Physical Compact Model for Three-Terminal SONOS Synaptic Circuit Element

Su-in Yi, A. Alec Talin, Matthew J. Marinella, and R. Stanley Williams*

1. Introduction

Improvements in computing performance and efficiency recently opened up opportunities\textsuperscript{[1]} in artificial intelligence (AI), represented by the seminal work of Alex Krizhevsky et al.\textsuperscript{[2]} in 2012 where graphic processing units (GPUs) combined with big data excelled on an image classification task (top-5 error rate of 15.3% achieved by artificial neural networks vs 26.2% from the runner-up algorithms). This achievement reignited the attention for deep neural networks dating back to the 1980s,\textsuperscript{[3,4]} in particular for convolutional neural networks (CNNs). Further efforts to optimize computing hardware are being actively pursued, such as using field-programmable gate arrays (FPGAs) and application-specific integrated circuits (ASICs),\textsuperscript{[5,6]} including the tensor processing units (TPUs)\textsuperscript{[7,8]} developed by Google and system on a chip (SoC)-scale AI accelerators used in mobile phones. This recent progress in computing hardware is still bound to CMOS technology, where the von Neumann bottleneck\textsuperscript{[9]} limits speed and power efficiency. In-memory analog computing based on Ohm’s and Kirchhoff’s laws\textsuperscript{[10–12]} utilizing nonvolatile memories (NVMs), such as resistive random access memory (RRAM),\textsuperscript{[13–15]} phase-change memory (PCM),\textsuperscript{[16,17]} and flash memory,\textsuperscript{[18–21]} is a promising route to eliminate the von Neumann bottleneck and implement neuromorphic computing circuits for future AI systems.\textsuperscript{[22–26]}

Here we present a dynamical compact model for three-terminal silicon–oxide–nitride–oxide–silicon (SONOS) synaptic circuit elements. Three-terminal memory devices such as floating gate (FG) or SONOS flash memories have a long history for manufacturable data storage applications, but their compact models have relied on static metal–oxide–semiconductor field-effect transistor (MOSFET) behavior.\textsuperscript{[27–29]} Moreover, a significant number of three-terminal synaptic devices are now being reported with various material systems,\textsuperscript{[30–32]} but most publications describe experimental characterization without supplying a compact model.\textsuperscript{[33]} So there exists a significant gap between the device community and circuit designers. Our work addresses this gap by constructing a well-posed compact model for three-terminal synaptic circuit elements,\textsuperscript{[34–36]} beginning with the technologically mature SONOS device, which is a staple for NAND flash memory products.\textsuperscript{[37–40]} In this work, we utilized technology computer-aided design (TCAD) physics-based calculations within Synopsys Sentaurus to simulate the physics of a device and identified a key state variable $Q_M$, the amount of charge in the SONOS trap layer, to guide us in constructing a compact model. Subsequently, we validated the model through a circuit simulation using Cadence Spectre to compare with the experimentally measured behavior of source–drain current after applying a wide range of voltage pulse amplitudes on the gate of a SONOS device. The use of both physics-based simulations and experimental data to identify the state variables and calibrate the model was crucial, as the critical state variable $Q_M$ is extremely difficult to measure experimentally (see the Video S1, Supporting Information).

A well-posed physics-based compact model for a three-terminal silicon–oxide–nitride–oxide–silicon (SONOS) synaptic circuit element is presented for use by neuromorphic circuit/system engineers. Based on technology computer-aided design (TCAD) simulations of a SONOS device, the model contains a nonvolatile memristor with the state variable $Q_M$ representing the memristor charge under the gate of the three-terminal element. By incorporating the exponential dependence of the memristance on $Q_M$ and the applied bias $V$ for the gate, the compact model agrees quantitatively with the results from TCAD simulations as well as experimental measurements for the drain current. The compact model is implemented through VerilogA in the circuit simulation package Cadence Spectre and reproduces the experimental training behavior for the source–drain conductance of a SONOS device after applying writing pulses ranging from $-12 \, \text{V}$ to $+11 \, \text{V}$, with an accuracy higher than 90%.

S. Yi, R. S. Williams
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX 77843, USA
E-mail: rstanleywilliams@tamu.edu

A. A. Talin
Department of Materials Physics
Sandia National Laboratories
Livermore, CA 94550, USA

M. J. Marinella
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ 85281, USA

© 2022 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH.
Figure 1. a) The equivalent circuit diagram for a three-terminal SONOS synaptic circuit element based on physical state variables $Q_M$ and $Q$. The state variable $Q_M$ is the total amount of charge that has passed through an effective memristor $M$, equal to the charge on the series capacitor $C_M$ (the branch on the left), and controls the slow dynamics. The memristance is a function of both the state variable $Q_M$ and the applied voltage, $V_{GS} - V_T - v_D$, with strong nonlinearity. The parallel branch on the right, composed of a series resistor $(R)$ and capacitor $(C)$, is responsible for the short-term dynamics with a state variable $Q$, which represents the typical switching behavior of a MOSFET. For simplicity, the capacitor $C$ is described as a linear capacitor, which holds when $V_{GS} - V_T > 0$, where the influence of charge depletion is negligible. $R$ describes an effective resistance including the contact resistance of the gate electrode, the scattering by the ionized impurities in the depletion region, etc. The conductances and currents through the channel of the SONOS device are represented by variable resistors with identical resistances $R_{CH}$ and depend on the two state variables $Q$ and $Q_M$. b) Schematic 2D cross section perpendicular to the gate of the SONOS device used in the TCAD simulations and experiments (the study by Agarwal et al. [33]), with the doping concentration shown in different colors. Inset shows three exemplary programmed states with colors indicating the trapped electron concentration in the Si$_3$N$_4$ layer.

Figure 1a depicts the equivalent circuit diagram for the well-formed compact model of a three-terminal SONOS synaptic circuit element$^{[34-36]}$ for neuromorphic applications, which consists of three components. The first is found on the left vertical branch, composed of a series capacitor $C_M$ and an extended memristor$^{[41]} M$, defined as

$$M = f(Q_M, V)$$

(1a)

$$\frac{dQ_M}{dt} = I_M = g(Q_M, V)$$

(1b)

where we are not directly interested in the current $I_M$ through the memristor, which is in general too small to be measured experimentally, but rather how the memristor charge controls the long-term SONOS dynamics for synaptic weight modulation or NVM through time dependence

$$Q_M = C_M(V_G - V_T - v_D)(1 - e^{-t/(MC_M)})$$

(2)

which is discussed further. The second vertical branch on the right has a series resistor $R$ and a capacitor $C$, which control the short-term dynamics represented as

$$Q = C(V_G - V_T - v_D)(1 - e^{-t/(RC)})$$

(3)

and in turn the drain current, $i_D$ is

$$i_D = \mu I_D \frac{1}{L} Q$$

(4)

Equation (4) is essentially quasistatic$^{[42,43]}$ without an explicit time dependence

$$i_D = \mu I_D \frac{1}{L} C(V_G - V_T - v_D) = \mu I_D \frac{W}{L} C_M(V_G - V_T - v_D)$$

(5)

due to the small time constant $\tau = RC \ll 1 \text{ns}$. We performed circuit simulations based on a SONOS device with $W = 1.2 \mu m$, $L = 7 \mu m$, $\mu = 350 \text{ cm}^2 \text{V}^{-1} \text{s}^{-1}$, and $C = 26 \text{fF}$ as shown in Figure 1b combined with $R = 10^8 \Omega$, chosen to manage the simulation speed with a time constant $\tau = 2.6 \mu \text{s}$, which permits the use of a timestep as large as 1 ns in Cadence Spectre to examine both short- and long-term dynamics. For most practical circuit simulations, the short-term dynamics can be safely eliminated by making $Q$ a parameter instead of a state variable$^{[18]}$ as discussed in Section S2, Supporting Information. The last component comprises two identical variable resistors, derived from $v_D/I_D$ as

$$R_{CH} = \left[\frac{1}{2\tau Q} \left( Q - C_M \right) \right]^{-1}$$

(6)

where $\chi$ stands for the percentage of $Q_M$ contributed by the silicon channel. The memristor conductance is a function of the $Q_M$ history as described later.

We performed TCAD simulations to determine the quantitative dynamics of the state variables $Q_M$ and $Q$, representing charge transport via tunneling and electron drift diffusion$^{[44]}$ because they are not easily experimentally accessible. A memristor circuit element was necessary to describe the dynamics of $Q_M$, whereas a linear $R-C$ circuit was suitable for $Q$. The results in Figure 2 provided the information used to derive the compact model in Figure 1. To focus on the dynamics of synaptic weight modulation, the drain and the source were grounded and denoted as “Si,” which eliminated the channel current. A bias of $V = 8 \text{ V}$ was applied to the gate at $t = 0$, with Figure 2 representing two times after bias application, $t = 1 \mu \text{s}$ and $t = 200 \mu \text{ms}$. The resistor in the short-term branch $(R)$ in Figure 1 was not included because $C$ was fully charged within 1 $\mu \text{s}$ in the TCAD simulations. The amount of charge under 8 $\text{ V}$ bias at $t = 1 \mu \text{s}$ (left, Figure 1 (a) working) was $Q = 17.5 \times 10^{-12} \text{ e} \text{ cm}^{-2}$, corresponding to $2.8 \mu \text{ C cm}^{-2}$, which shows good agreement with the expected value of $2.5 \mu \text{ C cm}^{-2}$ calculated from 8 $\text{ V}$ on a capacitor with $0.31 \mu \text{F cm}^{-2}$ (3 nm SiO$_2$, 8 nm Si$_3$N$_4$, and 4 nm SiO$_2$) visualized schematically in Figure 2b. The area density of
electrons was used as the unit for $Q$ and $Q_M$ for more intuitive analysis in this article. A slightly larger amount of charge compared with the expected capacitor value is attributed to the negative threshold voltage $V_T = -0.01$ V and nonlinear $Q-V$ curve near the depletion region ($V_C \approx V_T$) of a MOSFET. The corresponding band diagram across the material stack is shown in Figure 2c. The trap layer ($\text{Si}_3\text{N}_4$) remains electrostatically neutral until $t = 1$ ms with $Q_{M} = 0$ and provides a tunneling barrier thickness of 3 nm due to the 9 eV forbidden gap of $\text{SiO}_2$. After allowing enough time for tunneling, at $t = 200$ ms (Figure 2 right column), $Q_M = 5.8 \times 10^{12}$ e$^{-} \text{cm}^{-2}$, and $Q_{C} = 20 \times 10^{12}$ h$^{+} \text{cm}^{-2}$, hence, $Q_{M} - Q_{C} = 14.2 \times 10^{12}$ e$^{-} \text{cm}^{-2}$ because of charge neutrality ($Q_{C} = Q_{M} + Q_{\chi} M$). Consequently, the synaptic weight of the SONOS decreased based on the net charge changing from $17.5 \times 10^{12}$ e$^{-} \text{cm}^{-2}$ at $t = 1$ ms to $14.2 \times 10^{12}$ e$^{-} \text{cm}^{-2}$ at $t = 200$ ms. In other words, $57\%$ of the increased $Q_M (0 \rightarrow 5.8 \times 10^{12}$ e$^{-} \text{cm}^{-2}$) is contributed by the electron charge on $C (= Q_{M} - Q_{C} = 17.5 \times 10^{12}$ e$^{-} \text{cm}^{-2} \rightarrow 14.2 \times 10^{12}$ e$^{-} \text{cm}^{-2}$), realizing the synaptic weight depression. Although the tunneled charge is $5.8 \times 10^{12}$ e$^{-} \text{cm}^{-2}$, intuitively implying $\Delta (Q_{C} - Q_{M}) = -5.8 \times 10^{12}$ e$^{-} \text{cm}^{-2}$, the loss of charge on $C$ is spontaneously complemented by the gate electrode ($Q_{C} = 17.5 \times 10^{12}$ h$^{+} \text{cm}^{-2} \rightarrow 20 \times 10^{12}$ h$^{+} \text{cm}^{-2}$) due to the electrostatic force from the trapped charge located at the middle of dielectric stack (O–N–O). The Supplementary Video provides a dynamic circuit illustration of this mechanism and the relevant discussion on $\chi$ follows. The underlying physics of the memristor in our compact model is revealed in the band diagram at $t = 200$ ms (right panel of Figure 2c). The slightly lifted conduction band edge (or electrostatic potential) due to the injected negative space charge (trapped electrons) alters the tunneling thickness so that an electron, from the conduction band edge of silicon interfaced with the $\text{SiO}_2$ tunneling oxide, encounters a thicker tunneling barrier of 5 nm compared with 3 nm at $t = 1$ ms when $Q_M = 0$. Consequently,
the tunneling current decreases, that is, \( \frac{dQ_M}{dt}_{Q_M = 5.8 \times 10^{12}} < \frac{dQ_M}{dt}_{Q_M = 0} \). Figure 2 presents results at two discrete times, but the memristance changed smoothly and continuously with the state variable \( Q_M \).

Figure 3a,b shows the simulation results of \( Q_M \) as a function of time under constant biases: \(+9\), \(+8\), and \(+7\) V in (a) and \(-8\), \(-7\), \(-6\), and \(-5\) V in (b). The dynamics are qualitatively reminiscent of an \( R-C \) circuit, where the initial current, \( i_{\text{in}} \), is equal to \( V/R \) and \( Q \) increases inverse exponentially until it is saturated at \( Q = CV \) but is quantitatively different. For example, the two curves with \(+9\) and \(-8\) V in Figure 3a show that \( +9/|\frac{dQ_M}{dt}|_{t=0} \) from the former is not identical to \( -8/|\frac{dQ_M}{dt}|_{t=0} \) from the latter, immediately identified by the slope of the curves at \( t = 0 \). Therefore, the resistive component in series with the capacitor has a voltage dependence. Second, the slope of the tangent decreases exponentially (not linearly) with \( Q_M \), which is evident from the replotted curves on a log scale in Figure 3c, demonstrating that the resistance is also a function of \( Q_M \), as forecast by Equation (1). The \( y \)-axis represents \( Q_{M,\text{max}}/Q_M \) for \( V > 0 \) and \( Q_M \) for \( V < 0 \), because \( \ln(CV - Q) + D \) for \( V > 0 \) and \( \ln(Q - CV) + D \) for \( V < 0 \), where \( D \) is a constant (see Equation (S8), Supporting Information). For a usual \( R-C \) circuit, \( \ln(Q - CV) \) for \( V < 0 \) (or \( \ln(CV - Q) \) for \( V > 0 \)) is a linear function of time with a slope \( 1/\tau \), where \( \tau \) is the time constant.

Figure 3. Physics model (TCAD) simulation results of a SONOS device and corresponding compact model. (a and b, positive and negative gate biases, respectively). In (a), the state variable \( Q_M \) increases with time and saturates at \( Q_M = Q_{M,\text{max}} = 8 \times 10^{12} \text{ e}^{-}\text{cm}^{-2} \), which is defined by the trap concentration \( (10^{19} \text{ cm}^{-3}) \) chosen for the 8 nm-thick Si\(_3\)N\(_4\) layer. c) Logarithmic-scale plots of \( Q_M \) for \( V > 0 \) (or \( Q_{M,\text{max}} - Q_M \) for \( V < 0 \)), conveying the identical information of (a) and (b), but emphasizing the nonlinearity of the response. The decaying slopes \( = -1/\tau \) with respect to time show that the characteristic time is not constant. d,e) DRMs extracted from (a) and (b), respectively. The nearly linear relations between \( \log(|\frac{dQ_M}{dt}|) \) and \( Q_M \) demonstrate that the memristance depends exponentially on \( Q_M \). The compact model linear approximations to the DRMs are denoted with dotted lines. The eccentric behaviors observed for \( 9 \) V with \( Q_M > 5 \times 10^{12} \text{ e}^{-}\text{cm}^{-2} \) and \(-8 \) V with \( Q_M < 10^{12} \text{ e}^{-}\text{cm}^{-2} \) are due to \( Q_M \) approaching \( Q_{M,\text{max}} = 8 \times 10^{12} \text{ e}^{-}\text{cm}^{-2} \) and \( Q_{M,\text{min}} = 0 \), respectively. f) Memristance as a function of voltage. Upward and downward triangles represent the memristance values at two extremes, \( Q_M = 0 \) and \( Q_M = 8 \times 10^{12} \text{ e}^{-}\text{cm}^{-2} \), respectively, from (d and e). The overall trend of the memristance \( M \) versus \( Q_M \) and \( V \) exhibits the dependence on the state variable (governed by a parameter \( \gamma \)) and a strong nonlinearity with \( V \) (governed by a parameter \( \beta \)) in Equation (8).
The memristor characteristics for an applied sinusoidal voltage. a) Current versus time (blue: \(i_{M}\) and red: \(i_{C0}\)) as a result of an AC voltage of 8 V and 1 Hz. The dashed line is \(i_{M}\) from the compact model of Equation (7) and (8), whereas the solid lines are from the physics simulation of Synopsys Sentaurus. b) Current–voltage-pinched hysteresis curve of the memristor, \(i_{M}\), and the experimentally observable long-term gate current, \(i_{C0}\), for the same conditions as (a). The trajectories of the hysteresis curve are always clockwise. c) Magnified view of the pinched hysteresis curve of the memristor near the zero crossing. d) Magnified view of the avoided crossing of a series memristor and capacitor near 0,0.

\[
d\frac{d Q_M}{dr} = -\frac{1}{M(Q_M, V)C_M} Q_M + \frac{V}{M(Q_M, V)} \tag{7}
\]

This equation breaks down when \(Q_M\) approaches the extrema at 0 or \(Q_{M,max}\), which both require extreme times to reach. As long as the SONOS operation range is limited to intermediate states with moderate values of \(Q_M\), the compact model featuring the exponential correlation between \(M\) and \(Q_M\) can emulate the behavior with only small errors. The compact model calibrated to an experimental SONOS device showed that \(Q_M\) was bound between 20 and 60 fC even after many pulses of various magnitudes, because \(d Q_M/dr\) changes exponentially with \(Q_M\). The fitted memristance values at \(Q_M = 0\) and \(Q_M = Q_{M,max} = 8 \times 10^{-12} \times e^{-cm^{-2}}\) are marked as upward and downward triangles, respectively, in Figure 3f for seven different bias points performed in our physics simulation. Seven sets of memristance values as a function of \(Q_M\) were interpolated and extrapolated to establish the following analytic equation.

\[
M(Q_M, V) = \alpha \times \beta^{\alpha-\mid V\mid} \times \gamma^{\alpha-\mid V\mid} \times (V+\mid V\mid)^{-n} \tag{8}
\]

where \(\alpha = 50, \beta = 70, \gamma = 50, \delta = 25,\) and \(n = 0\) for \(V > 0\) and 1 for \(V < 0\). The dependences on \(V\) and \(Q_M\) are described by \(\beta\) and \(\gamma\), respectively. In addition, the sensitivity of the memristance on \(Q_M\) (slopes of the dotted lines in Figure 3d,e) also exponentially increases with \(V\), which is reflected by the fitting parameter \(\delta\). Although the curves from the analytic equation denoted by lines produce some errors compared with the physics simulation results (symbols), it can account for all possible combinations of \(Q_M\) and \(V\) so that it can stand alone as a circuit element. A discontinuity in \(M\) between \(V = 0-\) and \(V = 0+\) is observed when \(Q_M = Q_{M,max}\) for which the physical mechanism is attributed to a change in the charge tunneling dynamics. While a larger \(Q_M\) provides a thicker tunneling barrier in the case of trapping into Si3N4 (\(V > 0\)), it fosters detrapping when \(V < 0\) because the electrostatic potential of the trap is higher compared with when \(Q_M = 0\).
The compact model and physics simulation present good agreement, although they have a quantitative mismatch owing to the simple analytical form of Equation (8) that can be improved at the cost of complexity. The red solid curves in Figure 4a,b represent the long-term gate current, $I_{G,0}$, which is the sole way to deduce $I_M$ through experimental measurements. The shape of $I_{G,0}$ is almost identical to $I_M$, but with different magnitudes, approximately half of $I_M$, which is consistent with the illustration in the Supplementary Video and justifies the parameter $\chi$ ranging from 0 to 1. The detailed procedure for extracting the long-term gate current ($I_{G,0}$) from the total gate current ($I_{G,0} + I_{G,\infty}$) is available in Figure S2, Supporting Information. The memristor has both nonlinearity and asymmetry in its $I$–$V$ characteristic, which leads to a “history erase effect” or loss of initial condition, as shown in simulations of our compact model under sinusoidal voltage (Figure S3, Supporting Information).

We next calibrated the compact model extracted from TCAD simulations using experimental data from a SONOS device.[33] As an accurate compact model requires dynamical information from the target, we designed a protocol to extract the state variable from measurements that can actually be performed,[46] as shown in Figure 5, for a general three-terminal synaptic device.[30–32,47,48] For SONOS, the measured channel currents were converted to the change in threshold voltage ($\Delta V_T$) so that the desired state variable $Q_M$ could be obtained. Figure 5a shows the channel current of a SONOS device potentiated by $-12$ V and suppressed by $+11$ V from the work of Agarwal et al.[33] Utilizing the rigid shape of $i_D$–$V_G$ regardless of the shift in $V_T$, $\Delta V_T$ can be deduced from the current based on the measured $i_D$–$V_G$ curve at an arbitrary state, as shown in Figure 5b, to obtain Figure 5c.

While a simple correlation between the amount of fixed charge ($\Delta Q_{NOT}$) at the interface of the silicon channel and SiO$_2$ and the shift in threshold voltage ($\Delta V_T$) holds from $\Delta V_T = \Delta Q_{NOT}/C$ for a MOSFET, when it comes to a SONOS device with $\Delta Q_M$ (most charges exist in Si$_3$N$_4$) instead of $\Delta Q_{NOT}$, the relation becomes $\Delta V_T < \Delta Q_M/C$.[19,49] Due to electrostatic equilibration, the trapped charge closer to the channel results in the larger $\Delta V_T$, mainly dictated by the amount of charge at the channel. Likewise, trapped charge closer to the gate eventually draws more charge from the gate rather than from the channel (because the gate supplies the charge to the channel, a mechanism animated in Video S1, Supporting Information), so that $\Delta V_T$ is reduced. To estimate $\chi$, ranging from 0 to 1, for a generalized correlation of $\Delta V_T = \chi \Delta Q_M/C$, a series of additional TCAD simulations were performed with three different scenarios for the trap concentrations in Si$_3$N$_4$: $1 \times 10^{19}$, $3 \times 10^{19}$, and $10 \times 10^{19}$ cm$^{-2}$. As shown

![Figure 5](https://example.com/figure5.png)

**Figure 5.** Facile calibration protocol applicable to general three-terminal synaptic circuit elements utilizing the threshold voltage $V_T$ of a MOSFET. The experimentally measured $i_D$ after every pulse of $V_C = -12$ V and $+11$ V of the SONOS device from the study by Agarwal et al. [33] is used to illustrate the procedure. a) Channel current, $i_D$ (with $V_D$–$V_S = 0.1$ V and $V_G = 2.4$ V during reading) with time under repeating training pulses of $V_G = -12$ V (left) or $V_G = +11$ V (right) on the gate. b) The one-to-one mapping of the current, $i_D$, versus $\Delta V_T$ is possible because the $i_D$–$V_G$ curve is simply shifted on the $\nu_C$ axis, while the entire shape is negligibly deformed. c) Based on the conversion relation in (b), $\Delta V_T$ versus time can be obtained.
in Figure 6a, all three cases require $\Delta Q_M = 2.5 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$ to induce $\Delta V_T = 1 \text{V}$; hence, the resultant $\chi$ is calculated to be $0.77$, because $C = 0.31 \mu \text{F} \text{cm}^{-2}$. For the lowest trap concentration, $10^{19} \text{cm}^{-3}$ (Figure 6b), a slightly larger amount of trapped charge, $\Delta Q_M = 2.65 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$, is required to achieve $\Delta V_T = 1 \text{V}$ (i.e., $\chi = 0.73$) and the incremental efficiency becomes worse for the larger $Q_M (> 4 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$). This is because of the limited capacity of the trap layer so that for $Q_M > 4 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$, for instance, the trapped charges reside not only near the interface of $\text{Si}_3\text{N}_4$, but also in the bulk region of $\text{Si}_3\text{N}_4$, which is closer to the gate and induces a smaller $\Delta V_T$. With the higher trap concentrations in $\text{Si}_3\text{N}_4$ ($3 \times 10^{19} \text{cm}^{-3}$ and $10 \times 10^{19} \text{cm}^{-3}$), as shown in Figure 6c,d, the newly trapped electrons always occupy the interface between $\text{Si}_3\text{N}_4$ and the tunneling $\text{SiO}_2$ layer, so that the efficiency remains the similar with $\chi = 0.77$. Based on the analysis with different trap concentration scenarios, a conversion factor of $3 \times 10^{12} \text{e}^{-} \text{cm}^{-2} \text{V}^{-1}$ was chosen to extract $Q_M$ from the experimental measured data of $\Delta V_T$ in Figure 5c.

Figure 7a,b shows the resultant DRMs from $Q_M$ of the experimentally measured channel currents versus the number of square writing pulses on the gate[3] with seven different bias conditions: +11, +10, and +9 V for depression and −12, −11, 10, and −9 V for potentiation. Although the experimental points are noisy compared with those from TCAD simulations, the exponentially changing memristance with $Q_M$ is consistent with the expected capacitance of the ONO layer ($= 0.31 \mu \text{F} \text{cm}^{-2}$), which governs the correlation between $\Delta V_T$ and $\chi$. For a MOSFET, $\Delta Q_M = 1.93 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$ is required to induce $\Delta V_T = 1 \text{V} (= [0.31 \mu \text{F} \text{cm}^{-2}]/[1.6 \times 10^{-19} \text{C}])$ with $\chi = 1$ when the fixed charges exist at the interface between the silicon channel and $\text{SiO}_2$. For the SONOS device, for which the trapped charge is dominantly in the $\text{Si}_3\text{N}_4$ layer, $\chi = 0.73$ was derived from the simulation results: $\Delta Q_M = 2.65 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$ for every $\Delta V_T = 1 \text{V}$. c) When the trap density increases to $3 \times 10^{19} \text{cm}^{-3}$, the trapped charge has a narrower distribution, spatially closer to the channel rather than the gate. As a result, $\Delta Q_M$ contributes 77% of $\Delta Q_M$ (i.e., $\chi = 0.77$) and the conversion factor slightly decreased to $2.52 \times 10^{12} \text{e}^{-} \text{cm}^{-2} \text{V}^{-1}$. d) Further increase to $10 \times 10^{19} \text{cm}^{-3}$ does not create a notable change compared with $3 \times 10^{19} \text{cm}^{-3}$ from (c) (i.e., $\chi = 0.77$), hence the conversion factor remains as $2.52 \times 10^{12} \text{e}^{-} \text{cm}^{-2} \text{V}^{-1}$.

Figure 6. Relationship between $V_T$ and $Q_M$ under various conditions of $\text{Si}_3\text{N}_4$ trap density to obtain $Q_M$ from the experimentally obtained $V_T$ data. a) $V_T$ linearly increases by 1 V for every $\Delta Q_M = 2.5 \times 10^{12} \text{e}^{-} \text{cm}^{-2}$. A detailed analysis for the case of trap density equal to $10^{12} \text{e}^{-} \text{cm}^{-2}$ is shown in Figure 7b. b) A detailed analysis for the case of trap density equal to $10^{19} \text{cm}^{-3}$ is shown in Figure 7c. c) When the trap density increases to $3 \times 10^{19} \text{cm}^{-3}$, the trapped charge has a narrower distribution, spatially closer to the channel rather than the gate. As a result, $\Delta Q_M$ contributes 77% of $\Delta Q_M$ (i.e., $\chi = 0.77$) and the conversion factor slightly decreased to $2.52 \times 10^{12} \text{e}^{-} \text{cm}^{-2} \text{V}^{-1}$. d) Further increase to $10 \times 10^{19} \text{cm}^{-3}$ does not create a notable change compared with $3 \times 10^{19} \text{cm}^{-3}$ from (c) (i.e., $\chi = 0.77$), hence the conversion factor remains as $2.52 \times 10^{12} \text{e}^{-} \text{cm}^{-2} \text{V}^{-1}$.
0.36 \text{ m}. This is likely due to the absence of trap-assisted tunneling (TAT) in the TCAD simulations that becomes prominent at smaller $V$.\textsuperscript{19} The nonlinearity parameter $\beta$ for the experimental SONOS device is ten times lower, so that the low-bias dynamics is faster compared with the simulation model. This may also be caused by the absence of TAT through defect states in the forbidden bands of SiO$_2$ and Si$_3$N$_4$ and thus worse predictability for low-bias cases. This is a known problem for industrial SONOS TCAD models, where the pass disturb simulations with $V = 7$–8 V typically underestimate $Q_M$ compared with fabricated devices,\textsuperscript{19} because defects in SiO$_2$ and Si$_3$N$_4$ are difficult to model. The parameter $\gamma$, responsible for the sensitivity to $Q_M$, was found to be the same for both the physics simulations and the experimental data. Finally, $\delta$, which handles the $Q_M$ sensitivity of $M$ depending on $V$, was found to be the same for both the physics simulations and the experimental data. Figure 7d shows the improved fits with distinct parameters for potentiation ($V < 0$) and depression ($V > 0$), such that little error remains between the extracted memristance values and the analytic equation, as also found in two-terminal compact models.\textsuperscript{146}

We deployed the compact model in a commercial circuit simulation package, Cadence Spectre, using VerilogA to assess the agreement with the measured drain current (with $V_D = 0.1$ V and $V_C = 2.4$ V during reading)\textsuperscript{131} after various training pulses ($V_C$) of $-12, -11, -10, -9, +9, +10, +11$ V, as shown in Figure 8a. Despite the remarkable simplicity of the compact model, good agreement with the measured drain current was achieved.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure7}
\caption{Figure 7. Extracted DRMs and corresponding memristance value maps of the experimental SONOS device in the study by Agarwal et al.\textsuperscript{33} a) and b) Under both positive and negative biases for depression and potentiation, respectively, a linear approximation for $\log(dQ_M/dt)$ versus $Q_M$ holds reasonably well, consistent with the physics simulation results and the compact model. c) Selected experimental $M$ values (triangles, red is $Q_M = 0$ and blue is $Q_M = Q_{M,\text{max}}$, similar to Figure 3f to build the dotted lines in (a) and (b). The analytical function with a parameter set ($\alpha$, $\beta$, $\gamma$, $\delta) = (70,000, 7, 50, 1)$ simultaneously fits both potentiation and depression, but with significant errors possibly due to different physics for charge trapping and detrapping. d) A modified function for $M$, where the parameters are different for potentiation ($\alpha_p$, $\beta_p$, $\gamma_p$, $\delta_p) = (10^5, 2.5, 3000, 1)$ and depression ($\alpha_d$, $\beta_d$, $\gamma_d$, $\delta_d) = (5500, 9, 900, 1)$, models the experimental SONOS data with negligible error.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure8}
\caption{Figure 8. a) Cadence Spectre simulation results for SONOS current, $i_D$, versus the accumulated learning time $t$ compared with the experimental results of Agarwal et al.\textsuperscript{33} under $V_C = 2.4$ V and $V_D = 0.1$ V for reading the channel current. While the measurements of $i_D$ for the experimental SONOS devices were conducted after every writing pulse ($-9$ to $-12$ V for potentiation, $+9$ to $+11$ V for depression on $V_C$ with $10 \mu$s width while $V_D = V_T = 0$), our Cadence Spectre simulations were sampled after every $100 \mu$s. Slight errors are attributed to the deviations from the perfectly exponential increase of $M$ with $Q_M$ assumed in our compact model. b) The corresponding evolution of the state variable $Q_M$ is available from the simulation, which demonstrates the strong correlation between $\Delta Q_M$ and $\Delta i_D$.}
\end{figure}
agreement between the simulation and the experiment was confirmed with the accuracy higher than 90%, proving that our compact model captures the essential physics in a simple way. Figure 8b shows the corresponding evolution of the key state variable $Q_M$, where a linear correlation is observed between $\Delta Q_M$ and $\Delta N_D$ as foreseen by the description of $R_{CH}$ in Equation (6). The linearity between $\Delta Q_M$ and $\Delta N_D$ can be further exploited for linear and symmetric writing, which is ideally desired for autonomous weight updates without the necessity for a time-consuming “write and verify” scheme in neuromorphic learning. Fuller et al.\[32\] reported the use of current pulses instead of voltage pulses to mitigate the nonlinearity and asymmetry in writing that is commonly observed in memristors or NVM devices including SONOS devices, as shown in Figure 3a,b, 5a,c, and 8a,b. The mathematical representation of our compact model as represented by Equation (6) is consistent with the empirical observation of Fuller et al.\[32\] and thus three-terminal synaptic circuit elements are expected to be promising candidates for the building unit of neuromorphic learning systems.

In conclusion, this work presented a well-posed compact model of a SONOS three-terminal synaptic circuit element that can be readily utilized by circuit designers for neuromorphic computing circuits/systems. This compact model offers the rare combination of simplicity and good predictability, stemming from physics-driven state variables. The basis of the compact model is rooted on the carrier concentration modulation in the conducting channel, so that it is readily applicable to other three-terminal synaptic circuit elements besides SONOS devices.

**Supporting Information**

Supporting Information is available from the Wiley Online Library or from the author.

**Acknowledgements**

S.Y. and R.S.W. were supported by the Air Force Office of Scientific Research (AFOSR) under grant number AFOSR-FW9550-19-0213, titled “Brain Inspired Networks for Multifunctional Intelligent Systems in Aerial Vehicles.” R.S.W. acknowledges the X-Grants Program of the President’s Excellence Fund at Texas A&M University. A.A.T. and M.J.M. were supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the USA. Department of Energy’s National Nuclear Security Administration under contract DE-NA-0003525, and by the DOE Office of Science Research Program for Microelectronics Codesign (sponsored by ASCR, BES, HEP, NP, and FES) through the Abisko Project, PM Robinson Pino (ASCR). Note: Name of the author was changed from Richard Stanley Williams to R. Stanley Williams on September 22, 2022, after initial publication online.

**Conflict of Interest**

The authors declare no conflict of interest.

**Data Availability Statement**

The data that support the findings of this study are available in the supplementary material of this article.

**Keywords**

compact models, memristors, neuromorphic circuit designs, SONOS, three-terminal synaptic devices

Received: March 12, 2022
Revised: April 15, 2022
Published online: May 22, 2022

[1] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[2] A. Krizhevsky, I. Sutskever, G. Hinton, *Adv. Neural Inform. Process. Syst.* **2012**, *25*, 1097.
[3] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *J. Proc. IEEE* **1998**, *86*, 2278.
[4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. Jackel, *J. Neural Comput.* **1989**, *1*, 541.
[5] A. Shafiee, A. Nag, N. Muralimanohar, R. Balasubramonian, J. P. Strachan, M. Hu, R. S. Williams, V. Srikumar, *J. ACM Sigarch Comput. Architect. News* **2016**, *44*, 14.
[6] Y.-H. Chen, T. Krishna, J. S. Emer, V. Sze, *IEEE J. Solid-State Circuits* **2016**, *52*, 127.
[7] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *Nature* **2017**, *550*, 354.
[8] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers, in *Proc. of the 44th Annual Int. Symp. on Computer Architecture*, 2017, pp. 1–12.
[9] R. S. Williams, *J Comput. Sci. Eng.* **2017**, *19*, 7.
[10] D. B. Strukov, G. S. Snider, D. R. Stewart, R. S. Williams, *Nature* **2008**, *453*, 80.
[11] L. Chua, *IEEE Trans. Circuit Theory* **1971**, *18*, 507.
[12] L. O. Chua, S. M. Kang, *J. Proc. IEEE* **1976**, *64*, 209.
[13] J. J. Yang, M.-X. Zhang, J. P. Strachan, F. Miao, M. D. Pickett, R. D. Kelley, G. Medeiros-Ribeiro, R. S. Williams, *Appl. Phys. Lett.* **2010**, *97*, 232102.
[14] K. M. Kim, R. S. Williams, *IEEE Trans. Circuits Syst. I: Regul. Papers* **2019**, *66*, 4348.
[15] M. Prezioso, F. Merrikk-Bayat, B. Hoskins, G. C. Adam, K. K. Likharev, D. Strukov, *Nature* **2015**, *521*, 61.
[16] S. Kim, M. Ishii, S. Lewis, T. Perri, M. Brightsly, W. Kim, R. Jordan, G. Burr, N. Sosa, A. Ray, in *IEEE Int. Electron Devices Meeting (IEDM)*, IEEE, Piscataway, NJ **2015**, pp. 17.1–17.1. 4.
[17] S. Ambrogio, P. Narayanan, H. Tsai, R. M. Shelby, I. Boybat, C. Di Nolfo, S. Sidler, M. Giordano, M. Bodini, N. Farinha, *Nature* **2018**, *558*, 60.
[18] S.-I. Yi, S. Kumar, R. S. Williams, *IEEE Trans. Circuits Syst. I: Regul. Papers* **2021**, *68*, 4970.
[19] S.-I. Yi, J. Kim, *Micromachines* **2021**, *12*, 584.
[20] H. Kim, J. Park, M.-W. Kwon, J.-H. Lee, B.-G. Park, *IEEE Electron Dev. Lett.* **2016**, *37*, 249.
[21] M. R. Mahmoudi, D. Strukov, in *Proc. of the 55th Annual Design Automation Conf.*, **2018**, pp. 1–6.
[22] M. Hu, C. E. Graves, C. Li, Y. Li, N. Ge, E. Montgomery, N. Davila, H. Jiang, R. S. Williams, J. J. Yang, *Adv. Mater.* **2018**, *30*, 1705914.
[23] Z. Wang, C. Li, W. Song, M. Rao, D. Belkin, Y. Li, P. Yan, H. Jiang, P. Lin, M. Hu, *Nat. Electron.* **2019**, *2*, 115.
[24] C. Li, Z. Wang, M. Rao, D. Belkin, W. Song, H. Jiang, P. Yan, Y. Li, P. Lin, M. Hu, Nat. Machine Intell. 2019, 1, 49.

[25] C. Li, D. Belkin, Y. Li, P. Yan, M. Hu, N. Ge, H. Jiang, E. Montgomery, P. Lin, Z. Wang, Nat. Commun. 2018, 9, 1.

[26] Z. Wang, S. Joshi, S. Savel’ev, W. Song, R. Midya, Y. Li, M. Rao, P. Yan, S. Asapu, Y. Zhuo, Nat. Electron. 2018, 1, 137.

[27] L. Danial, V. Gupta, E. Pikhay, Y. Roizin, S. Kvatinsky, in 2020 Design, Automation & Test in Europe Conf. & Exhibition (DATE), IEEE, Piscataway, NJ 2020, pp. 472–477.

[28] C. Zhang, S. R. Hasan, Analog Integr. Circuits Signal Process. 2019, 101, 1.

[29] M. Kim, S. Kim, H. Shin, IEEE Trans. Electron Dev. 2020, 67, 3095.

[30] E. J. Fuller, S. T. Keene, A. Melianas, Z. Wang, S. Agarwal, Y. Li, Y. Tuchman, C. D. James, M. J. Marinella, J. J. Yang, Science 2019, 364, 570.

[31] Y. van de Burgt, E. Lubberman, E. J. Fuller, S. T. Keene, G. C. Faria, S. Agarwal, M. J. Marinella, A. A. Talin, A. Salleo, Nat. Mater. 2017, 16, 414.

[32] E. J. Fuller, F. E. Gabaly, F. Léonard, S. Agarwal, S. J. Plimpton, R. B. Jacobs-Gedrim, C. D. James, M. J. Marinella, A. A. Talin, Adv. Mater. 2017, 29, 1604310.

[33] S. Agarwal, D. Garland, J. Niroula, R. B. Jacobs-Gedrim, A. Hsia, M. S. Van Heukelom, E. Fuller, B. Draper, M. Marinella, IEEE J. Expl. Solid-State Comput. Dev. Circuits 2019, 5, 52.

[34] B. Widrow, Adaptive” adaline” Neuron Using Chemical” memistors”, 1960.

[35] S. P. Adhikari, H. Kim, Memristor Networks, Springer, New York 2014, pp. 95–112.

[36] H. Kim, S. P. Adhikari, IEEE Circuits Syst. Mag. 2012, 12, 75.

[37] J. Jiang, H.-S. Kim, W. Cho, H. Cho, J. Kim, S. I. Shim, J.-H. Jeong, B.-K. Son, D. W. Kim, J.-J. Shim, in 2009 Symp. on VLSI Technology, IEEE, Piscataway, NJ 2009, pp. 192–193.

[38] E. Nowak, J.-H. Kim, H. Kwon, Y.-G. Kim, J. S. Sim, S.-H. Lim, D. S. Kim, K.-H. Lee, Y.-K. Park, J.-H. Choi, in Symp. on VLSI Technology (VLSIT), IEEE, Piscataway, NJ 2012, pp. 21–22.

[39] D. Kang, M. Kim, S. C. Jeon, W. Jung, J. Park, G. Choo, D.-K. Shim, A. Kavala, S.-B. Kim, K.-M. Kang, in IEEE Int. Solid-State Circuits Conf. (ISSCC), IEEE, Piscataway, NJ 2019, pp. 216–218.

[40] S. Lee, C. Kim, M. Kim, S. M. Joe, J. Jang, S. Kim, K. Lee, J. Kim, J. Park, H. J. Lee, in IEEE Int. Solid-State Circuits Conf (ISSCC), IEEE, Piscataway, NJ 2018, pp. 340–342.

[41] L. Chua, Handbook of Memristor Networks, Springer, New York 2019, pp. 89–157.

[42] A. S. Sedra, K. C. Smith, T. C. Carusone, V. Gaudet, Microelectronic Circuits, Vol. 4, Oxford University Press, New York 2004.

[43] B. G. Streetman, S. Banerjee, Solid State Electronic Devices, Vol. 10, Pearson/Prentice Hall, Upper Saddle River 2006.

[44] Synopsys, Sentaurus Manual S-Device, L-version, 2016.

[45] A. Ascoli, R. Tetzlaff, L. O. Chua, J. P. Strachan, R. S. Williams, IEEE Trans. Circuits Syst. I: Regul. Papers 2016, 63, 389.

[46] M. D. Pickett, D. B. Strukov, J. L. Borghetti, J. J. Yang, G. S. Snider, D. R. Stewart, R. S. Williams, J. Appl. Phys. 2009, 106, 074508.

[47] C. D. Danesh, C. M. Shaffer, D. Nathan, R. Shenoy, A. Tudor, M. Tadayon, Y. Lin, Y. Chen, Adv. Mater. 2019, 31, 1808032.

[48] Y. Li, E. J. Fuller, J. D. Sugar, S. Yoo, D. S. Ashby, C. H. Bennett, R. D. Horton, M. S. Bartsch, M. J. Marinella, W. Lu, Adv. Mater. 2020, 32, 2003984.

[49] C. M. Compagnoni, A. Goda, A. S. Spinelli, P. Feeley, A. L. Lacaia, A. Visconti, Proc. IEEE 2017, 105, 1609.

[50] M. J. Marinella, S. Agarwal, A. Hsia, I. Richter, R. Jacobs-Gedrim, J. Niroula, S. J. Plimpton, E. Ipek, C. D. J. James, IEEE J. Emerg. Select. Topics Circuits Syst. 2018, 8, 86.