nature nanotechnology

Article

https://doi.org/10.1038/s41565-023-01399-y

Scalable CMOS back-end-of-line-compatible AlScN/two-dimensional channel ferroelectric field-effect transistors

Received: 11 October 2022

Accepted: 13 April 2023

Published online: 22 May 2023

Check for updates

Kwan-Ho Kim[1], Seyong Oh[2], Merrilyn Mercy Adzo Fiagbenu[1], Jeffrey Zheng[3], Pariasadat Musavigharavi[1,3], Pawan Kumar[1], Nicholas Trainor[4], Areej Aljarb[5], Yi Wan[6], Hyong Min Kim[1], Keshava Katti[1], Seunguk Song[1], Gwangwoo Kim[1], Zichen Tang[1], Jui-Han Fu[7], Mariam Hakami[7], Vincent Tung[6,7], Joan M. Redwing[4], Eric A. Stach[3], Roy H. Olsson III[1] ✉ & Deep Jariwala[1] ✉

Three-dimensional monolithic integration of memory devices with logic transistors is a frontier challenge in computer hardware. This integration is essential for augmenting computational power concurrent with enhanced energy efficiency in big data applications such as artificial intelligence. Despite decades of efforts, there remains an urgent need for reliable, compact, fast, energy-efficient and scalable memory devices. Ferroelectric field-effect transistors (FE-FETs) are a promising candidate, but requisite scalability and performance in a back-end-of-line process have proven challenging. Here we present back-end-of-line-compatible FE-FETs using two-dimensional MoS_2 channels and AlScN ferroelectric materials, all grown via wafer-scalable processes. A large array of FE-FETs with memory windows larger than 7.8 V, ON/OFF ratios greater than 10^7 and ON-current density greater than 250 μA um^{-1}, all at ~80 nm channel length are demonstrated. The FE-FETs show stable retention up to 10 years by extension, and endurance greater than 10^4 cycles in addition to 4-bit pulse-programmable memory features, thereby opening a path towards the three-dimensional heterointegration of a two-dimensional semiconductor memory with silicon complementary metal–oxide–semiconductor logic.

The generation of vast amounts of data by electronic devices and the sensors within them presents an acute need for high-speed and energy-efficient data storage and processing. In conventional processor-centric computing, a processor core has to traverse various levels of memory at varying distances and speeds, resulting in a data bottleneck and inefficient data handling[1–3]. Recently, compute-in-memory architectures with vertically stacked, dense, high-efficiency and tightly integrated memory have been suggested to overcome such data-handling bottlenecks[4,5]. In contrast with conventional compute-in-memory architectures, which are primarily front end of line (that is, the memory is co-located with the Si logic transistors and peripheral circuits on the same layer), a memory array vertically stacked directly over the front end of line can provide a huge advantage in areal density and energy efficiency as well as reduce latency[6]. At a

[1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. [2]Division of Electrical Engineering, Hanyang University ERICA, Ansan, South Korea. [3]Department of Materials Science and Engineering, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Materials Science and Engineering, Pennsylvania State University, University Park, PA, USA. [5]Department of Physics, King Abdulaziz University, Jeddah, Saudi Arabia. [6]Department of Physical Science and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [7]Department of Chemical System Engineering, University of Tokyo, Tokyo, Japan. ✉e-mail: rolsson@seas.upenn.edu; dmj@seas.upenn.edu

single-unit level, this scheme requires a fast, reliable and low-energy non-volatile memory (NVM) device that can be easily integrated with the processing transistors without occupying precious space on the logic layer[7]. This drives the need for materials and devices compatible with back-end-of-line (BEOL) processing. Therefore, monolithic three-dimensional (M3D) integration of NVM devices and Si complementary metal–oxide–semiconductor (CMOS) logic is desirable not only from the perspective of bringing memory closer to the processing unit but this approach can also reduce the data bottleneck issue and increase chip-level integration density[8,9].

With recent advances in ferroelectric (FE) materials such as $Hf_xZr_{1-x}O_2$ (HZO), a ferroelectric field-effect transistor (FE-FET) is considered to be one of the most promising, compact and energy-efficient NVM candidates for M3D integration, as it allows a non-destructive read operation[6,10]. Although HZO FE-FETs have made substantial advances, FE properties superior to HZO have recently been discovered in aluminium scandium nitride (AlScN)[11,12]. AlScN not only exhibits a high remnant polarization (P_r) of greater than 110 μC cm^{-2}, which is more than three times that of HZO, but also has a BEOL-compatible growth temperature of 350 °C with reliable growth demonstrated even as low as room temperature (Supplementary Fig. 1)[13]. This makes it a very attractive candidate for not only BEOL-compatible FE-FETs (process temperature should be less than 400 °C)[14] but also flexible device applications that should be processed at a temperature of less than 200 °C. In addition, AlScN has a comparably low dielectric constant of around 12–18; excellent thermal stability of FE properties even at temperatures greater than 1,000 °C (ref. 15), which is critical for high-temperature applications; $Al_xSc_{1-x}N$ (for $x < 0.43$) only exists in the wurtzite phase, which is FE;[16] and it does not require a post-deposition annealing process to achieve the required single-phase FE structure. The fact that AlScN readily forms the desired single-phase FE structure is highly advantageous for uniform device performance at scale. In contrast, HZO has various metastable phases out of which only the orthorhombic or rhombohedral phases are known to be FE, which requires careful post-annealing processes to increase the proportion of the FE phase[17,18].

Here we report an array of BEOL-compatible two-dimensional (2D)/3D heterogeneous FE-FETs with 80–500 nm channel lengths using thin-film AlScN (20, 45 and 100 nm) over large areas. Selecting atomically thin monolayer MoS_2 as the FE-FET channel provides a key advantage by exploiting the high P_r of AlScN as an FE gate dielectric because of the ability of MoS_2 to support high carrier densities[19], which is critical for a high ON-current density and high-speed operation as well as alleviating short-channel effects. We present three notable advances. (1) By reducing the thickness and increasing the scandium content in AlScN, we demonstrate control over the memory window (MW) and switching voltage of the BEOL-compatible 2D channel FE-FETs, bringing them closer to conventional flash memory devices[20,21]. (2) Our devices are highly scaled with channel lengths (L_{CH}) as small as 78 nm, which concurrently have an ON/OFF current ratio of 10^7 and a current density of 252 μA μm^{-1}. (3) We demonstrate stable retention (10 years by extension), endurance (>10^4 cycles) and pulse-programmed 4-bit memory operation in addition to their 7-bit operation as artificial synapses.

Figure 1a shows the schematic of a MoS_2/AlScN FE-FET. It is fabricated using 20-, 45- or 100-nm-thick $Al_{1-x}Sc_xN$ FE dielectric films deposited on four-inch Pt(111)/Ti/SiO_2/Si or Al(111)/sapphire wafers. The substrate temperature during the deposition of AlScN was maintained at 350 °C, a BEOL-compatible thermal budget. Large-area single-layer MoS_2 was used as the channel material of the FE-FETs. The large-area MoS_2 films were prepared via three different methods, namely, chemical vapour deposition (CVD) 1 (ref. 22), metal–organic chemical vapour deposition (MOCVD)[23] and CVD 2 (Methods) on two-inch sapphire wafers and were transferred onto 20/45/100 nm $Al_xSc_{1-x}N$ films for device fabrication and testing (Methods). The device surface morphology and interface structure were confirmed through scanning electron microscopy (SEM) and cross-sectional transmission electron microscopy (TEM). As shown in Fig. 1b,c, the FE-FET has a channel width of 20 μm and L_{CH} of 500 nm. This L_{CH} is further aggressively scaled down to ~78 nm (Fig. 2a, low-magnification bright-field TEM image of the semiconductor/dielectric interface), and the phase-contrast lattice image of the MoS_2 and AlScN interface combined with elemental analysis (energy-dispersive X-ray spectroscopy mapping in the scanning TEM mode) shows a single layer of MoS_2 on top of crystalline AlScN. There is no evidence of an oxide layer on AlScN (Fig. 1d,e). It should be noted that minimizing the oxidation of the AlScN top surface that forms the interface with the semiconductor is important to avoid serious performance degradation of the FE-FETs (Supplementary Figs. 2 and 3).

Next, the current density–electric field (J–E) hysteresis loops of the as-sputtered $Al_{1-x}Sc_xN$ samples were first measured on a metal–AlScN–metal structural capacitor to study its coercive field (E_c). A value of −4.5/5.1 MV cm^{-1} was extracted under 10 kHz excitation (Fig. 1f). Polarization-dependent leakage was observed in the loop as in previous reports[14]. To minimize the effect of the leakage current and estimate an accurate P_r, positive-up–negative-down (PUND) pulsed measurements were carried out using short pulse widths of 500 ns (Supplementary Fig. 3). The PUND results show a saturated P_r^+ and P_r^- of around 135 μC cm^{-2} obtained by $(P_r^+ + P_r^-)/2$. Detailed information about the J–E loop and PUND measurements can be found in the Methods section.

All the as-fabricated long-channel (L_{CH} = 500 nm) FE-FETs (based on three different types of large-area MoS_2 and 100 nm $Al_{0.68}Sc_{0.32}N$ film) show counterclockwise hysteretic I_D–V_G plots with a very large MW of about 18 V, a high ON/OFF ratio of 10^7 and an ON-current density of 71 μA μm^{-1} (W_{CH} = 20 μm) at V_{DS} = 1 V (Fig. 1g). Under a positive (negative) gate voltage above E_c, the FE polarization is switched in the direction pointing towards the channel (opposite of the channel), and consequently, electrons are accumulated (depleted) in the channel, causing a low-threshold-voltage (LVT) (high-threshold-voltage (HVT)) state. For low-energy consumption and M3D integration of the FE-FETs with Si CMOS, the switching voltage must be reduced. One way to achieve this is by reducing the AlScN thickness and increasing the Sc alloying concentration[16]. As shown in Fig. 1h, the FE switching voltage for the maximum MW is reduced from 20 V for 100-nm-thick $Al_{0.72}Sc_{0.28}N$ to 10 V for 45-nm-thick $Al_{0.68}Sc_{0.32}N$ and 5–6 V for 20-nm-thick $Al_{0.68}Sc_{0.32}N$. Consequently, the MW also reduces from 21.0 V to 7.8 V and 1.0–4.0 V, respectively. Arrays of FE-FETs fabricated using 20 nm AlScN and various types of MoS_2 are shown in Supplementary Fig. 4. The MWs are obtained by subtracting an LVT from an HVT that are extracted at a current level of $(W_m/L_m) \times 10^{-7}$ A, where W_m and L_m are the channel width and length of the mask, respectively. To corroborate the evidence of FE switching observed in the FE-FETs, the transfer curves are compared between 50 nm SiO_2/MoS_2 FETs and 45 nm AlScN/MoS_2 FE-FETs that are fabricated using the same CVD MoS_2, fabrication process and device dimensions, except for the gate insulator. As shown in Fig. 1i, the transfer curve of SiO_2/MoS_2 FETs shows a clockwise hysteresis loop that originates from charge trapping[24], whereas that of AlScN/MoS_2 FE-FETs have a counterclockwise hysteresis loop. In addition, even if the sweep range of the gate voltage is three times narrower (−10 to 10 V), the current level and ON/OFF ratio are approximately 10^4 and 5×10^4 times larger, respectively, in AlScN/MoS_2 FE-FET compared with SiO_2/MoS_2 FET. This observation confirms FE switching in the AlScN/MoS_2 structure. Figure 1j shows the output curves of the devices that demonstrate an ON-current density of 252 μA μm^{-1} at V_{DS} of 3 V. To the best of our knowledge, this is among the highest current density values obtained without any channel doping or contact resistance engineering in a 2D channel FET, further highlighting the importance of high-P_r FE materials like AlScN.

Next, the L_{CH} values of the FE-FETs are aggressively scaled from 500 to 78 nm (Fig. 2a) as the channel width is maintained. Further, the evaluation of the device metrics over an array of devices is also performed (Supplementary Figs. 6 and 7). The devices show significant overlap

https://doi.org/10.1038/s41565-023-01399-y

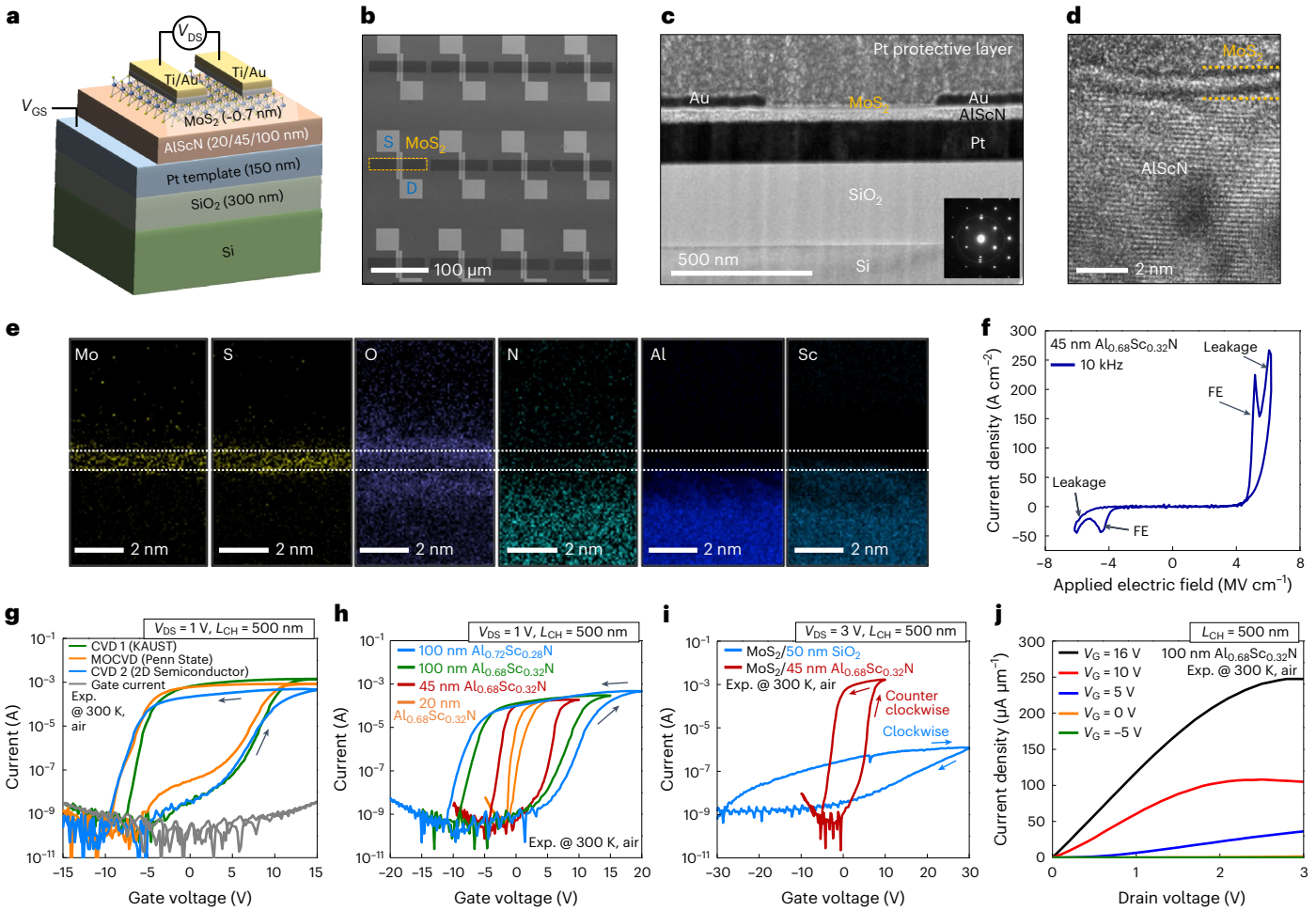

Fig. 1 | AlScN/MoS_2 FE-FET device structure and electrical characteristics and the FE property of AlScN. a, Schematic of a MoS_2/AlScN FE-FET. **b**, SEM image of the FE-FET array. S, source; D, drain. **c**, Cross-sectional bright-field TEM image of the MoS_2/AlScN FE-FET. **d**, Phase-contrast lattice image of the MoS_2/AlScN interface. Inset shows an electron diffraction pattern of the film stack on the zone axis of the Si (100) substrate. **e**, Energy-dispersive X-ray spectroscopy maps of the FE-FET. **f**, The *J*–*E* hysteresis loops of $Al_{0.68}Sc_{0.32}N$ under 10 kHz. **g**,**h**, Semi-logarithmic-scale transfer characteristics at room temperature of a representative AlScN/MoS_2 FE-FET based on MoS_2 grown using three different methods: CVD 1 (green), MOCVD (orange) and CVD 2 (blue) (**g**), and by changing the alloying concentration of Sc (28% and 32%) and thickness of AlScN (20, 45 and 100 nm), namely, 100 nm $Al_{0.72}Sc_{0.28}N$ (blue curve), 100 nm $Al_{0.68}Sc_{0.32}N$ (green curve), 45 nm $Al_{0.68}Sc_{0.32}N$ (red curve) and 20 nm $Al_{0.68}Sc_{0.32}N$ (orange curve) (**h**). The transfer characteristics of the FE-FETs are recorded at 10 Hz rate with 0.2 V gate-voltage spacing. **i**, Semilogarithmic-scale transfer characteristics at room temperature of a representative 45 nm AlScN/MoS_2 FE-FET (red) and 50 nm SiO_2/MoS_2 (blue). **j**, Linear-scale output characteristics of a representative 100 nm $Al_{0.68}Sc_{0.32}N$/MoS_2 FE-FET at various gate voltages (V_G).

in the transfer characteristics and MWs (Fig. 2b), making the case for viability at technology levels with further developments. It is worth noting that the FE-FETs maintain a large MW of ~8 V and an ON/OFF ratio greater than 10^6 even after the aggressive scaling of L_{CH} and AlScN thickness. This is because of the high P_r value of AlScN, which noticeably keeps the OFF current low. This reliable memory operation of MoS_2/AlScN FE-FETs is maintained even as the width is reduced from 20 μm to 800 nm (Supplementary Fig. 8). These results are due to the atomically thin body of MoS_2, which permits superior electrostatic control. The advantages and motivations for adopting MoS_2 are described in Supplementary Figs. 9 and 10. For a fair assessment of the FE-FET performance, a figure that compares a normalized MW and ON-state conductivity that are extracted from previous reports is included (Fig. 2c). Since the MW increases with the thickness of the FE, an MW normalized to the FE material thickness is the fairest metric for comparison. Similarly, the ON-state conductivity is also a normalized metric to the channel width/length and drain voltage. As evident from this figure, both normalized MW and ON-state conductivity of the FE-FETs are among the highest compared with other 2D channel FE-FETs. It is also noteworthy that both normalized MW and ON-state conductivity values are maintained even when the thickness of AlScN is reduced from 100 to 45 nm (the left and right red stars (Fig. 2c) correspond to 45 and 100 nm AlScN, respectively). This suggests the possibility for the further scaling down of AlScN thickness in future works without performance degradation. Then, statistical analyses of the MW (Fig. 2d) and ON/OFF ratios (Fig. 2e,f) are also shown. It is worth noting that the variation in HVT is larger than that in the LVT, and this is possibly related to the larger resistive leakage current in p^+-polarized AlScN (ref. 25). Furthermore, the mean values of ON/OFF_{MAX} and ON/OFF ratios at V_G of 0 V (Fig. 2e,f) were observed to be 1.6×10^6 and 2.3×10^5, respectively. These variations in device-to-device transfer curves primarily originate from university-laboratory-/cleanroom-level device fabrication processing and channel inhomogeneities, which can be reduced with advanced foundry-based fabrication processes as well as with improved MoS_2 synthesis and AlScN deposition processes. Also, the extracted mobility and subthreshold swing from the transfer curves can be found in Supplementary Fig. 12. The experimental *I*–*V* curves of the FE-FET devices are further verified via technology computer-aided design simulations (Supplementary Fig. 13). These simulations suggest a low level of fixed charge at the interface in the devices.

https://doi.org/10.1038/s41565-023-01399-y

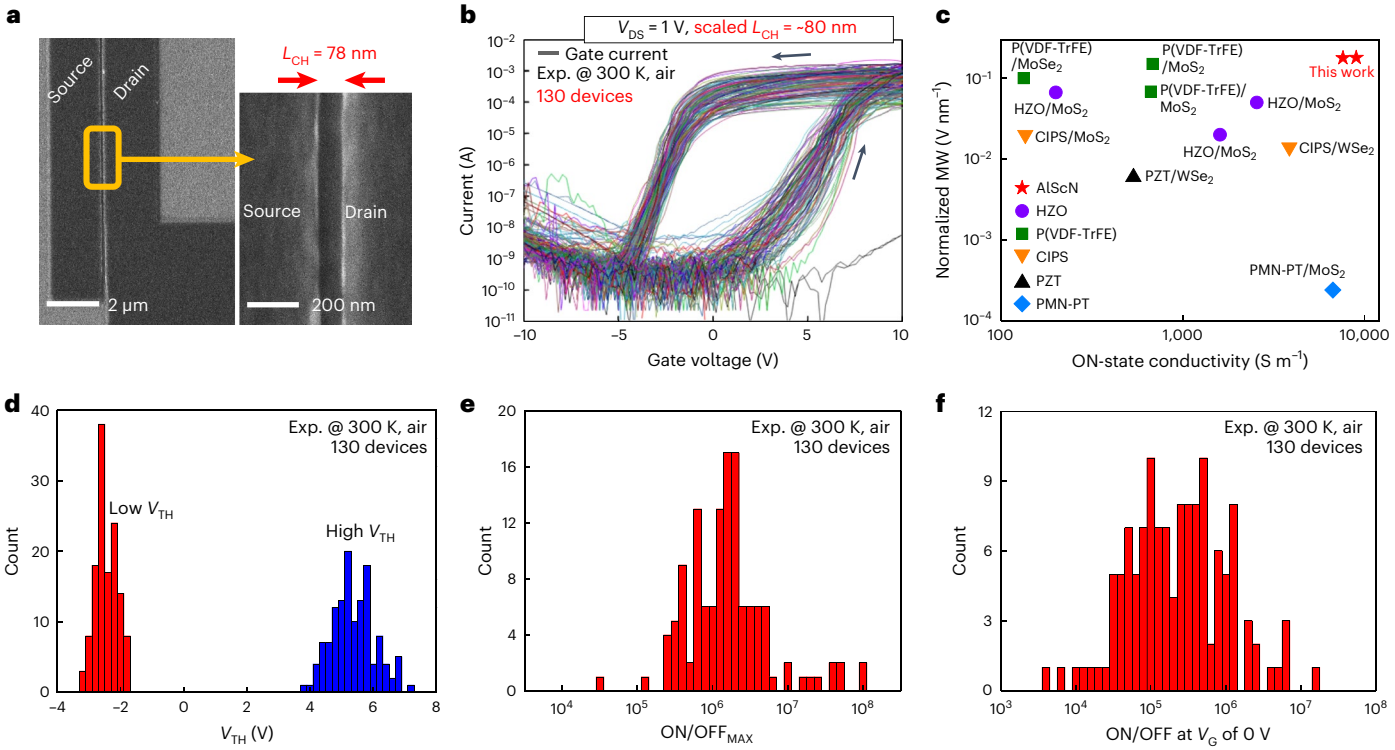

Fig. 2 | Array of scaled AlScN/MoS_2 FE-FETs. a, Magnified SEM image utilizing in-lens backscattered detector to confirm the channel length of the FE-FET. **b**, Semi-logarithmic-scale transfer characteristics at room temperature of an array of 45 nm $Al_{0.68}Sc_{0.32}N/MoS_2$ FE-FETs with channel lengths of around 80 nm (total, 130 devices). **c**, Comparison of normalized MW and ON-state conductivity from the reported 2D channel FE-FETs in the literature with various FE materials[28–37]. The left and right red stars in the graph correspond to 45 nm and 100 nm AlScN/MoS_2 FE-FETs, respectively. **d**–**f**, Distributions of LVT and HVT states (**d**), maximum of the ON/OFF current (ON/OFF$_{MAX}$) (**e**) and ON/OFF current at V_G of 0 V (**f**) of the array of FE-FETs.

Next, the voltage-pulse-induced FE switching of the FE-FETs by applying various pulse amplitudes and widths is investigated. Voltage-pulse-induced FE switching is important since FE-FETs operate in a circuit application based on pulsed programming and erasing of the resistance states. As shown in Fig. 3a, after applying a programming (PRG) or erasing (ERS) pulse, the transfer characteristics of the scaled FE-FET are measured using narrow d.c. gate-voltage sweeps ranging from 6 to −5 V in which the voltage range is lower than the switching voltage (Fig. 3a); the threshold voltage (V_{TH}) of the FE-FETs can be controllably changed by these pulses from the initial state (black) to the LVT (red) or HVT (blue). The width and amplitude of the PRG and ERS pulses used in this study range from (500 ns, 34 V) to (40 ms, 12 V). It should be noted that, as in many FE materials, there is a notable trade-off relation between the pulse width and pulse amplitude for FE switching[20], namely, a shorter pulse width requires a higher pulse amplitude and vice versa, and the relationship between the two is a power of frequency (Fig. 3c). The trade-off between a pulse width and pulse amplitude is further confirmed through frequency-dependent *I*–*E* hysteresis loops (Supplementary Fig. 15). Not only programming and erasing but also understanding and evaluating the non-volatile retention of these states is equally important. In the FE-FET devices, the retention of LVT/HVT of the scaled FE-FET was measured at room temperature in air (Fig. 3b). The FE-FETs exhibit stable retention characteristics, showing a large MW of more than 3 V even after an extension of the trend to a timescale of 10 years. It is confirmed that this stable retention is reproducible even with a shorter PRG and ERS pulse width (Supplementary Fig. 18). These retention measurements further solidify the evidence of FE switching as there is no retention observed in SiO_2/MoS_2 FETs, even when the same CVD MoS_2 and fabrication process are used (Supplementary Fig. 19). Aside from time-dependent retention, the devices also exhibit stable switching endurance for more than 10,000 cycles and maintain the ON/OFF ratio when a pulse with 10 V amplitude and 40 ms width is used (Fig. 3e,f). To confirm the statistical endurance of the FE-FETs, four more devices are measured, and it is observed that the endurance cycle ranges from 8,000 to 10,000 in all the devices (Supplementary Fig. 20). An additional comparison of the device performance of the AlScN/MoS_2 FE-FETs to other memory, storage and emerging devices is presented in Supplementary Fig. 14.

Voltage-pulse-induced resistance and V_{TH} switching are important attributes of FE-FETs since these can be made tunable as a function of voltage amplitude and voltage duration to induce partial switching of the FE domains underneath the channel. This property of partial switching is due to the stochastic nature of FE-domain switching. This attribute of the FE-FET memory devices is explored in terms of multibit operation. To increase the effective data density per NVM cell, multibit operation is a key feature of modern memories. Although the device performance of individual FE-FETs has reached or surpassed floating-gate FETs in flash technology[20], multibit demonstrations in FE-FETs are still in their infancy even for HfO_x-based FE-FETs and has never been demonstrated before for nitride FE materials. The large MW of the FE-FETs is due to the large E_c of AlScN, which is a favourable attribute for the demonstration of multibit storage in BEOL-compatible FE-FETs. Figure 4a shows the successful demonstration of 2-bit operation measured from 30 scaled FE-FETs. It is worth nothing that the four memory states have a relatively tight distribution in the 30 measured devices, a key requirement for a scalable and reliable multibit memory technology. The different levels of V_{TH} shift are also obtained by applying a different number, width and amplitude of pulses (Supplementary Fig. 22). As shown in Fig. 4b, the obtained 2-bit V_{TH} states also show stable retention up to 10^3 s. By splitting the V_{TH} values more finely, even

https://doi.org/10.1038/s41565-023-01399-y

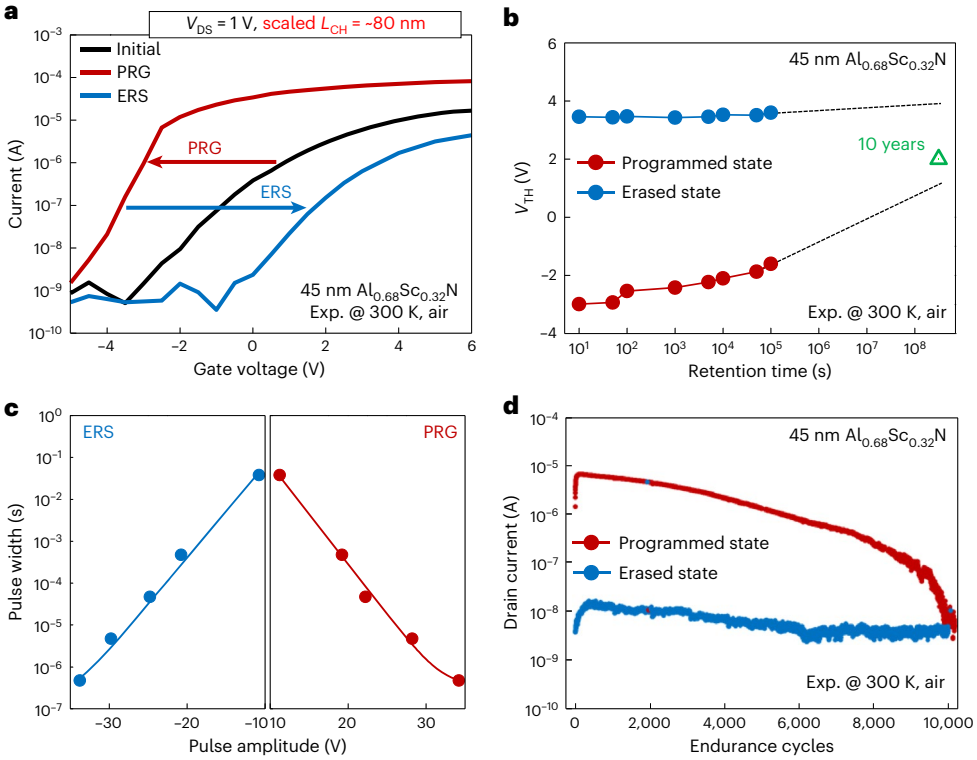

Fig. 3 | Electrical characterization of voltage-pulse-induced FE switching in scaled FE-FETs. a, Semilogarithmic-scale transfer characteristics of the 45 nm AlScN/MoS_2 FE-FETs with a channel length of around 80 nm after applying a PRG or ERS pulse with a width of 40 ms and amplitude of ±12 V. **b**, Retention measurement of the extracted V_{TH} after applying a PRG/ERS pulse of 40 ms and ±12 V up to 10^5 s. The pulse scheme for the retention measurements can be found in Supplementary Fig. 16. **c**, Trade-off relation between pulse width and pulse amplitude for FE switching for achieving identical V_{TH} shift. **d**, Write endurance measurements of the extracted drain current at V_{GS} of 0 V and V_{DS} of 0.01 V versus the number of cycles using input pulses with 40 ms width and 10 V amplitude for more than 10,000 cycles. The pulse scheme for the endurance measurements can be found in Supplementary Fig. 16.

4-bit operation can be demonstrated in the FE-FET (Fig. 4c). Furthermore, it is confirmed that the 4-bit V_{TH} states can also be programmed using shorter pulse widths of 1 µs at the expense of PRG amplitude (Supplementary Fig. 21). This multibit operation indicates that multiple FE domains are contained under the channel of the FE-FET, and the domains can be partially polarized by PRG pulses[25,26]. This partial FE switching is confirmed once again because the different shifts in the threshold voltage depend on the pulse number and pulse amplitude (Supplementary Fig. 22). Due to the accumulative FE switching property, when pulses with a sub-coercive voltage are consecutively applied, the threshold voltage shifts (Supplementary Fig. 22). To the best of our knowledge, this is the first demonstration of multistate programming in FE nitrides and in BEOL-compatible FE-FETs at this scale. These results suggest the foundation for a scalable M3D integration of memory with logic.

As a final demonstration of another application, 7-bit (128) conductance states are shown for the pulse-programmed operation of the FE-FET as an artificial synapse (Supplementary Fig. 26 shows the magnified images). Such fine programming of the conductance suggests the presence of multiple FE domains in the channel, out of which a small number switch stochastically with progressive numbers of pulses. Figure 4d shows the synaptic weight update of the FE-FETs from each of the 128 consecutive potentiation (V_P) and depression (V_D) pulses. Long-term potentiation (LTP) and long-term depression (LTD) was observed when a pulse with a width of 150 µs and 15 V amplitude was applied to the gate at 4 kHz speed (Supplementary Fig. 23). This LTP/LTD behaviour is observed to be reliable in extending the performance without degradation over at least 8,000 PRG pulses (Fig. 4e). In addition, a difference in excitatory postsynaptic output current was observed to stably maintain up to at least 100 s after applying various numbers of V_P pulses (Supplementary Fig. 24). Finally, a multilayer perceptron (MLP)-based artificial neural network (ANN) simulation is performed using the open-source code NeuroSimV3.0 (Fig. 4f)[27]. Here the designed ANN consisted of 400 input, 100 hidden and 10 output neurons, and each neuron was fully connected via artificial synapses that contained the nonlinear parameters of the device. The Modified National Institute of Standards and Technology dataset of black-and-white handwritten digit patterns with a size of 20 × 20 was used for training (60,000) and testing (10,000). As a result, the maximum accuracy based on the LTP/LTD curve reached a very high value of 94.26% (96.19% for software-based simulation). A comparison table for MLP-based ANN simulation using various devices is presented in Supplementary Fig. 27. It should be noted that due to the high linearity of the LTP/LTD curve, the high training accuracy could be maintained even when there are several fluctuations in each state of the 7-bit operation (Supplementary Fig. 25).

Conclusions

In summary, we have demonstrated NVM applications of a scalable and CMOS BEOL-compatible FE AlScN with less than 1-nm-thick 2D channels down to ~80 nm channel length in an array of devices. The stable memory performance of the FE-FETs combined with their scalability and low-temperature integration make a promising case for vertical heterointegration with Si CMOS logic transistors. Our work enables the demonstration of high-performance, stable, scalable and BEOL-compatible 2D + FE memory technology, a key development for both 2D memory technology and FE materials. Further, our work opens the door to replace flash memory and ultimately also high-bandwidth

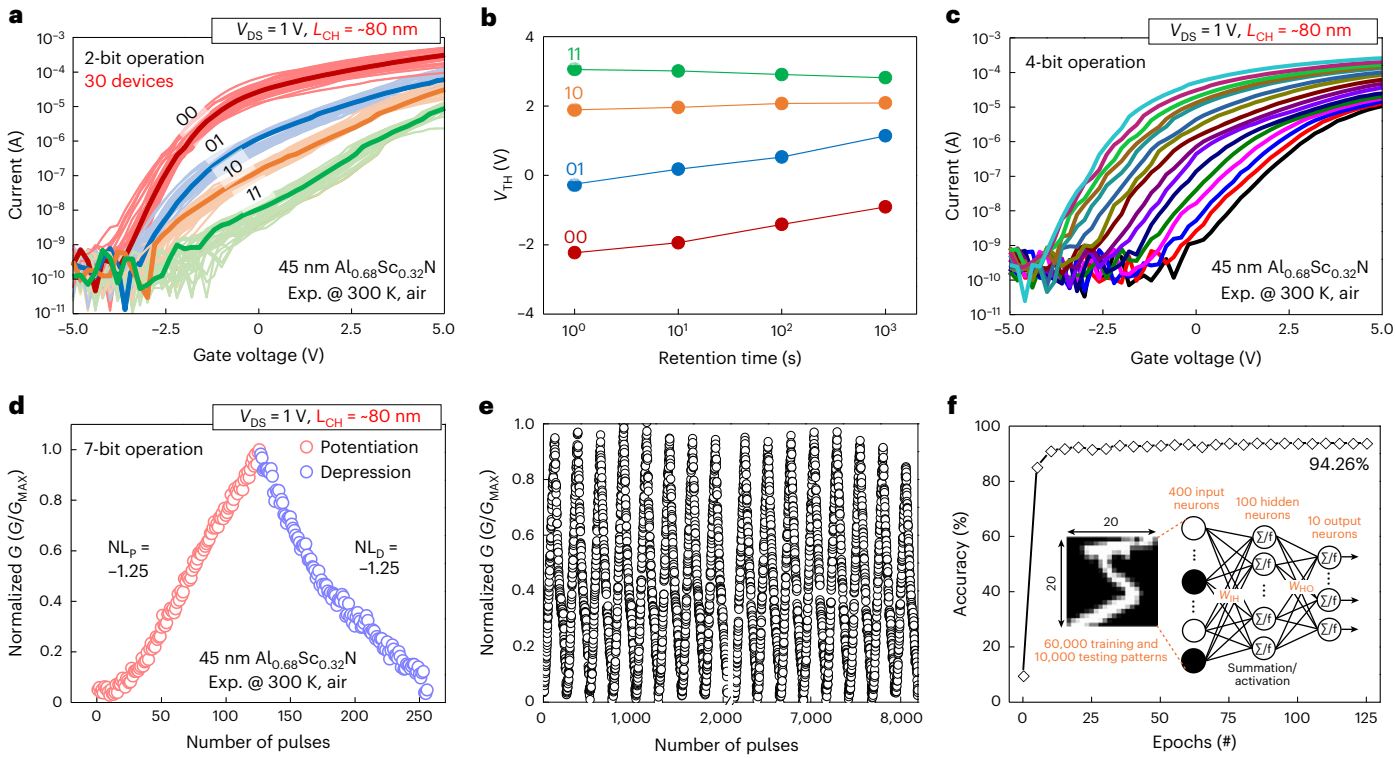

Fig. 4 | Multibit operation of scaled FE-FET devices. a, A 2-bit operation measured from 30 scaled FE-FETs after applying PRG or ERS pulses with an amplitude of 10–12 V and a width of 40 ms. **b**, Retention of the 2-bit V_{TH} states up to 1,000 s. **c**, A 4-bit operation measured from the scaled FE-FETs after applying PRG or ERS pulses with an amplitude of 9–12 V and a width of 40 ms. **d**, Normalized 7-bit LTP/LTD curves obtained from the scaled FE-FET. A positive pulse with 150 µs width and 15 V amplitude is used for LTP and a negative pulse with 150 µs width and −1 to −15 V amplitude is used for LTD. NL_P and NL_D refer to potentiation non-linearity and depression non-linearity, respectively. **e**, Cycle-to-cycle variations of the LTP/LTD curve for over 30 cycles (total, over 8,000 input pulses). **f**, Recognition rate as a function of the number of training epochs based on the LTP/LTD curve in **d**, and the inset shows a schematic of an MLP-based ANN with a size of 400 × 100 × 10. W_{IH} and W_{HO} refer to weight of input to hidden layer and weight of hidden layer to output layer, respectively.

volatile memory such as dynamic random-access memory by monolithically building dense, high-performance and fast NVM on Si logic layers.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41565-023-01399-y.

References

1. Migliato Marega, G. et al. Logic-in-memory based on an atomically thin semiconductor. *Nature* **587**, 72–77 (2020).
2. Wang, Z. et al. Resistive switching materials for information processing. *Nat. Rev. Mater.* **5**, 173–195 (2020).
3. Yang, R. et al. Ternary content-addressable memory with MoS_2 transistors for massively parallel data search. *Nat. Electron.* **2**, 108–114 (2019).
4. Shulaker, M. M. et al. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature* **547**, 74–78 (2017).
5. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
6. Dutta, S. et al. Monolithic 3D integration of high endurance multi-bit ferroelectric FET for accelerating compute-in-memory. In *2020 IEEE International Electron Devices Meeting (IEDM)* 36.4.1–36.4.4 (IEEE, 2020).
7. Khan, A. I., Keshavarzi, A. & Datta, S. The future of ferroelectric field-effect transistor technology. *Nat. Electron.* **3**, 588–597 (2020).
8. Akinwande, D. et al. Graphene and two-dimensional materials for silicon technology. *Nature* **573**, 507–518 (2019).
9. Polyushkin, D. K. et al. Analogue two-dimensional semiconductor electronics. *Nat. Electron.* **3**, 486–491 (2020).
10. Ni, K. et al. Ferroelectric ternary content-addressable memory for one-shot learning. *Nat. Electron.* **2**, 521–529 (2019).
11. Wang, D. et al. Ferroelectric switching in sub-20 nm aluminum scandium nitride thin films. *IEEE Electron Device Lett.* **41**, 1774–1777 (2020).
12. Liu, X. et al. Post-CMOS compatible aluminum scandium nitride/2D channel ferroelectric field-effect-transistor memory. *Nano Lett.* **21**, 3753–3761 (2021).
13. Tsai, S.-L. et al. Room-temperature ÿdeposition of a poling-free ferroelectric AlScN film by reactive sputtering. *Appl. Phys. Lett.* **118**, 082902 (2021).
14. Wang, D. et al. Sub-microsecond polarization switching in (Al,Sc)N ferroelectric capacitors grown on complementary metal-oxide-semiconductor-compatible aluminum electrodes. *Phys. Status Solidi RRL* **15**, 2000575 (2021).
15. Islam, M. R. et al. On the exceptional temperature stability of ferroelectric $Al_{1-x}Sc_xN$ thin films. *Appl. Phys. Lett.* **118**, 232905 (2021).
16. Fichtner, S., Wolff, N., Lofink, F., Kienle, L. & Wagner, B. AlScN: a III-V semiconductor based ferroelectric. *J. Appl. Phys.* **125**, 114103 (2019).

17. Lederer, M. et al. Local crystallographic phase detection and texture mapping in ferroelectric Zr doped HfO_2 films by transmission-EBSD. *Appl. Phys. Lett.* **115**, 222902 (2019).
18. Dragoman, M. et al. Ferroelectrics at the nanoscale: materials and devices—a critical review. *Crit. Rev. Solid State Mater. Sci.* 1–19 (2022).
19. Siao, M. D. et al. Two-dimensional electronic transport and surface electron accumulation in MoS_2. *Nat. Commun.* **9**, 1442 (2018).
20. Mulaosmanovic, H. et al. Ferroelectric field-effect transistors based on HfO_2: a review. *Nanotechnology* **32**, 502002 (2021).
21. Mikolajick, T. et al. Next generation ferroelectric materials for semiconductor process integration and their applications. *J. Appl. Phys.* **129**, 100901 (2021).
22. Aljarb, A. et al. Ledge-directed epitaxy of continuously self-aligned single-crystalline nanoribbons of transition metal dichalcogenides. *Nat. Mater.* **19**, 1300–1306 (2020).
23. Sebastian, A., Pendurthi, R., Choudhury, T. H., Redwing, J. M. & Das, S. Benchmarking monolayer MoS_2 and WS_2 field-effect transistors. *Nat. Commun.* **12**, 693 (2021).
24. Zhang, Y., Brar, V. W., Girit, C., Zettl, A. & Crommie, M. F. Origin of spatial charge inhomogeneity in graphene. *Nat. Phys.* **5**, 722–726 (2009).
25. Liu, Y.-S. & Su, P. Variability analysis for ferroelectric FET nonvolatile memories considering random ferroelectric-dielectric phase distribution. *IEEE Electron Device Lett.* **41**, 369–372 (2020).
26. Lederer, M. et al. Ferroelectric field effect transistors as a synapse for neuromorphic application. *IEEE Trans. Electron Devices* **68**, 2295–2300 (2021).
27. Luo, Y et al. MLP+NeuroSimV3.0: improving on-chip learning performance with device to algorithm optimizations. In *ICONS '19: Proc. International Conference on Neuromorphic Systems* 1–7 (ACM, 2019).
28. Ko, C. et al. Ferroelectrically gated atomically thin transition-metal dichalcogenides as nonvolatile memory. *Adv. Mater.* **28**, 2923–2930 (2016).
29. Xu, L. et al. Ferroelectric-modulated MoS_2 field-effect transistors as multilevel nonvolatile memory. *ACS Appl. Mater. Interfaces* **12**, 44902–44911 (2020).
30. Young Tack Lee, H. K. et al. Nonvolatile ferroelectric memory circuit using black phosphorus nanosheet-based field-effect transistors with P(VDF-TrFE) polymer. *ACS Nano* **9**, 10394–10401 (2015).
31. Jiang, X. et al. Ferroelectric field-effect transistors based on WSe_2/$CuInP_2S_6$ heterostructures for memory applications. *ACS Appl. Electron. Mater.* **3**, 4711–4717 (2021).
32. Si, M., Liao, P. Y., Qiu, G., Duan, Y. & Ye, P. D. Ferroelectric field-effect transistors based on MoS_2 and $CuInP_2S_6$ two-dimensional van der Waals heterostructure. *ACS Nano* **12**, 6700–6705 (2018).
33. Wang, X. et al. Ferroelectric FET for nonvolatile memory application with two-dimensional $MoSe_2$ channels. *2D Mater* **4**, 025036 (2017).
34. Liu, L. et al. Electrical characterization of MoS2 field-effect transistors with different dielectric polymer gate. *AIP Adv* **7**, 065121 (2017).
35. Jiawen, X. et al. Experimental demonstration of HfO_2-based ferroelectric FET with MoS_2 channel for high-density and low-power memory application. In *2021 Silicon Nanoelectronics Workshop (SNW)* 1–2 (IEEE, 2021).
36. Huang, K. et al. $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric embedded dual-gate MoS_2 field effect transistors for memory merged logic applications. *IEEE Electron Device Lett.* **41**, 1600–1603 (2020).
37. Zhang, S. et al. Low voltage operating 2D MoS_2 ferroelectric memory transistor with $Hf_{1-x}Zr_xO_2$ gate structure. *Nanoscale Res. Lett.* **15**, 157 (2020).

Methods

Synthesis of MoS_2 by CVD 1

The monolayer MoS_2 films were synthesized on sapphire substrates on a two-inch wafer using a conventional CVD approach, which was adopted from previous work[22]. In essence, high-purity MoO_3 powders (Sigma-Aldrich, 99.90%) and S powders (Sigma-Aldrich, 99.99%) were used as the precursor. The MoO_3 powders were placed in a ceramic crucible located at the centre of the furnace, whereas the S powders were annealed by a heating tape at the upstream side. During the reaction, the furnace was heated to 800 °C and the heating tape was heated to 140 °C; meanwhile, 70 s.c.c.m. Ar gas was introduced into the furnace to transport the precursors to the downstream side where the sapphire substrates were located. The reaction was maintained for 15 min at 30 torr.

Synthesis of MoS_2 by MOCVD

Fully coalesced MoS_2 monolayers were also grown on two-inch *c*-plane sapphire substrates using a horizontal MOCVD system from CVD Equipment. $Mo(CO)_6$ and H_2S were served as the Mo and S sources, respectively, with the former supplied from a stainless steel bubbler maintained at 10 °C and 650 torr. During the growth, 1.1×10^{-3} s.c.c.m. $Mo(CO)_6$, 400 s.c.c.m. H_2S and 4,100 s.c.c.m. H_2 were fed into the reactor for 12 min, which was held at 900 °C and 50 torr. After the growth, the film was annealed for 10 min under H_2S before cooling down.

Synthesis of MoS_2 by CVD 2

For sulfurization, H_2S (g), H_2 (g) and MoO_{3-x} (g) were used at around 690 °C. The temperature has various steps to initiate nucleation followed by enhancing lateral growth. Sapphire samples are subject to even higher-temperature treatments before the growth to ensure predictable nucleation characteristics.

AlScN deposition

The depositions of 45 and 100 nm AlScN were performed on four-inch Pt(111)/Ti/SiO_2/Si or Al(111)/sapphire wafers via 150 kHz pulsed d.c. co-sputtering with 20 s.c.c.m. N_2 flow under 8.3×10^{-4} mbar in an Evatec CLUSTERLINE 200 II pulsed d.c. sputtering system. The chamber temperature was maintained at 350 °C, a BEOL-compatible thermal budget.

PUND and *P*–*E* loop measurement

For the PUND and *P*–*E* loop measurement, metal–FE–metal (Pt/AlScN/Al) capacitors were used. The PUND measurements were taken on 20-µm-radii capacitors using a Keithley 4200A-SCS analyser. The voltage waveform of the PUND test consists of four monopolar pulses with a pulse width of 500 ns, rise or fall times of 140 ns and delay (interval between subsequent pulses) of 1 µs. Voltages (positive or negative) were applied to the bottom electrode of the metal–FE–metal capacitors and current was sensed from the top electrode to minimize current transients in the data due to parasitic capacitances. The Keithley analyser integrates the current measured in each pulse and returns the associated change in charge, which is related to polarization through a division by the area of the capacitor. In post-measurement processing using MATLAB, the remnant polarization for the applied positive voltages (that is, P and U pulses) was obtained from the PUND measurements by subtracting the U polarization from the P polarization, whereas the remnant polarization for the applied negative voltages (that is, the N and D pulses) was obtained by subtracting the D polarization from the N polarization. The motivation behind such a measurement is first that switching from an upward-polarization (saturated) state to a downward-polarization (saturated) state requires a change in polarization of $-2 \times P_r$, where P_r is the average remnant polarization of the FE material, and the reverse requires a change in polarization of $+2 \times P_r$. The second motivation is that by subtracting the U pulse from the P pulse, or the D pulse from the N pulse, one can reduce the contribution of leakage current—which becomes large as the applied voltages approach the breakdown limit—as well as of the capacitive current from the measured polarization of the FE material.

The *P*–*E* loop measurements were performed on 20-µm-radii capacitors using a Radiant Precision Premier II FE tester. The applied voltage utilized the standard bipolar profile: the voltage was monotonically (linearly) ramped from 0 V to the maximum voltage V_{MAX}, and then decreased monotonically to $-V_{MAX}$ before being increasing back to 0 V. The period of the voltage profile was 0.1 ms, corresponding to a measurement frequency of 10 kHz. The current density–voltage characteristics were then extracted from the polarization–voltage data by dividing the change in polarization at successive measurement points by the time interval between those measurements. The electric field was then obtained by dividing the applied voltage by the thickness (45 nm) of the metal–FE–metal capacitors.

Fabrication of large-area MoS_2/AlScN FE-FETs

First, large-area MoS_2 having a size of around 1 cm^2 grown on sapphire was transferred by a wet transfer method. The MoS_2-transferred sample was dried in a glove box for a day. Then, the sample was coated using PMMA A4 and PMMA A8 or PMGI SF 5S and ZEP 520A followed by source/drain patterning by electron-beam lithography. After being developed using methyl isobutyl ketone or PMGI 101A, 10 nm Ti and 30 nm Au were deposited as the source/drain contact metal and pad metal, respectively, using electron-beam evaporation. The samples were immersed in acetone or Remover 1165 for approximately 20 min, gently shaken to lift the metal and then rinsed with isopropyl alcohol and deionized water. Next, to define the channel area, a second patterning step by photolithography was done after coating the photoresists (LOR 3A and S1813), followed by developing with AZ 300MIF. Finally, the exposed area of MoS_2 was etched using oxygen reactive ion etching.

Electrical measurement of FE-FETs

Electrical measurements were performed in air at ambient temperature in a Lake Shore probe station using a Keithley 4200A semiconductor characterization system. For the d.c. *I*–*V* characteristics, the source measurement unit connection was used, whereas both phasor measurement unit and source measurement unit connections were used for the voltage-pulse-induced FE switching measurements.

Scanning TEM and SEM characterizations

Scanning TEM characterization and image acquisition were carried out on a JEOL F200 instrument operated at 200 kV acceleration voltage. Energy-dispersive X-ray spectroscopy analysis was performed using a JEOL NEOARM TEM instrument operated at a voltage of 200 kV, with a point resolution better than 0.08 nm. This is a powerful technique, which can detect differences in composition on the atomic scale. All the captured scanning TEM images were collected/calculated using DigitalMicrograph software version GMS 3.5 (Gatan). A BrightBeam SEM instrument was used inside the dual-beam plasma focused-ion-beam system (TESCAN S8000X). The SEM instrument operated at 5 keV utilized an in-lens backscattered middle detector (MD) to capture the high-resolution image for a better analysis of the ~80-nm-channel-length devices.

TEM/scanning TEM sample preparation

The TEM cross-sectional sample was prepared by a Xe^+ plasma focused-ion-beam (TESCAN S8000X) system. The sample surface was coated with electron-beam- and ion-beam-deposited Pt protection layers to prevent damaging the top surfaces and heating effects during focused-ion-beam milling. The focused-ion-beam lamella was milled at 30 keV and further in situ lift-out technique was used with the help of a Kleindiek probe manipulator. The final thinning and cleaning of the lamella was performed at 10 and 5 keV, respectively.

Wet transfer

MoS_2 grown on a two-inch wafer was first cut to 1 cm^2 size. The 1-cm^2-sized MoS_2 was coated by PMMA 4 at 2,000 r.p.m. for 60 s. After soaking the sample in 90 °C water for about 10 min, potassium hydroxide was used to separate MoS_2 from sapphire. The detached MoS_2 was floated on deionized water for about 15 min to clean the potassium hydroxide. Finally, MoS_2 was transferred on AlScN in deionized water.

MoS_2 etching using reactive ion etching

For the defined channel area, MoS_2 was etched using a March Jupiter II etcher. The power of the plasma was 100 W, O_2 gas flow was at 450 s.c.c.m. and the run time was 30 s.

Data availability

All data needed to evaluate the conclusions of this study are present in the Article and its Supplementary Information.

Acknowledgements

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) TUFEN program under agreement no. HR00112090046. The work was carried out in part at the Singh Center for Nanotechnology at the University of Pennsylvania, which is supported by the National Science Foundation (NSF) National Nanotechnology Coordinated Infrastructure Program (NSF grant NNCI-1542153). H.M.K., K.K. and D.J. acknowledge partial support from the Penn Center for Undergraduate Research and Fellowships. We gratefully acknowledge the use of the facilities and instrumentation supported by NSF through the University of Pennsylvania Materials Research Science and Engineering Center (MRSEC) (DMR-1720530). P.K., E.A.S. and D.J. also acknowledge partial support from the NSF DMR Electronic Photonic and Magnetic Materials (EPM) core program (grant no. DMR-1905853) as well as the University of Pennsylvania Laboratory for Research on the Structure of Matter, a Materials Research Science and Engineering Center (MRSEC) supported by the NSF (no. DMR-1720530). A.A., Y.W. and V.C.T. are indebted to the support from the King Abdullah University of Science and Technology (KAUST) Solar Center and Office of Sponsored Research (OSR) under award no. OSR-2018-CARF/CCF-3079. The MOCVD-grown MoS_2 monolayer samples were provided by the 2D Crystal Consortium—Materials Innovation Platform (2DCC-MIP) facility at the Pennsylvania State University, which is funded by the NSF under cooperative agreement nos. DMR-1539916 and DMR-2039351. S.S. acknowledges support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant no. 2021R1A6A3A14038492). K.K. acknowledges support from the NSF Graduate Research Fellowship Program (GRFP), Fellow ID: 2022338725. N.T. acknowledges that this material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE1255832.

Author contributions

D.J., R.H.O. and K-H.K. conceived the idea of using large-area 2D semiconductors with AlScN to make the CMOS BEOL-compatible FE-FETs at scale and with scaled dimensions. K.-H.K. designed the experiments and performed the device fabrication and characterization of the samples. K.-H.K. and D.J. wrote the manuscript. M.M.A.F. and K.-H.K. conducted the *P–E* loop and PUND measurements and R.H.O. supervised them. J.Z. and K.-H.K. performed sputtering to prepare the various AlScN substrates and R.H.O. supervised them. P.M. and P.K. performed the TEM and SEM characterizations, respectively. P.K. prepared the cross-sectional lamella for the subsequent TEM observation. E.A.S. supervised the microscopy efforts. N.T. prepared the two-inch wafer-scale MoS_2 using MOCVD and J.M.R. supervised it. CVD-based two-inch wafer-scale MoS_2 was prepared by A.A., Y.W., J.-H.F. and M.H., and V.T. supervised them. K.-H.K. and P.K. performed the wet transfer of MoS_2 on AlScN and SiO_2. S.O. and K.K. contributed to the MLP-based artificial neural network simulation and technology computer-aided design simulation. K.-H.K. and H.M.K. performed the electrical measurements of 130 FE-FET arrays. S.S. and G.K. performed the device fabrication for revision. Z.T. developed the recipe for electron-beam lithography. All the authors contributed to the discussion and analysis of the results.

Competing interests

D.J., R.H.O., K.-H.K. and E.A.S. are co-inventors on a patent (US Patent App. 17/354,256) based on this work. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41565-023-01399-y.

Correspondence and requests for materials should be addressed to Roy H. Olsson or Deep Jariwala.

Peer review information *Nature Nanotechnology* thanks Kah-Wee Ang and Weida Hu for their contribution to the peer review of this work.