nature communications

Article

https://doi.org/10.1038/s41467-026-71906-x

CMOS compatible probabilistic computing hardware with cointegrated reconfigurable p-bits and synapse arrays

Received: 29 August 2025

Accepted: 31 March 2026

Published online: 11 April 2026

Check for updates

Jun-Young Park[1], Jae-Hyun Lee[1], Jeong-Min Lee[1], Ki-Nam Park[2], Sungho Kim[3] & Joon-Kyu Han[1]

Probabilistic computing (p-computing) has emerged as a promising approach to solve combinatorial optimization problems (COPs) at room temperature, offering a practical alternative to quantum computing. Scalable p-computing hardware requires integration of probabilistic bits (p-bits) and synapses, two components with inherently contrasting device requirements. Here, we present a charge-trap field-effect transistor (CT-FET) with a silicon-oxide-nitride-silicon gate stack that functions reconfigurably as both a p-bit and a synapse, derived from a commercialized flash memory structure. Gate voltage control of trap energy enables stochastic shallow trap operation for p-bits and stable deep trap operation for synapses. P-bits and synapse array are cointegrated on a single wafer using standard CMOS processes, enabling scalable p-computing hardware. Experimental validation shows successful solution of a 4-node weighted max-cut problem including negative weights, while simulations confirm efficient solution of a 20-node weighted max-cut problem including negative weights via chaotic simulated annealing. These results demonstrate a CMOS compatible pathway toward practical and scalable probabilistic processors.

A combinatorial optimization problem (COP) is characterized as the task of determining the optimal solution to a given objective function over an exponentially large search space[1,2]. Many critical challenges in science and industry, such as very-large-scale integration chip design[3,4], logistics optimization[5,6], protein structure prediction[7,8], and portfolio management[9,10], can be formulated as COPs. The scale and complexity of COPs have grown explosively with the expansion of big data and artificial intelligence (AI)[11,12]. However, conventional deterministic computing, which sequentially evaluates possible solutions, faces limitations in finding the optimal solution within a feasible time. In contrast, quantum computing, using the superposition of qubits to explore all possibilities in parallel, has emerged as a promising approach for solving COPs. However, maintaining the superposition and entanglement essential for quantum computing requires a cryogenic environment, demanding bulky and energy-intensive refrigeration. In addition, the wiring bottleneck between room-temperature CMOS circuits and cryogenic qubits hinders large-scale implementation[13,14]. These challenges have highlighted probabilistic computing (p-computing), which operates at room temperature, as a practical alternative to quantum computing.

In p-computing hardware, the probabilistic bit (p-bit) and synapse constitute the two critical components. A p-bit randomly outputs 0 or 1 with a tunable probability governed by the intrinsic randomness of the device, while a synapse implements the connection strength between two p-bits. Figure 1a illustrates the overall framework for solving COPs using p-computing hardware. Similar to quantum computing, p-computing maps the target COP into the energy (E) of an

[1]Department of Materials Science and Engineering and Inter-university Semiconductor Research Center, College of Engineering, Seoul National University, Seoul, Republic of Korea. [2]Department of Electronic Engineering, College of Engineering, Hanyang University, Seoul, Republic of Korea. [3]Division of Electronic and Semiconductor Engineering, College of Engineering, Ewha Womans University, Seoul, Republic of Korea. e-mail: joonkyuhan@snu.ac.kr

Ising model, as expressed in Eq. (1)[15,16].

$$E = -\left(\sum_{i,j} J_{ij} x_i x_j + \sum_i h_i x_i\right) \quad (1)$$

x_i and x_j denote variables encoded as p-bits. J_{ij} represents the coupling strength between variables, while h_i denotes the external field applied to each variable, both implemented via synapses. A configuration of p-bits that minimizes E is the optimal solution to the COP. As shown in Eq. (2), the input function (I_i), defined as the negative partial derivative of E with respect to each variable, is applied as the controllable probability input to each p-bit.

$$I_i = -\frac{\partial E}{\partial x_i} = \sum_j J_{ij} x_j + h_i \quad (2)$$

P-computing hardware operates by repeating probabilistic sampling of p-bits and calculation of I_i in synapses, enabling effective exploration of the global minimum of E while avoiding local minima. This requires high stochasticity in p-bits and high stability in synapses, which are two inherently contrasting device characteristics.

To date, research on p-computing hardware has focused on implementing individual p-bits. Early work reported p-bits based on magnetic tunnel junction[17], followed by diffusive memristors[18], oscillators[19], 2D materials[20], and floating body metal-oxide-semiconductor field-effect transistor (MOSFET)[21] with inherent stochastic properties. However, synapses have typically been implemented on separate hardware, such as a field-programmable gate array or a host PC, and connected to p-bits via interface circuits such as a digital-to-analog converter and an analog-to-digital converter[17,22–24]. Implementing p-bits and synapses on separate chips causes delays and extra power use from data transfer, which limits large-scale integration and expansion. Furthermore, many prior studies have fabricated p-bits using materials with low CMOS compatibility, making cointegration with CMOS circuits challenging[17–20]. This necessitates chip-to-chip bonding or complex interconnect structures, adding manufacturing cost[25–29]. While some recent work has explored devices capable of supporting both synaptic and stochastic behavior within a single element, these demonstrations were limited to achieve CMOS-level cointegration[30].

In this work, we present a charge-trap field-effect transistor (CT-FET) that can operate as both a p-bit and a synapse, which are cointegrated on a single wafer. The device adopts a silicon-oxide-nitride-silicon (SONS) gate structure consisting of a polycrystalline Si gate, blocking SiO_2, charge-trap Si_3N_4, and single-crystalline Si channel, which is derived from the widely commercialized silicon-oxide-nitride-oxide-silicon (SONOS) flash memory, with the tunneling oxide layer removed. By controlling trap energy, the stochasticity and retention of charges trapped in Si_3N_4 can be tuned, thereby making the device reconfigurable. Since CT-FETs can be fabricated based on standard CMOS processes, they enable direct cointegration with CMOS circuits, including memory, counters, and transimpedance amplifiers (TIAs). This enables implementation of a p-computing system on a single chip, improving integration density, reducing chip cost, and simplifying fabrication[31–34]. Trap energy control is achieved by adjusting the gate voltage (V_G) pulse applied to the CT-FET: strong negative V_G engages shallow traps for high-stochasticity p-bit operation, while weak negative V_G engages deep traps for stable synapse operation. In a synapse array capable of storing both positive and negative weights, the output current from the synapse array (representing I_i) is converted into a voltage using a TIA and voltage scaler, and then applied to the p-bit. This demonstrates the seamless interaction between p-bits and synapse array cointegrated on the same wafer, enabling experimental solution of a 4-node weighted max-cut problem including negative weights. Moreover, chaotic simulated annealing (CSA) is employed to solve more complex problems, as the shallow traps of the synapse array can realize transiently decaying weights. Simulations based on measured data demonstrate that the proposed p-computing hardware can successfully solve the 20-node weighted max-cut problem including negative weights, when CSA is applied.

Results

Characteristics of charge-trap field-effect transistor

Figure 1b shows an optical microscope image of the cointegrated p-bits and synapse array on an 8-inch wafer. The fabrication process based on standard CMOS technology is described in Supplementary Fig. 1. As shown in Fig. 1c, both the p-bits and synapse array were implemented using CT-FETs with a channel width (W_{ch}) of 200 nm and a gate length (L_G) of 400 nm, featuring a SONS gate structure. By applying different V_Gs, both p-bit and synapse operations can be realized, as illustrated in Fig. 1d. For p-bit operation, a strong positive V_G followed by a strong negative V_G is applied to randomly modulate charges in shallow traps, thereby inducing a stochastic drain current (I_D). For synapse operation, a weak negative V_G is applied to reliably control charges in deep traps with long retention, producing a deterministic I_D.

Figure 2a–c show 100 I_D measurements when a sequence of positive and negative V_G pulses with equal pulse width (t_{pulse}) was applied, where a drain voltage (V_D) of 1 V was applied after V_G pulses to read I_D for 1 s. The positive V_G pulse was applied to induce hole detrapping for initialization, whereas the negative V_G pulse was applied to induce hole trapping in the charge-trap Si_3N_4 and increase I_D. As shown in the insets, Fig. 2a–c used (positive V_G, negative V_G, t_{pulse}) of (9 V, −8 V, 100 μs), (4 V, −4 V, 1 ms), and (2 V, −2 V, 100 ms), respectively. The values of V_G and t_{pulse} were chosen such that I_D exhibited no overall increasing or decreasing trend with cycle number. Figure 2d shows that I_D measured 1 s after V_G pulse application follows a Gaussian distribution, and that the standard deviation (σ) increases with larger $|V_G|$, while the mean (μ) shows no clear trend. The measurements in Fig. 2a–c were repeated 10^4 times, as shown in Supplementary Fig. 2a, d, g. σ decreased from 515 nA at $V_G = -8$ V (Supplementary Fig. 2b) to 274 nA at −4 V (Supplementary Fig. 2e) and 131 nA at −2 V (Supplementary Fig. 2h). Figure 2e plots σ/μ for each V_G pulse condition, along with the decay in the average I_D measured over 100 cycles in 1 s. The σ/μ was 2.7% for $V_G = -8$ V, but decreased to 1.4% and 0.7% for $V_G = -4$ V and −2 V, respectively. The I_D decay was 2.25 μA for $V_G = -8$ V, but dropped to below 0.25 μA for $V_G = -4$ V and −2 V. These results indicate that increasing $|V_G|$ enhances the cycle-to-cycle variation of I_D and accelerates the I_D decay.

Figure 2f, g illustrate how V_G modulates the trap energy in the CT-FET through band diagrams. Under a negative V_G pulse, the accumulated holes in the p-type Si channel are injected into the charge-trap Si_3N_4 via direct tunneling[35–37]. With a stronger negative V_G pulse (Fig. 2f), the slope of the Si_3N_4 valence band (E_{v,Si_3N_4}) increases, bringing the hole tunneling path closer to E_{v,Si_3N_4} and resulting capture in shallow traps, from which holes are more readily detrapped by thermal excitation. Conversely, a weaker negative V_G pulse (Fig. 2g) shifts the tunneling path farther from E_{v,Si_3N_4}, leading to capture in deep traps with suppressed detrap. Thus, a strong negative V_G pulse (Fig. 2a) enables access to shallow traps, resulting in large cycle-to-cycle I_D variations that are suitable for stochastic p-bit operation while inducing a time-dependent I_D decay that is later used for CSA. In contrast, a weak negative V_G pulse (Fig. 2c) predominantly engages deep traps, yielding stable I_D with long retention, which is suitable for synapse operation. Based on this mechanism, Fig. 2h presents the trap analysis using the trap spectroscopy by charge injection and sensing (TSCIS) method[38,39]. The results indicate the presence of various trap energies along the depth, enabling distinct electrical characteristics through trap energy control. In this work, a shallow trap was employed

https://doi.org/10.1038/s41467-026-71906-x

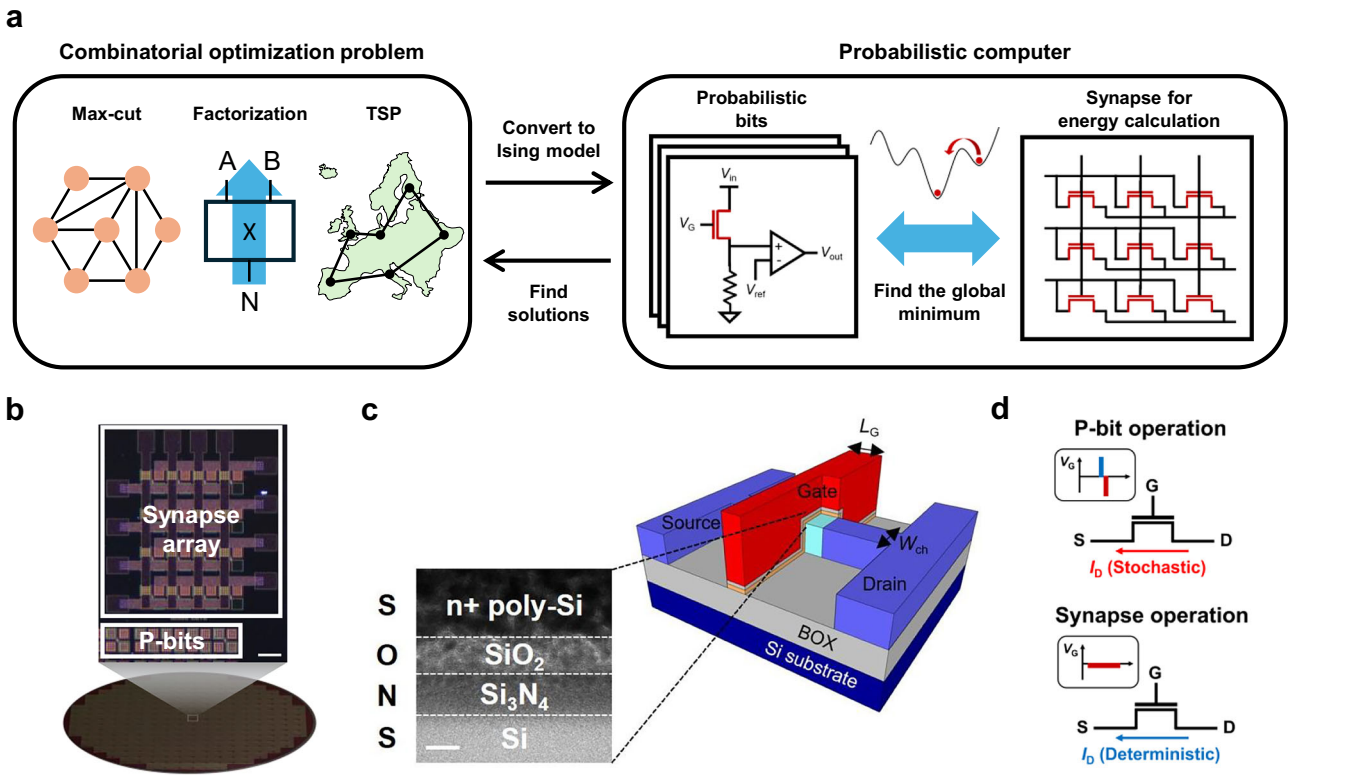


Fig. 1 | Concept of p-computing hardware with cointegrated p-bits and synapse array. a Schematic of a p-computing hardware composed of p-bits and synapses for solving combinatorial optimization problems (COPs). **b** Optical image of the fabricated 8-inch wafer with cointegrated p-bits and synapse array. (Scale bar = 200 μm) **c** Device structure of the charge-trapping field-effect transistor (CT-FET). Both p-bit and synapse operations are implemented using CT-FETs with a silicon-oxide-nitride-silicon (SONS) gate stack, as confirmed by the cross-sectional TEM image. (Scale bar = 5 nm) **d** Operation diagram of p-bit and synapse modes, where the gate voltage (V_G) and drain current (I_D) serve as inputs and outputs, respectively. Distinct V_G pulse schemes enable stochastic shallow trap operation for p-bits and stable deep trap operation for synapses.

for stochastic p-bit operation, while a deep trap was utilized for stable synapse operation. Detailed TSCIS procedures are provided in Supplementary Fig. 3, Supplementary Table 1, and Supplementary Note 1.

While a CT-FET with the SONS gate structure supports both p-bit and synapse operations, the SONOS gate structure for conventional charge trap flash memory, with its charge-trap Si_3N_4 sandwiched between two SiO_2 layers for stable charge storage, is suitable only for synapse operation and not for p-bit implementation[40–42]. Supplementary Fig. 4 presents the results of the same measurements as in Fig. 2, performed on a SONOS FET fabricated by inserting a 3 nm tunneling SiO_2 into the CT-FET. As shown in Supplementary Fig. 4e, the I_D decay remained below 30 nA regardless of V_G, and σ/μ stayed below 0.15%, both significantly smaller than those of the CT-FET. This is because, at the same V_G, the electric field in the charge-trap Si_3N_4 of the SONOS structure is weaker than that of the SONS structure, making it more difficult for holes to access shallow traps. Moreover, even if holes trapped in the charge-trap Si_3N_4 are thermally emitted, they must overcome the energy barrier of the tunneling SiO_2 to escape from the charge-trap Si_3N_4. Consequently, the CT-FET, owing to the absence of a tunneling SiO_2, can perform both p-bit and synapse operations, unlike the SONOS FET.

Probabilistic bit operation

To utilize the CT-FET as a p-bit, a digitizing circuit that converts the fluctuation of I_D into '0' or '1' is required. Figure 3a shows the p-bit circuit composed of a CT-FET, a resistor, and a comparator (HA17393, Renesas, Japan), while Fig. 3b illustrates its operating scheme. The p-bit operation consists of a probabilistic switching step, where strong positive and negative V_G pulses are alternately applied to randomly vary the hole occupancy of shallow traps, and an amplifying & digitizing step, where an input voltage (V_{in}) is applied to measure the CT-FET conductance determined by the trap states. In the probabilistic switching step, V_G = 10.7 V and −9.6 V were applied for 100 μs each, where V_G was adjusted to maximize cycle-to-cycle variation without causing a systematic increase or decrease in I_D. Although the CT-FET fabricated in this work requires relatively large $|V_G|$ to drive holes toward shallow traps during probabilistic switching, this voltage can be reduced through gate-stack engineering. As demonstrated in Supplementary Fig. 5, decreasing the blocking oxide thickness (t_{ox}) enhances the electric field at the Si/Si_3N_4 interface and enables effective shallow trap activation at substantially lower $|V_G|$. During this process, V_{in} was set to 0 V to prevent hot carrier injection (HCI) degradation. In the amplifying & digitizing step, V_{in} was applied for 100 μs, allowing the stochastic I_D to flow from the CT-FET drain through the resistor. When the resistance is R, the resulting node voltage (V_{node}) is determined by Ohm's law (I_D·R). If V_{node} exceeds the reference voltage (V_{ref}), the comparator output (V_{out}) is 5 V (state '1'); otherwise, it is 0 V (state '0'). Although a comparator was used for conceptual demonstration, the stochastic I_D of the CT-FET can also be binarized using simple domino logic, which eliminates static comparator power and enables a more energy efficient p-bit circuit. As V_{in} increases, I_D also increases, raising the probability that V_{node} surpasses V_{ref}, thereby enabling the output probability of the p-bit to be tuned by V_{in}. Figure 3c presents the probability of generating state '1' as a function of V_{in}, while Fig. 3d and Supplementary Fig. 6 show V_{out} variations under different V_{in} conditions. The measured probability curve exhibits a good fit to a sigmoid function, which arises from the shift of

https://doi.org/10.1038/s41467-026-71906-x

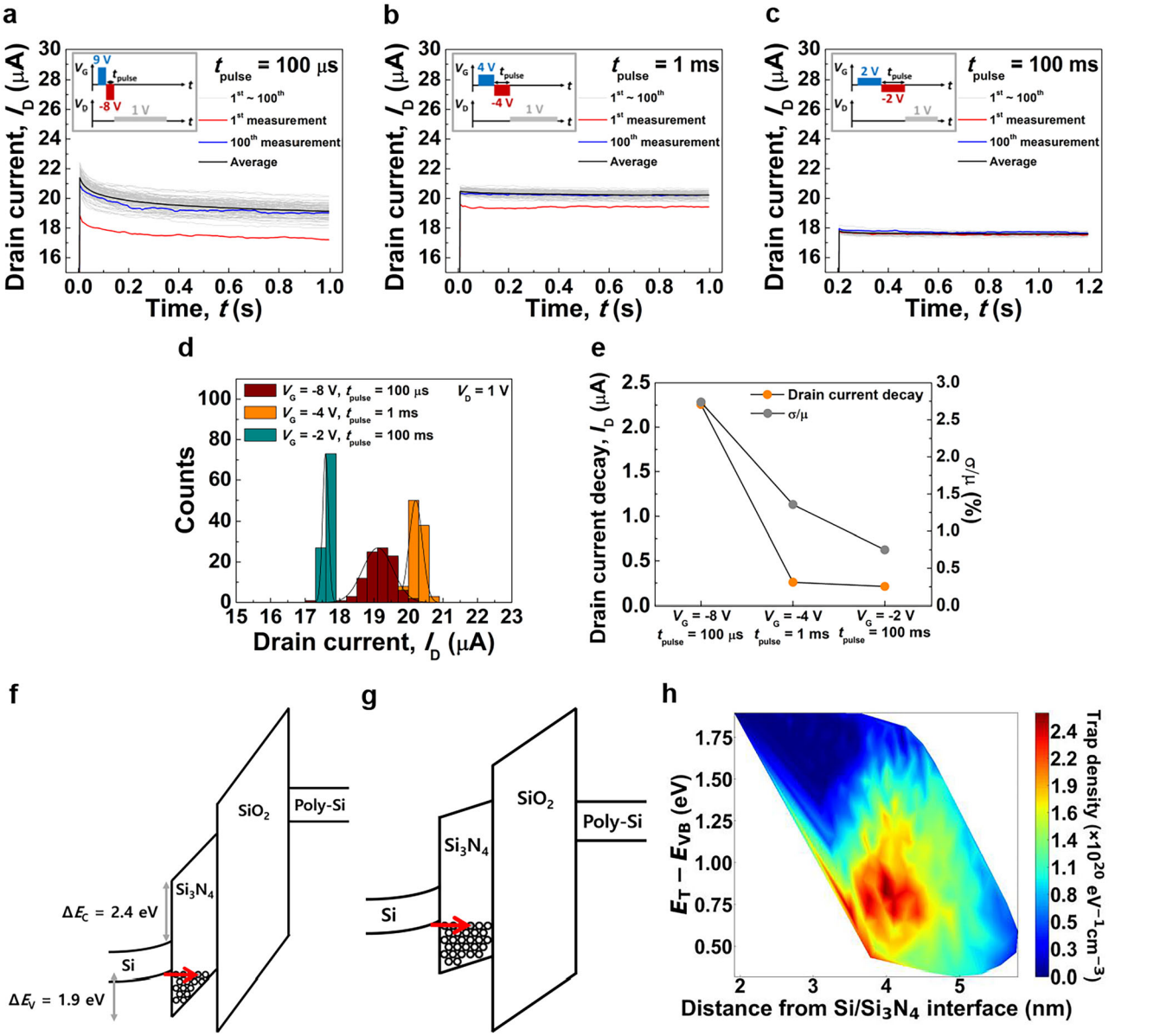


Fig. 2 | Gate voltage (V_G) tunable stochastic characteristics of the CT-FET. **a**–**c** Iteratively measured I_D after consecutive initializing V_G (>0) pulse and hole trapping V_G (<0) pulse, showing the stochasticity modulation under various V_G conditions. Insets show the applied V_G pulse schemes, with (positive V_G, negative V_G, t_{pulse}) of (9 V, −8 V, 100 μs), (4 V, −4 V, 1 ms), and (2 V, −2 V, 100 ms) for (**a**), (**b**), and (**c**), respectively. **d** Histogram of I_D measured 100 times, taken 1 s after applying V_G pulses in (**a**–**c**). **e** Decay of the average I_D measured 1 s after applying V_G in (**a**–**c**), together with the ratio of σ/μ for each Gaussian distribution in (**d**). Direct tunneling path of holes from the Si channel to the charge-trap Si_3N_4 under **f** a strong negative V_G and **g** a weak negative V_G. **h** Trap density distribution of Si_3N_4 in the CT-FET analyzed by trap spectroscopy by charge injection and sensing (TSCIS) method.

the Gaussian distribution of I_D with varying V_{in}. The measurement setup, external speed limitations, and intrinsic nanosecond response of the CT-FET p-bit are discussed in Supplementary Fig. 7, Supplementary Fig. 8, and Supplementary Note 2.

R used in the p-bit circuit was set to 47 kΩ, which was identified as the optimal value for stable and energy efficient operation. When R decreased below 47 kΩ (Supplementary Fig. 9a–c), the series resistance of the CT-FET and resistor decreased, leading to higher I_D at the same V_{in}. As a result, the voltage drop across the CT-FET increased, V_{node} decreased, and the output probability to generate state '1' decreased. Although power consumption in the resistor ($P_{resistor}$) decreased with smaller R, power consumption in the CT-FET ($P_{transistor}$) increased, resulting in a higher total average power (P_{avg}) (Supplementary Fig. 9d). Moreover, a larger V_{in} was required to maintain probabilistic operation, further increasing power consumption. In contrast, increasing R to 4.7 MΩ (Supplementary Fig. 10a, b) raised V_{node} and drastically narrowed the V_{in} range for probabilistic output from 1.2–2.5 V to 0.378–0.379 V (Supplementary Fig. 10c). This makes the p-bit more sensitive to V_{in} fluctuations and device-to-device variation, thereby undermining the reliability of p-computing. Therefore, R = 47 kΩ represents the optimized resistance, balancing power consumption and reliable p-bit operation. Under this optimized resistance condition, Supplementary Fig. 11 shows that the p-bit exhibits a total average power consumption of approximately 1.25 μW. Further analysis of $P_{transistor}$, $P_{resistor}$, and the gate-related power (P_G) reveals that P_G is significantly smaller than the other components because the CT-FET, fabricated using a standard CMOS-compatible process, exhibits negligible low gate-leakage current.

https://doi.org/10.1038/s41467-026-71906-x

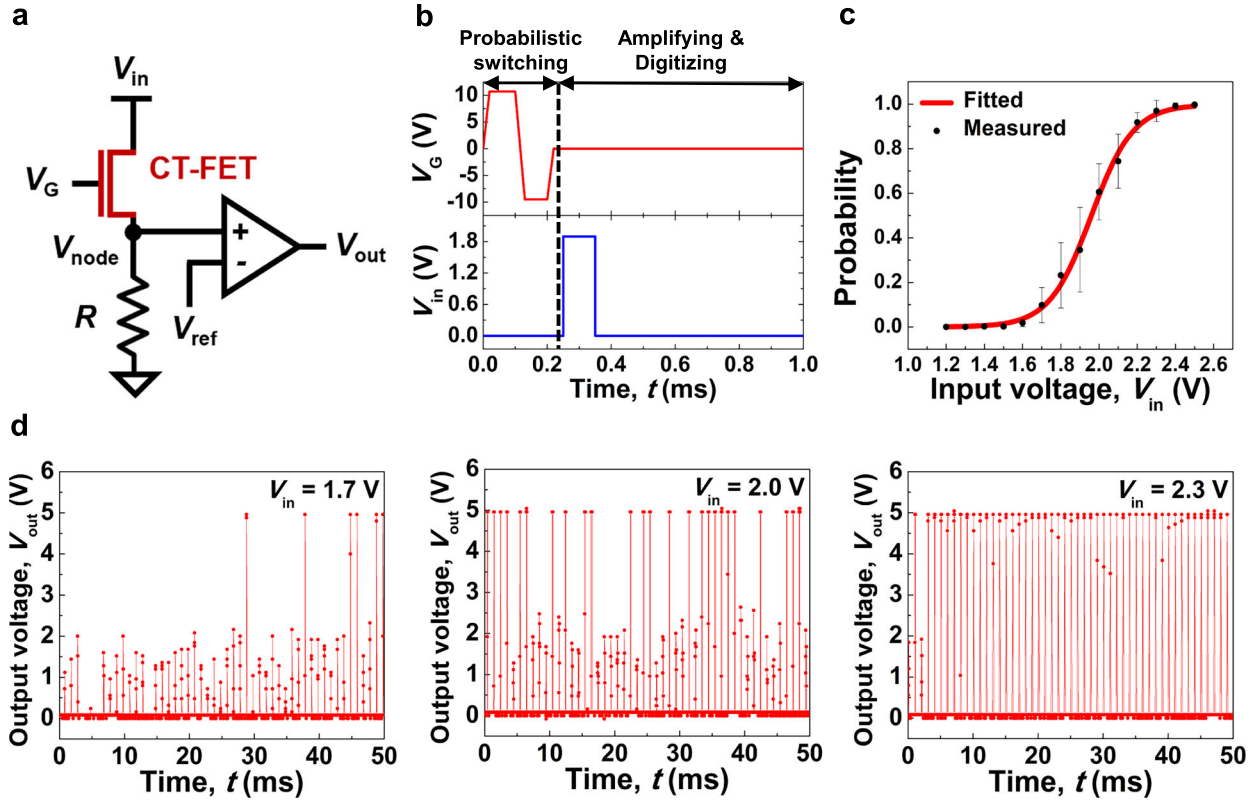


Fig. 3 | P-bit operation of the CT-FET. a P-bit circuit consisting of a CT-FET, a resistor, and a comparator. The stochastic I_D induced by shallow traps serves as the random source of the p-bit. **b** Voltage bias scheme of V_G and input voltage (V_{in}) during one cycle of p-bit operation. **c** Output probability characteristics to generate state '1' as a function of V_{in}. For each V_{in}, measurements were performed over 50 cycles × 10 repetitions, and the error bars represent the standard deviation of the 10 averaged results. **d** Stochastic V_{out} over 50 cycles of p-bit operation at V_{in} = 1.7 V, 2.0 V, and 2.3 V, showing the probabilistic switching behavior.

Supplementary Fig. 12a, b show the effect of cumulative p-bit operation on the subthreshold swing (SS) of the CT-FET and the output probability characteristics. If SS degrades, the sigmoid function can be altered by shifting the threshold voltage (V_{th}) of the FET. However, Supplementary Fig. 12a shows that SS remains nearly unchanged up to 10^7 cycles. This stability arises because V_{in} of 0 V was applied during the probabilistic switching step, thus avoiding HCI. Furthermore, unlike SONOS FETs, where Fowler Nordheim tunneling through SiO_2 is highly vulnerable to interface trap generation, the CT-FET employs charge-trap Si_3N_4 directly in contact with the Si channel, which is less susceptible to further degradation[43,44]. As a result, the output probability characteristics were well preserved after 10^7 cycles, enabling reliable p-bit operation, as shown in Supplementary Fig. 12b. Moreover, variations arising from device-to-device differences and temperature can also be effectively calibrated through V_G adjustment, as detailed in Supplementary Note 3 and Supplementary Figs. 13 and 14.

Synapse array operation

To realize a complete p-computing hardware system, a synapse that computes I_i based on the outputs of p-bits is required. Figure 4a illustrates the synapse array that computes I_i according to E corresponding to the COP, together with the auxiliary circuit that converts I_i into V_{in} (Supplementary Note 4). The 8 × 8 synapse array was implemented with 64 CT-FETs, fabricated on the same wafer and by the same process as CT-FETs used for the p-bits. In this array, CT-FETs in the same row share the drain and source, while CT-FETs in the same column share the gate. V_G of the ith column ($V_{G,i}$) represents the output of the ith p-bit and determines the on/off state of the CT-FETs in that column. The synapse output current corresponding to I_i ($I_{syn,i}$) is determined by the drain voltage of the ith row ($V_{D,i}$) and the conductance of the j^{th} column CT-FET (G_{ij}), as expressed in Eq. (3).

$$I_{syn,i} = V_{D,i} \sum_j G_{ij} \tag{3}$$

Since G_{ij} is always positive, all terms are restricted to have the same sign as $V_{D,i}$. To overcome this limitation, the sources of two rows with opposite $V_{D,i}$ polarities are connected, allowing the calculation of I_i as expressed in Eq. (4).

$$I_{syn,i} = V_{D,i+} \sum_j G_{ij+} + V_{D,i-} \sum_j G_{ij-} \tag{4}$$

For the conversion of I_{syn} into V_{in}, two OPAMPs were configured as a TIA and a voltage scaler. The detailed implementation of the auxiliary circuit and the measurement setup are described in Supplementary Note 4 and Supplementary Fig. 15.

Figure 4b illustrates the on/off operation of the CT-FET, where the programmed conductance is either reflected (on) or not reflected (off) depending on $V_{G,i}$. The CT-FET employs deep traps in the charge-trap Si_3N_4 to finely tune V_{th}, thereby enabling multi-level characteristics. As a result, the device remains in the on-state and reflects G_{ij} to I_{syn} when $V_{G,i}$ = 0 V and in the off-state when $V_{G,i}$ = −0.5 V, regardless of G_{ij}. Figure 4c shows the array voltage bias conditions required for weight (G_{ij}) programming and to calculate I_i based on the stored weights. During the read operation, $V_{G,i}$ was applied as 0 V (state "1") or −0.5 V (state "0") depending on the p-bit output. These two voltage conditions ensure stable on/off switching of the CT-FET while preventing unintended trapping or detrapping. Positive weights and negative weights

https://doi.org/10.1038/s41467-026-71906-x

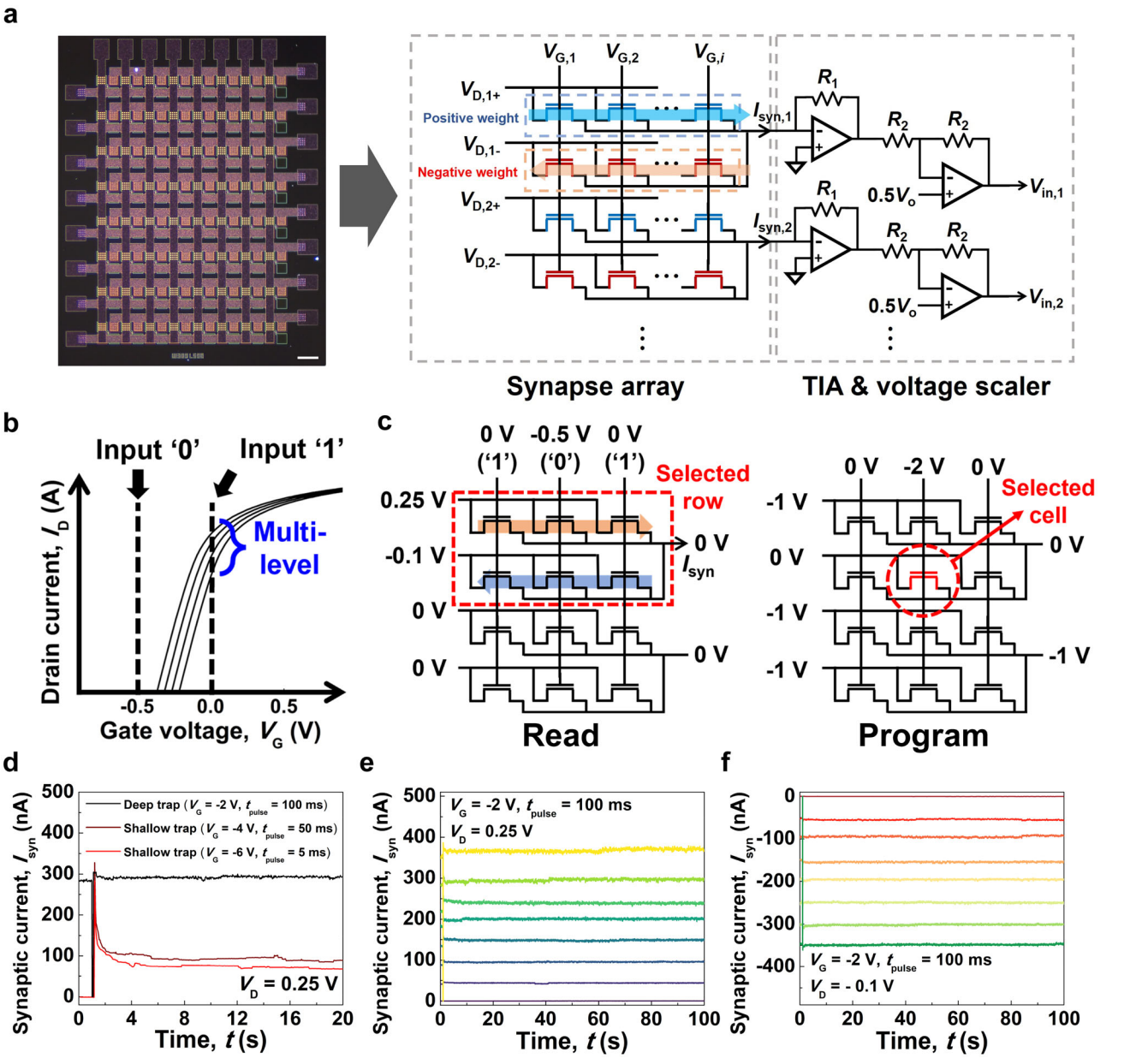


Fig. 4 | Synapse array operation of the CT-FET. a Optical image of the 8 × 8 synapse array composed of CT-FETs (Scale bar = 200 μm), along with the circuit configuration where the synapse array is connected to peripheral circuits (transimpedance amplifier (TIA) and voltage scaler) to convert synapse output current (I_{syn}) into p-bit input voltage (V_{in}). To implement both positive and negative weights, two adjacent rows share the same source line. **b** Schematic I_D-V_G curve illustrating multi-level weight implementation using deep traps in CT-FETs. **c** Voltage bias scheme for read and program operations in the synapse array. **d** I_{syn} over time during weight programming using deep trap pulses (V_G = −2 V) versus shallow trap pulses (V_G = −4, −6 V), showing that deeper traps yield more stable retention. **e**, **f** Multi-level characteristics of a positive weight (V_D = 0.25 V) and negative weight cell (V_D = −0.1 V) under repeated programming pulses.

were represented with $V_{D,i+}$ = 0.25 V and $V_{D,i-}$ = −0.1 V, respectively. In addition, 0 V was applied to the positive input of the OPAMPs connected to the sources of synapse array, thereby fixing the source voltage (V_S) of all rows at 0 V.

Programming operations require that only the selected cell stores the desired conductance. Before programming, V_G = 9 V and $V_D = V_S$ = 0 V were applied to initialize the synapse array. To achieve high-retention multi-level characteristics, cells were programmed by applying V_G = −2 V for 100 ms to the selected column, while the selected row was biased at $V_D = V_S$ = 0 V to enable hole trapping. Although the same V_G pulse is distributed along the column, unintended programming of other cells is prevented because all unselected rows are simultaneously biased at $V_D = V_S$ = −1 V, which suppresses hole capture into deep traps. This row selective drain and source biasing ensures that each synapse cell can be independently programmed even when the gate line is shared. Figure 4d presents I_{syn} over time under different V_G pulse conditions. When a deep trap programming pulse of V_G = −2 V for 100 ms was applied 24 consecutive times, I_{syn} reached 300 nA and remained stable without decay. In contrast, when a shallow trap pulse of V_G = −4 V for 50 ms was applied, I_{syn} immediately reached 300 nA after a single pulse but decayed to below 100 nA within 5 s. Under an even stronger condition (V_G = −6 V for 5 ms), I_{syn} decayed to 75 nA within 3 s, showing even more unstable retention. This instability arises because strong electric fields predominantly capture holes into shallow traps, which are easily detrapped by thermal excitation. Therefore, for long-retention synapse operation, a deep

https://doi.org/10.1038/s41467-026-71906-x

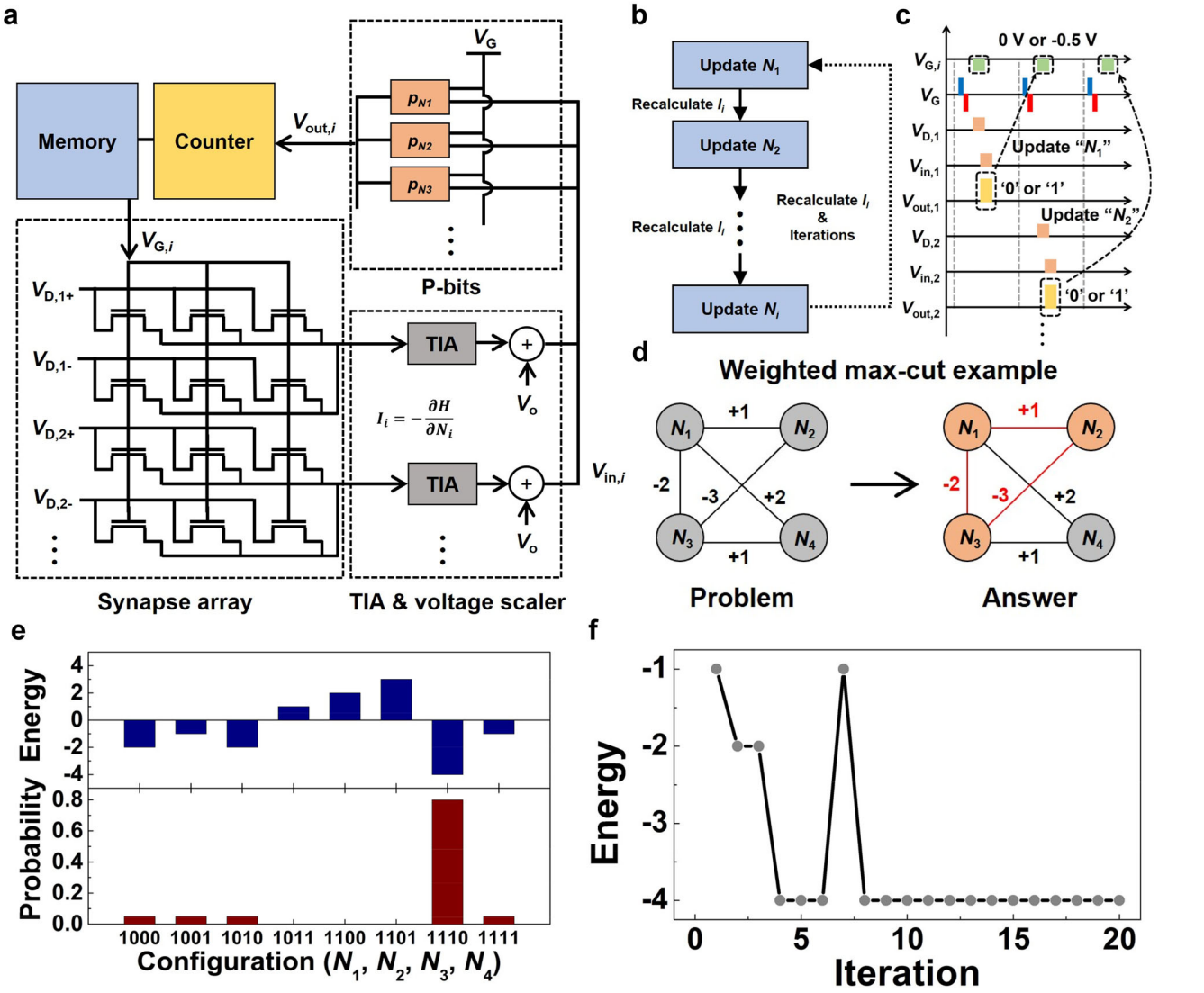


Fig. 5 | Experimental results of the weighted max-cut problem based on cointegrated p-bits and synapse array hardware. a Full p-computing hardware, consisting of p-bits and the synapse array based on CT-FETs. **b** Schematic illustration of sequential sampling. **c** Timing diagram of the proposed p-computing hardware based on sequential sampling. **d** Experimental demonstration of a 4-node weighted max-cut problem including negative weights, solved within 20 iterations. **e** Probability distribution and E of all possible configurations in the example problem, with N_1 fixed at state '1'. (f) E versus iterations, showing stable convergence to the global minimum.

trap-based programming pulse scheme (small $|V_G|$) is suitable. Although deep-trap programming under small $|V_G|$ inevitably requires latency in the order of tens of milliseconds, this step is performed only once during problem embedding and does not influence the iterative sampling for p-computing, which relies solely on fast shallow trap in the p-bits. Furthermore, this programming latency is not a fundamental limitation, as the trap density in Si_3N_4 can be increased through material-level engineering to facilitate faster hole capture[45].

Figure 4e, f present the transient response of I_{syn} as a function of the cumulative number of programming pulses applied to the CT-FET configured for positive and negative weights, respectively. As described in Supplementary Note 4, $I_{syn} = 50$ nA was defined to correspond to $I_i = 1$, and a program-and-verify scheme was employed in which programming pulses were repeatedly applied and monitored. As a result, 3-bit multi-level operation was successfully achieved under both $V_{D,i+} = 0.25$ V (positive weight) and $V_{D,i-} = -0.1$ V (negative weight) conditions. Furthermore, no significant decay of I_{syn} was observed over 100 s, confirming the stable operation of the synapse. The detailed measurement setup and power consumption analysis of array operations are provided in Supplementary Note 5 and Supplementary Fig. 16. Supplementary Fig. 17 shows Sentaurus technology computer-aided design (TCAD) simulations comparing trap energy distributions in Si_3N_4 under identical deep trap programming pulses. When traps are set near shallow levels, only a few trap sites are available along the tunneling path, leading to limited hole trapping and rapid detrapping. In contrast, with deep-level traps, holes are well retained, indicating that deep traps in the charge-trap Si_3N_4 are the primary factor enabling long-term retention during synapse programming.

Experimental result of weighted max-cut problem

Figure 5a presents the schematic of the complete p-computing hardware, consisting of p-bits and the synapse array based on CT-FETs. To enable problem solving through the interaction between p-bits and the synapse array, a TIA and a voltage scaler are needed to convert I_{syn} into V_{in}, while memory and counters store the sampled outputs of p-bits over iterations, and apply the corresponding $V_{G,i}$ to the array. Importantly, all of these components, together with CT-FETs, can be cointegrated on a single wafer using standard CMOS processes. Figure 5b

https://doi.org/10.1038/s41467-026-71906-x

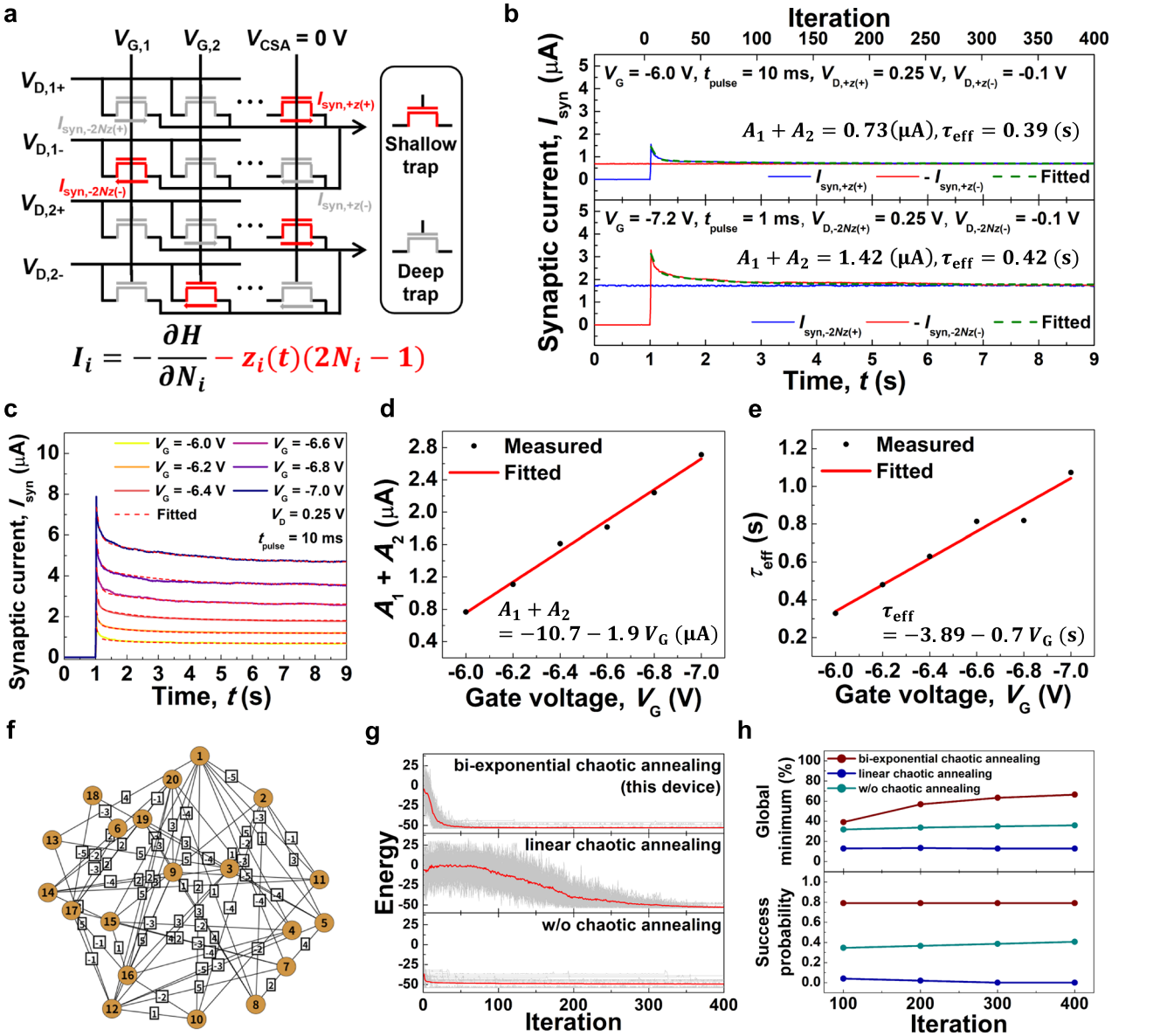


Fig. 6 | Simulation results of the weighted max-cut problem with chaotic simulated annealing (CSA). a Schematic of CSA implementation in a synapse array consisting of CT-FETs. Current decay of CT-FET by shallow traps was used to compute two self-feedback terms for the CSA. **b** Time-dependent current ($I_{syn,-2Nz(+)}$, $I_{syn,+z(-)}$, $I_{syn,-2Nz(-)}$, and $I_{syn,+z(+)}$) of programmed cell for CSA when $N_i = 1$, showing a decaying profile well fitted by a bi-exponential function. **c** I_{syn} decay with t_{pulse} fixed at 10 ms under $V_G = -6$ to -7 V. The profiles at each condition are well fitted by a bi-exponential function. **d**, **e** Extracted $A_1 + A_2$ and τ_{eff} from the bi-exponential fitting in (**c**), plotted as a function of V_G. Both parameters exhibit linear dependence on V_G. **f** Weighted max-cut problem consisting of 20 nodes and 75 edges. **g** Iteration-dependent E profiles for the problem in (**f**) under three different annealing schemes: bi-exponential chaotic annealing (this device), linear chaotic annealing, and without chaotic annealing, each averaged over 100 trials. **h** Global minimum rate and success probability versus iteration for the three annealing schemes.

illustrates sequential sampling for p-computing. In each cycle, the synapse array calculates I_i according to the current p-bit configuration and updates the ith p-bit. After all p-bits are updated once in sequence, the process is defined as one iteration, and the configuration at each iteration is stored in memory. The necessity of sequential sampling is discussed in detail in Supplementary Note 6, and the detailed procedure of p-computing, including the Python simulation code, is described in Supplementary Note 7.

Figure 5c illustrates the timing diagram for the proposed p-computing hardware based on sequential sampling. To initiate probabilistic switching, V_G was applied to the p-bit for 200 μs. During the subsequent 800 μs digitizing step, $V_{D,i+}$ and $V_{D,i-}$ of the ith p-bit to be updated, along with the $V_{G,i}$s were applied to the synapse array for 100 μs. The synapse array then output $I_{syn,i}$, which was converted into V_{in} by the TIA and voltage scaler, producing V_{out} of the ith p-bit. The counter recorded the number of completed iterations, and the memory updated the current p-bit configuration. This operation proceeded in a synchronized manner under a clock, with each interval determined by the p-bit pulse scheme shown in Fig. 3b. As discussed in Supplementary Note 2, cointegration of the pulse generator with the CT-FETs can reduce the probabilistic switching time below 200 μs. Moreover, full CMOS-based integration of all components in Fig. 5a can further minimize delays in digitizing, memory, and counter operations.

The max-cut problem is a representative non polynomial (NP)-hard COP, in which a graph is partitioned into two subsets to maximize the number of edges between them. By mapping each node to a p-bit (N_i or N_j) and expressing the connections as w_{ij}, the corresponding E

can be formulated like Eq. (5).

$$E = \sum_{i>j} w_{ij}(2N_iN_j - N_i - N_j + \frac{1}{2}) \quad (5)$$

Unlike the conventional bipolar formulation, where s_i is −1 or 1, the binary representation used in Eq. (5), where N_i is 0 or 1, introduces an additive constant from the algebraic variable transformation. It should be noted that this constant does not affect the energy ordering and preserves the equivalence of the formulation.

Extending w_{ij} beyond binary values (0 or 1) to arbitrary numbers represents the weighted max-cut problem including negative weights. This explicitly incorporates negative interactions and provides a more realistic model than the conventional max-cut problem[46–49]. In this study, we experimentally solved a 4-node weighted max-cut problem including negative weights using the proposed p-computing hardware. Figure 5d shows the example problem and its solution, while Supplementary Table 2 presents the corresponding E and I_i. Supplementary Fig. 18a illustrates that the programmed synapse array corresponds to the example problem, and Supplementary Fig. 18b–d verify accurate computation of I_i and conversion to V_{in} by measuring $I_{syn,i+}$, $I_{syn,i-}$, and V_{in} under all possible cases of N_2, N_3, N_4, while N_1 was fixed at state '1'. Figure 5e shows E of all configurations and the results from 20 iterations of sequential sampling. The correct solution, corresponding to the lowest energy state, was identified with a pass value above 0.2, which was empirically chosen as the minimum threshold that reliably captures all degenerate optimal solutions while excluding transient or noisy cases. As shown in Fig. 5f, the global minimum of E was reached within four iterations and stably maintained, with occasional fluctuations due to the stochastic nature of p-bits that allow escape from local minima. To further assess the scalability of the proposed p-computing hardware beyond the experimentally demonstrated array size, we performed large-scale max-cut simulations on G-set graphs of up to 3,000 nodes using the measured CT-FET p-bit characteristics. As summarized in Supplementary Table 3, the simulated results consistently achieved accuracies comparable to those reported in prior hardware-based Ising machines, demonstrating that the measured CT-FET characteristics are sufficient to support large-scale probabilistic computing.

Intrinsic chaotic simulated annealing in synapse array

P-computing can overcome the energy barrier near local minima due to the stochastic nature of p-bits, but as the energy landscape becomes more complex, escaping local minima becomes difficult. To address this, various annealing strategies such as simulated annealing[50,51] and parallel tempering[52,53] have been proposed to gradually decrease the system temperature and prevent being trapped in local minima. In this work, CSA was directly implemented within the synapse array by exploiting the shallow trap of CT-FETs, without requiring additional hardware. As represented in Eq. (6), CSA introduces a transient self-feedback chaos into I_i, broadening the search space initially and gradually shrinking it over time to guide the system toward the global minimum[54–56].

$$I_i = -\frac{\partial E}{\partial N_i} - z_i(t)(2N_i - 1) \quad (6)$$

Here, z_i(t) is a decaying self-feedback term that initially suppresses state fixation to broaden exploration, and gradually vanishes to guide the system toward the global minimum.

Figure 6a illustrates the implementation of CSA in the synapse array. Applying CSA introduces two additional self-feedback terms to I_i, $-2N_iz_i$(t) term programmed into the negative weight cell of ith column and $+z_i$(t) term programmed into a positive weight cell of bias column, where V_G for bias column (V_{CSA}) fixed at 0 V. Since z_i(t) must decay over time, both terms were programmed using shallow traps. Ideally, the change in I_{syn} induced by shallow traps should fully decay to zero. In practice, however, a residual component remains because some holes become trapped in deep levels. To compensate, the residual component was preprogrammed into the adjacent row with opposite $V_{D,i}$ polarities using deep trap pulses, so that the positive and negative contributions cancel out, ensuring complete decay of the self-feedback term. In Fig. 6a, $I_{syn,-2Nz(-)}$ and $I_{syn,+z(+)}$ represent I_{syn} used to compute $-2N_iz_i$(t) term and $+z_i$(t) term in Eq. (6), respectively, and were programmed using shallow traps. $I_{syn,-2Nz(+)}$ and $I_{syn,+z(-)}$ represent I_{syn} used to compensate residual components, and were preprogrammed using deep traps. Figure 6b shows $I_{syn,-2Nz(-)}$ and $I_{syn,+z(+)}$ over time, along with $I_{syn,-2Nz(+)}$ and $I_{syn,+z(-)}$ when $N_i = 1$. The transient decay of $I_{syn,-2Nz(-)}$ and $I_{syn,+z(+)}$ can be fitted with a bi-exponential function, as described in Eq. (7).

$$I_{syn} = y_0 + A_1e^{-\frac{t-t_0}{\tau_1}} + A_2e^{-\frac{t-t_0}{\tau_2}} \quad (7)$$

Let A_1 and A_2 as amplitudes and τ_1 and τ_2 as time constants of two exponential functions, $A_1 + A_2$ represents the initial value of z_i(t), while τ_{eff}, the weighted average of τ_1 and τ_2, indicates the effective decay time. For programming, $-2N_iz_i$(t) was implemented with $V_G = -6$ V for 10 ms, and $+z_i$(t) with $V_G = -7.2$ V for 1 ms, resulting in currents of the same decay profile but with a twofold magnitude difference. The fitting results confirm that $I_{syn,-2Nz(-)}$ yields $A_1 + A_2 = 0.73$ μA and $\tau_{eff} = 0.39$ s, while $I_{syn,+z(+)}$ yields $A_1 + A_2 = 1.42$ μA and $\tau_{eff} = 0.42$ s.

By tuning V_G and t_{pulse} of the shallow trap pulse, the key parameters of the current decaying profile, $A_1 + A_2$ and τ_{eff}, can be controlled. Figure 6c shows I_{syn} with different V_G at $t_{pulse} = 10$ ms, together with the corresponding bi-exponential fitting. Figure 6d, e summarize the dependence of $A_1 + A_2$ and τ_{eff} on V_G. Both parameters increased linearly with V_G, as a larger V_G increased the fraction of shallow traps, leading to a larger detrapped charge ($A_1 + A_2$) and longer detrapping time (τ_{eff}). In addition, as discussed in section "Characteristics of charge-trap field-effect transistor", increasing t_{pulse} at the same V_G results in more holes being trapped in shallow levels. Thus, by controlling V_G and t_{pulse}, the decaying profile of z_i(t) can be precisely controlled.

Based on the measurement in Fig. 6b, CSA was applied to a Python simulation to solve a 20-node weighted max-cut problem including negative weights (Fig. 6f). Supplementary Table 4 presents E and I_i of this problem. As assumed in Fig. 5c, the update time for a single p-bit is 1 ms, so one iteration of sequential sampling requires 20 ms. Thus, the 8 s current decay observed in Fig. 6b corresponds to 400 iterations of z_i(t) in the simulation. Incorporating the I_{syn} decaying profile induced by shallow traps in the CT-FET, three different annealing schemes were compared: bi-exponential chaotic annealing, linear chaotic annealing, and without chaotic annealing. In the linear case, the schedule was obtained by linearly connecting the initial and final values of the bi-exponential schedule (Supplementary Fig. 19). Figure 6g shows E profiles with iteration for the three cases. Without chaotic annealing, E did not decrease consistently and becomes trapped in different local minima. In linear chaotic annealing, E eventually converged to the global minimum but required more than 350 iterations. By contrast, bi-exponential chaotic annealing enabled convergence to the global minimum within 100 iterations, after which it was stably maintained. Figure 6h compares the global minimum rate and success probability for the three cases. The global minimum rate was defined as the fraction of iterations sampling the global minimum, while the success probability was defined as the fraction of trials in which the global minimum rate exceeds the pass value of 0.2. Linear chaotic annealing performed even worse than without chaotic annealing case, indicating that excessive self-feedback hindered global minimum search. In contrast, bi-exponential chaotic annealing consistently yielded the

https://doi.org/10.1038/s41467-026-71906-x

highest performance. Although the global minimum rate was about 40% at 100 iterations, the success probability was 0.8, showing that even with fewer iterations, the global minimum rate reliably exceeded the pass value in each trial. These results demonstrate that exploiting the inherent current decaying profile of CT-FET by shallow traps enables efficient self-feedback scheduling without additional circuitry, allowing stable convergence to the global minimum with a reduced number of iterations.

Discussion

In this work, cointegration of p-bits and synapse array was demonstrated on a single wafer by utilizing CT-FETs that can function reconfigurably as both p-bits and synapses. The trap energy in the charge-trap Si_3N_4 can be tuned by controlling the gate voltage, thereby enabling reconfigurable shallow trap operation for stochastic p-bits and deep trap operation for stable synapses. With additional CMOS peripheral circuits, both p-bit and synapse array operation were experimentally verified, and a 4-node weighted max-cut problem was successfully solved. Furthermore, by exploiting the inherent transient current decay of CT-FETs, CSA was directly applied without extra circuitry, allowing efficient solution of more complex max-cut problems. Fully CMOS compatible p-bits based on CT-FETs exhibited a wide input margin and high endurance compared to previously reported p-bit devices, as summarized in Table 1. In addition, we evaluated a CPU-based software p-bit on a 13th-Gen Intel Core i7−13700F, which required about 4.7 mJ and 188 µs per operation. This further highlights the energy-efficiency advantage and high speed potential of the proposed hardware, given the CT-FET's intrinsic response time of 50 ns (Supplementary Note 2 and Supplementary Fig. 8), despite the measurement being limited to 1 ms by the experimental setup. This work highlights the feasibility of integrating all essential components for p-computing within a single CMOS chip, offering a scalable pathway toward practical p-computing hardware.

Methods

Device fabrication

The CT-FETs were fabricated on a 8-inch p-type (100) SOI wafer. After active region patterning, a 6 nm Si_3N_4 layer serving as the charge-trapping layer was deposited by low pressure chemical vapor deposition (LPCVD), followed by the deposition of an 8 nm blocking SiO_2. Heavily doped polycrystalline Si layer for gate electrode was then deposited and patterned to define the gate region. Self-aligned ion implantation by P and rapid thermal annealing at 1000 °C were performed to form the source and drain regions. Then, a tetraethyl orthosilicate (TEOS) interlayer dielectric was deposited and contact vias were etched. Al layer was subsequently deposited by sputter to connect the sources and drains of CT-FETs. Next, TEOS intermetal dielectric was deposited and contact vias were etched. Finally, an additional Al layer was deposited by sputter to connect the gates of CT-FETs, enabling the formation of synapse arrays.

Electrical measurement

Electrical characteristics of the cointegrated p-bits and synapse array were measured using a B1500 semiconductor parameter analyzer (Keysight). Current sampling over time of the p-bits and synapse array was performed by voltage source current measurement mode of source measurement units (SMU, Keysight). Waveform generator fast measurement units (WGFMU, Keysight) was used for TSCIS analysis of CT-FET, probabilistic switching of p-bit, and weight programming of synapse array. To evaluate the intrinsic speed of the CT-FET, a pulse generator (PG, Agilent) was used to apply fast, high $|V_G|$ pulses directly to the CT-FET without connecting an OPAMP-based amplifier. V_{out}, V_{node}, and V_{ref} were measured using a TDS 2025 C oscilloscope (Tektronix).

Table 1 | Comparison of the proposed CT-FET p-bit with various p-bit devices

	Magnetic Tunnel Junction-based[17]	Diffusive memristor-based[18]	Oscillator-based[19]	2D material-based[20]	Si MOSFET-based[21]	This work
Device structure	Ru/Ta/CoFeB/MgO/CoFeB/Ta/[Co/Pt]$_D$/Co/Ru/[Co/Pt]$_B$/Co/Pt/Ta	$Cu_{0.2}Te_{0.8}/HfO_2$/Pt	TiN/NbO_x/TiN	h-BN/SnS_2 transistor	Floating body MOSFET	Charge trapping FET
Mechanism	Low-barrier nanomagnet	Volatile threshold switching	Metal-insulator transition	Charge trapping	Single transistor latch	Charge trapping
CMOS compatibility	Partial	Partial	Partial	Partial	Full	Full
Auxiliary circuit	Resistor, NMOS, comparator	Resistor, comparator	Resistor, capacitor, counter	Active switch, buffer	Resistor, comparator	Resistor, comparator
Power consumption for generating p-bit	10 µW	154 nW @ 5.4 V	0.5 mW	67 nW	110 nW @ 3.4 V	740 nW @ 1.2 V
Endurance	Not reported	10^6	Not reported	> 10^8	<10^7	>10^7
Control input & its margin	Voltage pulse level 1.4–2.4 V (1 V)	Voltage pulse level 4.8–5.6 V (0.8 V)	Voltage pulse level 1.32–1.46 V (0.14 V)	Voltage pulse width 10 µs–10 ms (9.99 ms)	Voltage pulse level 3.3–4.0 V (0.7 V)	Voltage pulse level 1.2–2.4 V (1.2 V)
Inference latency	350 ns	1 ms	288 ns	10 µs–10 ms	1 ms	1 ms
Large scale cointegration of p-bit and synapse	-	-	-	-	Partial (only p-bit on 8 in. wafer)	Yes (p-bit and synapse array on 8 in. wafer)
Reconfigurability	-	-	-	-	-	Yes
Annealing based on inherent device characteristics	-	-	-	-	-	Chaotic simulated annealing

Simulations

The device simulation was conducted using Sentaurus TCAD (Synopsys) with device parameters matched to the fabricated CT-FETs to study the effect of trap energy in the charge-trap Si_3N_4. The software simulation of a 20-node weighted max-cut problem with negative weights under CSA was implemented in Python. The p-bit outputs versus V_{in} and the transient synaptic weight decay were modeled by incorporating the experimental results from Fig. 3c and Fig. 6b, respectively. V_{in} calculation through the synapse array, TIA, and voltage scaler, as well as the memory and counter modules, was also implemented in Python.

Data availability

The data generated in this study are provided in the Source Data file. All the data supporting the findings of this study are available from the corresponding author upon request. Source data are provided with this paper.

Code availability

The Python code for the p-computing simulation used in this study is provided in the Supplementary Information. All other code is available from the corresponding authors upon request.

References

1. Crawford, B. et al. Binary secretary bird optimization algorithm for the set covering problem. *Mathematics* **13**, 2482 (2025).
2. Goodarzimehr, V., Fanaie, N. & Mirjalili, S. Computer-aided mathematical and structural optimization using an advanced stochastic algorithm. *Mech. Based Des. Struct. Mach.* **53**, 7485–7512 (2025).
3. Held, S., Korte, B., Rautenbach, D. & Vygen, J. Combinatorial optimization in VLSI design. In *Combinatorial Optimization: Methods and Applications* (ed. Chvátal, V.) 33–96 (IOS Press, Amsterdam, 2011).
4. Anand, S., Saravanasankar, S. & Subbaraj, P. Customized simulated annealing based decision algorithms for combinatorial optimization in VLSI floorplanning problem. *Comput. Optim. Appl.* **52**, 667–689 (2012).
5. Mu, N. et al. Multi-objective combinatorial optimization analysis of the recycling of retired new energy electric vehicle power batteries in a sustainable dynamic reverse logistics network. *Environ. Sci. Pollut. Res.* **30**, 47580–47601 (2023).
6. Ren, Y. et al. A review of combinatorial optimization problems in reverse logistics and remanufacturing for end-of-life products. *Mathematics* **11**, 298 (2023).
7. Wei, Y., Thompson, J. & Floudas, C. A. CONCORD: a consensus method for protein secondary structure prediction via mixed integer linear optimization. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **468**, 831–850 (2011).
8. Istrail, S. & Lam, F. Combinatorial algorithms for protein folding in lattice models: a survey of mathematical results. *Commun. Inf. Syst.* **9**, 303–346 (2009).
9. George, E. D. & Farid, S. S. Stochastic combinatorial optimization approach to biopharmaceutical portfolio management. *Ind. Eng. Chem. Res.* **47**, 8762–8774 (2008).
10. Ni, H., Xu, H., Ma, D. & Fan, J. Contextual combinatorial bandit on portfolio management. *Expert Syst. Appl.* **221**, 119677 (2023).
11. Schwab, K. *The Fourth Industrial Revolution* (World Economic Forum, 2016).
12. Routray, S. K. Cloud and edge computing for industry 4.0. *IT Prof.* **27**, 48–53 (2025).
13. Jeong, J. et al. Cryogenic III-V and Nb electronics integrated on silicon for large-scale quantum computing platforms. *Nat. Commun.* **15**, 10809 (2024).
14. Genssler, P. R. et al. Cryogenic embedded system to support quantum computing: from 5-nm FinFET to full processor. *IEEE Trans. Quantum Eng.* **4**, 1–11 (2023).
15. Lucas, A. Ising formulations of many NP problems. *Front. Phys.* **2**, 5 (2014).
16. Mohseni, N., McMahon, P. L. & Byrnes, T. Ising machines as hardware solvers of combinatorial optimization problems. *Nat. Rev. Phys.* **4**, 363–379 (2022).
17. Borders, W. A. et al. Integer factorization using stochastic magnetic tunnel junctions. *Nature* **573**, 390–393 (2019).
18. Woo, K. S. et al. Probabilistic computing using Cu0.1Te0.9/HfO2/Pt diffusive memristors. *Nat. Commun.* **13**, 5762 (2022).
19. Rhee, H. et al. Probabilistic computing with NbOx metal-insulator transition-based self-oscillatory pbit. *Nat. Commun.* **14**, 7199 (2023).
20. Han, J. et al. Highly durable and energy-efficient probabilistic bits based on h-BN/SnS2 interface for integer factorization. *InfoMat* **7**, 7 (2025).
21. Maeng, H. et al. Homeothermic p-bit computing hardware with stochastic operations beyond limit of non-stochastic materials. *Adv. Funct. Mater.* **35**, 12 (2024).
22. Pervaiz, A. Z., Datta, S. & Camsari, K. Y. Probabilistic computing with binary stochastic neurons. In *Proc. IEEE BiCMOS and Compound Semiconductor Integrated Circuits and Technology Symposium (BCICTS)* 1–6 (IEEE, 2019).
23. Kaiser, J. et al. Hardware-aware in situ learning based on stochastic magnetic tunnel junctions. *Phys. Rev. Appl.* **17**, 014016 (2022).
24. Chowdhury, S. et al. A full-stack view of probabilistic computing with p-bits: devices, architectures, and algorithms. *IEEE J. Explor. Solid-State Comput. Devices Circuits* **9**, 1–11 (2023).
25. Atabaki, A. H. et al. Integrating photonics with silicon nanoelectronics for the next generation of systems on a chip. *Nature* **556**, 349–354 (2018).
26. Qin, Y., Howlader, M. & Deen, M. Low-temperature bonding for silicon-based micro-optical systems. *Photonics* **2**, 1164–1201 (2015).
27. Huang, W.T. et al. Designing an efficient silicon photonics system: direct integrating photonic crystal surface emitting laser with inverse-designed perfect vertical grating coupler for optimal optical communication applications. *J. Light. Technol.* **43**, 3338–3347 (2025).
28. Ohashi, K. et al. On-chip optical interconnect. *Proc. IEEE* **97**, 1186–1198 (2009).
29. Thraskias, C. A. et al. Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications. *IEEE Commun. Surv. Tutor.* **20**, 2758–2783 (2018).
30. Heo, J., Kim, S., Kim, S. & Kim, M. Configurable synaptic and stochastic neuronal functions in ZnTe-based memristor for an RBM neural network. *Adv. Sci.* **11**, 2405768 (2024).
31. Mikhaylov, A. N. et al. Neuromorphic computing based on CMOS-integrated memristive arrays: current state and perspectives. *Supercomput. Front. Innov.* **10**, 2 (2023).
32. Zhu, Y. et al. CMOS-compatible neuromorphic devices for neuromorphic perception and computing: a review. *Int. J. Extrem. Manuf.* **5**, 042010 (2023).
33. Zhang, W. et al. Neuro-inspired computing chips. *Nat. Electron.* **3**, 371–382 (2020).
34. Majumdar, S. Back-End CMOS compatible and flexible ferroelectric memories for neuromorphic computing and adaptive sensing. *Adv. Intell. Syst.* **4**, 2100175 (2021).
35. Baek, S., Lee, J., Park, I., Baek, R.H. & Lee, J.S. Systematic analysis of oxide trap distribution of 4H-SiC DMOSFETs using TSCIS and its correlation with BTI and SILC behavior. *Solid-State Electron.* **140**, 18–22 (2018).
36. An, C.-H., Lee, M. S. & Kim, H. Effects of Si3N4 thickness on the electrical properties of oxide-nitride-oxide tunneling dielectrics. *J. Electrochem. Soc.* **155**, G247 (2008).

37. Krick, D. T., Lenahan, P. M. & Kanicki, J. Nature of the dominant deep trap in amorphous silicon nitride. *Phys. Rev. B* **38**, 8226–8229 (1988).
38. Degraeve, R. et al. Trap spectroscopy by charge injection and sensing (TSCIS): a quantitative electrical technique for studying defects in dielectric stacks. In *Proc. IEEE International Electron Devices Meeting* 1–4 (IEEE, 2008).
39. Cho, M. et al. A consistent model for oxide trap profiling with the Trap Spectroscopy by Charge Injection and Sensing (TSCIS) technique. *Solid-State Electron.* **54**, 1384–1391 (2010).
40. Lim, J.G. et al. High performance SONOS flash memory with in-situ silicon nanocrystals embedded in silicon nitride charge trapping layer. *Solid-State Electron.* **140**, 134–138 (2018).
41. Jeong, J.K. et al. Physical and electrical analysis of poly-Si channel effect on SONOS flash memory. *Micromachines* **12**, 1401 (2021).
42. Han, J.K. et al. Cointegration of single-transistor neurons and synapses by nanoscale CMOS fabrication for highly scalable neuromorphic hardware. *Sci. Adv.* **7**, 32 (2021).
43. Yu, J.-M., Han, J.K. & Choi, Y.K. Mimicking biological synaptic plasticity with a leaky charge-trap FinFET. *J. Mater. Chem. C* **10**, 9961–9967 (2022).
44. Dong-Il Moon et al. A novel FinFET with high-speed and prolonged retention for dynamic memory. *IEEE Electron Device Lett.* **35**, 1236–1238 (2014).
45. Kim, H., Aziz, J., Chavan, V. D. & Kim, D. Effect of mechanical stress on the traps in silicon nitride thin films. *Curr. Appl. Phys.* **57**, 59–64 (2024).
46. Goemans, M. X. & Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM* **42**, 1115–1145 (1995).
47. Poljak, S. & Rendl, F. Solving the max-cut problem using eigenvalues. *Discret. Appl. Math.* **62**, 249–278 (1995).
48. Faria, L., Klein, S., Sau, I. & Sucupira, R. Improved kernels for signed max cut parameterized above lower bound on (r,l)-graphs. *Discret. Math. Theor. Comput. Sci.* **19**, 1 (2017).
49. Etscheid, M. & Mnich, M. Linear kernels and linear-time algorithms for finding large cuts. *Algorithmica* **80**, 2574–2615 (2017).
50. Kirkpatrick, S., Gelatt, C. D. Jr. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
51. Ingber, L. Simulated annealing: practice versus theory. *Math. Comput. Model.* **18**, 29–57 (1993).
52. Hukushima, K. & Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **65**, 1604–1608 (1996).
53. Earl, D. J. & Deem, M. W. Parallel tempering: theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **7**, 3910 (2005).
54. Chen, L. & Aihara, K. Chaotic simulated annealing by a neural network model with transient chaos. *Neural Netw.* **8**, 915–930 (1995).
55. Bao, H. et al. Energy efficient memristive transiently chaotic neural network for combinatorial optimization. *IEEE Trans. Circuits Syst. I: Regul. Pap.* **71**, 3708–3716 (2024).
56. Yang, K. et al. Transiently chaotic simulated annealing based on intrinsic nonlinearity of memristors for efficient solution of optimization problems. *Sci. Adv.* **6**, 33 (2020).

Acknowledgements

This work was supported by the National Research Foundation of Korea (Grant No. RS–2024-00334953, J.-Y.P., J.-H.L., J.-M.L. and J.-K.H.). This work was also supported by Korea Institute of Science and Technology (KIST) Institutional Program (2E33581, J.-Y.P., J.-H.L., J.-M.L. and J.-K.H.). The EDA tool was supported by the IC Design Education Center (IDEC), Korea.

Author contributions

J.-Y.P. and J.-K.H. designed the study concept and approach. J.-K.H. fabricated the device. J.-Y.P. performed electrical measurements. J.-Y.P. performed simulations. J.-Y.P., J.-H.L., J.-M.L., K.-N.P. and S.K. analyzed the data. J.-Y.P. and J.-K.H. wrote the manuscript. J.-K.H. supervised this work. All authors discussed the results and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-71906-x.

Correspondence and requests for materials should be addressed to Joon-Kyu Han.

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

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

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

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

© The Author(s) 2026