nature communications

Article

https://doi.org/10.1038/s41467-023-41363-x

Simultaneously ultrafast and robust two-dimensional flash memory devices based on phase-engineered edge contacts

Received: 25 May 2023

Accepted: 31 August 2023

Published online: 13 September 2023

Jun Yu[1], Han Wang[1], Fuwei Zhuge[1], Zirui Chen[2], Man Hu[1], Xiang Xu[1], Yuhui He[2], Ying Ma[1], Xiangshui Miao[2] & Tianyou Zhai[1]

As the prevailing non-volatile memory (NVM), flash memory offers mass data storage at high integration density and low cost. However, due to the 'speed-retention-endurance' dilemma, their typical speed is limited to ~microseconds to milliseconds for program and erase operations, restricting their application in scenarios with high-speed data throughput. Here, by adopting metallic 1T-Li_xMoS_2 as edge contact, we show that ultrafast (10–100 ns) and robust (endurance>10^6 cycles, retention>10 years) memory operation can be simultaneously achieved in a two-dimensional van der Waals heterostructure flash memory with 2H-MoS_2 as semiconductor channel. We attribute the superior performance to the gate tunable Schottky barrier at the edge contact, which can facilitate hot carrier injection to the semiconductor channel and subsequent tunneling when compared to a conventional top contact with high density of defects at the metal interface. Our results suggest that contact engineering can become a strategy to further improve the performance of 2D flash memory devices and meet the increasing demands of high speed and reliable data storage.

In the era of digital technology, massive data awaiting processing and storing keeps calling for high-density, ultrafast, and robust memory technology[1–3]. In the last decade, as the elemental device in solid state disk, flash memory had evolved markedly to fulfill the requirement of high-volume density by moving from planar configuration to 3D packing[4,5]. However, for decades, the operation speed for the prevailing flash memory products in market barely improved due to the dilemma of 'speed-retention-endurance' in optimizing their performance. Increasing the operation speed of flash memory, e.g., by large operation voltage, typically sacrifices its endurance and retention[6]. This is because even though charge injection efficiency for program/erase could be theoretically improved by high electric field, the cell tends to fail early due to the accelerated defect generation and growth in tunneling dielectric[7,8]. Unintentional interfacial roughness in device also focuses local electric field and adversely degrades cell lifetime[9]. Though channel or source side hot carrier injection had been successfully exploited to enhance the charge injection efficiency without using high voltage, fast operation speed ~sub-microsecond is only achieved for program, and the erase speed is still slow ~milliseconds[10,11].

Ideal Schottky contacts had been predicted with the ability to inject hot carriers to semiconductor channel under gate bias modulation, and hold the potential to dramatically boost the performance of flash memory[12]. However, in conventional Si technology, achieving an abrupt Schottky contact is challenging due to issues related to interdiffusion and Fermi level pinning (FLP)[13–15]. With their atomically flat layer structure, the emerging two-dimensional van der Waals (2D vdW) materials have numerous ways to engineer an ideal Schottky barrier, including using phase-engineered or vdW contact[16,17]. They could also be integrated at high packing density yet with low defect density and

[1]State Key Laboratory of Materials Processing and Die and Mould Technology, School of Material Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China. [2]Hubei Yangtze Memory Laboratory; School of Integrated circuits, Huazhong University of Science and Technology, Wuhan 430074, China. e-mail: zhugefw@hust.edu.cn; yingma@hust.edu.cn; zhaity@hust.edu.cn

https://doi.org/10.1038/s41467-023-41363-x

clean interfaces[18,19], which are desirable for ultrafast and robust memory operation by providing an ideal charge injection interface and durable tunneling layer[20]. However, the performance of 2D flash memory in previous literatures falls behind of the expectation in spite of various channel and float gate used[21–27]. Until recently, 20–160 ns superior operation speed was achieved in InSe and MoS_2 flash memories based on the clean interface in vdW heterostructures or hot carrier injection directly though the ultrathin 2D material[28,29]. However, a competing endurance lifetime to the prevailing Si flash technology (>10^5 cycles)[30,31] was rarely demonstrated for a simultaneously ultrafast and robust flash memory. To this end, a recent example that adopts bipolar WSe_2 as the channel displays the potential in achieving well-balance memory performances by combine the Lucky-electron injection mechanism, with an oxide charge trapping layer structure for endurance enhancement[32]. On the other hand, though an ideal Schottky contact could theoretically enhance charge injection efficiency in a flash memory, the vital role of contact interface in leveraging memory performance had been overlooked in the past, especially considering that the ultrathin two-dimensional crystal lattice is extremely sensitive to direct metal deposition in the conventional top contact configuration.

Here, we demonstrate the realization of simultaneously ultrafast (program/erase speed ~10/100 ns) and super-robust (endurance lifetime >10^6 cycles) based on phase-engineered edge contacts to 2D MoS_2 flash memory. If compared to traditional top metal contacts that have rich lattice and electronic defects due to metal-induced gap states (MIGS) or interdiffusion impurities[16], lateral edge contacts exhibit highly tunable Schottky barrier and render efficient hot carrier injection into the semiconductor channel during program/erase (P/E) operation. This markedly improves the charge injection efficiency to float gate and guarantees ultrafast and super-robust memory operation at the same time, which was been rarely reported in 2D flash memory. The comprehensively optimized key figure of merits over the prevailing commercial flash memory makes our edge-contacted 2D flash memory a viable option for high-speed and durable memory in the future.

Results

Ultrafast P/E speed in memory cells with edge contact

Figure 1a illustrates the structure of our flash memory made of a vdW stacking of MoS_2/hBN/few-layer graphene (FLG) on top of a thermal oxidized Si substrate. The phase-engineered edge contact is made by

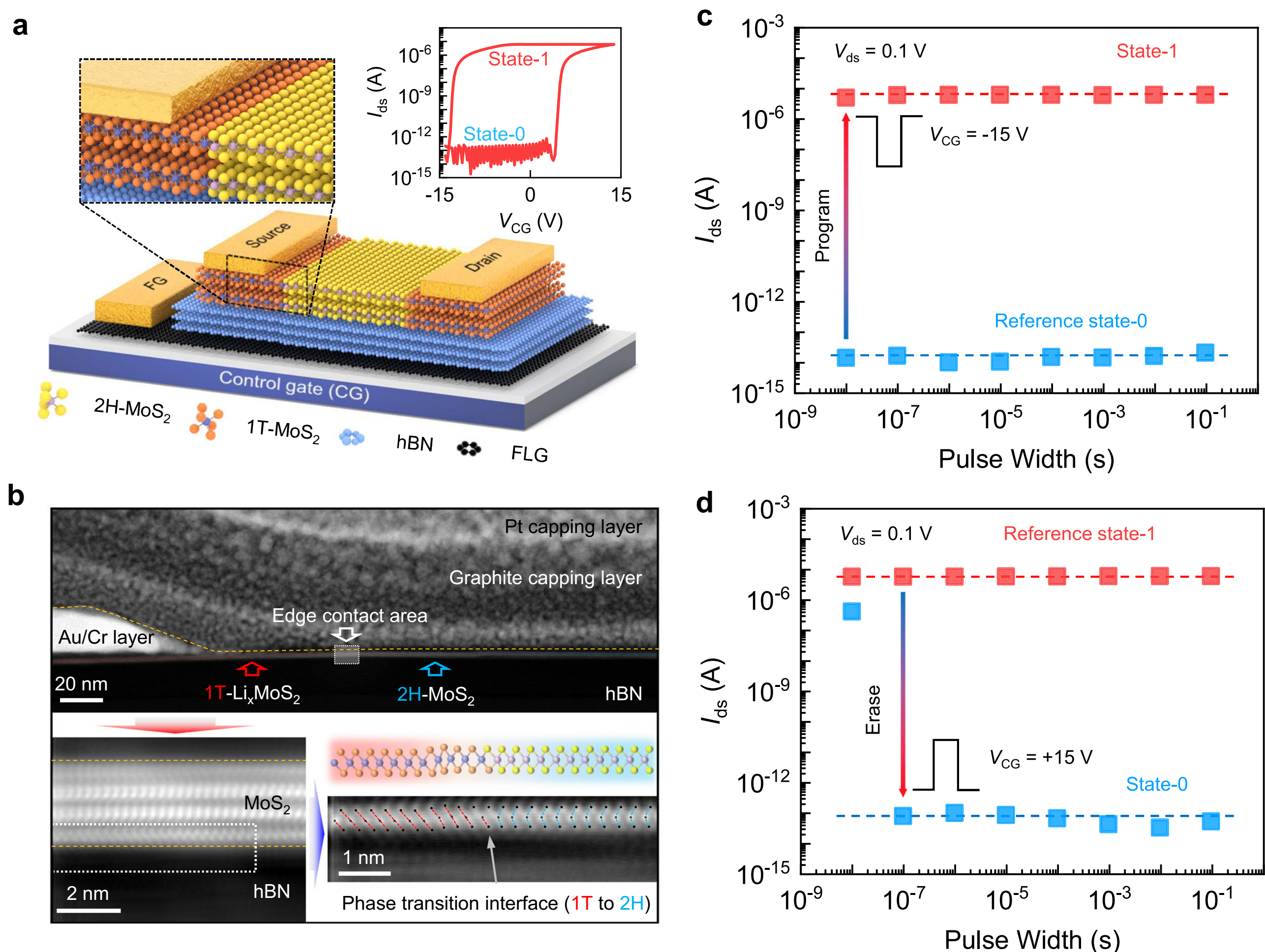

Fig. 1 | Ultrafast MoS_2 flash memory with phase-engineered edge contact. **a** Illustration of the flash memory based on vdW heterostructure of MoS_2/hBN/few-layer graphene (FLG), the edge contact is formed by patterned 1T-Li_xMoS_2 under conventional Cr metal contact. The inset hysteresis loop shows clearly the switch of readout current (I_{ds}) between state-0 and state-1 via sweeping the control gate bias (V_{CG}). **b** Atomic-scale observation of the edge contact in MoS_2, the high-angle annular dark field scanning transmission electron microscopy (HAADF-STEM) image validates the phase transformation of 2H-MoS_2 into a distorted 1 T phase. The atomic arrangement of S-Mo-S in the bottom layer is marked using red and blue dash lines to highlight the 1 T and 2H phase transition interface. **c**, **d** reveal the program and erase performance when varying the width (10 ns to 100 ms) for applied V_{CG} pulse (−15 V for program, and 15 V for erase). The reference states (state-1 in **c** and state-0 in **d**) were set by initializing voltage pulses with 100 ms width. Red and blue dash lines are guidelines that indicate the level of ON and OFF states by a successful program/erase operation.

patterned lithium intercalation in n-butyl lithium solution[33,34], which transforms the semiconductor MoS_2 (2H) layers into metallic phase (1T-Li_xMoS_2) (see method and Supplementary Note 1 for fabrication details). Cr/Au contact is followingly made for electrical connection to the transformed metallic phase. The typical thickness of MoS_2 and hBN in the prepared memory cells is 3–5 nm and 11–12 nm (Supplementary Note 2). Using high-angle annular dark field scanning transmission electron microscopy (HAADF-STEM) image (Fig. 1b), the vdW interfaces and edge contact via phase transition in the fabricated heterostructure are confirmed. A transition from 2H-MoS_2 in channel to a distorted 1T-MoS_2 structure could be identified near the contact pad, forming the lateral edge contact. Due to the interlayer diffusion of lithium, the 1T-2H interface extrudes slightly from the lithography patterned area in to the channel, and has slight structure distortion in plane due to intercalation-induced strain. In our design, the phase transformation in MoS_2 is ensured to reach the bottom layer from the top surface, which avoids intercalation at the hetero-interface, e.g., the one between MoS_2 and hBN. The extinction of photoluminescence spectra of 2H-MoS_2 is used in experiment to verify the degree of phase transformation (Supplementary Note 3). When sweeping the voltage bias applied to control gate (CG), the device displays an apparent memory window in the inset of Fig. 1a, which stem from charge trapping in floating gate (FG) rather than hBN or SiO_2 dielectric layer (Supplementary Note 4).

Using ultrafast electric pulses applied to control gate, we successfully program/erase the above memory cell within 10/100 ns (pulse waveform discussed in Supplementary Note 5). As indicated in Fig. 1c, when starting from a reference OFF state (state-0, erased by $V_{CG} = 15$ V for 100 ms), the memory cell is fully programmed into state-1 by 10 ns pulse at $V_{CG} = -15$ V, reaching a high ON/OFF ratio ~10^7. At a lower voltage of -7V, the memory was still successfully programmed with an on/off ratio>10^4. The significantly lower operation voltage than previous reports suggested high charge injection efficiency in present memory cell, which is vital for later robust endurance behavior. Reversely, when starting from a reference ON state (state-1, programmed by $V_{CG} = -15$ V for 100 ms), the memory cell is fully erased into state-0 within 100 ns ($V_{CG} = 15$ V). If compared to the prevailing silicon flash, the above ultrafast P/E speed is markedly improved by 2–4 orders respectively. According to the tunneling barrier for electrons ($\Phi^{e}_{tB} = 2.8$eV, for erase) and holes ($\Phi^{h}_{tB} = 2.0$eV, for program) from MoS_2 conduction band (CB) and valence band (VB) through hBN (Supplementary Note 6), such ultrafast P/E operation speed could theoretically be achieved considering a strong electric field in hBN ($E_{hBN} > 10$ MVcm^{-1}) during the applied voltage pulse (Supplementary Note 7), which is met in our device based on its high gate coupling ratio ~0.9 (Supplementary Note 8). However, it is worth noting that the high operation voltage itself does not guarantee high P/E speed in practical devices, and the edge contact by 1T-Li_xMoS_2 is considered as the other key factor in improving the charge injection efficiency for ultrafast operation.

To validate the critical role of edge contact in enhancing the operation speed, we fabricated paired devices on the same MoS_2/hBN/graphene vdW heterostructure using edge contact (1 T) and conventional top face contact (Cr) respectively, as indicated in Fig. 2a and b. The paired devices have identical thickness for each constituting layers (MoS_2, 4.2 nm; hBN, 15.8 nm; Gr, 2.8 nm, Supplementary Note 9) and gate coupling ratio (GCR), therefore comparable electrical field strength in tunneling hBN layer. In Fig. 2c, d, their P/E performance is compared directly by varying both the amplitude and width of applied P/E pulses. For each measurement, the memory was initially set to a saturated reference state-0 or state-1 using 100 ms electric pulses. Despite of additional etching process in fabricating paired memory cells, the obtained edge-contacted memory cell have identical P/E characteristic to the one from direct vdW stacking (Supplementary Note 10). If taking an on/off ratio of 10^4 as the criteria for successful P/E operation, the edge contacted cell has 2–3 orders faster P/E speed than top contacted memory cell at the same operation voltage, which clearly indicated the superior charge injection efficiency offered by edge contact. In the laboratory, we have fabricated and evaluated more than 49 memory cells (26 cells with 1 T edge, and 23 cells with Cr top contact) to compare their P/E performances. The thickness of hBN layer in these memory cells distributed within the range of 8–17 nm with an average value of ~11 and ~12 nm for top and edge contacted memory cells (Supplementary Note 2). In Fig. 2e, f, the P/E speed of all memory cells is summarized by the attained ON/OFF ratio (relative to the reference state) after applying 10/100 ns P/E pulses of varying amplitudes. Despite the device-to-device (D2D) variation by fabrication, statistically, we could find that replacing Cr contact using 1T-Li_xMoS_2 markedly improves the yield of high-speed memory cells. According to the cumulative analysis, >90% memory cells with edge contact can be successfully programmed (defined as when an acceptable ON/OFF ratio ≥10^2 is attained) within 10 ns at the normalized operation voltage of 14 V for 10 nm hBN layer (V_{norm} = $V_{pulse}/t_{hBN} \times 10$ nm, where t_{hBN} represents the thickness of hBN layer in device). Comparatively, only ~40% is achieved for top-contacted memory cells. For erase, the statistical yield at 100 ns is also improved from 20% to 40%. The lower rate for erase is related to the large tunneling barrier through hBN for erase, and may be improved via designing the band alignment in vdW heterostructure for a more balanced P/E performance.

Edge contact facilitated charge tunneling injection to float gate

Hot carrier injection had been known occur through Schottky barrier in metal contacted field-effect transistors, and had been exploited to boost the program speed of float gate memory to microseconds[11,35]. Though almost all 2D float gate transistors are made with Schottky contacts, fast operation speed was not always guaranteed[21–27]. In Fig. 3a, b, we illustrate the tunneling pathways for edge and top contacted memory cells and the related potential barrier for hole tunneling, which is responsible for program in our memory (Supplementary Note 6). Previously, hole tunneling had been discussed responsible for the operation of 2D float gate transistors that use MoS_2 as the channel and graphene as float gate[36]. In our case, the dramatic enhancement of P/E speed via contact engineering to MoS_2 suggest that electron/hole tunneling from MoS_2 to graphene is most likely responsible for the memory operation. Here, hole tunneling is used for illustration considering the apparent improvement to program speed if compared to erase. Although MoS_2 displays n-type conduction with electrons as the majority carriers, the high field modulation under positive gate bias during program is sufficient to induce significant amount of holes at the level of 10^{13} cm^{-2} if considering a field strength of 10 MV/cm across hBN.

Recent studies on MoS_2 memory cells have shown that vertical charge tunneling from metal contact to bottom 2D semiconductor is critical to generate hot carriers for ultrafast memory operation[29]. However, in our case, the measured tunneling current density via 1T-MoS_2/hBN/Gr pathway is ~2 order lower than that through 2H-MoS_2 due to the higher tunneling barrier from the Fermi level of 1T-MoS_2 (Supplementary Note 11). By comparing directly the tunneling current across edge contacted 2H-MoS_2/hBN/graphene heterostructure, we found ~2 orders enhancement to conventional top contacted one at the electric field of 8 MV/cm (Supplementary Note 11). This suggest that the edge contact interface plays a critical role in enhancing the charge injection efficiency. In addition, replacing graphene with MoS_2, a symmetric tunneling structure of MoS_2/hBN/ MoS_2 with respective edge and top contact to channel and float gate displays highly asymmetric P/E speed, showing an ultrafast program within 10 ns but slow erase >100 ms (Supplementary Note 12). This again confirms the adoption of edge contact to MoS_2 channel contributes to the observed difference in charge injection efficiency. Considering the edge contact

https://doi.org/10.1038/s41467-023-41363-x

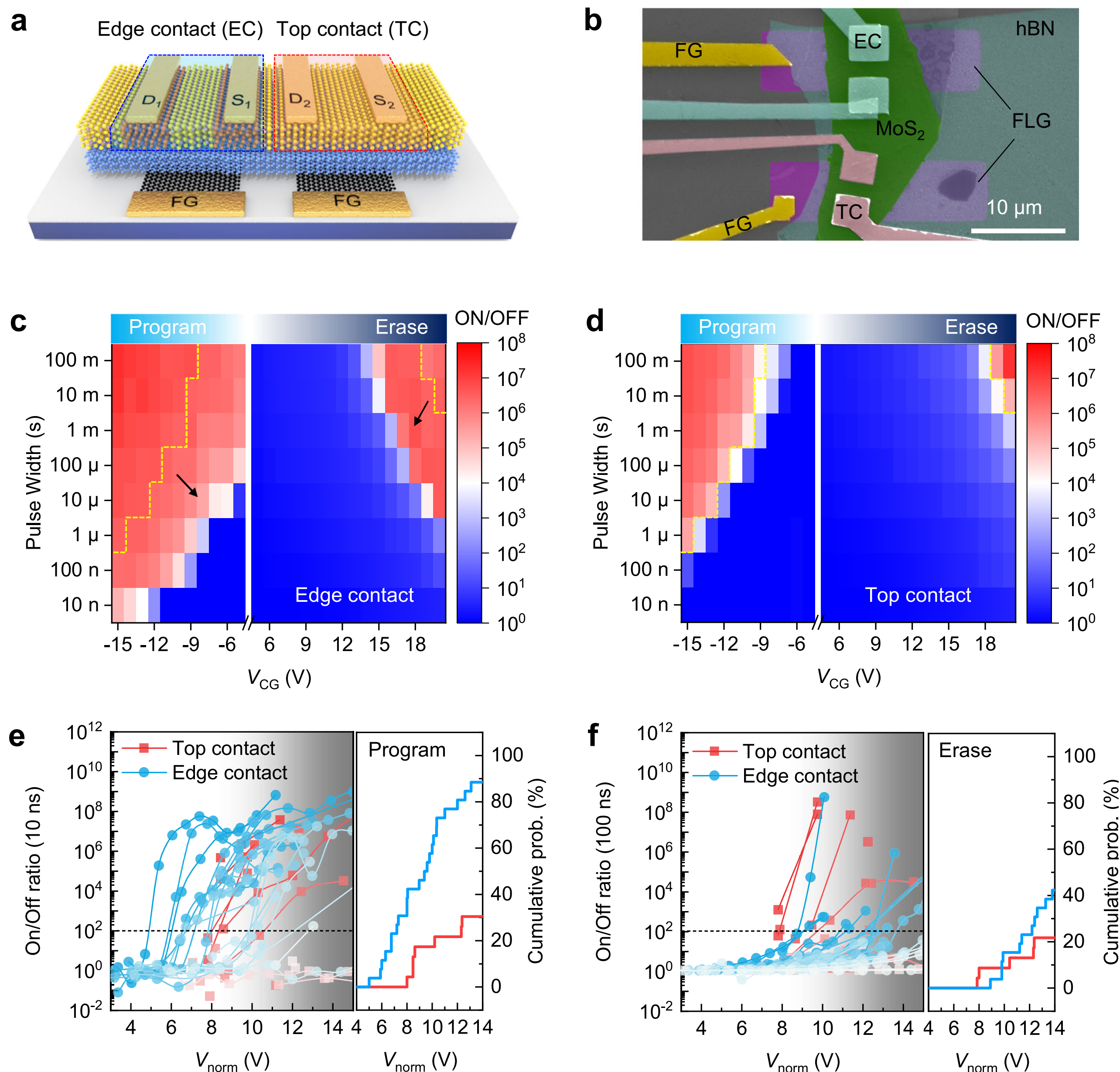

Fig. 2 | Comparison of the program/erase (P/E) performance of memory cells with 1 T edge or Cr top contacts. a, **b** Schematic illustration (**a**) and false color scanning electron microscopy (SEM) image (**b**) of the paired memory cells on the same vdW heterostructure. The float gate (FG) is made by etching an exfoliated few-layer graphene. Thus, the paired memory cells exhibit identical thickness combination for each layer and differ only in contact configuration using 1 T edge (EC) or Cr top contact (TC). **c**, **d** Map of the attained ON/OFF ratio of memory under different voltage pulse conditions when changing both the amplitude and pulse width: 1 T edge contact (**c**) and Cr top contact (**d**). The P/E condition that guarrantees a high ON/OFF ratio = 10^4 in top contacted memory cell is marked in both figures for a guidline (dash line) when comparing their operation speed. **e**, **f** P/E behavior (**e**, **f**) of >20 memory cells under ultrafast electric pulse (10 ns for program, 100 ns for erase) and varying operation voltage. The operation voltage was normalized according to the thickness of tunneling hBN to reflect the electric field strength at the tunneling layer. On the right panel of **e**, **f**, the cumulative probability is counted if an ON/OFF ratio $\geq 10^2$ is attained by the applied ultrafast P/E pulses.

configuraiton, we expect that instead of tunneling from contact region, the charge injection in edge contacted memory cell tends to initiate via a lateral pathway, which crosses the phase change interface from 1 T to 2H-MoS_2. In this case, the band bending (Φ_s) in 2H-MoS_2 accelerates the injected carriers from Schottky contact and efficiently lowers the tunneling barrier through hBN layer via hot carrier effect ($\Phi^h_{tB} = \Phi^h_{tB0} - \Phi_s$). In comparison, for the case of conventional top contact, the Fermi level in MoS_2 tends to be pinned by the trap states under contact affected (CA) area, which reduces Φ_s under gate modulation[16,37]. By transforming the CA area in to a metallic phase, the present 1 T contact extends from the defined top contact area by lithium diffusion in interlayer space, thus making the charge injection via lateral Schottky junction immune to manufacturing defects.

The low trap density in edge contacted memory cell is confirmed by the frequency dependence of the capacitance-voltage (C-V) characteristic (Fig. 3c), which is measured between the MoS_2 channel and float gate. Within the frequency range of 4–100 kHz, the measured capacitance (C_p) for edge contacted memory cell displays slight frequency dependence, while C_p for Cr top contacted memory cell reduces apparently with increasing the frequency, especially when V_{FG} is above the flat band voltage (V_{FB}). Using the high-low frequency method, the interface trap density was determined according to

https://doi.org/10.1038/s41467-023-41363-x

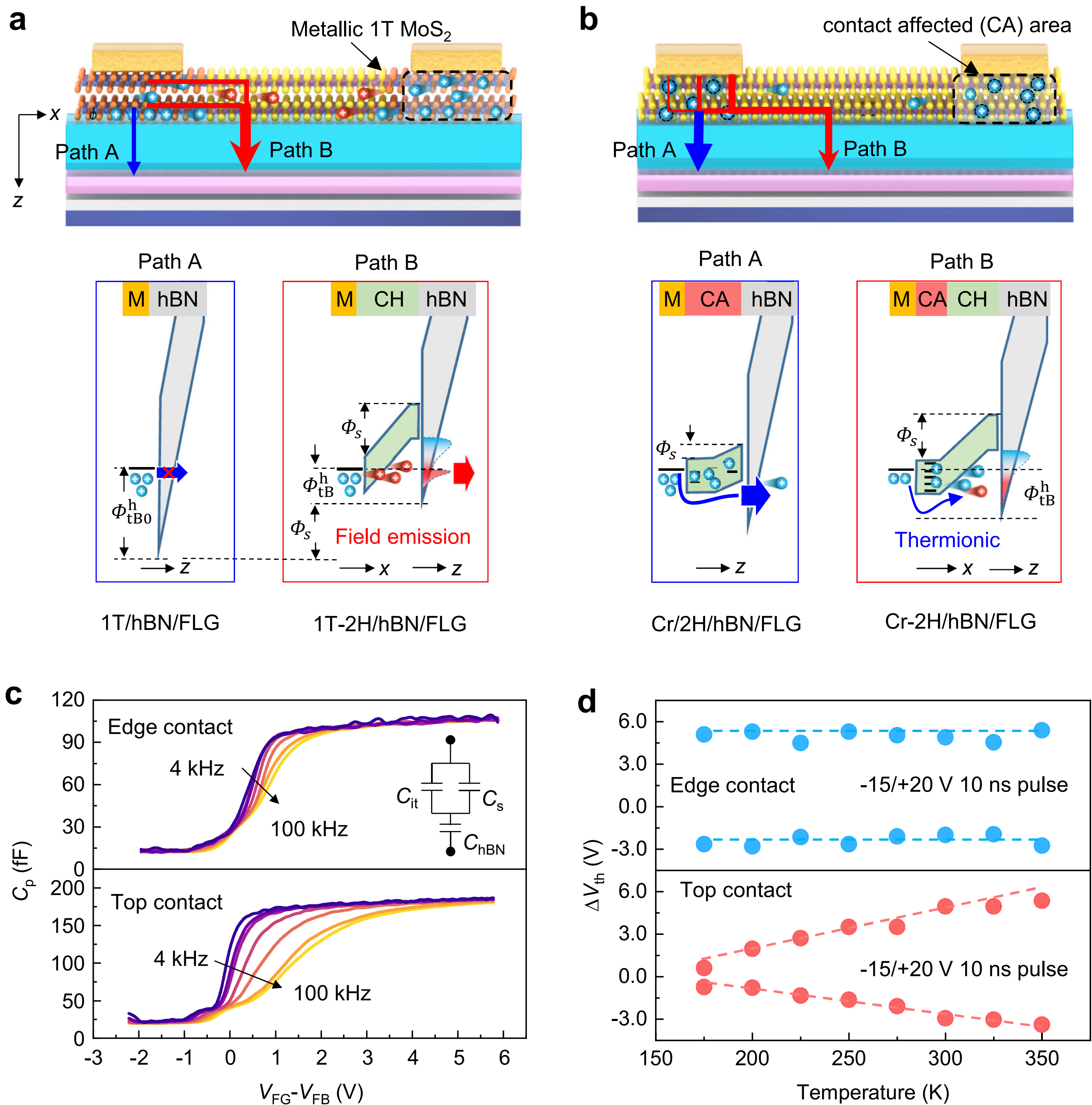

Fig. 3 | The tunneling pathway and charge injection behavior for edge or top contacted flash memory. Illustration of the energy band diagram for charge injection during program operation in **a** edge and **b** top contacted memory cells, the tunneling pathways through contact region and semiconductor channel are compared (M: metal contact, CH: channel, CA: contact affected area). When considering the band bending in semiconductor channel (Φ_s) under field modulation, the hole tunneling barrier (Φ^{h}_{tB}) though the valance band of hBN is apparently reduced from the initial value (Φ^{h}_{tB0}). If compared to the field emission (indicated by red arrow) in edge contact, rich trap states under contact affected area (CA) in top contact configuration lead to thermionic process governed charge emission at contact (blue arrow). **c** Capacitance-voltage characteristic and **d** temperature-dependent charge injection behavior of edge and top contacted memory cell. The measured capacitance is intepreated using the inset equalivent circuit, which considers capacitance associated with interface states (C_{it}), semiconductor channel (C_s), and hBN dielectric layer (C_{hBN}). **d** The charge injection efficiency is reflected by the threshold shift under ultrafast (10 ns) P/E operation, in which the pulse amplitude are respectively −15 V and 20 V for program and erase. Their distinct temperature dependence is indicated by the dash guidelines.

$qD_{it} = \left[\left(\frac{1}{C_{LF}} - \frac{1}{C_{hBN}}\right)^{-1} - \left(\frac{1}{C_{HF}} - \frac{1}{C_{hBN}}\right)^{-1}\right]$, in which C_{LF} and C_{HF} are respectively the areal capacitance at low and high-frequency limit, and q the elementary charge[38]. For top and edge contacted memory cell, the integrated trap density was ~$2.2 \times 10^{13}\,\mathrm{cm}^{-2}$ and $8 \times 10^{11}\,\mathrm{cm}^{-2}$ (Supplementary Note 13), respectively. The high trap density in conventional top contacted memory cells may come from both electronic traps generated from electron-beam lithography and also lattice disorders caused by depositing top metal contact (Supplementary Note 14). Since the response time of these trap states lag significantly behind of the applied 10–100 ns pulse duration for ultrafast P/E operation, they act as fixed space charges that increase the potential barrier for charge injection[39]. The additional capacitance related to interfacial trap states also tend to reduce the potential drop in MoS_2,

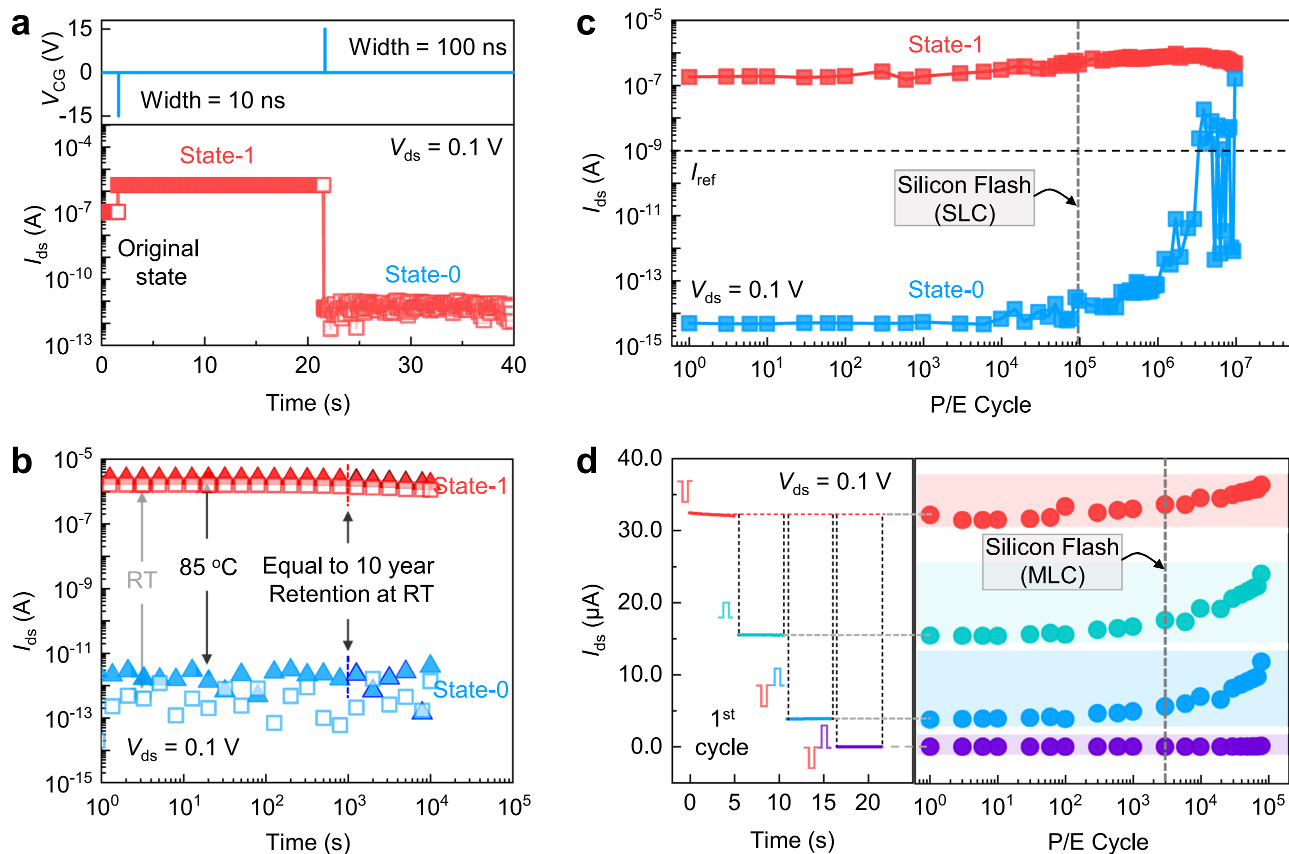

Fig. 4 | Ultrafast operation speed, long-term retention, and robust endurance characteristics simultaneously attained in edge-contacted MoS_2 flash memory. **a** Monitored change of memory state during sequentially applied ultrafast P/E pulses. **b** Retention performance of the memory under RT (open square), and 85 °C (filled triangle) acceleration, with an acceleration ratio of ~3.3×10^5 at 85 °C, 10-year data retention is expected at the vertical dash line. **c** Endurance characteristic of the memory as single-level cell (SLC). The programming and erasing operations are performed by (−8 V, 10 ns) and (15 V, 100 ns) operation pulses, respectively. Memory failure is confirmed if the readout at erase state become higher than the defined I_{ref} = 1 nA, indicated as horitontal dash line. **d** Endurance characteristic of the memory as multi-level cell (MLC). Four memory states marked in shaded area is achieved during the repeated P/E operation. As SLC and MLC, the endurance lifetime is ~3×10^6 cycles and 8×10^4 cycles respectively. **c**, **d** Vertical dash lines that indicate the commercial standard of endurance lifetime for Silicon flash (10^5 cycles as SLC, and 10^3 cycles as MLC) are included for comparison.

thus counteracts the barrier lowering for hot carrier injection. In present float gate memory cells, the reduced trap density in edge contact renders highly tunable Schottky barrier under gate modulation (Supplementary Note 15), which is essential for hot carrier injection via Schottky emission at the metal-semiconductor interface. Consistently, when the edge contact is placed away from float gate coupling (Supplementary Note 9), the absence of Schottky barrier modulation at the metal contact interface during memory operation deteriorates apparently the charge injection efficiency by edge contact. This suggests that gate coupling to edge contact interface is crucial for achieving superior charge injection efficiency.

The essential role of trap effects in impeding ultrafast memory operation is further supported by the distinct temperature effect to memory operation in the paired memory cells. Figure 3d displays the extracted threshold voltage shift (ΔV_{th}, read at I_{ds} = 20 μA with V_{ds} = 1 V) of memory at the temperature range from 100 to 300 K after ultrafast P/E operation. By fixing the P/E pulses (P: −15 V, 10 ns; E: 15 V, 100 ns), ΔV_{th} reflects directly the charge injection efficiency in memory. For edge contacted memory cell, ΔV_{th} is found independent of T, which agrees with the behavior of FN tunneling limited charge injection that is weakly influenced by T[40]. After a short program pulse, the measured ΔV_{th} reaches as high as 7–9 V for all temperature, corresponding to a high charge density of ~5.7×10^{11} cm^{-2} in float gate. In comparison, the paired Cr contacted memory cell displayed slow operation speed and apparent T dependence of ΔV_{th}, which suggested a thermal activation behavior for charge injection. An Arrhenius fitting to the result yields shallow activation energy of 58 meV and 40 meV for trapped charges during program and erase operation respectively[36].

Robust endurance of edge contacted flash memory

The significantly improved charge injection efficiency in the above via edge contact essentially allows us to realize ultrafast P/E performance without sacrificing retention or endurance characteristics. Figure 4a, b displays respectively the ultrafast P/E performance and long data retention characteristics of an edge-contacted memory cell. The memory states are rapidly switched with a superior ON/OFF ratio >10^7 via 10 ns program (−15 V) and 100 ns erase (15 V) pulses. Notably, the obtained state-0 and state-1 in above memory display retention >10^4 s at both room temperature (RT) and 85°C acceleration (Fig. 4b). According to a leakage barrier of ~1.95 eV and related acceleration ratio (AR) of 3.3×10^5 (Supplementary Note 16), this value suggests an equivalent data retention lifetime >10 years (~3.3×10^8 s) at RT, which stems from the excellent charge trapping ability of graphene and low defect states in hBN[22,28,29].

In Fig. 4c, the endurance behavior of an edge-contacted memory cell as a single-level cell (SLC) is revealed. The memory state is switched between state-0 and state-1 by P/E pulses of (−8 V, 10 ns) and (15 V, 100 ns). Notably, if compared to previously reported ultrafast InSe (20 V, 20 ns) and MoS_2 (30 V, 20 ns) flash memory cells that adopted top metal contacts[28,29], our memory is operated at

considerably lower voltage based on its superior charge injection efficiency. This endows the memory cell an exceeding endurance lifetime ~3×10^6 cycles (Fig. 4c). The barely drifted readout current for state-0 and state-1 before 10^6 cycles suggest well-suppressed stress in hBN layer. As discussed in Supplementary Note 17, the memory after 10^5 endurance cycles still keeps long data retention over years. Comparatively, top contacted memory cells working at the same operation voltage require longer pulse duration to reach the same on/off ratio, and results in typical endurance lifetime ~10^4 cycles (Supplementary Note 18). According to the analysis of optical images of failed devices and time-to-breakdown of the heterostructure (Supplementary Note 18), we associate the superior endurance lifetime in edge-contacted memory cells to the suppressed interfacial roughness, which in top-contacted memory cells may be introduced from the lattice distortion in MoS_2 by direct metal evaporation. The present edge contact strategy based on in-situ phase transition transforms 2H-MoS_2 under metal contact into its metallic 1 T phase, as an interfacial layer, it avoids undesired roughness at contact interface. With an endurance lifetime >10^6 cycles, the edge-contacted memory cell meets the requirement of the prevailing silicon flash memory (~10^5 cycles for SLC)[41–44], while having 2–4 orders' faster P/E speed (Supplementary Note 19).

Finally, we show the potential of edge-contacted memory cells as multi-level cell (MLC) to increase the bit density. It is worth noting that though multiple memory states (e.g., 4-bits) can be written in memory by altering the pulse condition (Supplementary Note 20), the aggressively narrowed margin among states is prone to deteriorate the attained endurance lifetime due to the sensitive drift of stored states by stress effect in hBN layer. Though hBN is exfoliated as a single crystal, it still exhibits a certain level of trap states that could store charges during repeated P/E cycles. Under the high electric field, more traps will be created by the defect generation, which in turn exacerbates the drift of memory states[45]. Thus, we instead choose to realize an MLC with 2-bit storage, with state-0 written via a program pulse (−8 V/10 ns for state-0), and states-1, 2, and 3 by erase pulses of varying amplitudes (8, 10, 12 V/100 ns) following the initial state-0. As indicated in Fig. 4d, the stored states are well separated in linear space, and could still be clearly distinguished after 8×10^4 cycled P/E operation. Such performance is ~1 order better than the commercial standard (~3×10^3 cycles for MLC) while having the ultrafast P/E speed[44]. In the future, further study on the integration of such memory cells in NAND or NOR architectures could promote its application in high-density mass-storage memory[46]. Meanwhile, for high-density integration of dramatically scaled memory cells with reduced channel and contact size, precise control of the position and quality of phase-engineered contact would be critical. As mass-storage memory, such dramatically improved P/E speed to <100 ns and multi-bit performance is favored by scenarios that demand high throughput data processing and storage, i.e., cloud servers, video recording, and analysis, etc.

Discussion

In summary, by introducing phase-engineered edge contact, we have demonstrated the realization of simultaneously ultrafast P/E speed ~10/100 ns, stable data retention >10 years, and super-robust endurance lifetime (>3×10^6 cycles for SLC, and 8×10^4 for MLC) in MoS_2 flash memory. We associate our device performance to the lateral hot carrier injection by edge contact under gate modulation, enabling a markedly lower P/E voltage for robust endurance lifetime while maintaining a decent charge injection efficiency for high-speed operation. The comprehensively improved key figures of merit over the existing commercial flash memory devices indicate a potential strategy to break the 'speed-retention-endurance' dilemma in conventional charge-based memory, making our 1 T contacted memory cell a viable option for high-speed and robust flash memory in the future.

Methods

Fabrication of 2D flash memory and characterization

The vdW heterostructures were made by sequentially stacking graphene, hBN, and MoS_2 obtained via mechanical exfoliation on a 300 nm SiO_2/Si substrate using PDMS-assisted dry transfer processes. The thickness of each layer was characterized via atomic force microscopy (Dimension Icon, Bruker). First, few-layer graphene (FLG) nanosheets were transferred onto the highly doped p-type silicon substrate that was cleaned by three-minute oxygen plasma in advance. After that, hBN and MoS_2 nanosheets were transferred to the top of FLG successively by dry transfer method with the help of PDMS (polydimethyl siloxane). As PDMS residues were sticky, an annealing process was carried out to remove residues and obtain a clean interface. Next, PMMA (polymethyl methacrylate) was spin-coated onto the fabricated heterostructures, and patterned by electron-beam lithography. Cr/Au (10 nm/40 nm) electrodes were then deposited by thermal evaporation. For edge-contacted devices, the heterostructures were immersed in 3 ml of n-butyl lithium (n-BuLi, Aladdin) in a sealed container at room temperature in a glovebox (ref. 33). Lithium intercalation into the interface between MoS_2 and hBN layer should be avoided. After soaking in the n-BuLi for 2.5 h, the samples were washed with hexane to remove excess n-BuLi. The sample was evaluated by Raman and photoluminescence (Alpha 300 R, WITec) to confirm the phase transition before depositing metal contact. For the fabrication of paired memory cells, FLG nanosheet transferred onto substrates was etched by Ar plasma to create two separate FLG flakes with identical shape and area.

Electrical characterization of memory

The electrical testing of the memory devices was performed at room temperature and in vacuum (if not specified) in the probe station (TTPX, Lake Shore). A semiconductor characterization system (B1500A, Keysight) is used to perform the electrical measurements. The direct current signals were generated using the source/monitor unit in the B1500A, while ultrafast electric pulses were generated using a Pulse/Pattern Generator (81110 A, Agilent). The waveform of generated pulse was recorded with an oscilloscope with a bandwidth of 2.5 GHz (DSO9254A, Keysight). All the electrical testing was performed in a vacuum condition to avoid the effect of the atmosphere.

Data availability

The Source Data underlying the figures of this study are available at https://doi.org/10.6084/m9.figshare.23994789. All raw data generated during the current study are available from the corresponding authors upon request.

References

1. Badaroglu, M. International Roadmap for Devices and Systems 2021 (IEEE, 2021); https://irds.ieee.org/editions/2021.
2. Zidan, M. A., Strachan, J. P. & Lu, W. D. The future of electronics based on memristive systems. *Nat. Electron.* **1**, 22–29 (2018).
3. Wong, H. S. & Salahuddin, S. Memory leads the way to better computing. *Nat. Nanotechnol.* **10**, 191–194 (2015).
4. Park, J. W. et al. A 176-stacked 512Gb 3b/Cell 3D-NAND flash with 10.8Gb/mm^2 density with a peripheral circuit under cell array architecture. *In:* 2021 IEEE International Solid-State Circuits Conference (ISSCC), 422-423 (IEEE, 2021).
5. Yosuke, K. et al. Disturbless flash memory due to high boost efficiency on BiCS structure and optimal memory film stack for ultra high density storage device. *In:* 2008 IEEE International Electron Devices Meeting (IEDM), 1-4 (IEEE, 2008).
6. Jeong, J. et al. Dynamic erase voltage and time scaling for extending lifetime of NAND flash-based SSDs. *IEEE Trans. Comput.* **66**, 616–630 (2017).

7. Lombardo, S. et al. Dielectric breakdown mechanisms in gate oxides. *J. Appl. Phys.* **98**, 121301 (2005).
8. Palumbo, F. et al. A review on dielectric breakdown in thin dielectrics: silicon dioxide, high-k, and layered dielectrics. *Adv. Funct. Mater.* **30**, 1900657 (2019).
9. Park, H. et al. Impact of SiO_2/Si interface micro-roughness on SILC distribution and dielectric breakdown: a comparative study with atomically flattened devices. *In:* 2017 IEEE International Reliability Physics Symposium (IRPS) (IEEE, 2017).
10. Houdt, J. V. et al. Analysis of the enhanced hot-electron injection in split-gate transistors useful for EEPROM applications. *IEEE Trans. Electron Devices* **39**, 1150–1156 (1992).
11. Sung-Jin, C. et al. Enhancement of program speed in dopant-segregated Schottky-Barrier (DSSB) FinFET SONOS for NAND-type flash memory. *IEEE Electron Device Lett.* **30**, 78–81 (2009).
12. Uchida, K. et al. Enhancement of hot-electron generation rate in Schottky source metal–oxide–semiconductor field-effect transistors. *Appl. Phys. Lett.* **76**, 3992–3994 (2000).
13. Panciera, F. et al. Ni(Pt)-silicide contacts on CMOS devices: Impact of substrate nature and Pt concentration on the phase formation. *Microelectron. Eng.* **120**, 34–40 (2014).
14. Chen, L. J. Metal silicides: An integral part of microelectronics. *JOM* **57**, 24–30 (2005).
15. Lin, L., Guo, Y. & Robertson, J. Metal silicide Schottky barriers on Si and Ge show weaker Fermi level pinning. *Appl. Phys. Lett.* **101**, 052110 (2012).
16. Liu, Y. et al. Approaching the Schottky-Mott limit in van der Waals metal-semiconductor junctions. *Nature* **557**, 696–700 (2018).
17. Yang, Z. et al. A fermi-level-pinning-free 1D electrical contact at the intrinsic 2D MoS_2–metal junction. *Adv. Mater.* **31**, 1808231 (2019).
18. Liu, Y., Huang, Y. & Duan, X. Van der Waals integration before and beyond two-dimensional materials. *Nature* **567**, 323–333 (2019).
19. Novoselov, K. S., Mishchenko, A., Carvalho, A. & Castro Neto, A. H. 2D materials and van der Waals heterostructures. *Science* **353**, aac9439 (2016).
20. Giusi, G., Marega, G. M., Kis, A. & Iannaccone, G. Impact of interface traps in floating-gate memory based on monolayer MoS_2. *IEEE Trans. Electron Devices* **69**, 6121–6126 (2022).
21. Bertolazzi, S., Krasnozhon, D. & Kis, A. Nonvolatile memory cells based on MoS_2/graphene heterostructures. *ACS Nano* **7**, 3246–3252 (2013).
22. Choi, M. S. et al. Controlled charge trapping by molybdenum disulphide and graphene in ultrathin heterostructured memory devices. *Nat. Commun.* **4**, 1624 (2013).
23. Li, D. et al. Nonvolatile floating-gate memories based on stacked black phosphorus-boron nitride-MoS_2 heterostructures. *Adv. Funct. Mater.* **25**, 7360–7365 (2015).
24. Wang, S. P. et al. New floating gate memory with excellent retention characteristics. *Adv. Electron. Mater.* **5**, 1800726 (2019).
25. Chen, Y. et al. An asymmetric hot carrier tunneling van der Waals heterostructure for multibit optoelectronic memory. *Mater. Horiz.* **7**, 1331–1340 (2020).
26. Migliato Marega, G. et al. Logic-in-memory based on an atomically thin semiconductor. *Nature* **587**, 72–77 (2020).
27. Wang, Y. A. et al. Band-tailored van der Waals heterostructure for multilevel memory and artificial synapse. *Infomat* **3**, 917–928 (2021).
28. Wu, L. et al. Atomically sharp interface enabled ultrahigh-speed nonvolatile memory devices. *Nat. Nanotechnol.* **16**, 882–887 (2021).
29. Liu, L. et al. Ultrafast non-volatile flash memory based on van der Waals heterostructures. *Nat. Nanotechnol.* **16**, 874–881 (2021).
30. Lin, Y. R. et al. Comparison with nitride interface defects and nanocrystals for charge trapping layer nanowire gate-all-around nonvolatile memory performance. *IEEE Trans. Electron Devices* **65**, 493–498 (2018).
31. Fang, H. K. et al. Operation characteristics of gate-all-around junctionless flash memory devices with Si_3N_4/ZrO-based stacked trapping layer. *IEEE Trans. Electron Devices* **67**, 3626–3631 (2020).
32. Huang, X. et al. An ultrafast bipolar flash memory for self-activated in-memory computing. *Nat. Nanotechnol.* **18**, 486–492 (2023).
33. Kappera, R. et al. Phase-engineered low-resistance contacts for ultrathin MoS_2 transistors. *Nat. Mater.* **13**, 1128–1134 (2014).
34. Voiry, D. et al. Covalent functionalization of monolayered transition metal dichalcogenides by phase engineering. *Nat. Chem.* **7**, 45–49 (2015).
35. Shih, C. H. & Luo, Y. X. Effects of dopant-segregated profiles on Schottky barrier charge-trapping flash memories. *IEEE Trans. Electron Devices* **61**, 1361–1368 (2014).
36. Sasaki, T. et al. Material and device structure designs for 2D memory devices based on the floating gate voltage trajectory. *ACS Nano* **15**, 6658–6668 (2021).
37. Shen, P. C. et al. Ultralow contact resistance between semimetal and monolayer semiconductors. *Nature* **593**, 211–217 (2021).
38. Zhao, P. et al. Evaluation of border traps and interface traps in HfO_2/MoS_2 gate stacks by capacitance–voltage analysis. *2D Mater.* **5**, 031002 (2018).
39. Nowicki, R. Influence of space-charge on potential barrier in field emission. *Surf. Sci.* **8**, 357–369 (1967).
40. Gehring, A. & Selberherr, S. Modeling of tunneling current and gate dielectric reliability for nonvolatile memory devices. *IEEE Trans. Electron Devices* **4**, 306–319 (2004).
41. Li, Y. et al. A novel dual-doping floating-gate (DDFG) flash memory featuring low power and high reliability application. *IEEE Electron Device Lett.* **28**, 622–624 (2007).
42. Chen, C., Chang-Liao, K., Wu, K. & Wang, T. Improved erasing speed in junctionless flash memory device by HfO_2/Si_3N_4 stacked trapping layer. *IEEE Electron Device Lett.* **34**, 993–995 (2013).
43. Chen, G. et al. Metal floating gate memory device with SiO_2/HfO_2 dual-layer as engineered tunneling barrier. *IEEE Electron Device Lett.* **35**, 744–746 (2014).
44. Advani, R. N. 3D Flash Memories Ch. 1 (Springer Netherlands, 2016).
45. Hattori, Y., Taniguchi, T., Watanabe, K. & Nagashio, K. Impact ionization and transport properties of hexagonal boron nitride in a constant-voltage measurement. *Phys. Rev. B* **97**, 045425 (2018).
46. Liu, Y. et al. Promises and prospects of two-dimensional transistors. *Nature* **591**, 43–53 (2021).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 21825103 (T.Z.) and No. U21A2069 (T.Z.), Item “large-scale and energy-efficient in-memory computing systems” of National Key Research and Development Program of China (Y.H.), and Ministry of Science and Technology of China under the grant No. 2021YFA1200500 (Y.M.). We also thank the technical support from the Analytical and Testing Center at Huazhong University of Science and Technology.

Author contributions

F.Z. and T.Z. conceived and supervised the project. J.Y. and H.W. contributed equally to this work, by performing most of the device fabrication and electrical characterization. M.H., X.X. participated partly in device fabrication. Z.C., Y.H., and X.S.M. provided support to the simulation and endurance measurements. F.Z., J.Y., and H.W. conducted data analysis by discussing with all authors. F.Z., J.Y., H.W., Y.M., and T.Z. co-wrote the manuscript and revised it by discussion with all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-41363-x.

Correspondence and requests for materials should be addressed to Fuwei Zhuge, Ying Ma or Tianyou Zhai.

Peer review information *Nature Communications* thanks Fei Xue, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2023