nature electronics

Article

https://doi.org/10.1038/s41928-022-00840-9

Superconducting optoelectronic single-photon synapses

Received: 25 April 2022

Accepted: 22 August 2022

Published online: 6 October 2022

Check for updates

Saeed Khan[1,3], Bryce A. Primavera[1,2,3], Jeff Chiles[1], Adam N. McCaughan[1], Sonia M. Buckley[1], Alexander N. Tait[1], Adriana Lita[1], John Biesecker[1], Anna Fox[1], David Olaya[1], Richard P. Mirin[1], Sae Woo Nam[1] and Jeffrey M. Shainline[1] ✉

Superconducting optoelectronic hardware could be used to create large-scale and computationally powerful artificial spiking neural networks. The approach combines integrated photonic components that offer few-photon, light-speed communication with superconducting circuits that offer fast, energy-efficient computation. However, the monolithic integration of photonic and superconducting devices is needed to scale this technology. Here we report superconducting optoelectronic synapses that are created by monolithically integrating superconducting nanowire single-photon detectors with Josephson junctions. The circuits perform analogue weighting and the temporal leaky integration of single-photon presynaptic signals. Synaptic weighting is implemented in the electronic domain allowing binary, single-photon communication to be maintained. Records of recent synaptic activity are locally stored as current in superconducting loops, and dendritic and neuronal nonlinearities are implemented with a second stage of Josephson circuitry. This hardware offers synaptic time constants spanning four orders of magnitude (hundreds of nanoseconds to milliseconds). The synapses are responsive to presynaptic spike rates exceeding 10 MHz and consume approximately 33 aJ of dynamic power per synapse event before accounting for cooling. This demonstration also introduces new avenues for realizing large-scale single-photon detector arrays.

The performance of a neural system typically improves with increasing number and connectivity of processing primitives, both in biology[1–3] and artificial intelligence[4,5]. Additionally, analogue processing with spiking communication in systems exhibiting complex temporal dynamics is physically efficient and computationally powerful[3,6–9]. Spiking neural networks, which are inspired by biological systems, are of particular interest for their exploitation of the temporal domain and suitability for efficient implementation in analogue hardware. With recent advances in training algorithms[10–12], spiking networks are becoming competitive with conventional neural networks and are preferable in some applications[13–15]. However, implementing sophisticated spike-based processing in large, highly interconnected networks remains challenging using current hardware.

Biological neurons are capable of directly fanning out to tens of thousands of synapses, but electronic systems struggle with physical fan-outs greater than a few in number and typically resort to digital multiplexing[16]. Multiplexed communication systems inevitably introduce trade-offs between network size and latency. Direct connections between neurons are, therefore, desirable but require innovative hardware. Multiplexed communication can potentially be avoided through

[1]National Institute of Standards and Technology, Boulder, CO, USA. [2]Department of Physics, University of Colorado Boulder, Boulder, CO, USA. [3]These authors contributed equally: Saeed Khan, Bryce A. Primavera. ✉e-mail: jeffrey.shainline@nist.gov

the use of integrated optical receivers and transmitters, which do not suffer from charge-based parasitics and can achieve dedicated connections from each neuron to thousands of synaptic recipients. Dense photonic waveguide networks[17,18] enable high fan-out neurons with light-speed communication. At the same time, superconducting hardware is potentially uniquely capable of creating large-scale spiking neural networks with sophisticated processing units[19–21]. Superconducting single-photon detectors (SPDs) enable the optical communication of synaptic events at the physical limit of energy efficiency, whereas the speed, nonlinearity and low power of Josephson junctions (JJs) make them attractive for implementing neural behaviour[22–29].

In this Article, we report superconducting optoelectronic synapses created by monolithically integrating SPDs and JJs. These synapses detect single-photon events and implement analogue signal weighting in the electronic domain. In addition to such minimum synaptic-processing requirements, the synapses perform the leaky integration of events over time, and the non-dissipative nature of superconductivity allows the leak rates to be chosen over a wide range of timescales covering hundreds of nanoseconds to milliseconds. This capability is promising for leveraging temporal dynamics over many orders of magnitude in complex, adaptive networks. Synaptic circuits are also shown to inductively couple to neuronal and dendritic circuit blocks for further nonlinear processing in a manner that has been theoretically shown to enable high fan-in[30]. As the synapse is the fundamental computational element of neural systems[31], this result could aid the development of large-scale superconducting optoelectronic networks.

The synaptic circuit

The basic synaptic circuit is shown in Fig. 1a and is discussed in detail elsewhere[19,32,33]. The example behaviour (Fig. 1b–j) was calculated with the circuit model described in Supplementary Section 1. We use the notation i_x to refer to currents normalized by the JJ critical current, that is, $i_x = I_x/I_c$. The circuit comprises an SPD receiver at the left and a signal integrator at the right. We refer to the integrator as the synaptic integration (SI) loop. Each time the SPD detects one or more photons, current (I_{si}) is added to the SI loop, and the amount of current added is independent of the number of photons detected, yielding binary photonic communication. This current is added in discrete increments called fluxons[34,35] by the JJ circuit. In the quiescent state, the bias I_{sy} is chosen so that the current flowing through the first junction (J_1) is below that junction's critical current (I_c) and the junction provides a superconducting path to ground. On detection of a photon, the SPD transitions to a resistive state and diverts current I_{spd} into J_1. When $I_{sy} + I_{spd}$ exceeds I_c, J_1 produces a train of fluxons that propagates through a Josephson transmission line and accumulates in the SI loop. The synapse makes use of a passive reset, as the supercurrent returns to the SPD with a time constant set by $L_{spd}/r_{spd} \approx 40$ ns. The number of fluxons added to the SI loop per photon detection event depends on the duration of time for which J_1 is driven above I_c. This behaviour is illustrated in Fig. 1b, where $I_{sy} + I_{spd}$ is plotted for two different values of I_{sy}. The dashed horizontal line in Fig. 1b represents the junction I_c. Fluxons are produced for as long as $I_{spd} + I_{sy}$ remains above I_c. For a low value of I_{sy}, the combined currents exceed I_c only briefly, whereas a high value drives J_1 above I_c for nearly the entire duration of the SPD pulse. Additionally, the rate of fluxon production increases with the current flowing through the junction. Example fluxon trains for the two synaptic weights are shown in Fig. 1c,d, and their corresponding contributions of current to the SI loop are illustrated in Fig. 1e,f. Only a brief subsection of the fluxon train is shown in the high weight case. The difference in fluxon rates is evident, as seven fluxons are produced within 3 ns by the weak synaptic weight, whereas only 200 ps is required to produce seven fluxons for the high synaptic weight. The total number of fluxons added to the SI loop in the case of the synapse event with a high synaptic weight is 1,346.

The role of fluxon trains has a direct analogy with biological synapses, where the strength of synaptic connections is determined by the number of synaptic vesicles containing neurotransmitters that are passed across the synaptic cleft. In our synapses, the strength of synaptic connections is determined by the number of fluxons passed into the SI loop. Analogous to biology, this low-level, discrete picture can often be disregarded in favour of a simpler, essentially analogue description of a high-level synaptic operation.

Figure 1g–j shows the integration of multiple photon detection events in the SI loop. Figure 1g shows the presynaptic spike trains of photon detection events at two different frequencies. Each photon detection event adds current to the SI loop, and the integrated current in the SI loop decays with a time constant set by $\tau_{si} = L_{si}/r_{si}$. We show that the passive elements L_{si} and r_{si} can be engineered over many orders of magnitude. In this way, the SI loop exhibits leaky integrator behaviour desired of spiking neural computational primitives (Fig. 1h). In Fig. 1g–j, τ_{si} is 250 ns; here the model used an accelerated-time synapse with 5 ns SPD recovery to facilitate numerical efficiency. The slower of the two input photonic pulse trains (Fig. 1g) is 4 MHz ($1/\tau_{si}$), whereas the faster input train is twice this frequency. Figure 1h shows that the higher-frequency series of ten pulses results in an appreciably larger integrated signal, a feature that will be prominent in the measured data shown in the 'Experimental characterization' section. The magnitude of current stored in the SI loop is, thus, a record of recent synaptic activity that can be used in subsequent computations, including local weight-update circuits.

The synapse as a whole can be modelled with a leaky integrator equation of the form

$$\frac{\mathrm{d}I_{si}}{\mathrm{d}t} = I_{fq}\,R_{fq}(t) - \frac{I_{si}}{\tau_{si}}, \qquad (1)$$

where $I_{fq} = \Phi_0/L_{si}$ is the current of a single flux quantum entering the integration loop and $R_{fq}(t)$ is the rate of flux-quantum production. At a fixed frequency of input photonic pulses, current accumulates in the SI loop. Yet, during each inter-spike interval, some of this signal will exponentially decay with time constant τ_{si}. A quasi-steady state is reached when the signal added with each photonic pulse counters the signal decay between pulses. Here we use the phrase 'quasi-steady state' to refer to the circumstance where the time-averaged signal between evenly spaced incident photon pulses is constant from one photon pulse to the next. In this case, the time average of $\mathrm{d}I_{si}/\mathrm{d}t = 0$ and equation (1) informs us that $\bar{I}_{si} = \tau_{si}\,I_{fq}\,\bar{R}_{fq}$, where $\bar{I}_{si}$ and $\bar{R}_{fq}$ indicates the time averages. The time-averaged rate at which fluxons are produced depends on the synaptic weight, the rate at which photons are incident, and the value of I_{si}. The emergence of a steady-state behaviour is illustrated in Fig. 1i and its frequency and synaptic weight dependence is illustrated in the transfer function shown in Fig. 1j. This behaviour is experimentally validated in the 'Experimental characterization' section.

The SI loop is also inductively coupled to a superconducting quantum interference device (SQUID) to implement nonlinear transfer functions and transduce I_{si} into a measurable voltage. We refer to this SQUID as the dendritic receiving (DR) loop. The DR transfer function in the present study results in a roughly sigmoidal response and can be tuned with the final bias current. The shape of the response is the culmination of at least three factors: the DR transfer function (Supplementary Section 4), the saturating behaviour of the SI loop (Fig. 1i,j) and the dead time of the SPD at high frequencies (~20 MHz).

Although we study individual synapses here, the theoretical results suggest that these circuits are well suited for neurons with very high fan-in using transformers for coupling[19]. The demonstrated inductive coupling between the SI and DR loops could be extended such that several SI loops are coupled into a single DR loop. Further layers of inductively coupled DR loops could funnel the synaptic information to the neuron cell body in a tree structure, just as that in biological dendritic arbours. This situation is examined elsewhere[30] where it

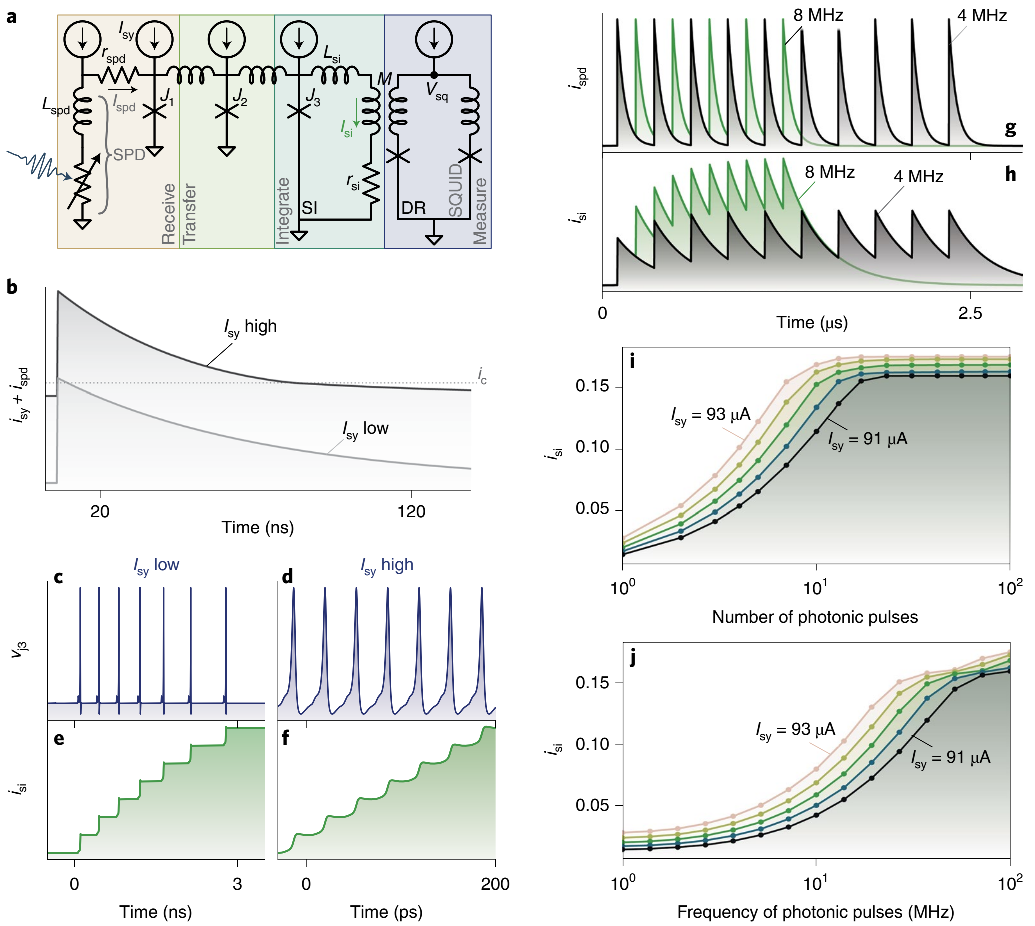

Fig. 1 | Synapse concept. a, Synaptic circuit diagram showing four main circuit blocks. **b**, Simulation of the current flowing into J_1 on detection of a photon for two different values of I_{sy}. A train of fluxons is produced for as long as these traces exceed the critical current I_c of the junction. **c**–**f**, Fluxon trains for low (**c**) and high (**d**) synaptic weights and the resulting additions to the SI loop at low (**e**) and high (**f**) synaptic weights. The high synaptic weight case only shows a brief fraction of the pulse train. **g**, SPD response to a train of single-photon pulses at different frequencies. **h**, SI current in response to the two photonic pulse trains shown in **g**. The higher-frequency input pulse train results in a higher current in the integration loop. **i**, Peak value of SI current (normalized by the JJ critical current) as a function of the number of pulses in a pulse train, demonstrating the emergence of steady-state behaviour and synaptic weighting capability. **j**, Maximum SI current as a function of the frequency of the pulse train, illustrating the saturating behaviour of the SI loop. Time-accelerated SPDs are used in **i** and **j** to reduce the numerical simulation time. Supplementary Section 1 provides details of the circuit model.

is shown that even biological levels of fan-in (>1,000 synapses per neuron) may be achievable with this hardware. A confirmation of this hypothesis should be demonstrable without any changes to the present fabrication process.

Although these synapses have many desirable properties for future large-scale systems, the monolithic integration of JJs and SPDs is the engineering achievement that enabled this demonstration and will also stimulate advances in several applications outside of neuromorphic computing. We now discuss the fabrication process before presenting the measured data.

Fabrication

A 14-layer fabrication process was developed for this demonstration and supports Nb/amorphous silicon (a-Si)/Nb JJs[36] externally shunted with PdAu resistors and MoSi SPDs[37,38]. The high kinetic inductance of the MoSi thin film was also leveraged in conjunction with Au resistors to realize the leaky integrating loops. Electron-beam lithography was used for the SPD step, whereas all other patterning was accomplished with photolithography using a 365 nm i-line stepper. A complete process flow can be found in the Methods section. In brief, the MoSi SPD/high-kinetic-inductance layer is deposited and patterned on a pristine oxidized silicon surface. Contact is made from this layer to Nb. The SPD layer is separated from the JJ layers by a SiO_2 insulator layer and a Nb ground plane. Contact is made from the SPDs to the JJs with etched and backfilled vias. An additional Nb wiring layer is included above the JJs to enable the transformers that couple synapses to SQUIDs. The PdAu and Au resistor layers are patterned last.

A synapse layout and microscopy images of the key components are shown in Fig. 2. Five synapse designs were fabricated with different synaptic time constants and storage capacities. The synapse areas range from 0.32 to 0.52 mm^2 excluding wiring pads, although no effort was made to make the circuits compact in this work. In a mature process, various device layers will be placed atop one another with planarization performed between them. Previous analysis found that similar synapses will fit in 30 μm × 30 μm (ref. [21]). The results presented here are from the first wafer run with this process. The immediate yield is suggestive that the process will be robust.

Experimental characterization

Measurements were performed between 800 and 900 mK in a closed-cycle sorption pump ^{4}He cryostat. Due to wiring limitations, each synapse was measured during a separate cool-down process. A fibre-coupled 780 nm pulsed laser source was used to flood illuminate the chip and serve as the presynaptic input. The laser pulse width was

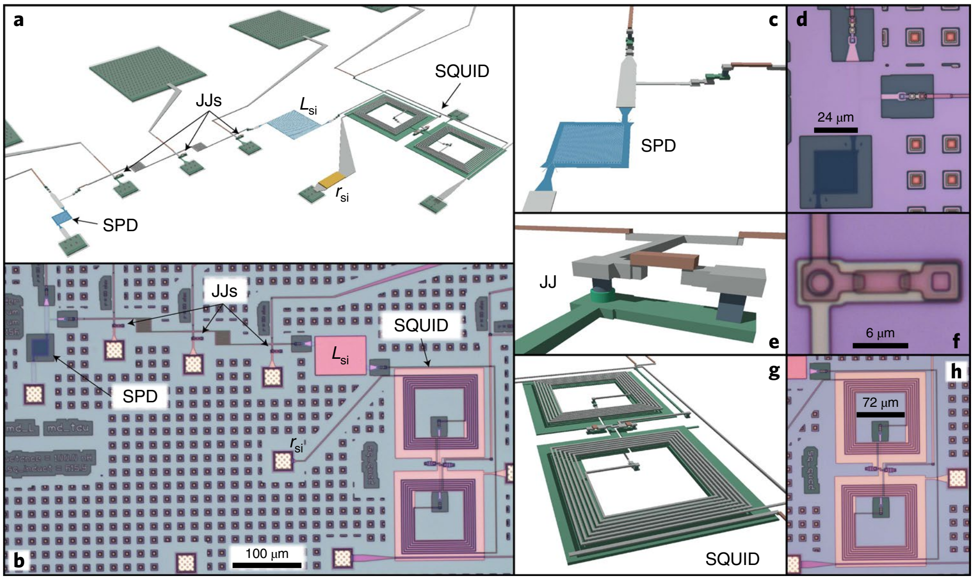

Fig. 2 | Layouts and completed circuits. a, Three-dimensional layout of the entire synapse circuit. **b**, Microscopy image of the completed fabrication. **c**, Layout of SPD. **d**, Fabricated SPD. **e**, Layout of JJ and shunt resistor. **f**, Fabricated JJ and shunt. **g**, Layout of SQUID used for the DR loop. **h**, Fabricated DR SQUID.

approximately 480 ps. This is much shorter than the SPD recovery time; therefore, multiple detection events per pulse are unlikely, as supported by the synapse count rate measurements (Supplementary Section 3). Although each optical pulse contains multiple photons to guarantee detection for each presynaptic event in this free-space configuration, we confirmed the synapses' ability to detect single photons with a linearity measurement under very low light levels (Supplementary Section 3). The detector response is also independent of photon number[39]; therefore, the fact that more than one photon was used as the synaptic input for these measurements is ultimately how the detectors are intended to behave in a network context, and identical dynamics should be expected for few-photon pulses in future waveguide-integrated circuits. As we have argued elsewhere, it may be advantageous for presynaptic spikes to arrive with an average of seven photons to ensure 99% successful communication despite unavoidable Poisson variability[20,21].

The voltage across the DR loop, V_{sq} (Fig. 1a), is on the order of 10 μV and is read out with a room-temperature amplifier. Two different amplifiers (voltage gain, 40 and 60 dB) were used to accommodate the wide range of timescales under investigation. Averaging was required on most traces to counteract the low-frequency line noise disturbing the microvolt signals. An unaveraged trace is shown in Supplementary Section 5. All the biases were generated outside the cryostat and were individually tuned for each synapse to maximize the signal amplitude and account for device variation. Supplementary Fig. 2 shows an exhaustive list of experimental parameters used in the presented data.

Figure 3 presents the characterization of a single synapse designed with a time constant of 6.25 μs and inductance of 2.5 μH. The measurements suggest the actual values to be 8.06 μs and 3.2 μH, respectively. The cause of the increased inductance is described in the Methods section. Synaptic weighting is shown in Fig. 3a, where the response to a single optical pulse is plotted for different values of I_{sy}. The DR response was tuned to be approximately linear over this relatively small signal range, allowing the current in the SI loop, I_{si}, to be estimated from V_{sq} (Supplementary Section 4). The inset in Fig. 3a compares the measured data with the circuit model shown in Supplementary Section 1 and shows that the weighting function measured experimentally agrees with the prediction of our theoretical model. Here I_{sy} is shown to modulate the height of V_{sq} in response to a single optical pulse by at least a factor of 28, and the ability to resolve small weights was limited by noise.

Figure 3b shows the integrating capability of the synapse in response to a pulse train of fixed frequency and pulse number for five values of synaptic weight. For sufficiently long pulse trains, the synapse reaches a steady state that can be tuned with I_{sy} and depends on the frequency of the input pulse train. We refer to such an operation as the 'rate-coding' domain. Figure 3c demonstrates an activity level that is better regarded as the 'burst-coding' domain[40]. Here the synaptic weight is fixed and the different traces correspond to different numbers of pulses as the steady state is approached. Figure 3d shows the synaptic response to ten pulses at three different frequencies, analogous to the theoretical traces (Fig. 1h). Figure 3e,f summarizes these behaviours by plotting the maximum value of V_{sq} as a function of the photonic pulse number and frequency, respectively, for several values of synaptic weights in a manner analogous to the theoretical traces (Fig. 1i,j). Figure 3e shows the transition from the burst-coding regime for low numbers of pulses to the rate-coding regime where I_{si} reaches a steady-state level. Figure 3f displays the desired roll-over behaviour with pulse frequency discussed previously. Both figures also demonstrate the ability of I_{sy} to tune the synaptic response over a wide range. Sigmoidal fits for these plots are presented in Supplementary Section 5. Although the pulse number and frequency transfer functions capture the essential analogue computation for the burst- and rate-coding domains, these synapses could be used with spike timing as well[33] and make use of the extremely low jitter (picoseconds) of superconducting nanowire detectors[41].

The synaptic transfer functions can be engineered over a wide range of timescales (Fig. 4). Figure 4a shows the temporal response of synapses with four different time constants to a single photonic pulse. Moderate synaptic weights were chosen for each synapse. We see that the temporal dynamics range from the sub-microsecond to several milliseconds. The dotted lines (Fig. 4a) show the exponential fits on the semilogarithmic plot. Figure 4b,c shows the integrating behaviour for synapses with the fastest (271 ns) and slowest (5.83 ms) time constants, respectively, demonstrating no degradation in the leaky integrating behaviour at either extreme. Figure 4d–f shows the frequency response functions for those same two synapses as well as one with an intermediate time constant designed to be 6.25 μs and measured to be 8.06 μs. We once again see similar behaviour across

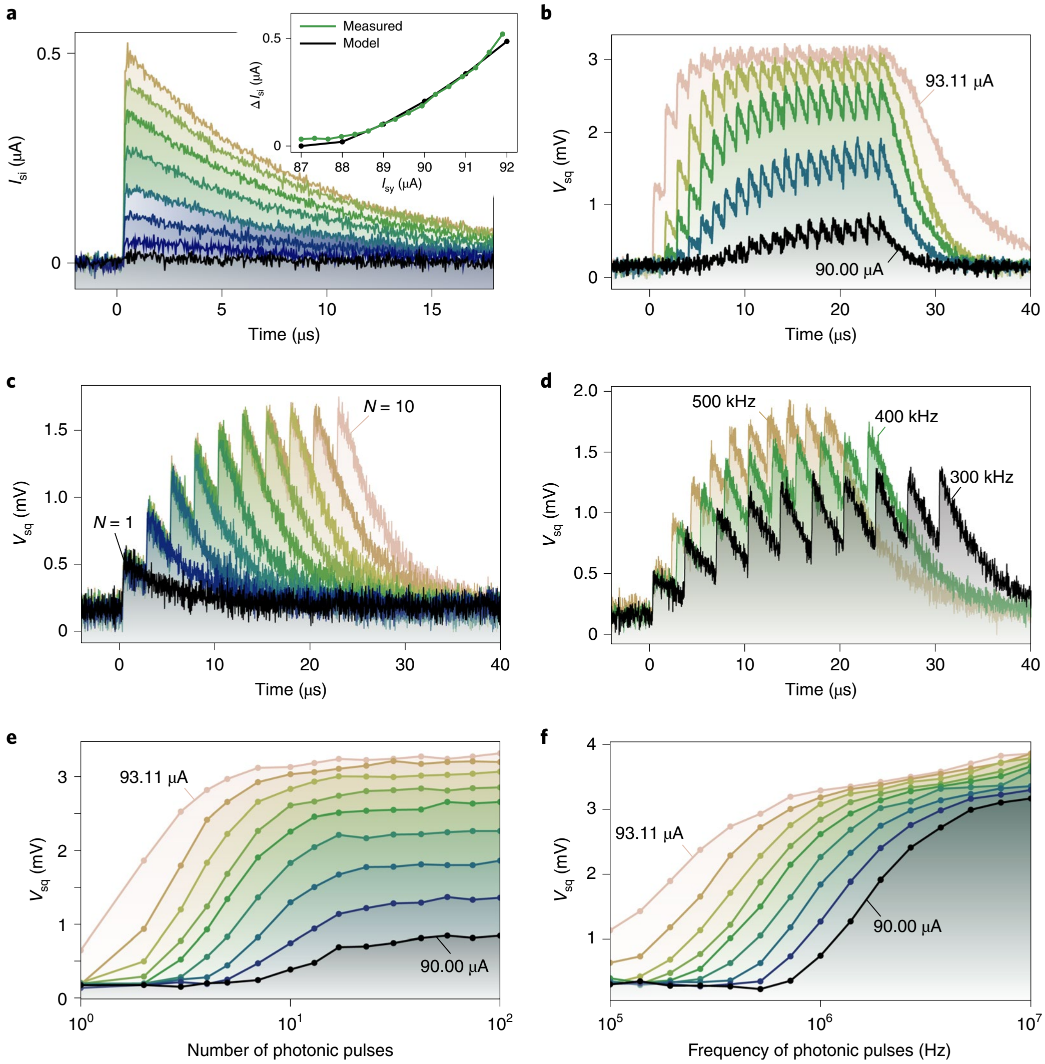

Fig. 3 | Detailed characterization of 6.25 μs, 2.5 μH synapse. **a**, Response of I_{si} to a single optical pulse for different values of I_{sy}. The inset shows a comparison to the model shown in Supplementary Section 1 (1,000 averages). **b**, Response to an 800 kHz train of 20 pulses for different synaptic weights. **c**, Response at fixed I_{sy} to 400 kHz pulse trains containing different numbers of pulses. **d**, Response at fixed I_{sy} to pulse trains at three different frequencies (compare with Fig. 1h). **e**, Transfer function versus number of pulses in a pulse train at different synaptic weights and fixed frequency (1 MHz) (compare with Fig. 1i). **f**, Transfer function versus pulse train frequency at different synaptic weights for 100 pulses (compare with Fig. 1j).

the timescales and observe that the onset of integration can be tuned from 100 Hz to almost 10 MHz. Figure 4 also illustrates the interplay between the parameters fixed in hardware (L_{si} and r_{si}) and those that can be dynamically reconfigured (I_{sy}).

The vast parameter space (pulse frequency, pulse number, I_{sy}, τ_{si} and L_{si}) exhibited by these synapses is promising for fostering complex dynamics in large-scale networks. Figure 5 captures the synaptic response across a large range of this space. The temporal traces (Figs. 3 and 4) are once again reduced to single data points corresponding to the maximum change in voltage recorded for each trace, as in the transfer functions. For fixed I_{sy}, these values of V_{sq} are plotted as two-dimensional heat maps with the frequency of a pulse train along the *x* axis and the number of pulses in a train along the *y* axis (Fig. 5a). Both *x* and *y* axes are spaced with a geometric progression. These data are presented for six different values of I_{sy} (rows) and five different synapses (columns) (Fig. 5c), and the colour bars in Fig. 5b give the scale for each column. Due to the use of different amplifiers and different bias conditions on the synapses measured on different cool downs, the variation in colour axis was unavoidable. The synapses are labelled by their designed time constant and SI loop inductance. Comparisons between the different synapses (columns) should be taken somewhat qualitatively given the fabrication and biasing variations. Nonetheless, we can clearly see great diversity in behaviour, and several notable trends can be observed. The turn-on frequency increases inversely with τ_{si}, as expected for an *LR* filter. Both frequency and number responses can be broadly adjusted with the synaptic weight. The 250 ns synapse responds far more strongly to frequencies greater than 1 MHz, and this is striking in comparison to the microsecond synapse that successfully integrates at much lower frequencies.

The third and fourth columns in Fig. 5c show two synapses with the same designed time constant, but two different values of L_{si}. Here L_{si} sets the amount of current added to the SI loop per fluxon. This is a different weighting mechanism than I_{sy}, which sets the number of fluxons added per photon detection. A larger L_{si} value corresponds to lower current added to the loop per fluxon; therefore, the apparently slower turn on of the 2.5 μH synapse with I_{sy} is expected. However, we caution that some of this discrepancy may be from different biasing conditions including a reduced SPD bias on the 2.5 μH synapse to account for a lower SPD I_c. Investigating the effect of different values of L_{si} is intriguing because it sets the ultimate number of

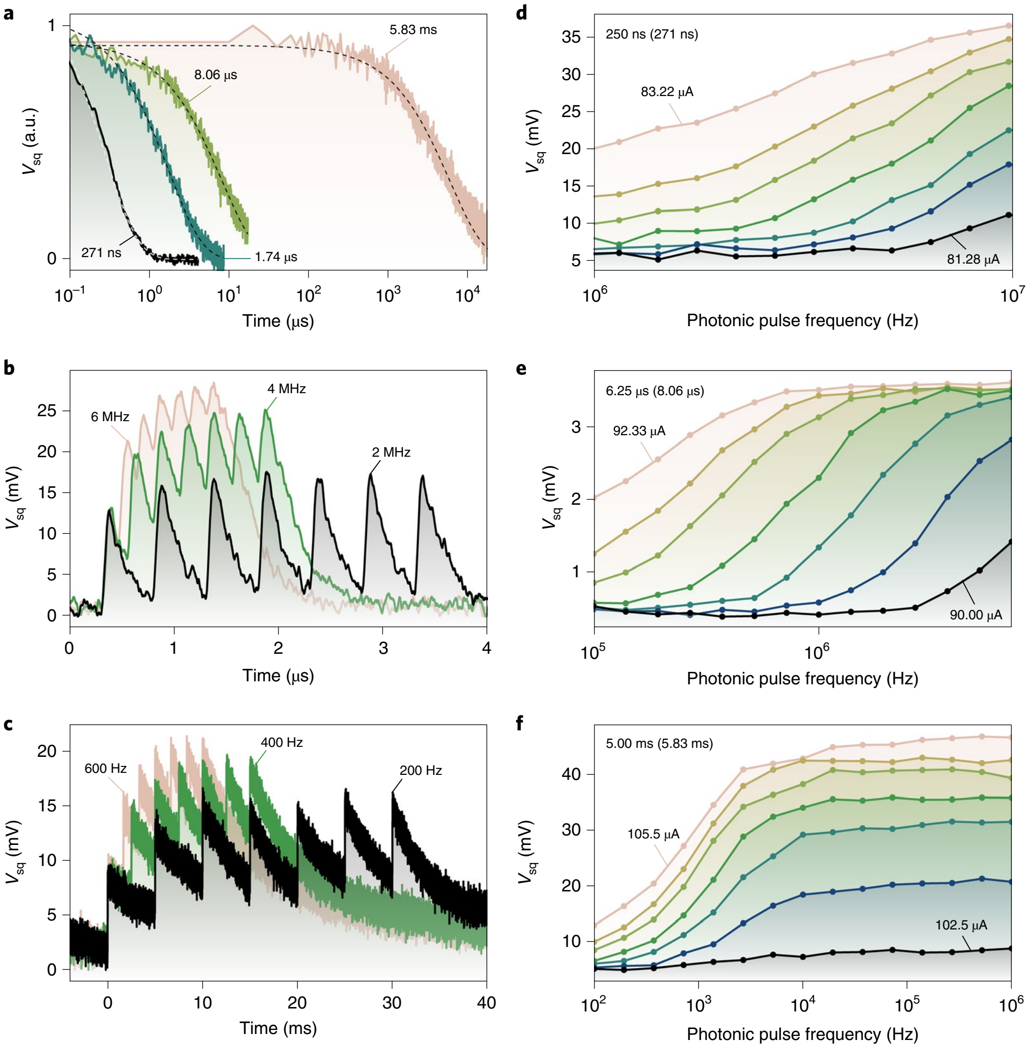

Fig. 4 | Comparison of synapses with varying time constants. a, Response to a single optical pulse at moderate synaptic weight for synapses with four different designed time constants (250 ns, 1.25 μs, 6.25 μs and 5 ms). The dotted lines show the exponential fits with the extracted time constants. **b**, The 250 ns synapse response to seven optical pulses at three different megahertz-range frequencies. **c**, The 5 ms synapse response to seven optical pulses at three different sub-kilohertz frequencies. **d**–**f**, Frequency transfer functions with varying synaptic weights for synapses with three different designed time constants: 250 ns (**d**), 6.25 μs (**e**) and 5.00 ms (**f**). Panels **d** and **e** show the response to 100 input pulses, while **f** is for 10 input pulses.

events that can be stored by the SI loop, but a full analysis requires further study.

The millisecond synapse in the far right column (Fig. 5c) is a curious case, as it is independent of the frequency over this measured range (100 kHz to 10 MHz). The current in the SI loop does not have time to decay between pulses, making this synapse essentially a photon-counting device over this range. This slow-leak behaviour is illustrated in Fig. 6, where the millisecond synapse is shown responding to pulse trains of 500 kHz, 1.0 MHz and 1.5 MHz. Each pulse train is identical in length (15 events). We see that the final magnitude of V_{sq} is independent of frequency, and is instead determined only by I_{sy} and the number of pulses in the train. We anticipate this regime to be useful in long-term plasticity mechanisms for modulating the behaviour of faster synapses, as well as for applications in SPD imaging arrays.

Conclusions

Through the monolithic integration of SPDs and JJs, we have created tunable single-photon optoelectronic synapses with numerous temporal filtering properties that are useful for spiking neural systems. The synapses are responsive to presynaptic spiking events encoded as single-photon signals, which makes them of potential use in large-scale networks that use direct optical connections between neurons. Unlike electrical communication, optical signals are immune to the parasitics that create challenges for electronic fan-out approaches. Direct communication is superior to multiplexed systems in large, highly interconnected networks, as latency is essentially decoupled from network complexity. Latency in this system is limited only by the time it takes light to travel between nodes.

The demonstrated synapses are more than simple weighting elements, as they also perform temporal integration and other analogue computational primitives observed in biology. A system with integrating synapses (and dendrites) is more dynamic than the one in which only the neurons perform integration. The synapses reported here make full use of the temporal advantages of spiking networks. The range of decay times of hundreds of nanoseconds to several milliseconds is another advantage of the approach, and is valuable for use in networks that are matched to applications that have to operate over a variety of timescales—from accelerated simulations and precise control systems to interactions with humans. The millisecond synapse achieves a biologically relevant timescale, which has been a major area

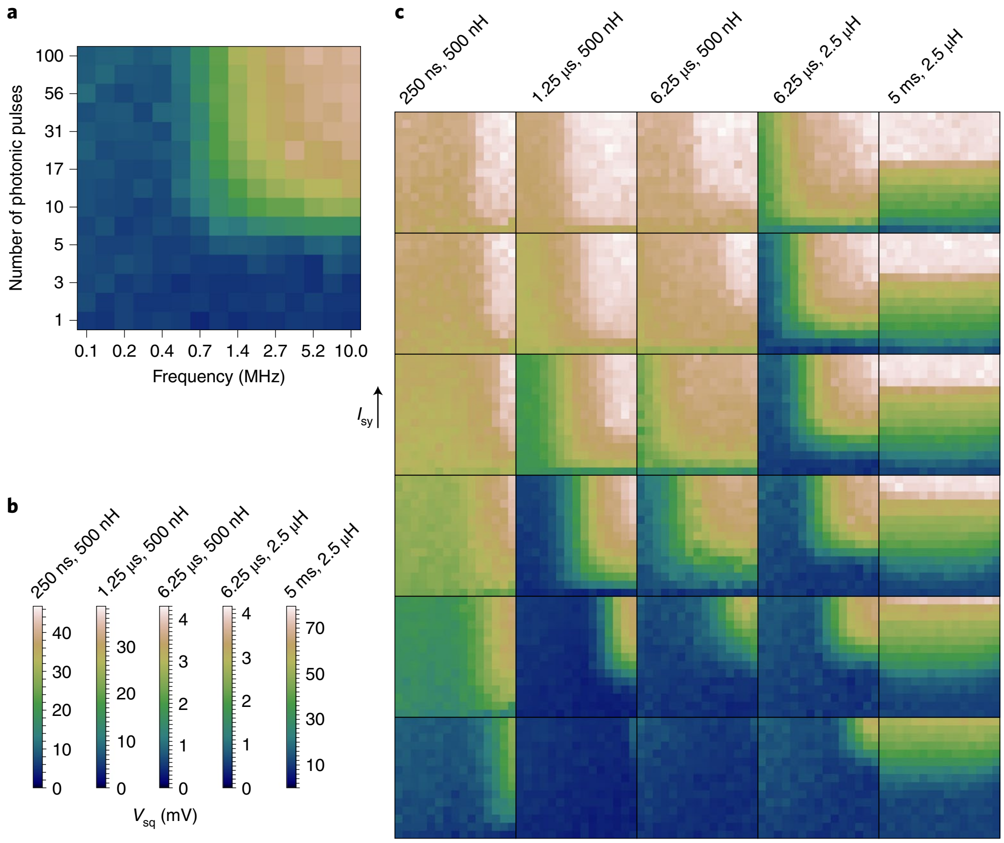

Fig. 5 | Heat maps of synaptic response. **a**, Sample map at a fixed synaptic weight showing frequency and number axes (fixed for all the plots). **b**, Colour bars showing the change in V_{sq}, normalized for each synapse. **c**, Grid showing the evolution of synaptic response with I_{sy} (increasing vertically) for five different synapses.

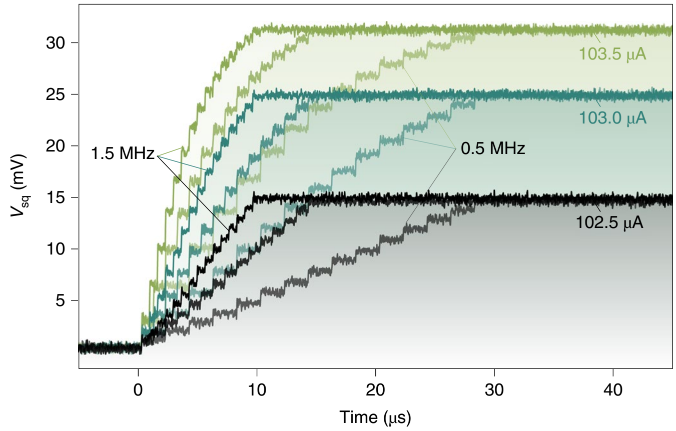

Fig. 6 | The 5 ms synapse operating in the number-counting regime. There are 15 optical pulse inputs at 500 kHz, 1.0 MHz and 1.5 MHz. We observe that the final value of V_{sq} is independent of frequency for this timescale.

of research in analogue complementary metal–oxide–semiconductor implementations[42,43].

Mixing synapses with different time constants in the same network may also be advantageous, as networks with time constants spanning orders of magnitude may be more suitable to develop temporal dynamics with power-law statistics—the signature of critical behaviour that has been studied for its important role in cognition[44–46]. Integrating behaviour is also useful for plasticity mechanisms that rely on the recent history of synaptic activity to update weights, such as spike-timing-dependent plasticity[47,48]. We envision local plasticity circuits coupled to the SI loop at every synapse. Many of these plasticity (or homoeostatic) mechanisms will also be desired to operate at timescales substantially longer than those of the synapses themselves, which appears feasible.

There are a number of directions in which these circuits could be developed further. The current fabrication process should be compatible with operating temperatures of up to 2.7 K, but other SPD materials (probably NbTiN) may prove to be better suited for operation at 4.0 K where liquid-helium immersion cooling can be used. This would improve energy efficiency as well as simplify the measurement apparatus and remove compressor noise. Energy efficiency of the devices themselves could also be improved. The main source of energy consumption here is through the discharging of detector inductance. Free-space operation required relatively large detector areas and thus large inductance (≈825 nH). This energy ($\frac{1}{2}LI^2$) corresponds to 33 aJ per event (or 33 fJ if a factor of 1,000 is included for the cooling overhead). A low-inductance waveguide-integrated synapse could improve this metric by an order of magnitude[17]. Optical communication is still likely to dominate power consumption even in this case; therefore, thousands of fluxons can be produced per synaptic operation before computational power begins to dominate over communication.

Another deficiency is the use of an external bias to provide the synaptic weight. A local memory would be superior. One possibility is to store the weight as a persistent current in another superconducting loop that is inductively coupled to the I_{sy} bias line[19]. Such memory should be achievable without any changes to the process flow. Both unsupervised Hebbian-based learning[12,49] and recently developed supervised algorithms for local gradient descent[11,50] could be implemented with local analogue plasticity circuits that adjust the current stored in such memory loops. In the near term, the synapses could be interfaced with bias-generator circuits for either hardware-in-the-loop training or for implementing a fixed network in inference tasks. Further work remains to combine these synapses with light-emitting neuron circuits. Substantial progress in cryogenic integrated light sources[51,52] and superconducting optical transmitter circuits[53] has occurred, indicating that full superconducting optoelectronic neurons may not be far away.

Although this platform was designed for high-performance neuromorphic computing, there are many application areas that could benefit from this monolithic SPD–JJ integration. The exact synaptic circuits we have presented could be used as single-photon integrating pixels for advanced imaging applications. To date, large-scale SPD arrays have been limited by read-out technologies, but the integrating character of these circuits allows photon detection events to be stored and read out at later convenience. The demonstrated millisecond retention times are particularly suitable to large arrays. Single-photon to single-fluxon transduction with long integration times should be feasible, enabling accurate photon counting similar to that described in another work[54]. The fabrication process should also be applicable to Josephson circuits incorporating single-flux-quantum logic, presenting the possibility of digital processing of single-photon events[55,56] for applications including qubit control or post-processing of images acquired with SPD arrays. Thus, we expect monolithic SPD–JJ integration to offer new opportunities in fields as diverse as quantum information and communication, biomedical imaging and broad-spectrum astronomical observations[57], potentially creating an entirely new field of integrated superconducting optoelectronic hardware.

Methods

Fabrication details

The fabrication of the SPD–JJ synaptic circuits combines previously developed SPD and JJ fabrication processes. MoSi SPDs have been described in refs. [37,38], and Nb JJs have been described in refs. [36,58]. The process began with a thermally oxidized silicon wafer. Thin Nb contacts (45 nm) were deposited via sputtering, and a liftoff process was employed to realize a very shallow sidewall angle. This defined the first metal layer, M1. Next, MoSi was sputtered to a thickness of 5 nm and capped with a 2 nm a-Si layer. This defined the superconducting thin-film (STF) layer. The Nb liftoff process was used for M1 so the thin MoSi layer could make the superconducting contact. In a mature fabrication process, it will be desirable to deposit Nb wires with a damascene process followed by chemical–mechanical planarization. MoSi would be deposited on the planarized surface and would make contact with the planarized Nb. In the present work, we avoided this planarization step as it is non-trivial in the National Institute of Standards and Technology cleanroom. We have performed experiments with MoSi deposited on planarized oxidized wafers with less than 0.5 nm root-mean-square roughness and measured less than 0.2 K change in the film critical temperature compared with a film deposited on a thermally oxidized substrate.

The SPDs were defined on the STF layer through electron-beam lithography. The wires defining the SPDs were 200 nm wide on the mask with 50% fill factor, resulting in wire widths near 180 nm on the wafer. A 100 kV column and 2 nA beam current were used in patterning. The inductors comprising the SI loops were formed from the same STF layer but were defined with i-line photolithography. The feature sizes on this layer were 1 μm or larger. All the features on the STF were dry etched with SF_6 chemistry.

Following M1 and STF, the first SiO_2 insulator (I1) was deposited using electron cyclotron resonance plasma-enhanced chemical vapour deposition. I1 was 200 nm thick, as were all the insulating layers. The first via layer (V1) was etched through I1, terminating on the Nb contacts formed in M1. V1 was etched with CHF_3 chemistry with O_2 added to increase the sidewall slope to facilitate via formation without a damascene process. M2 was then deposited by sputtering Nb, and the wires were formed with i-line photolithography and dry etching using an inductively coupled plasma with SF_6 chemistry. M2 serves as the superconducting ground plane for the circuits. I2 was then deposited using the same process as I1. V2 was opened using the same process as V1.

The JJs were fabricated next using an Nb/a-Si/Nb trilayer. The lower-electrode Nb layer (JJ1) made contact with M2 through V2. The a-Si tunnelling barrier was approximately 5 nm thick. Self-shunted JJs with Nb-doped Si barriers have been previously used in single-flux-quantum circuits[36]. However, in this work, undoped a-Si barriers and external shunts were used as the JJ area was not a concern in these circuits. The target critical current density of the trilayer was 1 kA cm^{-2}, but the resulting value was more than twice as large due to an unidentified change in the process. Subsequent wafers have returned to the previous value. The fact that the synaptic circuits remained highly functional when the critical currents of all the JJs on the wafer were more than twice as large as designed is evidence for the robustness of the circuit concepts. The JJ top electrode (JJ2) and tunnelling barrier were dry etched with SF_6 chemistry, stopping on JJ1 to leave a metal wiring layer. The JJ1 layer was also used as the SQUID washer body, which served as the pickup component of the transformer between the SI and DR loops. I3 was then deposited at 200 nm with the same SiO_2 as the other insulators, and V3 was etched to contact the JJ top and bottom electrodes.

Upper Nb metal M3 was then deposited and etched just as M2. This wiring layer was used for superconducting interconnects and to form the input coil for the transformer from the SI to DR loop. M1 was also employed in this transformer to cross under JJ1. All the SQUIDs on this wafer leveraged quadrupole designs (Fig. 2g,h) to avoid sensitivity to stray uniform magnetic fields. Some SQUIDs on a diagnostic chip included resistors coupled to the washer to damp *LC* resonances. However, for the SQUIDs employed to interface with the synapses, an abundance of caution drove us to omit those resistors to our detriment. *LC* resonances driven by the room-temperature amplifier had deleterious effects on the SQUID response curves (Supplementary Section 4), limiting the dynamic range we were able to use for the measurements.

The JJ shunt resistors were formed from PdAu due to that material's low residual resistivity ratio, whereas the resistors used in the SI loops to set the leak rate were formed from Au due to its low resistivity. This was particularly important for the long-time-constant synapses utilizing interdigitated structures for very low resistances. These resistors can require large areas if the sheet resistance is high. These two layers were deposited in separate liftoff steps. Again, a mature foundry would be able to straightforwardly adapt these steps to a damascene process. A final insulator was deposited just as the others, and large vias were opened to enable wire bonding to the top Au layer.

In addition to the JJ tunnelling barrier being thinner than desired leading to higher I_c than designed, one other processing step was sub-optimal. Before depositing I1 over STF, a radio-frequency clean was conducted in the deposition chamber. This plasma clean slightly thinned the STF film resulting in suppressed critical temperature and increased inductance. All the circuits were designed to operate at 2.7 K, with the MoSi film having a critical temperature above 5.5 K, but this film degradation required operation below 1.0 K to support the designed SPD I_c values. The increased inductance is probably the cause of the SI loop inductances being larger than design. This issue has since been resolved by depositing a slightly thicker a-Si capping layer on STF.

All the data presented in this paper were acquired from the first wafer fabricated with this combined SPD–JJ process.

Experimental details

A schematic of the experimental setup is presented in Supplementary Section 2. Each synapse requires seven coaxial connections for input/

output. Four current biases (namely, I_{spd}, I_{sy}, I_{jtl} and I_{sc}) are required by the synaptic block (Supplementary Fig. 1). The DR block requires a bias for the SQUID (I_{dr}), an 'add-flux' bias (I_{af}) for tuning the operating point of the DR loop and a line for reading the voltage (V_{sq}) across the loop. Fabrication variation between the synapses required each of the six current biases to be adjusted to maximize the signal amplitude. Due to current sharing on the JJs between the biases, there are many possible biasing conditions that behave in a qualitatively similar manner. To maintain consistency, the biases I_{sy}, I_{jtl} and I_{sc} were chosen with similar values when possible. The add-flux line was chosen so that V_{sq} just started to increase with I_{si}, except for where more linear operation was desired (Figs. 3a and 4a). The biases for Figs. 1–6 in the main text are provided in Supplementary Fig. 2.

All the measurements were performed between 800 and 900 mK in a closed-cycle sorption pump ^{4}He cryostat. Two concentric cylindrical mu-metal shields reduce external magnetic noise. An optical fibre was positioned to flood illuminate the entire chip. A function generator triggered the 780 nm laser to produce bursts of pulses of a given number and frequency. Here I_{dr} was supplied by a commercially available current source, whereas all the other current biases came from resistors in series with isolated voltage sources. Two different amplifiers were used in these experiments to read V_{sq}. One was a home-built amplifier with 40 dB voltage gain and 1 MHz bandwidth. The other was a commercial amplifier with 60 dB gain and 10 MHz bandwidth. The bandwidth limitations of the amplifier are visible for the 250 ns synapse (Fig. 4b). The time traces were recorded with a 1 GHz oscilloscope triggered by the function generator.

Data availability

The data that support the findings of this study are publicly available via Figshare at https://doi.org/10.6084/m9.figshare.21277845.

References

1. Dicke, U. & Roth, G. Neuronal factors determining high intelligence. *Phil. Trans. R. Soc. B* **371**, 20150180 (2016).
2. Herculano-Houzel, S. The human brain in numbers: a linearly scaled-up primate brain. *Front. Hum. Neurosci.* **3**, 31 (2009).
3. Sterling, P. & Laughlin, S. *Principles of Neural Design* (MIT Press, 2015).
4. Hestness, J. et al. Deep learning scaling is predictable, empirically. Preprint at https://arxiv.org/abs/1712.00409 (2017).
5. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
6. Koch, C. & Segev, I. The role of single neurons in information processing. *Nat. Neurosci.* **3**, 1171–1177 (2000).
7. Laughlin, S. B. & Sejnowski, T. J. Communication in neuronal networks. *Science* **301**, 1870–1874 (2003).
8. Schemmel, J. et al. A wafer-scale neuromorphic hardware system for large-scale neural modeling. In *2010 IEEE International Symposium on Circuits and Systems (ISCAS)* 1947–1950 (IEEE, 2010).
9. Indiveri, G. et al. Neuromorphic silicon neuron circuits. *Front. Neurosci.* **5**, 73 (2011).
10. Zenke, F. & Ganguli, S. Superspike: supervised learning in multilayer spiking neural networks. *Neural Comput.* **30**, 1514–1541 (2018).
11. Kaiser, J., Mostafa, H. & Neftci, E. Synaptic plasticity dynamics for deep continuous local learning (DECOLLE). *Front. Neurosci.* **14**, 424 (2020).
12. Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T. & Maida, A. Deep learning in spiking neural networks. *Neural Netw.* **111**, 47–63 (2019).
13. Davies, M. et al. Advancing neuromorphic computing with Loihi: a survey of results and outlook. *Proc. IEEE* **109**, 911–934 (2021).
14. Beer, M., Urenda, J., Kosheleva, O. & Kreinovich, V. Why spiking neural networks are efficient: a theorem. In *Information Processing and Management of Uncertainty in Knowledge-based Systems* 59–69 (Springer, 2020).
15. Indiveri, G. & Sandamirskaya, Y. The importance of space and time for signal processing in neuromorphic agents: the challenge of developing low-power, autonomous agents that interact with the environment. *IEEE Signal Process. Magazine* **36**, 16–28 (2019).
16. Liu, S.-C., Delbruck, T., Indiveri, G., Whatley, A. & Douglas, R. *Event-based Neuromorphic Systems* (John Wiley & Sons, 2014).
17. Chiles, J. et al. Multi-planar amorphous silicon photonics with compact interplanar couplers, cross talk mitigation, and low crossing loss. *APL Photon.* **2**, 116101 (2017).
18. Chiles, J., Buckley, S. M., Nam, S. W., Mirin, R. P. & Shainline, J. M. Design, fabrication, and metrology of 10 × 100 multiplanar integrated photonic routing manifolds for neural networks. *APL Photon.* **3**, 106101 (2018).
19. Shainline, J. M. et al. Superconducting optoelectronic loop neurons. *J. Appl. Phys.* **126**, 044902 (2019).
20. Shainline, J. M. Optoelectronic intelligence. *Appl. Phys. Lett.* **118**, 160501 (2021).
21. Primavera, B. A. & Shainline, J. M. Considerations for neuromorphic supercomputing in semiconducting and superconducting optoelectronic hardware. *Front. Neurosci.* **15**, 732368 (2021).
22. Harada, Y. & Goto, E. Artificial neural network circuits with Josephson devices. *IEEE Trans. Magn.* **27**, 2863–2866 (1991).
23. Hidaka, M. & Akers, L. An artificial neural cell implemented with superconducting circuits. *Supercond. Sci. Technol.* **4**, 654 (1991).
24. Mizugaki, Y., Nakajima, K., Sawada, Y. & Yamashita, T. Implementation of new superconducting neural circuits using coupled SQUIDs. *IEEE Trans. Appl. Supercond.* **4**, 1–8 (1994).
25. Rippert, E. D. & Lomatch, S. A multilayered superconducting neural network implementation. *IEEE Trans. Appl. Supercond.* **7**, 3442–3445 (1997).
26. Kondo, T., Kobori, M., Onomi, T. & Nakajima, K. Design and implementation of stochastic neurosystem using SFQ logic circuits. *IEEE Trans. Appl. Supercond.* **15**, 320–323 (2005).
27. Hirose, T., Asai, T. & Amemiya, Y. Pulsed neural networks consisting of single-flux-quantum spiking neurons. *Physica C* **463**, 1072–1075 (2007).
28. Crotty, P., Schult, D. & Segall, K. Josephson junction simulation of neurons. *Phys. Rev. E* **82**, 011914 (2010).
29. Schneider, M. et al. SuperMind: a survey of the potential of superconducting electronics for neuromorphic computing. *Supercond. Sci. Technol.* **35**, 053001 (2022).
30. Primavera, B. A. & Shainline, J. M. An active dendritic tree can mitigate fan-in limitations in superconducting neurons. *Appl. Phys. Lett.* **119**, 242601 (2021).
31. Shepherd, G. M. *The Synaptic Organization of the Brain* (Oxford Univ. Press, 2004).
32. Shainline, J. M. et al. Circuit designs for superconducting optoelectronic loop neurons. *J. Appl. Phys.* **124**, 152130 (2018).
33. Shainline, J. M. Fluxonic processing of photonic synapse events. *IEEE J. Sel. Top. Quantum Electron.* **26**, 1–15 (2019).
34. Tinkham, M. *Introduction to Superconductivity* (Courier, 2004).
35. Duzer, T. V. & Turner, C. W. *Principles of Superconductive Devices and Circuits* 2nd edn (Prentice Hall, 1998).
36. Olaya, D. et al. Planarized process for single-flux-quantum circuits with self-shunted Nb/Nb$_x$Si$_{1-x}$/Nb Josephson junctions. *IEEE Trans. Appl. Supercond.* **29**, 1–8 (2019).
37. Verma, V. B. et al. High-efficiency superconducting nanowire single-photon detectors fabricated from MoSi thin-films. *Opt. Express* **23**, 33792–33801 (2015).

38. Lita, A. E., Verma, V. B., Chiles, J., Mirin, R. P. & Nam, S. W. 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).
39. Buckley, S. M. et al. Integrated-photonic characterization of single-photon detectors for use in neuromorphic synapses. *Phys. Rev. Appl.* **14**, 054008 (2020).
40. Zeldenrust, F., Wadman, W. J. & Englitz, B. Neural coding with bursts—current state and future perspectives. *Front. Comput. Neurosci.* **12**, 48 (2018).
41. Korzh, B. et al. Demonstration of sub-3 ps temporal resolution with a superconducting nanowire single-photon detector. *Nat. Photon.* **14**, 250–255 (2020).
42. Chicca, E., Stefanini, F., Bartolozzi, C. & Indiveri, G. Neuromorphic electronic circuits for building autonomous cognitive systems. *Proc. IEEE* **102**, 1367–1388 (2014).
43. Mayr, C. et al. A biological-realtime neuromorphic system in 28 nm CMOS using low-leakage switched capacitor circuits. *IEEE Trans. Biomed. Circuits Syst.* **10**, 243–254 (2015).
44. Beggs, J. M. The criticality hypothesis: how local cortical networks might optimize information processing. *Phil. Trans. R. Soc. A* **366**, 329–343 (2008).
45. Cocchi, L., Gollo, L. L., Zalesky, A. & Breakspear, M. Criticality in the brain: a synthesis of neurobiology, models and cognition. *Prog. Neurobiol.* **158**, 132–152 (2017).
46. Tomen, N., Herrmann, J. M. & Ernst, U. *The Functional Role of Critical Dynamics in Neural Systems* Vol. 11 (Springer, 2019).
47. Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J. & Masquelier, T. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Netw.* **99**, 56–67 (2018).
48. Dan, Y. & Poo, M.-m Spike timing-dependent plasticity of neural circuits. *Neuron* **44**, 23–30 (2004).
49. Diehl, P. U. & Cook, M. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* **9**, 99 (2015).
50. Bourdoukan, R. & Deneve, S. Enforcing balance allows local supervised learning in spiking recurrent networks. *Adv. Neural Inf. Process. Syst.* **28**, 982–990 (2015).
51. Buckley, S. et al. All-silicon light-emitting diodes waveguide-integrated with superconducting single-photon detectors. *Appl. Phys. Lett.* **111**, 141101 (2017).
52. McDonald, C. et al. III-V photonic integrated circuit with waveguide-coupled light-emitting diodes and WSi superconducting single-photon detectors. *Appl. Phys. Lett.* **115**, 081105 (2019).
53. McCaughan, A. N. et al. A superconducting thermal switch with ultrahigh impedance for interfacing superconductors to semiconductors. *Nat. Electron.* **2**, 451–456 (2019).
54. Onen, M. et al. Single-photon single-flux coupled detectors. *Nano Lett.* **20**, 664–668 (2019).
55. Yabuno, M., Miyajima, S., Miki, S. & Terai, H. Scalable implementation of a superconducting nanowire single-photon detector array with a superconducting digital signal processor. *Opt. Express* **28**, 12047–12057 (2020).
56. Ortlepp, T. et al. Demonstration of digital readout circuit for superconducting nanowire single photon detector. *Opt. Express* **19**, 18593–18601 (2011).
57. Steinhauer, S., Gyger, S. & Zwiller, V. Progress on large-scale superconducting nanowire single-photon detectors. *Appl. Phys. Lett.* **118**, 100501 (2021).
58. Olaya, D., Dresselhaus, P. D. & Benz, S. P. 300-GHz operation of divider circuits using high-J_c Nb/Nb_xSi_{1-x}/Nb Josephson junctions. *IEEE Trans. Appl. Supercond.* **25**, 1101005 (2015).

Acknowledgements

We appreciate the recommendations on the read-out SQUID design from B. Mates, M. Durkin and J. Aumentado. We appreciate insights on the experimental characterization from M. Castellanos-Beltran and D. Rampini. This work was made possible by the institutional support from the National Institute of Standards and Technology (award no. 70NANB18H006; B.A.P.) and the effort to advance hardware for artificial intelligence and by the DARPA Invisible Headlights Program (HR0011149863 and HR0011151332; S.K. and J.M.S.).

Author contributions

S.K. contributed to the design of experiment, circuit concepts, circuit modelling, layout, fabrication, experimental apparatus and experimental characterization. B.A.P. contributed to the circuit modelling, experimental apparatus, experimental characterization and data analysis. J.C. contributed to the fabrication and experimental characterization. A.N.M. contributed to the circuit concepts and experimental characterization. S.M.B. contributed to the circuit concepts and experimental apparatus. A.N.T. contributed to the circuit concepts and experimental apparatus. A.L., J.B., A.F. and D.O. contributed to the fabrication process development. R.P.M. contributed to the project management. S.W.N. contributed to the project management and experimental characterization. J.M.S. contributed to the experiment design, circuit concepts, circuit modelling, layout, fabrication, data analysis and project management. All the authors contributed to the manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41928-022-00840-9.

Correspondence and requests for materials should be addressed to Jeffrey M. Shainline.

Peer review information *Nature Electronics* thanks Robert Dynes, Robert Hadfield and Taro Yamashita for their contribution to the peer review of this work.