Article

Subnanosecond flash memory enabled by 2D-enhanced hot-carrier injection

https://doi.org/10.1038/s41586-025-08839-w

Received: 26 September 2024

Accepted: 25 February 2025

Published online: 16 April 2025

Open access

Check for updates

Yutong Xiang[1,3], Chong Wang[1,3], Chunsen Liu[1,3]✉, Tanjun Wang[1], Yongbo Jiang[1], Yang Wang[1,2], Shuiyuan Wang[1] & Peng Zhou[1,2]✉

The pursuit of non-volatile memory with program speeds below one nanosecond, beyond the capabilities of non-volatile flash and high-speed volatile static random-access memory, remains a longstanding challenge in the field of memory technology[1]. Utilizing fundamental physics innovation enabled by advanced materials, series of emerging memories[2–5] are being developed to overcome the speed bottleneck of non-volatile memory. As the most extensively applied non-volatile memory, the speed of flash is limited by the low efficiency of the electric-field-assisted program, with reported speeds[6–10] much slower than sub-one nanosecond. Here we report a two-dimensional Dirac graphene-channel flash memory based on a two-dimensional-enhanced hot-carrier-injection mechanism, supporting both electron and hole injection. The Dirac channel flash shows a program speed of 400 picoseconds, non-volatile storage and robust endurance over 5.5×10^6 cycles. Our results confirm that the thin-body channel can optimize the horizontal electric-field (E_y) distribution, and the improved E_y-assisted program efficiency increases the injection current to 60.4 pA μm^{-1} at $|V_{DS}| = 3.7$ V. We also find that the two-dimensional semiconductor tungsten diselenide has two-dimensional-enhanced hot-hole injection, but with different injection behaviour. This work demonstrates that the speed of non-volatile flash memory can exceed that of the fastest volatile static random-access memory with the same channel length.

In light of the accelerated advancement of artificial intelligence, there is an urgent requirement for non-volatile data memory technology with a sub-1-ns speed to overcome the bottleneck of computing energy efficiency[1,11,12]. The fastest static random-access memory (SRAM) can achieve sub-1-ns speed, but its stored data are volatile and its has poor energy efficiency and storage density. Although the mainstream non-volatile flash memory offers large non-volatile storage capacity and low manufacturing cost[13–15], the program speed of flash memory is still far from that of the volatile memory. To overcome the speed bottleneck of non-volatile memory, series of emerging non-volatile memories, including phase-change memory[2,3], resistive switching memory[16], ferroelectric memory[4,17,18] and magnetic random-access memory[5], are being developed. All of the emerging memories indicate the presence of innovative fundamental mechanisms in advanced materials, enabling unprecedented speeds. However, it should be noted that these technologies also have their own set of challenges. The International Roadmap for Devices and Systems[19] indicates several key challenges, including thermal disturbance for phase-change memory, statistical fluctuation for resistive switching memory, complementary metal–oxide–semiconductor processing degradation for ferroelectric memory, and difficult miniaturization for magnetic random-access memory. Although many emerging memory technologies have been proposed, it is widely believed that the most extensively applied flash memory cannot work at a program speed below 1 ns.

The fundamental physics of flash memory is based on the electric-field design, including both the channel-to-gate direction (vertical electric field, $E_{dielectric}$) and the source-to-drain direction (horizontal electric field, E_y). The $E_{dielectric}$ and E_y design correspond to the Fowler–Nordheim (FN) tunnelling and hot-carrier-injection mechanisms of flash memory, respectively. Specifically, the FN tunnelling efficiency of conventional silicon-based flash is limited by the barrier height, which restricts its speed to a range of 10–100 μs (ref. 6). Recently, it has been shown that in two-dimensional (2D) semiconductors, the FN tunnelling speed can be increased to 10–20 ns at about 15 V by reducing the effective barrier height[7–9,20–22], but this speed remains insufficient and the operating voltage is still high. Another important approach is to exploit the E_y of the channel to accelerate the carriers to a high-energy state, allowing these 'hot' carriers to be injected across the barrier and improve the program speed of silicon flash memory[10,23]. However, in the silicon material system, the limited E_y acceleration efficiency is only capable of supporting a program speed of tens of nanoseconds.

In this study, we found a channel-thickness-modulated E_y distribution effect, which exploits the atomically thin properties of 2D materials to effectively increase the maximum value of E_y ($E_{y,max}$) and facilitate 2D-enhanced hot-carrier injection (2D-HCI). The injection current was

[1]State Key Laboratory of Integrated Chips and Systems, College of Integrated Circuits and Micro-Nano Electronics, Frontier Institute of Chip and System, Zhangjiang Fudan International Innovation Center, Fudan University, Shanghai, China. [2]Shaoxin Laboratory, Zhejiang, China. [3]These authors contributed equally: Yutong Xiang, Chong Wang, Chunsen Liu. ✉e-mail: chunsen_liu@fudan.edu.cn; pengzhou@fudan.edu.cn

observed to be orders of magnitude higher in the 2D material than in the silicon material system. Furthermore, distinct injection behaviours were observed on a 2D Dirac material (graphene) and a 2D semiconductor (tungsten diselenide (WSe_2)). Utilizing the 2D-HCI mechanism, we developed sub-1-ns flash memory with different channel lengths (L_{ch}), and the injection efficiency increased with the scaling down of devices. The graphene device with a short channel (L_{ch} = 0.2 μm) showed a program speed of 400 ps while maintaining a low program voltage of $|V_{PROG}|$ = 5 V.

The material-dependent hot-carrier injection

It is known that the hot-carrier-injection effect exists in silicon technology. This phenomenon entails carriers to keep gaining energy by accelerating from the source along the channel under E_y. When the acquired energy surpasses the barrier between the channel and the dielectric, a portion of the 'hot' (high energy) carriers will be injected into the gate under $E_{dielectric}$, corresponding to channel-to-gate injection.

Interestingly, we found that 2D semiconductor and Dirac transistors show a material-dependent behaviour of hot-carrier injection, which differs from previous studies[4]. Figure 1a,b shows schematic diagrams of the 2D transistor structures used in this work, comprising a bottom control gate, a hexagonal boron nitride (hBN) dielectric and a thin-body channel. A semimetal technology (antimony (Sb)/platinum (Pt)) was used to obtain a p-type quasi-ohmic contact in the WSe_2 transistors, as shown in Fig. 1a. The holes are continuously accelerating to gain more energy but this process will lose part of its energy owing to scattering. Different from 2D semiconductors, the 2D Dirac material shows a carrier effective mass approaching zero. In addition, it has a longer mean free path (approximately mircometres) and a lower carrier scattering probability[24]. Consequently, both electrons and holes in the graphene transistor are accelerating along the channel with scattering suppressed (Fig. 1b). Detailed information on the device fabrication process and characterization are provided in Supplementary Information section 1.

The detailed carrier acceleration process is dependent on the E_y and the horizontal potential (V_y) distributions. The E_y and V_y of the WSe_2 channel and graphene channel in the transistor are illustrated in Fig. 1c,d. When applying a negative gate–source bias (V_{GS}) larger than the sum of the drain–source bias (V_{DS}) and the threshold voltage (V_{th}) in the WSe_2 transistor (Fig. 1c), for example, $|V_{GS} - V_{th}| \approx |V_{DS}|$, the entire channel becomes conductive. Given that the conductivity of the channel decreases from the source to the drain, it can be observed that $|E_y|$ shows an increasing trend in this direction. As the integral of E_y in the channel direction is V_y, it follows that $|V_y|$ exhibits a nonlinear increasing trend. In a graphene transistor (Fig. 1d), the E_y and V_y distributions show some differences from the WSe_2 channel transistor. In the graphene device, here we apply the same electrical conditions as Fig. 1c, denoted as $|V_{GS} - V_{Dirac}| \approx |V_{DS}|$, where V_{Dirac} is the gate voltage at the lowest drain–source current (I_{DS}). At this time, E_y increases more slowly, resulting in a quasi-linear V_y distribution. This is because the high-resistance region would not appear in the graphene channel, as has been observed in other reported studies[25]. The $|E_y|$ value of the graphene transistor in most of the channel region is greater than that of WSe_2, corresponding to better front acceleration (80% of L_{ch}).

To quantify the injection current of the carriers from channel to gate, we measured the injection current under different V_{GS} and V_{DS} configurations in the WSe_2 and graphene transistors, as shown in Fig. 1e,f. For the p-type WSe_2 transistor (Fig. 1e), V_{th} is determined by the linear extrapolation method[26–28]. When a negative bias is applied to the drain, as ($V_{GS} - V_{th}$) becomes more negative, the hole injection current initially increases. This is because a decrease in ($V_{GS} - V_{th}$) is conducive to generating more accelerating holes and the enhanced $E_{dielectric}$ will also collect the holes more efficiently. When $|V_{GS} - V_{th}| = |V_{DS}|$, the hole injection current reaches a maximum value (Supplementary Fig. 11). The further reduction in the ($V_{GS} - V_{th}$) will lead to the disappearance of the high-resistance pinch-off region, quickly decreasing the maximum value of E_y ($E_{y,max}$) and the injection current, which is similar to that of silicon transistors[29].

The Dirac transistor shows behaviour that is distinct from that of the semiconductor transistors. As shown in Fig. 1f, both holes and electrons can be efficiently accelerated with either a negative or positive V_{DS}. For the hole acceleration, as ($V_{GS} - V_{Dirac}$) decreases, the injection current increases monotonically, because the increasing density of holes leads to more hot holes. It should be noted that the graphene transistor has no high-resistance pinch-off region, which means that ($V_{GS} - V_{Dirac}$) decreases will not render a significant reduction of the $E_{y,max}$. Therefore, the injection current of the graphene transistor shows a monotonic trend as a function of ($V_{GS} - V_{Dirac}$), and has no maximum value at the $|V_{GS} - V_{Dirac}| = |V_{DS}|$ condition. For the electron acceleration, a similar monotonic characteristic is observed.

We have also found that the scaling down of the device can improve the injection efficiency (Supplementary Fig. 15). In Fig. 1e,f, the channel length of the silicon transistor[29] (L_{ch} = 0.8 μm) is slightly shorter than that of the WSe_2 transistor (L_{ch} = 1 μm) and the graphene transistor (L_{ch} = 1.5 μm). The injection behaviour of the 2D semiconductor and Dirac materials is different, but all the 2D material transistors show a much higher injection current than that of the bulk silicon transistor.

The 2D-enhanced hot-carrier-injection mechanism

To further explore the factors that make the injection efficiency of 2D materials transistors better than that of silicon transistors, we employed a quasi-2D model[30,31] for analysis. As Fig. 2a shows, the structure of this physical model is consistent with that of our fabricated 2D transistor. The x direction is defined as the channel surface to gate vertically and the y direction is defined as source to drain horizontally, where the zero point of the x and y axes is the channel surface and source, respectively. $V(y)$ is the horizontal potential in the channel, t_{ch} is the channel thickness, E_x and E_y are the vertical and horizontal electric field in the channel, respectively, $E_{dielectric}$ is the vertical electric field in the dielectric, and E_{sat} and V_{DSsat} are the E_y and $V(y)$ at the leading edge of the velocity saturation region. Different from the triangle pinch-off region in the conventional bulk silicon, the thin body ($t_{ch} \to 0$) of the 2D material channel results in a square pinch-off region, denoted as Gaussian box, and E_y only varies with horizontal position. Furthermore, the value of E_x at $x = 0$ is close to zero, which allows us to replace $\frac{\partial E_x}{\partial x}$ with the average value given by $\frac{E_x(t_{ch}, y)}{t_{ch}}$. The 2D Poisson equation can be simplified as:

$$\frac{dE_y}{dy} + \frac{E_x(t_{ch}, y)}{t_{ch}} = \frac{\rho(x, y)}{\varepsilon_0 \varepsilon_{ch}} \quad (1)$$

where $\rho(x, y)$ is the charge density enclosed in the Gaussian box, ε_0 is the vacuum permittivity and ε_{ch} is the channel relative permittivity. According to the boundary condition of the electric field, $E_x(t_{ch}, y) = \frac{\varepsilon_{dielectric}}{\varepsilon_{ch}} E_{dielectric}(t_{ch}, y)$, equation (1) is simplified as

$$\frac{dE_y}{dy} = \frac{V(y) - V_{DSsat}}{\lambda^2}, \text{ where } \lambda^2 = \frac{\varepsilon_{ch} t_{dielectric} t_{ch}}{\varepsilon_{dielectric}} \quad (2)$$

$\varepsilon_{dielectric}$ is the dielectric relative permittivity, $t_{dielectric}$ is the dielectric thickness and λ is the effective length of the Gaussian box. It is noted that λ is also known as scale length[32]. Given the strong sensitivity of hot-carrier injection to $E_{y,max}$ (ref. 30), we have further applied the boundary conditions at the leading edge of the Gaussian box: $V(y) = V_{DSsat}$, and $E_y = E_{sat}$, and find that $|E_{y,max}|$ is given by

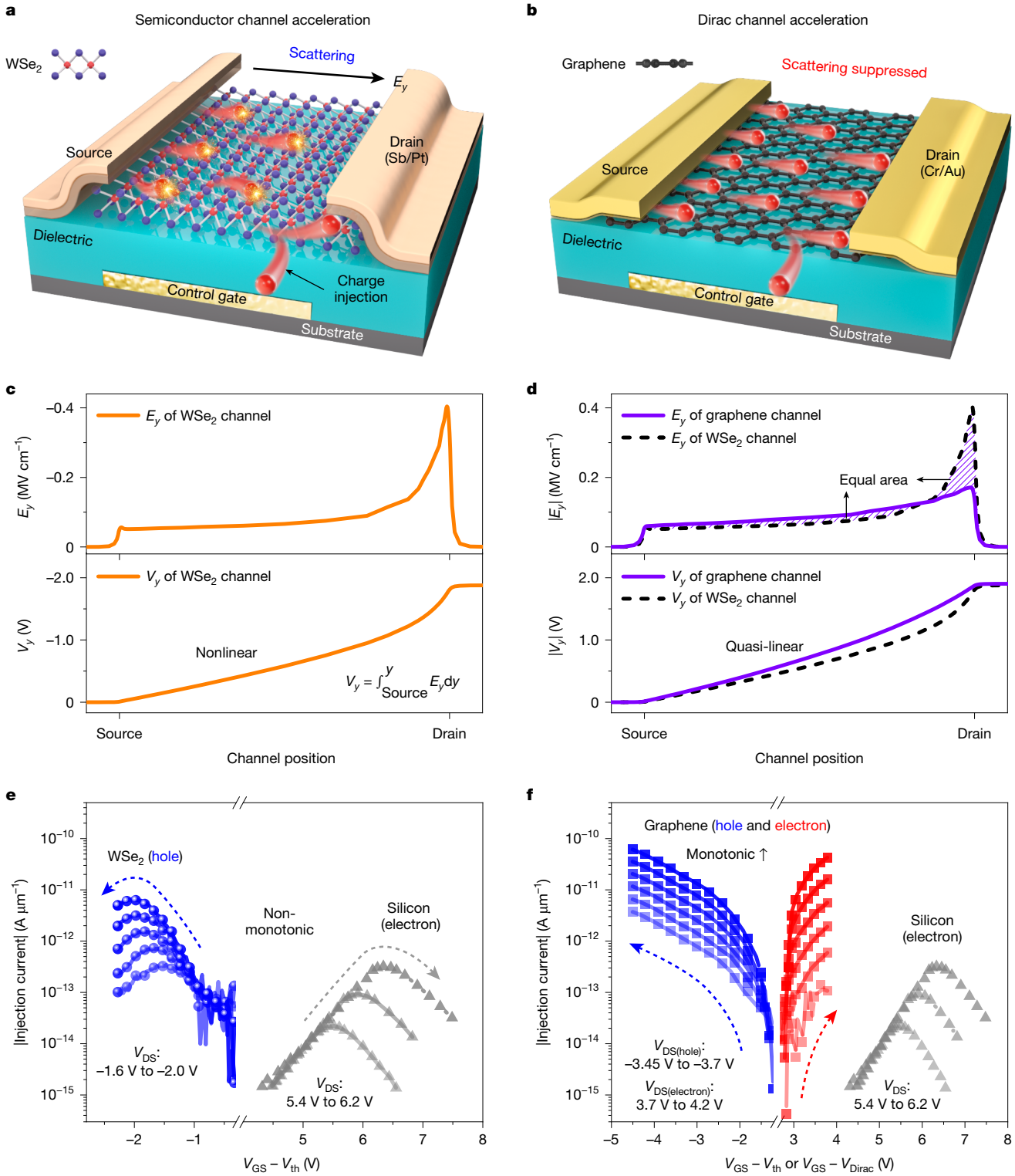

Fig. 1 | Material-dependent behaviour of hot-carrier injection. a, Schematic diagram showing the WSe_2 transistor with an Sb/Pt contact. The holes flow from the source to the drain with scattering under E_y, with some of them injected to the gate. **b**, Schematic of the graphene transistor structure. There are more carriers in the channel, and both electrons and holes can be accelerated with scattering suppressed. **c**, E_y and V_y distribution of a simulated WSe_2 transistor under the $|V_{GS} - V_{th}| \approx |V_{DS}|$ condition. V_y, denoted as the integral of E_y along the channel, shows a nonlinear change from the source to the drain. **d**, E_y and V_y distribution of a simulated graphene transistor under the same electrical conditions as in **c**. The black dashed line is from **c** for reference. The area of the shadow for the two parts in the E_y distribution is equal owing to the identical applied $|V_{DS}|$, and V_y shows a quasi-linear increase. **e**, Injection current of the hot hole in the WSe_2 transistor (L_{ch} = 1 μm, Sb/Pt contact; blue spheres). V_{DS} is −1.6 V to −2 V and the step is −0.1 V, from bottom to top. For the silicon hot electron (L_{ch} = 0.8 μm, grey triangles), V_{DS} is 5.4 V to 6.2 V and the step is 0.4 V (refs. 29,49), from bottom to top. The injection current shows a non-monotonic behaviour. **f**, Injection current of hot electron (red squares) and hole (blue squares) in the graphene transistor (L_{ch} = 1.5 μm, chromium (Cr)/gold (Au) contact; $V_{DS(hole)}$ is −3.45 V to −3.7 V with step of −0.05 V and $V_{DS(electron)}$ is 3.7 V to 4.2 V with a of step, 0.1 V, from bottom to top). The injection current shows a monotonic behaviour.

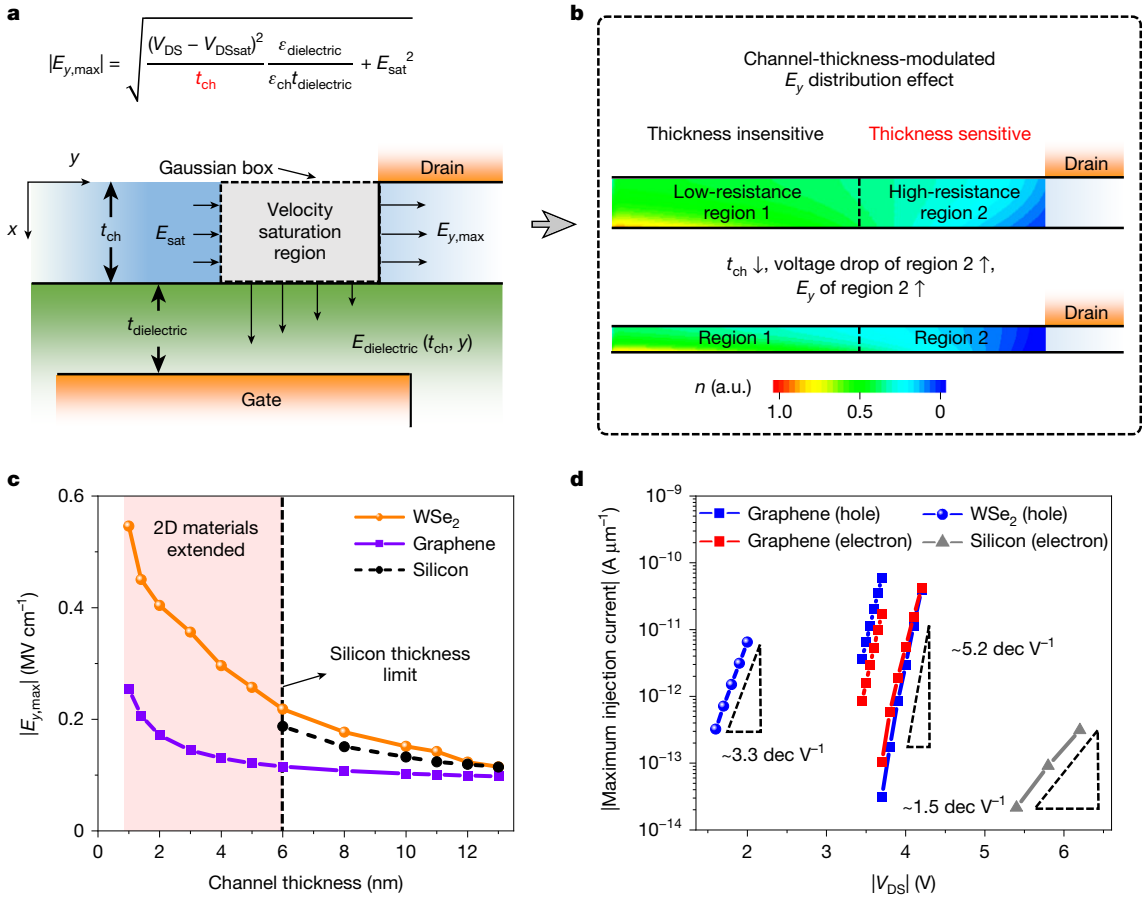

Fig. 2 | The 2D-HCI mechanism enabled by channel-thickness-modulated E_y distribution effect. a, Schematic diagram of the quasi-2D model. **b**, Schematic diagram of the channel-thickness-modulated E_y distribution effect, including carrier density (n) contour plots for a simulated WSe_2 device under different channel thickness conditions. The n of low-resistance region 1 is thickness insensitive, but thickness sensitive in the high-resistance region 2. **c**, Simulation of the relationship between $E_{y,max}$ and the channel thickness in the graphene transistor (purple line), WSe_2 transistor (orange line) and silicon transistor (black dashed line). **d**, Maximum injection currents versus $|V_{DS}|$ for WSe_2, graphene and silicon transistors. The injection current of the graphene and WSe_2 transistor is up to 60.4 pA μm^{-1} at $|V_{DS}| = 3.7$ V and 8.3 pA μm^{-1} at $|V_{DS}| = 2.25$ V, respectively, which is much higher than that of silicon transistors (313 fA μm^{-1} at $|V_{DS}| = 6.2$ V). It can be observed that even with a lower V_{DS}, the injection current of 2D transistors is orders of magnitude larger than that of silicon. Furthermore, the channel acceleration efficiency of 2D transistors is much higher than that of the silicon transistor. The spheres, boxes and triangles represent different materials (WSe_2, graphene and silicon); red and blue denote electrons and holes, respectively.

$$|E_{y,max}| = \sqrt{\frac{(V_{DS} - V_{DSsat})^2}{t_{ch}} \frac{\varepsilon_{dielectric}}{\varepsilon_{ch} t_{dielectric}} + E_{sat}^2} \quad (3)$$

For a detailed theoretical derivation, see Methods. In the silicon transistors, the previous simulation and experiment ($t_{ch} \geq 50$ nm) have shown a similar relationship[33–35].

Equation (3) shows that t_{ch} scaling will increase the value of $|E_{y,max}|$, and Fig. 2b further reveals the physics of this phenomenon. For a transistor in the ON state, the resistance of the channel increases from the source to the drain. As the velocity saturation region region has a higher resistance than the other part of the channel, we define this region as the high-resistance region 2, and the other part is defined as the low-resistance region 1. A reduction in t_{ch} results in an increase in the resistance of the channel but also increases the gate-control ability. As the control ability of V_{GS} in region 1 is much stronger than that in region 2, the resistance increase of region 1 is slowed by the gate generating new carriers. Because the resistance of region 2 is more sensitive to t_{ch} change, a reduction in t_{ch} results in a further increase in the resistance in region 2, leading to an increase in the voltage drop and E_y of region 2. This phenomenon is referred to as the channel-thickness-modulated E_y distribution effect. Both the Dirac material and semiconductor are applicable to this effect, and the only difference is that the region 2 of Dirac graphene is more conductive than that of the semiconductor.

We have further utilized the technology computer-aided design tool to numerically calculate and quantify the relationship between $|E_{y,max}|$ and t_{ch}. As Fig. 2c shows, we found that $|E_{y,max}|$ increases as the channel thickness decreases, where $|E_{y,max}| \propto t_{ch}^{-2/5}$ in graphene (purple line) and $|E_{y,max}| \propto t_{ch}^{-1/2}$ in WSe_2 (orange line). Although the $|E_{y,max}|$ in the graphene channel is smaller than that in the semiconductor channel, the graphene channel has better front-acceleration efficiency owing to its E_y distribution morphology. Figure 2c also illustrates that $|E_{y,max}|$ in the graphene transistor is more sensitive to thickness alterations when the material thickness is less than 3 nm. The specific simulation parameters are detailed in Supplementary Information section 3. According to the International Roadmap for Devices and Systems, the silicon thickness limit is around 6 nm in the sub-5-nm node[36], which limits its ultimate injection efficiency. Below this thickness, 2D materials are necessary to attain greater $|E_{y,max}|$ and realize 2D-HCI.

Here we present a comparison of the maximum injection current in 2D materials and bulk silicon (Fig. 2d). For WSe_2 transistors, the peak currents of the curves in Fig. 1e were selected, whereas, for graphene transistors, the maximum injection currents at every V_{DS} illustrated in Fig. 1f and Supplementary Fig. 14 were chosen. The maximum injection

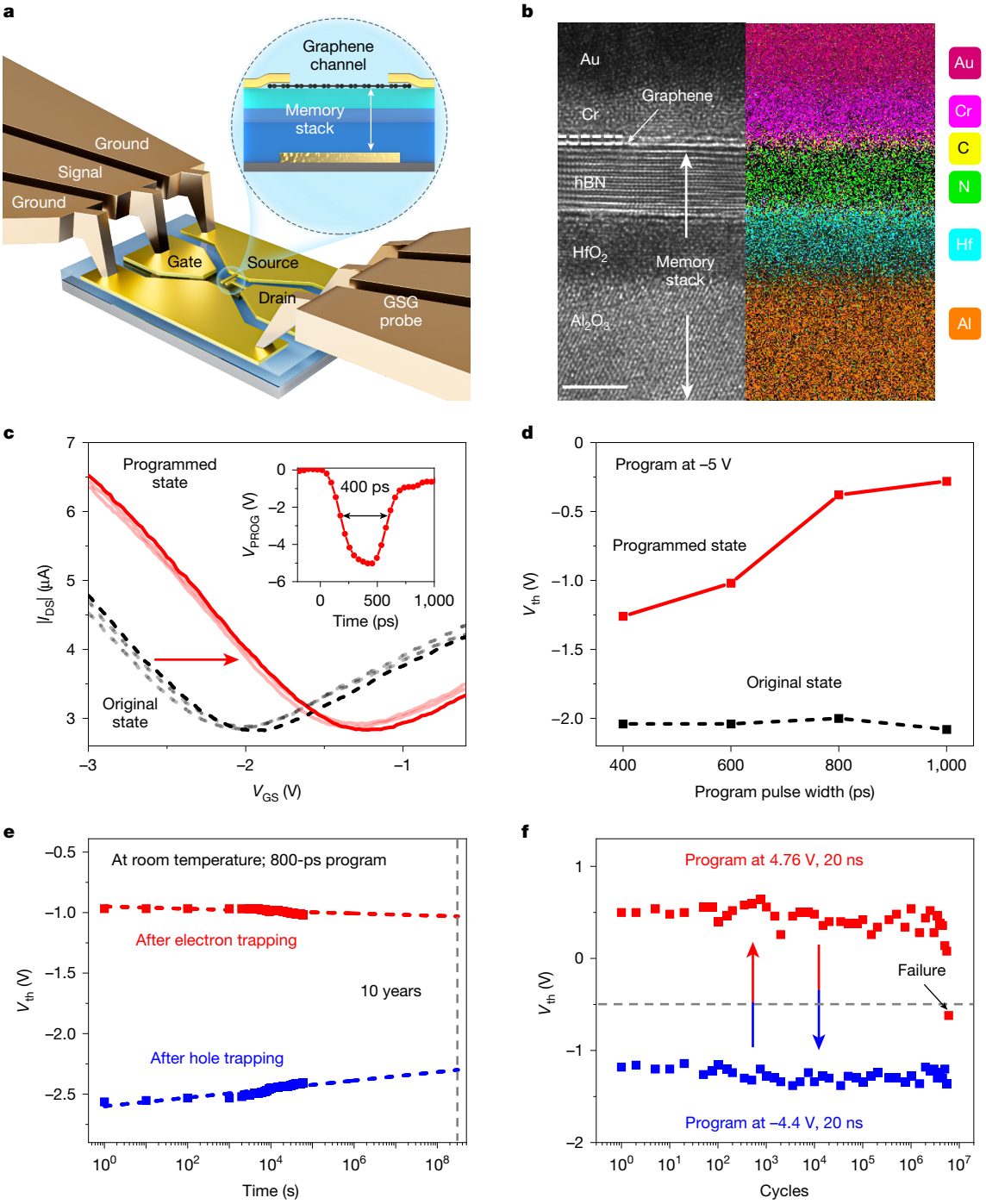

Fig. 3 | Memory performance of the sub-1-ns flash memory. a, Schematic diagram of the configuration of the GSG probes in conjunction with the flash device. Inset: a charge-trapping flash structure consisting of a control gate, a memory stack of hBN/HfO_2/Al_2O_3 and the graphene channel. **b**, Transmission electron microscopy image and elemental mapping image of the sub-1-ns flash device. The thicknesses of Al_2O_3, HfO_2, hBN and graphene are 20 nm, 5 nm, 6 nm and bilayer, respectively. Scale bar, 5 nm. **c**, The 400-ps programming performance of the device. The test was repeated three times (black and grey curves represent the original state; red and light red curves represent the programmed state). The transfer curves show a large memory window achieved by $V_{D,PROG}$ = −5 V, 400 ps with the grounded gate and source. Inset: the 400-ps program voltage waveform. **d**, Modulation of V_{th} through varying program pulse widths at $V_{D,PROG}$ = −5 V with the grounded gate and source. The black and red curves represent a typical original and programmed state, respectively. **e**, Data retention of the graphene flash memory. The device was measured at room temperature and programmed by $V_{G,PROG}$ = 3.5 V, $V_{D,PROG}$ = −6.8 V (for electron trapping) and $V_{G,PROG}$ = −1.8 V, $V_{D,PROG}$ = 6.6 V (for hole trapping), 800 ps, with source grounding. **f**, Endurance test of the device. The device was programmed using $V_{G,PROG}$ = $V_{D,PROG}$ = 4.76 V, 20 ns, and $V_{G,PROG}$ = $V_{D,PROG}$ = −4.4 V, 20 ns, respectively, with source grounding.

current shows a positive correlation with $|V_{DS}|$. A detailed discussion of the correlation between the injection current and V_{DS} (E_y) is provided in Supplementary Information section 4. The injection current of thin-body channel transistors increases more quickly with $|V_{DS}|$, 5.2 dec V^{-1}, 3.3 dec V^{-1} and 1.5 dec V^{-1} for the graphene transistor, WSe_2 transistor and silicon transistor, respectively. Dirac material graphene shows better injection performance because of its suppressed scattering effect and high carrier density. The injection current of the graphene

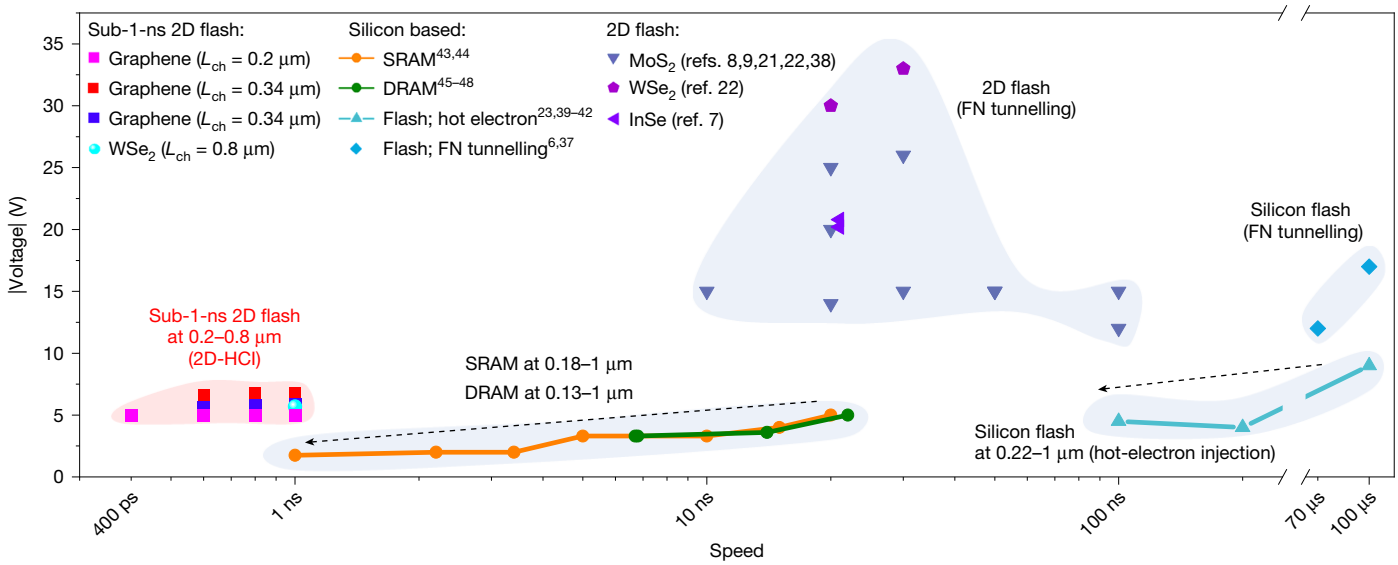

Fig. 4 | Benchmark for program voltage and speed in charge-based memory. The are five areas: silicon FN tunnelling flash[6,37], 2D flash based on the FN tunnelling mechanism[7–9,21,22,38], silicon flash based on the hot-electron injection mechanism[23,39–42], high-speed volatile memory[43–48], and sub-1-ns 2D flash based on the 2D-HCI mechanism. High-speed volatile memory (SRAM and DRAM) and silicon flash programmed by hot electrons are shown for different channel lengths, with the arrows indicating the advanced nodes. For our work, the solid pink squares represent graphene flash (L_{ch} = 0.2 μm) with 2D-enhanced hot-electron injection; the solid red and blue squares represent graphene flash (L_{ch} = 0.34 μm) with 2D-enhanced hot-electron and hot-hole injection, respectively; and the solid sky-blue sphere represents the WSe_2 flash (L_{ch} = 0.8 μm) with 2D-enhanced hot-hole injection. The voltage selections of the flash for comparison are: $V_{G,PROG}$ for flash based on the FN tunnelling mechanism, and $V_{D,PROG}$ for this work and silicon hot-electron program flash.

and WSe_2 transistor is up to 60.4 pA μm^{-1} at $|V_{DS}|$ = 3.7 V and 8.3 pA μm^{-1} at $|V_{DS}|$ = 2.25 V, respectively, which is much higher than that of silicon transistors (21.6 fA μm^{-1} at $|V_{DS}|$ = 5.4 V; 313 fA μm^{-1} at $|V_{DS}|$ = 6.2 V). It is observed that even at a lower V_{DS}, the injection current of 2D materials is orders of magnitude higher than that of silicon.

Sub-1-ns flash-memory performance

Utilizing the 2D-HCI mechanism, it is anticipated that the flash memory can break its sub-1-ns program speed bottleneck. The simultaneous occurrence of hot-electron and hot-hole channel acceleration in graphene makes it an ideal material for demonstrating the superior performance of the mechanism in flash memory. We have fabricated graphene flash memory based on a hBN/HfO_2/Al_2O_3 memory stack. To deliver sub-1-ns measurement, we used a radio-frequency probe with a ground–signal–ground (GSG) structure, where the signals are connected to the gate and drain terminal and the ground to the source terminal, as shown schematically in Fig. 3a. Two GSG probes are shorted by the calibration substrate to ensure that the high-speed test system can show a sub-1-ns program voltage (V_{PROG}) waveform (Supplementary Information section 5). The enlarged part of Fig. 3a depicts the detailed 2D flash structure, wherein the carriers can inject through the hBN layer into the HfO_2 layer by manipulating the gate program pulse ($V_{G,PROG}$) and drain program pulse ($V_{D,PROG}$). Figure 3b shows a transmission electron microscopy image of a typical bilayer graphene device, which shows that the fabrication process achieves an atomically flat interface in the heterostructure.

Figure 3c,d shows the performance of the flash memory with the 2D-HCI mechanism, exhibiting a program speed that can break the 1-ns bottleneck. When applying $V_{D,PROG}$ = −5 V, 400 ps with the grounded gate and source, the electrons in the channel can quickly achieve enough energy to inject into the HfO_2 trapping layer by the channel acceleration. As Fig. 3c shows, owing to the immense injection efficiency enabled by the thin-body channel, there are enough stored electrons to generate a large non-volatile memory window. The inset shows the 400-ps V_{PROG} waveform captured by the oscilloscope. The test was repeated three times and shows a stable 400-ps response. The relationship between the memory window and the program speed is illustrated in Fig. 3d. As the pulse width is reduced from 1 ns to 400 ps, the memory window undergoes a corresponding decrease, ranging from 1.8 V to 0.78 V. The hot holes can also support the sub-1-ns program and the detailed sub-1-ns performance of batches of graphene devices is provided in Supplementary Information section 6. Using both 2D-enhanced hot-electron and hot-hole injection, graphene flash can realize bidirectional V_{th} shift with sub-1-ns speed. In addition, the minimum program time is discussed in Supplementary Information section 7. We have also demonstrated that based on the 2D-enhanced hot- hole injection mechanism, WSe_2 flash memory shows a program speed up to 1 ns with an ON/OFF ratio of about 10^3 (Supplementary Fig. 27).

Figure 3e confirms the non-volatile data retention capacity of the flash device. The stability of both states was evaluated at room temperature. Transfer curves were measured at different time intervals and the V_{th} retention after electron and hole trapping was extracted to demonstrate that the device remains stable even after 60,000 s. Linear extrapolation of V_{th} indicates that the flash device still has a large memory window after 10 years at room temperature. The robust endurance of the flash memory is shown in Fig. 3f. Following a series of programming cycles, the device can repeatedly switch between two states and work well within 5.5 × 10^6 cycles. This robust endurance characteristic benefits from low program voltage and short accumulated stress time, demonstrating the advancement of the 2D-HCI mechanism. The endurance tests of more devices are provided in Supplementary Fig. 32. We also provide more experimental data to support the 2D-HCI mechanism (Supplementary Information section 9). In the future, the use of high-quality chemical-vapour-deposition materials and large-scale integration process will further improve the uniformity of our devices and pave the way to practical applications.

The benchmark of sub-1-ns flash memory

Figure 4 presents a comparison of the program speed and voltage of charge-based memory. We can divide it into five areas: silicon flash based on FN tunnelling[6,37], 2D flash based on FN tunnelling[7–9,21,22,38], silicon flash based on hot-electron injection[23,39–42], high-speed volatile memory[43–48], and our sub-1-ns 2D flash based on the 2D-HCI mechanism. In terms of traditional FN tunnelling flash, the program speed is approximately 100 µs. Recently, research into 2D ultrafast flash has yielded a significant increase in speed (10–20 ns), but still requires relatively high voltages (15–30 V). Another means for improving injection efficiency is the hot-carrier program in silicon flash memory. The program voltage is strongly influenced by L_{ch}: the smaller L_{ch}, the less $V_{D,PROG}$ required. Nevertheless, the speed of these developments is still far behind high-speed volatile memory, such as SRAM and dynamic random-access memory (DRAM). The speed of transistor-based volatile memory is also size dependent, for example, SRAM can achieve a working speed of 1 ns at the 0.18-µm node. The utilization of the 2D-HCI mechanism enables the programming of our non-volatile flash devices at a speed below 1 ns. The graphene flash (L_{ch} = 0.2 µm) operated at $V_{D,PROG}$ = −5 V, 400 ps, and the WSe_2 flash (L_{ch} = 0.8 µm) worked at $V_{G,PROG}$ = $V_{D,PROG}$ = −5.7 V, 1 ns. A comparison of the sub-1-ns flash (0.2–0.8 µm) with SRAM (L_{ch} = 0.18–1 µm) and DRAM (L_{ch} = 0.13–1 µm) reveals that the non-volatile flash memory has outperformed the volatile memory in terms of speed. It is noteworthy that this sub-1-ns speed breakthrough matches with the enhanced injection current (Fig. 2d) in the order of magnitude. It can be reasonably inferred that sub-1-ns 2D flash memory will continue to demonstrate superior performance with the scaling down of devices. A comparison between the emerging sub-1-ns non-volatile memory and sub-1-ns 2D flash memory is provided in Supplementary Table 4.

Conclusion

On the basis of the atomic thickness of 2D materials, we found a channel-thickness-modulated E_y distribution effect, which is used to improve the carrier acceleration efficiency and realize a 2D-HCI mechanism. We further built 2D graphene flash devices and verified that the 2D-HCI mechanism could lead to 400-ps program speed, which breaks the sub-1-ns program speed bottleneck of non-volatile memory. This mechanism shows robust endurance and supports both 2D Dirac material and 2D semiconductors, which indicates the reliability of the 2D-HCI mechanism. In the future, it is expected that the performance of the device will be further advanced by reducing the channel length. Our findings provide a mechanism to achieve sub-1-ns program speed in flash memory, providing a path to achieve high-speed non-volatile memory technology.

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/s41586-025-08839-w.

1. Lu, A. et al. High-speed emerging memories for AI hardware accelerators. *Nat. Rev. Electr. Eng.* **1**, 24–34 (2024).
2. Loke, D. et al. Breaking the speed limits of phase-change memory. *Science* **336**, 1566–1569 (2012).
3. Rao, F. et al. Reducing the stochasticity of crystal nucleation to enable subnanosecond memory writing. *Science* **358**, 1423–1427 (2017).
4. Yasuda, K. et al. Ultrafast high-endurance memory based on sliding ferroelectrics. *Science* **385**, 53–56 (2024).
5. Song, M. et al. High speed (1 ns) and low voltage (1.5 V) demonstration of 8Kb SOT-MRAM array. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)* 377–378 (IEEE, 2022).
6. Kouchi, T. et al. 13.5 A 128Gb 1b/cell 96-word-line-layer 3D flash memory to improve random read latency with t_{PROG} = 75 µs and t_R = 4 µs. In *2020 IEEE International Solid-State Circuits Conference (ISSCC)* 226–228 (IEEE, 2020).
7. Wu, L. et al. Atomically sharp interface enabled ultrahigh-speed non-volatile memory devices. *Nat. Nanotechnol.* **16**, 882–887 (2021).
8. Yu, J. et al. Simultaneously ultrafast and robust two-dimensional flash memory devices based on phase-engineered edge contacts. *Nat. Commun.* **14**, 5662 (2023).
9. Jiang, Y. et al. A scalable integration process for ultrafast two-dimensional flash memory. *Nat. Electron.* **7**, 868–875 (2024).
10. Feng, Y. et al. Design-technology co-optimizations (DTCO) for general-purpose computing in-memory based on 55 nm NOR flash technology. In *2021 IEEE International Electron Devices Meeting (IEDM)* 12.11.11–12.11.14 (IEEE, 2021).
11. Wong, H.-S. P. & Salahuddin, S. Memory leads the way to better computing. *Nat. Nanotechnol.* **10**, 191–194 (2015).
12. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R. & Eleftheriou, E. Memory devices and applications for in-memory computing. *Nat. Nanotechnol.* **15**, 529–544 (2020).
13. Makarov, A., Sverdlov, V. & Selberherr, S. Emerging memory technologies: trends, challenges, and modeling methods. *Microelectron. Reliab.* **52**, 628–634 (2012).
14. Burr, G. W. et al. Overview of candidate device technologies for storage-class memory. *IBM J. Res. Dev.* **52**, 449–464 (2008).
15. Ishimaru, K. Future of non-volatile memory—from storage to computing. In *2019 IEEE International Electron Devices Meeting (IEDM)* 1.3.1–1.3.6 (IEEE, 2019).
16. Zhang, Y. et al. Three-dimensional perovskite nanowire array-based ultrafast resistive RAM with ultralong data retention. *Sci. Adv.* **7**, eabg3788 (2021).
17. Ma, C. et al. Sub-nanosecond memristor based on ferroelectric tunnel junction. *Nat. Commun.* **11**, 1439 (2020).
18. Luo, Z. et al. High-precision and linear weight updates by subnanosecond pulses in ferroelectric tunnel junction for neuro-inspired computing. *Nat. Commun.* **13**, 699 (2022).
19. *International Roadmap for Devices and Systems* (IEEE, 2023); https://irds.ieee.org/editions.
20. Liu, C. et al. A semi-floating gate memory based on van der Waals heterostructures for quasi-non-volatile applications. *Nat. Nanotechnol.* **13**, 404–410 (2018).
21. Liu, L. et al. Ultrafast non-volatile flash memory based on van der Waals heterostructures. *Nat. Nanotechnol.* **16**, 874–881 (2021).
22. Huang, X. et al. An ultrafast bipolar flash memory for self-activated in-memory computing. *Nat. Nanotechnol.* **18**, 486–492 (2023).
23. Mahapatra, S., Shukuri, S. & Bude, J. CHISEL flash EEPROM. I. Performance and scaling. *IEEE Trans. Electron Devices* **49**, 1296–1301 (2002).
24. Fugallo, G. et al. Thermal conductivity of graphene and graphite: collective excitations and mean free paths. *Nano Lett.* **14**, 6109–6114 (2014).
25. Meric, I. et al. Current saturation in zero-bandgap, top-gated graphene field-effect transistors. *Nat. Nanotechnol.* **3**, 654–659 (2008).
26. Shen, P.-C. et al. Ultralow contact resistance between semimetal and monolayer semiconductors. *Nature* **593**, 211–217 (2021).
27. Cheng, Z. et al. How to report and benchmark emerging field-effect transistors. *Nat. Electron.* **5**, 416–423 (2022).
28. Guan, S.-X. et al. Monolithic 3D integration of back-end compatible 2D material FET on Si FinFET. *npj 2D Mater. Appl.* **7**, 9 (2023).
29. Takeda, E., Suzuki, N. & Hagiwara, T. Device performance degradation to hot-carrier injection at energies below the Si–SiO_2 energy barrier. In *1983 International Electron Devices Meeting* 396–399 (IEEE, 1983).
30. Einspruch, G. N. & Gildenblat G. *Advanced MOS Device Physics* (Academic, 1989).
31. Cheng, Y. & Hu, C. *MOSFET Modeling & BSIM3 User's Guide* (Springer Science & Business Media, 2007).
32. Yan, R.-H., Ourmazd, A. & Lee, K. F. Scaling the Si MOSFET: from bulk to SOI to bulk. *IEEE Trans. Electron Devices* **39**, 1704–1710 (1992).
33. Yoshimi, M. et al. Analysis of the drain breakdown mechanism in ultra-thin-film SOI MOSFETs. *IEEE Trans. Electron Devices* **37**, 2015–2021 (1990).
34. Cristoloveanu, S., Gulwadi, S., Ioannou, D., Campisi, G. & Hughes, H. Hot-electron-induced degradation of front and back channels in partially and fully depleted SIMOX MOSFETs. *IEEE Electron Device Lett.* **13**, 603–605 (1992).
35. Wann, H.-J., King, J., Chen, J., Ko, P. K. & Hu, C. Hot-carrier currents of SOI MOSFETs. In *Proc. 1993 IEEE International SOI Conference* 118–119 (IEEE, 1993).
36. Zeng, S., Liu, C. & Zhou, P. Transistor engineering based on 2D materials in the post-silicon era. *Nat. Rev. Electr. Eng.* **1**, 335–348 (2024).
37. 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).
38. Zha, J. et al. A 2D heterostructure-based multifunctional floating gate memory device for multimodal reservoir computing. *Adv. Mat.* **36**, 2308502 (2024).
39. Chan, T., Young, K. & Hu, C. A true single-transistor oxide-nitride-oxide EEPROM device. *IEEE Electron Device Lett.* **8**, 93–95 (1987).
40. Chung, S. S. et al. N-channel versus p-channel flash EEPROM-which one has better reliabilities. In *2001 IEEE International Reliability Physics Symposium Proceedings. 39th Annual (Cat. No. 00CH37167)* 67–72 (IEEE, 2001).
41. Wu, J.-Y. et al. A novel P-poly gate PNOS device featuring high 2nd-bit operation window for multi-bit/cell flash memory applications. In *2007 International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA)* 1–2 (IEEE, 2007).
42. Choi, S.-J. et al. High injection efficiency and low-voltage programming in a dopant-segregated Schottky barrier (DSSB) FinFET SONOS for NOR-type flash memory. *IEEE Electron Device Lett.* **30**, 265–268 (2009).
43. Intel product specifications. *Intel* https://ark.intel.com/content/www/us/en/ark.html#@Processors (2024).
44. Sawada, K. et al. A 5 ns 369 kb port-configurable embedded SRAM with 0.5 µm CMOS gate array. In *1990 37th IEEE International Conference on Solid-State Circuits* 226–227 (IEEE, 1990).
45. Lu, N.-C. et al. A 22-ns 1-Mbit CMOS high-speed DRAM with address multiplexing. *IEEE J. Solid State Circuits* **24**, 1198–1205 (1989).

46. Kirihata, T. et al. A 14-ns 14-Mb CMOS DRAM with 300-mW active power. *IEEE J. Solid State Circuits* **27**, 1222–1228 (1992).
47. Kimuta, T. et al. 64 Mb 6.8 ns random ROW access DRAM macro for ASICs. In *1999 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC First Edition (Cat. No. 99CH36278)* 416–417 (IEEE, 1999).
48. Noh, K. et al. A 130 nm 1.1 V 143 MHz SRAM-like embedded DRAM COMPILER with dual asymmetric bit line sensing scheme and quiet unselected IO scheme. In *2004 Symposium on VLSI Circuits. Digest of Technical Papers (IEEE Cat. No. 04CH37525)* 190–191 (IEEE, 2004).
49. Takeda, E. & Suzuki, N. An empirical model for device degradation due to hot-carrier injection. *IEEE Electron Device Lett.* **4**, 111–113 (1983).

Methods

Device fabrication

The bottom control gates (5/15 nm Cr/Au) were patterned by e-beam lithography with a bilayer photoresist process followed by deposition of metal by e-beam evaporation on silicon dioxide (300 nm)/silicon substrates. For the flash memory, following the preparation of the bottom gates, an Al_2O_3 blocking oxide and an HfO_2 trap layer were grown by thermal atomic-layer deposition at 250 °C. During the atomic-layer-deposition process, trimethylaluminium and tetrakis(ethylmethylamino)hafnium reacted with water to form Al_2O_3 and HfO_2, respectively. This atomic-layer-deposition process is necessary only for flash fabrication. WSe_2, graphene and hBN bulk crystals were purchased from HQ Graphene and the heterostructure of the 2D materials was prepared using mechanical exfoliation and a dry-transfer approach. The hBN flake was first transferred to the bottom gate, and then the thin-body material was transferred onto the hBN flake. The adhesion of the heterostructure on the substrate was improved by heat annealing for more than 2 h at 200 °C under a nitrogen atmosphere. Next, drain–source contacts were patterned by e-beam lithography, and a metal stack (WSe_2, 10/12 nm Sb/Pt; graphene, 5/60 nm Cr/Au stack) was deposited using e-beam evaporation. After the standard lift-off approach, devices with the Sb/Pt contact were annealed by the hot plate in the glove box (200 °C, 2 h) to form the alloy. To conduct the sub-1-ns flash memory test, metal wires and pads matched with the radio-frequency probes were patterned by e-beam lithography, and 5/60 nm Cr/Au was deposited by e-beam evaporation.

Material characterization

Atomic force microscopy for the devices was measured by a MFP-3D Origin+ (Asylum Research, Oxford Instruments) system. The transmission-electron-microscopy-ready sample was prepared using an in situ focused ion-beam lift-out technique on a Thermo Scientific Helios Eurofins EAG lab G4 HX or UC Dual Beam focused ion-beam/scanning electron microscope. The sample was plated with iridium and capped with electron-beam Pt and ion-beam Pt before milling. The transmission electron microscopy image was captured with a Thermo Scientific Tecnai F20 transmission electron microscope operated at an accelerating voltage of 200 kV. Energy-dispersive spectroscopy was performed on the Super X FEI System under scanning transmission electron microscopy mode.

Electrical measurements

In this work, the devices were measured at room temperature and under atmospheric conditions in a probe station (MPI, TS200-SE). The 2D-HCI mechanism verification under various temperatures was performed using a Lake Shore vacuum low-temperature probe station under a vacuum of $<10^{-4}$ mbar. The d.c. signals were generated using a source measure unit (B1500A, Keysight). The voltage pulses (≥20 ns) were generated using a semiconductor pulse generator unit (B1500A, Keysight). The sub-1-ns measurement was conducted based on our homemade high-speed system (Supplementary Fig. 19). The waveforms of the voltage pulse were captured by the oscilloscope (DPO 5204, Tektronix). The electrical test was conducted using a semiconductor device parameter analyser (B1500A, Keysight).

Theoretical derivation of quasi-2D model

We employed a quasi-2D model with the same structure as our fabricated 2D transistor for analysis. The x direction is defined as the channel surface to gate vertically, and the y direction is defined as source to drain horizontally, where the zero point of the x and y axes is the channel surface and source, respectively. y_0 is at the leading edge of the velocity saturation region, $V(y)$ is the horizontal potential in the channel, t_{ch} is the channel thickness, E_x and E_y are the vertical and horizontal electric field in the channel, $E_{dielectric}$ is the vertical electric field in the dielectric, and E_{sat} and V_{DSsat} are the E_y and $V(y)$ at the leading edge of the velocity saturation region, respectively. To obtain the E_y distribution, it is necessary to solve Poisson's equation and the current transport equation simultaneously in the velocity saturation region. It is expressed as

$$\frac{\partial^2 V(x,y)}{\partial x^2}+\frac{\partial^2 V(x,y)}{\partial y^2}=\frac{\rho(x,y)}{\varepsilon_0\varepsilon_{ch}} \tag{4}$$

where $\rho(x,y)$ is the charge enclosed in the Gaussian box, ε_0 is the vacuum permittivity and ε_{ch} is the channel relative permittivity. Under the conditions of (1) the thin body ($t_{ch}\to 0$) of the 2D material channel results in a square pinch-off region, denoted as a Gaussian box, (2) E_y only varies with horizontal position, and (3) the value of E_x at $x=0$ is close to zero, which allows us to replace $\frac{\partial E_x}{\partial x}$ with the average value given by $\frac{E_x(t_{ch},y)}{t_{ch}}$, we apply Gauss' law

$$\begin{aligned}&-E_{sat}t_{ch}+E_y(y)t_{ch}+\frac{\varepsilon_{dielectric}}{\varepsilon_{ch}}\int_{y_0}^{y}E_{dielectric}(t_{ch},k)\mathrm{d}k\\&=\frac{\rho(x,y)}{\varepsilon_0\varepsilon_{ch}}t_{ch}(y-y_0)\end{aligned} \tag{5}$$

where $E_{dielectric}$ is the electric field in the gate dielectric and $\varepsilon_{dielectric}$ is the dielectric relative permittivity. Differentiating equation (5) with respect to y, we have

$$t_{ch}\frac{\mathrm{d}E_y(y)}{\mathrm{d}y}+\frac{\varepsilon_{dielectric}}{\varepsilon_{ch}}E_{dielectric}(t_{ch},y)=\frac{\rho(x,y)}{\varepsilon_0\varepsilon_{ch}}t_{ch} \tag{6}$$

We can express $E_{dielectric}(t_{ch},y)$ as

$$E_{dielectric}(t_{ch},y)=\frac{[V_{GS}-V_{FB}-2\varphi_B-V(y)]}{t_{dielectric}} \tag{7}$$

where V_{GS} is the gate–source bias, V_{FB} is the flat-band voltage, φ_B is the surface potential and $t_{dielectric}$ is the dielectric thickness. As the boundary conditions at the leading edge of the Gaussian box: $V(y=y_0)=V_{DSsat}$ and $E_{dielectric}(y=y_0)=[V_{GS}-V_{FB}-2\varphi_B-V_{DSsat}]/t_{dielectric}$, we can obtain that $\rho(x,y)=\frac{\varepsilon_0\varepsilon_{dielectric}}{t_{ch}t_{dielectric}}(V_{GS}-V_{FB}-2\varphi_B-V_{DSsat})$. Then equation (6) is simplified to

$$\frac{\mathrm{d}E_y(y)}{\mathrm{d}y}=\frac{V(y)-V_{DSsat}}{\lambda^2},\ \text{where}\ \lambda^2=\frac{\varepsilon_{ch}t_{dielectric}t_{ch}}{\varepsilon_{dielectric}} \tag{8}$$

where λ is the effective length of the Gaussian box, also known as the scale length. By applying the boundary conditions at the leading edge of the Gaussian box: $V(y=y_0)=V_{DSsat}$, and $E_y(y=y_0)=E_{sat}$, we can write the $V(y)$ and $E_y(y)$ in the Gaussian box as

$$V(y)=V_{DSsat}+\lambda E_{sat}\sinh\left(\frac{y-y_0}{\lambda}\right) \tag{9}$$

and

$$E_y(y)=E_{sat}\cosh\left(\frac{y-y_0}{\lambda}\right) \tag{10}$$

At the drain end of the channel where the field peaks

$$E_{y,max}=E_y(y=y_0+\Delta L)=E_{sat}\cosh\left(\frac{\Delta L}{\lambda}\right) \tag{11}$$

and

$$V_{DSmax} = V_{DSsat} + \lambda E_{sat} \sinh\left(\frac{\Delta L}{\lambda}\right) \quad (12)$$

where ΔL is the region that extends from the channel pinch-off point to the drain. Equations (11) and (12) can be combined to yield

$$\Delta L = \lambda \ln \frac{\left\{\frac{[V_{DS} - V_{DSsat}]}{\lambda}\right\} + E_{y,max}}{E_{sat}} \quad (13)$$

and

$$|E_{y,max}| = \sqrt{\frac{(V_{DS} - V_{DSsat})^2}{t_{ch}} \frac{\varepsilon_{dielectric}}{\varepsilon_{ch} t_{dielectric}} + E_{sat}^2} \quad (14)$$

This equation shows that the $E_{y,max}$ is highly sensitive to t_{ch} scaling.

Data availability

Source data are provided with this paper. Other data that support the plots in this paper and other findings of this study are available from the corresponding authors upon request.

Code availability

The codes used for the simulation are available from the corresponding authors upon request.

Acknowledgements This work was supported by the National Key Research and Development Program of China (2024YFA1208400 and 2021YFA1200500), the National Natural Science Foundation of China (61925402, 62322405, 62374042 and 62090032), the Shanghai Pilot Program for Basic Research-Fudan University 21TQ1400100 (21TQ011), the Shanghai Rising-Star Program (22QA1400700), the New Cornerstone Science Foundation through the XPLORER PRIZE (62488101), and the young scientist project of the MOE innovation platform.

Author contributions C.L. and P.Z. conceived of the idea. C.L., Y.X. and C.W. designed and conducted the experiment. C.L. constructed the device mechanism theory. Y.X. and C.W. provided valued discussion and technology computer-aided design simulation support. T.W., Y.J., Y.W. and S.W. provided experimental assistant and paper discussion. C.L. and Y.X. co-wrote the paper. C.W. and P.Z. made valued input on the discussion and revision of the paper.

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/s41586-025-08839-w.

Correspondence and requests for materials should be addressed to Chunsen Liu or Peng Zhou.

Peer review information *Nature* thanks Jonghwa Eom and Yann-Wen Lan for their contribution to the peer review of this work.