ULTRA**RAM**: Toward the Development of a III–V Semiconductor, Nonvolatile, Random Access Memory

D. Lane, P. D. Hodgson, R. J. Potter, R. Beanland, and M. Hayne

***Abstract*— ULTRARAM is a III–V compound semiconductor memory concept that exploits quantum resonant tunneling to achieve nonvolatility at extremely low switching energy per unit area. Prototype devices are fabricated in a 2 × 2 memory array formation on GaAs substrates. The devices show 0/1 state contrast from program/erase (P/E) cycles with 2.5 V pulses of 500-μs duration, a remarkable switching speed for a 20 μm gate length. Memory retention is tested for 8 × 10^4 s, whereby the 0/1 states show adequate contrast throughout, whilst performing 8 × 10^4 readout operations. Further reliability is demonstrated via program-read-erase-read endurance cycling for 10^6 cycles with 0/1 contrast. A half-voltage array architecture proposed in our previous work is experimentally realized, with an outstandingly small disturb rate over 10^5 half-voltage cycles.**

***Index Terms*— InAs/AlSb, memory, non-volatile memory (NVM), non-volatile RAM (NVRAM), resonant tunneling.**

I. Introduction

A ''universal memory'' should combine the best aspects of dynamic random access memory (DRAM) and flash. In essence, it must have very robust logic states that can, nevertheless, be easily changed. As the nature of these requirements appears to be contradictory, the widely accepted view is that universal memory is unfeasible [1] or almost impossible [2]. ULTRA**RAM**[1] is a novel, III–V compound-semiconductor memory that utilizes the unusual band offsets of the 6.1-Å semiconductor family (InAs, AlSb, and GaSb) [3]. In particular, the extraordinarily large conduction-band offset of InAs/AlSb (2.1 eV) delivers electron barriers akin to those of dielectrics to achieve nonvolatility. In common with flash, the logic state is defined by charge (electrons) stored within a floating gate (FG). However, in ULTRA**RAM** electrons are transported into and out of the FG via triple-barrier resonant tunneling (TBRT) structure formed from InAs/AlSb heterojunctions [4]. This resolves the paradox of universal memory, as the tunneling structure provides a high-energy barrier when there is no bias applied, but allows resonant-tunneling (i.e., transparent barriers) at program/erase (P/E) voltages of around 2.5 V, approximately ten times lower than flash. These characteristics are predicted by simulations of quantum transport [5] and have previously been demonstrated in single devices at room temperature [6]. The intricate physics of the tunneling mechanism used here and a comparison of ULTRA**RAM** with current and emerging memory technologies are described in detail in our previous work for the interested reader [5]. Additionally, the devices out-perform other resonant-tunneling-based memories in endurance benchmarks with at least a similar logic retention time [7], [8]. Most importantly, the FG design allows for high-density array architectures and the possibility of vastly improved readout (1/0) contrast [5]. Moreover, the current through the gate during P/E cycles is extremely small, significantly reducing memory power consumption by comparison.

Initial prototype single-cell devices [6] exhibited a limited endurance despite the extraordinary InAs/AlSb conduction band offset and switching at extremely low voltages. This was undoubtedly the result of a large (milliampere) hole leakage current passing from the control gate (CG) terminal to the source/drain (S/D) terminals due to the low valence band offset of the InAs/AlSb heterojunction of just 0.1 eV. Here, the design is amended to include an Al_2O_3 gate dielectric formed via atomic layer deposition (ALD). This layer provides the necessary band offsets with InAs to block all carrier flow through the CG [9] but requires the memory tunneling structure to be reversed such that tunneling for P/E cycles occurs from the source of the cell (Fig. 1).

II. Fabrication

The ULTRA**RAM** memory heterostructure (Fig. 1) was grown on 2-in highly doped (Si, $n \sim 2 \times 10^{18}$ cm^{-3})

Manuscript received December 10, 2020; revised January 27, 2021 and February 15, 2021; accepted March 4, 2021. Date of publication March 25, 2021; date of current version April 22, 2021. This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC), U.K., via the 2017–2020 Impact Acceleration Account (IAA) funding allocation to Lancaster University under Grant EP/R511560/1 and Grant EP/N509504/1, in part by the Future Compound Semiconductor Manufacturing Hub under Grant EP/P006973/, in part by the ATTRACT Project funded by the European Commission (EC) under Grant 777222, and in part by the Joy Welch Educational Charitable Trust. The review of this article was arranged by Editor J. Kang. *(Corresponding author: D. Lane.)*

D. Lane, P. D. Hodgson, and M. Hayne are with the Department of Physics, Lancaster University, Lancaster LA1 4YB, U.K. (e-mail: d.lane@lancaster.ac.uk; p.hodgson1@lancaster.ac.uk; m.hayne@lancaster.ac.uk).

R. J. Potter is with the Department of Mechanical, Materials and Aerospace Engineering, University of Liverpool, Liverpool L69 3GH, U.K. (e-mail: rjpott@liverpool.ac.uk).

R. Beanland is with the Department of Physics, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: r.beanland@warwick.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2021.3064788.

Digital Object Identifier 10.1109/TED.2021.3064788

[1]Trademarked.

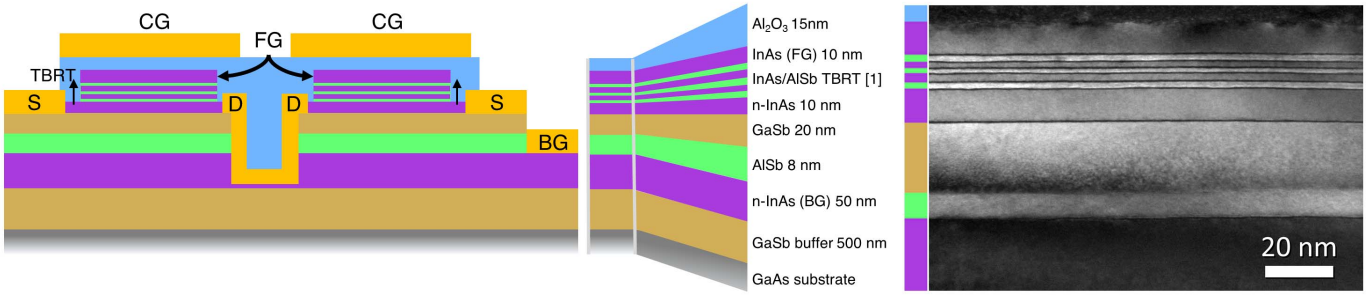

Fig. 1. Left: cross-sectional schematic of ULTRA**RAM** and material layers. Right: dark field $g = 002$ TEM image of the epitaxial structure. Layer thicknesses for the TBRT region can be found in Table I.

TABLE I
CROSS-SECTIONAL TEM MEASUREMENTS

	Target layer thickness (nm)	Measured Layer thickness (nm)
AlSb barrier 1	1.8	2.1
InAs QW 1	2.4	1.8
AlSb barrier 2	1.2	1.6
InAs QW 2	3.0	2.3
AlSb barrier 3	1.8	2.1

Layers of the tunnelling region are ordered top down from the surface of the wafer. Target thicknesses are based on detailed simulations of the device physics [5].

GaAs wafers by molecular beam epitaxy (MBE) on a Veeco GENxplor system. The 7.8% GaSb/GaAs lattice mismatch was mitigated by the use of an interfacial misfit array between the substrate and GaSb buffer layer [10] before the growth of the GaSb/InAs/AlSb memory structure. Layer thicknesses are measured via cross-sectional transmission electron microscopy (TEM) with the crucial TBRT structure thicknesses listed in Table I.

Memory arrays were processed on the MBE-grown wafer using a top-down approach (Fig. 1). Devices were fabricated using standard photolithography techniques. Inductively coupled plasma (ICP) etching with $BCl_3/Cl_2/Ar$ gas mixtures was used to access the back gate (BG) layer. *In situ* reflectance monitoring allowed etching to cease accurately in the desired layer. In order to reveal the channel layer, an alternating selective wet etch was employed to etch each layer in succession. Microposit MF-319 (tetramethylammonium hydroxide) was used to selectively etch AlSb and GaSb over InAs [11], and a citric-acid-based etchant ($C_6H_8O_7$:H_2O_2:H_2O) was used to selectively etch InAs over AlSb and GaSb. Contacts joining D-BG-D along with S terminals were fabricated via Ti-Au sputtering through lift-off resist windows. The memory design utilizes a gate-last approach, where the ALD-Al_2O_3 layer was deposited over the surface prior to metal CG layers being added. This was followed by further SiO_2 passivation via plasma-enhanced chemical vapor deposition. Last, the device CG, S, and BG terminals were revealed once more by buffered HF etching of the Al_2O_3 and SiO_2 layers, before depositing metal Ti–Au contact pads. A scanning electron microscope image of the fabricated arrays is shown in Fig. 2, where word lines (WLs) connecting CG terminals pass across the array horizontally. Bit lines (BLs) connecting S terminals are situated vertically in the image, separated from the underlying WL contact by the SiO_2 layer.

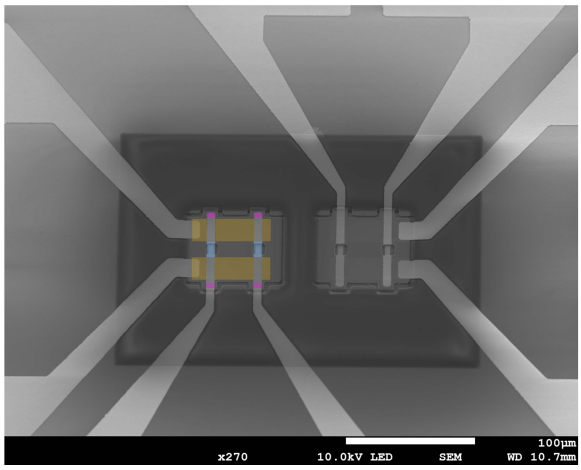

Fig. 2. Scanning electron microscope image of two ULTRA**RAM** 2 × 2 memory arrays. The false coloring on the array pictured to the left indicates the extent of the WLs (yellow) and the etched access for the BL contacts to the S terminals (pink). The buried back-to-back D contacts, which are connected to the BG and isolated from the BL, are in the center of each pair of devices (blue).

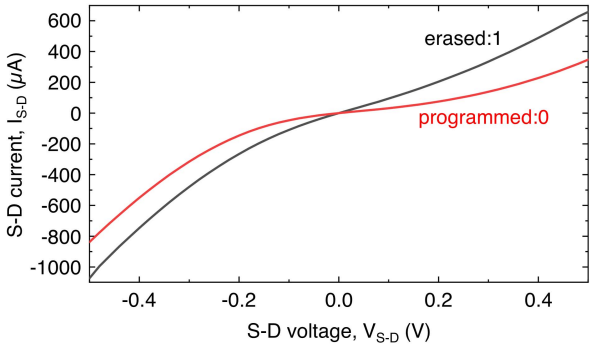

Fig. 3. Sweep of S-D voltage with measured S-D current after an erase cycle of 2.5 V, 500 μs duration (black line) and a program cycle of −2.5 V, 500 μs duration.

III. Low Voltage P/E

Fig. 3 presents the current flow from S to D during an S–D voltage sweep after a 500-μs, −2.5-V program cycle (red) and

a 500-μs, +2.5–V erase cycle (black) applied to the S terminal of a 20-μm-gate-length cell within a 2 × 2 memory array. Such a P/E cycle corresponds to a 10^2 and 10^3 reduction in switching energy per unit area compared to DRAM and NAND flash, respectively [12]. There is clear state contrast between 0/1 following the P/E cycles. Overall current is significantly reduced compared to the previous iteration of the technology [6], due to the introduction of the Al_2O_3 gate dielectric. Moreover, CG-D resistance is improved from 10^3 to $>10^{10}\Omega$ (the limit of measurement). Within the array, architecture S-D current (I_{S-D}) is measured via the BG terminal as the D terminals are buried within the random access memory (RAM) architecture [5], as shown in Fig. 1.

The speed of the P/E cycle is noteworthy, and is 2000× faster than previous devices [6]. As the speed of quantum tunneling is in the sub-picosecond scale [13], the switching speed is limited by the *RC* time constant, and is, therefore, subject to Dennard's scaling law (scaling linearly with the area) [14]. Thus, for a 20-nm gate length device with ideal scaling sub-nanosecond switching speed is predicted–significantly faster than DRAM and comparable to static RAM (SRAM) [1], [2], [12]. However, rigorous testing on small-scale devices is required to confirm this.

P/E cycles at ±2.5 V were carried out by applying the voltage pulse to the BL whilst grounding the WL of the target device. The other cell on the array which shares this BL is undisturbed as its CG terminal is floating. Previously, a half-voltage architecture was proposed in which individual memory cells are selected by applying half of the required P/E voltage to the WL and the other half to the BL [5]. It is found that the same 0/1 contrast can be obtained using this P/E scheme, whereby ±1.25 V pulses applied to BL and WL are used to cycle the memory state. A disturbance test consisting of an uninterrupted ±1.25 V bias was applied separately to BL and WL in both 0 and 1 states for 120 s, equivalent to 10^5 P/E cycles, and did not perturb the memory state from a 0 or 1 logic position.

The results presented in Fig. 3 show a clear, measurable difference between the 0 and 1 states. However, if potentially 1000's of cells are to be connected in a single BL in the future, a dramatic improvement in read contrast (0/1) is of paramount importance [12]. Fortunately, the insufficient read contrast is not an indication of logic state weakness, but rather due to the simplicity of the channel construction. The channel of the memory cells is formed from an n-doped InAs layer and is, therefore, normally-ON. It is partially depleted by the presence of FG charges, resulting in a measurable, but limited, change in channel conductivity. Work is ongoing to incorporate the normally-OFF InGaAs channel design described in our previous work [5] to address this issue. Producing a threshold-voltage-based readout scheme should dramatically improve readout contrast allowing larger memory arrays. Although the P/E cycling scheme for the RAM architecture presented in [5] has been confirmed, the readout contrast and uniformity in channel conductivity are not sufficient to reliably test device-to-device switching in the array formation. While the insufficient contrast is due to the channel construction, the variation in the channel conductivity is a result of a suboptimal etching procedure (discussed in Section IV). As such, tests are carried out on a single device within the array formation, with surrounding devices ignored.

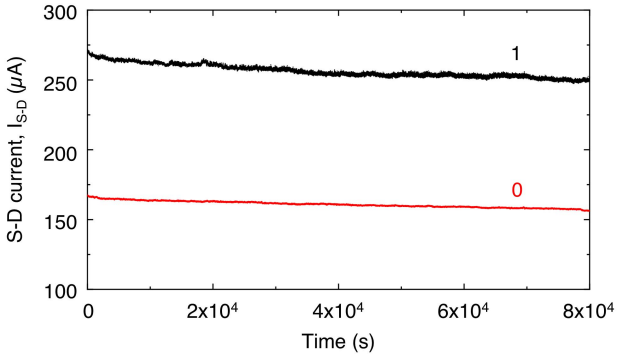

Fig. 4. Retention data for a 20-μm gate length cell in a 2 × 2 memory array. Memory logic is programed (0) and erased (1) with 500 μs pulses of −2.5 and +2.5 V on the S terminal, respectively. Read-out current is measured with a 0.5 V source voltage every second.

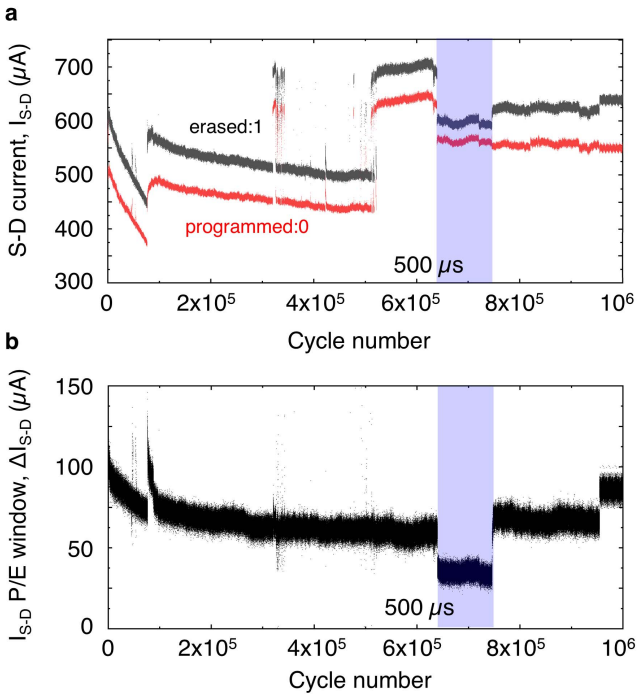

Fig. 5. Endurance data for an ULTRA**RAM** memory cell in an array. (a) S-D current after a +2.5 V erase cycle (gray), and a −2.5 V program cycle (red). Pulse duration was set to 5 ms, except for those data points with blue shading where a 500-μs pulse duration was used. (b) S-D current difference calculated by subtracting erase and program current from consecutive cycles.

IV. Reliability

Fig. 4 demonstrates the stability of the memory logic over time. Both programed and erased states show stable contrast over 8 × 10^4 s and 8 × 10^4 readout cycles at 0.5 V S-D bias. The nonvolatility of the memory is a result of the 2.1 eV barriers from the InAs/AlSb heterostructures on the one side of the FG, and the 3.1 eV barrier from the InAs/Al_2O_3 interface on the other [9].

Typically, FG-storage memories such as flash suffer from poor endurance (i.e., degradation due to many P/E cycles), such that wear-leveling is required to prolong their lifetime [15]. Wear leveling is unsuitable for RAM, which requires superior endurance properties, with individual cells being programed and erased with each computational operation. In this work, ULTRA**RAM** cells withstood 10^6 P/E (P-read-E-read) switching cycles [Fig. 5(a)], whilst maintaining a clear 0/1 state contrast. P/E cycling was performed at a rate of 200 cycles per minute with 5 ms P/E pulses, except for the blue shaded region, where it was shortened to 500 μs, reducing the 0/1 contrast. The reason for this reduction is the significant *RC* time constant due to the device feature size (i.e., the gate-stack potential does not reach 2.5 V within the pulse). The tunneling mechanism itself is intrinsically extremely fast [5].

In this first-ever test, endurance is at least an order of magnitude higher than flash memory [2]. There is, however, movement of the 0/1 window throughout the process. The reason for this is currently unknown, but it is thought that it may be a result of an inconsistent channel contact that is sensitive to temperature or vibrations. Atomic force microscopy of the wet-etched channel surface shows significant etch pitting, which could cause intermittent contact with the underlying layers. An ICP etch process to create a smooth surface for consistent contact to the thin (10 nm) channel material is currently being developed in response. Fluctuations in I_{S-D} offset aside, a memory is realized for over 10^6 cycles (Fig. 5). Moreover, the difference in current between 0 and 1 [ΔI_{S-D}, Fig. 5(b)] persists throughout the endurance test with the P and E states tracking each other. Despite the inconsistencies in overall current, Fig. 5(b) shows a significant 0/1 state contrast over the 10^6 logic-switching cycles.

V. Conclusion

We have experimentally confirmed the principles required for a RAM using the III–V ULTRA**RAM** memory concept within cells of 2 × 2 arrays. Cells can be programed and erased at extremely low switching energy (per unit area) using a half-voltage architecture in which the P/E voltage is split between BL (S) and WL (CG). The logic states of cells within this architecture are shown to be disturb-free for the equivalent of at least 10^5 cycles. An up to 2000× improvement in switching speed compared with previous devices is demonstrated, with P/E at ≥500 μs for a 20 μm gate length. Assuming capacitive scaling, this predicts sub-nanosecond operation at the 20 nm node. Highly robust retention of both states is established for 8×10^4 s with 8×10^4 reads, limited only by the length of the experiment. Memory cells can withstand 10^6 P/E cycles without degradation, thus the benchmark for endurance exceeds that of flash and many resistive-memory technologies. As a result, fast, ultraefficient, nonvolatile, random access ULTRA**RAM** memories are a real possibility.

Acknowledgment

The data in the figures of this article are openly available from Lancaster University data archive in [16].

References

[1] 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.

[2] S. Yu and P.-Y. Chen, "Emerging memory technologies: Recent trends and prospects," *IEEE Solid State Circuits Mag.*, vol. 8, no. 2, pp. 43–56, Spring 2016, doi: 10.1109/MSSC.2016.2546199.

[3] 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.

[4] M. Hayne, "Electronic memory devices," U.S. Patent 10 243 086 B2, Dec. 7, 2017.

[5] D. Lane and M. Hayne, "Simulations of ultralow-power nonvolatile cells for random-access memory," *IEEE Trans. Electron Devices*, vol. 67, no. 2, pp. 474–480, Feb. 2020, doi: 10.1109/TED.2019.2957037.

[6] O. Tizno, A. R. J. Marshall, N. Fernández-Delgado, M. Herrera, S. I. Molina, and M. Hayne, "Room-temperature operation of low-voltage, non-volatile, compound-semiconductor memory cells," *Sci. Rep.*, vol. 9, no. 1, p. 8950, Dec. 2019.

[7] M. Nagase, T. Takahashi, and M. Shimizu, "Resistance switching memory operation using the bistability in current–voltage characteristics of GaN/AlN resonant tunneling diodes," *Jpn. J. Appl. Phys.*, vol. 55, no. 10, pp. 100301–100304, 2016.

[8] J. Denda, K. Uryu, K. Suda, and M. Watanabe, "Resistance switching memory characteristics of Si/CaF_2/CdF_2/CaF_2/Si resonant-tunneling quantum-well structures," *Appl. Phys. Exp.*, vol. 7, no. 4, Apr. 2014, Art. no. 044103.

[9] H.-Y. Chou *et al.*, "Band offsets and trap-related electron transitions at interfaces of (100) InAs with atomic-layer deposited Al_2O_3," *J. Appl. Phys.*, vol. 120, no. 23, Dec. 2016, Art. no. 235701, doi: 10.1063/1.4971178.

[10] A. P. P. J. H. H. Craig Carrington Liu and A. R. J. Marshall, "Characterization of 6.1 Å III–V materials grown on GaAs and Si: A comparison of GaSb/GaAs epitaxy and GaSb/AlSb/Si epitaxy," *J. Cryst. Growth*, vol. 435, pp. 56–61, Feb. 2016.

[11] C. Gatzke, S. J. Webb, K. Fobelets, and R. A. Stradling, "*In-situ* monitoring of the selective etching of antimonides in GaSb/AlSb/InAs heterostructures using Raman spectroscopy," in *Proc. IEEE 24th Int. Symp. Compound Semiconductors*, Sep. 1997, pp. 337–340.

[12] K. Prall, "Benchmarking and metrics for emerging memory," in *Proc. IEEE Int. Memory Workshop (IMW)*, May 2017, pp. 1–5.

[13] M. Feiginov, "Frequency limitations of resonant-tunnelling diodes in sub-THz and THz oscillators and detectors," *J. Infr., Millim., Terahertz Waves*, vol. 40, pp. 365–394, Mar. 2019.

[14] R. H. Dennard, F. H. Gaensslen, H.-N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, Oct. 1974, doi: 10.1109/JSSC.1974.1050511.

[15] Micron Technology. (2008). *TN-29-42: Wear-Leveling Techniques in NAND Flash Devices Introduction*. [Online]. Available: https://www.micron.com/-/media/client/global/documents/products/technical-note/nand-flash/tn2942_nand_wear_leveling.pdf

[16] *ULTRA**RAM**: Towards the Development of a III-V Semiconductor, Non-Volatile, Random-Access Memory Dataset*. Accessed: Mar. 15, 2021. [Online]. Available: https://www.10.17635/lancaster/researchdata/420