SOLID-STATE IONICS

Nanoscale protonic programmable resistors for analog deep learning

Murat Onen1,2, Nicolas Emond2,3, Baoming Wang2,3, Difei Zhang1,2, Frances M. Ross2,3, Ju Li2,3,4, Bilge Yildiz2,3,4, Jesús A. del Alamo1,2,4

Nanoscale ionic programmable resistors for analog deep learning are 1000 times smaller than biological cells, but it is not yet clear how much faster they can be relative to neurons and synapses. Scaling analyses of ionic transport and charge-transfer reaction rates point to operation in the nonlinear regime, where extreme electric fields are present within the solid electrolyte and its interfaces. In this work, we generated silicon-compatible nanoscale protonic programmable resistors with highly desirable characteristics under extreme electric fields. This operation regime enabled controlled shuttling and intercalation of protons in nanoseconds at room temperature in an energy-efficient manner. The devices showed symmetric, linear, and reversible modulation characteristics with many conductance states covering a 20× dynamic range. Thus, the space-time-energy performance of the all–solid-state artificial synapses can greatly exceed that of their biological counterparts.

aqueous ions with mobile Na+, K+, Ca2+, and other ions underpin biological information processing. In neurons and synapses, the action potential of magnitude ∼100 mV evolves over a characteristic time scale of milliseconds, which fundamentally limits the speed of thinking and reflexes of animals. Because liquid water decomposes at voltages >1.23 V, a weak action potential is understandable. However, with human-made solid-state neurons and synapses (1–3) for analog machine learning, research efforts are no longer limited by the stability window of aqueous electrolyte. Furthermore, it is also possible to fabricate devices that are much smaller than biological neurons, by a length-scale shrinkage factor of 103 (from ~10 μm to ~10 nm). This possibility raises a fundamental question about how much faster we can train such “artificial synapses”—i.e., what is the ultimate speed limit to solid-state ionic–based analog deep learning? Also, when we approach that speed limit, how energy intensive would the training be, given that dissipative processes generally produce more entropy per task the faster that task is accomplished?

Fundamental ionic arguments seem to call for high voltage and small length scales—that is, an extreme programming field approach (4–10). Transport of ions (such as H+) inside a solid electrolyte (SE) layer and a mixed ionic–electronic conductor (MIEC) conductance channel layer, as well as charge-transfer reactions at the SE/MIEC interfaces, scale monotonically and nonlinearly with the applied voltage. Therefore the speed of change in the nonvolatile state of artificial neurons should increase with increasing electric field. However, there is a dielectric breakdown limit (11) to solids, as well as a thermodynamic electrochemical stability window for SEs. Therefore, the programming field should be raised as high as possible, but not so high as to permanently damage the SE. In this work, we demonstrate that such an approach produced exceedingly fast nanionic devices (5 ns), at least 104 times as fast as biological synapses, when the voltage was on the order of +10/−8.5 V, yielding an extremely high electric field of ~1 V/μm across the proton-conducting phosphosilicate glass (PSG) SE. Surprisingly, we also show that under appropriate operating parameters, this nonlinear ionic-electronic device was robust and reversible, operating successfully over millions of cycles. Although the layout of our three-terminal device is analogous to that of a solid-state battery, it operated 1010 times as fast as conventional solid-state batteries. Finally, even when approaching the ultimate speed limit of our devices, the heat generated per programming task was still favorable compared with that generated by a human synapse (~10 fJ per state).

Interest in engineering the ideal programmable resistor for analog computing applications has skyrocketed owing to increasing workloads of deep learning problems (12–14) and diminishing expectations for transistor performance improvements through size scaling (15). Highly optimized digital application-specific integrated circuits that run reduced precision arithmetic operations can still manage inference tasks (16); however, the resolution requirements for training tasks do not allow sufficient bit reduction (17) and make beyond-digital approaches imperative. The core idea behind analog training accelerators is to process information locally using physical device properties instead of conventional Boolean arithmetic—i.e., using Ohm’s and Kirchhoff’s laws for the matrix inner product (18) and threshold-based updating for the outer product (19). However, the performance benefits attained by analog training processors are dependent on a set of very strict properties: The device must be fast, energy-efficient, nonvolatile, and reversible, and it must show symmetric conductance modulation with many conductance states across a large dynamic range (20, 21). Unfortunately, readily available memory technologies [e.g., phase-changing (22, 23), filamentary (24), bulk-switching (25–28), and ferroelectric (29) resistive memories] do not completely satisfy these requirements. As a result, demonstrations so far have required additional control circuitry and serial operations (22–24) to correct such imperfections, at the expense of substantially diminished space-time-energy performance.

Because the core idea of neural network training is to tune the state parameters (weights) through many small incremental modifications (12, 13), it is reasonable that devices originally designed for information storage purposes (i.e., memory technologies) do not perform well in this information processing application. Devices for analog deep learning (i.e., training) must be primarily optimized for state transition (i.e., modulation) rather than long-term state preservation (i.e., nonvolatility). The retention of the weight parameters is obviously important such that the information learned from previous inputs is not lost faster than the ongoing optimization. However, given that the purpose of the network is not to memorize correct labels for the training dataset but to find generalized features instead, retaining exact values for extended durations of time loses relevance. The best example for this argument can be found in digital neural network training, for which it is a common practice to simulate weight decay on values that could otherwise be retained for indefinite time (30). Because the properties that ensure long retention times contradict those that enable high-speed and low-energy modulation (30), we conclude that for training implementations the focus should be on the latter two, trading retention down to the ~100-s–to–1000-s range for nanosecond and subpicosecond timescales [see the supplementary materials (SM)].

Any memristive technology that aims to realize practical analog processors must contain a high density of devices built on top of standard complementary metal-oxide semiconductor (CMOS) logic and therefore must rely on CMOS-compatible materials and involve back-end-of-line-compatible processes.

1Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA. 2MIT-IBM Watson AI Lab, 75 Binney St., Cambridge, MA 02142, USA. 3Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA. 4Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA. *Corresponding author. Email: monen@mit.edu (M.O.); alamoa@mit.edu (J.A.d.A.); byldiz@mit.edu (B.Y.); liju@mit.edu (J.L.)
been extensively modulated in the ultrafast regime. A top-view SEM image of a fabricated device with a 30-nm–by–60-nm channel. (C) Transmission electron microscopy (TEM) cross-sectional image of a protonic programmable resistor that had previously been extensively modulated in the ultrafast regime.

Moreover, device operation should not depend on unconventional environmental conditions such as high temperature or humidity. With these concepts in mind, we developed an integration-friendly technology with inorganic solid materials that are native to conventional Si processing: WO$_3$ as the active channel material, nanoporous PSG as the protonic SE, and Pd as the hydrogen reservoir and controlling gate. The basic operating principle of the device relies on modulating the channel conductance via the electrochemically controlled intercalation of protons (chosen as the ion with the smallest radius and lightest mass) into WO$_3$ (32, 33).

The rationale for choosing these materials is as follows: WO$_3$ is known for its conductance modulation and electrochromism upon ion intercalation (25, 32–34) with well-established dynamics (32, 33). For analog deep learning applications, the channel material must also have high base resistivity (19), proton insertion rate, and sensitivity. We found that superior modulation characteristics result from reactive sputtering of WO$_3$ at room temperature, followed by a 400°C annealing process that both oxidizes and crystallizes the material (see SM for process details). Regarding the electrolyte, the material must simultaneously show very high electrochemical resistivity and high proton conductivity. In our previous work, we identified nanoporous PSG as an outstanding SE (36–38) that displays both properties at room temperature without the need for hydration (i.e., humidity) (33). Finally, Pd was chosen as the gate metal and hydrogen reservoir (PdH$_x$), owing to its capability to take up and store hydrogen (39).

The three-terminal protonic programmable resistors studied in this work (Fig. 1A) were fabricated through a series of electron-beam lithography processes. A self-aligned gate structure was used to scale down device dimensions, avoiding mask alignment limitations. In a key design aspect of this process, the Pd layer was overlaid across a large region (Fig. 1B), which was then used as an etch mask for the PSG layer beneath. The height of the channel electrodes (Au) was calibrated such that the PSG layer could cover the sidewalls and the Pd layer that overlapped with the channel electrodes was disconnected from the gate electrode (Fig. 1C), avoiding the charging of those regions and unnecessary capacitance. This configuration was intended to maximize energy efficiency. The complete fabrication flow, additional metrology results, and yield details can be found in the SM.

Before testing, the devices were exposed to forming gas (3% H$_2$ in N$_2$) at room temperature for protonation of the Pd reservoir and then pumped down to vacuum. Future iterations of these protonic devices will be encapsulated to avoid environmental preconditioning steps. Figure 2A shows the channel conductance modulation of a 50-nm–by–150-nm device with 10-nm-thick PSG for 1000 protonating voltage pulses ($V^+ = 10$ V) followed by 1000 deprotonating pulses ($V^- = -8.5$ V). Between successive pulses, the channel conductance was read under drain-source voltage ($V_{DS}$) = 0.1 V and gate current ($I_G$) = 0 conditions and averaged for ~1 s. The devices displayed nearly ideal characteristics in terms of (i) high modulation speed, responding to 5-ns voltage pulses; (ii) nearly linear and symmetric behavior for incremental and decremental changes; (iii) conductance retention characteristics over durations longer than ~10$^{10}$ times the unit pulse time (Fig. 2B); (iv) dynamic conductance range of 20x; (v) optimal base resistance of 88 megohms for readout (39, 32); and (vi) preservation of these ideal properties without any degradation over extended time and use (Fig. 2C).

Moreover, the devices showed excellent energy efficiency under this ultrafast operation, the gate current supplied during each pulse being too small to be precisely measured for the small devices. The energy consumption during the transients was estimated to be ~2.5 fJ per pulse (see SM), which is a technology-agnostic overhead related to charging and discharging the gate capacitance. On the other hand, the energy consumed in proton transfer while the 5-ns voltage pulse was at its peak value was estimated to be ~15 aJ per pulse for the device whose performance is shown in Fig. 2. This value was based on a dc gate current measurement of ~30 nA at 10 V for a 750-nm–by–1000-nm device. This latter value is associated with the efficient shuttling of ions within the

Fig. 1. Nanosecond protonic programmable resistors. (A) Three-dimensional illustration of the protonic programmable resistors studied in this work. Au (yellow), WO$_3$ (green), PSG (magenta), and Pd (gray) layers are indicated. As a result of an engineered sidewall, the Pd layer that overlaps with the channel electrodes is isolated from the remainder of the gate electrode. G, gate; D, drain; S, source. (B) False-colored top-view SEM image of a fabricated device with a 30-nm–by–60-nm channel. (C) Transmission electron microscopy (TEM) cross-sectional image of a protonic programmable resistor that had previously been extensively modulated in the ultrafast regime.

Fig. 2. Ultrafast and energy-efficient modulation characteristics of protonic programmable resistors. (A) Modulation performance of a 50-nm–by–150-nm protonic device with 10-nm PSG, showing fast (5 ns per pulse), nearly linear, and symmetric characteristics. W, width; L, length. (B) Retention behavior of the protonic device for ~100 s at different conductance levels over the full dynamic range. (C) Endurance characterization of the protonic device, displaying nondegrading modulation over 10$^3$ pulses conducted over 30 hours.
gate stack under the high electric field. Additional data and calculations related to the energy consumption can be found in the SM.

As a result, the combined material, processing, and performance benefits of the all-solid-state protonic devices demonstrated here exceed those of nonvolatile memory technologies.

To explain these desirable modulation characteristics, we have developed a model for device operation that consists of two key parts: (i) proton transport in the PSG and (ii) proton-coupled electron-transfer reaction rates at the PSG/electrode interfaces. Both processes have qualitatively similar formalisms and dependencies—the former is governed by hopping conduction in a disordered solid with random site energies (4–7, 10), whereas the latter is determined by the Butler-Volmer charge-transfer kinetics (8, 9). Conventionally, a high activation energy for protons to be freed from their sites (e.g., Si–O–H or P–O–H) results in low ionic conductivity (40) and ultimately limits the operation speed at room temperature. For amorphous SiO₂, this activation energy was reported to be 0.38 eV, which is much larger than the thermal energy in ambient conditions (25.9 meV) (41).

In the presence of a high electric field, the energy barrier to ion conduction is lowered in the field direction, yielding an enhanced proton hopping current

\[ I_G(V_{\text{pulse}}) \propto \sinh \left( \frac{q a V_{\text{pulse}}}{2 k_B T d_{\text{PSG}}} \right) \]  

where \( q \) is the electron charge, \( k_B \) is the Boltzmann constant, \( T \) is the temperature, \( a \) is the apparent hopping distance, and \( d_{\text{PSG}} \) is the thickness of the electrolyte (5). Owing to the high resistance of the PSG layer, all of the pulse voltage is assumed to drop across the electrolyte (\( V_{\text{PSG}} \approx V_{\text{pulse}} \); see SM for details). Figure 3A shows the experimentally observed conductance change per pulse (\( \Delta G_{\text{channel}} \)) as a function of pulse amplitude (\( V_{\text{pulse}} \)) and pulse time (\( t_{\text{pulse}} \)) for a device with \( d_{\text{PSG}} = 5 \text{ nm} \). Over a range of electric fields similar to that detailed in Fig. 2, the results shown in Fig. 3 closely followed Eq. 1, as indicated by the lines. Furthermore, experiments for devices with different \( d_{\text{PSG}} \) values allowed us to extract a hopping distance for the PSG of 5.6 Å (Fig. 3B). This result is in good agreement with previously reported values for amorphous silica glasses (41–43).

In our case, the electric field (\( \approx 1 \text{ V/nm} \)) across the hopping distance (\( a = 0.5 \text{ nm} \)) was so high that it might completely remove the activation barrier within the PSG (\( \approx 0.4 \text{ eV} \)). This effect resolves the bottleneck of low proton conduction at room temperature, thus enabling high-speed operation. Such an effect was predicted in simulations (44) but has not been previously observed experimentally, as the required conditions are beyond the breakdown field or the electrochemical stability window of traditional electrolyte materials (45). Instead, PSG allows a high critical field (8 to 15 MV/cm) (46, 47) in addition to a moderate base proton conductivity (36–38), making it an ideal electrolyte choice for this application. Moreover, under the classical statistical mechanics scenario, reducing the migration energy below a few \( k_B T \) boosts the likelihood of classical or quantum ballistic motion of protons. Indeed, in liquid water, proton motion is known to have a pronounced quantum character, with activationless quantum nuclear dynamics in some exchange events between water molecules (49). Thus, under such high electric fields, “quantum ballistic” transport of protons in solids may be realized.

Unlike the conductive filament devices controlled by angstrom-scale short-range ion motion at the tip of the filament (40), our protonic devices rely on long-range (\( \sim 10 \text{ nm} \)) and uniform partitioning of protons across the electrolyte with the protons then behaving as dynamic dopants in the entire \( \text{H}_3\text{WO}_4 \) channel (32, 35). Furthermore, conducting filament formation is assisted by Joule heating, with local temperatures rising as high as \( \sim 500 \text{ K} \) (50), whereas in our case ion transport is facilitated only by the electric field with negligible temperature increase (\( \Delta T \approx 2 \text{ K} \); see SM). As a result, although very-short-range ion transport at the tip of the filament can readily occur in sub-nanosecond time scales, the dependence of filament properties on the local microstructure and microchemistry makes those devices unpredictable and stochastic (2, 9, 40). On the contrary, devices studied in this work are modulated by macroscopic chemical and electronic modifications, which result in distinctly controllable and deterministic performance characteristics. These benefits make their acceleration to the nanosecond regime highly relevant.

The ability to rapidly shuttle protons within bulk PSG at extreme speeds shifts the bottleneck of proton transport to the interfaces. Although the hydrogen uptake rate of Pd is high (51) and the \( \text{PdH}_x/\text{PSG} \) interfacial reaction is likely also efficient, the same cannot be said for proton insertion into polycrystalline \( \text{WO}_3 \) (52), which shows up as a PSG/\( \text{H}_3\text{WO}_4 \)
interfacial charge-transfer resistance. Because protons do not have high diffusivity in H₂WO₃ (33), it is likely that the inserted species cannot quickly vacate their sites near the interface, thus further reducing the insertion rate in a self-limiting fashion. These factors result in an excess of highly energetic protons at the WO₃/PSG interface and may lead to H₂ gas formation and buildup at the interface (54).

The electrical signature and morphological consequences of long-pulse-time stressing of the devices are captured in Fig. 4. In Fig. 4A, the device previously characterized in Fig. 2 was tested under the same Vpulse but for increasing fpulse. Above 40 ns a cascading effect was apparent, in which the conductance change increased with each pulse, ultimately causing device failure for 90-ns pulses. A scanning electron microscopy (SEM) image of the device after this experiment is shown in Fig. 4B, whereas Fig. 4C shows an image of another device stopped at earlier stages (corresponding to the 40-ns pulse regime shown in Fig. 4A). These images show damage features that we suggest were consistent with H₂ gas evolution at the PSG/H₂WO₃ interface, nanobubble formation, and stress buildup.

Most importantly, as shown in Fig. 2C, no degradation was observed under high-speed (5 ns) operation. To provide further evidence of good endurance characteristics in the ultrafast regime, we performed electron energy-loss spectroscopy (EELS) for the SE in fresh and tested devices. Figure 4D shows that the Si L₂,3 energy-loss spectra were similar in active and inactive regions across the PSG layer, which indicates that there was no stoichiometry change or that the change was too subtle to be detected. Considering that the widest electrochemical stability window for room temperature protonic electrolytes was previously ~3.35 V (55), we attribute our devices’ ability to operate stably and reversibly at 10 V/8.5 V to the operation at nanosecond transients, where the PSG does not have to conform to the thermodynamic stability requirement for electrolytes (established after quasi-static wait times). Furthermore, these short time scales may simply be too fast for oxygen motion (slower than that of protons), which would otherwise cause degradation of material properties. We believe these are two key physical dynamics underlying the no-damage “protonic breakdown,” a transient phenomenon that we took advantage of in this novel ultrafast long-range transport regime.

In summary, we explored the limits of all-solid-state ionics to answer the question of how much faster an electrochemical artificial synapse can operate relative to its biological counterpart. Under extreme electric field conditions, we demonstrated ultrafast modulation of nanoscale protonic programmable resistors with outstanding energy efficiency. Moreover, the devices exhibited many nonvolatile channel conductance states with good retention across a large and ideal dynamic range, which could be programmed reversibly, repeatedly, and symmetrically. Operation of the devices at extreme electric fields across the stack without any material degradation is the key breakthrough to enable such performance. Under these conditions, the activation energies that govern proton transport as well as charge-transfer reactions at the interface were substantially lowered, thus resulting in operation speed enhanced by six orders of magnitude at room temperature. Beyond artificial synapses, these findings open up possibilities in fields that require fast ion motion, such as microbatteries, artificial photosynthesis, and light-matter interactions.

REFERENCES AND NOTES
32. X. Yao et al., Nat. Commun. 11, 3134 (2020).

ACKNOWLEDGMENTS

We thank J. Rozen, T. Todorov, P. Solomon, T. Gokmen, and M. Rasch from the IBM T. J. Watson Research Center for program support and fruitful discussions. Device fabrication took place at the facilities of MIT.nano and the MIT Electron-Beam Lithography Facility. Materials characterization took place at the facilities of MIT’s Materials Research Laboratory. Funding: This project is funded by the MIT-IBM Watson AI Lab. N.E. would like to thank the Fonds de recherche du Québec et Nature et technologies (FRQNT) for financial support. Author contributions: M.O. conceived the original idea, fabricated devices, and performed experimental characterization. N.E. optimized the channel material and performed metrology studies. M.O. and N.E. designed experiments and analyzed data. B.W. designed and conducted TEM characterizations. All authors contributed to developing the theory and writing the manuscript. Competing interests: None declared. Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper or the supplementary materials and can be found in Zenodo (56).

License information: Copyright © 2022 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse
Nanosecond protonic programmable resistors for analog deep learning
Murat Onen, Nicolas Emond, Baoming Wang, Difei Zhang, Frances M. Ross, Ju Li, Bilge Yildiz, and Jess A. del Alamo

Science, 377 (6605), 6
DOI: 10.1126/science.abp8064

Speed limit for solid-state nanoionics
The speed of biological information processing in neurons and synapses is limited by the aqueous medium through which weak action potentials of about 100 millivolts propagate over milliseconds. Above 1.23 volts, liquid water decomposes. Artificial solid-state neurons are not limited by such time and voltage constraints and can also be fabricated at the nanoscale, 1000 times smaller than their biological counterparts. Using complementary metal-oxide semiconductor–compatible materials, Onen et al. prototyped nanoscale protonic programmable resistors that can withstand high electric fields of around 10 megavolts per centimeter and which have energy-efficient modulation characteristics at room temperature. The proposed devices are 10,000 times faster than biological synapses and offer a promising direction for implementing various applications that can benefit from fast ionic motion.—YS

View the article online
https://www.science.org/doi/10.1126/science.abp8064
Permissions
https://www.science.org/help/reprints-and-permissions