Mimicking Biological Synaptic Functionality with an Indium Phosphide Synaptic Device on Silicon for Scalable Neuromorphic Computing

Debarghya Sarkar,† Jun Tao,† Wei Wang, Qingfeng Lin, Matthew Yeung, Chenhao Ren, and Rehan Kapadia*‡

Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, United States

ABSTRACT: Neuromorphic or “brain-like” computation is a leading candidate for efficient, fault-tolerant processing of large-scale data as well as real-time sensing and transduction of complex multivariate systems and networks such as self-driving vehicles or Internet of Things applications. In biology, the synapse serves as an active memory unit in the neural system and is the component responsible for learning and memory. Electronically emulating this element via a compact, scalable technology which can be integrated in a three-dimensional (3-D) architecture is critical for future implementations of neuromorphic processors. However, present day 3-D transistor implementations of synapses are typically based on low-mobility semiconductor channels or technologies that are not scalable. Here, we demonstrate a crystalline indium phosphide (InP)-based artificial synapse for spiking neural networks that exhibits elasticity, short-term plasticity, long-term plasticity, metaplasticity, and spike timing-dependent plasticity, emulating the critical behaviors exhibited by biological synapses. Critically, we show that this crystalline InP device can be directly integrated via back-end processing on a Si wafer using a SiO2 buffer without the need for a crystalline seed, enabling neuromorphic devices that can be implemented in a scalable and 3-D architecture. Specifically, the device is a crystalline InP channel field-effect transistor that interacts with neuron spikes by modulation of the population of filled traps in the MOS structure itself. Unlike other transistor-based implementations, we show that it is possible to mimic these biological functions without the use of external factors (e.g., surface adsorption of gas molecules) and without the need for the high electric fields necessary for traditional flash-based implementations. Finally, when exposed to neuronal spikes with a waveform similar to that observed in the brain, these devices exhibit the ability to learn without the need for any external potentiating/depressing circuits, mimicking the biological process of Hebbian learning.

KEYWORDS: synaptic device, neuromorphic computing, metaplasticity, spike timing-dependent plasticity, templated liquid-phase growth, indium phosphide, Si back-end processing

The arithmetic/logic unit (ALU) in any modern day central processing unit (CPU) is the fundamental building block responsible for accurate computation and logic operations. While the requirement of accuracy and precision in computation is necessary for many applications, a dramatic increase in the need to extract general relations from large sets of high-dimensional data has driven the use of machine learning and multidimensional optimization algorithms in fields such as computer vision, voice recognition, natural language processing, and autonomous systems. To obtain useful results from these approaches, it is typically necessary to train the algorithms on large clusters with large amounts of data, requiring hours or even days for commercially relevant applications. However, in this respect, the human brain supersedes any supercomputer, being 6–9 orders of magnitude more power efficient than today’s digital logic computers for these types of applications. A distinctive feature of the human brain is the closely intertwined nature of the memory and logic units which enables this efficient computing, unlike today’s computers where the CPU and memory units, though themselves being sufficiently fast, are largely restricted by data transport among them, commonly referred to as the von Neumann architecture bottleneck. To this end, researchers have started developing computing architectures that mimic the
functioning of the brain, referred to as neuromorphic computing. While the exact functioning of the brain is an area of active neuroscientific research, to a good approximation, it may be said that the neurons act as the logic units operating in a three-dimensional (3-D) multiconnected network supported by their end connections called synapses, acting as the memory units.1

Neural networks have been long studied and implemented in software,1 and some prototypes of hardware neuromorphic systems have already been built using existing CMOS technology.6–12 Neuronal spiking13–15 and synaptic behavior16–20 have been emulated using CMOS-based circuits containing 6–12 transistors depending on the specific functionality and robustness of the design. But an omnipresent usage would require hardware implementation with reduced footprint and energy consumption, and increased parallelism and interconnections, by several orders of magnitude.5

Architectures based on single devices acting as individual units of the neuromorphic computing block have been heavily studied.1–4 While there have been fewer reports of neuronal behavior emulation with a single device,17,22 emulation of synaptic activity with a single device has been widely demonstrated, pioneered by Diorio et al. in 1996 based on hot-electron injection and electron tunneling in a floating gate Si MOS transistor.23 Additionally, memristor-based synaptic devices from phase-change materials,24 ferroelectricity,25,26 ferrimagnetism,27 and nanionics28–30 have been demonstrated. More recently, time-dependent threshold voltage shift arising from charging and discharging of traps in the semiconductor,31 at the insulator-semiconductor interface,32,33 or ion migration in the gate dielectric34 in MOSFETs have been utilized for synaptic emulation.

However, due to the fundamental limitations on the growth of high-quality crystalline material on amorphous materials, current 3-D approaches often use low-mobility solution deposited materials,35,36 nanocrystalline channels,24,41 or physical transfer processes42,43 that are not scalable. Here, we demonstrate a scalable, back-end compatible, artificial synapse with a crystalline indium phosphide (InP) channel grown directly on Si/SiO2 wafer with templated liquid-phase (TLP) growth.47–50 We show these synaptic devices emulate a range of behaviors such as elasticity, short-term and long-term plasticity, metaplasticity, and spike timing-dependent plasticity (STDP).

RESULTS AND DISCUSSION

A synapse31 is a 20–40 nm junction between two neurons (schematically shown in Figure 1a) which aids in the communication of signals between neurons. These connections are not hard-coded but evolve over time and are plastic, depending on the activity of the connecting neurons. Based on the nature of pre- and postsynaptic activity, the synaptic weight may remain unchanged (elasticity), increase (potentiation), or decrease (depression). The change in the synaptic connectivity (also called synaptic plasticity) is regarded as a key ingredient in learning and memory. Synaptic plasticity leads to future data processing along the same pathway to be accordingly affected: A potentiated synapse would enable better transduction of signals from presynaptic action potentials, and a depressed synapse would lead to a constrained transduction of signals. The difference in postsynaptic current (PSC) determines the synaptic weight change that occurred owing to the coupling neuronal activity. The synaptic weight change is also time dependent, with a gradual decay over time. Based on the lifetime of the plasticity, two categories of plasticity are defined: short-term plasticity and long-term plasticity. While the exact times scales defining each is not precisely defined, short-term plasticity often refers to behavior on the milliseconds to seconds time frame, while long-term plasticity refers to behavior on the time frame of seconds to years.

The device architecture used in this work to emulate the synaptic behavior is schematically shown in Figure 1b. The device is a top-gated 100 nm-thick InP nanowire transistor fabricated on Si/SiO2 wafer with 60 nm Al2O3 as the gate dielectric. The Al2O3-InP stack may be regarded as the synapse connecting the cell membrane of the presynaptic neuron (gate electrode) to the cell membrane of the postsynaptic neuron (drain electrode). Presynaptic action potentials are applied as pulses to the gate electrode, which leads to a PSC in the drain electrode based on the strength of the synapse. The synaptic
strength is the conductance of the semiconductor channel. Application of a voltage pulse on the gate leads to charging or discharging of traps in the oxide or at the oxide/semiconductor interface. This in turn changes the electrostatics of the channel and leads to a shift in the threshold voltage, so that the drain current (PSC) at the same constant gate voltage before and after the gate voltage pulse may be different. The magnitude of the hysteresis depends on the pulse amplitude and width. The amplitude determines the electric field within the oxide and the population of charge carriers in the semiconductor, with higher amplitudes accessing higher energy traps. The pulse width determines the length of time under which the oxide is under tension, changing the number of traps that are charged or discharged. The synaptic strength evolves as a function of the nature of the pulse and the strength of the synapse before the application of the pulse.

Here we demonstrate the implementation of these artificial synapses using a scalable platform, which enables arrays of crystalline InP devices to be grown on Si/SiO₂ wafers with near unity yield. To carry this out, we utilize TLP growth. Briefly, an array of 100 nm-thick In nanowires and capping SiO₂ layer is defined on the Si/SiO₂ substrate using photolithography, evaporation, and lift-off processes, schematically shown in Figure 2a–b. A thin 4–6 nm layer of molybdenum oxide (MoOₓ) is deposited below the indium to aid in wetting of In. At the growth temperature, the indium nanowires turn to liquid but maintain their geometry due to the capping layer. In the presence of phosphine and hydrogen, InP precipitates out of the In liquid, leading to an array of InP nanowires directly grown on an Si/SiO₂ substrate, as shown in Figure 2c. After the nanowires are grown, we fabricate device arrays, as shown schematically in Figure 2d. Using lithography, we carry out source–drain metallization with Ge-Au-Ni contacts, followed by a 380 °C rapid thermal annealing (RTA). 60 nm of Al₂O₃ is deposited using atomic layer deposition (ALD) as the gate dielectric, and Ni is used as the gate metal. Details of the TLP growth of InP field-effect transistor (FET) channels and device fabrication are discussed in the Methods section. The output

Figure 2. (a–d) Schematic of TLP growth of crystalline InP nanowire arrays on Si/SiO₂ and subsequent fabrication of nanowire FETs.

Figure 3. Spike amplitude-dependent plasticity. (a) Hysteretic transfer characteristics showing depression for +5 V and potentiation for −5 V. (b) Hysteretic transfer characteristics showing elasticity for +0.7 V and potentiation for −0.7 V. (c) Transient PSC before and after application of presynaptic pulse leading to elasticity (0.1 V), potentiation (−5 V), and depression (+5 V). (d) Short-term and long-term synaptic weight change for different values of presynaptic voltage pulse. All curves reported with Vᵣᵣ = 3 V.
Characteristics of a representative 2 × 4 array of devices are shown in Figure S1a–h, with an average extracted effective mobility of ∼200 cm²/V·s at 3 V gate voltage overdrive with peak effective mobilities of ∼500 cm²/V·s. This is more than 10 times higher than most CNT network devices, which exhibit peak mobilities of ∼5–20 cm²/V·s. The top row of devices all have channel lengths of L = 10 μm, while the bottom row all have channel lengths of L = 25 μm. Critically, we demonstrate the ability to deterministically place and grow devices on an amorphous substrate with high device yield.

The effect of amplitude of gate pulse in changing the electrostatics of the channel and thus the behavior of the artificial synapse is shown in Figure 3. We interpret the amplitude of gate pulses to be proportional to the probability of exocytosis for neurotransmitter release. Thus, a higher amplitude pulse signifies higher probability of release of neurotransmitters than that of a lower amplitude pulse. Physically, a higher gate voltage pulse causes higher energy traps to be accessed. Figure 3a–b shows the transfer characteristics of the synaptic device at 3 V source–drain voltage for different peak gate voltages (5 and 0.7 V, respectively), starting from 0 V. It is found that a peak gate voltage of +5 V increases the threshold voltage, lowering I_d (Figure 3a), while a peak gate voltage of −5 V decreases the threshold voltage, increasing I_d. From this, we show that positive gate voltage pulses lead to depression, a reduction in the PSC, while negative voltage pulses lead to potentiation. However, as shown in Figure 3b, it is seen that a peak voltage of +0.7 V has nearly no effect on the PSC, leading to elastic behavior of the synapses, where a pulse is passed, however no synaptic weight change occurs. To show the dynamics of the synapse, representative PSC curves are plotted in Figure 3c. The current for t < 0 is the PSC before arrival of the spikes, for t > 0 is the PSC after arrival of the spikes, and the dotted line indicates the time of spike arrival. Here, we use a rectangular pulse train of 40 pulses of width 500 μs and 50% duty cycle with 0 V baseline. The three behaviors were obtained by the application of pulses with varying amplitudes. A pulse amplitude of 0.1 V results in elasticity, −5 V gives potentiation, and +5 V gives depression. We also plot the synaptic weight change for different peak presynaptic potentials in Figure 3d.

We define the short-term plasticity (STP) as the difference between the average PSC over the first 1 s after pulse, and that before pulse, normalized to the PSC before pulse. The long-term plasticity (LTP) is defined as the difference between the normalized average PSC between 10 and 40 s after pulse, and that before pulse. The magnitude of plasticity increases with increasing voltage amplitude and can be modeled with exponential functions of the form:

$$\Delta w = A \times e^{V_g/V_a}$$

where $V_g$ is the peak gate voltage, and $V_a$ is the activation voltage for the traps. The physical implication of a nonzero plasticity is that a future signal would be transduced with the modified weight of the synapse. This leads to selectivity of pathways for transduction of a signal to converge to an optimal solution in a neural network.

It may be noted that the synaptic weight change at time $t$ is actually a function of the synaptic weight at $t$ and the neuronal activity at that time. Mathematically, it may be represented as

$$\Delta w_t = f(w_t, \theta_t)$$

where $w_t$ is the synaptic weight, and $\theta_t$ is the neuronal activity at time $t$.

Thus, for the same neuronal activity, the synaptic weight change would depend on the initial condition of the synapse. In other words, for the same gate voltage pulse(s), the change in the drain current (arising from the change in the electrostatics of the channel) would depend on the initial electrostatic condition of the channel and the threshold voltage of the device. This phenomenon is referred to as metaplasticity and is one of the most important characteristics of biological synapses. It allows a highly nonlinear optimization of the synaptic weight based on the sequence of signals it transduces, to in turn give rise to an optimal selectivity of the signals being transduced. Metaplasticity of our synaptic device was emulated by applying the same gate voltage pulse but preceded by different “priming” gate voltage pulses to emulate prior brain activity. The gate voltage pulse trains consisted of 10 pulses of 5 ms width and 6 ms period. As a demonstration, we chose +2.5 V pulse as a depressing priming pulse and −2.5 V pulse as a potentiating priming pulse. The PSC plots for two representative potentiating and depressing cases are depicted in Figure 4a–b. The same −4.5 V pulse train gives a significantly higher relative change of PSC when it follows depressing priming than when it succeeds a potentiating priming. Similarly, a depressing priming causes a smaller depression for a + 4.5 V pulse train compared to a potentiating priming.
priming. Figure 4c shows the short-term synaptic weight change extracted from the PSCs for gate voltages ranging between $-5$ V and $+5$ V, for three different initial conditions of no priming, depressing priming, and potentiating priming. It is observed that a reduction in potentiation weight change occurs for an initially potentiated synapse, while an increased potentiation occurs for an initially depressed synapse. Similarly, an initially depressed synapse leads to having less depression than an initially potentiated synapse. As may be realized, the feedback mechanism embedded in metaplasticity allows the synapse to optimally tune its weight such that it regressively stays within bounds instead of diverging toward either extreme. For the same initial state of the synapse, the synaptic weight change is also dependent on the number of neurotransmitters being transmitted simultaneously. The number of neurotransmitters is in turn dependent on the number of action potentials that have reached the presynaptic terminal in quick succession (each action potential typically releasing same number of neurotransmitters). To emulate the pulse number-dependent plasticity of synapses, we excite the synaptic device with $+5$ V peak (for depression) or $-5$ V peak (for potentiation) and 0 V baseline pulse trains of 5 ms width and 5.5 ms period, consisting of varying number of pulses, the results of which are shown in Figure 5. Some representative potentiation and depression curves (for pulse trains consisting of 1, 20, and 100 pulses) are plotted in Figure 5a. A general trend of increased potentiation or depression is observed as we increase the number of incident action potentials. The PSC in each case was fitted with a double exponential function given by

$$PSC = I_{ss} + 0.5 \times (I_{t=0} - I_{ss}) \times (e^{-t/c} + e^{-t/d})$$

The PSC approaches the steady-state value $I_{ss}$ from $I_{t=0}$ at a rate given by the double exponentials: One of them is used to fit the very fast decay right after the pulse, while the other gives the relatively slower decay rate until the PSC virtually reaches the steady-state value. The PSC decay time constant (inverse of the slower decay rate), referred to in the graph as the synaptic relaxation time constant, is plotted in Figure 5b. It is also seen that the relaxation rate initially decreases and gradually saturates with increase in the number of action potentials stimulating the synapse.

Figure 5c,d, respectively shows the long-term and short-term synaptic weight change, where the weight change definitions are the same as earlier. As expected, a gradual increase in the plasticity of the synapse is observed with increasing number of action potentials, as more number of traps are populated, until gradually all or most of the traps that can be affected by the pulse amplitude are influenced, so that the plasticity gradually saturates. The same reasoning may be applied for the trend observed for the relaxation rate as well. Considering that faster traps are populated earlier than slower traps, with increasing number of pulses, we gradually excite higher number of slower traps, increasing the decay time constant.

Once an action potential is generated in the postsynaptic neuron, it travels down the length of the axon. However, a fraction of this signal is reflected and impinges on the output terminal of the synapse, potentially changing the synaptic strength. Synaptic modification arising from the correlated activity of the presynaptic and postsynaptic neuron was first introduced by Hebb in 1949 and later refined by other.
researchers such as Bienenstock–Cooper–Munro in 1982. One of the most well-known versions of Hebbian (or BCM) learning is STDP, some of the first demonstrations of which were in classic neuroscience experiments such as by Markram et al., Bi and Poo, and modeled by van Rossum et al. STDP has been emulated by the neuromorphic device community over the years, mostly based on memristor technology. However, STDP in MOSFET-based synapses has not clearly been demonstrated using solely time-correlated activity of the pre- and postsynaptic neuron. Instead, due to the nature of the pulse waveforms used and the behavior of MOSFETs, it is necessary to use differing pulses for depression and potentiation, which would require external circuitry to enable learning.

While many different STDP behaviors exist, a commonly implemented STDP behavior is the potentiation of the synapse when the postsynaptic action potential succeeds the reflected presynaptic action potential, and depression when the reflected postsynaptic action potential precedes the presynaptic action potential. The highest magnitude of potentiation and depression occurs for the shortest time gap ($\Delta t$) between the pre- and postsynaptic actions, and the synapse tends to stay elastic for increasing time spacing between the pre- and postsynaptic action potentials. Such an asymmetric and time-correlated behavior may well be reasoned based on causality principles. Here we utilize engineering of hysteresis in a family of transfer characteristics to demonstrate STDP in our synaptic device. The presynaptic action potential is designed as a pulse that goes from 0 to $+2.75 \text{ V}$ to $-5 \text{ V}$ with a total width of 10 ms. The reflected postsynaptic action potential is designed as a pulse with a baseline of 1.5 V, ramping up to 2.1 V in 100 ms, ramping down to 0.9 V in 20 ms, and again ramping up to 1.5 V in 100 ms. The peak values of the presynaptic potential were chosen such that it results in elasticity at the 1.5 V baseline postsynaptic potential (Figure S2b), which is the boundary condition for high $\Delta t$. For postsynaptic potentials $<1.5 \text{ V}$, the same presynaptic pulse gives depression, with increasing plasticity for lower potentials (representative hysteresis shown in Figure S2a). For postsynaptic potentials $>1.5 \text{ V}$, the presynaptic pulse gives potentiation, with increasing plasticity for higher potentials (representative hysteresis shown in Figure S2c). This effect has been utilized to accordingly design the shape of the postsynaptic potential. The ramp times have been chosen to closely match that found in STDP of typical synapses. The upper panel of Figure 6a shows the waveform simulating the reflected postsynaptic action potential, while the lower panel depicts the presynaptic pulse. The relative positioning of the two pulses can be directly interpreted as positive and negative $\Delta t$ as has been shown in Figure 6a. The long-term synaptic weight change for different values of time offset between the pre- and postsynaptic action potentials.

**CONCLUSION**

In summary, we demonstrated the scalable fabrication of an array of nanowire FETs integrated on Si/SiO$_2$ wafer by direct growth of InP nanowire channels and reported their synaptic functionality for spiking neural network-based neuromorphic computation. Modeling the gate electrode as the presynaptic axon terminal, the drain electrode as the postsynaptic dendrite, and the gate oxide-semiconductor channel as the synapse, we have controlled the charging and discharging of interfacial traps in the MOS structure to modulate the channel conductance interpreted as the synaptic weight. Nonlinear synaptic weight change dependent on pre- and postsynaptic neuronal activity, such as elasticity, short- and long-term plasticity, metaplasticity, spike number-dependent plasticity, and STDP, has been emulated by engineering the hysteresis of the channel conductance. Future work may involve realization of a single device spiking neuron and integration of the synapse and neuron arrays to form an on-chip compact neuromorphic circuit. Further innovation in the synaptic device architecture may also be done to expand the synaptic behavioral space such as to include selectivity and responsivity of synapses to different pre- and postsynaptic activity.

**METHODS**

**Templated Liquid-Phase Growth.** The Si/SiO$_2$ substrate was thoroughly cleaned with acetone, isopropyl alcohol, and deionized...
water and dried with dry nitrogen. FET channel templates were photolithographically patterned on the substrate.

4–6 nm MoOx followed by 100 nm indium (99.9999%) was thermally evaporated using alumina covered tungsten boats, on the patterned samples connected to a cryo-cooled (liquid nitrogen) stage. 100 nm SiOx was electron-beam evaporated using SiOx pellets (99.99%), as the capping layer. FET templates of MoOx/In/SiOx were then obtained using standard lift-off procedure.

TLP growth of InP was done in a single-zone hot-wall tube furnace at furnace temperatures in the range 550–575 °C. 99.9999% PH3, diluted ex situ with 99.999% H2 to achieve the desired PH3 partial pressure, was used as the P precursor.

**Device Fabrication.** The capping SiOx layer on TLP grown InP was etched in 1:10 hydrofluoric acid (HF, 48–50%) dissolved in water. Source–drain pads were then patterned using photolithography. Six nm Ge, 20 nm Au, and 80 nm Ni were electron-beam evaporated, followed by lift-off. A 380 °C RTA was done to reduce contact resistance by Ge alloying with InP at the source–drain contact regions. 60 nm of Al2O3 was then deposited using ALD as the gate dielectric. After that, gate electrodes were patterned using photolithography, followed by 80 nm Ni electron-beam evaporation and lift-off. Source–drain contact vias were then opened by photolithography and etching the Al2O3 on the contact pads.

**Device Characterization.** The synaptic devices were excited with pulses generated from a Tektronix AWG 2021, and a time-dependent measurement was done with an Agilent B1500A semiconductor device analyzer.

**ASSOCIATED CONTENT**

- Supporting Information
- Output characteristics of an array of transistors are shown in Figure S1; hysteresis engineering of transfer characteristics showing depression, elasticity, and potentiation for different Vgs is shown in Figure S2 (PDF)

**AUTHOR INFORMATION**

**Corresponding Author**

*E-mail: rkapadia@usc.edu.

**ORCID**

Debarghya Sarkar: 0000-0002-5411-7066
Wei Wang: 0000-0001-7007-0088
Rehan Kapadia: 0000-0002-7611-0551

**Author Contributions**

†These authors contributed equally.

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

R.K. acknowledges funding from the National Science Foundation (award no. 1610604). D.S. thanks the support by the USC Annenberg Graduate Fellowship. J.T. thanks the support by the USC Provost Graduate Fellowship. The authors gratefully acknowledge the use of the ALD facility in Dr. Stephen Cronin’s lab.

**REFERENCES**

(52) Cao, C.; Andrews, J. B.; Kumar, A.; Franklin, A. D. Improving Contact Interfaces in Fully Printed Carbon Nanotube Thin-Film Transistors. ACS Nano 2016, 10, S221–S229.