Hardware Acceleration of Bayesian Network based on Two-dimensional Memtransistors

Yikai Zheng

Pennsylvania State University

Harikrishnan Ravichandran

Pennsylvania State University

Thomas Schranghamer

Pennsylvania State University

Nicholas Trainor

Pennsylvania State University

Joan Redwing

Saptarshi Das (✉ sud70@psu.edu)

Pennsylvania State University <https://orcid.org/0000-0002-0188-945X>

Article

Keywords:

Posted Date: February 1st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1196768/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on September 23rd, 2022. See the published version at <https://doi.org/10.1038/s41467-022-33053-x>.

Hardware Acceleration of Bayesian Network based on Two-dimensional Memtransistors

Yikai Zheng¹, Harikrishnan Ravichandran¹, Thomas F Schranghamer¹, Nicholas Trainor^{2,3},

Joan M Redwing^{2,3}, and Saptarshi Das^{2,3,4,5,}*

¹Engineering Science and Mechanics, Penn State University, University Park, PA 16802, USA

²Materials Science and Engineering, Penn State University, University Park, PA 16802, USA

³Materials Research Institute, Penn State University, University Park, PA 16802, USA

⁴Electrical Engineering and Computer Science, Penn State University, University Park, PA 16802, USA

Abstract: Long before Bayes proposed his landmark theorem on the probability of an event, based on prior knowledge of other events, natural intelligence has adopted Bayesian inference as a tool to ensure survival for almost all species. The fact that living beings must make critical decision for finding food, avoiding predators, and locating mates, based on information gathered by their sensory organs with limited sensitivity and under noisy surroundings, emphasizes the importance of probabilistic computing for evolutionary success. While the anatomy of neural hardware that accomplishes such task is far from being known, it is clear that stochastic computing is a fundamental aspect of natural intelligence, and Bayesian networks (BNs) are powerful mathematical constructs for the same. Interestingly, BNs also find widespread application in many real-world probabilistic problems including diagnostics, forecasting, computer vision, etc. While the concept of BN is well known, there are very limited hardware realizations of BN. CMOS [1, 2] based BNs

require massive hardware resources (thousands of transistors), whereas, memristor [3-5] and spintronics [6-8] based BNs necessitate hybrid design with CMOS peripherals limiting the area and energy efficiency [9]. Here, we circumvent these challenges by introducing a compact and low-power BN architecture embedded in memory based on 2D memtransistors.

Introduction

Animals gather information from their surroundings with the help of their sensory organs and process the same using their brain to make decisions, which aid them to locate mates, find preys, and escape from predators. Better decision making leads to better survival, which naturally translates into the evolutionary success for the species. However, it is often very difficult for animals to gather accurate information either due to the limitations of their sensory organ or because of the information being obscured by noisy environment. For example, visual cues are unreliable source of information for freshwater fishes like the rainbow trout since these can be both misleading and manipulated by a predator. In contrast chemical cues released into the water from the epidermis of an injured fish are more reliable indicators of predatory events [10]. The decision to invoke an alarm responses, i.e. less swimming and more time to resume foraging, therefore, depends on how the brain of the rainbow trout processes the visual and the chemical cues based on their relative probability of success from prior experiences. While the neural basis of such computation is relatively unknown, the mathematical construct can be represented using a Bayesian network (BN) with theoretical foundation in Bayes' theorem.

BN is a probabilistic graphic network used to estimate and infer the probability of events, where events are interdependent on each other [11]. **Fig. 1a** shows the basic building block of a BN comprising of a parent node, A , a child node, B , and an edge connecting the two. Each node

represents an event, e.g., presence of chemical cue or A and presence of a predator or B , and the connection represents how two events are mutually dependent. The dependence is provided in a conditional probability table (CPT), which contains the conditional probability or likelihood values, i.e. $P(B/A)$ and $P(B/A^c)$, where A^c is the complement of the event A . In the present example, these represents the likelihood of presence of a predator given that a chemical cue is present (A) or absent (A^c), respectively. Knowing the probability of occurrence for the event A , i.e., $P(A)$, the marginal probability of occurrence of the event B , i.e. $P(B)$ can be evaluated using Bayes' theorem following Eq. 1.

$$P(B) = P(B/A)P(A) + P(B/A^c)P(A^c) = P(B/A)P(A) + P(B/A^c)[1 - P(A)] \quad [1a]$$

$$P(A) + P(A^c) = 1 \quad [1b]$$

In a generic BN, a child node can have multiple parent nodes, and a parent node can have multiple children. For example, **Extended Data Fig. 1a** shows a BN where the child node, B is connected to n parent nodes, $A_1, A_2, \dots, A_n$. Note that the CPT in this instance contains $N = 2^n$ entries, which are the conditional probability or likelihood for the occurrence of the event B for all possible combinations of occurrence of the events A_i 's, $i = 1, 2, \dots, n$. Similarly, **Extended Data Fig. 1b** shows a BN where the parent node, A is connected to m children, $B_1, B_2, \dots, B_m$. In this case, there are m CPTs each with $N = 2$ entries.

Note that the probability estimation for a child node with n parent require $n2^n$ multiplications, $n2^{n-1}$ subtractions, and $2^n - 1$ additions or a total of $2^{n-1}(3n + 2) - 1$ arithmetic operations. Needless to mention that the number of arithmetic operations increases as the number of nodes in the BN increases. This makes hardware acceleration of BN using conventional silicon complementary metal oxide semiconductor (CMOS) technology resources extensive since

arithmetic operations require circuits consisting of hundreds of transistors, which have large footprints and consume significant amount of energy. Furthermore, the von Neumann bottleneck necessitates storing of the CPT in the memory, which is physically separated from the arithmetic core and, therefore, requires frequent data shuttling between the two increasing the energy burden. In contrast, even the tiniest brains with very limited number of neurons can perform such apparently complex computational tasks with miniscule energy expenditure. The success of biological brains in implementing BN lie in the inherently stochastic nature of neural computation.

Figure 1. Bayesian networks (BNs). a) Schematic of basic building block of a BN comprising of a parent node, A , a child node, B , and an edge connecting the two. Each node represents an event, and the connection represents how two events are mutually dependent. The dependence is provided in a conditional probability table (CPT), which contains the conditional probability or likelihood values, i.e., $P(B/A)$ and $P(B/A^c)$, where A^c is the complement of the event. Knowing the probability of occurrence for the event A , i.e., $P(A)$, the marginal probability of occurrence of the event B , i.e., $P(B)$ can be evaluated using Bayes' theorem. b) Acceleration of 2-node BN in (a) using three stochastic bit (s -bit) generators and one 2×1 multiplexer (MUX) circuit. c) Examples of BN architecture that represent real-life situations from ecology to forecasting and drug discovery highlighting its usefulness in decision making.

Drawing inspiration from biology, stochastic computing (SC) has been explored for the hardware acceleration of BN [12]. The key difference is that unlike classical computing where information is presented in the form of binary values (1's and 0's), SC encodes information using stochastic bits (s-bits) that are interpreted as probabilities that fall in the interval $[0,1]$. For instance, the bit-stream, $S = [1\ 0\ 0\ 1\ 0\ 1\ 0\ 0]$ encodes the value $P(S) = 3/8$, i.e., the probability of finding '1' in the bit-stream, S . An attractive feature of SC is that arithmetic operations to be performed using simple logic gates [13, 14]. For example, the 2-node BN in **Fig. 1a** can be realized using a multiplexer (*MUX*) circuit as shown in **Fig. 1b**. The output, B , of a *MUX* with two input variables, X_1 , X_2 , and a select line, A , is given by Eq. 2.

$$B = AX_1 + A^cX_2 \quad [2]$$

Instead of being digital variables, if X_1 , X_2 , and A represent stochastic variables with $P(X_1)$, $P(X_2)$, and $P(A)$ being the probability of obtaining '1' in their respective bit-streams, then B also transforms into a random variable whose probability is given by Eq. 3.

$$P(B) = P(A)P(X_1) + P(A^c)P(X_2) \quad [3]$$

Note that, if $P(X_1) = P(B/A)$ and $P(X_2) = P(B/A^c)$, then Eq. 3 transforms into Eq. 1. Therefore, hardware acceleration of a child node with single parent can be accomplished by using 3 s-bit generators and a 2×1 *MUX*. Interestingly, the *MUX* architecture can be scaled to accelerate any BN. For example, hardware acceleration of BN in **Extended Data Fig. 1a** can be achieved by using n s-bit generators to obtain the A_i 's, another $N = 2^n$ s-bit generators to obtain the CPT, and one $N \times 1$ *MUX* with n select lines as shown in **Extended Data Fig. 1c**. Similarly, **Extended Data Fig. 1d** shows the hardware architecture for the BN in **Extended Data Fig. 1b** consisting of 1 s-bit generator to obtain A , another $2m$ s-bit generators to obtain the m CPTs, and m 2×1 *MUX*s.

Note that BN architecture can be used to represent many real-life situations as shown in **Fig. 1c**. For example, in the case of the rainbow trout, the events A_1 and A_2 represent the presence of visual and chemical cues, respectively, which are independent of each other, the event B represents the presence of a predator, and the events C_1 and C_2 represent the decision taken by the rainbow trout to stop swimming and stop foraging, respectively, which are also independent of each other but depend on B . Similarly, in forecasting, the events A_1 and A_2 represent the probability of a day being cloudy and windy, respectively, the event B represent the probability of rain, and the events C_1 and C_2 may represent the decision to purchase an umbrella or drink coffee, respectively. Finally, a third example is derived from genetics and drug discovery, where the events A_1 and A_2 may represent the probability of expressing *gene 1* and *gene 2* when intervening with a specific drug, the event B represents the activation of a critical signaling pathway, and the events C_1 and C_2 represent production of specific hormones or antibodies. The above discussion exemplifies the usefulness of BNs in depicting causal relationship using acyclic graphs, which can subsequently be used to predict outcomes based on prior knowledge and likelihood. For example, to predict the relative effectiveness between drug-1 and drug-2 that influence expression for *gene 1* and *gene 2*, respectively, the only experiments that one needs to do is to obtain respective prior, i.e., $P(A_1)$ and $P(A_2)$. BN can then be used to obtain marginal likelihoods, i.e., $P(C_1)$ and/or $P(C_2)$ to assess the relative effectiveness of the two drugs.

The fundamental computing primitive for BN is a s-bit generator, which allows control of the output probability of obtaining ‘1’ in a given bit-stream. So far, probabilistic CMOS [15], field-programmable gate arrays (FPGA) [16-18], memristors [3, 4], and spintronic devices [19-22] have been successfully used for BN acceleration. However, CMOS and FPGA based BN architecture

require hundreds of transistors to generate s-bits due, which limits its area and energy efficiency [9]. In contrast, memristors offer inherent stochasticity in their switching dynamics, which can be exploited to obtain random bits. However, memristor-based BN architectures heavily rely on CMOS peripherals to translate random bits into s-bits and for subsequent logic operations using those s-bits. Recently, spintronic device such as magnetic random access memory (MRAM) [23] and magnetic tunnel junctions (MTJs) [24] have shown potential for BN acceleration since s-bits can be obtained by controlling the probability of spin-flip through externally driven current. However, temperature and supply voltage fluctuations can impact the spin-flip probability, which necessitates additional CMOS-based peripheral circuits to remove the bit-bias. In addition, spin-based devices still require CMOS-based logic circuits for BN acceleration.

In this work, we demonstrate hardware acceleration of BN using a monolithic memtransistor technology based on two-dimensional (2D) semiconductors such as monolayer MoS₂. Memtransistors are three terminal devices with the gate terminal allowing non-volatile and analog programming of the conductance states, which can be readout by applying source-to-drain bias. Our main contributions in this work are 1) the design of an area and energy efficient s-bit generator circuit comprising of six memtransistors to achieve tunable probability of obtaining ‘1’ in the bit-stream in the range [0,1] and 2) integration of s-bit generators with 2D memtransistor based 2×1 *MUX* that consists of three *NAND* gates and one as *NOT* gate for BN acceleration. In brief, we exploit the inherent stochasticity in the charge trapping and detrapping processes in the gate dielectric of the memtransistor as the source of randomness. Our in-memory compute approach based on 3-terminal 2D memtransistors not only overcomes the von Neumann limitations of

conventional digital CMOS, but also eliminates the need for peripherals, which is inescapable for emerging memristor- and spin-based 2-terminal stochastic devices for BN acceleration.

Fabrication and characterization of 2D memtransistors

Fig. 2a-b, respectively, show the 3D schematic and optical image of a representative 2D memtransistor based on monolayer MoS₂, which are locally back-gated using a stack comprising of atomic layer deposition (ALD) grown 50 nm Al₂O₃ on sputter deposited 40/30 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO₂/p⁺⁺-Si substrate. As we will discuss later, the analog, non-volatile, and stochastic programming capability offered by the Al₂O₃/Pt/TiN gate stack is central to our BN architecture. The monolayer MoS₂ used in this work was grown using metal organic chemical vapor deposition (MOCVD) technique on sapphire substrate at 950 °C [25, 26]. Use of epitaxial substrate and elevated growth temperature ensure uniform and high quality 2D film, which is critical for the successful demonstration of our BN architecture that involves many 2D memtransistors. The monolayer MoS₂ film was transferred from the growth substrate to the SiO₂/p⁺⁺-Si substrate with predefined islands of Al₂O₃/Pt/TiN for subsequent 2D memtransistor fabrication. Details on monolayer MoS₂ synthesis, film transfer, and fabrication of the local back-gate gate islands, MoS₂ memtransistors, and BN architecture can be found in the *Methods* section.

The film quality and device performance were assessed using optical and electrical measurements. Raman spectra (**Extended Data Fig. 2a**) obtained for a representative 2D memtransistor shows two characteristics monolayer MoS₂ peaks at 383 cm⁻¹ and 404 cm⁻¹ corresponding to the in-plane E_{2g}^1 and out-of-plane A_{1g} modes, respectively, with the expected peak separation of ~20 cm⁻¹ for

2D memtransistors for hardware acceleration of Bayesian network (BN)

Figure 2. 2D memtransistors. a) 3D schematic and b) optical image of a representative 2D memtransistor based on monolayer MoS₂, which are locally back-gated using a stack comprising of atomic layer deposition (ALD) grown 50 nm Al₂O₃ on sputter deposited 40/30 nm Pt/TiN. All back-gate islands were placed on a commercially purchased SiO₂/p⁺⁺-Si substrate. c) Transfer characteristics, i.e., source-to-drain current (I_{DS}) versus local back-gate voltage (V_{BG}) measured using source-to-drain bias, $V_{DS} = 1$ V, in linear and logarithmic scale for a representative MoS₂ memtransistor with channel length, $L = 1$ μm , and channel width, $W = 5$ μm . d) Output characteristics, i.e., I_{DS} versus V_{DS} for different V_{BG} for the same MoS₂ memtransistor. e) Post-programmed and f) post-erased transfer characteristics of a representative 2D memtransistor. g) Non-volatile retention for 4 representative post-programmed and post-erased conductance states (G_{MT}) for 100 seconds.

monolayer MoS₂ [27]. Similarly, the photoluminescence (PL) spectra (**Extended Data Fig. 2b**) shows peak at 1.83 eV corresponding to the direct bandgap of monolayer MoS₂. The transfer characteristics, i.e. source to drain current (I_{DS}) versus local back-gate voltage (V_{BG}) measured using source-to-drain bias, $V_{DS} = 1$ V, in linear and logarithmic scale for a representative MoS₂ memtransistor with channel length, $L = 1$ μm , and channel width, $W = 5$ μm is shown in **Fig. 2c**. As expected, n-type transport is observed in MoS₂, which is attributed to the pinning of the metal Fermi level near the conduction band [28-30]. Nevertheless, MoS₂ memtransistor exhibits excellent electrostatic gate control with current on/off ratio ($r_{ON/OFF}$) $> 10^5$, subthreshold slope (SS) < 400 mV/decade averaged over 3 orders of magnitude change in I_{DS} , minimal gate hysteresis when measured in air, and low gate leakage current. The threshold voltage (V_{TH}) was found to be ~ 2.2 V extracted at iso-current of 10 nA/ μm and the electron field effect mobility (μ_{FE}) extracted

from the peak trans-conductance was found to be $6 \text{ cm}^2/\text{V}\cdot\text{s}$. **Fig. 2d** shows the output characteristics, i.e. I_{DS} versus V_{DS} for different V_{BG} for the same MoS_2 memtransistor. The on current (I_{ON}) reached as high as $\sim 15 \text{ }\mu\text{A}/\mu\text{m}$ for an inversion carrier density of $\sim 1 \times 10^{12}/\text{cm}^2$ at $V_{\text{DS}} = 5 \text{ V}$. These results indicate reasonably good quality monolayer film growth using MOCVD, relatively damage-free film transfer, and clean memtransistor fabrication processes.

The post-programmed and post-erased transfer characteristics of a representative 2D memtransistor after being subjected to negative “Write” (V_P) and positive “Erase” (V_E) voltage pulses of different amplitudes applied to the local back-gate electrode, each for a duration of $\tau_{P/E} = 100 \text{ ms}$ is shown in **Fig. 2e**. The negative and positive shift in the respective transfer characteristics can be ascribed to electron trapping and detrapping at and near the $\text{MoS}_2/\text{Al}_2\text{O}_3$ interface, respectively. The trapping and de-trapping processes were found to be non-volatile as shown in **Fig. 2f** for 4 representative post-programmed and post-erased conductance states (G_{MT}), respectively, for 100 seconds. We also examined long-term memory retention for the 2D memtransistors and found that states remain indistinguishable even after 3 hrs. Memory retention is important to store the CPT and the memtransistors demonstrate adequate memory performance for the acceleration of BN using SC.

Programming stochasticity in 2D memtransistor and design of s-bit generator

Design of hardware for high-quality random bit generation is central to the acceleration of BN. Here, we exploit the cycle-to-cycle variation in the post-programmed and post-erased conductance states (G_{MT}) of 2D memtransistor as the source of true randomness. **Fig. 3a** shows the transfer characteristics of a representative MoS_2 memtransistor, which is measured each time after applying

$V_P = -10$ V and $V_E = 10$ V for $\tau_s = 100$ ms, for a total of 100 cycles and **Fig. 3b-c**, respectively, show the histogram of post-programmed and post-erased G_{MT} extracted at $V_{BG} = 0$ V. Clearly, G_{MT} follow Gaussian random distributions. To translate the stochastic conductance fluctuation into s-bits, we deploy a circuit consisting of six memtransistors ($MT1$, $MT2$, $MT3$, $MT4$, $MT5$, and $MT6$) as shown using the circuit diagram and the corresponding optical image in **Fig. 3d-e**, respectively. The voltage waveforms applied to the nodes, $N1$, $N2$, i.e., V_{N1} , V_{N2} , respectively, are shown in **Fig. 3f**. Note that during each clock cycle (τ_{clk}), V_{N1} switches between 0 V, 0 V, and 2 V and V_{N2} switches between $V_P = -7$ V, $V_E = 10$ V, and $V_R = 1$ V. Voltages applied to nodes, $N3$, and $N4$, i.e., V_{N3} , and V_{N4} are held constant at 1 V and 0 V, respectively. This allows programming and reset of $MT1$ during each τ_{clk} . The voltage readout at node, $N5$, i.e., V_{N5} is shown in **Fig. 3g**, which exhibits stochastic fluctuation. Note that the series connection of memtransistors, $MT1$ and $MT2$ represents a voltage divider circuit, and hence V_{N5} is determined by their respective conductance values, i.e., G_{MT1} and G_{MT2} . Since G_{MT1} fluctuates from cycle-to-cycle owing to programming and reset voltages applied to its local back-gate terminal, i.e., $N2$, so does V_{N5} . In other words, the voltage divider translates conductance fluctuation into voltage fluctuation. **Fig. 3h** shows the histogram of V_{N5} , which, as expected, follows a random Gaussian distribution with mean, $\mu_{VN5} = 0.37$ V and standard deviation, $\sigma_{VN5} = 0.05$ V.

Next the Gaussian distribution is broadened by using an inverting amplifier constructed using $MT3$ and $MT4$. Note that the local back-gate of $MT3$ is shorted to its source at node, N_6 . This ensures that $MT3$ operates as a depletion mode (normally on) transistor or as a load resistor. **Fig. 3i** shows the output, V_{N6} , as a function of the input, V_{N5} . The slope of the curve is referred to as the gain of the amplifier, and higher the gain wider is the broadening of the Gaussian. We achieved a gain of

Figure 3. 2D memtransistor based s-bit generator. a) Transfer characteristics of a representative 2D memtransistor, measured each time after the application of $V_P = -10$ V and $V_E = 10$ V each for $\tau_s = 100$ ms, for a total of 100 cycles. Distribution of b) post-programmed and c) post-erased conductance (G_{MT}) measured using $V_{BG} = 0$ V. d) Circuit diagram and e) corresponding optical image for the proposed s-bit generator consisting of six memtransistors (MT1, MT2, MT3, MT1, MT2, MT3). f) Voltage waveforms applied to the nodes, N1, N2, i.e., V_{N1} , V_{N2} . During each clock cycle (τ_{clk}), V_{N1} toggles between 0 V, 0 V, and $V_{DD} = 2$ V and V_{N2} toggles between $V_P = -7$ V, $V_E = 10$ V, and $V_R = 1$ V. Voltages applied to nodes, N3, and N4, i.e., V_{N3} , and V_{N4} are held constant at 1V and 0 V, respectively. g) Voltage readout at node, N5, i.e., V_{N5} . h) Distribution of V_{N5} over 200 τ_{clk} follows a random Gaussian distribution with mean, $\mu_{VN5} = 0.37$ V and standard deviation, $\sigma_{VN5} = 0.05$ V. i) Output, V_{N6} , of an inverting amplifier constructed using MT3 and MT4 as a function of the input, V_{N5} with a gain of ~ 8 . j) V_{N6} corresponding to V_{N5} shown in (g). k) Distribution of V_{N6} follows a random Gaussian distribution with mean, $\mu_{VN6} = 0.74$ V and an increased standard deviation of $\sigma_{VN6} = 0.3$ V. l) Output, V_{N7} , of a thresholding inverter constructed using MT5 and MT6 as a function of the input, V_{N6} for different inversion threshold, V_{IT} . m) V_{N7} corresponding to V_{N6} shown in (i) for different V_{IT} . n) Probability of obtaining '1' in the bit stream (p_s) as a function of V_{IT} .

~8. The gain can be increased by cascading multiple amplifiers; however it adds area and energy overhead. **Fig. 3j** shows V_{N6} corresponding to V_{N5} obtained in **Fig. 3g**. Clearly, the histogram of V_{N6} shown in **Fig. 3k** exhibit a Gaussian distribution with mean, $\mu_{V_{N6}} = 0.74$ V and an increased standard deviation of $\sigma_{V_{N6}} = 0.3$ V.

Finally, to transform the analog fluctuations seen in V_{N6} into s-bits, a thresholding inverter with programmable inversion threshold, V_{IT} , is constructed using $MT5$ and $MT6$. **Fig. 3l** shows the output, V_{N7} , as a function of the input, V_{N6} for different V_{IT} . Note that V_{IT} is the magnitude of V_{N6} for which V_{N7} reaches $V_{DD}/2$, i.e. 1 V in the present case. The programmability of V_{IT} is a critical feature that distinguishes 2D memtransistors based inverters from conventional CMOS-based inverters and allows us to seamlessly obtain the s-bits. **Fig. 3m** shows V_{N7} corresponding to V_{N6} obtained in **Fig. 3j** for different V_{IT} and **Fig. 3n** shows the corresponding probability of obtaining ‘1’ in the bit stream, i.e., p_s as a function of V_{IT} . As expected, if V_{IT} is too low, then almost all V_{N6} values translate into $V_{N7} \approx 0$ V, which is reflected as near zero p_s . Similarly, if V_{IT} is too high, then almost all V_{N6} values translate into $V_{N7} \approx 2$ V leading to $p_s = 1$. Between these two extremes, p_s increases monotonically with V_{IT} . This clearly shows that we are able to convert the cycle-to-cycle random conductance fluctuations in 2D memtransistor into s-bits with reconfigurable p_s that lie between [0,1] using the circuit based on 6 memtransistors. The average energy expenditure for s-bit generation (E_{s-bit}) was calculated using Eq. 3.

$$E_{s-bit} = \frac{1}{2} C_G [V_P^2 + V_E^2 + V_{DD}^2]; C_G = \frac{\epsilon_0 \epsilon_{ox} W L}{t_{ox}} \quad [3]$$

In Eq. 3, C_G is the gate capacitance, $\epsilon_0 = 8.85 \times 10^{-12}$ F/m is the vacuum permittivity, $\epsilon_{ox} = 10$, and $t_{ox} = 50$ nm are, respectively, the relative permittivity and thickness of Al_2O_3 . We found that $E_{s-bit} < 2$ pJ/cycle, which supports our claim on energy efficient s-bit generation. The active

footprint for the s-bit generator is $\sim 30 \mu\text{m}^2$ since each memtransistor has an active device area of $\sim 5 \mu\text{m}^2$ excluding the large contact pads used for probing. Since monolayer 2D materials offer aggressive dimensional scalability, it is possible to reduce the footprint of s-bit generators even further. Nevertheless the use of only 6 memtransistors is the key towards the realization of area and energy efficient s-bit generator circuit.

2D memtransistor based digital circuits and BN acceleration

As described earlier, stochastic multiplexers (*MUX*s) can be used for computing the marginal probability values at any BN node. **Fig. 4a** shows the circuit configuration of a 2×1 *MUX*, which consists of one inverter and three 2-input *NAND* gates. **Fig. 4b** show the optical image and corresponding circuit configuration of a 2-input *NAND* gate comprising of 3 memtransistors, *MT1*, *MT2*, and *MT3* connected in series with *MT1* serving as the depletion load. The supply voltage, $V_{DD} = 2 \text{ V}$, is applied to the drain terminal of *MT1* at node N_1 , whereas the source terminal of *MT3*, i.e., node N_5 is kept grounded. **Fig. 4c** shows the input waveforms, V_{N3} and V_{N4} , which are applied to the local back-gate terminals of *MT2* and *MT3* at nodes N_3 and N_4 , respectively, and the corresponding output waveform, V_{N2} , which is obtained at node N_2 . Clearly, the circuit operates as a *NAND* gate.

Fig. 4d-e, respectively, show the optical image and corresponding circuit configuration for hardware acceleration of a 2-node BN consisting of 3 s-bit generators and a 2×1 *MUX* with a total of 29 memtransistors. The V_{IT} values for the s-bit generators for X_1 and X_2 can be pre-programmed using the CPT for the nodes A and B . **Fig. 4f** shows the representative stochastic bit-streams for the random variables A , X_1 , and X_2 with $P(A) = 0.12$, $P(X_1) = P(B/A) = 0.26$, and $P(X_2) =$

Figure 4. Hardware acceleration of BN. a) Circuit schematic for hardware acceleration of BN using three s-bit generator and one 2×1 MUX. The MUX consists of one inverter and three 2-input NAND gates. b) Optical image and corresponding circuit configuration of a 2-input NAND gate comprising of 3 memtransistors, MT₁ MT₂, and MT₃ connected in series with MT₁ serving as the depletion load. c) Input waveforms, V_{N3} and V_{N4}, which are applied to the local back-gate terminals of MT₂ and MT₃ at nodes N₃ and N₄, respectively, and the corresponding output waveform, V_{N2}, which is obtained at node N₂. d) Optical image and e) corresponding circuit configuration for hardware acceleration of a 2-node BN consisting of 3 s-bit generators and a 2×1 MUX with a total of 29 memtransistors. The V_{IT} values for the s-bit generators for X₁ and X₂ can be pre-programmed using the CPT for the nodes A and B. f) Representative stochastic bit-streams for the random variables A, X₁, and X₂ with $P(A) = 0.12$, $P(X_1) = P(B/A) = 0.26$, and $P(X_2) = P(B/A^c) = 0.36$. g) Correlation coefficient (CC) values between A, X₁, and X₂ confirm mutual independence of the s-bit generator modules. h) Stochastic bit-streams obtained at the output node, B. The measured and expected values for $P(B)$ are 0.56 and 0.54.

$P(B/A^c) = 0.36$. Note that accurate estimation of $P(B)$ requires that the stochastic input variables to the MUX, i.e., A, X₁, X₂ must be mutually independent. **Fig. 4g** shows the correlation coefficient (CC) between these three variables. The CC values were found to be close to zero, which confirm mutual independence of the s-bit generator modules. **Fig. 4h** shows the stochastic bit-streams

obtained at the output node, B . The measured and expected values for $P(B)$ are 0.56 and 0.54.

Extended Data Fig. 3 shows the results for three more sets of measurement. In all instances, we found that our 29 memtransistor module is able to accelerate a 2-node BN with relatively high accuracy. The average energy expenditure for the BN acceleration is miniscule ~ 1.2 nJ, when 200 τ_{clk} are used. Certainly, the energy expense can be reduced by reducing the length of the s-bit streams at the cost of reduced precision.

Conclusion

In conclusion, we have exploited the cycle-to-cycle variability in the programmed conductance of 2D memtransistors and transcribed the same into s-bits with reconfigurable probability of obtaining ‘1’ in the bit-stream using a circuit that comprises of only 6 memtransistors and by spending < 2 pJ per clock-cycle. We subsequently combined the s-bit generator with 2D memtransistor based 2×1 *MUX* to demonstrated hardware acceleration of BN. The BN architecture comprises of total 29 memtransistors and require ~ 1.2 nJ energy for precise computation. Our demonstration of memtransistor based standalone in-memory compute fabric shows the potential for emerging 2D materials and devices.

Methods

Fabrication of local back-gate islands: To define the back-gate island regions, the substrate 285 nm SiO₂ on p⁺⁺-Si was spin coated with bilayer photoresist consisting of Lift-Off-Resist (LOR 5A) and Series Photoresist (SPR 3012) baked at 185 °C and 95 °C, respectively. The bilayer photoresist was then exposed to Heidelberg Maskless Aligner (MLA 150) to define the island and developed using MF CD26 microposit, followed by a de-ionized (DI) water rinse. The back gate electrode of 20/50 nm TiN/Pt was deposited using reactive sputtering. The photoresist was removed using acetone and Photo Resist Stripper (PRS 3000) and cleaned using 2-propanol (IPA) and DI water. Atomic layer deposition (ALD) process was then implemented to grow 50 nm Al₂O₃ on the entire substrate including the island regions. To access the individual Pt back-gate electrodes etch patterns were defined using the same bilayer photoresist consisting of LOR 5A and SPR 3012. The bilayer photoresist was then exposed to MLA 150 and developed using MF CD26 microposit. 50 nm Al₂O₃ was subsequently dry etched using the BCl₃ chemistry at 5 °C for 20 seconds, which was repeated four times to minimize heating in the substrate. Next the photoresist was removed to give access to the individual Pt electrodes.

Large area monolayer MoS₂ film growth: Monolayer MoS₂ was deposited on epi-ready 2” c-sapphire substrate by metalorganic chemical vapor deposition (MOCVD). An inductively heated graphite susceptor equipped with wafer rotation in a cold-wall horizontal reactor was used to achieve uniform monolayer deposition as previously described [31]. Molybdenum hexacarbonyl (Mo(CO)₆) and hydrogen sulfide (H₂S) were used as precursors. Mo(CO)₆ maintained at 10°C and 650 Torr in a stainless-steel bubbler was used to deliver 1.1×10^{-3} sccm of the metal precursor for the growth, while 400 sccm of H₂S was used for the process. MoS₂ deposition was carried out at

1000°C and 50 Torr in H₂ ambient, where monolayer growth was achieved in 18 min. The substrate was first heated to 1000°C in H₂ and maintained for 10 min before the growth was initiated. After growth, the substrate was cooled in H₂S to 300°C to inhibit decomposition of the MoS₂ films. More details can be found in our earlier work [26, 32, 33].

MoS₂ film transfer to local back-gate islands: To fabricate the 2D memtransistors, MOCVD grown monolayer MoS₂ film was transferred from the sapphire to SiO₂/p⁺⁺-Si substrate with local back-gate islands using PMMA (polymethyl-methacrylate) assisted wet transfer process. First, MoS₂ on sapphire substrate was spin coated with PMMA and then baked at 180 °C for 90 s. The corners of the spin-coated film were scratched using a razor blade and immersed inside 1 M NaOH solution kept at 90 °C. Capillary action causes the NaOH to be drawn into the substrate/film interface, separating the PMMA/ MoS₂ film from the sapphire substrate. The separated film was rinsed multiple times inside a water bath and finally transferred onto the SiO₂/p⁺⁺-Si substrate with local back-gate islands and then baked at 50 °C and 70 °C for 10 min each to remove moisture and residual PMMA, ensuring a pristine interface.

Fabrication of 2D memtransistors: To define the channel regions for the memtransistors, the substrate was spin-coated with PMMA and baked at 180 °C for 90 s. The resist was then exposed to electron beam (e-beam) and developed using 1:1 mixture of 4-methyl-2-pentanone (MIBK) and 2 propanol (IPA). The monolayer MoS₂ film was subsequently etched using sulfur hexafluoride (SF₆) at 5 °C for 30 s. Next, the sample was rinsed in acetone and IPA to remove the e-beam resist. To define the source and drain contacts, sample is then spin coated with methyl methacrylate (MMA) followed by A3 PMMA. Then using e-beam lithography source and drain contacts are

patterned and developed by using 1:1 mixture of MIBK and IPA for 60s. 40 nm of Nickel (Ni) and 30 nm of Gold (Au) are deposited using e-beam evaporation. Finally, lift-off process is performed to remove the evaporated Ni/Au except from the source/drain patterns by immersing the sample in acetone for 30 min followed by IPA for another 30 mins. Each island contains one memtransistor to allow for individual gate control.

Monolithic Integration: To define the connections between the respective memtransistors the substrate was spin coated with MMA and PMMA, followed by the e-beam lithography and developing using 1:1 mixture of MIBK and IPA, and e-beam evaporation of 60 nm Au. Finally, the e-beam resist was rinsed away by lift-off process using acetone and IPA.

Electrical Characterization: Electrical characterization of the fabricated devices are performed using Lake Shore CRX-VF probe station under atmospheric condition using a Keysight B1500A parameter analyzer.

Data Availability: The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code Availability: The codes used for plotting the data are available from the corresponding authors on reasonable request.

AUTHOR INFORMATION

Corresponding Author

sud70@psu.edu, das.sapt@gmail.com

Author Contributions

S.D conceived the idea and designed the experiments. Y.Z., H.R., and T. F. S. fabricated the memtransistors. Y.Z., H.R., and S.D performed the measurements, analyzed the data, discussed the results, and agreed on their implications. N. T. grew MOCVD MoS₂. All authors contributed to the preparation of the manuscript.

Competing Interest

The authors declare no competing interests

Acknowledgement

The work was supported by Army Research Office (ARO) through Contract Number W911NF1920338. Authors also acknowledge the materials support from the National Science Foundation (NSF) through the Pennsylvania State University 2D Crystal Consortium–Materials Innovation Platform (2DCCMIP) under NSF cooperative agreement DMR-1539916.

Reference

1. Smithson, S.C., et al., *Efficient CMOS invertible logic using stochastic computing*. IEEE Transactions on Circuits and Systems I: Regular Papers, 2019. **66**(6): p. 2263-2274.
2. Ardakani, A., et al., *VLSI implementation of deep neural network using integral stochastic computing*. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2017. **25**(10): p. 2688-2699.
3. Knag, P., W. Lu, and Z. Zhang, *A native stochastic computing architecture enabled by memristors*. IEEE Transactions on Nanotechnology, 2014. **13**(2): p. 283-293.
4. Gaba, S., et al. *Memristive devices for stochastic computing*. in *2014 IEEE International Symposium on Circuits and Systems (ISCAS)*. 2014. IEEE.
5. Gaba, S., et al., *Stochastic memristive devices for computing and neuromorphic applications*. Nanoscale, 2013. **5**(13): p. 5872-5878.
6. Venkatesan, R., et al. *Spintastic: Spin-based stochastic logic for energy-efficient computing*. in *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. 2015. IEEE.
7. Finocchio, G., et al., *The promise of spintronics for unconventional computing*. arXiv preprint arXiv:1910.07176, 2019.
8. Hu, J., et al., *Spin-hall-effect-based stochastic number generator for parallel stochastic computing*. IEEE Transactions on Electron Devices, 2019. **66**(8): p. 3620-3627.
9. Yang, K., et al. *16.3 A 23Mb/s 23pJ/b fully synthesized true-random-number generator in 28nm and 65nm CMOS*. in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*. 2014. IEEE.
10. Brown, G.E. and R.J.F. Smith, *Conspecific skin extracts elicit antipredator responses in juvenile rainbow trout (*Oncorhynchus mykiss*)*. Canadian Journal of Zoology, 1997. **75**(11): p. 1916-1922.
11. Puga, J.L., M. Krzywinski, and N. Altman, *Bayesian networks*. Nature Methods, 2015. **12**(9): p. 799-800.
12. Alaghi, A. and J.P. Hayes, *Survey of stochastic computing*. ACM Transactions on Embedded computing systems (TECS), 2013. **12**(2s): p. 1-19.
13. Gaines, B.R. *Stochastic computing*. in *Proceedings of the April 18-20, 1967, spring joint computer conference*. 1967.
14. Poppelbaum, W., C. Afuso, and J. Esch. *Stochastic computing elements and systems*. in *Proceedings of the November 14-16, 1967, fall joint computer conference*. 1967.
15. Weijia, Z., G.W. Ling, and Y.K. Seng. *PCMOs-based Hardware Implementation of Bayesian Network*. in *2007 IEEE Conference on Electron Devices and Solid-State Circuits*. 2007.
16. Cai, R., et al., *VIBNN: Hardware Acceleration of Bayesian Neural Networks*, in *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems*. 2018, Association for Computing Machinery: Williamsburg, VA, USA. p. 476-488.
17. Kulesza, Z. and W. Tylman. *Implementation Of Bayesian Network In FPGA Circuit*. in *Proceedings of the International Conference Mixed Design of Integrated Circuits and System, 2006. MIXDES 2006*. 2006.
18. Zermani, S., et al. *FPGA implementation of Bayesian network inference for an embedded diagnosis*. in *2015 IEEE Conference on Prognostics and Health Management (PHM)*. 2015.
19. Debashis, P., et al., *Hardware implementation of Bayesian network building blocks with stochastic spintronic devices*. Scientific Reports, 2020. **10**(1).
20. Faria, R., et al., *Hardware Design for Autonomous Bayesian Networks*. Frontiers in Computational Neuroscience, 2021. **15**(14).

21. Shim, Y., et al., *Stochastic Spin-Orbit Torque Devices as Elements for Bayesian Inference*. Scientific Reports, 2017. **7**(1): p. 14101.
22. Faria, R., K.Y. Camsari, and S. Datta, *Implementing Bayesian networks with embedded stochastic MRAM*. AIP Advances, 2018. **8**(4): p. 045101.
23. Jaiswal, A., X. Fong, and K. Roy, *Comprehensive scaling analysis of current induced switching in magnetic memories based on in-plane and perpendicular anisotropies*. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2016. **6**(2): p. 120-133.
24. Sengupta, A., et al., *Magnetic tunnel junction mimics stochastic cortical spiking neurons*. Scientific reports, 2016. **6**(1): p. 1-8.
25. 2DCC. *2d-crystal-consortium*. Available from: <https://www.mri.psu.edu/2d-crystal-consortium/user-facilities/thin-films/list-thin-film-samples-available>.
26. Sebastian, A., et al., *Benchmarking monolayer MoS₂ and WS₂ field-effect transistors*. Nature Communications, 2021. **12**(1): p. 693.
27. Li, H., et al., *From bulk to monolayer MoS₂: evolution of Raman scattering*. Advanced Functional Materials, 2012. **22**(7): p. 1385-1390.
28. Das, S., et al., *High performance multilayer MoS₂ transistors with scandium contacts*. Nano letters, 2013. **13**(1): p. 100-105.
29. Schulman, D.S., A.J. Arnold, and S. Das, *Contact engineering for 2D materials and devices*. Chemical Society Reviews, 2018. **47**(9): p. 3037-3058.
30. Chuang, S., et al., *MoS₂ p-type transistors and diodes enabled by high work function MoO_x contacts*. Nano letters, 2014. **14**(3): p. 1337-1342.
31. Xuan, Y., et al., *Multi-scale modeling of gas-phase reactions in metal-organic chemical vapor deposition growth of WSe₂*. Journal of Crystal Growth, 2019. **527**.
32. Jayachandran, D., et al., *A low-power biomimetic collision detector based on an in-memory molybdenum disulfide photodetector*. Nature Electronics, 2020. **3**(10): p. 646-655.
33. Dodda, A., et al., *Stochastic resonance in MoS₂ photodetector*. Nature Communications, 2020. **11**(1): p. 4406.

Extended Data Figure 1

Extended Data Figure 1. a) A BN where the child node, B is connected to n parent nodes, $A_1, A_2, \dots, A_n$. b) A BN where the parent node, A is connected to m children, $B_1, B_2, \dots, B_m$. c) Hardware acceleration of BN shown in (a) can be achieved by using n s -bit generators to obtain the A_i 's, another $N = 2^n$ s -bit generators to obtain the CPT, and one $N \times 1$ MUX with n select lines d) Hardware acceleration of BN shown in (b) can be achieved by using 1 s -bit generator to obtain A , another $2m$ s -bit generators to obtain the m CPTs, and m 2×1 MUXs.

Extended Data Figure 2

Extended Data Figure 2. a) Raman spectra obtained for a representative 2D memtransistor shows two characteristics monolayer MoS_2 peaks at 383 cm^{-1} and 404 cm^{-1} corresponding to the in-plane E_{2g}^1 and out-of-plane A_{1g} modes, respectively, with the expected peak separation of $\sim 20 \text{ cm}^{-1}$. b) Photoluminescence (PL) spectra for a representative 2D memtransistor shows peak at 1.83 eV corresponding to the direct bandgap of monolayer MoS_2 .

Extended Data Figure 3

Extended Data Figure 3. Three examples of representative stochastic bit-streams for the random variables A , X_1 , and X_2 , correlation coefficient (CC) values between A , X_1 , and X_2 , and stochastic bit-streams obtained at the output node, B for the 2-node BN. The measured and expected values for $P(B)$ are very similar confirming high precision hardware acceleration of BN.