to write or erase the target device without compromising the data stored in the surrounding devices (disturb). It is important to note that the BG terminal serves as a common ground for all devices in the array and that devices are back to back in pairs with grounded drain contacts, permitting a highly efficient architecture (see Fig. 4).

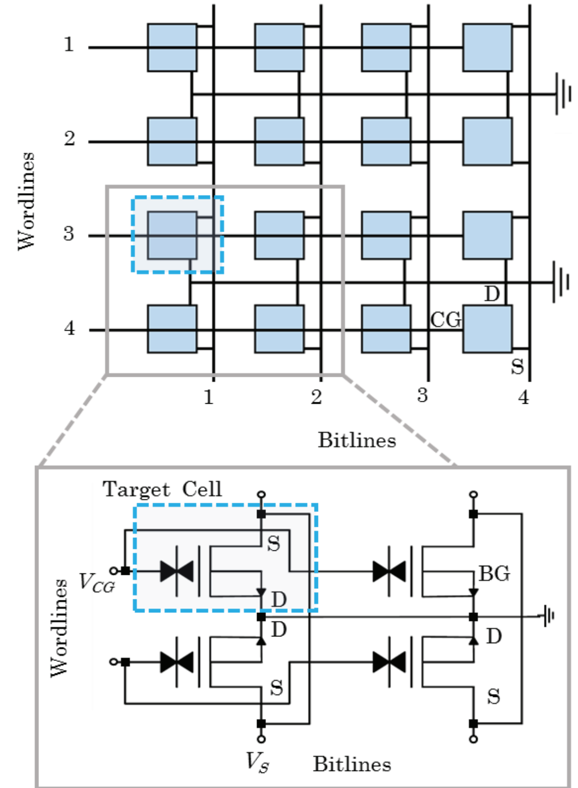

Fig. 4. Schematic of the proposed architecture for low-power, low-disturb NVRAM. Individual cells are addressed by the application of half-voltages to the appropriate wordlines and bitlines, without disturbing the other cells. For the example shown here, wordline 3 and bitline 1 are used to address the target cell (indicated by the dashed box).

The read operation is otherwise identical to that found in NOR-Flash memory and permits the reading of individual devices with this architecture [22]. This is achieved by applying a read voltage, V_{REF}, between CG and BG (CG and ground), to the appropriate wordline, a small voltage, e.g., <0.5 V, to the appropriate bitline, and testing for channel conductivity (current flow). Note that since the devices are normally off, current will only flow if the particular device that is addressed is in a logical-1 state. V_{REF} should be chosen such that it falls between the two threshold voltages of the

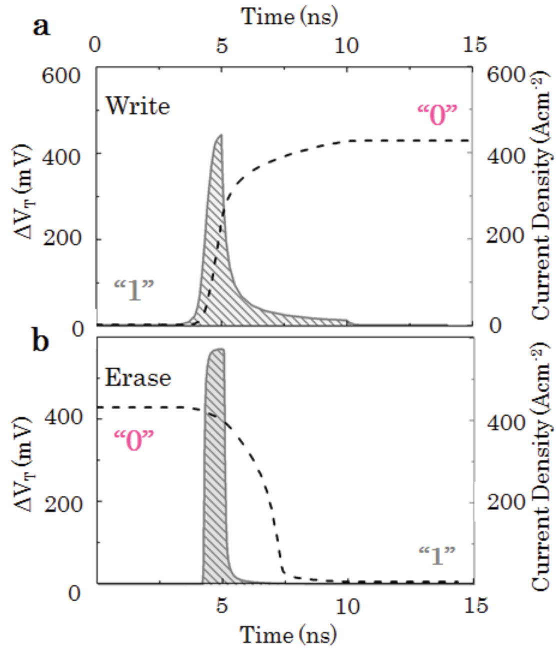

Fig. 5. Transient simulation for the change in threshold voltage (black dashed line) during the voltage pulse with the corresponding current density through the tunneling region (gray line) for (a) write cycle (top) and (b) erase cycle (bottom). In both cases, the logic state is changed within 10 ns.

0 and 1 states, e.g., 0.6 V [see Fig. 2(c)]. The ability to target individual devices (bits) lends itself toward RAM applications due to the speed of addressing an individual bit at random. Unlike the dominant RAM technology, DRAM, this memory will be nonvolatile with nondestructive read, but with similar (or improved) performance capabilities in other respects.

VII. Fast Low-Energy NVRAM

The modeling indicates that such an NVRAM can operate at low voltage, low energy, and high speeds. A transient simulation for the write cycle with a 5-ns rise time and 5-ns duration, demonstrating the potential speed of the device, is shown in Fig. 5(a). This gives a total pulse time of 10 ns, similar to the speed of DRAM [23]. There is a dependence on both the rise time and duration of pulse for the threshold shift; thus, they were set equal for the purposes of investigating the device speed. The 5-ns rise-time voltage pulse was selected specifically with DRAM in mind, where this speed limitation is a result of capacitive charging within a memory array. Thus, the choice of the voltage pulse considers capacitive limitations brought about by implementation in a hypothetical array. The figure depicts the change in the threshold voltage in real time during the pulse, along with the corresponding tunneling current density, i.e., the current density tunneling into the FG during the write pulse [see Fig. 5(a)]. The charge density stored in the FG is, therefore, the area under this plot and is the sole reason for the change in the threshold voltage in accordance with (1). Fig. 5(b) shows the same plot for the erase cycle, operating at similar speed and voltage, although not exactly the same, as the voltages have been optimized for minimal disturbances and an exact return to the original state after the erase cycle, i.e., with equal area under the current density curves (see Fig. 5), as we discuss now.

TABLE II
Benchmarking Metrics

Technology	Cell Switching Energy (J)	Fundamental Particle	Number
DRAM [23]	$E=0.5\times CV^2$ $E=0.5\times 15fF\times 0.6V^2$ $E\sim 10^{-15}$	Electron	$15fF\times 0.6V/q$ $\sim 5\times 10^4$
3D NAND Flash [23,26]	$E=0.5\times CV^2$ $E=0.5\times 50aF\times 20V^2$ $E\sim 10^{-14}$	Electron	$\sim 10^4$
PCM [23]	E=IVt E=0.1mA×4V ×0.4μs $E\sim 10^{-10}$	Atomic bond Bond angle Bond coordination	DFT $\sim 2\times 10^4$
RRAM [23]	E=IVt E=50uA×3V ×50ns $E\sim 10^{-11}$	Cluster of oxygen vacancies or metal ions	DFT [25] 10-1000
This work	$E=0.5\times CV^2$ $E=0.5\times 8aF\times 2.3V^2$ $E\sim 10^{-17}$	Electron	$8aF\times 2.3V/q$ ~ 100

Benchmarking metrics of memory technologies with a 20 nm feature size, in both production and research phases. The metrics for our memory device (20 nm feature size) show that the switching energy is significantly lower than all other technologies, including DRAM (100× lower) and 3D Flash (1000× lower).

The four optimized half-voltage pulses are −0.85 V (CG-write), 0.90 V (S-write), −1.16 V (S-erase), and 1.16 V (CG-erase). The total voltage for the write and erase cycles is slightly larger than the voltages corresponding to the peak current density [see Fig. 1(e)]. This is due to the change in voltage on V_{FG1} during the write and erase processes, which screens some of the applied potential and must be compensated by a slightly higher voltage. The unique voltage amplitude to each bitline or wordline for write or erase is chosen such that the threshold shift for the write and erase processes is exactly opposite, ensuring there is no drift in the threshold voltages over many cycles. The half-voltages, when applied individually, have a negligible effect on surrounding cells. The greatest disturbance on the cells was from the −0.85-V write half-voltage applied to the wordline and was determined to be approximately one electron loss every 4000 10-ns pulses for a 20-nm feature size. The extremely low disturbance of cells is derived from the lack of tunneling current at low voltages. This is demonstrated directly from the current density simulations [see Fig. 1(f)], where the current density is under 1 Acm^{-2} in the 0.85–1.16-V range (compared with a 10^4 Acm^{-2} peak magnitude). For the read process, the model predicts an excellent 0/1 threshold contrast of 430 mV [see Fig. 2(c)].

If we now compare some of the important memory metrics for different types of memory cells with 20-nm feature size cell [23], [24], both in production and under development, we observe some interesting results (see Table II). The most notable is the switching energy, which is lower than both DRAM and 3-D NAND Flash by factors of 100 and 1000, respectively, and thus also significantly lower than other emerging memory technologies. This remarkable observation is a result of the combination of low voltages and small capacitance in our devices. Furthermore, it contradicts the

argument that nonvolatility requires the expenditure of more energy to change the states than a volatile memory, due to the energy required to overcome the barrier energy [23]. This is not the case for RT as there exists only very specific energy alignments at which the tunneling can occur, allowing us to have a high barrier energy but still observe tunneling at small voltages. The only issue that comes to light in the benchmarking metrics listed in Table II is the electron number, which is the downside of the small capacitance of the FG. With only 100 electrons in the FG for the written state (0) at this feature size, a leakage of 30–50 electrons could result in failure of that data cell. However, the simulated 0-V leakage currents are negligible at 300 K, with an extremely small disturb for half-voltage pulses, as previously discussed. Moreover, 2-D NAND Flash technologies of similar feature size have just 30–50 electrons per cell level [24]. This comparison, combined with the high barrier energy and low disturb rate, suggests that this low number of stored electrons is not a stumbling block, at least until the technology is scaled to feature sizes <10 nm.

VIII. Conclusion

We have demonstrated a III-V semiconductor NVRAM with startlingly low switching energy (10^{-17} J for a 20-nm feature size) that operates as an FG memory at significantly lower voltages than Flash ($\leq$2.3 V). Positive endurance and data retention results are expected due to the extremely low switching energy and large barrier energy (2.1 eV), although rigorous testing of this on experimental devices is required. The combination of nextnano.MSB, nextnano++, and SPICE simulations indicates that the device can operate virtually disturb-free at 10-ns pulse durations, a similar speed to the volatile alternative, DRAM. These advantages are derived from the triple-barrier RT mechanism used to transport the charge in and out of the device, which occurs at much lower voltages than other FG memories (i.e., Flash). The proposed device has a threshold voltage and threshold voltage shift due to charge storage, allowing a similar read process to that of FGMOSFET cells used in Flash memory. This is achieved using a broken gap (Type-III) conduction band alignment formed from an $In_{1-x}Ga_xAs$/GaSb heterojunction, where the $In_{1-x}Ga_xAs$ channel is a thin (12 nm) QW. An excellent contrast in threshold voltages between the 0 state and 1 state is achieved. The resemblance to Flash memory cells allows NAND or NOR Flash architectures to be directly implemented on the device to produce large arrays. The simulation results indicate that half-voltages can be used within a NOR-type architecture to target individual cells for write, erase, and read processes. This exclusive feature, combined with the increased speed suggested from the transient results of the 1-D model, predicts that the device can be implemented in large arrays as a low-power, nonvolatile, nondestructively read alternative to DRAM.

Acknowledgment

The data in the figures of this manuscript are openly available from Lancaster University data archive in [27].

References

[1] I. Bhati, M. Chang, Z. Chishti, S. Lu, and B. Jacob, "DRAM refresh mechanisms, penalties, and trade-offs," *IEEE Trans. Comput.*, vol. 65, no. 1, pp. 108–121, Jan. 2016.
[2] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature Nanotechnol.*, vol. 10, no. 3, pp. 191–194, Mar. 2015.
[3] M. Hayne, "Electronic memory devices," U.S. Patent 10 243 086 B2, Mar. 26, 2019.
[4] T. Nowozin, D. Bimberg, K. Daqrouq, M. N. Ajour, and M. Awedh, "Materials for future quantum dot-based memories," *J. Nanomater.*, vol. 2013, pp. 1–6, Jun. 2013.
[5] P. Greck, S. Birner, B. Huber, and P. Vogl, "Efficient method for the calculation of dissipative quantum transport in quantum cascade lasers," *Opt. Express*, vol. 23, no. 5, pp. 6587–6600, 2015.
[6] S. Birner. Nextnano website. Nextnano GmbH Company. Accessed: Dec. 19, 2019. [Online]. Available: http://www.nextnano.de
[7] A. Vladimirescu, *The SPICE Book*, New York, NY, USA: Wiley, 1994.
[8] I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, "Band parameters for III-V compound semiconductors and their alloys," *J. Appl. Phys.*, vol. 89, no. 11, pp. 5815–5875, 2001.
[9] O. Tizno, A. R. J. Marshall, N. Fernández-Delgado, M. Herrer, S. I. Molina, and M. Hayne, "Room-temperature operation of low-voltage, non-volatile, compound-semiconductor memory cells," *Sci. Rep.*, vol. 9, Jun. 2019, Art. no. 8950.
[10] Integrated Circuit Engineering Corporation. *ROM, EPROM, & EEPROM Technology*. Accessed: Dec. 19, 2019. [Online]. Available: https://web.eecs.umich.edu/prabal/teaching/eecs373f10/readings/romepromeeprom-technology.pdf
[11] P. Greck, "Efficient calculation of dissipative quantum transport properties in semiconductor nanostructures," Ph.D. dissertation, Technische Univ. München, München, Germany, 2012.
[12] S. Datta, *Electronic Transport in Mesoscopic Systems*. New York, NY, USA: Cambridge Univ. Press, 1997, pp. 246–266.
[13] Y. H. Kang and S. Hong, "A simple flash memory cell model for transient circuit simulation," *IEEE Electron Device Lett.*, vol. 26, no. 8, pp. 563–565, Aug. 2005.
[14] B. Kalyan and B. Singh, "Design and simulation equivalent model of floating gate transistor," in *Proc. Annu. IEEE India Conf. (INDICON)*, Dec. 2015, pp. 1–6.
[15] K. V. Noren and M. Meng, "Macromodel development for a FLOTOX EEPROM," *IEEE Trans. Electron Devices*, vol. 45, no. 1, pp. 224–229, Jan. 1998.
[16] J. Singh, *Electronic and Optoelectronic Properties of Semiconductor Structures*. Cambridge, U.K.: Cambridge Univ. Press, 2003, pp. 478–483.
[17] Y. Li, Y. Zhang, and Y. Zeng, "Electron mobility in modulation-doped AlSb/InAs quantum wells," *J. Appl. Phys.*, vol. 109, no. 7, 2011, Art. no. 073703.
[18] J. Davies, *The Physics of Low-Dimensional Semiconductors: An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 1998, pp. 118–142.
[19] J. Suñé, S. Lanzoni, R. Bez, P. Olivo, and R. Riccò, "Transient simulation of the erase cycle of floating gate EEPROMs," in *IEDM Tech. Dig.*, Washington, DC, USA, Dec. 1991, pp. 905–908.
[20] A. Kolodny, S. T. K. Nieh, B. Eitan, and J. Shappir, "Analysis and modeling of floating-gate EEPROM cells," *IEEE Trans. Electron Devices*, vol. ED-33, no. 6, pp. 835–844, Jun. 1986.
[21] B. Jacob, S. Ng, and D. Wang, *Memory Systems: Cache, DRAM, Disk*. Amsterdam, The Netherlands: Elsevier, 2007.
[22] R. Micheloni, G. Campardo, and P. Olivo, *Memories in Wireless Systems*. Berlin, Germany: Springer-Verlag, 2008.
[23] K. Prall, "Benchmarking and metrics for emerging memory," in *Proc. IEEE Int. Memory Workshop (IMW)*, Monterey, CA, USA, May 2017, pp. 1–5.
[24] K. Prall, "Scaling non-volatile memory below 30 nm," in *Proc. IEEE NVSMW*, Aug. 2007, pp. 5–10.
[25] S. Sills *et al.*, "A copper ReRAM cell for storage class memory applications," in *Symp. VLSI Technol. (VLSI-Technol.), Dig. Tech. Papers*, Honolulu, HI, USA, Jun. 2014, pp. 1–2.
[26] R. Micheloni, L. Crippa, C. Zambelli, and P. Olivo, "Architectural and integration options for 3D NAND flash memories," *Computers*, vol. 6, no. 3, p. 27, 2017.
[27] *Simulations of Ultra-Low-Power Non-Volatile Cells for Random Access Memory (Dataset)*. Accessed: Dec. 19, 2019. [Online]. Available: https://doi.org/10.17635/lancaster/researchdata/307