Published in final edited form as:

Supercond Sci Technol. 2018 ; 32(1): . doi:10.1088/1361-6668/aae50d.

A kinetic-inductance-based superconducting memory element with shunting and sub-nanosecond write times

Adam N McCaughan¹, Emily Toomey², Michael Schneider¹, Karl K Berggren², Sae Woo Nam¹

¹National Institute of Standards and Technology, 325 Broadway St, Boulder, CO 80305, United States of America

²Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, MA 02139, United States of America

Abstract

We present a kinetic-inductance-based superconducting memory element with non-destructive readout, femtojoule read and write energies, both read and write shunts, which is writeable with pulses shorter than 400 ps. The element utilizes both a high-kinetic-inductance layer made from tungsten silicide as well as a low-kinetic-inductance layer made from niobium. By using tungsten silicide—which has a long (20 ns) thermal time constant—and measuring bit error rates from 10 MHz to 1 GHz, we were able to verify that the thin-film elements could be operated at a data rate at least as fast as the material thermal time constant with a bit error ratio less than 10^{-6} . We also analyze the margins of the device, and outline the characteristics by which a more efficient device may be designed.

Keywords

nMem; yTron; SFQ; shunting; memory; constriction; nanowire

1. The need for a superconducting memory

As the world's computational needs increase, researchers have been looking to superconducting supercomputers as a viable means of breaking the barriers of Moore's law [1, 2]. However, one of the largest challenges facing superconducting computational architectures is the lack of a scalable, low-power memory element. Existing superconducting logic technologies largely rely on Josephson junctions which are extremely fast and power efficient, and whose output takes the form of ultrafast (~ 1 ps), low-amplitude (~ 2 mV) pulses. Due to these output characteristics, integrating Josephson junctions with non-superconducting (e.g. magnetic or CMOS) storage elements has proved to be challenging—but it has been equally challenging to build a scalable all-superconducting memory system which could be compatible with Josephson junction-based input.

adam.mccaughan@nist.gov.

Supplementary material for this article is available online

Presently, most purely-superconducting memory elements have relied on large geometric inductances [3], and as a result have a limited ability to be scaled to sub-micrometer sizes. Recent work on a nanowire-based memory element [4] (called an nMem) has shown that using kinetic inductance may be a way to scale to much smaller memory element sizes than geometric inductances can attain. However, making a superconducting loop with purely high-kinetic-inductance materials necessarily means creating a large loop inductance, which can mean larger required write energies. In this memory element presented here, we have reduced the inductance of the overall loop by using low-kinetic-inductance niobium wiring. Additionally, we have shunted the element's Dayem-bridge-style weak links, allowing us to further reduce the read and write energy, and have characterized the performance of the memory element at rates from 10 MHz to 1 GHz.

In this work we describe and characterize a memory element which should be scalable to sizes on the order of the coherence length of a superconductor, uses fJ-scale read and write energies, features a non-destructive read process, and can be written at >50 MHz with a bit error ratio (BER) below 10^{-6} . This element was fabricated using optical lithography, and uses multilayer construction—it uses both a high-kinetic-inductance layer made from tungsten silicide (WSi) and a low-kinetic-inductance layer made from niobium (Nb).

2. Operation

2.1. Memory operation overview

As shown in figure 1, the loop memory presented here consists of three components: a superconducting loop, a weak-link constriction, and a yTron used for non-destructive readout of the loop state. The loop contains the memory state, where a binary 0 corresponds to a clockwise current circulating in a loop, and a binary 1 corresponds to a counter-clockwise current in the loop. The loop includes a weak-link constriction made from tungsten silicide (WSi) that allowed fluxoids to be added or removed from the loop. The loop also includes an in-line yTron which is used as a non-destructive readout element as described in [5]. The terminal of the yTron that is separate from the loop was used to determine the value of the current flowing in the loop, and thus discriminate between the 0 state and the 1 state. These processes are explained in greater detail below.

In the following descriptions, voltage pulses were generated using a two-channel arbitrary waveform generator. The pulses were carried to a 960 mK cold stage via 50 Ω coaxial lines. Extended circuit diagrams can be found in the supplementary material, available online at stacks.iop.org/SUST/32/015005/mmedia.

2.2. Read operation

The state of the memory loop was read out non-destructively by the in-line yTron. The yTron is a three-terminal device in which currents entering one terminal (the sense arm) induce a change in the critical current of the other terminal (the bias arm). For clarity in this circuit, hereafter we refer to the sense arm as the loop arm, and the bias arm as the readout arm. By placing the yTron loop arm inline with the loop, the current circulating in the loop i_{loop} shifts the critical current of the readout arm. In a separate measurement of an isolated

yTron of identical dimensions, we found the response of the yTron to fit $I_c^y = 0.397i_{\text{loop}} + 142.1 \mu\text{A}$. This fit was extremely linear with a least-squares-fit residual of $1.7 \times 10^{-12} \text{ A}^2$, meaning the shift in the read port critical current was almost exactly proportional to i_{loop} . There is nothing stopping analog determination of the critical current in order to gain a greater bit-depth per element, but for this demonstration we used a binary encoding scheme to simplify the characterization of the BER.

To read the memory state, we applied a voltage pulse to the read port, which induced a read current i_r . When the memory state was a logical 0, then the clockwise loop current flowing in the memory loop increased the critical current of the read port. Under these conditions, no voltage pulse was created. In the case of a logical 1, however, the situation was reversed; that is, $i_r > I_c^y(i_{\text{loop}})$. This resulted in the creation of a hotspot in the readout arm, which forces i_r out of the readout arm and into the output. It is important to note here that although a hotspot was created in the readout arm, no hotspot is formed in the loop arm, thus ensuring the number of fluxoids trapped in the loop remains constant. To verify this non-destructive property, we performed a single write operation followed by 10^7 100 ns-long read operations and did not observe a change in the memory state.

2.3. Write operation

The write operation was performed by applying a positive or negative voltage pulse to the write-input terminal, generating a current i_w , shown in figure 1. Since L_y was four times larger than L_w (524 pH and 131 pH respectively), the majority of this current was inductively directed into the write constriction. The purpose of the write pulse was to exceed the critical current of the constriction so that fluxoids would be added or removed from the loop and thus change the value of i_{loop} . A voltage state was only formed in the constriction if the sum of the write and loop currents exceeded the critical current of the write constriction I_c^w . Specifically, fluxoids were only added to the loop if $ai_w > I_c^w + i_{\text{loop}}$, where a is equal to $L_y/(L_w + L_y)$. When the system was at rest, the finite critical current values of both the yTron and the write constriction meant the range of i_{loop} was restricted. Since the yTron was symmetric, both arms had a critical current I_c^y of $142.1 \pm 1.1 \mu\text{A}$ (std. dev.), while the write constriction critical current I_c^w was measured to be $227.8 \pm 2.6 \mu\text{A}$ (measured with a slew rate of 0.08 A s^{-1}). As a result, the range of i_{loop} was restricted to $(-I_c^y, I_c^y)$ —after the write constriction switches, i_{loop} could increase until it reached I_c^y , at which point the yTron loop arm could switch also, expelling some amount of fluxoids and reducing i_{loop} to $\leq I_c^y$. Given the possible range of initial values for i_{loop} , in order to guarantee that fluxoids were added or removed from the loop the magnitude of the write pulse i_w had to be greater than $(I_c^y + I_c^w)$. In future designs, it would be beneficial to ensure that the critical current of the yTron loop arm was significantly larger than that of the constriction—this design rule would ensure that the critical current of the yTron loop arm is not inadvertently exceeded during the write process.

As shown in figure 2, we measured the device response to varying write pulse amplitudes and durations. In this experiment, we reset the memory state by applying a 10 μm -long

negative write pulse of amplitude -2.6 mA, and allowing the element to rest for $10\ \mu\text{s}$ before measuring the critical current of the yTron. The yTron critical current I_c^y was measured by ramping the input current i_r until an output pulse was measured from the read port. By measuring the critical current of yTron, we could then derive i_{loop} . The response map shown in figure 2 reveals the necessary conditions for writing to the element. First, there must be enough current carried in the transmission line connecting the write port to exceed I_c^w . This cutoff is shown as the horizontal line at 6.0 mV (corresponding to $240\ \mu\text{A}$ at the input), below which the write pulse does not alter i_{loop} . Additionally, there must be enough flux applied to charge the inductor L_w , and as a result the amplitude of the applied voltage pulse must be increased at shorter timescales. The dashed line in figure 2 represents a constant-flux line of $3300\Phi_0$. The number of fluxons required to write to this memory element is larger than necessary, primarily due to additional wirebond inductance. Given the approximate wirebond inductance L_{wb} of ~ 3 nH, the flux in the write pulse was divided between the wirebond and the constriction according to L_{wb}/L_w . As a result of this inductive flux divider, approximately $3150\Phi_0$ of the write pulse was used just to charge the wirebond inductance. In addition, the presence of reflections off the $50\ \Omega$ -to-inductor transition also increased the amount of required write flux. Fluxoid storage in the loop was limited by the yTron loop arm critical current I_c^y of $142.1\ \mu\text{A}$, and the maximum number of fluxoids that could be stored in the loop was calculated using $I_c^y(L_w + L_y)/\Phi_0$ and was found to be 46. Each fluxoid contributed $3.1\ \mu\text{A}$ of current to the loop, or equivalently a change of $1.2\ \mu\text{A}$ in the critical current of the yTron readout arm.

The purpose of the shunt resistor R_w on the write constriction was to damp the weak-link write constriction in a process similar to damping a Josephson junction [6]. In similar unshunted devices on the same chip, we observed that the regimes shown in figure 2 were not as clearly delineated. With the addition of the shunt resistor, however, these values became stable as is shown in figure 2. In these thin resistive films a Joule-heated hotspot is usually formed when a constriction's critical current is exceeded. However, in the presence of sufficiently small shunt resistance flux may instead enter a phase-slip-center state [7–9]. In our device, we used a shunt resistance of $0.64\ \Omega$, which we discovered was not quite small enough to induce this change of operation. This change of operation would have meant that the constriction became non-hysteretic (resembling an overdamped weak-link Josephson junction), but we observed that the I - V curves for our read and write constrictions curve were still hysteretic. At present, it is not entirely clear why in figure 2 there is a large decrease in trapped loop current above the horizontal cutoff line. This loss may be the result of overheating the constriction with higher-amplitude write pulses, which could allow fluxoids to slip out before the constriction cools. Alternatively, it may stem from the non-ideal wire-widths of this particular device; in this element, the the loop current was limited by the critical current of the yTron loop arm, rather than the write constriction.

2.4. Error ratio measurements

The BER measurements shown in figure 3 were performed using a two-channel arbitrary waveform generator. Each channel was subject to 30 dB of attenuation before reaching the memory element. One generator channel was used for generating write pulses, the other was

used for read pulses. Additionally, there was a third line connected to the read port which was used to capture the response of the read port. Shown in figure 3(a), each datapoint of BER measurement used alternating read and write pulses of duration τ , separated by delays also of length τ . The number of errors per trial was captured with a counter using the method described in [4]. Each datapoint in figure 3(b) was acquired by writing then reading $N = 10^7$ pseudorandom bits and counting the number of errors n . The two types of errors—write 0, read 1 errors (W0R1), and write 1, read 0 errors (W1R0)—were measured independently from one another while maintaining the pseudorandom write bit sequence. To measure W1R0 errors, we only applied read pulses after a logical 1 was written, and counted the number of logical 0 reads. For W0R1 errors, the situation was reversed. We then calculated the total BER by dividing the sum of the two error types by the number of bits. Confidence intervals for the BER were computed numerically by using a Poissonion approximation to the binomial error process. Specifically, for each BER datapoint we computed the 95% confidence interval by numerically solving for the BER r in the relation $\sum_0^n e^{-rN} (rN)^n / (n!) = C$, where $C = \{0.95, 0.05\}$ was used for the lower and upper bounds respectively.

As seen in figure 3(b), we found that the BER fell dramatically as the read and write times increased above 4 ns, reaching 10^{-6} at approximately $\tau = 15$ ns, and dropping below our measurement sensitivity at 40 ns. This is commensurate with the thermal time constants seen at 1 K in WSi-based superconducting nanowire single-photon detectors, which typically malfunction if operated faster than 20 ns [10]. This suggests that although the constrictions were shunted, they were not shunted sufficiently to reach a true phase-slip regime, and instead created hotspots, limiting their operational speed to approximately the thermal time constant of the material. The amount of energy input into the cryostat for each read pulse and write pulse at a data rate of $\tau = 20$ ns was 5.3 fJ and 39.4 fJ, respectively—the actual energy received by the device was less due to impedance mismatches and cable losses. In particular, we note that the amount of energy actually absorbed by the device is likely significantly smaller—since the device has inductive inputs but the connections to it are 50 Ω coaxial cables, a large fraction of the power in the input pulses was reflected. The termination mismatch also meant that the current induced in the device was twice that of the current carried in the cabling—the values induced in the device are the ones reported here.

Shown in figure 4 is the effect of changing the read pulse voltage on the two types of errors. We found that at lower datarates ($\tau = 100$ ns), there was a margin in which both types of errors fell below the measurement sensitivity. At higher datarates ($\tau = 10$ ns), however, the margin disappeared and the write-read process could not achieve a BER lower than 10^{-4} . It is likely that the plateau seen in figure 4 for the 10 ns data (W1R0) is a reflection of the thermal time constant, as for datarates above $\tau = 20$ ns the plateau disappears.

2.5. Fabrication

These devices were fabricated entirely using optical lithography with an I-line (365 nm) stepper. We first deposited a uniform thin (4nm) WSi film on a 75 mm Si wafer with 150 nm of thermal oxide. The thin film had a sheet resistance of $643.5 \Omega \text{ sq}^{-1}$. We then created a liftoff pattern using photolithography, and sputtered 80 nm of niobium to create the low-

inductance wiring layer. After liftoff of the niobium, we then patterned and etched the WSi layer. An etch was performed in an ICP etcher using a ratio of 40:1 sccm of Ar: SF₆. The resistors were deposited last, again using a liftoff process. We patterned the resistor layer in SPR-660 photoresist and deposited 5 nm Ti followed by 360 nm of AuPd.

Test structures were created on the sample to characterize the properties of the various material layers. We found that the resistance of the AuPd write constriction shunt resistor was equal to 0.64 Ω . The critical current of the WSi film was measured by testing the critical current of various-width constriction test structures (ranging from 0.5 to 2.0 μm) and was found to be 12.0 nAnm⁻² (1.20 MA cm⁻²). Using a resonator structure patterned on the same sample, we found the inductance of the WSi film to be 262 pH sq⁻¹.

3. Discussion

Due to the presence of the thin-film superconductor in the loop, the majority of the inductance of the loop is kinetic inductance, rather than geometric inductance. The presence of this inductance is what allows this type of loop memory to potentially scale; as the scale of the device is reduced, the kinetic inductance of the constriction will remain approximately constant. This is in contrast to conventional superconducting electronics which use geometric inductances to achieve the necessary inductance for fluxoid quantization in a superconducting loop. This device was implemented and tested as a precursor of a random access memory cell and so the test layout was rather large—it was not designed to minimize area. However, the operating principles remain the same as the device area shrinks, with a few caveats.

Although the loop of the device may be scaled to sizes approaching the materials' superconducting coherence length and maintain its inductance values, there are other factors that must be considered when analyzing the scalability of the device. One of these is the thermal limitations of using a Dayem-bridge-style weak link. At sufficiently small scales, the process of readout in the readout arm may generate enough thermal energy such that the superconductivity of the loop is broken. However, with the presence of a shunt resistor on the weak-link constriction, the amount of power generated within the constriction can be significantly reduced. As the constriction width shrinks, this becomes increasingly true. When a constriction reaches its critical current i_c and induces a vortex, the crossing energy dissipated in the constriction is fixed ($i_c \Phi_0$). However, a shunt resistor can raise the effective drag η_{eff} of the vortex motion, increasing the total transit time and reducing its dissipation per unit time [11, 12]. When $i = i_c$, the crossing time for the vortex in a thin-film constriction is on the order

$$\tau_0 = \eta_{eff} \frac{w^2}{\Phi_0 I_c} = \frac{\Phi_0}{I_c} \left(\frac{(w/\xi)^2}{2\pi R_{\square}} + \frac{1}{R_{sh}} \right), \quad (1)$$

where w is the constriction width and was 4 μm in our case, ξ is the coherence length of the constriction material and was estimated to be 7 nm [13]. $R_{\square}$ is the sheet resistance which was measured to be 643.5 Ω sq⁻¹, and R_{sh} is the constriction shunt resistance, which for the write constriction was defined by R_w as 0.64 Ω . For this device, τ_0 was equal to 722 ps, but

would still have been 709 ps without the shunt—clearly the shunting would need to be increased to spread out the energy dissipation over a longer time period in this thermally-slow WSi material.

Although this device had its write constriction shunted, it did not quite reach the non-hysteretic regime that would be necessary to add fluxoids one at a time to the loop. This was well-evidenced by the pulse tests shown in figure 2, in which write pulses either added zero fluxoids, or added many flux-oids. As a result, it is difficult to determine whether further exploration of the well-shunted regime (or lower critical current materials) may reveal that it is possible to operate such a memory at significantly higher datarates or in the single-fluxoid regime.

This memory element has all of the features necessary to incorporated directly into the single-flux quantum circuits currently being investigated for superconducting supercomputers. It can be read from and written to with mV-scale pulses such as those produced by Josephson junctions, and has low enough write and read energies to be compatible with cryogenic computing systems. Although the results presented here are not on the same data rate scale as Josephson junction-based computers (which can reach the tens or hundreds of GHz), there is much room for improvement: either by reducing the thermal time constant (e.g. through a different material like NbN), or through avoiding the thermal runaway process (e.g. by applying single-flux-quantum read/write pulses or optimizing the shunt resistances).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Qing-Yuan Zhao and Reza Baghdadi for helpful discussions. Adam McCaughan was supported by a fellowship from the National Research Council. Emily Toomey was supported by a grant from the Cryogenic Computing Complexity (C3) program and by the National Science Foundation Graduate Research Fellowship Program under grant no. 1122374. This research is based on work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Contract No. W911NF-14-C0089. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the US Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. This is an official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

References

- [1]. Holmes DS, Ripple AL and Manheimer M A 2013 Energy-efficient superconducting computing power budgets and requirements IEEE Trans. Appl. Supercond 23 1701610
- [2]. Kirichenko AF, Mukhanov OA and Brock DK 1999 A single flux quantum cryogenic random access memory ISEC99 (Extended Abstracts ISEC) (Berkeley, CA, 6 1999) pp 124–7
- [3]. Numata H, Nagasawa S and Tahara S 1997 A vortex transitional memory cell for 1 Mbit/cm² density Josephson RAMs IEEE Trans. Appl. Supercond 7 2282–7
- [4]. Zhao Q-Y, Toomey EA, Butters BA, McCaughan AN, Dane AE, Nam S w and Berggren KK 2018 A compact superconducting nanowire memory element operated by nanowire cryotrons Supercond. Sci. Technol 31 035009

- [5]. McCaughan AN, Abebe NS, Zhao Q-Y and Berggren KK 2016 Using geometry to sense current Nano Lett 16 7626–31 [PubMed: 27960481]
- [6]. Toomey E, Zhao Q-Y, McCaughan AN and Berggren KK 2018 Frequency pulling and mixing of relaxation oscillations in superconducting nanowires Phys. Rev. Appl 9 064021
- [7]. Brenner MW, Roy D, Shah N and Bezryadin A 2012 Dynamics of superconducting nanowires shunted with an external resistor Phys. Rev. B 85 224507
- [8]. Lau C, Markovic N, Bockrath M, Bezryadin A and Tinkham M 2001 Quantum phase slips in superconducting nanowires Phys. Rev. Lett 87 19–22
- [9]. Baranov VV, Balanov AG and Kabanov VV 2013 Dynamics of resistive state in thin superconducting channels Phys. Rev. B 87 174516
- [10]. Verma VB, Lita AE, Stevens MJ, Mirin RP and Nam SW 2016 A thermal avalanche in bilayer superconducting nanowire single-photon detectors Appl. Phys. Lett 108 131108
- [11]. Bulaevskii LN, Graf MJ, Batista CD and Kogan VG 2011 Vortex-induced dissipation in narrow current-biased thin-film superconducting strips Phys. Rev. B 83 144526
- [12]. Bulaevskii LN, Graf MJ and Kogan VG 2012 Vortex-assisted photon counts and their magnetic field dependence in singlephoton superconducting detectors Phys. Rev. B 85 014505
- [13]. Zhang X, Engel A, Wang Q, Schilling A, Semenov A, Sidorova M, Hubers H-W, Charaev I, Ilin K and Siegel M 2016 Characteristics of superconducting tungsten silicide W_xSi_{1-x} for single photon detection Phys. Rev. B 94 174509

Figure 1.

(a) Equivalent circuit of the memory element. Circuit element values are $R_r = 3.6 \, \Omega$, $R_w = 0.64 \, \Omega$, $L_w = 131 \, \text{pH}$, and $L_y = 524 \, \text{pH}$. (b) Micrograph of the memory element. False-color highlighting has been applied to the AuPd resistors (red) and WSi elements (light blue). The uncolored areas are the low-inductance Nb wiring—the inductances of the light blue WSi wiring are much larger than that of the uncolored Nb areas.

Figure 2.

Write response map showing the effect of write pulse amplitude and write pulse length on memory state. The horizontal dotted line shows the minimum required write amplitude, while the dashed line represents a constant-flux line.

Figure 3.

(a) Input timing diagram and oscilloscope output from readout used during BER measurements. Shown are the write pulse and read pulse timings for $\tau=20$ ns, and the corresponding output generated by the readout pulse. In this experiment, the write pulses generated to $397.2 \mu\text{A}$ of current input to the device, and the read pulses corresponded to $145.5 \mu\text{A}$ of input. (b) BER versus data rate T. Bars represent the upper and lower 95% confidence intervals. The dotted line represents the measurement sensitivity threshold.

Figure 4. BER for each error type (W1R0 and W0R1) as a function of read pulse amplitude, showns for two different datarates $\tau = 100$ ns and $\tau = 10$ ns.