Programmable superconducting optoelectronic single-photon synapses with integrated multi-state memory

Cite as: APL Mach. Learn. 2, 026122 (2024); doi: 10.1063/5.0204469
Submitted: 21 February 2024 • Accepted: 9 May 2024 •
Published Online: 14 June 2024

Bryce A. Primavera,[1,2,a] Saeed Khan,[1,2] Richard P. Mirin,[1] Sae Woo Nam,[1,b] and Jeffrey M. Shainline[1]

AFFILIATIONS

[1] National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305, USA
[2] Department of Physics, University of Colorado Boulder, 390 UCB, Boulder, Colorado 80309, USA

[a] **Author to whom correspondence should be addressed:** bryce.primavera@nist.gov
[b] Deceased.

ABSTRACT

The co-location of memory and processing is a core principle of neuromorphic computing. A local memory device for synaptic weight storage has long been recognized as an enabling element for large-scale, high-performance neuromorphic hardware. In this work, we demonstrate programmable superconducting synapses with integrated memories for use in superconducting optoelectronic neural systems. Superconducting nanowire single-photon detectors and Josephson junctions are combined into programmable synaptic circuits that exhibit single-photon sensitivity, memory cells with more than 400 internal states, leaky integration of input spike events, and 0.4 fJ programming energies (including cooling power). These results are attractive for implementing a variety of supervised and unsupervised learning algorithms and lay the foundation for a new hardware platform optimized for large-scale spiking network accelerators.

I. INTRODUCTION

Computing performance has been limited by the von Neumann bottleneck for decades.[1] These memory access challenges, in conjunction with the rise of memory-intensive deep learning applications, have led to a reexamination of computing architecture in recent years. Neuromorphic architectures modeled after biological neural systems are candidates for the next generation of artificial intelligence hardware. Computational and architectural motifs, such as distributed analog computation, highly interconnected communication networks, and co-location of memory and information processing, are key to the impressive performance of biological neural systems. These principles can serve as broad guidelines for hardware engineers.

Superconducting optoelectronic networks (SOENs) were introduced to maximize scalability while adhering to such biologically derived principles.[2,3] With this hardware, high-speed, low-power processing is performed with superconducting analog spiking neural circuits based on Josephson junctions (JJs). These superconducting neurons are embedded in a highly interconnected optical network that enables direct communication between each neuron and thousands of downstream synapses. Spiking events are encoded as few-photon pulses of light that are directly transmitted between an integrated light source at each neuron and single-photon sensitive detectors at each synaptic connection. This single-photon sensitivity manifests itself in the extreme fan-out capability of superconducting optoelectronic neurons by placing the physically minimal performance requirements on the light sources. Direct synaptic connections ensure communication latency is independent of network scale and activity up to systems of billions of neurons.[2]

Synapses for SOENs were realized in Ref. 4, enabled by the monolithic integration of superconducting-nanowire single-photon detectors (SPDs) with JJs. While those synapses demonstrated single-photon sensitivity and biologically relevant computations, such as leaky integration and tunable synaptic weights, the synaptic weights were defined with current biases generated off-chip. The

requirement of an independent current source for each synaptic weight is an unscalable solution that contradicts the principle of co-location of processing and memory.

Like other neuromorphic platforms, SOENs stand to greatly benefit from a local, multi-state memory that can be programmed for hardware-in-the-loop training or updated based on network activity for on-chip learning. While room-temperature synaptic memory technologies remain an intensely active research area centered on materials development and integration (be it memristive, ferroelectric, or phase-change materials),[5] suitable superconducting memories are a decades-old technology that require no changes to standard superconducting fabrication processes.[6,7] These "superconducting loop memories" store information as circulating currents trapped in superconducting loops. With identically zero resistance in the loop, the circulating current persists indefinitely. Furthermore, such memories permit high bit-depths, low programming energy, high endurance, and programming pulses easily produced by integrated JJ circuitry.[8,9] In this work, we adapt the synapses of Ref. 4 for integration with superconducting loop memory to demonstrate programmable single-photon sensitive optoelectronic synapses.

II. SOEN BACKGROUND

This section serves as a brief introduction to hardware of SOENs and the superconducting devices used in this work. Interested readers are encouraged to consult Refs. 2 and 3 for more details and scaling analyses. Figure 1 provides a schematic of a complete SOEN neuron.

Communication between neurons is performed in the optical domain (wavy arrows), while computation within each neuron is performed by superconducting analog electronics. When the signal contained within the neuron's soma exceeds an electronically adjustable threshold, an optical pulse is emitted by an integrated optical transmitter (labeled T), analogous to a biological action potential generated by an axon hillock. This optical transmitter consists of a light source[10] and a driving circuit.[11] The optical signal is then routed through a network of integrated photonic waveguides for neurons on the same chip or optical fibers for connections between chips in larger systems.

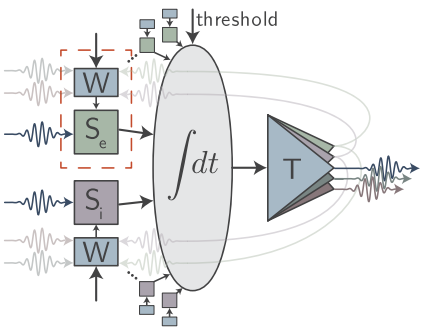

FIG. 1. Schematic of a mature SOEN neuron. The circuit demonstrated in this work is indicated with the red dashed box. Adapted from Ref. 2.

Each synapse is endowed with an SPD for transducing these faint optical pulses (on the order of a few photons) into electrical signals for further processing. The synapses are inductively coupled to the soma, and the sign of this coupling determines if they are excitatory (S_e) or inhibitory (S_i). A path toward coupling thousands of these synapses to the soma using dendritic compartments is presented in Ref. 12. Synaptic weights are maintained in superconducting memories contained within the circuit blocks labeled W. These weights determine the magnitude of magnetic flux that will be applied to the soma for each detected photon. The weight may be updated with either electrical programming signals (black arrow) or optical pulses from network activity (faded photonic signals). In this work, single-photon sensitive synapses are coupled to electronically programmable superconducting memories, realizing the components highlighted with the red dashed box in Fig. 1. For this demonstration, a pulsed laser is coupled from free space to the synapses to mimic the optical transmitter of an upstream neuron.

At the device level, the SPDs are superconducting nanowire detectors, which are an increasingly mature technology that have found wide application in recent years.[13] A number of features make them attractive for SOENs, including detection efficiencies exceeding 98%,[14] count rates approaching 1 GHz (although the detectors used in this work have count rates around 20 MHz),[15] high yield fabrication,[16,17] and waveguide compatibility.[18,19] Physically, the detectors are superconducting wires biased near their critical current. When a photon is absorbed, the wire transitions to a resistive state and redirects this bias through a parallel path.

The other key device for this demonstration is the Josephson junction.[20,21] JJs are the quintessential active element in superconducting electronics, which are used in SOEN hardware to perform a variety of bio-inspired operations, including leaky integration, thresholding, and synaptic weighting, which is the focus of this paper. A JJ is formed by two superconducting materials separated by a non-superconducting barrier. If the barrier is thin enough, superconducting current can tunnel from one side of the junction to the other. If the current through the junction exceeds a critical value, the device switches into a voltage state. This switching can occur on the order of picoseconds, making JJs particularly attractive for high-speed computation. Furthermore, Josephson electronics are commonly used with inductive coupling due to the interaction of magnetic fields with the superconducting wavefunction.[20] This is exploited repeatedly by the circuits described in Sec. III.

The cryogenic requirements of superconducting devices restrict SOENs to large-scale applications where cooling overhead will be acceptable.[22] The spike-based and analog nature of the hardware suggests that some of the first applications may be in temporal processing tasks where the continuous dynamics of the circuits prove especially useful. Possibilities include large-scale video analysis and high-speed control systems. In the longer term, SOENs are envisioned to benefit from the growing literature on bio-inspired machine learning, including local learning algorithms and spike-based computation.[23,24] Advances in these areas would open the technology to a more general set of applications and potentially provide a more energy efficient and scalable future for large-scale models.

III. CIRCUITRY

A. Superconducting loop memory

The phenomenon of flux quantization in superconducting loops is one of the most well-known manifestations of quantum mechanics at the macroscale.[25] The requirement that the wavefunction of the superconducting state be single valued ensures that the magnetic flux penetrating the loop must be an integer multiple of the magnetic flux quantum, Φ_0 (2.07×10^{-15} Wb). Equivalently, it is useful to think of the total flux as being composed of an integer number of flux quanta or fluxons, each carrying Φ_0 of flux. For an uninterrupted superconducting loop, the penetrating flux will for all time remain the same as when superconductivity was first established. However, if JJs are embedded in the loop, the amount of trapped flux can be changed in increments of a single fluxon. This is the basis of superconducting loop memory, where the number of trapped fluxons in the loop is used to define the state of the memory cell. Trapped fluxons are associated with an induced shielding current in the memory loop,

$$I_{\text{mem}} = \frac{N_\phi \Phi_0}{L_{\text{mem}}}, \tag{1}$$

where N_ϕ is the number of stored fluxons and L_{mem} is the loop inductance. This current will persist indefinitely, resulting in loop memory's exceptional retention times. In digital computing applications, it is common to employ a binary memory cell, representing a zero by the absence of trapped flux and a one by the presence of a single fluxon. In the present neural application, a many-state memory is desirable. In this work, we demonstrate three memory loops with N_ϕ = 8, 28, and 415, which we refer to as 3, 5, and 8 bit synapses, respectively.

While there are several variations of circuitry for programming the states of memory loops, we pursue the circuit in the lower part of Fig. 2(a) for its simplicity. In this case, the memory loop is

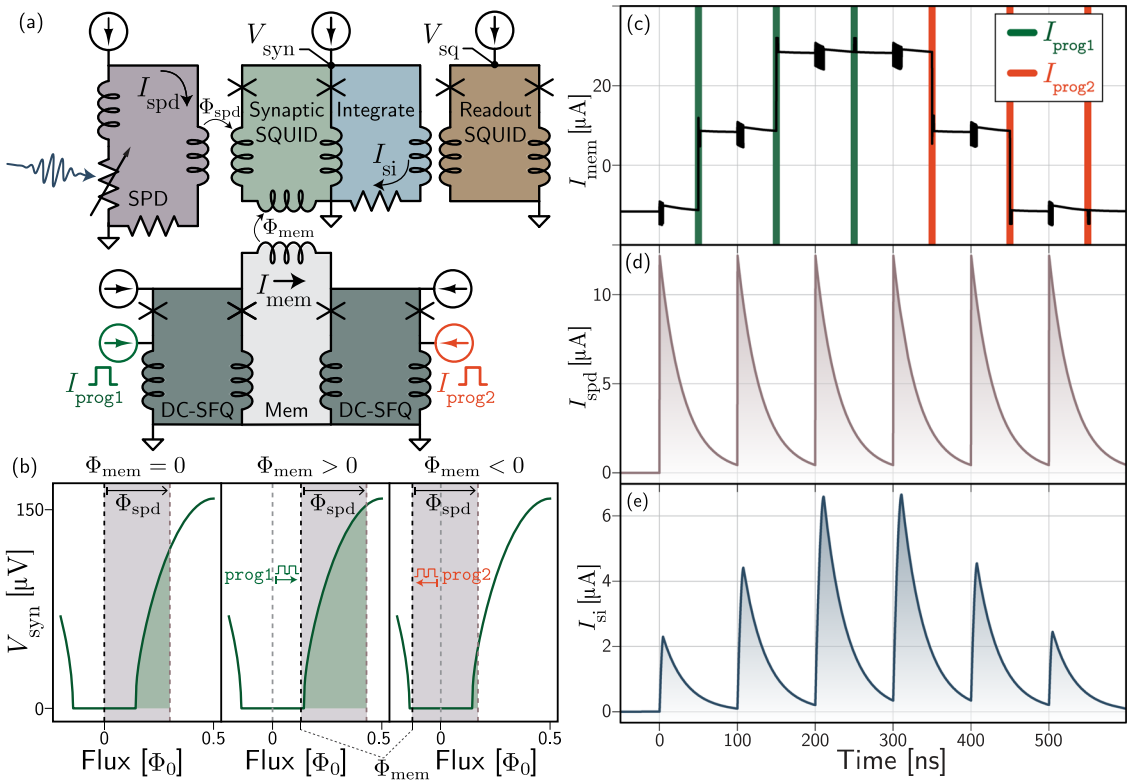

FIG. 2. Synapse concept and simulation. (a) Circuit diagram. Current I_{si} is added to the integration loop (blue) each time the SPD (purple) detects a photon. The amount of I_{si} produced is mediated by the synaptic SQUID (light green). A programmable memory loop (dark green and tan blocks) determines the quiescent point of the synaptic SQUID, which establishes the synaptic weight. The voltage across the readout SQUID (V_{sq}) is measured as a proxy for I_{si} in the experiments. (b) A section of the synaptic SQUID's flux-voltage transfer function is shown in green. The SQUID is initially biased below threshold. When the SPD detects a photon, the flux Φ_{spd} drives the SQUID into the active region of its response. Changing the state of the memory loop corresponds to shifting the initial flux penetrating the SQUID (Φ_{mem}). Higher values of Φ_{mem} result in the SPD driving the SQUID further into the voltage state and more current added to the integration loop. (c) Simulation of the stored current (I_{mem}) in a three-state memory loop. Unless the loop is saturated, each `prog1` pulse adds one fluxon to the loop and each `prog2` pulse subtracts one fluxon from the loop. (d) The current I_{spd} diverted away from the SPD each time a photon is detected. This waveform is determined by the optimal parameters for photon detection and is unaffected by changes to the synaptic weight. (e) The post-synaptic integrated current (I_{si}) at three different weights for six photon detection events, showing the effect of the memory loop. Note that the synaptic decay time is 30 ns in this simulation in order to present the SPD and integration loop dynamics on similar timescales. In the fabricated devices, the decay time is 6.25 μs.

inserted between two DC-SFQ circuit blocks. These DC-SFQ converters produce exactly one fluxon or single flux quantum (SFQ) each time the input current ($I_{\texttt{prog1}}$ or $I_{\texttt{prog2}}$) crosses a threshold set by the DC current bias.[20,25] By placing one on either side of the memory loop, fluxons can be added or subtracted one at a time by applying a programming pulse to the appropriate DC-SFQ converter. The operation of the DC-SFQ converters is independent of programming pulse width and largely independent of the precise programming pulse height, as long as the pulse amplitude exceeds the threshold set by the current bias (programming pulse height does affect the saturation level of the memory loop as described in the next section). The number of states available to the loop is ultimately limited by the loop inductance, with larger inductances supporting more states since the stored current per fluxon is inversely proportional to L_{mem}. In the data section, we demonstrate memories with three different inductances (154, 620 pH, and 22.5 nH designed values) to realize the three different storage capacities.

B. Synaptic circuits

In order to couple memory loops to synapses, the previous generation of synaptic circuits[4] underwent a major redesign [Fig. 2(a)]. In the new design, each synapse is based around a superconducting quantum interference device (SQUID) connected to an integration loop. A SQUID is a common device in superconducting electronics in which two Josephson junctions are current-biased in parallel (in this work, we refer specifically to the DC SQUID). The SQUID acts as a flux to voltage transducer, where the voltage across the device is a periodic function of the magnetic flux penetrating the loop. For these circuits, we restrict the range of inputs so that only one period of the response plays a role and the SQUID transfer function is a monotonic, nonlinear function of input flux, as shown in Fig. 2(b). The SPD is inductively coupled to the SQUID such that any photon detection event results in magnetic flux (Φ_{spd}) applied to the SQUID. Φ_{spd} is proportional to the current I_{spd} and decays with a time constant set by the detector (30 ns for this work). In this application, we bias the SQUID below the critical current so that it is in a state of zero voltage when no synaptic activity occurs. The flux coupled in by the SPD must exceed a bias-dependent threshold to activate the SQUID. When the SPD detects a photon, the SQUID will be driven into the voltage state where it produces a series of fluxons at a rate proportional to the voltage V_{syn}. These fluxons are stored as current I_{si} in the integration loop (blue). Adding a resistor to this integration loop causes the current to decay exponentially, resulting in leaky integrator behavior. I_{si} is analogous to the post-synaptic potential of a biological synapse and is the signal that will be fed into the neuron cell body (or dendritic tree).

The synaptic weighting mechanism demonstrated here functions by adjusting a flux bias to the synaptic SQUID. The memory loop couples flux (Φ_{mem}) into the SQUID, acting as an offset flux that can move the SQUID's quiescent point closer or further from its turn-on point. The total number of fluxons generated during the SPD response depends on the total flux coupled in, which is the sum of the contribution from the SPD (Φ_{spd}) and the contribution from the memory cell (Φ_{mem}). Incoming SPD flux will then drive the synaptic SQUID more or less strongly depending on where Φ_{mem} is placed relative to the turn-on point of the SQUID transfer function [Fig. 2(b)]. Φ_{mem} is programmed by placing an integer number of fluxons in the memory loop using a sequence of programming pulses to ports `prog1` and `prog2`.

In Figs. 2(c)–2(e), we simulate the current in various parts of the synapse in the time domain. In Fig. 2(c), I_{mem} is adjusted one fluxon at a time with a series of programming pulses. Recall I_{mem} is inductively coupled to the SQUID and therefore proportional to the flux offset that will ultimately determine the synaptic weight. We simulate a memory loop with only three states to illustrate the phenomenon of saturation. Saturation occurs when the memory loop has railed at either its maximum or minimum level of current. If one of the programming ports is pulsed repeatedly, current will be diverted from the DC-SFQ bias and into I_{mem} with each fluxon produced. Eventually, there is no longer enough bias current for the DC-SFQ converter to produce a fluxon in response to the next programming pulse, and the current in the memory loop will saturate. We observe positive and negative saturation after the third `prog1` pulse and third `prog2` pulse, respectively. Memory loop saturation is a beneficial behavior, as it is used to keep the synaptic SQUID operating in a useful range of its response. In Fig. 2(d), the SPD detects a series of six photons. Each of these detection events results in exactly the same amount of diverted current from the SPD, regardless of the state of the synapse. This allows the SPD to be biased for optimal detection efficiency at any synaptic weight and permits many SPDs to be biased in series. In contrast, the amplitude of I_{si} per detected photon is a function of the state of the memory loop and can be programmed as desired [Fig. 2(e)]. Note the increasing amplitude following each `prog1` pulse until saturation and the opposite behavior with each `prog2` pulse. This work used externally generated square programming pulses commensurate with hardware-in-the-loop training, but similar cells could be programmed directly with SPD pulses[26] for fully on-chip learning with local algorithms, such as spike-timing-dependent-plasticity.

IV. FABRICATION

Circuits were fabricated at the NIST Boulder Microfabrication Facility in a 15-layer process. The full process details are described in the Appendix of Ref. 4. The SPDs are made from MoSi and patterned into 200 nm-wide meandering wires using electron-beam lithography.[27,28] We also use MoSi as a high kinetic inductance material for the larger-valued inductors (namely, the synaptic integration inductor and the 22.5 nH 8 bit memory loop). The Josephson junctions are externally shunted Nb/a-Si/Nb tri-layers[29] with a target I_c of 100 μA. PdAu resistors were used for the JJ shunts, while Au was used for leak resistors in the synaptic integration loop. An integration time constant of 6.25 μs was targeted for all synapses.

A microscope image of the full 5 bit synapse is shown in Fig. 3(a). The synaptic SQUID, SPD, and memory loop are shown in Fig. 3(b). The synaptic SQUID uses a quadrupole configuration. This design both mitigates the effects of background magnetic fields varying over length-scales larger than the SQUID and provides a natural way to couple two independent input coils into the SQUID. The SPD drives the top input coil, while the memory loop drives the bottom. The SPD and a single DC-SFQ converter are shown in Figs. 3(c) and 3(d). Both JJs are visible in Fig. 3(d) (circles) along with external shunt resistors. No attempt was made to reduce the size of the circuits for these proof-of-principle experiments. The full synapse is

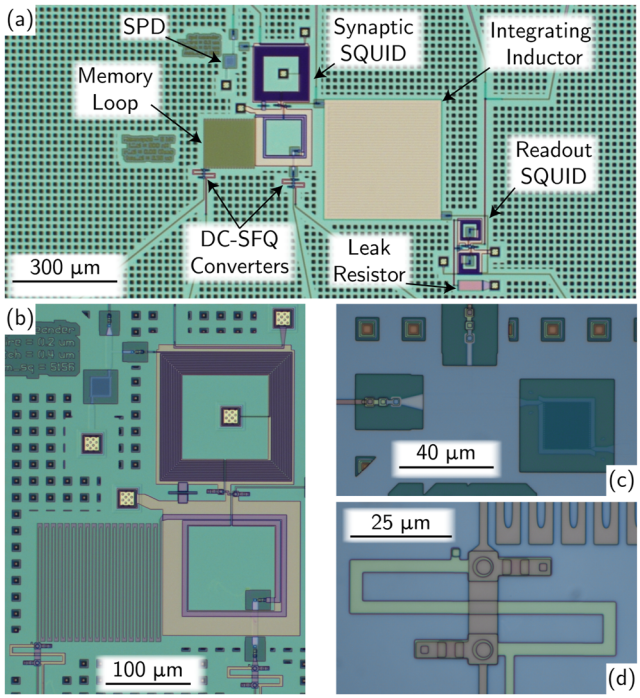

FIG. 3. Microscope images of the 5 bit synapse. (a) Full synapse. (b) Synaptic SQUID with SPD and memory loop inputs. (c) SPD (meander at right) along with vias and connections. (d) DC-SFQ converter for one side of the memory loop.

~840 × 700 μm^2. Similar synapses were estimated to occupy around 30 × 30 μm^2 in more advanced fabrication processes.[9]

V. EXPERIMENTAL CHARACTERIZATION

Measurements were performed by inductively coupling the synaptic integration loops to another SQUID, which we refer to as the readout SQUID [Fig. 2(a)]. This allows us to measure the voltage across the readout SQUID (V_{sq}) as a proxy for I_{si}. We note that V_{sq} is a somewhat distorted representation of I_{si}, particularly when significant current is stored in the integration loop, due to the nonlinear response of the readout SQUID. Nonetheless, this convenient readout mechanism allows us to measure changes in I_{si} at sub-microsecond timescales and is essentially identical to how we envision coupling these synaptic signals into dendritic and somatic structures in future work.[12,30] V_{sq} is amplified with a 60 dB room temperature amplifier and recorded on a 1 GHz oscilloscope. All plots report the amplified value of V_{sq}.

All measurements were performed at 2.3 K in a cryostat with a Gifford–McMahon cryocooler. The circuits were placed inside two concentric mu-metal shields to limit external magnetic noise. Nine coaxial cables were used for electrical input and output, and a single optical fiber was positioned above the test chip to flood-illuminate the entire sample. A 780 nm laser with 480 ps pulse width was used for optical input. The SPDs are not number-resolving detectors (except under special circumstances), and the laser pulse width is significantly shorter than the detector reset time (≈30 ns). Thus, even though the optical input is not in the single photon regime, the response of the SPD to a single laser pulse will not differ significantly from its single photon response. These responses are explored experimentally in Ref. 17 and in the supplementary material of Ref. 4, where we operated the previous generation of synapses under low-light conditions and confirmed single-photon sensitivity. Additionally, averaging was performed on all time traces to counteract electrical noise obscuring the microvolt signals. The number of averages is given in each figure caption. While readout noise inhibits the ability to test the variance of the synaptic response, the amplitude of an averaged trace represents the mean of the underlying probability distribution of synaptic weight.

There are two additional current bias lines not shown in Fig. 2(a) that are used to tune the two SQUIDs. We refer to these as the "addflux" biases. The addflux lines couple flux into the SQUIDs to set the initial operating points [the zero flux point in Fig. 2(b)]. These operating points were tuned by hand before measurement began. In the future, on-chip magnetic shielding could be added to ensure that all SQUIDs begin with zero flux penetrating the loop. Additionally, every synapse was initialized in its lowest weight state by repeatedly pulsing the `prog2` port before each measurement.

For the first experiment, we demonstrate how the post-synaptic response to a single optical pulse evolves with the programming history of the memory loop. In Fig. 4(a), we initially alternate between pulsing the laser and applying programming pulses to `prog1` for the 3 bit synapse. The post-synaptic current in the synaptic integration loop is allowed to decay to zero between successive laser pulses. We see that V_{sq} increases following each `prog1` programming pulse as desired. We then cease pulsing `prog1` and pulse the laser three times. As seen in Fig. 4(a), these three pulses are nearly identical in height, confirming that the memory loop is indeed retaining its programmed state. We then begin pulsing `prog2` between laser pulses and see that the synaptic weight can be reduced one step at a time before the memory loop saturates at its lowest level after eight pulses. The experiment is repeated for the 5 bit (620 pH memory loop) synapse in Fig. 4(b). We witness the same qualitative behavior as with the 3 bit synapse, but have significantly more states. `prog1` was pulsed 35 times in this experiment, although we observe that the post-synaptic height stops changing after about 28 pulses. This is due to the memory loop saturating at its maximum level slightly earlier than the designed 32 fluxons. It similarly takes about 28 `prog2` pulses to bring the synapse back to the minimal weight state, as expected.

In Fig. 5, we demonstrate that the circuits also exhibit an integrating ability inspired by biological synapses.[31] If the laser is pulsed at a high enough frequency such that I_{si} does not decay fully between pulses, multiple detection events can be integrated over time. The time constant of the leak is determined by the L/R value of the synaptic integration loop. Although not the subject of this study, the synaptic integration times can be engineered across at least four orders of magnitude (hundreds of nanoseconds to several milliseconds), as shown in Ref. 4. Figure 5(a) shows the 3 bit synapse responding to 15 laser pulses arriving at a 500 kHz frequency. Each trace corresponds to a different initialization of the memory loop. Before the laser is turned on, the memory loop is given a fixed number of `prog1` pulses (0–10 in this case). We observe the synaptic response growing with each additional programming pulse until saturation at 8 pulses (the 8, 9, and 10 programming pulse traces are

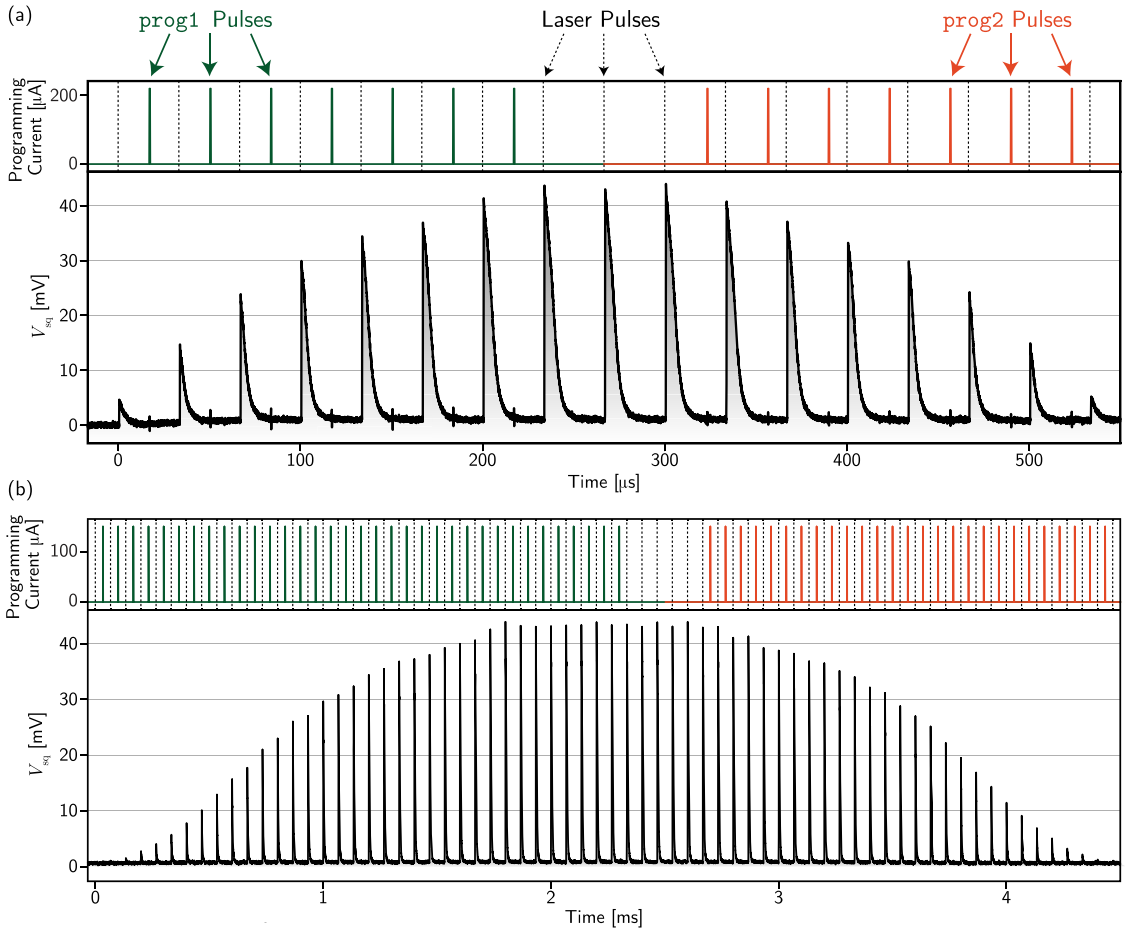

FIG. 4. Synaptic response to individual laser pulses evolving with programming history. (a) 3 bit synapse. The alternating sequence of programming and laser pulses is shown in the top plot, while measured V_{sq} data are shown in the bottom plot. Note how the three V_{sq} responses following the last `prog1` pulse maintain approximately the same peak value when no additional programming pulses are generated. (b) 5 bit synapse. The post-synaptic response rises with each `prog1` pulse until the memory loop is saturated after about 28 pulses. The V_{sq} traces in both (a) and (b) are averaged 1000 times.

on top of each other). This experiment was repeated with a laser pulse frequency of 1.1 MHz (b), where we observe a higher synaptic response for each programming condition than in (a), as expected from the leaky integrator.

In Figs. 5(c) and 5(d), the synaptic response is plotted as a function of the rate of incoming laser pulses (c) and the number of laser pulses in an input pulse train (d). In Fig. 5(c), the number of laser pulses is fixed at 100, but the frequency of those input pulses is swept from 100 kHz to 10 MHz. Each data point corresponds to the peak value of the synaptic response under those conditions [i.e., the peak value of a single trace of the type in (a) and (b)]. Once again, each curve corresponds to a different weight initialization (0–10 `prog1` pulses). The demonstrated sensitivity to frequency is of particular interest in burst- and rate-coded applications. The complimentary measurement is represented in Fig. 5(d), where the frequency of laser pulses is fixed at 10 MHz, but the number of pulses is varied from 1 to 100. The eventual leveling off is characteristic of a leaky integrator reaching steady-state for sufficiently long pulse trains. In both (c) and (d), the ability of the memory loop to tune the synaptic response curve is evident. Figures 5(e)–5(h) show the same data for the 5 bit synapse. In these plots, the number of `prog1` pulses is varied between 0 and 35 pulses. In (e) and (f), all 35 curves are plotted, but every fifth trace is shown with a darker line stroke for clarity. We see qualitatively similar data to the 3 bit case, but with much higher synaptic weight resolution, as expected.

The synapses presented here benefit from the ability of the DC-SFQ programming circuits to operate at timescales shorter than the synaptic integration dynamics. All experiments utilized programming pulse widths on the order of 100 ns. This is significantly shorter than the 6.25 μs synaptic integration time and allows the synaptic weight to be changed dynamically, even while the synapse contains signal in its integration loop. This is illustrated in Fig. 6 for the 3 bit synapse. The laser is pulsed with a frequency of 700 kHz. A series of programming pulses are input into the memory loop, and

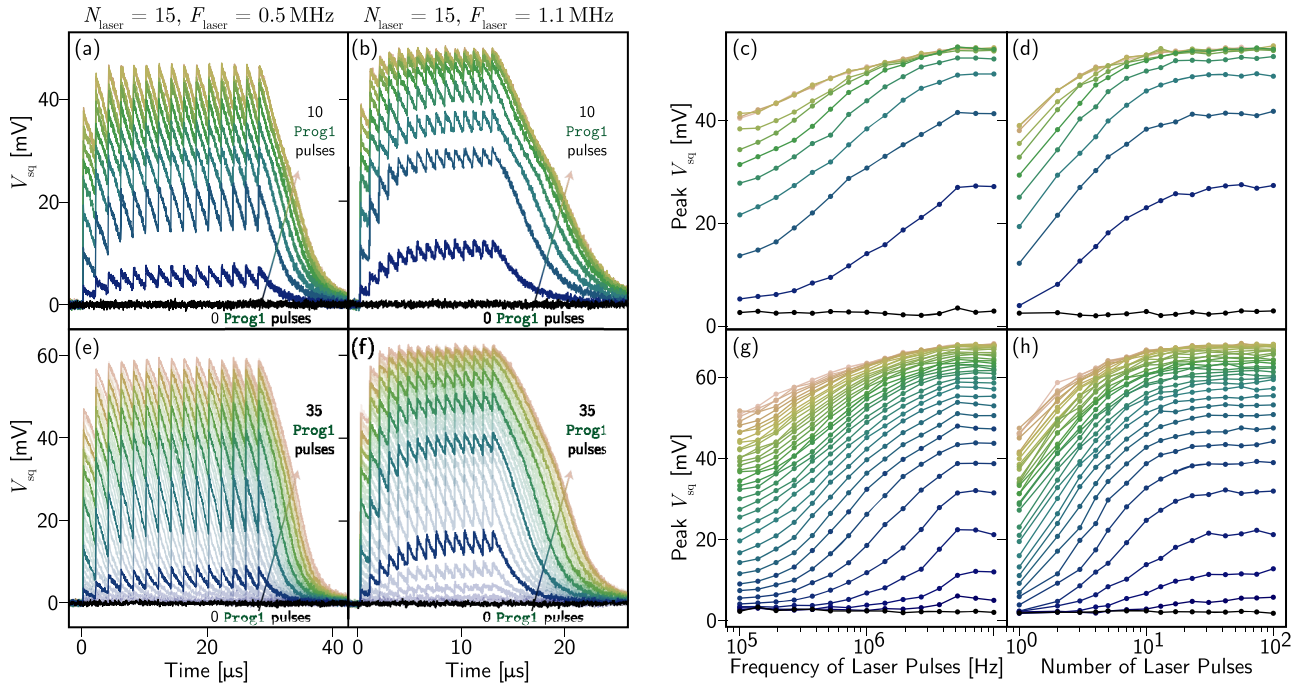

FIG. 5. Integrating ability for the 3 (a)–(d) and 5 (e)–(h) bit synapses. (a) Response to 15 laser pulses arriving at 500 kHz. Each trace corresponds to a different number of programming pulses sent into the memory loop at the beginning of the measurement. 500 averages. (b) Same as (a) except that the laser pulse rate is increased to 1.1 MHz. (c) Laser pulse frequency transfer function. The peak value of V_{sq} is plotted for different frequencies of optical pulses. The pulse number is fixed at 100 pulses. 200 averages. (d) Laser pulse number transfer function. The peak value of V_{sq} is plotted for numbers of optical pulses. The pulse frequency is fixed at 10 MHz. 200 averages. (e)–(h) Repetition of these plots for the 5 bit synapse. In (e) and (f), every fifth trace is bolded for clarity.

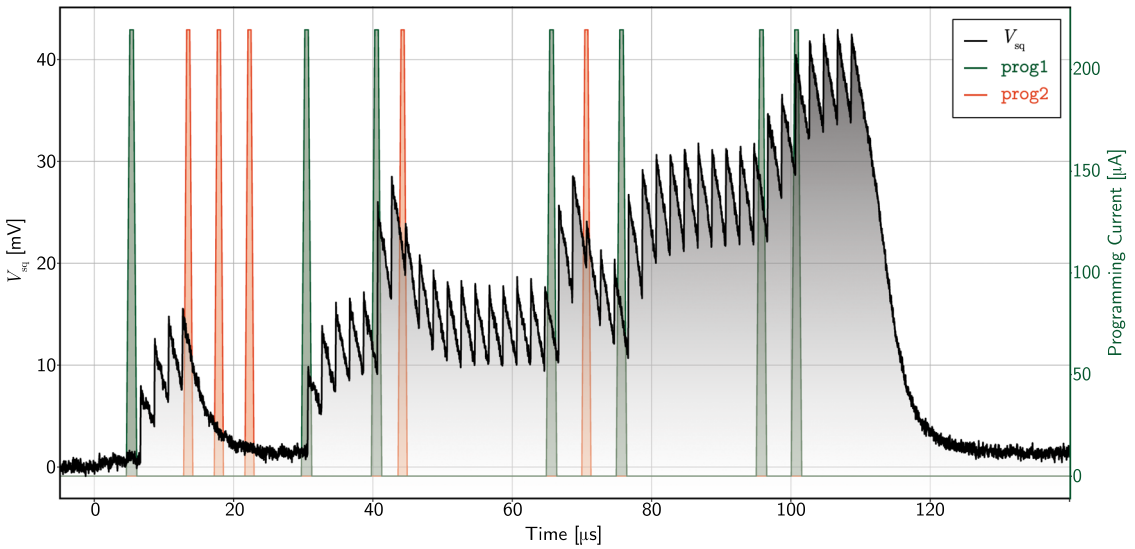

FIG. 6. Dynamically changing the synaptic weight of the 3 bit synapse while the synapse continues to receive and integrate signal. 500 averages. There is no observable cross-talk from the programming signals on the integrated synaptic current, allowing synaptic weights to be changed "on-the-fly."

we see the synaptic weight changing with every programming pulse (unless the loop is already saturated, as exhibited by the second and third prog2 pulses). There is no observable cross-talk between the programming pulses and the integrated current, as the integration loop is isolated from the memory circuitry by the synaptic SQUID. This ability to change the weight dynamically is promising for future implementations of short term plasticity and homeostatic mechanisms. Furthermore, memory updates can be completed in less time than the minimum inter-spike interval expected in SOEN hardware (≈30 ns). Future implementations of on-chip learning in a network

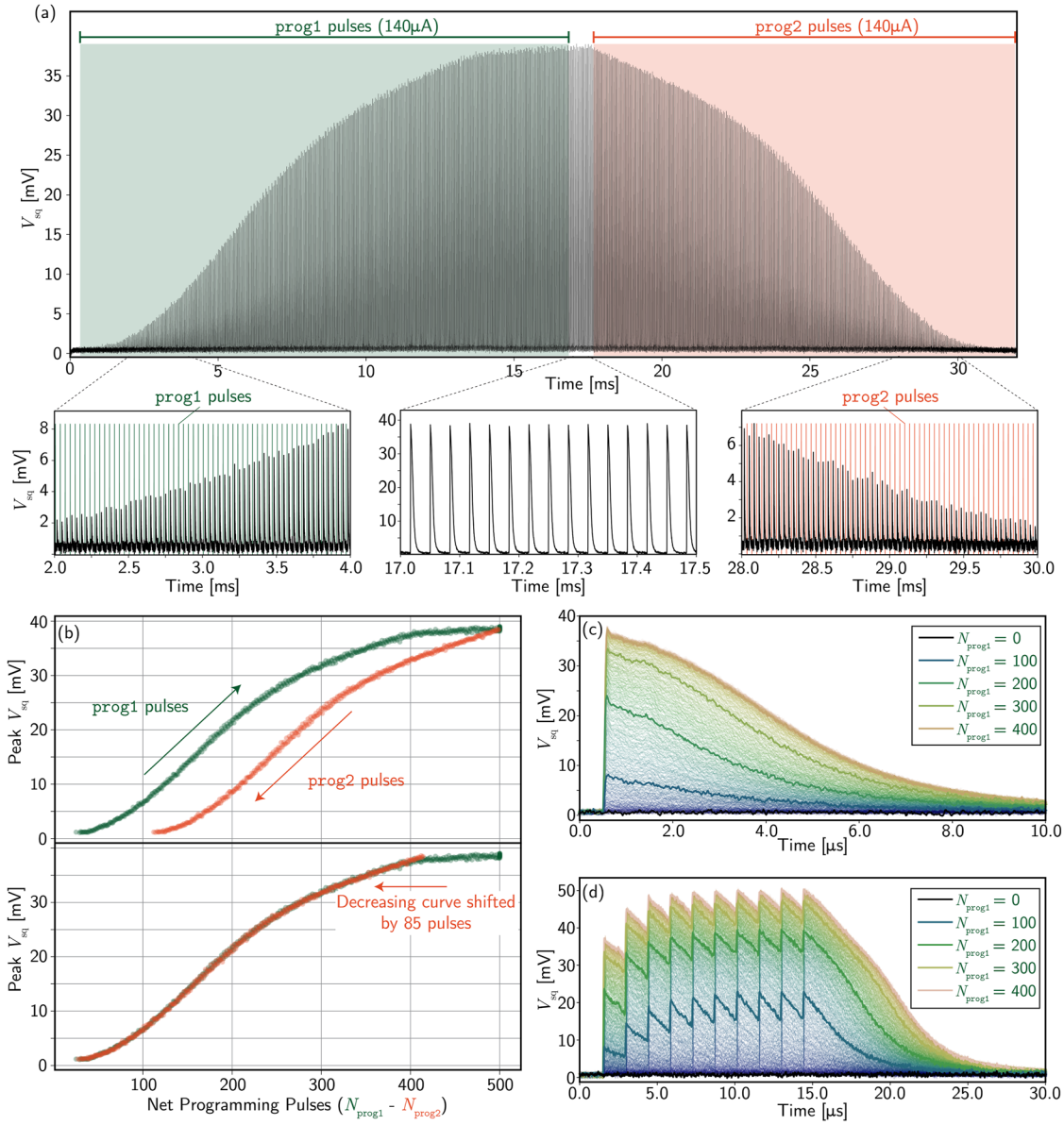

FIG. 7. 8 bit device. (a) Alternating between laser and programming pulses, analogous to Fig. 4. Zoom-ins below resolve individual programming pulses and synaptic responses. Laser pulses are omitted for clarity. The prog1 port is driven 500 times. 2500 averages. (b) Top: Peak V_{sq} as a function of the net number of programming pulses ($N_{prog1} - N_{prog2}$) applied to the memory loop. Green points are derived from the peaks while prog1 is pulsed, while red points are derived from pulsing prog2 following 500 prog1 pulses. Bottom: Same as above, except that the red curve is shifted 85 pulses left. We see that the rising path and falling path are nearly symmetric. The 85 pulse offset implies that the last 85 prog1 pulses occurred after the memory loop reached saturation and that there are ~415 states in the memory loop. (c) Response to a single laser pulse following different initializations of the memory loop (0–500 prog1 pulses). The 0, 100, 200, 300, and 400 pulse traces are bolded for clarity. 1000 averages. (d) Same as (c) except responding to a pulse train of 10 laser pulses at 700 kHz.

are therefore unlikely to ever be a bottleneck, even in the extreme case of a weight update after every synaptic event.

In Fig. 7, we present the results of the synapse coupled to the largest memory loop (≈22.5 nH). We refer to this as the 8 bit loop, although we estimate that there are actually over 400 internal states. In the post-synaptic response, the difference between adjacent states is less than the noise floor of our measurement, but we can discern the same qualitative behavior as in the 3 and 5 bit variants. In Fig. 7(a), we perform the same experiment as Fig. 4, alternating between individual laser and programming pulses. We pulse the potentiating port 500 times before reducing the synaptic weight back to zero. Zoomed-in portions show the synaptic weight increasing with the `prog1` pulses, remaining constant without any programming signals, and decreasing with `prog2` pulses. In Fig. 7(b), we plot the peak values of V_{sq} in part (a) following each laser pulse as a function of the net number of programming pulses, $N_{prog1} - N_{prog2}$. We break the plot into two parts: a green rising curve resulting from the region of `prog1` pulses and a red decreasing curve from the later region of `prog2` pulses. These two curves are nearly symmetric, but shifted by 85 pulses as shown in the bottom panel of Fig. 7(b). The symmetry between the curves is consistent with proper DC-SFQ operation and suggests that both programming ports are operating correctly. The 85-pulse offset between the two curves implies that 85 of the `prog1` pulses came after the memory loop had already reached saturation. This allows us to estimate that the memory loop has a capacity of ~500 − 85 = 415 states or 8.7 bits. This is also in visual agreement with the apparent flattening of the green curve after about 415 `prog1` pulses. The slight upward slope in the region between 415 and 500 pulses is likely due to a small amount of integration of multiple laser pulses occurring at high synaptic weights. Figures 7(c) and 7(d) show the synapse responding to a single laser pulse and a 700 kHz train of 10 laser pulses, respectively. While noise obscures the small differences between states, the behavior is as expected.

To quantify the volatility of the memory storage mechanism, the retention of one of the synapses was measured over a period of 48 h. In principle, the current stored in the memory loops should remain unchanged for as long as the circuit remains below the superconducting transition temperature. We test this on the 3 bit synapse by first taking two reference traces with one and three `prog1` pulses. We then re-initialize the synapse with two `prog1` pulses. No other programming pulses were applied for the rest of the experiment. The synapse was driven once an hour with bursts of 15 laser pulses at a frequency of 700 kHz. Once again, averaging was required to reduce readout noise. 500 such bursts were applied (5 kHz burst frequency) and averaged to accurately measure the synaptic weight at each hour. The results are represented in Fig. 8(a). The two gray traces correspond to the $N_{prog1} = 1$ and $N_{prog1} = 3$ responses at the beginning of the experiment. There is no visible difference in the response over time, and it never comes close to approaching either of the two adjacent states. In Fig. 8(b), we plot the χ^2 value of each trace as a function of time after programming. The following equation is used to calculate the statistic:

$$\chi^2(T) = \frac{\sum_{n=1}^{N_p} (V_{sq,T}(n) - \overline{V_{sq}}(n))^2}{\sum_{n=1}^{N_p} \overline{V_{sq}}(n)^2}, \quad (2)$$

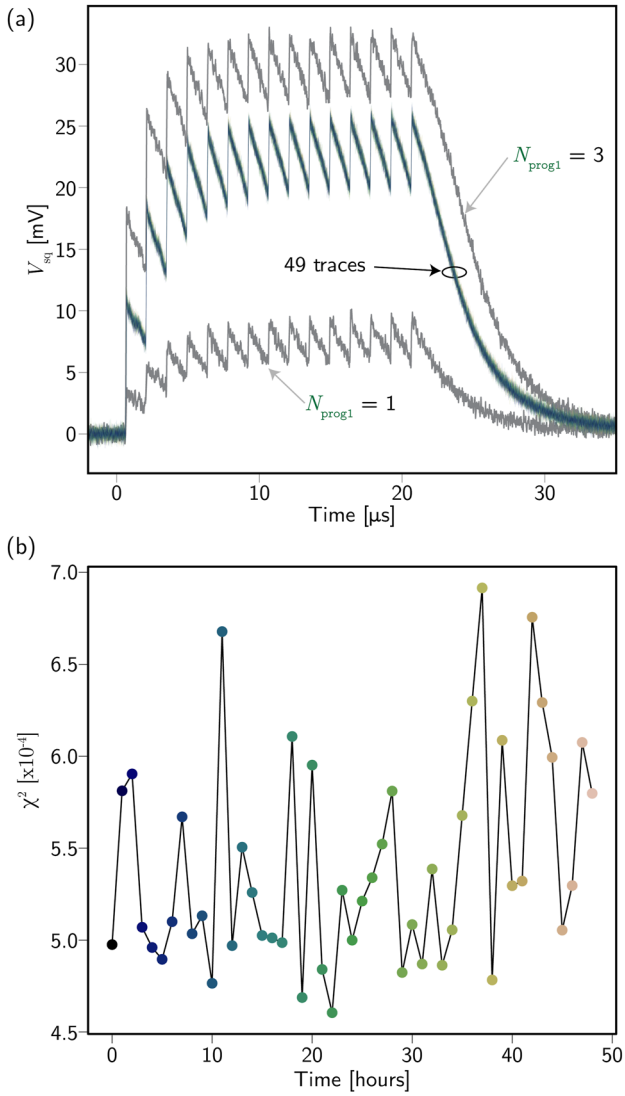

FIG. 8. Stability of the 3 bit synapse. (a) The gray traces correspond to the $N_{prog1} = 1$ and $N_{prog1} = 3$ states at the beginning of measurement period. All other traces correspond to the $N_{prog1} = 2$ state each taken one hour apart after a single initialization. 500 averages. (b) χ^2 value [Eq. (2)] of each trace as a function of time after initialization.

where $V_{sq,T}(n)$ is the value of the n^{th} point in the trace taken T hours after programming, N_p is the number of data points in each trace, and $\overline{V_{sq}}$ is the average of all 49 traces. There is no discernible trend over the 48 h measurement period, and we expect that much longer retention times are possible. As a point of reference, we anticipate superconducting optoelectronic neurons to achieve spiking rates beyond 20 MHz. Maximum spike rates in the brain are about 1 kHz (for chattering neurons; most pyramidal neurons do not exceed gamma bursts of 80 Hz). Using 20 MHz/1 kHz as a scale factor, 48 h of stable memory in this hardware is roughly commensurate with 110 years of human brain activity.

VI. DISCUSSION

We have demonstrated programmable superconducting single-photon optoelectronic synapses coupled to memory cells with over 400 states. For comparison, Intel's state-of-the-art digital neuromorphic chip Loihi uses 9-bit precision.[32] The programming energy for our synapses is $\sim 2\Phi_0 I_c \approx 0.4$ aJ. Accounting for cooling would inflate this to around 0.4 fJ. Assuming 1% light-source efficiency, each presynaptic photon will require about 13 fJ of energy. Programming energies are therefore unlikely to be a major contributor to the SOEN energy budget. Additionally, since the switching speeds of Josephson junctions (≈10 ps) are far faster than the reset times of the detectors (≈30 ns), these synapses lend themselves to experimentation with sophisticated bio-inspired "always-on" plasticity mechanisms that would be too costly (in terms of either energy or time) for other systems.

While loop memory has many attractive features, it has long been considered a weakness of superconducting digital computing due to its low area density.[33] This is a fair criticism in the digital computing space, but much less so in SOEN neuromorphic hardware. Since every synapse already requires a SQUID, the addition of the memory loop is not a major contribution to the total area. While no effort was made to reduce the area in this work, scaling analyses suggest that a single SOEN synapse with integrated memory should be realizable within a $30 \times 30\ \mu m^2$ area in more advanced fabrication processes where the different components fabricated here are stacked on top of each other.[9] There is little motivation to shrink this area further due to the size of passive photonic components on other fabrication layers. Over one million neurons and one billion synapses embedded in a network with a biologically realistic average path length can still be expected to fit on a 300 mm wafer.[9] An added advantage of superconducting electronics is that both the active computational elements (JJs) and the photon detectors (SPDs) can be monolithically fabricated on multiple layers.[34–36] Similar multi-layered fabrication of CMOS or semiconductor photodetectors is difficult due to the requirement of high-temperature steps for dopant activation, which cannot be performed after metal interconnects have been deposited. While it is conceivable (although not at all obvious) that neuromorphic hardware of similar complexity based on silicon microelectronics could be more dense, the overall system scalability is far more limited due to communication bottlenecks. Furthermore, superconducting loop memory outperforms emerging memristive technologies in many other metrics. While there is a wide range in reported performance of memristive devices,[37] programming energies are typically on the order 100 fJ/bit (more than two orders of magnitude greater than superconducting loop memory), write times are around 10 ns (three orders of magnitude slower than superconducting memories), less than 6 bit precision is typical,[38] and there are ongoing efforts to improve endurance and variability. Improved performance can be gained through more sophisticated programming protocols,[39] but the programming time and cost of such schemes are best-suited for inference applications rather than large-scale training or online learning systems. Thus, where superconducting digital computing is weak relative to semiconductors with regard to memory, it appears to be strong in the domain of neural memory.

Although the demonstrated synapses are a marked improvement over the previous generation, there are several areas that require further research. First, future work with on-chip magnetic shielding and a cryostat with improved isolation could better elucidate the limits of bit precision in the devices. For now, noise makes it difficult to assess how well adjacent synaptic weights are separated in the I_{si} domain. However, we note that the present devices are operated with a thermal noise parameter of $\Gamma = (2\pi kT)/(\Phi_0 I_c) \approx 9 \times 10^{-4}$, suggesting that the observed noise is not thermal in origin and likely external to the circuits.[21] Second, there was no effort in this work to linearize the synaptic weighting. While the quantization of magnetic flux ensures that the stored current in each memory cell is nearly linear with its programming history, the amount of current added to the integration loop per detected photon is a function of the SQUID response, which is nonlinear. However, there has long been interest in improving the linearity of SQUID responses for sensing applications, and only slightly more complicated three-JJ SQUID designs have shown nearly linear responses.[40] It may also be possible to design learning protocols where the nonlinearity does not pose an issue. Third, there is the key question of volatility. Superconducting loop memories do not require any power to maintain their state. Indeed, all the current biases to these synapses can be toggled off and on and the synapse will come back online in its previously programmed state. However, the temperature must remain below the critical temperature (T_c) of the superconducting materials (~9 K) or all information will be lost. There will likely be maintenance situations or power failures where the system must be warmed above T_c, and the ability to store the synaptic weights at room temperature would be beneficial. One possibility is to have a readout mechanism at every memory cell that could be used to measure the state of each memory cell and save this information digitally. The weight matrix could then be re-uploaded into the network once the system is cold again.

Local, programmable synaptic memory opens up a wide variety of potential applications. The present synapses could be useful in non-cognitive applications of spiking networks wherein a weight-adjacency matrix representing a specific problem is programmed into the network during an initialization phase. The network behavior is then allowed to evolve in time, with the resulting network dynamics encoding the solution to the problem. This scheme has been used to solve optimization problems and systems of differential equations.[41] In terms of artificial intelligence, the present synapses are already well-suited for inference applications utilizing a matrix of pre-learned weights. They are also amenable to hardware-in-the-loop training, in which a standard digital computer generates the appropriate programming pulses from observing the network's output and internal variables. While such training could be performed at a small scale with memory-less synapses, this scheme would require an independent current bias for every weight in the network, which is infeasible for large systems due to the heat load incurred by electrical lines. In contrast, these programmable synapses could be interfaced with cryo-CMOS control circuitry and an addressing system for large-scale programming using a limited number of lines between the cryostat and room temperature. The ultimate goal, however, is to remove the digital computer from the training loop. This will require the development of learning algorithms and plasticity circuits to update the synaptic memories using primarily local information.

We have developed a simulation framework for SOEN hardware[42] and recently simulated a neural network utilizing this synapse

design to solve a nine-pixel image classification problem.[43] That work employed a locally competitive architecture and introduced a learning rule that uses the signals stored in the synaptic integration loops along with a global error signal to update inductively coupled superconducting memory cells, such as those demonstrated here. Additionally, plasticity circuits implementing a spike-timing-dependent plasticity update rule with very similar memory elements were presented in Ref. 2. Such plasticity circuits must receive optical, rather electrical inputs, since the native spiking behavior in SOENs is performed in the optical domain. The single-photon-to-single-fluxon converters presented in Ref. 26 are an important step in this direction and can be considered optically programmable memory cells in their own right. Future work will focus on developing these plasticity circuits, investigating more sophisticated dendritic processing,[30] and demonstrating full superconducting optoelectronic neurons by integrating these synapses with on-chip light sources and passive integrated-photonic interconnection networks.

SUPPLEMENTARY MATERIAL

See supplementary material for additional details about the experimental setup, tests of the DC-SFQ converters, and further characterization of the synaptic responses.

ACKNOWLEDGMENTS

We thank Dr. Michael Schneider for useful feedback on the manuscript. This work was made possible by the institutional support from the National Institute of Standards and Technology (Award No. 70NANB18H006).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Bryce A. Primavera: Conceptualization (equal); Data curation (equal); Formal analysis (lead); Investigation (equal); Methodology (equal); Visualization (lead); Writing – original draft (lead); Writing – review & editing (lead). **Saeed Khan**: Data curation (equal); Investigation (equal); Methodology (equal); Writing – review & editing (supporting). **Richard P. Mirin**: Funding acquisition (equal); Project administration (equal); Resources (equal). **Sae Woo Nam**: Funding acquisition (equal); Project administration (equal); Resources (equal). **Jeffrey M. Shainline**: Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are publicly available via Figshare (https://figshare.com/s/893002d96b6ba54839c4).

REFERENCES

[1] A. Sebastian, M. Le Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," Nat. Nanotechnol. **15**, 529–544 (2020).
[2] J. M. Shainline, S. M. Buckley, A. N. McCaughan, J. T. Chiles, A. Jafari Salim, M. Castellanos-Beltran, C. A. Donnelly, M. L. Schneider, R. P. Mirin, and S. W. Nam, "Superconducting optoelectronic loop neurons," J. Appl. Phys. **126**, 044902 (2019).
[3] J. M. Shainline, "Optoelectronic intelligence," Appl. Phys. Lett. **118**, 160501 (2021).
[4] S. Khan, B. A. Primavera, J. Chiles, A. N. McCaughan, S. M. Buckley, A. N. Tait, A. Lita, J. Biesecker, A. Fox, D. Olaya *et al.*, "Superconducting optoelectronic single-photon synapses," Nat. Electron. **5**, 650–659 (2022).
[5] G. W. Burr, R. M. Shelby, A. Sebastian, S. Kim, S. Kim, S. Sidler, K. Virwani, M. Ishii, P. Narayanan, A. Fumarola *et al.*, "Neuromorphic computing using non-volatile memory," Adv. Phys.: X **2**, 89–124 (2017).
[6] J. W. Crowe, "Trapped-flux superconducting memory," IBM J. Res. Dev. **1**, 294–303 (1957).
[7] S. Tahara and Y. Wada, "A vortex transitional NDRO Josephson memory cell," Jpn. J. Appl. Phys. **26**, 1463 (1987).
[8] A. Schegolev, N. Klenov, I. Soloviev, and M. Tereshonok, "Learning cell for superconducting neural networks," Supercond. Sci. Technol. **34**, 015006 (2020).
[9] B. A. Primavera and J. M. Shainline, "Considerations for neuromorphic supercomputing in semiconducting and superconducting optoelectronic hardware," Front. Neurosci. **15**, 732368 (2021).
[10] S. Buckley, J. Chiles, A. N. McCaughan, G. Moody, K. L. Silverman, M. J. Stevens, R. P. Mirin, S. W. Nam, and J. M. Shainline, "All-silicon light-emitting diodes waveguide-integrated with superconducting single-photon detectors," Appl. Phys. Lett. **111**, 141101 (2017).
[11] A. N. McCaughan, V. B. Verma, S. M. Buckley, J. Allmaras, A. Kozorezov, A. Tait, S. Nam, and J. Shainline, "A superconducting thermal switch with ultrahigh impedance for interfacing superconductors to semiconductors," Nat. Electron. **2**, 451–456 (2019).
[12] B. A. Primavera and J. M. Shainline, "An active dendritic tree can mitigate fan-in limitations in superconducting neurons," Appl. Phys. Lett. **119**, 242601 (2021).
[13] I. Holzman and Y. Ivry, "Superconducting nanowires for single-photon detection: Progress, challenges, and opportunities," Adv. Quantum Technol. **2**, 1800058 (2019).
[14] D. V. Reddy, R. R. Nerem, S. W. Nam, R. P. Mirin, and V. B. Verma, "Superconducting nanowire single-photon detectors with 98% system detection efficiency at 1550 nm," Optica **7**, 1649–1653 (2020).
[15] J. Münzberg, A. Vetter, F. Beutel, W. Hartmann, S. Ferrari, W. H. Pernice, and C. Rockstuhl, "Superconducting nanowire single-photon detector implemented in a 2D photonic crystal cavity," Optica **5**, 658–665 (2018).
[16] E. E. Wollman, V. B. Verma, A. E. Lita, W. H. Farr, M. D. Shaw, R. P. Mirin, and S. Woo Nam, "Kilopixel array of superconducting nanowire single-photon detectors," Opt. Express **27**, 35279–35289 (2019).
[17] S. M. Buckley, A. N. Tait, J. Chiles, A. N. McCaughan, S. Khan, R. P. Mirin, S. W. Nam, and J. M. Shainline, "Integrated-photonic characterization of single-photon detectors for use in neuromorphic synapses," Phys. Rev. Appl. **14**, 054008 (2020).
[18] S. Ferrari, C. Schuck, and W. Pernice, "Waveguide-integrated superconducting nanowire single-photon detectors," Nanophotonics **7**, 1725–1758 (2018).
[19] J. M. Shainline, S. M. Buckley, N. Nader, C. M. Gentry, K. C. Cossel, J. W. Cleary, M. Popović, N. R. Newbury, S. W. Nam, and R. P. Mirin, "Room-temperature-deposited dielectrics and superconductors for integrated photonics," Opt. Express **25**, 10322–10334 (2017).
[20] A. M. Kadin, *Introduction to Superconducting Circuits* (Wiley-Interscience, 1999).
[21] J. Clarke and A. I. Braginski, *The SQUID Handbook* (Wiley Online Library, 2004), Vol. 1.
[22] D. S. Holmes, A. L. Ripple, and M. A. Manheimer, "Energy-efficient superconducting computing—Power budgets and requirements," IEEE Trans. Appl. Supercond. **23**, 1701610 (2013).
[23] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," Neural Networks **111**, 47–63 (2019).

[24]A. Ororbia, A. Mali, A. Kohan, B. Millidge, and T. Salvatori, "A review of neuroscience-inspired machine learning," arXiv:2403.18929 (2024).
[25]T. V. Duzer and C. Turner, in *Principles of Superconductive Devices and Circuits*, 2nd ed. (Prentice Hall, 1998).
[26]S. Khan, B. Primavera, R. P. Mirin, S. W. Nam, and J. Shainline, "Monolithic integration of superconducting-nanowire single-photon detectors with Josephson junctions for scalable single-photon sensing," Supercond. Sci. Technol. **37**, 035011 (2024).
[27]V. B. Verma, B. Korzh, F. Bussieres, R. D. Horansky, S. D. Dyer, A. E. Lita, I. Vayshenker, F. Marsili, M. D. Shaw, H. Zbinden *et al.*, "High-efficiency superconducting nanowire single-photon detectors fabricated from MoSi thin-films," Opt. Express **23**, 33792–33801 (2015).
[28]A. E. Lita, V. B. Verma, J. Chiles, R. P. Mirin, and S. W. Nam, "Mo_xSi_{1-x} a versatile material for nanowire to microwire single-photon detectors from UV to near IR," Supercond. Sci. Technol. **34**, 054001 (2021).
[29]D. Olaya, M. Castellanos-Beltran, J. Pulecio, J. Biesecker, S. Khadem, T. Lewitt, P. Hopkins, P. Dresselhaus, and S. Benz, "Planarized process for single-flux-quantum circuits with self-shunted Nb/Nb_xSi_{1-x}/Nb Josephson junctions," IEEE Trans. Appl. Supercond. **29**, 1101708 (2019).
[30]J. M. Shainline, "Fluxonic processing of photonic synapse events," IEEE J. Sel. Top. Quantum Electron. **26**, 7700315 (2019).
[31]J. C. Magee, "Dendritic integration of excitatory synaptic input," Nat. Rev. Neurosci. **1**, 181–190 (2000).
[32]M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," IEEE Micro **38**, 82–99 (2018).
[33]H. Hilgenkamp, "Josephson memories," J. Supercond. Novel Magn. **34**, 1621–1625 (2021).
[34]T. Ando, S. Nagasawa, N. Takeuchi, N. Tsuji, F. China, M. Hidaka, Y. Yamanashi, and N. Yoshikawa, "Three-dimensional adiabatic quantum-flux-parametron fabricated using a double-active-layered niobium process," Supercond. Sci. Technol. **30**, 075003 (2017).
[35]S. K. Tolpygo, V. Bolkhovsky, R. Rastogi, S. Zarr, A. L. Day, E. Golden, T. J. Weir, A. Wynn, and L. M. Johnson, "Planarized fabrication process with two layers of sis Josephson junctions and integration of SIS and SFS π-junctions," IEEE Trans. Appl. Supercond. **29**, 1101208 (2019).
[36]V. B. Verma, F. Marsili, S. Harrington, A. E. Lita, R. P. Mirin, and S. W. Nam, "A three-dimensional, polarization-insensitive superconducting nanowire avalanche photodetector," Appl. Phys. Lett. **101**, 251114 (2012).
[37]F. Zahoor, T. Z. Azni Zulkifli, and F. A. Khanday, "Resistive random access memory (RRAM): An overview of materials, switching mechanism, performance, multilevel cell (MLC) storage, modeling, and applications," Nanoscale Res. Lett. **15**, 90 (2020).
[38]T.-H. Kim, S. Kim, K. Hong, J. Park, Y. Hwang, B.-G. Park, and H. Kim, "Multilevel switching memristor by compliance current adjustment for off-chip training of neuromorphic system," Chaos, Solitons Fractals **153**, 111587 (2021).
[39]M. Rao, H. Tang, J. Wu, W. Song, M. Zhang, W. Yin, Y. Zhuo, F. Kiani, B. Chen, X. Jiang *et al.*, "Thousands of conductance levels in memristors integrated on CMOS," Nature **615**, 823–829 (2023).
[40]V. Kornev, I. Soloviev, N. Klenov, and O. Mukhanov, "Bi-SQUID: A novel linearization method for dc SQUID voltage response," Supercond. Sci. Technol. **22**, 114011 (2009).
[41]J. Aimone, P. Date, G. Fonseca-Guerra, K. Hamilton, K. Henke, B. Kay, G. Kenyon, S. Kulkarni, S. Mniszewski, M. Parsa *et al.*, "A review of non-cognitive applications for neuromorphic computing," Neuromorphic Comput. Eng. **2**, 032003 (2022).
[42]J. M. Shainline, B. A. Primavera, and S. Khan, "Phenomenological model of superconducting optoelectronic loop neurons," Phys. Rev. Res. **5**, 013164 (2023).
[43]R. O'Loughlin, B. Primavera, and J. Shainline, "Dendritic learning in superconducting optoelectronic networks," in *Proceedings of the 2023 International Conference on Neuromorphic Systems, ICONS'23* (Association for Computing Machinery, New York, 2023).