ULTRA**RAM**™: towards the development of a III-V semiconductor, non-volatile, random-access memory

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

Abstract—ULTRA**RAM**™ is a III-V compound semiconductor memory concept which exploits quantum resonant tunneling to achieve non-volatility 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 $8x10^4$ s, whereby the 0/1 states show adequate contrast throughout, whilst performing $8x10^4$ 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— Resonant tunneling, NVRAM, NVM, memory, InAs/AlSb.

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**™ is a novel, III-V compound-semiconductor memory which 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 non-volatility. 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 a 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 10 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-tunnelling 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 (mA) 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-inch highly doped (Si, n ~ $2x10^{18}$ cm^{-3}) GaAs wafers by molecular beam epitaxy (MBE) on a Veeco GENxplor system. The 7.8% GaSb/GaAs lattice mismatch was mitigated by use of an interfacial misfit array between the substrate and GaSb buffer layer [10] before growth of the

D. Lane is with the Department of Physics, Lancaster University, Lancaster LA1 4YB, United Kingdom (e-mail: d.lane@lancaster.ac.uk).

P. D. Hodgson is with the Department of Physics, Lancaster University, Lancaster LA1 4YB, United Kingdom (e-mail: p.hodgson1@lancaster.ac.uk).

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

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

M. Hayne is with the Department of Physics, Lancaster University, Lancaster LA1 4YB, United Kingdom (e-mail: m.hayne@lancaster.ac.uk).

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

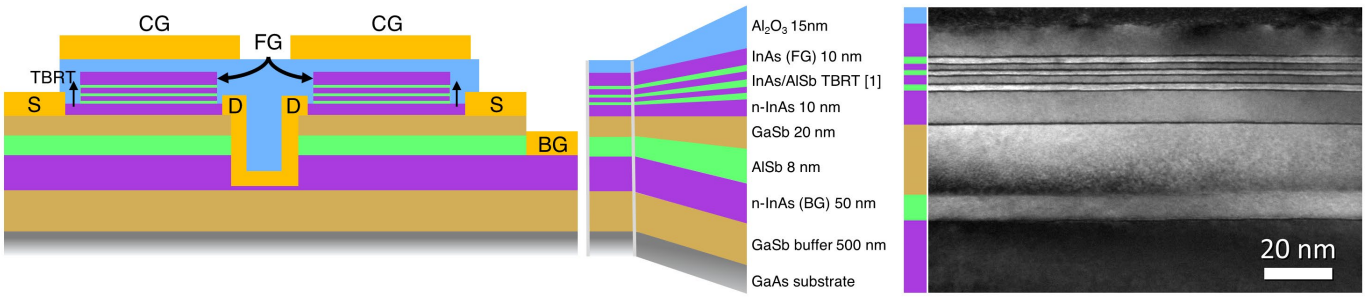

Fig. 1. Left: cross-sectional schematic of ULTRA**RAM**™ and material layers. Right: dark field g = 002 TEM image of the epitaxial structure. Layer thicknesses for the TBRT region can be found in Table 1.

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].

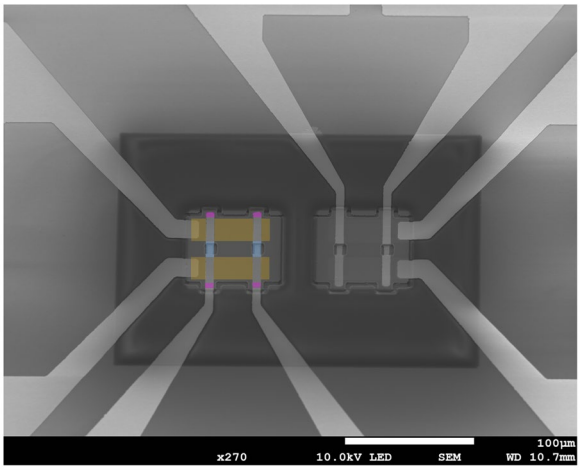

Fig. 2. Scanning electron microscope image of two ULTRA**RAM**™ 2x2 memory arrays. The false colouring on the array pictured to the left indicates the extent of the WLs (yellow) and the etched access for the BL contacts to the S terminals (pink). The buried back-to-back D contacts, which are connected to the BG and isolated from the BL, are in the center of each pair of devices (blue).

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 1.

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 control gate (CG) layers being added. This was followed by further SiO_2 passivation via plasma-enhanced chemical vapour deposition. Lastly, 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.

III. LOW VOLTAGE P/E

Fig. 3 presents the current flow from S to D during a 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

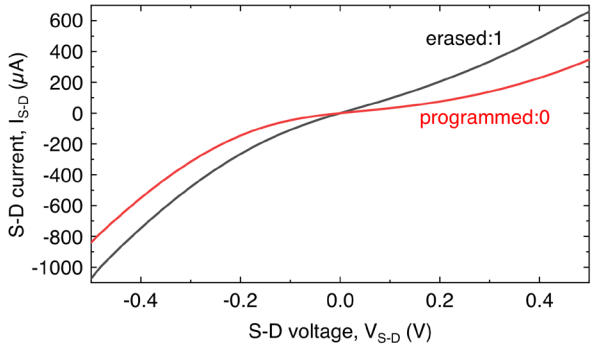

Fig. 3. Sweep of S-D voltage with measured S-D current after an erase cycle of 2.5 V, 500 µs duration (black line) and a program cycle of -2.5 V, 500 µs duration.

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 Ω to $>10^{10}$ Ω (the limit of measurement). Within the array architecture S-D current ($I_{S\text{-}D}$) 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 tunnelling is in the sub-ps scale [13], the switching speed is limited by the RC time constant, and is therefore subject to Dennard's scaling law (scaling linearly with area) [14]. Thus, for a 20-nm gate length device with ideal scaling sub-ns switching speed is predicted – significantly faster than DRAM and comparable to static (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 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 is not sufficient to reliably test device-to-device switching in the array formation. While the insufficient contrast is a 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.

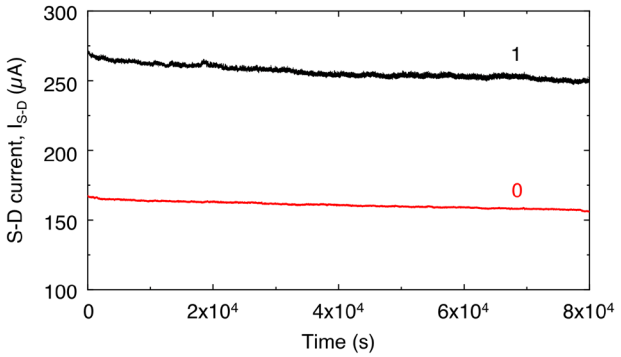

Fig. 4. Retention data for a 20 µm gate length cell in a 2×2 memory array. Memory logic is programmed (0) and erased (1) with 500 µs pulses of -2.5 V and +2.5 V on the S terminal respectively. Read-out current is measured with a 0.5 V source voltage every second.

IV. Reliability

Fig. 4 demonstrates the stability of the memory logic over time. Both programmed 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 non-volatility of the memory is a result of the 2.1 eV barriers from the InAs/AlSb heterostructures on 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-levelling is required to prolong their lifetime [15]. Wear leveling is unsuitable for RAM, which requires superior endurance properties, with individual cells being programmed 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 $I_{S\text{-}D}$ offset aside, a memory is realized for over 10^6 cycles (Fig. 5). Moreover, the difference in current

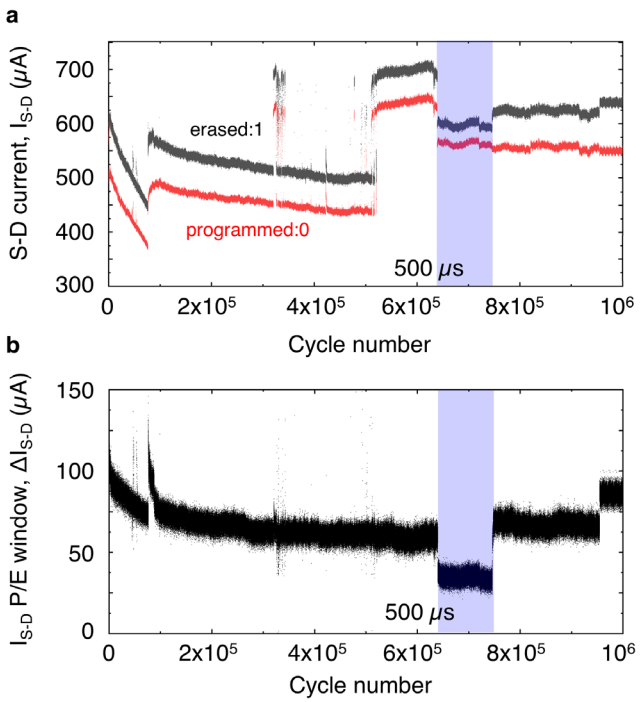

Fig. 5. Endurance data for an ULTRA**RAM**™ memory cell in an array. **a.** S-D current after a +2.5 V erase cycle (grey), and a -2.5 V program cycle (red). Pulse duration was set to 5 ms, except for those data points with blue shading where a 500-μs pulse duration was used. **b.** S-D current difference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

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 programmed and erased at extremely low switching energy (per unit area) using a half-voltage architecture in which the P/E voltage is split between bit line (S) and word line (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-ns 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, ultra-efficient, non-volatile, random-access ULTRA**RAM**™ memories are a real possibility.

Acknowledgment

This work was supported by the EPSRC, UK, via the 2017-2020 IAA funding allocation to Lancaster University under grant EP/R511560/1, a scholarship under grant EP/N509504/1 and the Future Compound Semiconductor Manufacturing Hub grant EP/P006973/, by the ATTRACT project funded by the EC under Grant Agreement 777222 and by the Joy Welch Educational Charitable Trust.

References

[1] H.S. Wong and S. Salahuddin, “Memory leads the way to better computing,” in *Nat. Nanotechnol*., vol. 10, pp. 191–194, 2015.

[2] S. Yu and P. Chen, "Emerging Memory Technologies: Recent Trends and Prospects," in *IEEE Solid-State Circuits Magazine,* vol. 8, no. 2, pp. 43-56, 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,” in *J. of App. Phys*., 89(11) pp. 5815-5875, 2001.

[4] M. Hayne, “Electronic memory devices.” US patent US10243086B2.

[5] D. Lane and M. Hayne, "Simulations of Ultralow-Power Nonvolatile Cells for Random-Access Memory," in *IEEE Trans. on Electron Devices*, vol. 67, no. 2, pp. 474-480, 2020, doi: 10.1109/TED.2019.2957037.

[6] 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,” in *Scientific Reports* vol. 9, 8950, 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" *in 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/CaF2/CdF2/CaF2/Si resonant-tunneling quantum-well structures” in *Appl. Phys. Express*, vol. 7, p. 044103, 2014.

[9] H.-Y. Chou, E. O'Connor, A. O'Mahony, I. M. Povey, P. K. Hurley, Lin Dong, P. D. Ye, V. V. Afanas'ev, M. Houssa and A. Stesmans, “Band offsets and trap-related electron transitions at interfaces of (100)InAs with atomic-layer deposited Al_2O_3,” in *J. Appl. Phys*., vol. 120, 235701, 2016, doi: https://doi.org/10.1063/1.4971178.

[10] A.P. Craig, P.J. Carrington, H. Liu, H. and A.R.J. Marshall, “Characterization of 6.1Å III–V materials grown on GaAs and Si: A comparison of GaSb/GaAs epitaxy and GaSb/AlSb/Si epitaxy,” in J. Cryst. Growth, vol. 435, pp. 56–61, 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 *Proceedings of the IEEE Twenty-Fourth International Symposium on Compound Semiconductor*s, pp. 337-340, 1997.

[12] K. Prall, “Benchmarking and Metrics for Emerging Memory,” in *2017 IEEE International Memory Workshop*, Monterey, pp. 1-5, 2017.

[13] M. Feiginov, “Frequency Limitations of Resonant-Tunnelling Diodes in Sub-THz and THz Oscillators and Detector,” in *J Infrared Milli Terahz Waves* 40, 365–394, 2019.

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

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

[16] doi:10.17635/lancaster/researchdata/420