Development of a Scalable Superconducting Memory

by

Brenden A. Butters

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
August 31, 2018

Signature redacted

Certified by .....
Karl K. Berggren
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Signature redacted

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Development of a Scalable Superconducting Memory

by

Brenden A. Butters

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering

Abstract

Superconducting computers promise very high computation speeds while also consuming far less power than their conventional counterparts. However, much of the progress in this field has been stymied by the lack of a scalable superconducting memory technology. In this thesis, I present the design of, and demonstrate the operation of, a superconducting nanowire-based memory cell. In contrast to existing designs, this cell operates by means of kinetic rather than geometric inductance. Thus, the cell size can be made much smaller than would otherwise be possible. With the successful operation of the single cell, paths to larger arrays are explored, and a small array demonstrated. The further development of the technology demonstrated in this work will allow for the production of large-scale superconducting processors, and the eventual development of superconducting supercomputers.

Thesis Supervisor: Karl K. Berggren
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Much of the work presented in this thesis would not be possible without the efforts of many of my colleagues. I would like to acknowledge:

My advisor, Prof. Karl Berggren, for his insights, his encouragement, and his dedication to the development of his students. His enthusiasm for our work really kept the project going – even when everything seemed to be going wrong.

Adam McCaughan, for his development of the nanocryotron devices, used extensively in this work, and for his initial efforts towards the superconducting memory project.

Qing-Yuan Zhao, for his contribution towards the non-destructive memory covered in chapter 2, and for the fabrication of the devices covered in chapters 2 and 3.

Reza Baghdadi, for the contributions towards, and fabrication of, the devices presented in chapter 4.

Emily Toomey, for her helpful discussions, and extensive experimental assistance.

Murat Onen, for his help with simulating designs, and experimental assistance.

Dorothy Fleischer, for helping keep me organized, and for her ever-helpful attitude and generosity.

Di Zhu, Marco Colangelo, and Andrew Dane, for their input, collaborations, and experimental assistance.

The work that I conducted in pursuit of this thesis, and other in other related projects, was only made possible by the support and advice that I received from my family, and friends. In particular, I would like to thank:

Dr. Raad Raad, for supporting my research, for all his help and encouragement, and for the many hours of discussions we have had over the years. Without his influence I would likely not have made it to where I am today.

Tony and Joe – truly the best MIT has to offer.

Finally, I would particularly like to thank my parents for their unwavering support and dedication. I am forever grateful for the sacrifices that they have made in order for me to pursue the path that has lead to this thesis being written.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Superconductivity and superconducting circuit components

Superconductivity was discovered in 1911 by Heike Kamerlingh Onnes [1]. Onnes was attempting to extend the resistance-temperature curve of mercury to lower temperatures using liquid helium which had just recently been produced. In his experiment, he found that his resistance measurements would suddenly drop to an immeasurably small value when the sample was cooled to below a certain temperature. What Onnes had just observed is the transition of mercury from a normal metal to a superconductor. In subsequent years, other materials were found to exhibit this same property; in particular niobium nitride (NbN), the material used in the experiments presented in this work, was found to superconduct at a temperature above 14 K in 1941 [2].

As stated above, superconductors are characterized by their ability to carry a current without any resistive dissipation, when cooled below their transition temperature. Below this temperature, the resistance of these materials is exactly zero. Superconductors should not be confused with perfect conductors. A hypothetical perfect conductor would have zero resistance and requires a constant magnetic flux within the conductor [3]. A superconductor, on the other hand, shares the zero resistance property, but imposes a strict requirement that the magnetic field within

the bulk must be exactly zero. In addition to this distinction, superconductors can only maintain their superconductivity provided a number of conditions are met, in particular restrictions on temperature, current density, and magnetic field exist. Additionally, the bulk properties of superconductors do not hold at the surface, or in thin films. These limitations are primarily a consequence of the finite number of superconducting charge carriers. In a perfect conductor none of these limitations exist.

Superconductors possess many properties that behave non-linearly. For example, the transition between the superconducting state and normal state is highly non-linear. The bulk material changes from a resistance of zero, to a resistance typically 10^{-3} to 10^{-1} that of its room temperature resistance, by only changing the temperature a fraction of 1 K. The existence of such non-linearities, and the fact that these can be modulated electrically, means that it is possible to build useful electronic components with superconductors. The following sections outline a number of these devices and their basic operation.

1.1.1 Josephson junctions

The Josephson junction (JJ) has been the main device used in superconducting circuits from the device's inception, to the current day [4]. These devices have a relatively simple construction. As shown in figure 1-1, a JJ comprises of three layers. The outer two layers are superconductors, and the inner layer is an insulator to form a superconductor-insulator-superconductor (SIS) junction, or a normal metal to form a superconductor-normal-superconductor (SNS) junction. There exists other constructions which feature more complex layer structures. While such devices are typically also referred to as JJs, we will only refer to SIS and SNS junctions as JJs.

In 1962, B. D. Josephson predicted that paired electrons (Cooper pairs), which are the means by which a superconductor carries a supercurrent, could tunnel through a thin, non-superconducting barrier [5]. Further, Josephson predicted that such a current could flow with no dissipation, that is, with no voltage drop – up until a certain critical current I_c. This operation is referred to as the DC Josephson effect.

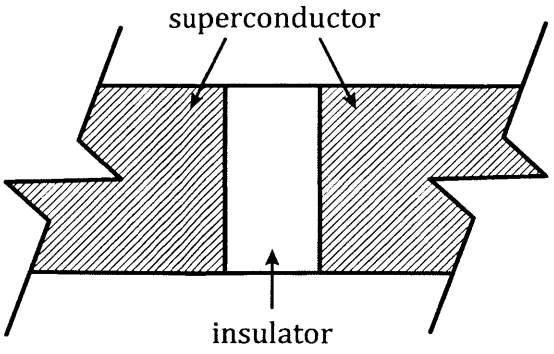

Figure 1-1: Schematic showing the basic structure of a Josephson junction. The distance between the two superconducting materials must be very small in order for tunneling to occur. When an insulator is used to separate the superconductors the junction is referred to as a SIS junction. It is also possible to use a normal metal, in place of the insulator, to crease a SNS junction, it is even possible to use a weaker superconductor to form a SsS junction.

This operation cab be seen graphically as the vertical line from $-I_c$ to I_c, in figure 1-2. If one were to bias a JJ with a voltage V, then the junction would begin oscillating at a frequency of $f = 2eV/h$, where e is the charge of an electron, V is the applied voltage, and h is Plank's constant. This operation is referred to as the AC Josephson effect [6].

Superconductivity is one of the most notable examples of a macroscopic quantum phenomenon. The supercurrent carrying Cooper pairs, consisting of two electrons, can be considered a compound boson. Treating Cooper pairs as such allows them to be described by the quantum mechanical wave function $\Psi = \Psi_0 \exp(j\theta)$, where θ is the temporal and spatially varying phase. The lack of a superconducting link between either side of a JJ allows for it to sustain a phase difference $\varphi := \theta_2 - \theta_1$, referred to as the JJ's phase. Josephson predicted that the supercurrent through the junction would be

$$I_s = I_c \sin(\varphi). \tag{1.1}$$

The superconducting wave function's phase can be related back to the voltage that

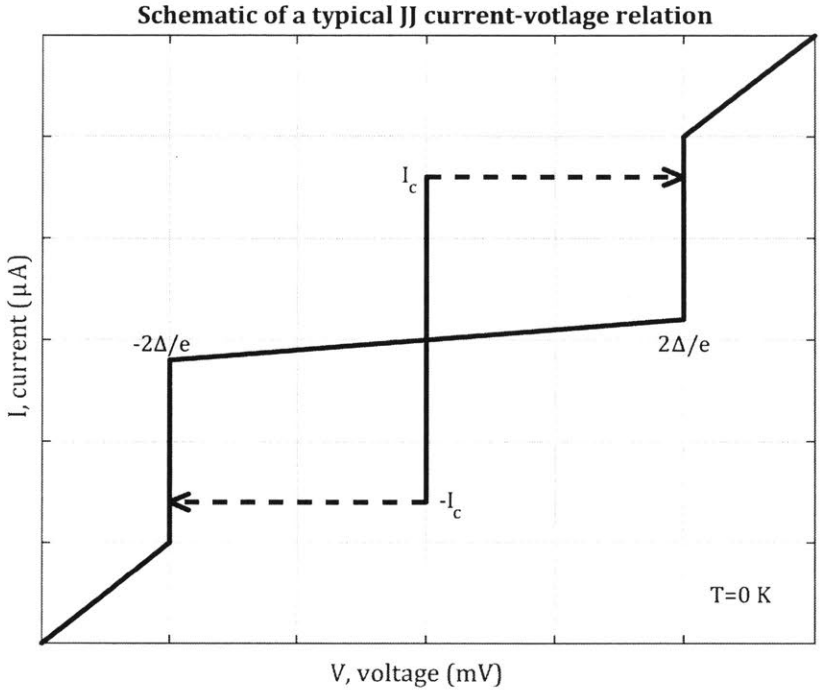

Figure 1-2: Current-voltage relation of a typical Josephson junction. The junction can pass a current with no voltage drop provided the current is less than the critical current I_c. When the critical current is exceeded the junction switches operating mode and presents a voltage $2\Delta/e$ where Δ is the superconducting gap, and e is the elementary charge. If the current is increased the junction behaved as a resistor. Upon decreasing the current, when close to zero the junction will return to a zero voltage drop. Note that the IV curve is rotationally symmetric, and that structure of the junction is intrinsically symmetric.

is developed across any two points in a superconductor (and across a JJ) as

$$V = \frac{\Phi_0}{2\pi}\frac{d\Delta\theta}{dt}, \tag{1.2}$$

where $\Phi_0 = h/2e$ is the magnetic flux quantum, $\Delta\theta$ is the phase difference between the two points of interest, and t is time.

One of the major consequences of superconductors being governed by a wave equation is that, for a supercurrent to be sustained, we must ensure coherence. This means that along any closed path in a superconductor, we must have the phase being zero modulo 2π. This has the most notable consequence that the flux through any

superconducting path is an integer multiple of Φ_0. This can be derived as such:

Extending equation 1.2 to the continuous case we have

$$E = \frac{\Phi_0}{2\pi}\frac{d^2\theta}{dtds}. \tag{1.3}$$

Now, we apply Faraday's law to find that the flux through any superconducting loop is

$$\Phi = -\int\oint E \cdot ds\, dt = -\int\oint \frac{\Phi_0}{2\pi}\frac{d^2\theta}{dtds}\, ds\, dt = -\frac{\Phi_0}{2\pi}\theta = n\Phi_0, \tag{1.4}$$

where $n \in \mathbb{Z}$. It should be noted that this holds for within the superconductor as well as for superconducting loops around a non-superconducting region.

1.1.2 Superconducting nanowires

Currently, the JJ is the dominant device for building superconducting circuits; however, devices based on the thermal switching of small superconducting wires are increasingly being considered for applications traditionally dominated by JJs. These small wires can have dimensions on the nano-scale, and are thus referred to as nanowires. The operation of these devices relies on the fact that superconductivity can only be sustained provided three conditions are met:

1. The temperature T is below its critical value $T_c(J, B)$,
2. The current density J is below its critical value $J_c(T, B)$, and
3. The magnetic field density B is below either $B_c(T, J)$ for a type I superconductor, or below $B_{c2}(T, J)$ for a type II superconductor.

It should be noted that each of these three critical values depends on the other two parameters. Superconducting nanowires take advantage the fact that if any of these three limits is exceeded, it will invoke a phase transition from the superconducting to the normal state. Once in the normal state, the device will exhibit a resistance – as any normal metal would.

1.1.3 Kinetic inductance in nanowires

All conductors have kinetic inductance [7]. The kinetic inductance L_{k}, is calculated by equating the kinetic energy of the charge carriers, with an equivalent inductive energy. For a normal metal this gives

$$\frac{1}{2}(mv^2)(nlA) = \frac{1}{2}L_{\mathrm{K}}I^2 \tag{1.5}$$

which, combined with $I = evnA$ yields

$$L_{\mathrm{K}} = \frac{ml}{Ane^2}, \tag{1.6}$$

where m is the charge carrier mass, v is the average carrier velocity, n is the density of carriers, l is the length of the conductor, A is the cross-sectional area of the conductor, and I is the applied current. For a superconductor, we have that the charge carriers are Cooper pairs. Thus, we have the same expression as above with the substitution $n \rightarrow n_{\mathrm{s}}$, $m \rightarrow 2m$, $e \rightarrow 2e$, which yields

$$L_{\mathrm{K}} = \frac{ml}{2An_{\mathrm{s}}e^2}. \tag{1.7}$$

The expressions in equation 1.6 and equation 1.7 are very similar, so one might expect that a superconductor and a normal metal would have similar kinetic inductance values. This assumption holds for many materials; however, for some materials, such as NbN, this is certainly not the case. The reason that kinetic inductance is typically not readily observed in normal metals is that for most materials the carrier density is very high, and so L_{K} is very small in comparison to the geometric inductance. The kinetic inductance of a normal metal only becomes large when the carrier density becomes very small, as this means that for the same current the velocity of the carriers must become very high. For a normal metal with such a low enough carrier density that a high kinetic inductance is easily observable, the resistance is typically so high that the inductance is negligible contribution to the total impedance, at least for frequencies less than $\sim$ THz. In contrast, a superconductor will have a zero

resistance provided there exists at least one Cooper pair. Thus, materials with low carrier concentration, and hence high kinetic inductance, can be used as to create inductances in circuits.

From examination of equation 1.7, we can see that L_K has the same dependence on l and A as the resistance does for a normal metal. Since we are dealing with thin films, we can use a inductive analog of sheet resistance, that is, a sheet inductance. For thin NbN films we have a sheet kinetic inductance of $L_K^{\square} \approx 14\,\mathrm{pH}/\square$ for a 20 nm film, and $L_K^{\square} \approx 60\,\mathrm{pH}/\square$ a 5 nm film. For most of the circuits presented here, the exact value of inductance is not of particular importance, rather the ratios of the inductance between two branches is important. Thus, it is sufficient to simply design circuit considering only the ratio of the number of squares – which can easily be determined.

1.1.4 Circuit model of superconductivity

As we saw in section 1.1.1 and section 1.1.2, our traditional approach to the design of electronic circuits is not sufficient for the design of superconducting circuits. On top of current and voltage, we need to take into account the superconducting wave function's phase. Like how the complexities of current flow through a resistor can be reduced to a lumped resistance value, the superconducting phase variance across a superconductor can be treated as a lumped inductance. Such a simplification, allows for the phase to be treated, along with voltage, as a nodal term. The particulars of superconducting devices, such as JJs and nanowires, can be treated as a black-box, and given a circuit symbol – as is done for conventional circuits. Thus, we can come up with a set of rules for designing and analyzing superconducting circuits as such

1. Kirchhoff's current law (KCL) – the sum of current into any node is zero;
2. Kirchhoff's voltage law (KVL) – the sum of voltage around any loop is zero;
3. Superconducting phase "law" – the sum of phase around any superconducting loop is zero modulo 2π.

As with conventional circuits, care should be taken when applying these rules. It should be noted that these rules are only true when lumped element models are applicable.

1.1.5 The cryotron

In the 1950s, there was a concerted effort to develop the device that would take over from the vacuum tube – in particular for digital computing applications. During this period, Dudley Buck invented an electrically actuated switch that operated by transitioning from a superconducting state to a resistive normal state by means of an applied magnetic field. He published the device along with a collection of logic circuits built with this device that Buck called the cryotron [8].

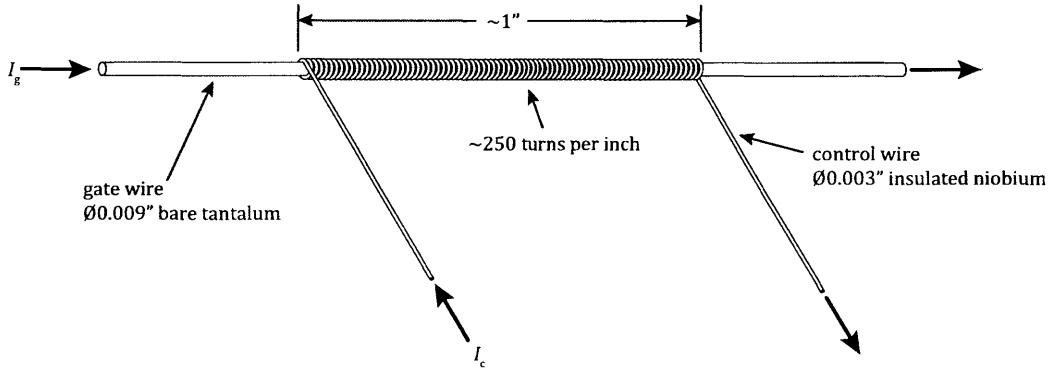

Figure 1-3: Schematic drawing of a typical cryotron used by Buck. The device consists of a $\varnothing 0.009$" uncoated tantalum gate wire wrapped with around 250 turns of a $\varnothing 0.003$" insulated niobium control wire. When a current I_c is applied to the control wire, a magnetic field is induced in the solenoid. This field suppresses superconductivity in the gate wire. As I_c is increased, the gate current I_g that the wire can support without transitioning to the normal region will progressively decrease. A greater suppression of superconductivity occurs in the gate wire than the control wire owing to the gate having a lower T_c, and lower B_c.

A drawing of the cryotron as used by Buck is shown in 1-3. The device operates by passing a control signal through the control wire. The passage of this current leads to the development of a magnetic field within the solenoid formed by the control wire. With the proper design, this field is substantial enough that it can suppress

superconductivity in the gate wire. The gate and control wire see similar magnetic fields. Since it is desired that the control wire remains superconducting at all times, it is necessary for the control wire to be constructed from a superconductor that can withstand a higher field than the gate.

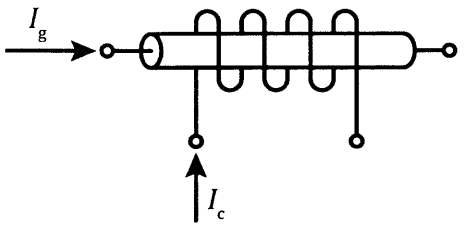

Figure 1-4: Schematic symbol for a cryotron, as proposed by Buck. This symbol is derived from the structure of the cryotron with the control wire depicted as a coil around a thicker conductor, that being the gate. Additionally, the symbol graphically demonstrates that the cryotron does not behave differently if the currents I_c and/or I_g are reversed in direction, as only their magnitudes is of importance.

Buck proposed the symbol shown in figure 1-4, which is inspired by the physical construction of the device. It should be noted that Buck utilized a naming scheme that, with the advent of field-effect transistors, became the opposite of the current conventions. Buck named the switched path of the cryotron, the gate, and the switching path, the control. In a field-effect transistor, the channel is the modulated path and the gate is the modulating potential. Here we will use the field effect naming convention, and only use Buck's convention when discussing the original cryotron.

The cryotron was pitted again the vacuum tube. While the cryotron was smaller and more power efficient than vacuum tubes, this came with many downsides. The cryotron was faster than electromechanical relays, but slower than contemporary vacuum tubes and transistors. With the device requiring cryogenic temperatures to operate, and with its slow speed, it was no match for transistors, which at the time were beginning to gain traction. Thus, the cryotron seemed relegated to the past at least until the advent of the nano-scale planar devices discussed in the following section.

1.2 Superconducting nanowire devices

The cryotron utilized the critical magnetic field density, and a control electromagnet, to induce a transition of a conductive channel from the superconducting state, to the normal state. Similar operations can be performed by exceeding the critical temperature or critical current density – both of these effects were avoided in the original cryotron. With the advent of modern lithography techniques, we can make such devices on the nano-scale. Reduction of the device size has a number of advantages, including reducing power dissipation and increasing the speed at which operations can occur. Devices that utilize the destruction of superconductivity in small features are collectively referred to as superconducting nanowire devices, and as a result of their similarity to the original cryotron, also referred to as cryotrons themselves.

In the context of the work presented here, a superconducting nanowire is a structure patterned into a superconducting thin film. Structure widths on the order of 100 nm and film thicknesses on the order of 10 nm are used here. While a strict definition of a nanowire would typically entail the device being able to be modeled as one dimensional, here we will use a relaxed definition, and consider two-dimensional devices as nanowires.

All of the devices we will be utilizing in this work are all composed of a single layer of superconducting material, in particular thin film NbN, which was deposited by means of reactive sputtering. In the latter part of this work, a second normal metal layer, which is insulated from the NbN by means of an oxide, will be placed on top of the NbN. NbN is a superconductor with a particularly high kinetic inductance in comparison to other materials. Such a high kinetic inductance is particularly useful for our applications as it gives us two desirable properties. First, high a kinetic inductance means that the devices and structures that we build with NbN are particularly insensitive to magnetic fields, as coupling to the loop magnetically is very difficult. Second, the high kinetic inductance means that structures can be very small, as we simply need to consider the number of squares to calculate the inductance, rather than requiring large areas to produce the desired geometric inductance. In fact, the

kinetic inductance of NbN is so much greater than geometric inductance that we can, for most applications, ignore the geometric inductance [9].

With our thin NbN films, it is possible to produce a number of different devices. As all of these devices operate by means of the destruction of superconductivity, and as such, can all be considered switches. The mechanism by which each device switches is dependent on the structure of the device. In the following sections, each of these devices is briefly introduced.

1.2.1 Constriction

While the nanowire constriction is not typically referred to as a cryotron, an understanding of it and its operation, is vital if the other cryotron devices are to be understood. The most trivial example of a nanowire device is a constriction. This device is simply a nanowire that has a region of reduced width, as shown in figure 1-5. The narrow portion of the wire must, of course, carry the same current as the rest of the wire. This means that at the narrow region, the current density will be higher than at other locations in the wire. If the current density in this narrow region is greater than the critical current density J_c, then the narrow region will transition to the normal state, leading to the development of a hotspot.

A sketch of a typical IV curve for a NbN nanowire is shown in figure 1-6. If we consider a nanowire that is in the superconducting state, then we can vary the current through the device within $-I_c < I < I_c$, and the wire will behave as an inductor with zero resistance; that is $V = L(I)i'(I)$, where $L(I)$ is the inductance of the wire at the current i. The inductance of the wire is the combination of a geometric inductance L_g, and a kinetic inductance $L_k(I)$, that is $L(I) = L_g + L_k(I)$. The variation in the kinetic inductance with current is often ignored, as doing so typically will not result in a significant error.

If the nanowire is in the superconducting state, and the current through the device exceeds I_c, then a hotspot will form in the nanowire. The hotspot is a region in which the material from which the nanowire is constructed behaves as a normal material. Once the hotspot has formed, a DC voltage can be sustained across the constriction.

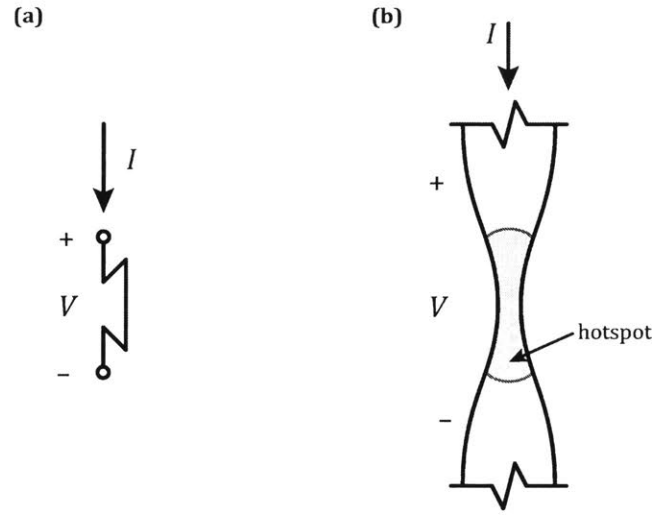

Figure 1-5: Schematic symbol for a superconducting nanowire constriction (a), and a sketch of the layout of a constriction (b). There is no generally accepted symbol for a nanowire, we will use this symbol throughout this work. The constriction shown in (b) has been switched to the "normal" state. While the vast majority of the wire is still superconducting, there exists a hotspot at the narrowest region of the wire – where the current density is the highest. As a current I is being applied to the constriction, a voltage V will be dropped across the normal region. The magnitude of the current I can cause the hotspot to grow or shrink. If the bias I is lowered sufficiently then the hotspot will vanish and the constriction return to the superconducting state. The jog lines indicate where the constriction would be connected to the rest of a circuit.

While the current through the wire is greater than the retrapping current I_r, the device will behave as a resistor (in series with an inductor) – at least in the ideal case, see section 2.4.3 for a description of non-ideal effects. It should be noted that the critical current I_c is the theoretical current at which a breakdown of superconductivity occurs, in practice this current is unobtainable due to many non-ideal effects including noise. The experimentally obtained current at which superconductivity breaks down is called the switching current $I_s < I_c$, to distinguish it from its theoretical counterpart.

When in the normal regime, if we reduce the current through the device to below I_r, then the device will transition back to the superconducting state. The reason that this hysteresis exists is due to the fact that Joule heating occurs within the hotspot. This heating leads to a rise in the local temperature of that region of the wire. Given that the critical current is dependent on the temperature, and in fact, reduces as

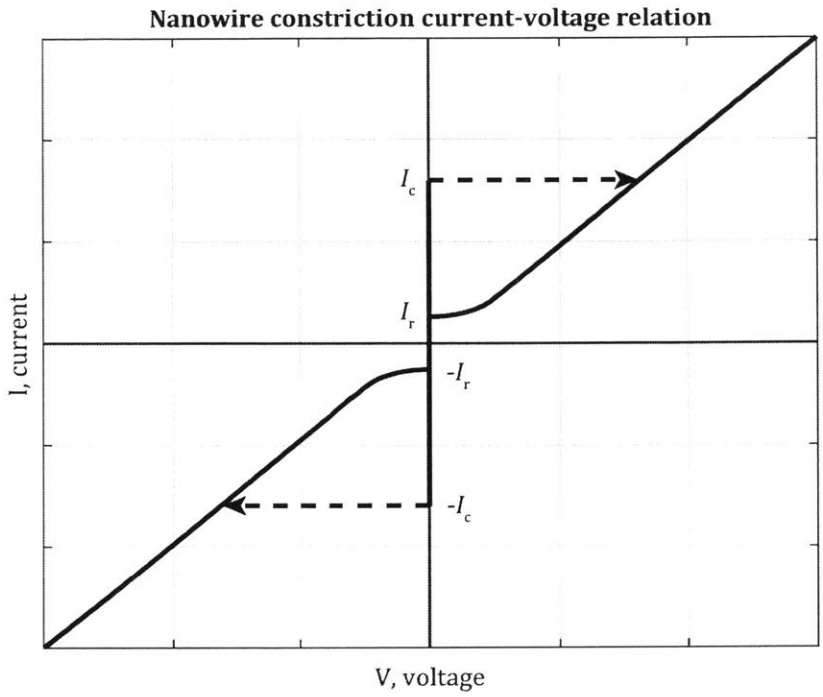

Figure 1-6: Sketch of the current-voltage relation for a single current-biased superconducting constriction. The constriction can carry any current without developing a voltage, provided that the magnitude of the current is less than the critical current I_c. If I_c is exceeded, then the device will enter the resistive state. Once in the resistive state, the device behaves similar to a resistor. From the resistive state, if the current is lowered to below the retrapping current I_r, then the device will return to the superconducting state. Note that the IV curve is rotationally symmetric, and that switching and retrapping only depends on the magnitude of the current.

temperature increases, then the retrapping current is lower than the critical current.

1.2.2 nTron

The nanocryotron (nTron) is a three terminal nanowire device [10]. This device is the first nanowire electrical component (other than the constriction) that was developed. The nTron can be considered a triggerable constriction. The device consists of a relatively wide channel and a narrow gate terminal that is attached to the channel to form a T-shape. The device symbol along with a typical layout are shown in figure 1-7. The channel behaves exactly as a constriction while no current is applied to

the gate. That is, the channel will remain superconducting while the applied current is less than the channel critical current. Provided the gate switching current is not exceeded, then the application of a current to the gate simply adds to the channel current, and provided this total is less than the channel critical current, then the device will remain superconducting. In this regime, the nTron simply behaves as a current combiner.

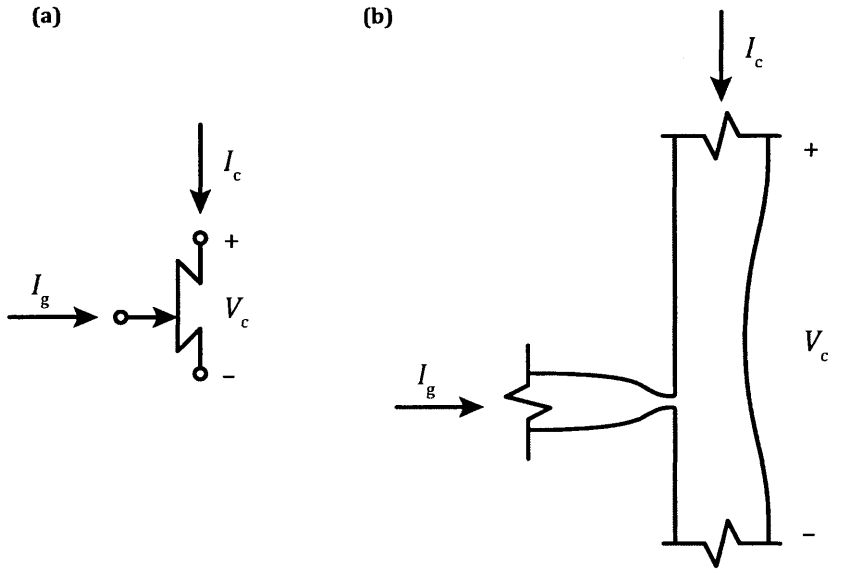

Figure 1-7: Schematic symbol for the nTron (a), and a sketch of a typical nTron layout (b). In the absence of a gate current I_g, a high channel current I_c can be sustained without the device switching and a channel voltage V_c forming. However, in the presence of a sufficient gate current, a hotspot will form at the gate. This hotspot reduces the effective width of the channel and increases the local temperature, thus leading to a reduction in the magnitude of current I_c that can be sustained without the channel switching. The narrowest region of the channel is located to one side of the gate such that during retrapping, the gate will become superconducting prior to the channel. The symbol for the device was designed such that it depicts which side of the narrow region of the channel the gate lies.

If the gate current is high enough that the gate constriction switches, then a substantial reduction in the channel switching current occurs. This switching current suppression is due to the fact that the gate constriction is in close proximity to the channel. This close proximity, in turn, means that heating that occurs at the gate constriction will lead to two effects. First, the gate hotspot can begin to expand

into the channel, thereby reducing its effective width. Secondly, the rise in the local temperature of the channel will lead to a reduction in the channel's switching current. This suppression of the channel switching current is plotted in figure 1-8. If the current being carried by the channel is less than the channel switching current with no gate bias $I_{c,s}(0)$, but greater than the plateau seen in figure 1-8, then the application of a sufficient gate current will lead to the switching of the channel.

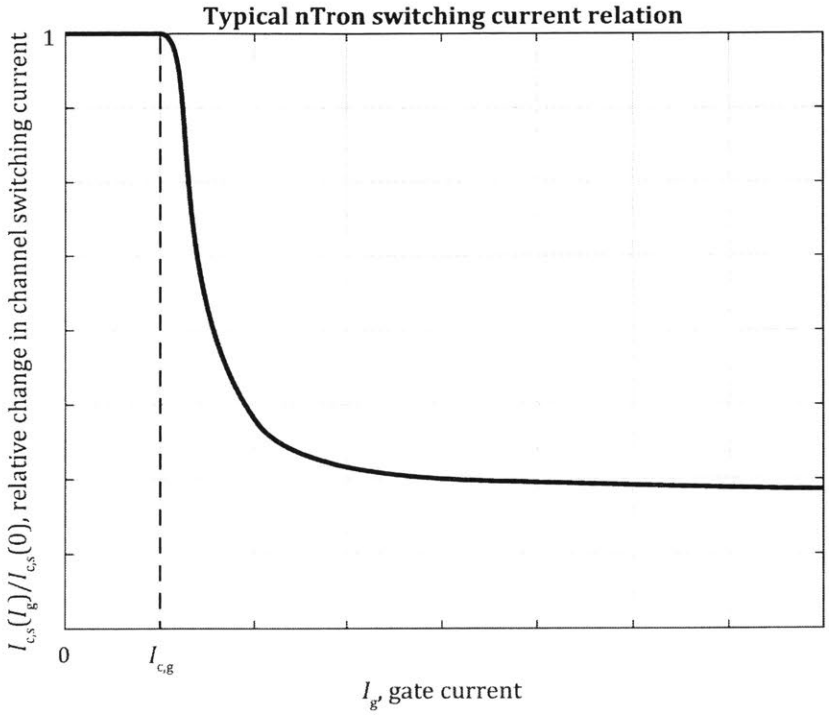

Figure 1-8: Sketch of a typical nTron suppression curve. The switching current of the channel $I_{c,s}$ is a function of the gate current I_g. The switching current of the channel is largely unaffected by the gate current while the gate is superconducting. That is, when the gate current is less than its critical value $I_{c,g}$ and no hot spot has formed, there is little modulation of the channel switching current by the gate current. Once the gate switches, and a hotspot forms, the channel is rapidly suppressed by the gate current. This rapid suppression leads to an operating region in which the nTron possesses a very high gain. This effect begins to diminish with the application of increasing gate current.

Like a constriction, after the channel switches, the current must be lowered to below the retrapping current in order for it to become superconducting again. During

the process of resetting the channel, it may be preferable to have the gate become superconducting prior to the channel. In order to achieve this operation, the narrowest region of the channel is made offset from the gate. This direction of this offset is typically chosen to be away from the terminal closest to ground. Through this design, when the channel bias is reduced, the hotspot will remain at the narrowest portion of the channel, while the region surrounding the location were the gate is attached has already become superconducting again.

The nTron can be seen to be similar to the original cryotron in that they are both electronically controlled switches. They differ in that the nTron is a three terminal device, and the cryotron is a four terminal device, and the nTron is thermally actuated, whereas the cryotron is magnetically actuated.

1.2.3 hTron

The heater-nanocryotron (hTron) is the thermal analog of the original cryotron [28]. Whereas the original cryotron utilized a magnetic field formed by the control current to suppress superconductivity, the hTron utilizes heat generated in a gate. In both devices the gating path (control wire in the cryotron, gate in the hTron) is galvanically isolated from the switched path (gate in the cryotron and channel in the hTron). This allows for more versatility in the circuits that can be constricted with these devices, in comparison to three terminal devices. The schematic symbol for the hTron is shown in figure 1-9.

The channel of the hTron will behave as a normal constriction when no current is applied to the gate. Upon the application of current to the gate – assuming a normal metal gate – heating will occur. This heating will result in a temperature rise of the channel, and thus a reduction in the maximum current it can carry before transitioning to the normal state. A sketch of the suppression characteristics of a hTron with a normal metal gate are shown in figure 1-10. It can be seen that progressively applying more heat to the gate results in a progressive reduction in the channel switching current.

It is possible to build a hTron with either a superconducting gate, or a normal

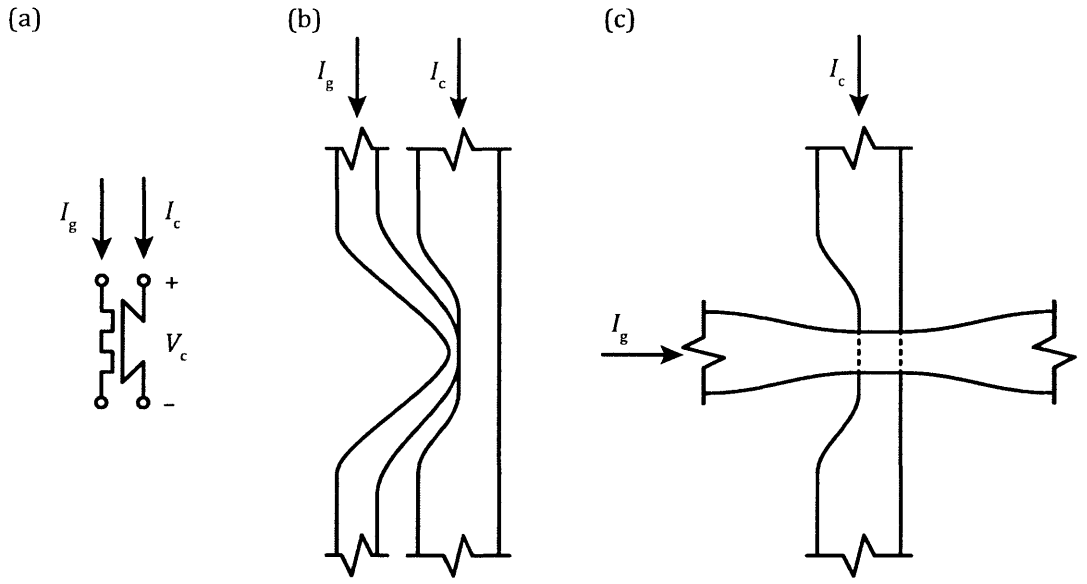

Figure 1-9: The schematic symbol (a) for the hTron, a sketch of an in-plane hTron layout (b), and a sketch of the layout for a multilayer, or stacked, hTron (c). The in-plane hTron is typically constructed from the same superconducting film. This allows the gate to be located relatively close to the channel – within 100 nm. The multilayer hTron can be fabricated with either a normal or superconducting gate. The gate can be within 30 nm of the channel as it is only separated by a dielectric layer. For the in-plane hTron the geometry of the channel and gate are limited by the need to keep them in close proximity, while also avoiding current crowding in the channel. For the multilayer hTron design, the channel and gate can have almost any geometry required, provided there is some location where they overlap – or at last come close to each other. While many different geometries are possible, typically, the gate and channel cross at a right angle, as shown here.

metal gate. Additionally, it is possible to build a hTron with in-plane heaters, or stacked heater. In this work two combinations are used, namely superconducting gates with in-plane heater are used in chapters 2 and 3, and normal metal heaters in a stacked arrangement are used in chapter 4. There are pros and cons of each approach. In comparison, a superconducting heater will produce a very high gain when the gate initially switches – similar to that seen in the nTron. This high gain is a result of the fact that the gate transitions from dissipating zero heat, to suddenly dissipating a substantial quantity with no intermediate level. In contrast, a normal heater allows for continuous control of the channel switching current, as seen in figure 1-10. For an in-plane design, the main advantage is that fabrication is simpler, in

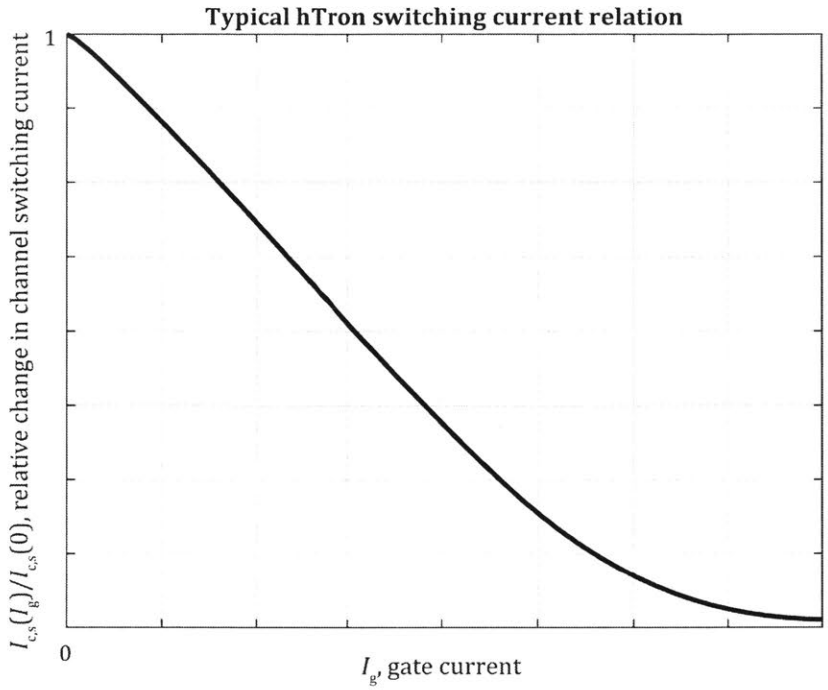

Figure 1-10: Sketch of a typical suppression curve of a hTron with a normal metal heater. As increasing gate current I_g is applied to the device, the channel switching current $I_{c,s}(I_g)$ decreases. After some gate current a regime of diminishing returns is entered where the higher gate currents are required to achieve greater suppression. In most devices normal metal gate hTrons that have been tested experimentally, it takes an very large gate current to totally suppress superconductivity, if it is ever achieved. On the other hand, superconducting gate hTrons can relatively easily achieve total suppression of the gate; however, a similar region of diminishing returns is witnessed. With a superconducting gate, there is a region of zero suppression from $I_g = 0$ to $I_g = I_{c,g}$, the gate critical current, in a similar manner to that of the nTron – see figure 1-8.

particular if the gate is made from the same film as the channel, then they can both be patterned at the same time. The stacked heater, on the other hand, requires a more complex fabrication process with the requirement for some dielectric separator. The advantage of a stacked header design is that the heat transfer is more efficient, and the time delay between the application of gate current and the modulation of channel current is shorter. In the following discussion a normal metal heater is considered.

The retrapping current $I_{c,r}(I_g)$ of the channel is the same as it would be for

an isolated constriction provided $I_{c,s}(I_g) > I_{c,r}$. That is, the retrapping current is not modulated to any great degree by the application of current to the gate. This operation is to be expected since when $I_{c,s}(I_g) > I_{c,r}$, as the heat from the gate is not sufficient to suppress superconductivity without the assistance of channel current. Thus, the temperature rise generated by a hotspot is greater than the temperature rise due to the heater. If, on the other hand, the gate current is sufficient to cause $I_{c,s}(I_g) < I_{c,r}$, then the retrapping current will be $I_{c,r} \approx I_{c,s}(I_g)$. This phenomenon occurs as the result of the fact that the gate is causing a temperature rise greater than that which can be obtained by the existence of a hotspot. So the heater is beginning, or already has, totally suppressed superconductivity by raising the local temperature above the critical temperature – even without the application of channel current. Thus, in general we have the relation $I_{c,r}(I_g) \approx \max(I_{c,r}(0), I_{c,s}(I_g))$, where $I_{c,r}(0)$ is the retrapping current of the channel without any gate current.

1.2.4 yTron

The yTron is a three terminal device that enables the non-destructive sensing of a supercurrent [11]. The device consists of two arms which meet and form a Y-shape – hence the device's name. A sketch of a yTron layout is shown in figure 1-11. When the yTron is used as intended, the sensed current should be unaffected by a read operation occurring. With the correct bias configuration, the yTron can be considered an inverter, where the bias port switches when no current is applied to the sense port, and remains superconducting when a sense current is present.

When a bias current I_b is applied to the bias port it will flow through the yTron, eventually making its way to ground. As it flows down the bias arm, it encounters the join between the two arms. At this point, it will experience current crowding (provided no bias is applied to the sense port). Current crowding occurs since the current sees an abrupt change in the wire width. As a result of this change, the current will attempt to take the shortest path, thus leading to a high current density on the yTron corner. The high current density on this corner results in a low switching current for the bias port. The current to be measured is applied to the sense port.

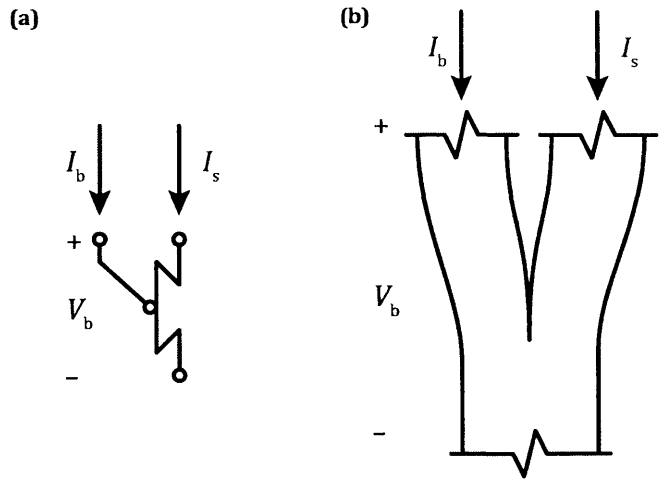

Figure 1-11: Schematic symbol for a yTron (a), and a sketch of a typical yTron layout. The sense current I_s, is the current to be measured non-destructively. If the yTron is operated correctly, this current should never be interrupted. The sense current is determined by applying a bias current I_b while monitoring the bias voltage V_b. With a low sense current, a relatively low bias current is required to switch the bias port, due to the high current crowding that occurs on the yTron corner. On the other hand, when a high positive sense current is applied, the bias current that is required to switch the bias port is comparatively higher than in the low bias case. This increase in switching current is due to the decreased current crowding that occurs when the biases applied to the ports are similar.

The sense current, I_s, passes through the yTron and flows to ground via the common port since the bias port presents a high impedance. When the bias current is applied to the bias port with the sense current flowing, a higher switching current is observed. This increase in switching current occurs because the existence of the sense current reduces the current crowding that occurs on the corner of the yTron. We could think about this in terms of a superposition of the crowding caused by the bias current and that caused by the yTron which occur in opposite directions. The opposing crowding effects lead to a comparatively lower current density at the corner than if any one were applied on its own. These effects lead to the sensitivity curve shown in figure 1-12.

Note that in [11] a different convention for the yTron symbol was used. While the operation of the physical device is not affected by if the sense and bias ports are switched (provided their widths are the same), since the layout is symmetric, the

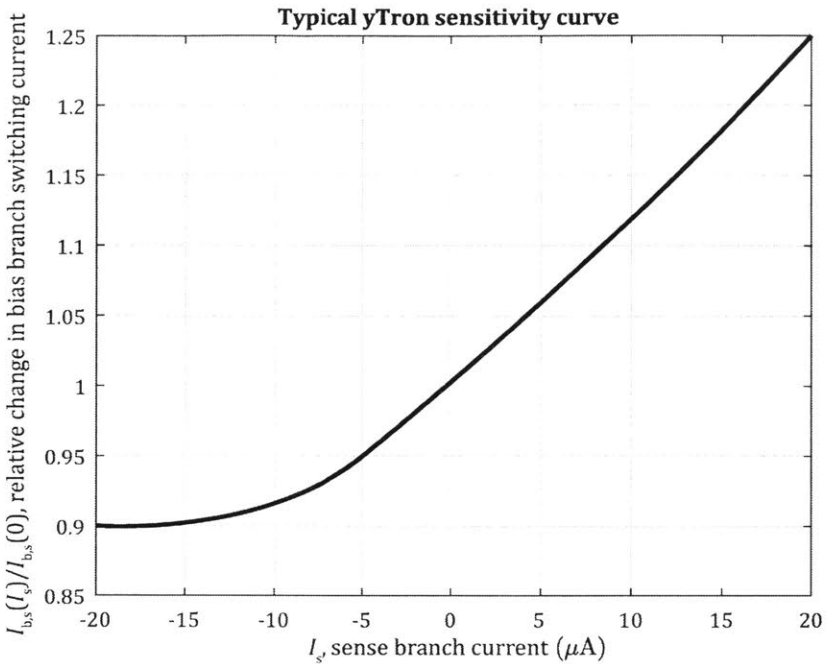

Figure 1-12: Sketch of the sensitivity curve of a typical yTron. This sketch is adapted from experiential results. In contrast to the other nanowire devices presented here, the switching current of the yTron increases with the increased application of an external current – the sense current. This effect will only occur over a region where the sense arm has not switched. Once the sense arm switches, then the yTron begins to behave as an nTron. While the sense arm is superconducting, we have that the bias arm switching current $I_{\mathrm{m,s}}(I_{\mathrm{s}})$ increases with increasing sense current I_{s}. With the application of a negative sense branch current, the switching current can be seen to decrease, but only to around 90% of its zero-bias value. The yTron is intended to operate with positive I_{s}. The sensitivity curve of the yTron is injective, and thus allows for the sense current to be measured without disrupting the supercurrent.

intent as to which arm switches during a read is important for the understanding of the circuit. Here we will use the convention shown in figure 1-11, where the bias-arm is switched and the sense-arm remains superconducting.

1.3 Superconducting memories

In order to gain an appreciation for why the development of superconducting memories is important, we must first consider their application. The primary application

of superconducting memories is for use in superconducting supercomputers [12–14]. Such supercomputers would be constructed with JJs, that are used as very fast (around 1 ps) and energy efficient (on the order of 10^{-19} J) switches. Logic families based on JJs, such as single-flux-quantum (SQF) and similar, could be used to implement extremely fast logic and power efficient logic gates. SFQ was developed in the 1980s and has progressively improved since then. More recently, gates have been demonstrated that can operate at speeds up to 770 GHz [15].

1.3.1 Need for a cryogenic-compatible memory

For a computer, the requirement for memory is fundamental. In order to build a computationally universal (Turing complete) computer, one of two requirements must be satisfied. We either need to have a circuit that is as large as the computation problem is complex. Or, we need to have a computer that, while only being capable of a limited number of logical operations, has access to sufficient memory to complete its task. The choice to pursue the second option is necessary for all but the most elementary of computation tasks.

As identified by [12], there are three main reasons SFQ computers have not been competitive. First, there exists no adequate means by which high bandwidth interconnects between the cryogenic computer and room temperature can be implemented. Second, fabrication limitations surrounding practical production of superconducting electronics. Finally, the lack of a scalable SFQ-compatible memory. Thus, there is a clear need for scalable superconducting memory technology to make major headway towards a superconducting supercomputer.

1.3.2 Existing memory technologies

There exist many memory designs implemented in SFQ [16–19]. Unfortunately, these memories are not suitable for a superconducting supercomputer's RAM. Drawing an analogy from complementary metal-oxide-semiconductor (CMOS) world, SFQ memories are similar to static-RAM (SRAM), whereas we are searching for a dynamic-RAM

(DRAM) technology [20]. Like SRAM, SFQ memories are built from the same logical elements as SFQ logic circuits, and as a result the cells are large, and many require a static bias. The technology we are searching for is akin to DRAM, where cells are extremely small and require no bias applied to retain their state (ignoring refresh cycles).

An additional drawback of SFQ-based memories is that they are typically, in some manner or other, based on a superconducting quantum interference device (SQUID). As such, the loop must be able to hold at least a single quanta of flux in order to store a state. This sets a limit on the minimum allowable cell inductance, and as most SQUIDs are designed to operate with geometric inductance, this limits the cell size to a few microns.

There have been a number of efforts to develop alternate memories, with many focusing on the incorporation of magnetic elements with JJs [21–25]. These designs typically require exotic materials and fabrication processes. Many of these designs are relatively large, and require substantial support electronics to enable array operation. Some of these designs are capable of nano-scale miniaturization; however, they are not purely superconducting, and require substantial process development as they incorporate new complex layer stackups. Superconducting, and in particular JJ, fabrication is a difficult task, with many existing fabrication issues, the last thing the industry needs is more complex materials and processes.

Recognizing the drawbacks of these designs, [26] developed a memory that operates utilizes high kinetic inductance nanowires to implement a cryogenic memory. The drawbacks of this design is the use of external magnetic fields in order to achieve memory operation, and the fact that no method of scaling the memory beyond a single cell is presented.

1.4 Thesis goal

The goal of this thesis is to design and demonstrate the operation of a scalable superconducting memory technology that is applicable to superconducting computers.

We intend to achieve this goal with a superconducting-loop-based memory that operates by means of kinetic rather than geometric inductance. In this work we are primarily interested in the operation of the memory cell, and less the interface to a microprocessor. It has, however, been demonstrated that with our nanowire devices interfacing to SFQ is possible [27].

In pursuit of our goal, two memory technologies were developed. The first of which has a relatively simple cell design, but proved to be difficult to scale. The second, had a deceptively simple cell structure, that proved to be difficult to design, but a exceptionally simple array architecture. Over the course of this work, a number of experimental setups were designed. Principal among winch is an integrated optimizer capable of automatically searching for low-error-rate operating points for our memory experiments.

1.5 Thesis outline

This thesis is broken into six chapters, each of which is summarized in the following sections:

Chapter 2 – Non-destructive readout memory

With the need for a scalable superconducting memory clear, chapter 2 presents some possible approaches to build such a memory. Upon comparison of the possible approaches one design, the non-destructive readout (NDRO) approach, is pursued in detail. Cell design, simulation, experimental setups, experimental results, and analysis are then provided for this cell.

Chapter 3 – NDRO array design

Chapter 3 explores the methods that could be used to build the NDRO cell into an array. Two approaches are explored, one of which is a modification of the NDRO cell that adds a read enable port, and the other of which is a design which multiplexes the cells in columns to form an array. The relative power consumption and area of these

cells is then compared. The design of a superconducting multiplexer is then presented and analyzed. Finally, a prototype multiplexer is fabricated and the experimental results presented.

Chapter 4 – Destructive readout cell and array design

With the complexity of integrating the NDRO cell into an array, a new method was explored which promises to make the array design far simpler at the expense of a somewhat more complex cell design. This new approach relies on the destructive readout of the cell's state, and is as such, is referred to as the destructive readout (DRO) design. This chapter chronicles the development of the DRO cell design, simulation and testing. Included in this process is the development of a new multilayer hTron, and the testing of such devices.

Chapter 5 – Experimental setup and apparatus design

Chapter 5 provides information on the experimental setup used throughout the memory development process. Of particular note is the integrated optimizer which was used to automatically find the optimal operating point in DRO cell and array experiments. Finally, the design and development of new experimental apparatus that will enable and accelerate future work on this, and similar projects, is presented.

Chapter 6 – Conclusion and future work

This chapter provides concluding remarks, and covers topics that could be the subject of future work.

Chapter 2

Non-destructive readout memory

With the need for a scalable superconducting memory established, and with a nanowire devices in hand, we set out to develop a basic memory cell design. Section 2.1 describes the approach to designing a cell with the superconducting devices and fabrication technology we have available. With a suitable design direction selected, the basic theory of operation is presented in section 2.2. To validate the cell design, a number of SPICE simulations were preformed and are covered in section 2.3. After the design was validated, devices were fabricated and tested. Section 2.4 covers the initial experimental procedure and the results of the experiments conducted. With what we learned from these experiments, the design was revised and new devices fabricated and tested. Their performance exceeded the capacity of the initial test setup used in the first experiments, so a new setup was designed. This new setup and the results of these experiments is presented in section 2.5. With the success of the revised design, the results were published in [28]. There exists substantial overlap between the publication and section 2.5, with the paper covering more device operation and metrics, and section 2.5 covering more of the process used to arrive at the final design, and measurement setup.

2.1 Memory approach

There are a number of effects present in superconductors, and other materials that are compatible with superconductors, that can be leveraged to create a memory. Essentially, all that is required to build a memory is some persistent – or reasonably slowly decaying – parameter that we can set, and a means by which we can later interrogate this parameter to determine the state. As we have seen, nanowires are intrinsically hysteric, so one may naively try to build a memory with a nanowire as the storage element.

Building a memory with a nanowire as the storage element seems perfectly reasonable at first. We simply bias the wire at $I_b \approx 0.6 I_c$ (roughly half-way between the critical and retrapping current), where I_c is the critical current. At this bias point we have two stable states, one where there exists no voltage dropped across the device, and one where a constant voltage is present. In this arrangement the memory cell exceptionally simple to operate, one need only apply a positive current pulse to the memory to set it to the "1" state, and a negative pulse to clear it to the "0" state – as shown in figure 2-1. Then the state of the memory is easily determined by reading the voltage across the wire. This memory seems to satisfy all our needs; it is exceptionally small, could be built into an array, and is trivial to operate. However, there is one drawback, the wire will be constantly dissipating power in order to maintain its state. Such constant dissipation is unacceptable for most superconducting computer applications – in particular for a computer's main memory.

Given that we cannot use the intrinsic hysteresis of a nanowire, we must search for alternate sources of hysteresis. One method is to take advantage of a material that exhibits some kind of conductive or magnetic hysteresis, such as memristor or ferromagnetic materials [29, 30]. Neither of these technologies would require a sustained power dissipation to store the state. However, the fabrication of such devices would be far more challenging than a single superconducting layer. Looking further, we can see the great success of charge based memories in the CMOS world. So we might want to implement something similar in a superconductor. Unfortunately,

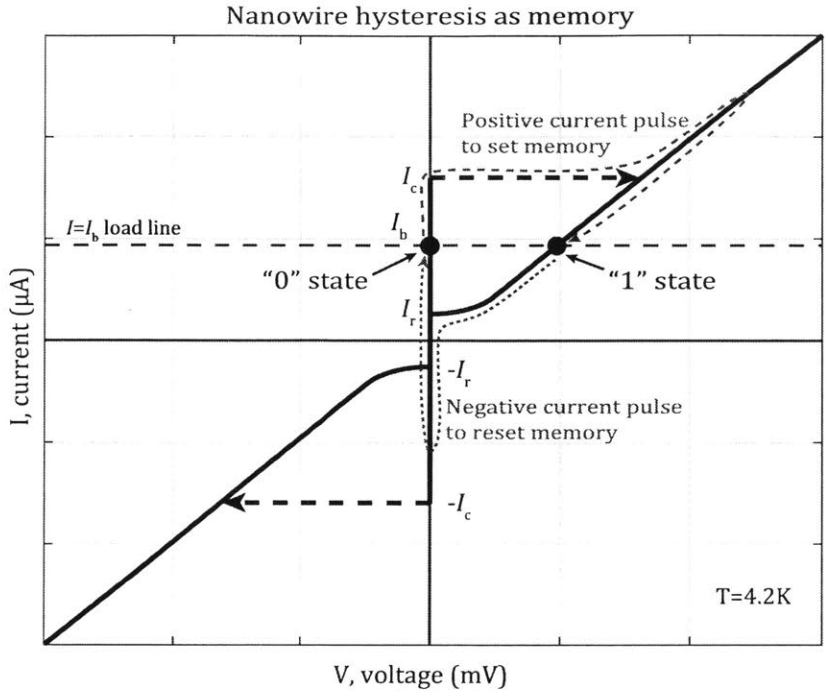

Figure 2-1: Schematic demonstrating how the IV curve of a hysteric nanowire can be exploited to implement a poor memory. Here, the wire is biased to $I_b = 0.6I_c$. At this bias, the load-line intersects the IV curve of the wire at two locations, thus there are two possible states, both of which happen to be stable. To switch between these two states a current pulse is applied. A positive pulse greater than $0.4I_c$ is applied to set the memory into the "1" state. A negative pulse greater than $0.4I_c$ is applied to reset the memory to the superconducting "0" state. When in the "1" state, the cell dissipates $0.36I_c^2R_h$ continuously, where R_h is the hot spot resistance (which is on the order of 200 Ω to 1000 Ω for NbN). For typical values, the power dissipation of such a memory can be estimated to be around 0.25 μW when in the "1" state. Hence, a memory constructed this way would not be acceptable in many superconducting applications.

superconductors, by their very nature, are not particularly sensitive to charge. The cell would require a large enough capacitor such that it could deliver sufficient current to switch a nanowire from the superconducting to the normal state. This would likely lead to a very large cell size, which would not be particularly practical.

With the charge-based operation in mind, we might consider what the dual of CMOS DRAM would be. Instead of utilizing a stored charge to encode a state, one would use a stored current. The storage element, instead of being a capacitor that,

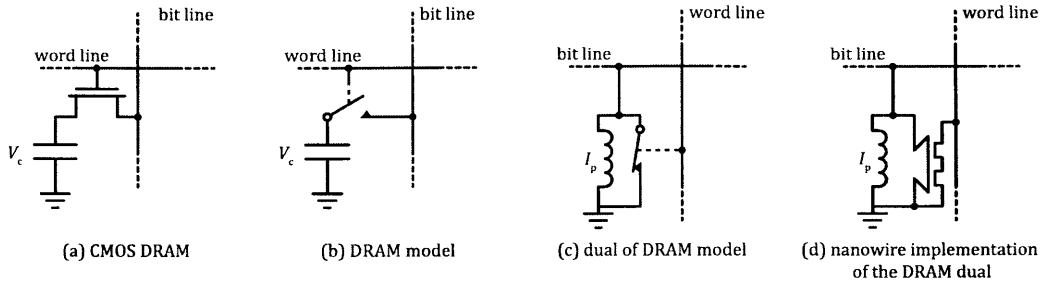

Figure 2-2: Derivation of a superconducting dual of a CMOS DRAM cell. (a) a typical DRAM cell as implemented in CMOS; (b) a simplified model of the CMOS DRAM cell where the enhancement mode MOSFET has been replaced by a normally open switch; (c) the dual of the simplified circuit, note that the switch is now normally closed. The capacitor voltage V_c which previously held the state of the cell is now a persistent current I_p, the magnitude and/or sign of which can now be used to store the state of the cell; (d) a nanowire implementation of the dual circuit where the normally closed switch has been replaced with a hTron.

when the cell not accessed, is left open circuit, would be an inductor that, when the cell is not addressed, would be short circuited, as depicted in figure 2-2. To access a DRAM cell, we use an enhancement mode MOSFET to connect to the capacitor to a bit line from which it is charged or read. In this configuration, the MOSFET can be considered a normally open switch, the dual of which is a normally closed switch. This situation is perfect for us, because all of our nanowire devices are essentially normally closed switches. So in our DRAM dual, we could use a controllable nanowire to open the loop and allow flux to enter and exit the loop. This approach is very promising, especially since it only requires a nanowire switch which in its non-addressed (zero power dissipation state) holds the memory's state. Additionally, we have one major advantage over DRAM. In a real DRAM implementation, the storage capacitor slowly discharges due to leakage; while a superconducting loop will store a supercurrent indefinitely without any loss – unless some part of the loop stops superconducting. Further to this, we can even make the readout of the cell non-destructive by utilizing a yTron's ability to sense a current non-destructively. This design is the first that was pursued in this project, and due to its use of a yTron for readout is referred to as the non-destructive readout (NDRO) cell.

2.2 NDRO cell operating principals

Following from the dual of a CMOS DRAM, the basic design of the NDRO cell consists of a hTron and a yTron, that are arranged such that they form a superconducting loop, as shown in figure 2-3. This loop is the means by which a persistent current is stored. The magnitude and direction of this current determines the state of the memory. When the cell is not being accessed, and is holding its state, the persistent current experiences no decay. Thus, the memory can be considered static. In contrast to CMOS SRAM, once the state is stored in the NDRO cell, all biases can be removed and the state will be retained indefinitely – provided the loop remains superconducting. A consequence of this operation is that the NDRO cell has no leakage and no quiescent current draw. Thus, the NDRO cell combines the advantages of both DRAM and SRAM, with the disadvantage that it must be kept cold in order to achieve this task. Additionally, the NDRO cell, is accessed asynchronously.

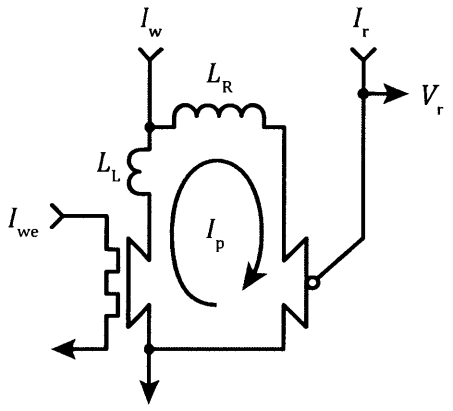

Figure 2-3: Simplified schematic of a basic NDRO cell. The device has three ports, namely a write enable, a write port, and a read port. The loop which carries the persistent current I_p, consists of the channel of the hTron, the yTron, and the two inductors L_L and L_R where $L_L < L_R$. The write enable port is galvanically isolated from the loop.

The hTron is used to write to the cell, and the yTron is used to readout the cell non-destructively. A single cell does not explicitly require a hTron to allow for the cell to be written to. It is possible to replace the hTron with a simple constriction. However, the hTron allows for the cell to be designed into an array – which is covered

in section 3.1. The inclusion of the hTron allows for more control of the current written to the loop. If a constriction were used rather than a hTron, then we could write to the loop by applying positive and negative current to the write port I_w. However, with the hTron, we do not require a bipolar write current. Similarly, for the yTron, we do not require a yTron to readout the memory. We could replace it with a constriction or hTron and readout the memory in the same fashion as a SQUID is read, that is we would read the switching current of the write port. As with a SQUID, the current circulating in the loop determines the overall switching current of the loop, thereby allowing us to determine the state of the persistent current. This method of reading the cell is more complex and requires a much more carefully designed cell. The major drawback of this approach is that reading the cell's state is destructive; however, this drawback comes with the benefit that forming the cell into an array is far simpler. The destructive readout approach is explored in chapter 4.

2.2.1 Writing to the memory

Writing to the memory requires applying two signals to the cell. One signal is the write enable I_{we}, which modulates the switching current of the hTron. The second signal is the write bias I_w. When a current is applied to the write port, it will inductively split between the left and right paths. The majority of the current will take the left path since the cell was designed such that $L_L < L_R$. This current, if sufficient in magnitude, would cause the hTron channel to switch. The hTron channel now presenting a non-zero resistance, will result in the majority of the bias current to be rediverted from the left side to the right side of the loop. Thus, when the write enable is deasserted, and the write bias removed, we are left with some persistent current I_p trapped in the loop.

The exact timing of a write operation is quite flexible. This means that there are a number of configurations that can be used to successfully access the NDRO cell. A timing diagram depicting a write and a read is shown in figure 2-4. The main timing limitation for a successful write to occur, is that the data must be held for the hold-time $T_h > 0$. This time must be greater than zero as the data can be considered

to be latched into the cell on the deactivation of the heater, that is the deassertion of the write enable port. The data hold time needs to be long enough for the channel of the hTron to cool down to around ambient temperature. This process is expected to take less than 100 ps; however, it has not been experimentally determined. Since the data is effectively latched on the falling edge of the write enable signal, there is no meaningful setup time as far as the cell is concerned. Provided the signal pulse is long enough for the correct current to be flowing through the cell, then the write should be successful. The minimum width of the write enable pulse T_w must be large enough for the heat to propagate to the hTrons channel from its heater. This delay is very dependent on the design of the hTron and the level of heat applied to the device.

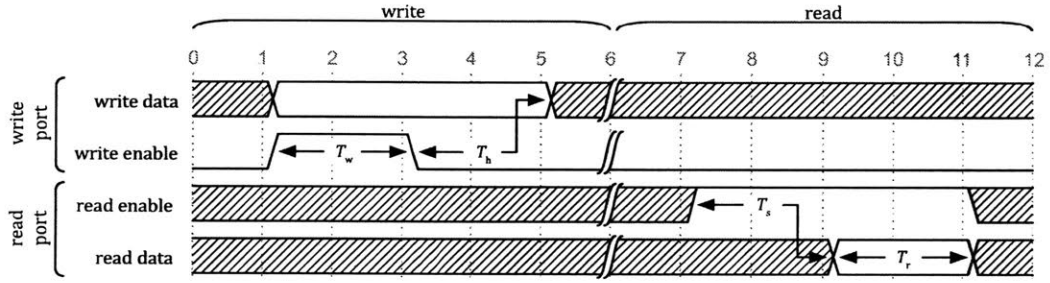

Figure 2-4: Timing diagram for writing to, and reading from, a NDRO memory cell. This diagram summarizes the logical sequence of accessing the memory. The levels of current and their signs will depend on the exact cell design and operating mode.

To illustrate a write procedure, consider a write bias I_W applied to the write port I_w. This current is not sufficient to cause any switching to occur in the loop. Rather, the current will simply inductively split between the left- and right-hand sides of the loop. We then apply a bias to the write enable port that is sufficient to fully suppress the hTron channel. The full suppression of the hTron channel will result in it remaining resistive, and so all of the write bias will be rediverted to the right-hand side. We then remove the write enable signal allowing it to cool back to ambient temperature and begin superconducting again. Finally, we remove the write bias. At this point a persistent current is trapped in the loop.

Of particular interest is the magnitude of the persistent current after a write. To simplify this calculation, we will again consider the case where a write enable signal

that is sufficient to fully suppress the hTron's channel is used. We will further assume that the moment before the hTron cools, the write bias was I_W, and we will ignore flux quantization. Under these conditions, when the hTron cools, all of the write bias is flowing through the right-hand side of the loop. Using the principal of superposition, we can determine the component of the bias current due to the persistent current I_p. First, we will consider the bias current. This current is inductively split between the left and right branches, and so in the left branch we have $I_{L,w} = I_W L_R/(L_L + L_R)$, and in the right branch we have $I_{R,w} = I_W L_L/(L_L + L_R)$. Now we will consider the loop current. In this case, the current will see the write port as open circuit, since it is current biased, and so we have $I_p = I_{R,p} = -I_{L,p}$. Thus, we have the right current will be $I_R = I_{R,p} + I_{R,w} = I_p + I_W L_L/(L_L + L_R) = I_W$, and the left will be $I_L = I_{L,p} + I_{L,W} = -I_p + I_W L_R/(L_L + L_R) = 0$. Solving, we find that both equations are satisfied by

$$I_p = I_W \frac{L_R}{L_L + L_R}. \tag{2.1}$$

Thus, in the ideal case where after a write no current will be flowing through the hTron, then the persistent current will be given by the above equation.

2.2.2 Reading from the memory

Reading from the cell is a relatively simple process. A bias current is applied to the read port (the yTron's bias port). Simultaneously, the voltage at this port is monitored. If the memory is in the "1" state, then no voltage is developed. On the other hand, if the memory is in the "0" state, then a voltage will be developed. The read process does not interfere with the persistent current, and as a result the read process is non-destructive. This process is summarized in the read portion of the timing diagram shown in figure 2-4.

Like the write process, the timing of the read process is fairly flexible. The read timing is characterized by two time periods, namely a setup time and a read time. The read enable must be asserted, that is the read current bias must be applied to the read port, for a setup time T_s before the output data becomes valid. This time is

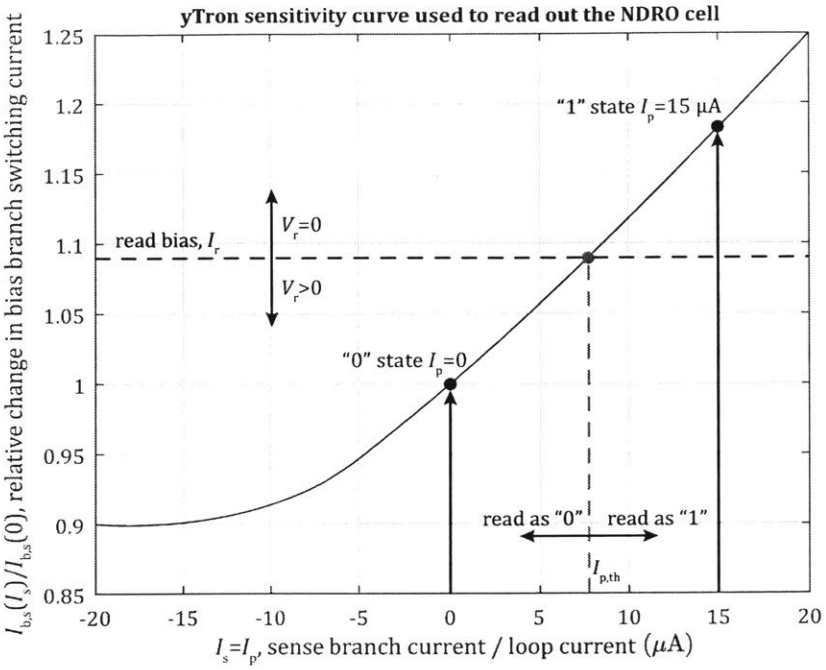

Figure 2-5: Sketch of a typical yTron sensitivity curve based on experimental results, that has been annotated to show the operation of the yTron memory readout. When the yTron is used to read the loop current, we have the sense current is the loop current, $I_s = I_p$. In this example, the yTron is biased at $I_r = 1.09 I_{b,s}(0)$, that is 9% over the zero bias switching current of the yTron. At this point we have the corresponding loop current $I_{p,th}$ is approximately mid-way between the "0" state loop current and the "1" state loop current. Thus, the yTron voltage V_r will be zero if the memory is in the "1" state, since the loop current will be $I_{b,s}(I_p) > I_r$. Conversely, if the memory is in the "0" state, then the loop current will be low and $I_{b,s}(I_p) < I_r$, so the read bias will be sufficient to switch the bias arm of the yTron and so $V_r > 0$.

very short, it is expected to less than 100 ps; however, it has not been experimentally determined. After the setup time, the data can be read, that is the voltage can be sampled. The data will remain valid while the read enable is asserted. There is no known limits on the read time T_r, it can be as short as the external readout circuitry will allow, or indefinitely long.

The read process relies on the operation of the yTron. To understand this operation, consider a loop that at one instance is in the "0" state with no persistent current $I_p = 0$ A, and at another instance is in the "1" state with $I_p = 15\,\mu$A. Under

these conditions, the operation of the yTron – and hence the read process – is summarized by figure 2-5. With the yTron curve shown, we would set the read current to $I_r = 1.09 I_{b,s}(0)$. Thus, when we are in the "0" state, we find ourselves at a point on the yTron curve below the bias point during the read. This means that the yTron will switch, and we will witness a voltage $V_r > 0$. However, if we were in the "1" state, then during a read we would find ourselves above the bias point on the yTron curve, as a result the yTron would remain superconducting, and we would have $V_r = 0$. Thus, we can determine the state of the loop by applying a bias to the yTron and reading the voltage.

In an ideal device, any loop current $I_p < I_{p,th}$ would result in a voltage at the read port. However, in practice the yTron has some switching distribution. This distribution can be considered as a variation in the threshold loop current $I_{p,th}$. Thus, for reliable operation, we need the difference between the zero state and one state persistent currents to be as large as possible. This reduction in operating margins is further compounded by the fact that the loop current will also likely have some distribution. This effects compound to further narrows the device operation margins, and mean that we require a large separation between the "0" state loop current and the "1" state loop current in order to achieve reliable operation.

2.2.3 Cell design limitations and trade-offs

In order to obtain the most sensitive readout possible, the right-hand loop inductance L_R must be substantially larger than the left-hand loop inductance L_L. This requirement is a consequence of two competing requirements. First, the yTron will pass, under the worst-case condition, the full write bias current I_w. Thus, we must have the sense branch of the yTron must be able to carry this current without switching, that is $I_{y,c} > I_W$. The second requirement we have is that I_p must be as large as possible in order to get the best separation between states. As we saw in section 2.2.1, the loop current after an ideal write will be given by equation 2.1. Ideally we would like I_p to be as close to $I_{y,c}$ as possible, as this will mean that our "0" and "1" states will be far apart in terms of yTron switching current $I_{b,s}(I_p)$. In comparing

our to requirements we find that we need to minimize the term $L_R/(L_L + L_R)$, this requirement is of course achieved by choosing $L_R \gg L_L$. Thus, with this requirement satisfied, the greatest separation between the yTron read current thresholds is obtained without switching the yTron during a read.

It is possible to obtain higher sensitivity from the yTron than the sense branch switching current $I_{y,c}$ would otherwise allow. This possibility comes about as a result of the yTron being a symmetric device, and there is no fundamental difference between the sense and bias ports other than their intended operation. This means that the sense branch switching current $I_{y,c}$ is in fact a function of the bias applied to the yTron's bias port (the opposite of the indented operation of the yTron). To write a current higher than the zero-bias sense branch switching current would allow, we must apply a bias to the read port during the read operation. The direction of this bias must be the same as that applied to the write port. The timing of such a bias can be the same as the write port bias. While this approach will allow a more sensitive readout, it makes the control circuitry that governs the write and read operations more complex. Additionally, a multiplexer would be required to switch the read port between its write and read modes. Thus, there exists a trade-off between complexity, cell size, and readout margins. The larger we make the cell the greater we can make L_R/L_L, which will give improved margins. The more complex the write circuitry, the less we need to rely on the inductor ratio to gain improved margins. After weighing these trade-offs, we decided that increasing write circuitry complexity was undesirable, and increasing the cell size in order to achieve improved margins was acceptable.

The calculations performed up to this point ignore flux quantization. We must determine what limitations the inclusion of these effects will have on our design. The memory loop, when in its superconducting state, must hold a quantized flux. This requirement implies that the persistent current I_p can only assume discrete values. That is, the current must be given by

$$I_p = n\frac{\Phi_0}{L_L + L_R}, \tag{2.2}$$

where $n \in \mathbb{Z}$. This requirements puts limitations on the loop inductances, and our hTron switching current. In terms of the loop inductance limitations, we need for the loop to be able to hold at least one fluxon. As it turns out, for NbN, being able to hold at least one fluxon in the loop is relatively easy to accomplish. Consider a hTron that, with no gate current applied, has a switching current of $I_{\mathrm{h,s}}(0) = 50\,\mu\mathrm{A}$. With the switching current set, we need $I_p < I_{\mathrm{h,s}}(0)$ to ensure that our persistent current does not trigger the hTron. Simultaneously, we need $I_{\mathrm{p}} > \Phi_0/\left(L_{\mathrm{L}} + L_{\mathrm{R}}\right)$, in order to have at least one fluxon in the loop. Thus, we have that we need $L_{\mathrm{L}} + L_{\mathrm{R}} > \Phi_0/I_{\mathrm{h,s}}(0) = 41\,\mathrm{pH}$. Such an inductance is easily obtained since the sheet inductance for our NbN films is $\sim 60\,\mathrm{pH}/\square$. Thus, a loop that consists of, say three or more squares, then we can – for the most part – ignore flux quantization.

The design of the hTron is largely limited by lithography and the desired channel characteristics. The width of the channel is set by the desired switching current $I_{\mathrm{h,s}}(0)$. The gate width is primarily set by the desired gate switching current. The last parameter we can vary is the spacing between the heater and channel. Ideally, we would like to have the gate as close to the channel as possible; however, there is a limit to how close these features can be. The closer the gate is to the channel, the shorter the gate to source trigger delay, and the lower the required gate power dissipation. The drawback is that, the closer the two features are to each other, the more difficult and unreliable the fabrication will be. We have found provided the separation between the gate and the channel is around 100 nm, then the device is very reliable.

As we have seen, the memory cell design is governed by ratios of inductances and switching currents. Thus, it is expected that the memory can be scaled to almost any size, provided all the ratios are maintained. In practice, we are limited by two major factors, namely lithography and noise. We are limited by how small we can make the loop by our lithographic techniques, and more importantly the variations in such. In terms of noise, we are limited by our experimental setup. Our devices have very low switching currents, and present a very low impedance, and so any noise – in particular voltage noise – will translate into a high current noise. These effects put

practical limits on the size of that cell that we can test and produce.

2.2.4 Cell layout

With a schematic level cell design ready, the next step is to turn this idea into a layout that can be written to a chip. The first cell designs tested here are the exact translation of the schematic shown in figure 2-3 to a layout, as shown in figure 2-6, with only minor modifications to address geometric effects. This layout was used for all initial device testing. The cell is constructed from a single layer of NbN sputtered onto a thermally oxidized silicon wafer. The NbN film is then patterned by means of electron-beam lithography and subsequent reactive-ion etching. A positive tone resist was used, and so only the outlines that isolate regions of NbN are drawn in the layout (the black lines shown in figure 2-6).

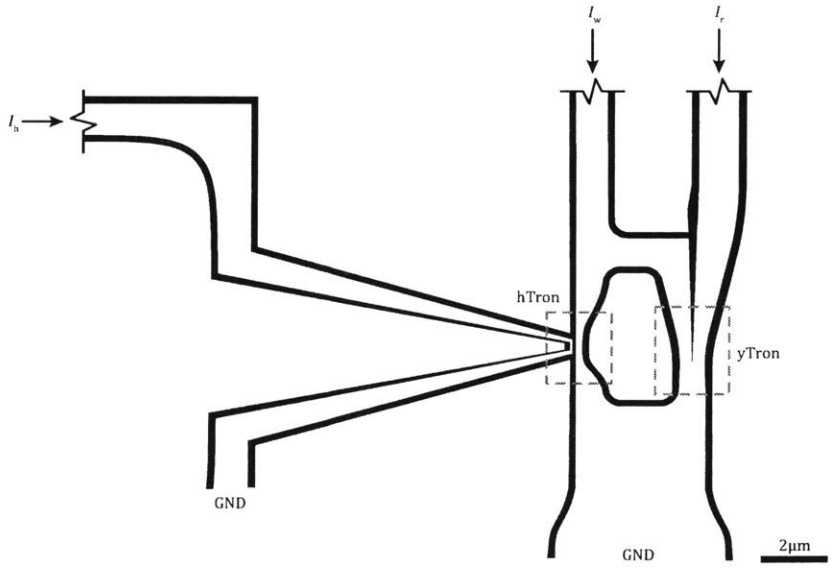

Figure 2-6: Layout of a compact NDRO cell used in the first set of measurements. The design shown is the exact layout of the device tested in section 2.4. The black area indicates where NbN has been etched away (leaving the bare substrate below), and the white area is where NbN remains. The drawing is shown this way because a positive tone resist was used, so the black area is where the resist was exposed. The jog lines indicate leads that extent to connection pads. Note, the hTron and yTron have been highlighted by a dashed box around each device.

The cell has been designed to avoid current crowding on all corners, with the exception for the yTron and hTron heater. To accomplish this, all inside corners are rounded to avoid currents making sharp turns. The heater was not designed to avoid current crowding since we intend for the gate to switch when current is applied, and we don't anticipate sourcing a high enough current to warrant avoiding current crowding in the leads.

The size of the cell is primarily limited by the lithography techniques. The finest feature in this design is the hTron gate. It needs to be as narrow as possible, so as to provide sufficient heating to the channel but not waste excessive power or require wide high-current-carrying leads. The gate width was chosen to be 60 nm, while this feature size is not trivial to fabricate, it is expected to yield well. Similarly, the spacing between the heater and the channel was selected to be 100 nm. Ideally, the channel would be as wide as possible from the point of view of increasing margins, but as narrow as possible from the point of view of reducing the required power dissipation in the gate. A compromise between the two factors was struck and the channel width was set to 190 nm. All outlines to isolate the devices from the rest of the NbN film were set to 160 nm.

Next, the inductance ratio was selected. The right-hand and left-hand sides of the loop each were designed to contain around ten squares which, assuming a calculated inductance per square of 60 pH/□, corresponds to $L_{\mathrm{L}} \approx L_{\mathrm{R}} \approx 0.6\,\mathrm{nH}$. From equation 2.1, this choice means that we can expect a write to have a write efficiency of $I_{\mathrm{p}}/I_{\mathrm{W}} \approx L_{\mathrm{R}}/(L_{\mathrm{L}}+L_{\mathrm{R}}) = 0.5$. While it would be ideal to have larger write efficiency, increasing L_{R} or decreasing L_{L} will result in a larger cell size.

Now, with the hTron channel width set and the inductance ratio set, the minimum size of the cell is fixed. This size is fixed because we know the write efficiency, and so the peak current through the loop. The value of this peak current fixes the minimum yTron width, and the loop trace widths. The smallest possible yTron arm width that we expected to be able to withstand the complete write bias was used – with some small margin added. All of the tapers between the hTron and yTron constrictions were added such that they adapt these narrow features to a width roughly three times that

needed to carry the expected current, while also ensuring excessive current crowding will not occur. The separation between the hTron and yTron was then chosen to give the desired inductance ratio. All ports were then connected, and we have the complete layout of the cell.

The cell is then placed into the inner region of a large pad stricture. This structure contains pads that are large enough to wire-bond to easily ($240\,\mu\text{m} \times 140\,\mu\text{m}$). The cell was electrically connected to the pads through $1\,\mu\text{m}$ traces, which featured rounded corners to avoid excessive current crowding. The pad design contains nine pads in total, which are arranged in a square. Thus, four devices could be placed within one set of pads. However, in the final design, it was decided to place only two cells within one set of pads, with the remaining pads being used for an array of test structures, in particular isolated yTrons, hTrons, and nanowires.

2.3 Cell simulation

With a layout complete, the design was verified by SPICE simulations before time and money was invested in a fabrication run. In order to simulate the memory in SPICE, a model for a nanowire was required. Two models, one for the hTron and one for the yTron, were used. Both of these models were originally based on the model presented in [31]. The original model on which our two models were based was intended for superconducting nanowire single-photon detectors (SNSPDs). As such, the original model is one dimensional, and does not take into account the non-constant width of our constrictions, or any kind of geometric effects. Thus, with these models we cannot fully replicate the actual memory; however, we can gain an understanding of the viability of a particular design.

LTspice was used for the simulations in this section. This software, like most versions of SPICE, cannot solve for many circuits that do not feature a resistance. Thus, it is not possible to directly simulate the memory cell as drawn in figure 2-3, instead a small resistance must be inserted into the memory loop as shown in figure 2-7. The inductor L_{R} was split into two inductors, L_{T} and L_{B}. This more accurately

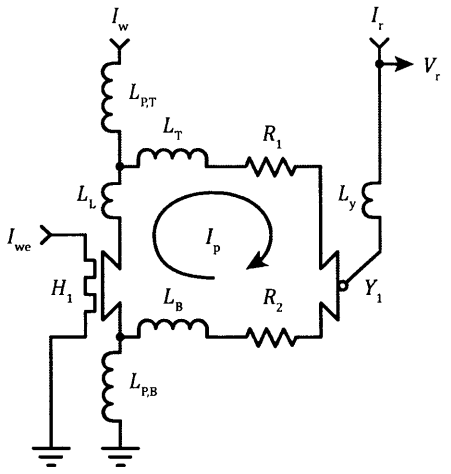

Figure 2-7: Schematic used in the LTspice simulation of the NDRO cell. Note the inclusion of the two resistors R_1 and R_2 which are not physical, but are included so that LTspice will reliably converge. This approximation is valid for small values of these resistors and for short time scales.

emulates the layout of the actual cell, as shown in figure 2-6. This modification to the simulation circuit is important as the current bias applied to the yTron will inductively split between both L_{T} and L_{B}. Finally, there are two more additional inductors, namely $L_{\mathrm{P,T}}$ and $L_{\mathrm{P,B}}$. These inductors model the leads to the device.

In order for the simulation to accurately model the device shown in figure 2-6, estimates of the corresponding component parameters were required. The resistors were set to $R_1 = R_2 = 1\,\mu\Omega$, so as to have as little effect on the simulation as possible. From our layout we have that $L_{\mathrm{R}} \approx 0.6\,\mathrm{nH}$. In the simulation L_{R} was split as $L_{\mathrm{T}} = 0.54\,\mathrm{nH}$ and $L_{\mathrm{B}} = 0.06\,\mathrm{nH}$, since there is roughly only one square below the yTron and ground in the layout. The left branch inductance was set to the value estimated from the layout, $L_{\mathrm{L}} = 0.6\,\mathrm{nH}$. The port inductances were set to $L_{\mathrm{P,T}} = L_{\mathrm{P,B}} = 1\,\mathrm{nH}$. The hTron and yTron parameter were set to match the layout. Finally, the current sources used to provide the biases to the various ports were modeled as an ideal current source with a Norton equivalent resistance of $R_{\mathrm{N}} = 50\,\Omega$. This equivalent resistance effectively emulates the signal generators that will be used in the experiments.

A number of simulations were performed, in particular experiments examining

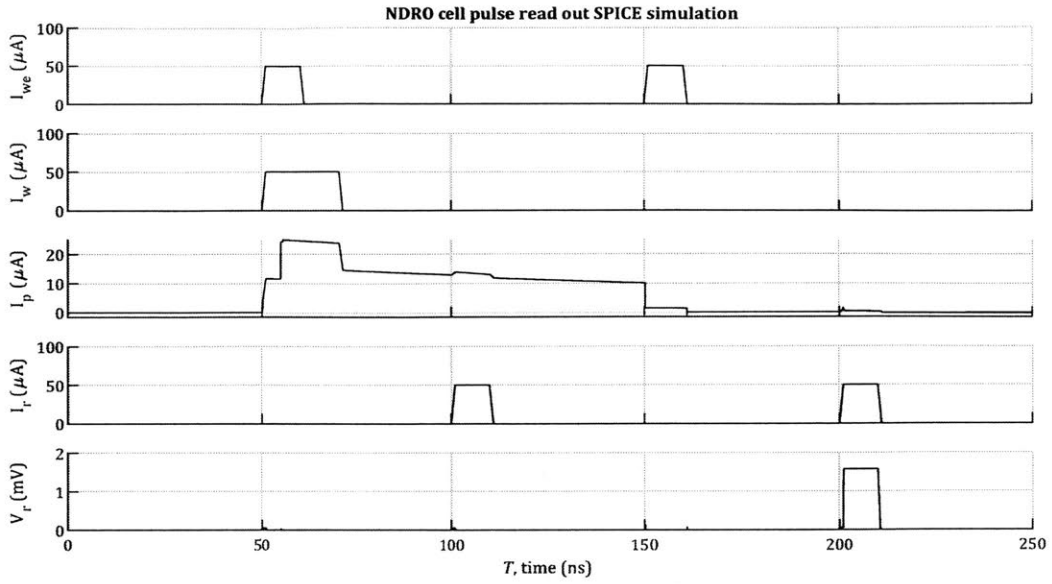

Figure 2-8: Results of a simple LTspice simulation of a single NDRO cell operated in the pulse readout modus. In this simulation, the cell is set (placed in the "1" state) at 50 ns. At this time, the loop current I_p can be seen to increase – indicating that the write was successful. The cell is then read at 100 ns. During the read, the yTron does not switch, as expected when the cell is set. The cell cleared (placed in the "0" state) at 150 ns. Again, the loop current can be seen to respond accordingly by reducing to zero. Finally, at 200 ns the memory is read again, but this time the yTron switches. The switching of the yTron indicates that the read was successful, and that, at least in this simulation, the memory is working as expected. Note that the loop current can be seen here to be decaying. This decay is an artifact of the simulation, specifically the resistors R_1, and R_2. In reality there is no decay in the loop current.

different write and read configurations were conducted. The outcome of these simulations influenced the methods used later in sections 2.4 and 2.5. The two methods of reading out the memory that were explored in the most detail, and were used in experiments are the *pulse readout* and the *ramp readout.* The pulse readout scheme is the method that would likely be used in a final application. This scheme utilizes a pulse whose amplitude is tuned such that the yTron only switches when the persistent current is low. The ramp readout, on the other hand, performs what is essentially the positive portion of an IV curve sweep to determine the switching current of the yTron. The latter approach is the most useful in initial tests as it removes one parameter that would otherwise need to be tuned. The ramp readout scheme was used

extensively in section 2.4. In later measurements, when the cell error rate was very low, and measurements speed critical, the pulsed readout was used, in particular in section 2.5.

Writing to the loop was performed in accordance to the timing diagram shown in figure 2-4. The write time was set, somewhat arbitrarily to $T_{\mathrm{w}} = 10\,\mathrm{ns}$, and the hold time similarly arbitrarily set to $T_{\mathrm{h}} = 10\,\mathrm{ns}$. In simulations, the hold time was varied between zero and many nanoseconds, and the cell performing identically at all values; however, any negative hold time would result in erroneous writes. For the pulse readout simulations, the read was performed exactly as shown in figure 2-4. The read time was set to match the write time, $T_{\mathrm{r}} = 10\,\mathrm{ns}$. The results of the simulation with these parameters is shown in figure 2-8.

It can be seen in figure 2-8 that the write process operates exactly as intended. The write bias is applied to the cell, and the assertion of the write enable causes the diversion of the current to the right-hand side of the loop, leading to an apparent high I_{p}. Once the write enable is deasserted and the write bias removed, the persistent current reduces to $I_{\mathrm{p}} = 18\,\mu\mathrm{A}$. From our simplistic analysis of the memory, we expected the loop current to be $I_{\mathrm{p}} = I_{\mathrm{W}} L_{\mathrm{R}}/(L_{\mathrm{L}} + L_{\mathrm{R}}) = 25\,\mu\mathrm{A}$. The discrepancy between this and the simulated value is likely primarily due to the apparent decay in the persistent current that is present in the simulation. This effect is a consequence of the fact that resistors were incorporated into the loop to allow LTspice to converge. So, it seems that our calculation and simulation broadly agree. The write to the "0" state can be seen to perform exactly as expected with the loop current dropping to close to zero as soon as the write enable signal is asserted in the absence of a write bias. Like the write process, the read process is also operating exactly as designed. In the case when the persistent current is high, the application of the read bias pulse resulted in no output voltage V_{r}. When the persistent current was low, the application of the read bias pulse caused the yTron to switch and a voltage to be generated at the V_{r} node.

The results of the ramp readout simulation shown in figure 2-9. It can be seen that the write operations were identical to those used in the pulse readout simulation.

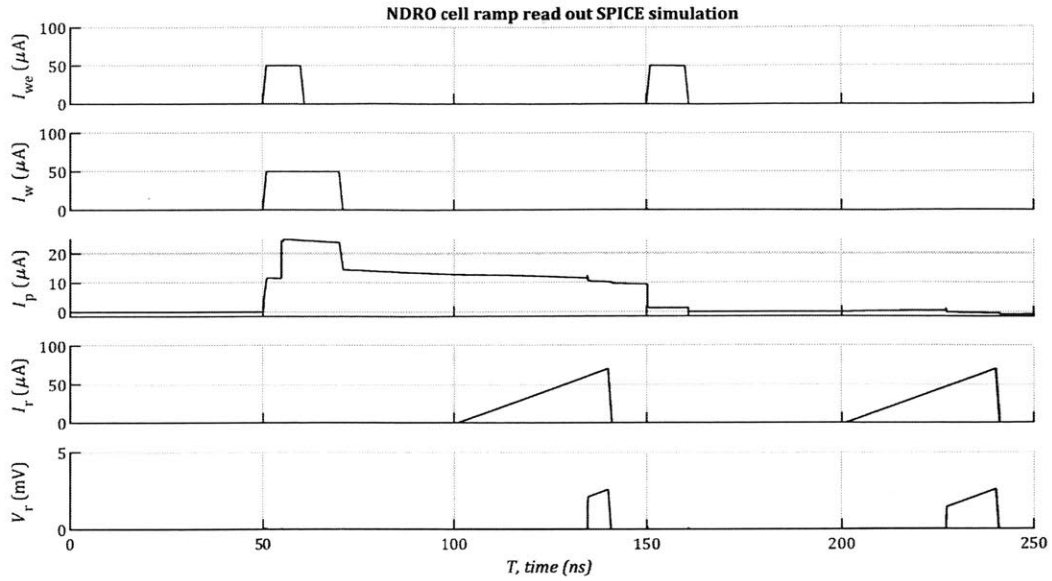

Figure 2-9: Results of a simple LTspice simulation of a single NDRO cell operated in the ramp readout modus. The write portions of this simulation are identical to those used in figure 2-8. The ramped readout current can be seen to show the switching current of the read port being modulated by the persistent current I_p. When the cell is in the "1" state with a high persistent current, the read port switching current can be seen to be higher than when the cell is in the "0" state with a zero persistent current.

The ramp readout starts from $I_r = 0\,\mu$A and linearly increase to $I_r = 75\,\mu$A. It can be seen that when the memory is in the "1" state, that is the persistent current is high, the yTron switches at a current around $60\,\mu$A. In contrast, when the cell is in the "0" state, that is the persistent current is low, the yTron switches at only $47\,\mu$A. This means that with the pulse readout regime, we could use a pulse height between $47\,\mu\text{A} < I_R < 60\,\mu\text{A}$. That is, a readout current $I_R = 53.5 \pm 6.5\,\mu$A would yield a successful readout. Such a readout regime corresponds to a 12% theoretical operating margin on the readout bias. This simulation has shown that the memory operates as expected, that the ramp readout process is viable for testing purposes, and gives us an estimate of the readout margins. Thus, we now have the confidence to proceed with device fabrication and testing.

2.4 Basic NDRO cell measurements

With the simulation results indicating that the memory design was sound, devices were fabricated and tested. As the design was modified and improved, the measurement procedures evolved. While the specific procedures changed from measurement-to-measurement and from device-to-device, the same general procedure was followed each time. These general measurement procedures – covered in section 2.4.1 – verify the basic properties of the device. If these initial tests indicate an issue with the device, then the lengthy device-specific measurements may be abandoned in favor of testing a different device.

The memory-specific measurement procedures change greatly based on the performance of the memory. The primary metric we are interested in when measuring a memory is the bit error rate (BER). A test to determine a BER is referred to as a bit error rate test (BERT). The BER value tells us what the probability of an error occurring in a specific reading is. There are two main means by which an error rate can be estimated. The first method is to directly measure the number of errors observed in a given number of write/read cycles. This method works well when the error rate is high, as we need around, say 1%, of the measurements to be erroneous in order to have confidence in the resultant error rate. The second method is to use a limited number of measurements, and then use the statistics of the results to extrapolate what the error rate could be. This means that we do not need to observe a single error in order to estimate the error rate. Initially, the memory had a relatively poor error rate, at around 10%. Measuring such memories requires a relatively simple setup, which is covered in section 2.4.2. With what was learned from this setup, a more sophisticated setup, that would perform a more accurate BERT using a pseudo-random bit sequence, was used. This updated setup, and the results generated using that setup, are covered in section 2.5.

2.4.1 General immersion measurement procedure

The superconductor used for the devices explored here is NbN. With thin-film NbN typically having $T_c \approx 10\,\mathrm{K}$, we need to cool the sample below around 6 K, or less, to perform measurements on the sample. There are two main methods for cooling a superconducting sample in order to perform measurements, namely the use of cryogen-free cryostats, and immersion of the sample in liquid helium (LHe). Cryogen-free systems are generally preferred as the sample is never exposed to LHe, rather it is held in a vacuum and mounted on a surface at the desired temperature. The major drawback of these systems is the cool-down and warm-up times – which can be around 24 hours. This wait time means that if a device is cooled down and a problem occurs with the device which requires the wire-bonds to be modified, or similar, then two days are wasted while the cryostat warms up and cools down again. In such situations LHe immersion measurements – also called dip measurements – become advantageous. A second advantage of LHe immersion measurements is that the temperature is stable at 4.2 K (at standard atmospheric pressure), whereas cryostat temperatures can vary. However, as will be covered in section 4.5.3, the boiling of LHe may cause temperature variations in the surface of the chip.

As immersion measurements are fast and the most flexible, they were performed for the majority of the tests covered here. The general procedure for performing these measurements involves the following steps:

1. Prepare sample – glue to PCB and wire-bond connections;
2. Measure room temperature resistances;
3. Attach device to cables and wrap with aluminum foil.
4. Verify resistances have not changed;
5. Vent dewar;
6. Measure helium level, and verify it is high enough for device coverage;
7. Lower device into dewar;
8. Seal top of dewar with aluminum foil;
9. Measure device resistances to verify they are as expected;

10. Measure IV curves of devices;

11. Proceed with device-specific measurements;

12. When the measurements are complete, remove the foil and carefully lift device from dewar;

13. Remove any ice from within the dewar neck;

14. Seal dewar and verify the valves are in the correct configuration.

The sample is prepared by gluing the chip, which is typically 10 mm × 10 mm, onto a small PCB using an acetone soluble varnish. This allows the sample to be removed from the PCB and the PCB reused at a later date. The PCB is custom designed for this task, and is small enough that it fits through the dewar's neck. The PCB features a number of SMP connectors. The PCB contains pads that are used for wire-bonding the device to, these pads are gold-plated through an electroless-nickel immersion-gold (ENIG) process. This plating is required as solder cannot be bonded to with the wire-bonding machine available, and bare copper will corrode. After the varnish securing the sample to the PCB cures, the device is bonded to the PCB pads using a wedge bonding machine with aluminum wire.

After bonding, the device resistances are measured with a multimeter. If the resistances are as expected then the device is covered with a 1" plastic cap which is taped to the PCB to prevent the bonds from being damaged. The PCB is then attached to the cables that will be used during the experiment, and wrapped with aluminum foil. The foil has two main purposes; it aids in removing the device after the measurement, and it prevents loss of the sample. The PCB can become caught on the edge of the neck of the dewar when the device is lifted out of the dewar. The PCB getting caught on the dewar would not be an issue if it were not for the fact that the PCB is only attached to the cabled, and those cables are only retained by a detent. Thus, the foil is added and shaped such that the edges of the PCB are less likely to get caught during removal from the dewar. On occasion, the sample becomes detached from the PCB, typically it is caught by the plastic cap, but as added security, the foil prevents the chip from falling into the dewar – from which it is nearly impossible to recover. Once the foil has been added, the resistances are

again measured to ensure that the connectors are correctly mated.

Next, the dewar is vented, the LHe level measured, the device lowered into the dewar, and the top of the dewar covered in foil. The neck of the dewar is covered in foil to restrict the flow of air into the dewar. This greatly reduces the contamination of the dewar. From experience, if no foil is used then within 30 minutes some ice will have built up within the neck of the dewar; however, with the use of foil, then after 8 hours no ice will be seen within the neck.

At this point the resistances are again measured. This time some devices will be expected to be superconducting. There is an issue with measuring the resistance with a standard multimeter. Most meters will, when auto-ranging is enabled, apply a progressively higher current to the device until the meter reads a resistance with the current range. Since most of our devices have switching currents around $100\,\mu$A, this will typically result in the device switching and the meter reading a high resistance. To prevent the meter switching the device, the meter is manually set to a range such that the current applied to the device is low enough that the device will not switch. The downside of this approach is that the required range is typically $600\,\text{k}\Omega$, and so at this range, the resistance will typically show as $\pm 0.1\,\text{k}\Omega$. Due to the low resolution at this range, a meaningful measurement cannot be performed. Additionally, in most measurements a voltage of $-420\,\mu$V can be witnessed between the signal and ground on each cable attached to a superconductor. This voltage is thought to be caused by thermoelectric effects since it can be seen with a cable terminated into a SMP connector shorted with tin-lead solder. This offset voltage often leads to a superconducting device showing an apparent negative resistance when measured with the multimeter.

The next measurements that all devices go through is the generation of current-voltage relation curves. These curves are measured by applying a triangle waveform with a frequency of 100 Hz to the device through a $10\,\text{k}\Omega$ resistor. The voltage either side of the resistor is monitored with an oscilloscope. This allows for the generation of a rough IV curve. From examination of the IV curve, a number of issues can be diagnosed. Most notable and common are highly suppressed devices (very low switching currents) and superconducting shorts (very high switching currents). Once

the IV curves have been generated and examined, the device-specific measurements are performed.

After all measurements are complete, the foil is removed from the top of the dewar and the device removed from the dewar. The device must be very carefully lifted from the dewar so as not the pull the PCB off of the connectors which would result in the device becoming lost in the dewar. Once the device is out of the dewar, a plastic tube slightly smaller than the neck of the dewar is inserted into the dewar to remove any ice buildup. Finally, the dewar is sealed and the connected to the helium recovery system, and the valve configuration is checked.

2.4.2 Initial NDRO measurement setup

The first cell design, which is shown in figure 2-6, was tested. First the IV curves were plotted, as shown in section 2.4.3. After performing these basic tests, the setup was changed to test the operation of the memory cell. The experimental setup used for these tests is presented in section 2.4.4. The results generated using this setup are then presented and analyzed in section 2.4.5.

2.4.3 IV Curves

The current-voltage relation for each of the three ports of the cell was measured. The results are as shown in figure 2-10. These experiments were performed in the usual manner, with a $10\,\mathrm{k}\Omega$ bias resistor and a $50\,\Omega$ termination on any unused ports. It can be seen that the devices are operating as expected. The heater is switching at a relatively low current, and the channel and yTron switch at around an order of magnitude higher – due to their widths being roughly ten times wider.

There are a number of features present in the IV curves that can tell us detailed information about the device and how it switched. Consider the IV curve of the hTron shown in figure 2-10. This plot roughly follows what we would expect from a nanowire (as shown in figure 1-6), with three main exceptions:

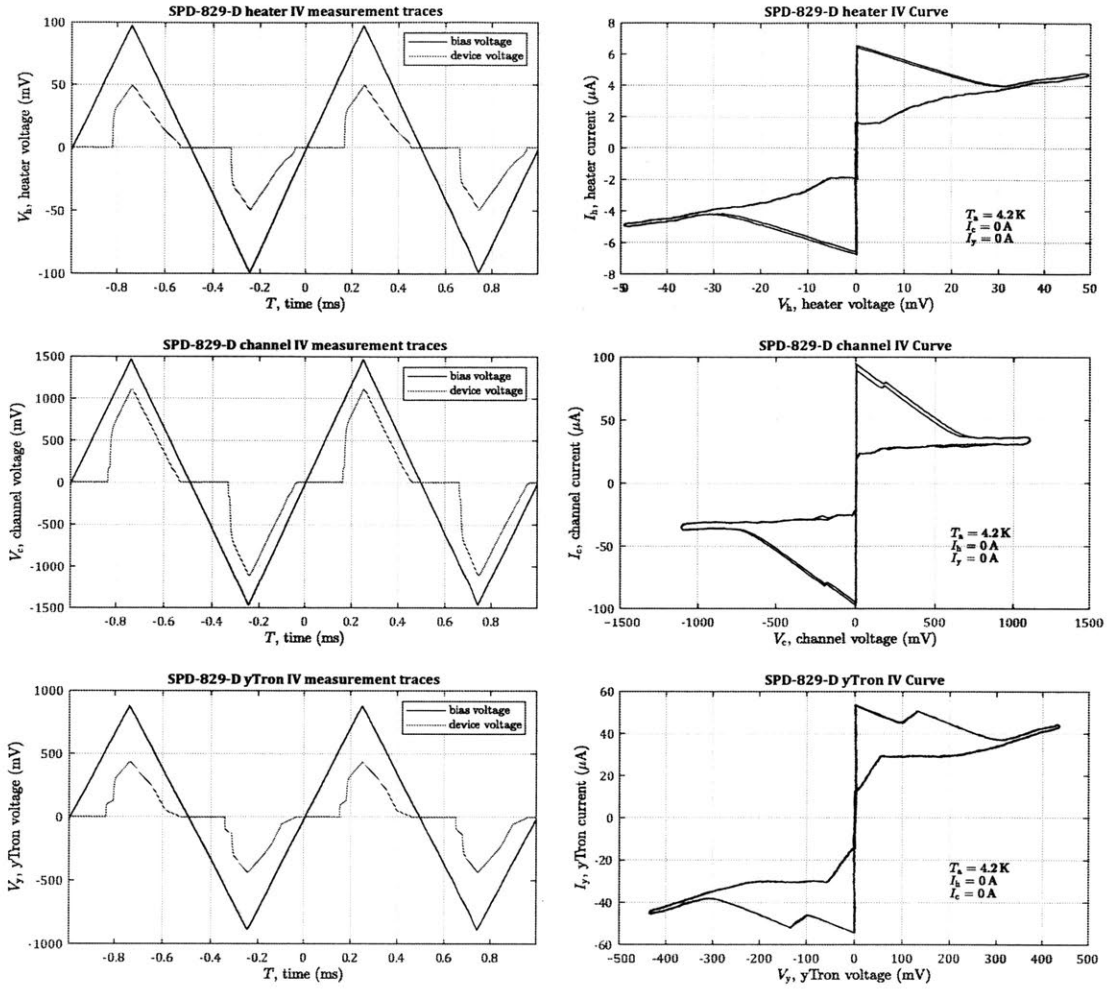

Figure 2-10: Results from an IV curve measurement made on the three ports of an NDRO cell. The figures on the left shows the time domain traces of the bias voltage and the device voltage. Note that a bias resistor of $10\,\mathrm{k}\Omega$ was used. The figures on the right show the IV curve of the device. Each of these experiments was conducted with all unused ports terminated into $50\,\Omega$. Note that in these plots the data has been decimated by a factor of 100.

Negative differential resistance We can see that the transition from the superconducting state to the normal state appears as a negative differential resistance. The existence of this region is primarily an artifact of the experimental setup. In the ideal case our current source would have an infinite Norton resistance; however, in our setup we have a resistance of $R_\mathrm{N} = R_\mathrm{b} = 10\,\mathrm{k}\Omega$. Thus, just before we transition to the normal state, we have $I_\mathrm{h} = I_\mathrm{h,sw}$, and the source voltage is $V_\mathrm{b} = R_\mathrm{N} I_\mathrm{h,sw}$, where

$I_{h,sw}$ is the heater's switching current. After the transition, we have approximately the same bias voltage V_b, but now we have a device resistance of R_h, and so the current is $I_h = V_b/(R_N + R_h) = R_N I_{h,sw}/(R_N + R_h)$. This second current is, of course, smaller than the original current $I_{h,sw}$. So we have transitioned from a zero voltage at a high current state, to a high voltage at a reduced current state – this manifests itself as a region of negative differential resistance. This effect can clearly be seen in all three IV curves, and will always occur – to varying extents – provided the source has some finite R_N.

Curves and steps in the normal region The appearance of curves and steps in the normal region is a consequence of two similar, but distinct effects. Both effects are the result of the geometry of the wire, in particular the variations in its width. The steps that occur after the device first switches are due to different regions with the device switching independently. In contrast, the curves are a result of the hot-spot growing and following the variations in width along the device. The reason that some of the variations in width result in a curve, and some result in a step is determined by if there exists a region of lower current density between two regions of high current density within the wire. If such a low-current density region exists, then we will see a step, and if no such region exists, then we will see a curve. To illustrate this distinction consider two devices shown in figure 2-11.

Variation in switching currents Each of the IV curves shown in figure 2-10 consist of two IV curves taken 1 ms after each other. In the channel IV curve, and to a lesser extent the heater IV curve, the switching current can be seen to be slightly different on the first and second sweep. This variation in the switching current is due to a number of effects, which cumulatively can be considered noise. One such effect, which is most likely dominant in these measurements, is the theorized formation of helium bubbles on the surface of the chip – this effect is covered in section 4.5.3. Additionally, the device is particularly sensitive to voltage noise when it is in the superconducting state. This sensitivity is a result of the device presenting a very low

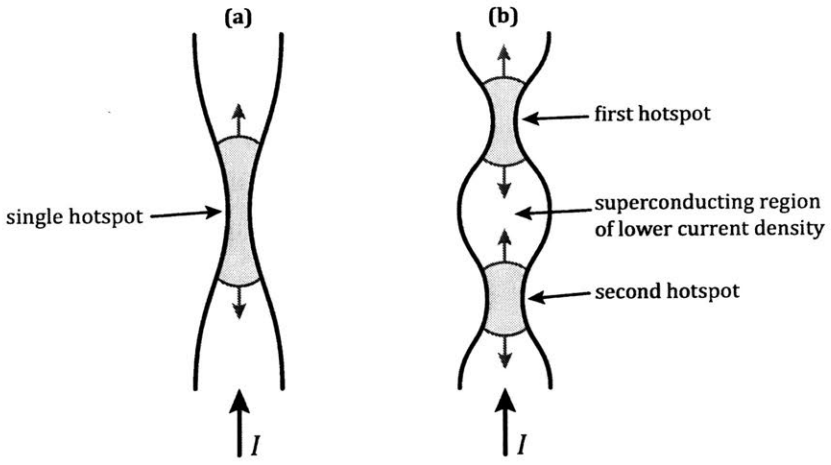

Figure 2-11: Depiction of the theorized hotspot growth in two different nanowire geometries. Both nanowires are exposed to a monotonically increasing current I. This current eventually leads to the switching of the single constriction in (a) and the first constriction in (b). As the current continues to increase, the hotspots will grow in the directions indicated by the gray arrows. In (a), the IV curve will not exhibit any steps other than the first switching event; however, due to the non-constant cross section of the nanowire, the IV curve will depict a curve. In (b), the IV curve will be similar to (a) at first, as the hot spot continues to grow. At some point however, the current density at the second construction will cause superconductivity to breakdown at this location, and a second hotspot will form. The formation of this second hotspot will result in a second step in the IV curve (separate from the first constriction switching). Thus, we can see how the non-constant width of the nanowire can lead to the IV curve exhibiting curves and steps.

impedance in this state, and so any pickup will result in a not-insignificant current, that will flow through the device. The sign and magnitude of this current can cause the switching current to appear higher or lower at the point that the device switches. This voltage-noise induced variation is likely a very minor effect in these measurements since the sweep was performed at a frequency of 1 kHz, and so only similarly low frequency noise could make this effect occur. High frequency noise will not cause this result, as it will simply manifest itself as an apparent suppression in the switching current.

2.4.4 Experimental setup

After the IV curves were measured and found to be consistent with what was expected, the device specific measurements were performed. There are a number of ways that these measurements could be performed. Ideally we would like to test the cell exactly as it would be used in practice. While such testing is possible, it would be extremely difficult to find the operating region. The device has three ports, each of which requires a specific bias level to operate, namely the write enable port current I_{WE}, the write current I_{W}, and the read current I_{E}. Other than the simulations covered in section 2.3, we did not have a real idea where to start in terms of timing of these signals or the levels required. Thus, it was desirable to reduce number of variables for initial measurements. The first method used to achieve a reduction in the parameter space we have to explore is the ramp readout.

The ramp readout is a means by which the switching current of the yTron can be determined. It is essentially the positive and monotonically increasing portion of the signal that was used to determine the IV curves in section 2.4.3. An example of the signals applied to the cell in these initial tests is shown in figure 2-12. The ramped current to the I_{r} can be seen at times 100 ns to 150 ns, and 250 ns to 300 ns. The switching current of the yTron can be determined by examining the read voltage V_{r}, and determining what current was required to cause the yTron to switch. This readout scheme means that we do not need to adjust the readout current level, we simply need to select a peak current that is sufficient to ensure the yTron switched under all conditions. In fact this approach gives us more information, as it allows us to determine the exact switching current rather than a binary switched/did not switch output.

With the read bias taken care of, we will now consider the two write signals. There is little that can be done to reduce the search-space for the remaining two parameters. Upper and lower limits were selected by choosing a set of operating rules for the cell. The write enable current was chosen to be positive – this decision is arbitrary since it is only a heater and the current direction does not make a difference. Originally, we

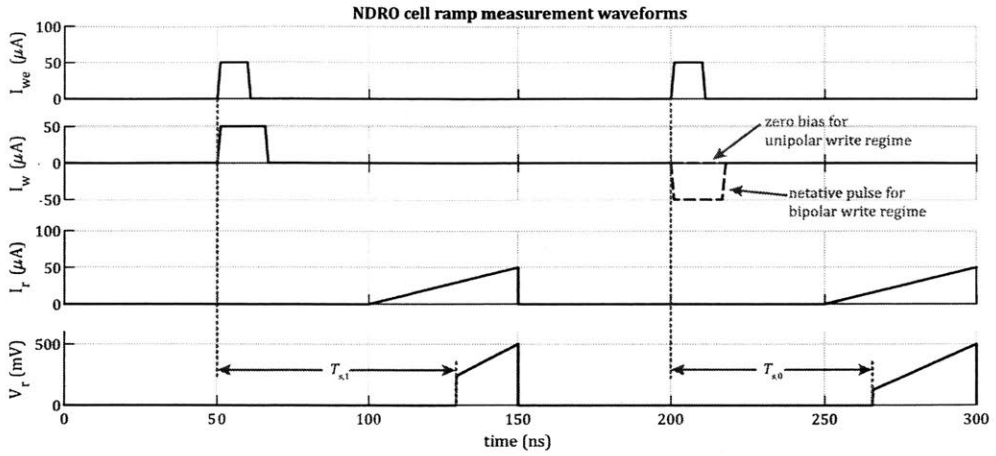

Figure 2-12: Schematic of the NDRO cell write and ramp readout scheme. This figure features two write/read cycles. The first operation, at 50 ns, is a set operation. Note that the channel bias is applied held after the deassertion of the write enable. The yTron switching current is them measured by applying a ramp to the read port. At some point the yTron switches and a voltage is seen at the read port. The time delay between the start of the write operation and the switching of the yTron is used to determine the current required to switch the yTron. At 200 ns the memory is cleared, for the unipolar wire scheme, the write bias is set to zero for this operation, and for the bipolar write scheme, the write bias is a negative pulse (the inverse of the set pulse shape). Again the memory is read out by applying a ramping current to the yTron. After a clear the switching current of the yTron should be lower than after a set, so we expect the skew times $T_{s,0} < T_{s,1}$. The time until the device switches is used as the oscilloscope has more time resolution than voltage resolution, and so a better estimate of the switching current can be obtained by this method.

chose to utilize only a positive write current, and so only positive write currents were considered. Originally, when we operated the memory, we utilized staggered pulses applied to the write. In this regime, to set the cell in to the one state, the write enable was asserted, followed by the write bias, then after some delay the write enable was deasserted, and finally the write bias was removed. Thus, a high persistent current would be imparted into the cell. In contrast, to write a zero we would first apply the write bias, followed by the write enable, then after some delay we remove the write bias, and finally deassert the write enable signal. Thus leaving a minimal persistent current circulating in the loop. This method was pursued since we were concerned that the hTron would not suppress the hTron switching current enough to trigger

the hTron channel without some assistance from an externally applied current. For the in-plane hTron design, this concern proved unfounded. While this method of writing to the loop was effective, it did require the signals to be timed correctly, and made writing to the memory a slower-than-necessary process. Thus, a system based purely on the levels of the write signals was used – of course still meeting out timing requirements, as outlined in section 2.2.1.

The writing regime that utilized only the levels of the write signals is relatively simple. An example of a set and reset of the NDRO cell is shown in figure 2-12. To set the cell to the "1" state, the write bias and write enable signals are applied roughly simultaneously – the exact timing is not important here. The write bias is positive, so as to impart a positive persistent current into the loop. To complete the write, the write enable signal is deasserted, and a short time later the write bias is removed. Thus, a positive persistent current will be written to the loop, setting it to the "1" state. For the clear to the "0" state, we have two options, we can either write a zero persistent current, or a negative persistent current. Originally, we chose to utilize only positive unipolar write current; however, with the initial cell design the margins necessitated a bipolar write scheme.

In order to apply the requisite signals to the cell, the experimental setup shown in figure 2-13 was used. Two arbitrary waveform generators (AWGs) and an oscilloscope were used. The AWGs used where one Keysight 33600A (AWG1), and one Keysight 33250A (AWG2). The oscilloscope used in these experiments was a LeCroy WaveRunner 620Zi. A low-noise amplifier (LNA) was used to amplify the signals from the yTron. The model of the LNA is LNA-2500 from RF-bay and provides a nominal gain of 25 dB. Additionally, a number of splitters and attenuator were used, all had a bandwidth from DC to in excess of 2.5 GHz.

It is desirable to perform a write and read within one oscilloscope acquisition window. This allows for the capture of all the data at once, without the need for some complex post-processing to determine if a specific acquisition contains a write or a read. Additionally, when setting up the signals for the first time, it is beneficial to be able to see all three signals on the oscilloscope. Thus, all signals, as well as the

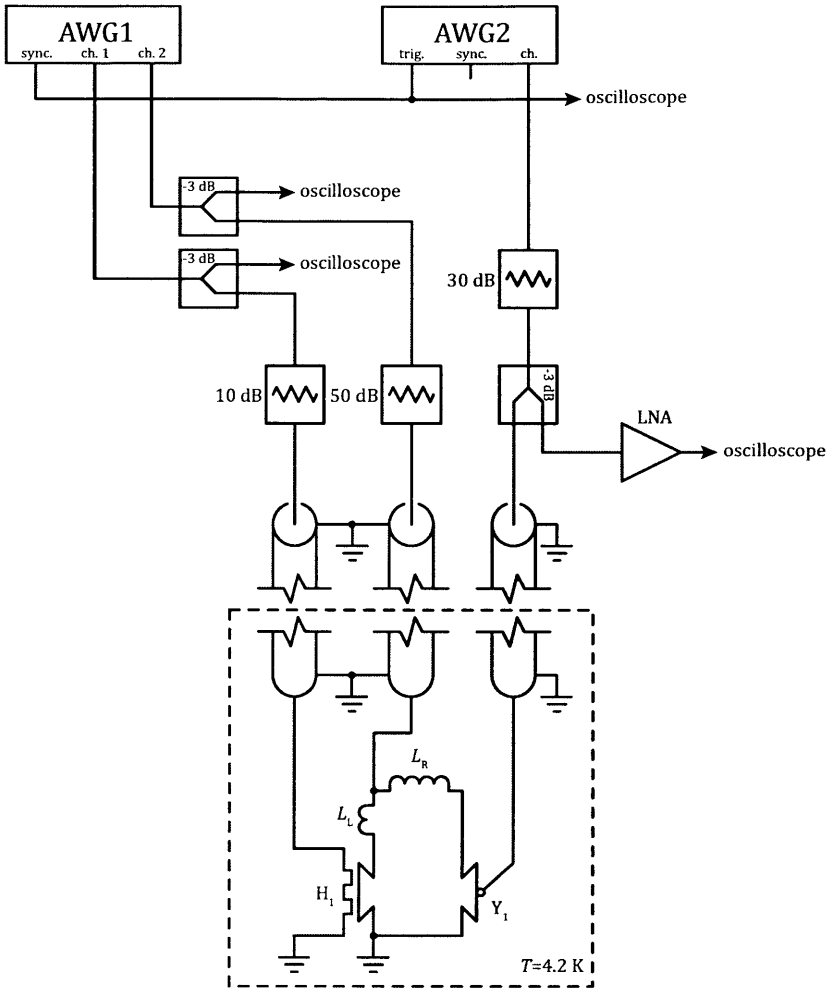

Figure 2-13: Experimental setup for basic NDRO measurements. AWG1 was used to control the write operation. AWG2 is triggered by AWG1, and after some delay, initiates the read operation. In order to minimize reflections, and enable high-speed operation, a system impedance of $Z_0 = 50\,\Omega$ was used. As the devices required relatively small currents to operate, attenuators were used to reduce signal amplitudes. To enable the oscilloscope to monitor the signals, splitters were added.

synchronization pulse from AWG1, were taken to the oscilloscope.

All of the room-temperature electronics were setup for a system impedance of $Z_0 = 50\,\Omega$. For this reason, all signals to the device were run through splitters to allow one branch to be monitored by the oscilloscope, while the other went to the device under test (DUT). The write enable and write bias signals were first generated by the AWG, then passed through splitters, with one branch heading to the oscilloscope,

and the other to the attenuators and finally the DUT. The reason the signal are split before attenuation is so that the signal will be large enough for the oscilloscope to be able to show easily. In contrast, for the yTron we first pass the output of the AWG through an attenuator, before it is sent to the splitter. One port of the splitter heads to the DUT, and the other to the LNA, which drives the oscilloscope. The reason for this configuration is that we are not interested in the bias applied to the yTron, rather we are only interested in when the yTron switches. It was found that the output voltage of the yTron was relatively low, and needed amplification in order for the oscilloscope to be able to easily detect the signal without quantization noise degrading the acquisition. To avoid reflections, all oscilloscope ports other than that connected to the synchronization signal, were set to a termination impedance of $50\,\Omega$. The synchronization signal's port was set to $1\,\mathrm{M}\Omega$.

We must extract the switching current of the yTron from the voltage pulse that the oscilloscope receives through the LNA. It would be difficult gain an accurate estimate of the current from the height of these pulses. Rather, a feature of the oscilloscope was used to extract the delay between the oscilloscope trigger signal (AWG1's synchronization signal) and the rising edge of the yTron output pulse. The time that the oscilloscope is measuring, when in this configuration, is demonstrated diagrammatically as $T_{\mathrm{s},1}$, and $T_{\mathrm{s},0}$ in figure 2-12. Downloading waveforms from the oscilloscope takes a few hundred milliseconds, and so it would be prohibitive to download one waveform for every write/read cycle. Instead, a second feature of the oscilloscope was used to track, and keep a record of the delay measurements. This feature on LeCroy oscilloscopes is referred to as a "trend". With the oscilloscope in analysis optimization mode, around 100 acquisitions and measurements can be performed per second. Thus, we can acquire the time-delay for many write/read cycles in a relatively short period.

Now that we have the raw time-delay between the synchronization pulse and yTron switch T_{s}, we need to convert this into a switching current I_{sw}. This can be achieved by using

$$I_{\mathrm{sw}} = I'_{\mathrm{r}}(T_{\mathrm{s}} - T_{\mathrm{h}}) \tag{2.3}$$

where T_h is the hold-off time between the synchronization and the start of the yTron current ramp, and I'_r is the slew rate of the read current ramp. This method provides a good estimate of the switching current, but it is not perfect. Errors in the hold off or ramp rate will lead to the estimated switching current becoming inaccurate. For our purposes however, these concerns as not an issue. We are not interested in the absolute switching current, rather we are only interested in an overlap between the switching current distribution for a memory in the "0" state and one in the "1" state. Since any errors in our assumed hold-off and slew rate will affect both reading a "0" and a "1" identically, we can be confident that our error rate estimates are accurate.

2.4.5 Memory operation results

With the memory-specific measurement setup ready, a number of measurements on the cell's operation were performed. The first measurements that were performed, used the basic write waveform shown in figure 2-12, and using a unipolar write bias. A clear pulse was sent after each measurement. This clear pulse was a large bias to the write port in the opposite polarity to the write signal. After tuning the write signal biases to obtain the best error rate, the operating point was as summarized in table 2.1.

Table 2.1: NDRO unipolar ramp readout operating point.

Parameter	Value
Write bias high level, I_W	$47\,\mu$A
Write bias pulse width, T_w	30 ns
Write bias edge times, $T_{w,rf}$	2.9 ns
Write enable high level, I_{WE}	$470\,\mu$A
Write enable pulse width, T_{WE}	25 ns
Write enable edge times, $T_{WE,rf}$	3.39 ns
Read ramp peak, I_R	$150\,\mu$A
Read ramp time, $T_{R,rf}$	150 ns

Using the values from table 2.1, two sets of trends were acquired. The first set was when the memory was written to the "1" state and read, then reset to some constant state with a large negative pulse. The second set was when the memory was written

to the "0" state and read, then reset to some constant state with a large positive pulse. Each trend contained 10,000 oscilloscope acquisitions, which results in 10,000 write "1" and read cycles, and another 10,000 write "0" and read cycles. The switching current of the read port during each of these acquisitions was plotted as a histogram, which were normalized to give an approximation to a probability density, are shown in figure 2-14.

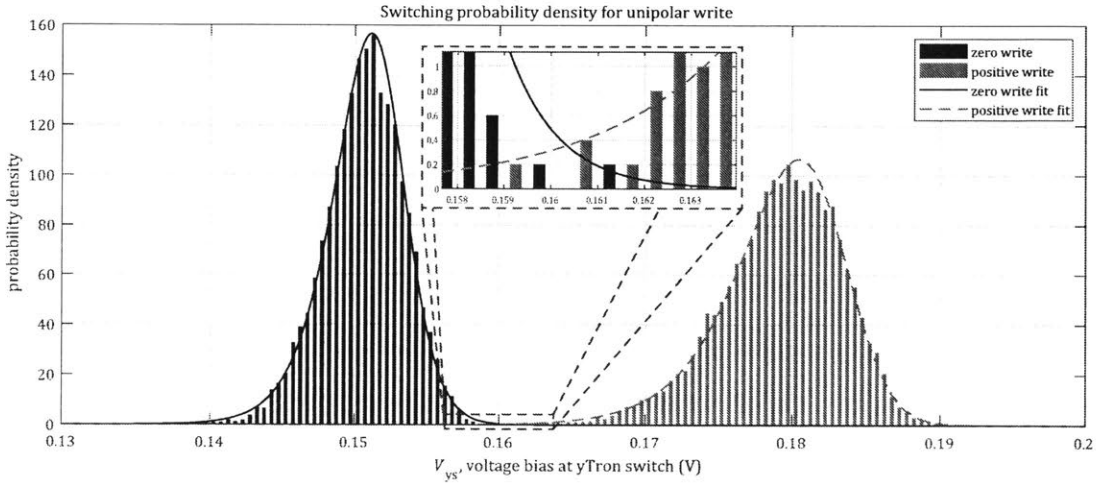

Figure 2-14: Switching probability density estimations, using a unipolar write pulse consisting of a positive write current for the set signal, and a zero write current for the reset signal. The horizontal axis represents the voltage bias that, through the bias network, resulted in a sufficient current to switch the yTron. The solid and dashed lines show a maximum likelihood fit of a Burr distribution to each histogram. The insert is a magnified section of the plot showing the overlap between the two distributions. It can be seen that, while the memory operates very well, there are a number of errors.

Two distributions were fitted to the read port switching current histograms, as shown in figure 2-14. These fits were Burr distributions, which were found to give the best goodness-of-fit. The fit parameters were determined through maximum likelihood estimation. Since the total number of errors observed in the data (provided a good threshold was selected) was very low, the observed error rate does not provide a good estimate of the actual error rate. For this reason the error rate was estimated using the fits.

To estimate the error rate from the fits, we will apply Bayesian hypothesis testing.

Let a set state be denoted W_1, a reset state be denoted W_0, a read that indicates the cell is in the "1" state be denoted R_1, and a read that indicated the cell is in the "0" state be denoted R_0. When we read the memory, we sample the switching current I_s, which will be our random variable. We will assume that the posterior probabilities $P(\mathrm{W}_0) = P(\mathrm{W}_1) = 0.5$, that is, we will assume that roughly half of writes are sets, and half are resets. In practice, this is very unlikely to be true, as values in a computer's memory tend to favor smaller numbers on average, and so depending on the coding scheme, we will tend to have more zero-bits or one-bits. Our objective is to take a reading from the memory, in the form of a switching current i_s, and determine if it should be denoted a zero or a one.

We want to choose the hypothesis that will minimize the probability of error $P_\mathrm{E} = P(\text{choose } \mathrm{W}_0|\mathrm{W}_1) + P(\text{choose } \mathrm{W}_1|\mathrm{W}_0)$. To achieve the minimum error rate, then if we are given a value i_s, we should choose the hypothesis that is the most likely, that is

$$P\left(W_1|I_\mathrm{s} = i_\mathrm{s}\right) \underset{\mathrm{W}_0}{\overset{\mathrm{W}_1}{\gtrless}} P\left(W_0|I_\mathrm{s} = i_\mathrm{s}\right). \tag{2.4}$$

Since we only have our fits for a positive and a negative write, $f_{I_\mathrm{s}}\left(i_\mathrm{s}|W_1\right)$ and $f_{I_\mathrm{s}}\left(i_\mathrm{s}|W_1\right)$, respectively, we need to use Bayes' rule to find $P\left(W_1|I_\mathrm{s} = i_\mathrm{s}\right)$ and $P\left(W_0|I_\mathrm{s} = i_\mathrm{s}\right)$. This is simply determined as

$$\begin{aligned} P\left(W_1|I_\mathrm{s} = i_\mathrm{s}\right) &= \frac{f_{I_\mathrm{s}}\left(i_\mathrm{s}|W_1\right) P(\mathrm{W}_1)}{f_{I_\mathrm{s}}(i_\mathrm{s})}, \\ P\left(W_0|I_\mathrm{s} = i_\mathrm{s}\right) &= \frac{f_{I_\mathrm{s}}\left(i_\mathrm{s}|W_0\right) P(\mathrm{W}_0)}{f_{I_\mathrm{s}}(i_\mathrm{s})}. \end{aligned} \tag{2.5}$$

Using this substitution, we find that our decision rule becomes

$$\frac{f_{I_\mathrm{s}}\left(i_\mathrm{s}|W_1\right)}{f_{I_\mathrm{s}}\left(i_\mathrm{s}|W_0\right)} \underset{\mathrm{W}_0}{\overset{\mathrm{W}_1}{\gtrless}} 1. \tag{2.6}$$

This is a simple maximum a posteriori (MAP) decision rule. As it turns out, our fits are well-behaved and have only one intersection. Thus, the above decision rule can

be termed as a threshold test on i_s, that is

$$i_s \underset{W_0}{\overset{W_1}{\gtrless}} I_{th}. \tag{2.7}$$

Our fits were performed with Burr distributions which have the distribution function

$$f(i_s; \alpha, c, k) = \frac{\frac{kc}{\alpha}\left(\frac{i_s}{\alpha}\right)^{c-1}}{\left(1+\left(\frac{i_s}{\alpha}\right)^{c}\right)^{k+1}}, \tag{2.8}$$

and as a result we found the threshold I_{th} by numerically solving $f(I_{th}; \alpha_1, c_1, k_1) = f(I_{th}; \alpha_0, c_0, k_0)$ [32]. After finding the threshold the write-one-read-zero and write-zero-read-one probabilities $P_{W1,R0} = P(\text{choose } W_0|W_1)$ and $P_{W0,R1} = P(\text{choose } W_1|W_0)$ were calculated using the Burr cumulative distribution function (CDF). Finally, the estimated error rate $P_{E,e}$ calculated. The results as shown in table 2.2.

With the promising results from the unipolar tests, we decided to explore a bipolar write in hopes that it would further separate the distributions, thereby allowing for lower error rates. The AWG configurations where modified to allow for this. The write "1" procedure is identical to that used in the unipolar write. The write "0" procedure was modified to allow for a negative write pulse, as shown in figure 2-12. The write "0" level was tuned to give the best separation possible while maintaining all the other setting the same as those used in the unipolar write. The timing for the write "0" was also chosen to be identical to that used when writing a "1", thus the only difference is the 'high level' $I_{W,0} = -I_W = -47\,\mu A$. The same measurements that were performed for the unipolar write, were again performed for the bipolar write. Due to the anticipated improved error rate for the bipolar write, more trials were acquired ($50,000$ acquisitions in place of the $20,000$ acquisitions used in the unipolar tests) to give a more accurate estimate of the error rate. The results were plotted in a similar manner, as shown in figure 2-15.

From the unipolar write results, we can see that in $N_T = 20,000$ trials we had a total of $N_{UP,E} = 3$ errors, provided we chose an ideal decision rule. Thus, for this

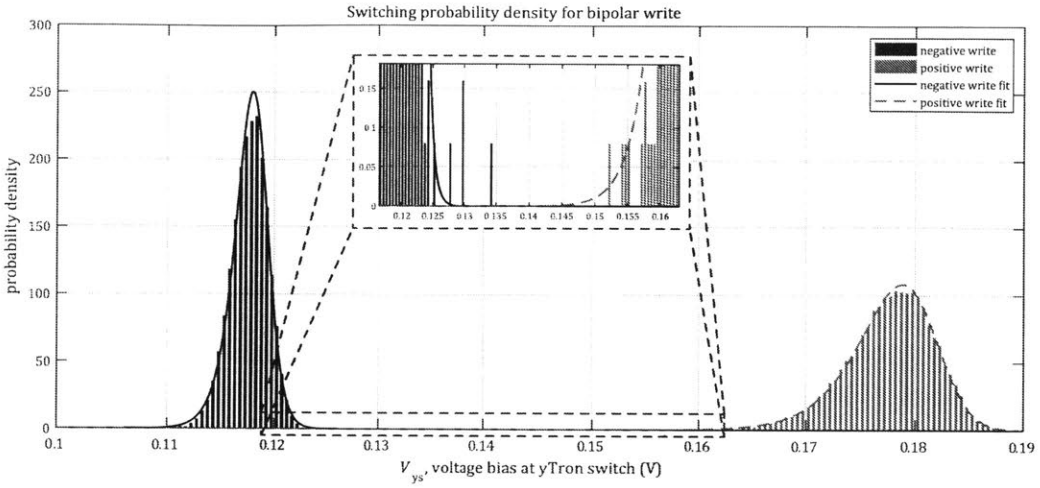

Figure 2-15: Switching probability density estimations, using a bipolar write pulse of a positive write current for the set signal, and a negative current pulse for the reset signal. The horizontal axis represents the voltage bias that, through the bias network, resulted in a sufficient current to switch the yTron. The solid and dashed lines show a maximum likelihood fit of a Burr distribution to each histogram. The insert is a magnified section of the plot showing the overlap between the two distributions. It can be seen that there were no errors observed, and that the overlap between the tails of the fits are very small, thus resulting in a very low fit-estimated error rate.

experiment an error rate of $P_{\mathrm{UP,E}} = 1.5 \times 10^{-4}$ was observed. On the other hand, for the bipolar write, and using the $N_{\mathrm{T}} = 50,000$, we observed no errors. Thus, the error rate in this modus could be estimated to be $P_{\mathrm{BP,E}} < 1/N_{\mathrm{T}} = 2 \times 10^{-5}$. Upon examination of these error rates, summarized in table 2.2, it can be seen that for the unipolar write, the directly observed error rates were somewhat lower than the trend-estimated error rates, while still being on the same order of magnitude. This discrepancy could be due to the low number of errors observed, and so high possibility of the estimate being incorrect, or could be due to the fit overestimating the number of errors. Whichever happens to be the case, the results are sufficiently similar that we can have confidence in the accuracy of the extrapolated results. Somewhat similarly, for the bipolar write we have that no errors were observed, so we estimated the error rate was $P_{\mathrm{E,o}} < 2.000 \times 10^{-5}$. The fit-estimated error rates back up this claim, as shown in table 2.2.

In order to gain an understanding of the operating margins for the two operating

Table 2.2: NDRO cell ramp readout error rate estimates.

	Unipolar write	Bipolar write
Directly observed error rate, $P_{\mathrm{E,o}}$	1.500×10^{-4}	$0\ (< 2.000 \times 10^{-5})$
Extrapolated probability of a false one, $P_{\mathrm{W1,R0}}$	4.670×10^{-4}	6.429×10^{-9}
Extrapolated probability of a false zero, $P_{\mathrm{W0,R1}}$	1.812×10^{-4}	2.207×10^{-9}
Extrapolated error rate from fit, $P_{\mathrm{E,e}}$	6.482×10^{-4}	8.636×10^{-9}

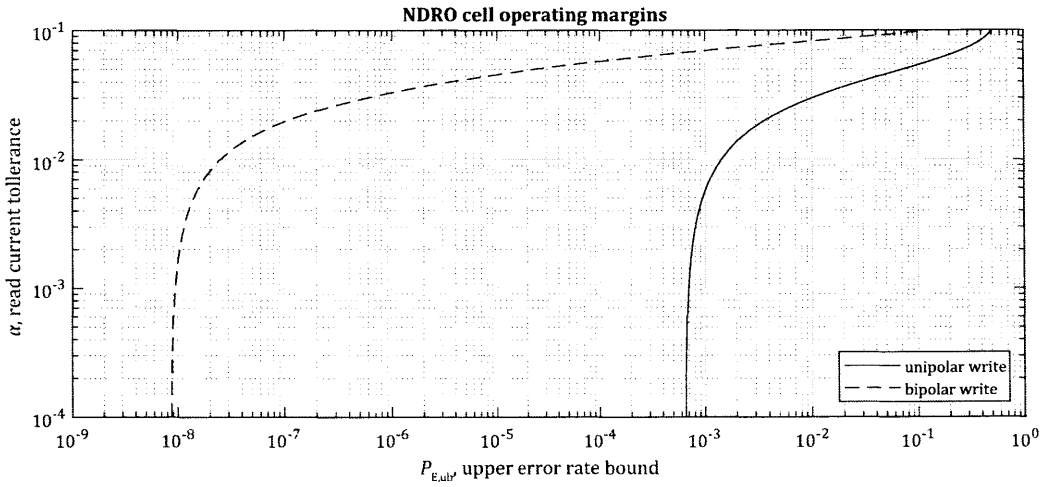

Figure 2-16: Extrapolated readout operating margins for both the unipolar and bipolar operation of the cell. These results were found using the fits shown in figures 2-14 and 2-15. The vertical axis indicated the relative tolerance in the yTron readout current. The horizontal axis provides an upper bound on the error rate. Examination of this graph allows for the determination of the readout operating margin for a desired upper bound on the error rate.

regimes, the fit results were analyzed in further detail. Using the CDF for the Burr distributions, and the fits shown in figures 2-14 and 2-15, the error rate, as a function of the read bias tolerance α, were found. In order to perform this calculation, the worst-case error rates at the extents of the tolerance interval, centered around the optimal current $I_{\mathrm{R,op}}$, were considered. That is, we calculated an upper bound on the error rate as

$$P_{\mathrm{E,ub}}(\alpha) = \max(P_{\mathrm{W0,R1}}((1 \pm \alpha)I_{\mathrm{R,op}})) + \max(P_{\mathrm{W1,R0}}((1 \pm \alpha)I_{\mathrm{R,op}})). \tag{2.9}$$

This provides only an upper bound; so for some tolerance α we will have that the error rate is $P_{\mathrm{E}}(\alpha) \leq P_{\mathrm{E,ub}}(\alpha)$. The results of this analysis are shown in figure 2-16. This figure shows that, in order to obtain an error rate close to the best that cell can provide, the read tolerance must be around $\pm 0.1\%$. Further increase in the read tolerance to $\pm 1\%$ results in an error rate on the same order of magnitude as the best that can be obtained from that cell. Any tolerance worse than $\pm 1\%$ will result in a rapidly deteriorating error rate.

2.5 Revised design

While the NDRO cell design presented in section 2.2.4 performed well, there are a number of metrics that could be improved. The error rates were poor when the cell was operated in the unipolar regime. Although the error rates improved when a bipolar write current was used, in a practical application, the generation of such bipolar biases would be more complex to implement than unipolar biases. Thus, it would be preferable if the cell operated with low error rates, even when a unipolar write pulse is used. Further, an increase in the operating margins would greatly improve system integration, and aid in scaling the cell into an array where we will also have to contend with process variations.

2.5.1 Revised cell design

One possible approach to increase the margins and decrease the error rate is to increase the magnitude of the persistent current. As per the calculations presented in section 2.2.1, in particular equation 2.1, we know that in order to write the maximum current to the loop, we want to maximize the ratio $L_{\mathrm{R}}/L_{\mathrm{L}}$. It is difficult to decrease the inductance of the hTron channel substantially, so L_{L} is somewhat fixed. It is relatively simple to increase the loop inductance by increasing the number of squares in the loop, as shown in figure 2-17. The downside of this approach is that the cell size increases substantially.

In order to increase the L_{R} inductance, a number of modifications to the layout

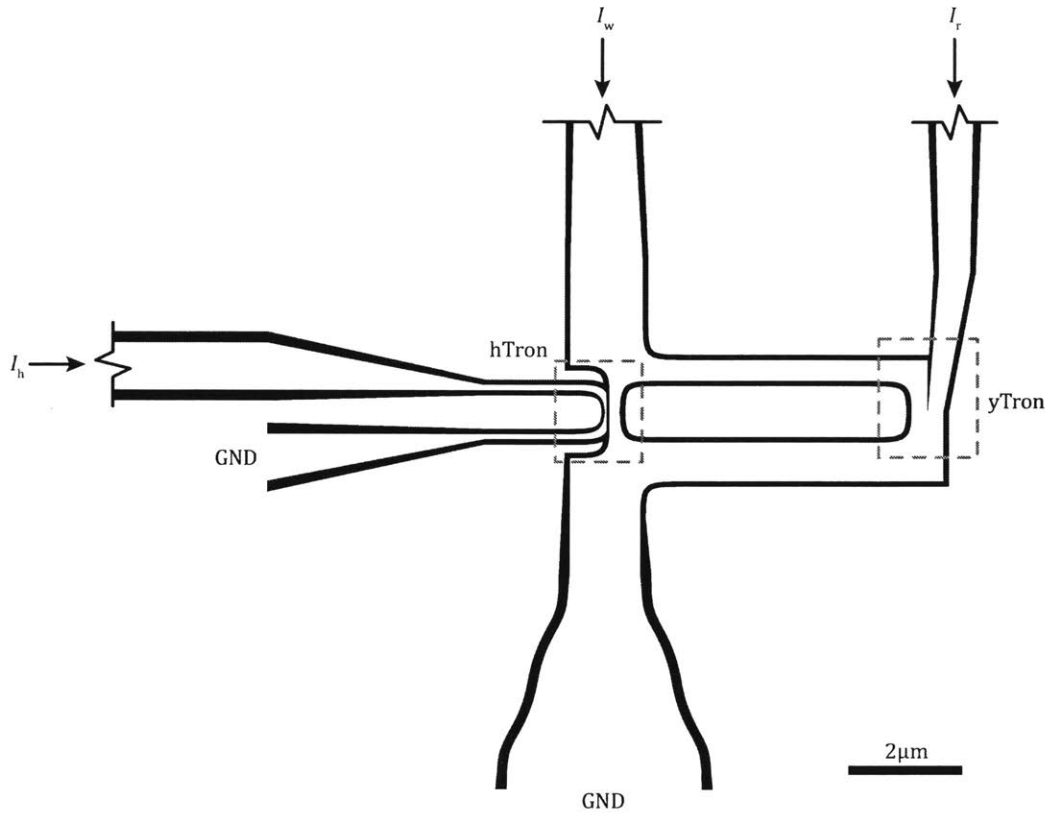

Figure 2-17: Layout for the revised NDRO cell used in the low error rate measurements. The design shown is the exact layout of the device that was tested in the following section. The black area indicates where NbN has been etched away (leaving the bare substrate below), the white area is where NbN remains. The jog lines indicate leads that extend to the connection pads. Note the hTron and yTron have been highlighted by a dashed line around each device. This layout can be seen to be a stretched version of the original layout shown in figure 2-6. Two additional changes were made. First, the ground connection was made narrower than the cell, and shifted below the hTron channel so as the further increase L_R. Second, the design of the hTron was modified in an attempt to reduce the hotspot size.

were made. The first and most evident, is that the loop was stretched such that the yTron and hTron are now farther apart. Stretching the cell in this way allowed for more square along the top part of the loop. In addition, the bottom of the loop was closed and a port added only to the left-hand side, further increasing the loop inductance. It should be noted that, the bottom of the loop was made wider that the top, such that no part of the loop other than the yTron and hTron will switch during

experiments – in particular if we choose to bias the yTron during a write. The design of the hTron was also modified in an attempt to localize the hotspot in the gate and hence the channel, and also to decrease L_{L}.

With this layout, and using a calculated kinetic inductance of $60\,\mathrm{pH}/\square$ we find that $L_{\mathrm{L}} = 0.37\,\mathrm{nH}$, and $L_{\mathrm{R}} = 1.37\,\mathrm{nH}$. With these values, we expect a single fluxon to correspond to a current of $\Delta I_{\mathrm{p}} = 1.19\,\mu\mathrm{A}$. We also expect a write efficiency of $I_{\mathrm{p}}/I_{\mathrm{W}} \approx L_{\mathrm{R}}/(L_{\mathrm{L}} + L_{\mathrm{R}}) = 0.79$. Thus, we will achieve a higher efficiency write than in prior experiments, and will be writing much more than a single fluxon to the loop. The hTron and yTron constrictions are the same as that presented in section 2.2.4, and are expected to perform similarly.

2.5.2 Low error rate measurement setup

The measurement setup used in section 2.4 is too slow to practically assess error rates less than about 10^{-5}. This is primarily due to the limited update rate of the oscilloscope ($\sim$ 100 updates per second). Thus, a new setup, exclusively for the assessment of the low error rate operation of the memory is needed. Since this setup will only be used for low error rate measurements, we can be somewhat more relaxed with the design of the BERT. In particular, we can assume that it is unlikely that we will encounter write-one-read-zero and write-zero-read-one errors simultaneously.

Taking advantage of the known-low error rates, we can use a counter to build a fast experimental setup – as shown in figure 2-18. While this setup cannot distinguish between an equal number of write-one-read-zero and write-zero-read-one errors. This setup does however, improve upon the setup used in section 2.4, as the new setup is capable of using a pseudo-random bit sequence (PRBS). This is adventitious, as a PRBS write better replicates a real-word use of the cell.

The revised experiential setup is a derivation of that covered in section 2.4.4. The new setup uses the same instruments as were used in the previous experiments, with the exception being that two Keysight 33250A AWGs were used (AWG2 and AWG3), and a Stanford Research Systems model SR400 counter was added. The operation waveforms are identical to those used in the previous experiments, with the exception

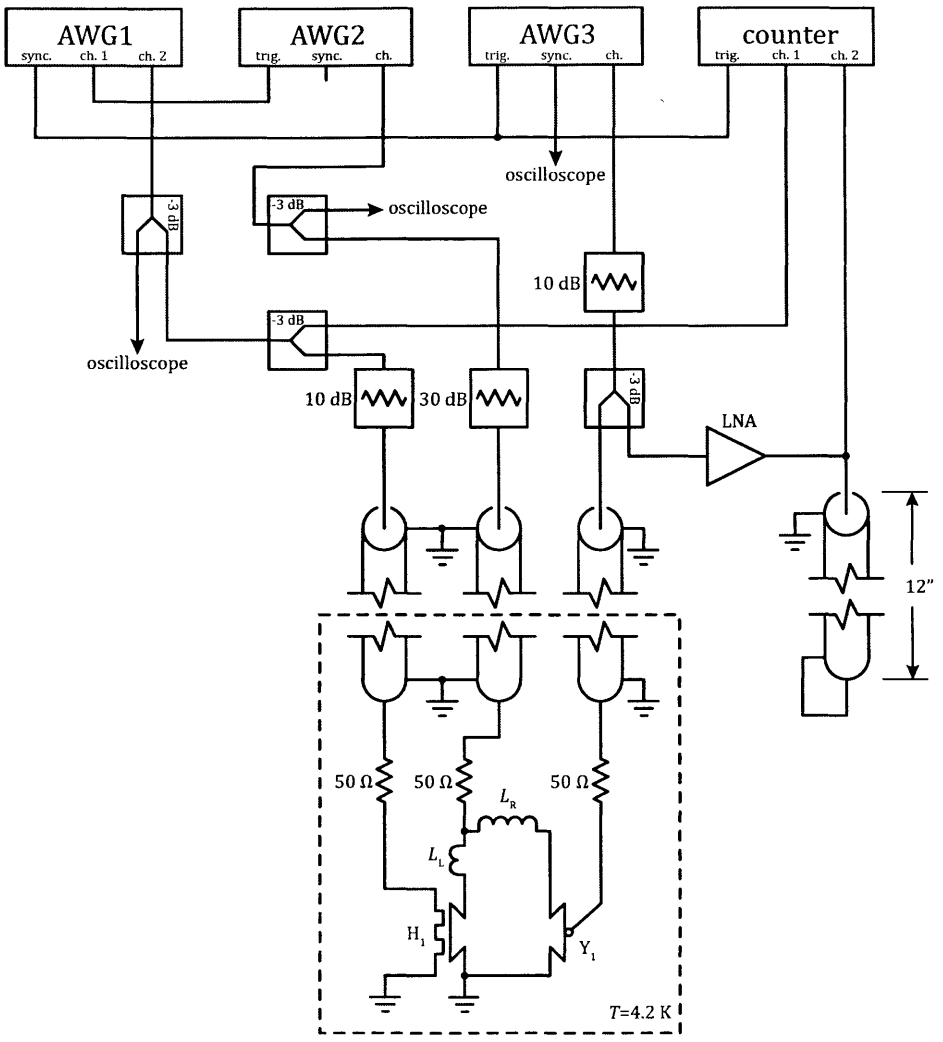

Figure 2-18: Experimental setup for the final, low BER NDRO cell measurements. A known PRBS is generated by AWG1 and used to trigger AWG2 which provides the write bias. The second channel of AWG1 provides the write enable signal. AWG3 is synchronized to AWG1, and provides the read bias pulse, which is swept during the experiment. A counter is used to track the number of writes, and the number of zeros read from the cell. From the counter's results, the error rate can be estimated since we know the intended number of zeros written. To reduce reflections 50 Ω series termination resistors were added close to the sample on the sample PCB.

that the yTron now utilizes a pulsed bias. The amplitude of the yTron pulse is chosen such that the yTron is intended to only fire if the cell is in the "0" state, and will stay superconducting if the cell is in the "1" state. Finding this level is difficult, and as a result, a sweep of this level was performed. An added benefit of performing this

sweep is that we can determine the yTron read current margins from the BER results.

AWG1 provides the PRBS, which is used as the trigger for AWG2. AWG2 provides the write current (write "1") when it is triggered by the rising edge of AWG1's channel 1, otherwise it provides a zero bias. The second channel on AWG1 is used to deliver the write enable signal. AWG3 is triggered by the synchronization signal from AWG1, and provides the read pulse to the yTron. The high level of the read pulse is varied, and the error rate monitored at each level. As with the previous setup, a system impedance of $50\,\Omega$ is maintained throughout, with the exception of the synchronization signal. Thus, power splitters and attenuators were used – again in a similar manner to those used in the former setup. In the new setup, series termination resistors were added to the sample PCB to reduce reflections.

Since the yTron read level is varied in the experiment, it was difficult for the counter to reliably trigger on each read pulse correctly. This issue was remedied by extracting only the rising edge of the pulse. This edge was extracted by using an approximately 12" length of coaxial cable terminated into a short. This effectively only passes the rising edge of the signal, before the reflection from the short interferes with the original signal, which leads to the counter only seeing an impulse when the yTron switches. The two channels of the counter are used to monitor the number of writes, and the number of times the yTron switches. Since we know the number of ones and zeros in the PRBS we can then extract an estimate of the error rate by looking at the number of times the yTron switched.

2.5.3 Experimental results

After preparing the sample, measuring resistances, and generating IV curves, in the manner covered in section 2.4, the first experiments on the new device were identical to those performed in section 2.4.5. Performing these measurements allowed us to verify that the device is functioning, and that the error rate is, in fact, low enough that the new setup will generate valid results. As expected, the cell performed better than the previous cell design, and we moved onto the new setup for low BER measurements.

The setup was changed to that shown in figure 2-18. The experimentally deter-

mined optimal operating levels were transferred from the initial tests, and further refined. The final operating parameters are shown in table 2.3. All measurements were performed with a unipolar write pulse, and a zero bias current when clearing the memory to the "0" state. For the readout, the limits of the yTron switching current were found, and the readout current was swept to include these two extremes. The final pulse amplitude sweep included values from $I_{\mathrm{R}} = 50\,\mu\mathrm{A}$ to $I_{\mathrm{R}} = 70\,\mu\mathrm{A}$ in steps of $0.2\,\mu\mathrm{A}$.

Table 2.3: Revised NDRO pulsed readout operating point.

Parameter	Value
Write bias high level, I_{W}	$32.0\,\mu\mathrm{A}$
Write bias pulse width, T_{w}	200 ns
Write bias edge times, $T_{\mathrm{w,rf}}$	50 ns
Write bit rate, F_{w}	300 kbps
PRBS length, L_{w}	32 b
Write enable high level, I_{WE}	$10.00\,\mu\mathrm{A}$
Write enable pulse width, T_{WE}	8.00 ns
Write enable edge times, $T_{\mathrm{WE,rf}}$	3.30 ns
Pulse read width, T_{R}	200 ns
Pulse read edge times, $T_{\mathrm{R,rf}}$	100 ns

After performing an acquisition, the resultant pulse counts were analyzed. At each read bias level, the counter gives us two numbers, namely the number of trials N_{T}, and the number of pulses is witnessed N_0 (which is the number of times a zero was read). In the ideal case, when the error rate is zero, we expect $N_0 = 3N_{\mathrm{T}}/4$. We expect this because the PRBS that was used contained an equal number of zeros and ones, and the AWG that provides the write bias only when it sees a rising edge, then the sequence written to the device is 75% zeros and 25%, ones on average. For the exact sequence we used, these vales hold.

When the cell is biased below the optimal read bias $I_{\mathrm{R,op}}$, the number of zero read is less than expected, so we have $N_0 - 3N_{\mathrm{T}}/4 < 0$. This bias condition corresponds to an excess of write-zero-read-one errors. On the other hand, when the bias is greater than the optimal value, we see too many zeros reads, that is $N_0 - 3N_{\mathrm{T}}/4 > 0$. This

bias condition corresponds to an excess of write-one-read-zero errors. To present these results in a more meaningful manner than a raw number of errors, these values were converted to an estimated error rate

$$P_{\mathrm{E,e}} = \frac{\left|N_0 - \frac{3}{4}N_{\mathrm{T}}\right|}{N_{\mathrm{T}}}. \tag{2.10}$$

The results of this analysis are shown in figure 2-19.

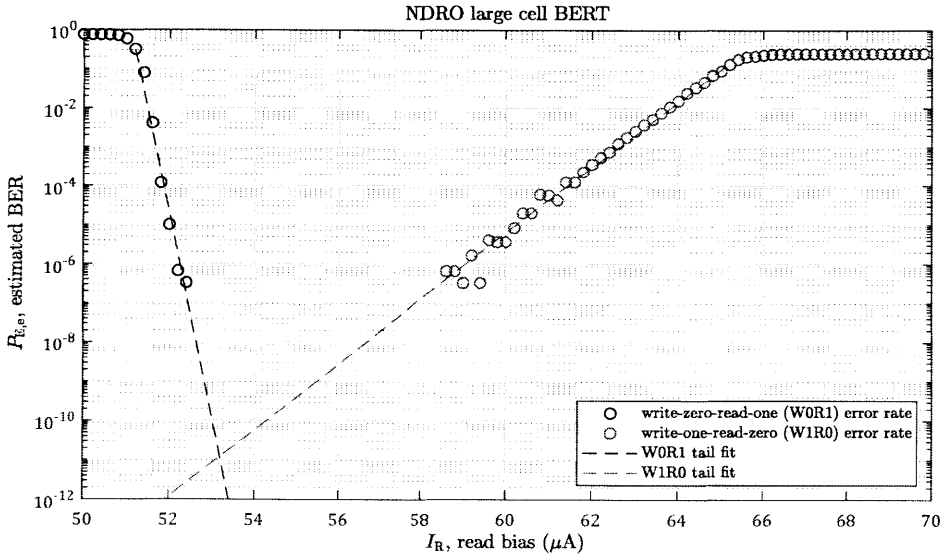

Figure 2-19: Results of the BERT on the revised cell. Each point represents one experiment at one read bias level. At each point, at least 3×10^5 write/read operations were performed. Two fits were added to the plot, one for the tail of the write-one-read-zero (W1R0) error and the other for the tail of the write-zero-read-one (W0R1) errors. The intersection of these fits predicts an ultimate error rate around $P_{\mathrm{E,min}} \approx 10^{-11}$. However, these fits lines are relatively, steep which means that the margins in the readout bias levels will be very small.

In order to predict the ultimate BER, two fits were generated for the tails of the write-zero-read-one and write-one-read-zero error rate tails. These fits predict an ultimate error rate of $P_{\mathrm{E,min}} \approx 10^{-11}$ at the optimal read bias point $I_{\mathrm{R,op}} = 53.2\,\mu\mathrm{A}$. This is a substantially lower error rate than what was obtained with the original cell. This new cell, with a unipolar write, achieves seven orders of magnitude lower error rates than the original cell operated with a unipolar write bias, and two order of

magnitude better error rates than the same cell operated with a bipolar write bias.

The error rate can be seen in figure 2-19 to be dependent on the read current level. If we choose for our error rate to be $P_{\mathrm{E}} = 10^{-10}$, then our read bias high level must be $I_{\mathrm{R}} = 53.675 \pm 0.575\,\mu\mathrm{A}$, that is the write tolerance must be $\pm 1.07\%$. This is a reasonable, albeit somewhat tight, operating margin; however, it remains to be seen if device-to-device variation is worse than this. If we choose to sacrifice our ultimate error rate, then wider margins can be attained, and scale linearly with respect to error rate exponent. For example, an error rate of around 10^{-7} would be obtained with a read current tolerance of $\pm 5\%$. With these low error rates, and reasonable operating margins, this experiment has shown that the revised cell design is capable of very low error rate operation, and is a suitable candidate for scaling into an array.

Chapter 3

NDRO array design

With the success of the single NDRO cell, our sights shift to scaling this design into an array. Two main array architectures were explored, and their various pros and cons compared in section 3.1. The result of this analysis was an array design that required the development of a superconducting multiplexer, the design, as well as the testing of a prototype device, is covered in section 3.2.

3.1 Array architecture

Due to the three-terminal design of the NDRO cell, there is a requirement for a method of selecting the cell, without also providing parasitic paths for supercurrents. There are a number of possible array architectures that could achieve isolation, and allow the cell to operate in an array. Of the many possible methods, two were investigated in detail. The first being a modified cell design that shunts parasitic currents through resistors, covered in section 3.1.1. The second design is a multiplexer approach which aims to eliminate the charging of parasitic loops, there by rendering thereby avoiding the issue, the method is covered in section 3.1.2.

3.1.1 Resistively isolated design

The resistively isolated design essentially adds a read enable port to the cell. This port enables and disables the operation of the yTron. Isolation is achieved by means of shunting the read port to ground through a resistor. The modified NDRO cell schematic is shown in figure 3-3. Two resistors are required to prevent parasitic paths from carrying persistent currents. The first resistor R_1, prevents a persistent current being trapped in the loop formed by the second hTron H_2, and the yTron (as well as in higher order paths such as those through the memory loop itself). The second resistor R_2, allows for the bias current to enter the cell when the H_2 is open, and prevents the bias from being shorted when the H_2 is closed.

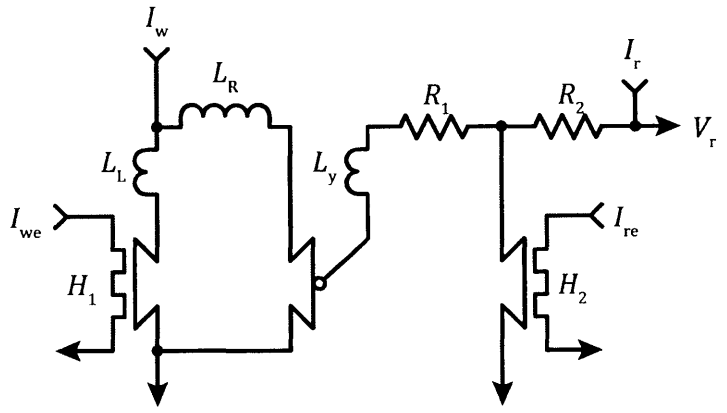

Figure 3-1: Schematic of a resistively isolated NDRO cell. The additional hTron is normally shorting the yTron port to ground. This prevents the write current I_r from being seen by the yTron, and so allows the read port to be common to all cells in the column. The two resistors prevent parasitic supercurrent paths from forming.

In order to write to the resistively isolated cell, a procedure identical to that required for the original NDRO cell is used. The resistor R_1, combined with the yTron parasitic inductance L_y, prevents any significant current from flowing through the yTron bias port during the write process. Reading from the cell requires two signals to be applied to the cell, and the resulting voltage sampled. The read bias and read enable current must be applied to the cell in order to perform a read operation. The order in which these signals are applied to the cell is arbitrary. With the read bias and read enable signals both applied, the read current will pass through R_2, and

resistively split between R_1, and the channel of the hTron. Since the resistor R_1 is chosen to be much smaller than the normal resistance of the hTron channel, the bias will mainly flow through the yTron. Once the bias is applied to the yTron, the same procedure that is described in section 2.2 will occur. That is, a voltage will either develop across the yTron, or it will remain superconducting. Given that the yTron used in the resistively isolated cell is of an identical design to that used the NDRO, then we can assume it requires the same read bias I_{R}. Thus, the bias that must be applied to the read port for a single cell is

$$I_{\mathrm{r}} = I_{\mathrm{R}} \left(1 + \frac{R_1}{R_{\mathrm{n,H}}} \right), \tag{3.1}$$

where $R_{\mathrm{n,H}}$ is the hTron channel normal resistance (typically on the order of $1\,\mathrm{k}\Omega$). When the cells are arranged in a column as shown in figure 3-3, the read bias current to the entire column of height n is simply $I_{\mathrm{r,col}} = nI_{\mathrm{R}}\,(1 + R_1/R_{\mathrm{n,H}})$.

When selecting the resistance values, there exists a trade-off between access speed and power dissipation. The smaller the resistance, the lower the power dissipation. The larger the resistance, the shorter the time constant $\tau = L/R$. Our inductances tend to be on the order of $10\,\mathrm{nH}$, and our time scales on the order of $1\,\mathrm{ns}$. Thus, we might choose the time constant to be ten times smaller than the expected time scale, that is $\tau = 100\,\mathrm{ps}$. With this time constant, we might choose resistance values on the order of $R = 100\,\Omega$. With $R_1 = R_2 = 100\,\Omega$, we have that the power dissipated during a read of a single cell is $P_{\mathrm{r}} \approx I_{\mathrm{R}}^2(R_1 + R_2)$, which for a typical read current of $I_{\mathrm{R}} \approx 50\,\mu\mathrm{A}$ corresponds to $P_{\mathrm{r}} = 0.5\,\mu\mathrm{W}$.

The timing diagram for a read operation is shown in 3-2. There are two timing limitations, both being setup times. The first is the read enable setup time $T_{\mathrm{s,re}}$, and the second is the read bias setup time $T_{\mathrm{s,b}}$. These times are expected to be very short – on the order of $100\,\mathrm{ps}$; however, no limit has been experimentally determined. It is expected that $T_{\mathrm{s,re}} > T_{\mathrm{s,b}}$, since the read enable signal must propagate thermally from the heater to the hTron channel. It is also expected that the read enable setup time is approximately equal to the write hold time, $T_{\mathrm{s,re}} \approx T_{\mathrm{h}}$.

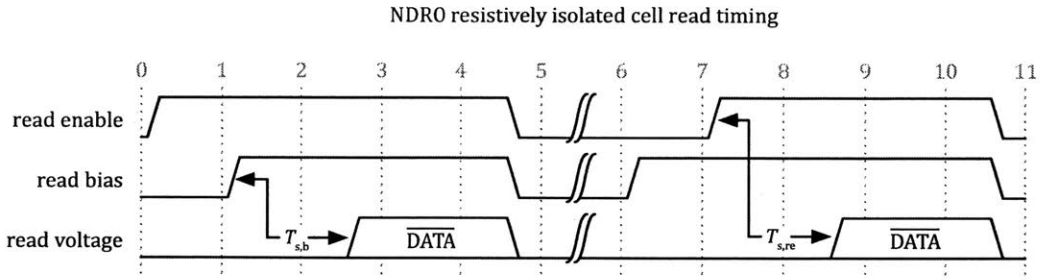

Figure 3-2: Timing diagram showing the two types of timing-limited read operations. On the left is a read-bias-setup-time $T_{s,b}$ limited operation, and on the right is a read-enable-setup-time $T_{s,re}$ limited operation. It can be seen that in both cases, there is a setup time between the application of the bias and the data becoming valid $T_{s,b}$, and between the application of the read enable signals and the data becoming valid $T_{s,re}$. It should be noted that regardless of the order in which the enable and the bias signals are applied, both setup times must be satisfied.

With the resistively isolated cell design, the array can simply be formed by tiling the cells, as shown in figure 3-3. Again, due to the possibility of forming parasitic loops, the cells cannot simply be connected in rows and columns, instead there are three different connection schemes used for the channel, the yTron, and the select lines. The channel is the most straightforward. The write port of the first cell is connected to the external memory controller. The common port of this cell is then connected to the write port of the cell in the following row. This is repeated until the final row where the common port is connected to ground. This arrangement is possible since there is no superconducting path that a current applied to the top of the column could take, other than through each write port. Thus, with the correct cell selection it is possible to access only the desired row. The yTron port has been isolated by the second hTron and the resistors. Thus, all the yTrons in a column can simply be connected in parallel. The resistively isolated design trades a simple cell access scheme, for an increased power dissipation, as discussed in section 3.1.3; however, this arrangement does allow for simple access to the desired cell. Row selection is achieved by means of only heating the hTrons in the desired row, thus allowing the yTron bias to only pass through the desired yTron. Then, by choosing the bias level correctly it is possible to have the yTron switch for one memory state,

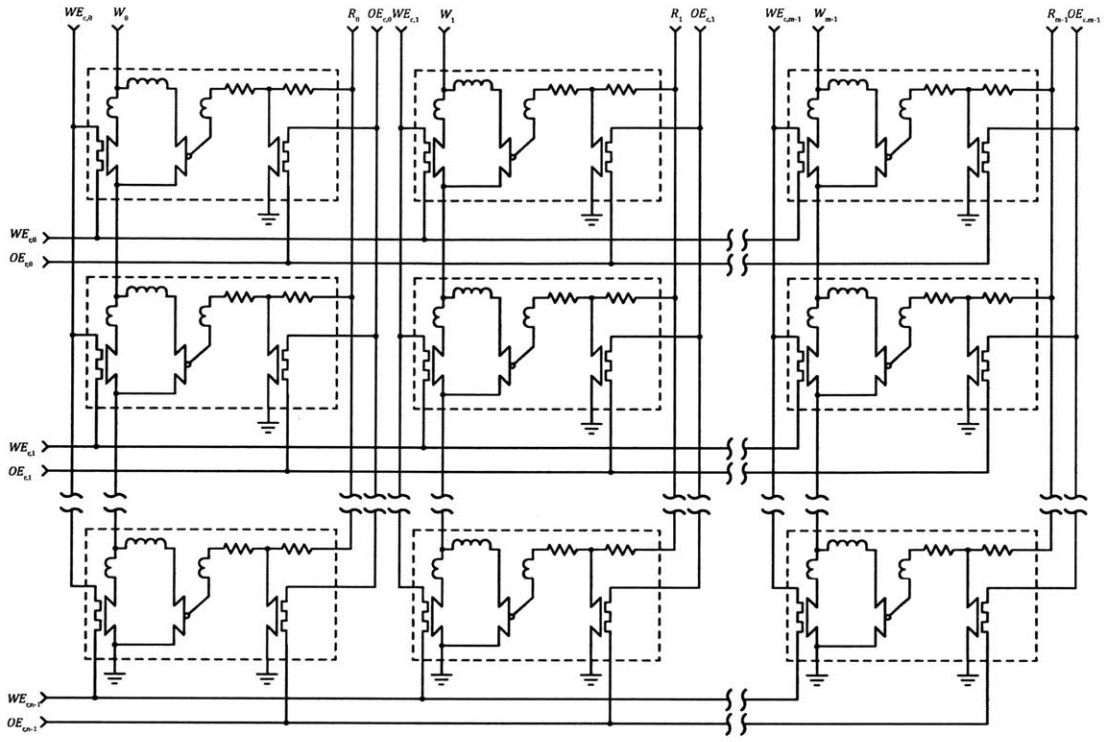

Figure 3-3: Schematic showing the connections between resistively isolated NDRO cells to form an m-bit word, n-row bank. In this figure the use of resistive hTron gates is assumed (provided bit-access is desired), if superconducting gates are used then series resistors are required. It can be seen that forming an array from the resistively isolated cells involves the hTron gates being connected in cross-bar arrangements, read ports connected in parallel along columns, and cell write ports stacked along the common ports in columns.

and remain superconducting for the other state. Finally, all the hTron gates (for both the write and read hTrons) are connected in a cross-bar fashion. If bit-access is desired, then the gates need to be either of a normal metal composition, or if they are superconducting, have a series resistor added. If only word-access is desired then the column write enable, and output enable signals can simply be connected to ground, and no resistances are required.

With the given array architecture, and a very small modification, both a bit-access and a word-access scheme are both possible. For the word-access scheme, writes and reads can be performed following the timing diagram shown in figure 3-4. In this scheme, the column write enable $\mathrm{WE_c}$ signals, along with the column read enable

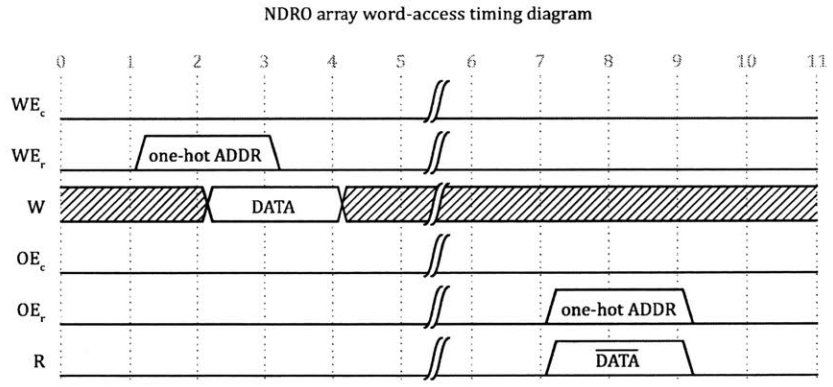

Figure 3-4: Timing diagram for word-access write and read to the NDRO cell array of either the resistively isolated or multiplexed design. The signal labels in this figure reference buses, specifically WE_c is a bus formed by the column write enable lines, WE_r is the row write enable lines, W is the write signals, OE_c is the column output enable lines, OE_r is the row output enable lines, and finally R is the read lines. For the word access scheme, the column lines, both write and read enables, are all grounded; alternatively, these could be held at a high potential and active-low row signals could be used. The first step in order to perform a write is to assert the desired rows write enable signal. The selection of the desired row is equivalent setting WE_r to be the one-hot address of the desired cell. Upon the falling edge of the write enable signal, the write data must be valid. That is, the data to be written to that row must be present at the W port. Once the write enable signal has been deasserted, the write port can be set arbitrarily. For a read operation, the desired row is selected by asserting the corresponding output enable signal. This operation is equivalent to setting the OE_r to the one-hot address of the desired row. The read bias current is also applied to the read port. The output enable and read bias signals can be applied in any order, or simultaneously – as shown here. The read port will then present the result of the read asynchronously. The presented output will be the complement value stored at the selected address. The data will be held, and is valid, for as long as the output enable and read bias is maintained. When the read is complete the biases are simply removed and the state of array will be unaffected.

signals OE_c are simply tied to ground. Alternatively, these signals could be tied to a high level, and the corresponding row signals made active-low. All signals are currents, and there need not be any high-impedance paths. The fact that the signals can be low-impedance means that parasitic capacitances are less likely to become an issue; however, parasitic inductances must be managed appropriately – that is, impedances must be high enough that the corresponding L/R time constants are short enough that inter-operation interference does not occur.

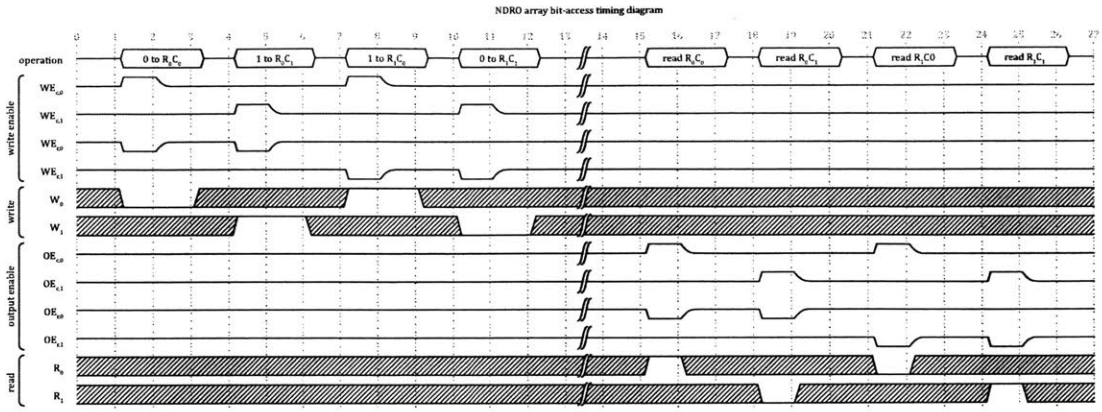

Figure 3-5: Timing diagram demonstrating the bit-access applied to a NDRO array of either the resistively isolated or multiplexed design. This timing diagram shows both write and read accesses to a 2×2 array, although the operation could be extended to an array of any size. This diagram shows four writes and four reads, one to each cell of the array. In order to achieve bit-access, tri-state write enable and/or column enable drivers are required (depending on if bit-access is only needed for writes, reads or both). A write operation to a single cell requires a current to be passed through that cell's write enable hTron gate. This can be achieved by the WE column and row drivers presenting a high impedance to all lines other than those corresponding to the desired cell. For the lines corresponding to the desired cell, one line, say the column signal, must be high, and the other, say the row, must be low. With the write enable signals set, the data to be written to the selected cell is applied to the corresponding column write port. Other columns can have any signals applied as they are not selected. The selected write enable lines are then either returned to the high impedance state, or all set to the same level. Once this is complete, the write bias can be removed, and the write operation is complete. A read operation is conducted in a similar manner. One of either the column or row output enable signals that corresponds to the selected cell is set high, and the other set low. The read bias is then applied to the read port corresponding to the column in which the desired cell resides. The voltage of this port can then be measured to determine the state of the cell. Like in the word-access scheme, the read voltage will be the complement of the data stored in that cell.

For a bit-access scheme, the timing diagram shown in figure 3-5 would be used along with modified cell constraints. There are two methods that could be used to implement a bit-access scheme. Both methods require some resistance in series with the hTron gates. The first scheme would consist of a superconducting gate with a series resistor. The signal levels would be chosen such that the current through the gate of the intersection of the desired column and row is high enough to switch the

gate without switching any other gates. For a large array, the unselected rows must carry a current around 67% that which was required for the selected cell to switch. The second possible cell design consists of hTrons that are built with resistive heaters. In this design, the unselected cells may be subject to a heater current 67% that which is supplied to the selected cell. Both of these cases are difficult to build in practice, with the former possibly being easier. If some device that enabled current flow in one direction, or at least some preference to flow in a given direction, where used then these requirements could be greatly eased.

The bit-access scheme requires the write and read enable signals to be tri-state, that is switched between a high potential, a low potential, and a high-impedance state. Such a driver is typically difficult to build; however, with the use of a hTron (or nTron), one could easily build a driver capable of such tri-state operation. In conventional circuits, tri-state drivers tend to be slow – due to the presence of high parasitic capacitance and low parasitic inductance. Here, we have little parasitic capacitance, but very high parasitic inductance, so a tri-state driver may in fact be very fast.

3.1.2 Multiplexed column design

The multiplexed column design can be considered a derivation of the resistively isolated design – although it was originally designed separately. This approach utilizes the original NDRO cell design and connects all of the read ports within a column to a multiplexer, the remaining port are connected as was done for the resistively isolated design, as shown in figure 3-6. The multiplexer would be implemented with nanowire devices, and as a result, when inactive would short the inputs together. The consequence of the normally-closed operation of nanowire devices is that we require a resistor in series with each input of the multiplexer. This method can be considered a rearrangement of the resistively isolated design, where instead of utilizing an hTron at each cell to achieve isolation, a multiplexer (likely constructed from hTrons) is used. The development of a multiplexer for this design is covered in section 3.2.

With the port names shown in figure 3-6, the timing diagram for this array, for

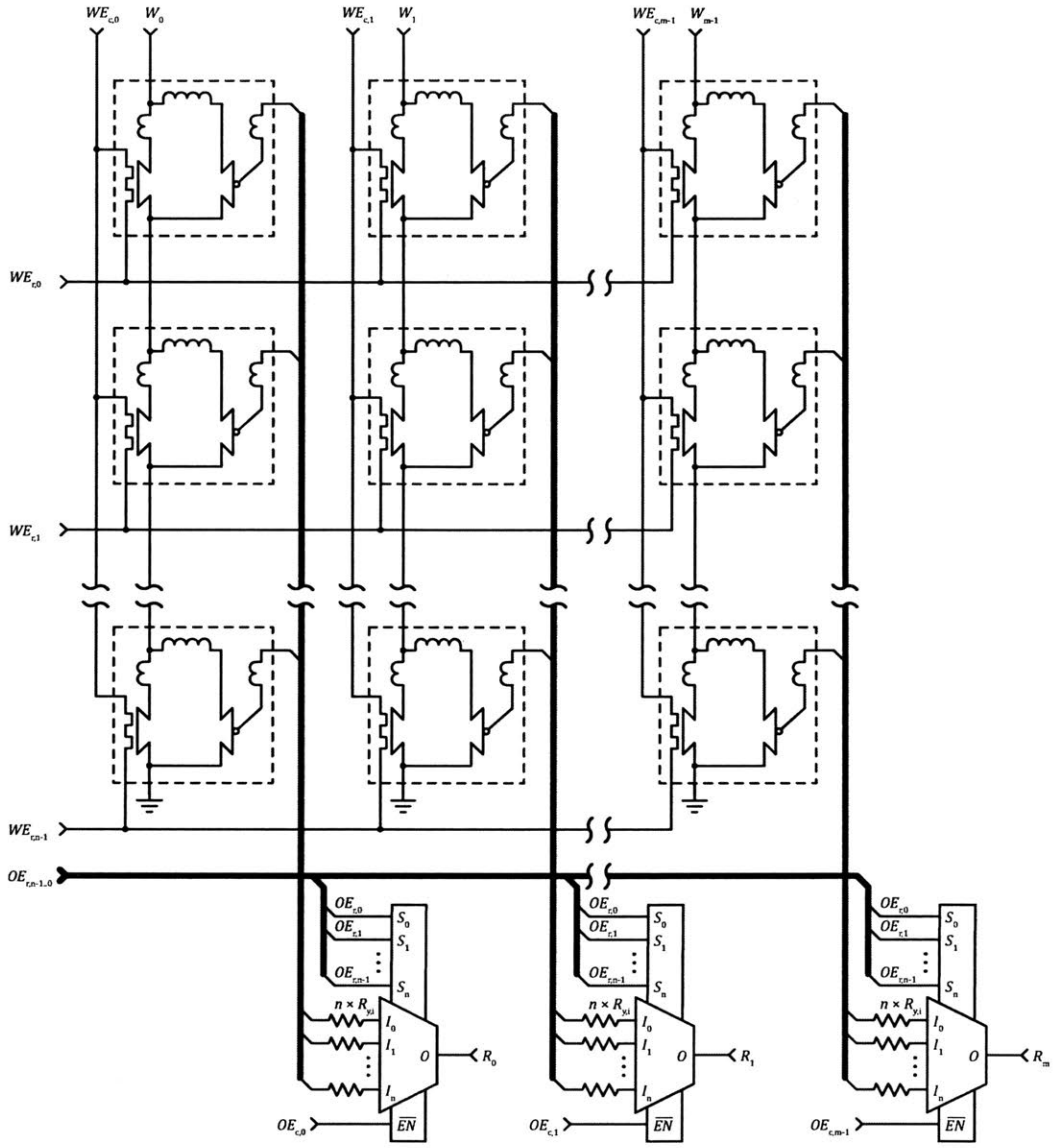

Figure 3-6: Schematic showing the connections between NDRO cells and hTron multiplexers to form an m-bit word, n-row bank. The bold lines indicate buses. In this array, the write enable heaters and write ports are connected in the same manner as that shown in figure 3-3, for the resistively isolated design. Here, the read ports of each cell are connected to multiplexers, which allow read access to the cells. As the multiplexers are constructed with nanowire devices, when no cell is selected, all read ports within a column will be shorted together. Thus, to prevent parasitic loops from forming through the multiplexer, a bank of resistors is placed at the input to each multiplexer.

both the word-access and bit-access schemes, are identical to those used for the resistively isolated design. Functionally, the multiplexed and resistively isolated design are very similar. This similarity is a result of the fact that in the resistivity isolated design, the second hTron that was added to each cell, if they were taken on their own, would be a multiplexer. The main distinction between the multiplexed column design and the resistively isolated design, is that we have a fixed multiplexer design in the resistively isolated array. Whereas the multiplexed column array allows for flexibility in the design of the multiplexer, and as covered in 3.2, much more efficient designs are possible.

3.1.3 Array design size and power comparison

In terms of array size, the only difference between the resistively isolated and multiplexed array architectures is the number of resistors and hTrons. Thus, to form a comparison between the two types, on the number of hTrons and resistors will be counted. The resistively isolated array requires two hTrons per cell – one for writing and one for reading. This array also requires two resistors per cell. Thus, for an $n \times m$ array, the number of hTrons and resistors is given by

$$n_{\text{res,hTrons}} = 2nm, \text{and} \tag{3.2}$$

$$n_{\text{res,resistors}} = 2nm. \tag{3.3}$$

Both array types write to the cell in a similar manner. The readout of each cell is the major distinguishing factor between each array type. For this reason, only the readout power dissipation will be considered. For the resistively isolated array, when reading out a cell, one hTron will be activated and a bias supplied to the read port. The read port bias will be split amongst the cells in the column – dissipating some power in the isolation resistors in each cell (R_2 in figure 3-3). In addition, some power will be dissipated in the yTron if it switches (when a "0" is read). The total power

for a read operation that results in a zero (that is when the yTron switches) is

$$P_{\text{res,read},0} = P_{\text{hTron}} + \frac{(nI_{\text{yTron,read}})^2}{\frac{n-1}{R_2} + \frac{1}{R_1+R_2+R_{\text{yTron}}}}, \tag{3.4}$$

where P_{hTron} is the power dissipated in the hTron (including the channel and gate), $I_{\text{yTron,read}}$ is the current required by the yTron for a read operation, R_1 and R_2 are the arm and isolation resistances from figure 3-3, and R_{yTron} is the normal resistance of the yTron when it switches. If the read results in a "1", and the yTron does not switch, then we have that the power dissipation is

$$P_{\text{res,read},1} = P_{\text{hTron}} + \frac{(nI_{\text{yTron,read}})^2}{\frac{n-1}{R_2} + \frac{1}{R_1+R_2}}. \tag{3.5}$$

If we assume that on average half of the reads result in a one and half of the reads result in a zero then the average power for a read is

$$\langle P_{\text{res,read}} \rangle = P_{\text{hTron}} + (nI_{\text{yTron,read}})^2 \left(\frac{R_2(R_1+R_2)}{(n-1)R_1 + nR_2} + \frac{1}{\frac{n-1}{R_2} + \frac{1}{R_1+R_2+R_{\text{yTron}}}} \right). \tag{3.6}$$

For the multiplexed array, each yTron is accessed independently, thus there needs only be one isolation resistor per cell. This prevents a fraction of the write current from exiting through the yTron and entering other cells. In terms of the number of hTrons, there are a number of means by which a multiplexer can be implemented. Here we will only consider the design presented in section 3.2. This design required $2(n-1)$ hTrons, where n is the number of multiplexed channels. An $n \times m$ memory is comprised of m, n-channel multiplexers for the read portion of the circuit. A further nm hTrons are required for the write portion of the circuit. If it is desired to have a single read, and a single write port, then an additional n-channel and an additional m-channel multiplexer would be required. However, since this would be required for both the resistively isolated and multiplexed designs, it will be ignored here. Thus, we have that the total number of resistors and hTrons required for the multiplexed

design is

$$n_{\text{mux,hTrons}} = m(3n - 2), \text{and} \tag{3.7}$$

$$n_{\text{mux,resistors}} = n. \tag{3.8}$$

The multiplexer covered in section 3.2 has been designed to minimize the number of hTrons that are on at any given time. By requiring the select lines to go high impedance (several hundred ohms is sufficient) when the line is not active, the total number of hTrons that are on for any given read operation has been reduced to $\log_2 n$. Thus, we get that the power dissipation when the yTron switches is

$$P_{\text{mux,read,0}} = m(P_{\text{hTron}} \log_2 n + I^2_{\text{yTronRead}} R_{\text{yTron}}), \tag{3.9}$$

and when the yTron does not switch we have

$$P_{\text{mux,read,1}} = m(P_{\text{hTron}} \log_2 n). \tag{3.10}$$

Thus, our average power dissipation, again assuming a roughly even number of reads resulting in a one and in a zero, is

$$\langle P_{\text{mux,read}} \rangle = m(P_{\text{hTron}} \log_2 n + \frac{1}{2} I^2_{\text{yTronRead}} R_{\text{yTron}}). \tag{3.11}$$

In order to facilitate a comparison between the memory architectures, the relative size and power dissipation of each array technique have been plotted in figures 3-7 and 3-8, respectively. The relative size of the array is calculated relative to a single cell without any circuitry required for accessing the cell in an array. These plots where calculated with the assumption that a resistor and a hTron have roughly the same area. A word size of $m = 32\,\text{b}$ was assumed. The power dissipation of each type of array has been calculated with nominal values for each parameter, namely $R_1 = 100\,\Omega$ and $R_2 = 100\,\Omega$, $I_{\text{yTronRead}} = 50\,\mu\text{A}$, $R_{\text{yTron}} = 1000\,\Omega$, and $P_{\text{hTron}} = 500\,\Omega(50\,\mu\text{A})^2 = 1.25\,\mu\text{A}$.

It can be seen from the figures 3-7 and 3-8, that for the same array size, the

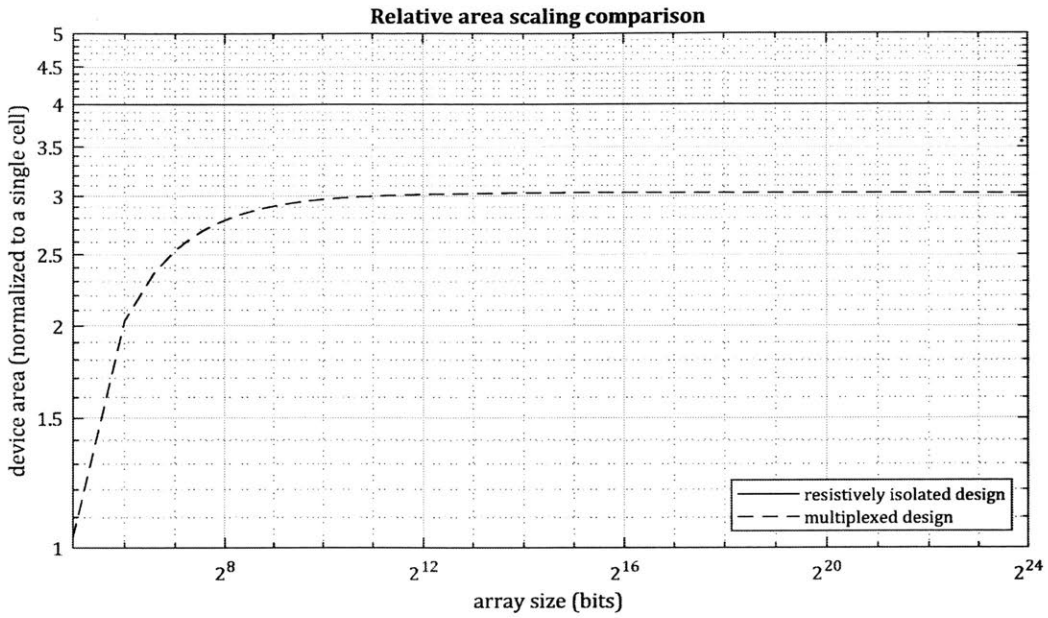

Figure 3-7: Comparison of the effective cell size for both a resistively isolated and a multiplexed column design. The relative size is computed with respect to the area of a single cell that contains a single hTron. Note that for this calculation, a word with of 32-bits was assumed. It can be seen that the resistively isolated array has a constant size of four times that of the single cell. In contrast, the multiplexed design starts the same size as a single cell for a bank containing one row, and grows as the bank size grows. The multiplexed design asymptotically approaches the limit of $3 + 1/m = 3.03125$ for an array of infinite size. Thus, the multiplexed array is always smaller than the resistively isolated array – when ignoring interconnects and the like.

resistivity isolated design requires more area and more power than the multiplexed design. For the resistively isolated design, the cell size relative to a single NDRO cell is constant at four times the size for all array sizes. In comparison, the multiplexed design starts the same size as the single cell at an array size of 32 b, and increases hyperbolically. After an array size of 2^{12} b $= 512$ B, the area of the multiplexed column cell reaches close to its limit of $3 + 1/m = 3.03125$. The power dissipation of the resistively isolated array scales as a square of the bank size n. This is to be expected since, for the resistively isolated array, the majority of the power for the read operation is dissipated in the unselected cells, the number of which scales linearly with array size. In contrast, the multiplexed column design scales logarithmically with n.

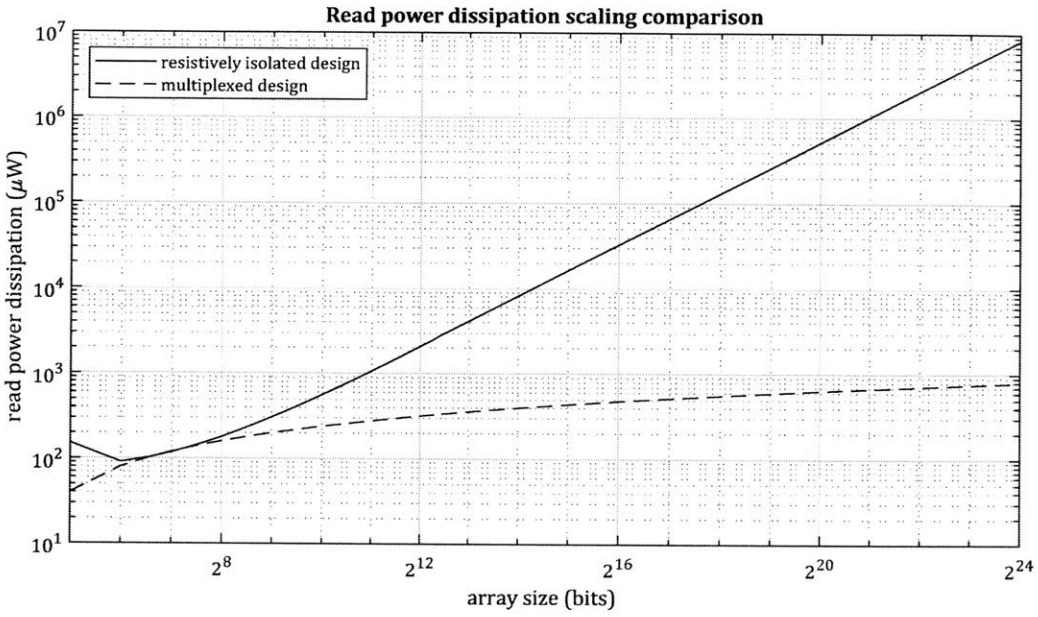

Figure 3-8: Comparison of the power dissipated during a read operation by the cell selection circuitry for both a resistively isolated and a multiplexed column array. Note that for this calculation, a word with of 32-bits was assumed, further it was assumed that the entire word was read, and that an equal number of bits were zero and one. For small bank sizes, the multiplex column and resistively isolated designs perform comparably. However, as bank size grows, the resistively isolated design's power dissipation grows as a square of bank size, whereas the multiplexed column design grows logarithmically. Thus, for large arrays the multiplex column design vastly out-performs the resistively isolated design for read power dissipation.

Thus, the multiplex column design, while requiring a more complex layout, would be preferable to the resistively isolated design.

3.2 Multiplexer design

With the resistively isolated cell design scaling poorly, we decided to pursue the multiplexer design further. However, in order to pursue this path, we need a design for a superconducting multiplexer. The following sections outline the development of such a multiplexer. The basic operation of the required circuit is covered in section 3.2.1. With an idea of the device requirements, we cover the superconducting implementa-

tion of this multiplexer in section 3.2.2. Finally, a prototype device was fabricated and is tested in section 3.2.3.

3.2.1 Multiplexer operation

There are a number of approaches that could be pursued in order to create a multiplexer using a four-terminal device such as the hTron. A conventional design, such as those that would be implemented in CMOS, would typically be comprised of two main parts, namely a decoder and an array of AND gates with their results combined in an OR gate, as shown in figure 3-9. It is of course possible to simplify the logic by merging the decoder and array of AND gates; however, for our purposes this form is more convenient. The array of AND gates with their results combined by an OR gate, can be considered to be a multiplexer of the same size as the original multiplexer; however, instead of a binary select input, it accepts a one-hot select input. A one-hot counting scheme is a unary scheme that has exactly one line high at any time – hence the name "one-hot". Any state where either more than two lines are high, or no lines are high is invalid.

The operation of such a multiplexer is fairly simple. The address, which in this case is a two bit number formed by $[A_1, A_0]$, is decoded to a one-hot signal. The output of the decoder is a signal requiring 2^n bits where n is the number of address lines – in our case $n = 2$, so the decoder has four output lines. The operation of the decoder is shown in the first two sections of table 3.1. The one-hot input multiplexer then takes the decoded signal and produced the logical AND with the corresponding input signal. Thus, copying the input signal and presenting it to the 2^n input OR gate. This gate, since there can be only one AND gate active at any one time, simply sets the output to be equal to the selected input signal. Thus, the operation of a conventional CMOS multiplexer is summarized by table 3.1.

This notion of a multiplexer can be simply expanded to an analog input/output device. That is, one where the input signals I_m, and the output signal O are analog signals, and the address lines A_m and select lines S_m are digital signals. In this case, the one-hot multiplexer portion of the circuit would resemble an array of switches

Table 3.1: Truth table for a four-to-one multiplexer.

A_1	A_0	S_3	S_2	S_1	S_0	O
0	0	0	0	0	1	I_0
0	1	0	0	1	0	I_1
1	0	0	1	0	0	I_2
1	1	1	0	0	0	I_3

2 to 4 decoder

one-hot MUX

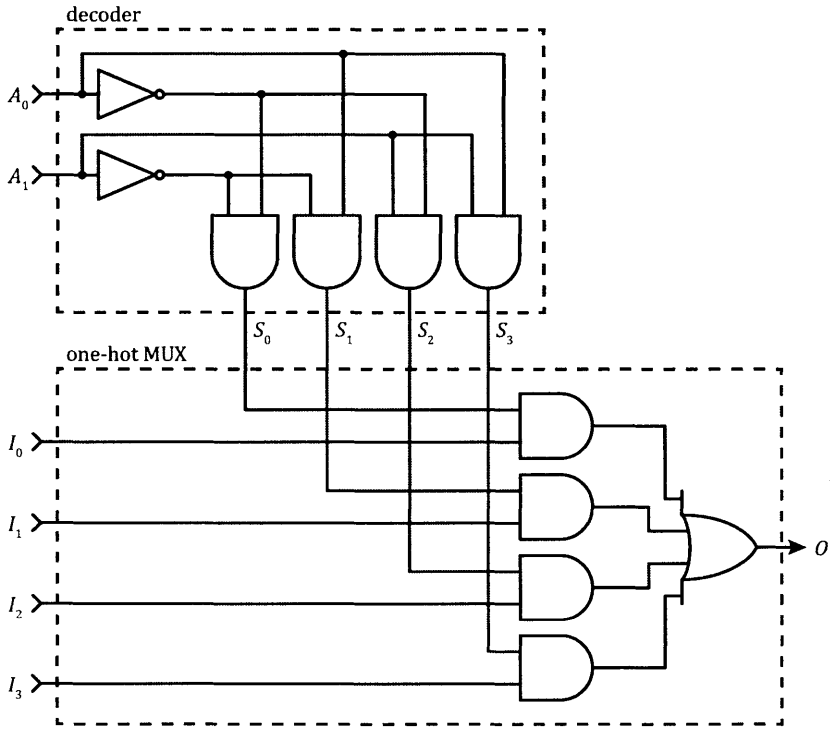

Figure 3-9: The logic that governs a typical four-to-one digital multiplexer. The schematic has been broken into two sections, with one being a two-input decoder, and the other being a four-to-one one-hot multiplexer. In this case, the select inputs to the multiplexer must be one-hot signal, and this requirement satisfied by the output of the decoder. Porting this design into nanowire devices would yield a very poorly performing device, for this reason an adaptation of this design is used instead.

as shown in figure 3-10. Is it particularly evident in the analog-case as to why the switch array can only accept a one-hot input, as if two select inputs were asserted simultaneously, then the corresponding analog inputs would be shorted together. The

activation of more than one input at any one time would short the selected inputs together. It is also notable that the switches are normally-open, and are only closed by the activation of the corresponding select line.

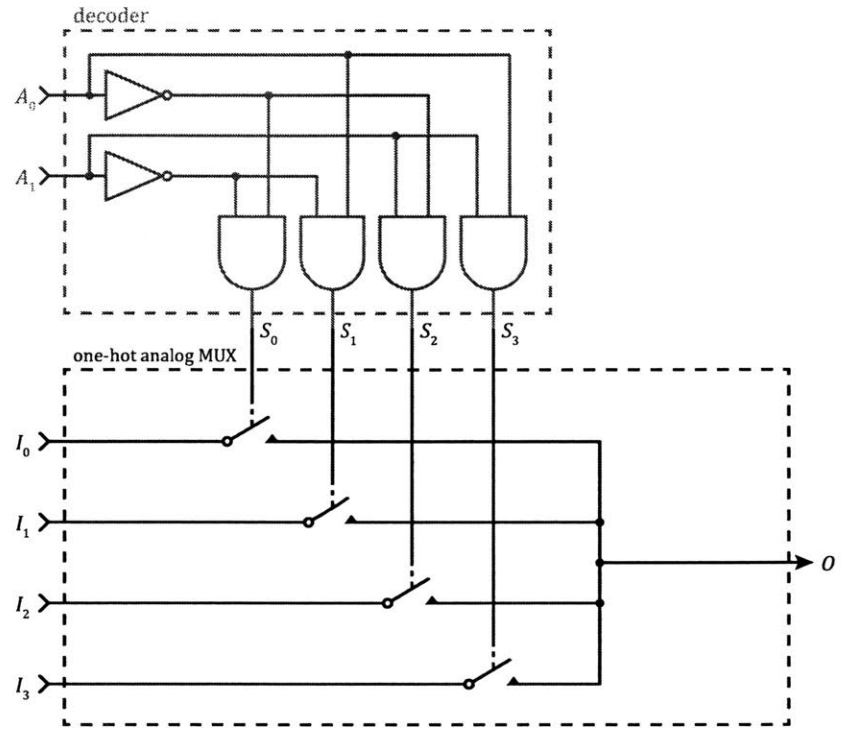

Figure 3-10: Schematic of a possible implementation of a four-to-one analog multiplexer. The parts of the circuit that are common between a digital and analog multiplexer were drawn in gray. In this analog variant of the multiplexer, the AND-gates are replaced with analog switches, and the OR-gate with a connection between all the switch outputs. For an analog multiplexer such as this, the output port "O" is also referred to as the common port.

3.2.2 Superconducting multiplexer implementation

Directly porting the digital design to superconducting devices is possible; however, it would result in a large design that is exceptionally power inefficient. Additionally, we do not have a device that can be used as a normally-open switch. The drawback of not having a normally-open switch is compounded by the fact that, of all the devices that we have which can be used as switches (nTron and hTron), require constant power dissipation in order to be kept in the open state. Thus, we must be very careful with

our design, and ensure that the least number of switches are on (open) at any one time, while also maintaining functionality.

We can achieve the same functionality as the analog multiplexer design shown in figure 3-10, by simple replacing the analog switches with hTrons, any by tying the one side of the heater to the logical high potential. Thus, since we have a one-hot input, we have no choice but to activate $n-1$ devices. There is no way to reconfigure the hTrons to get around this issue while also maintaining the same operation as the original analog multiplexer. This would result in a substantial amount of power being dissipated while the memory is not performing any operations. In fact, with this multiplexer design, we would have been better off with the original naive design presented in section 2.1. The only way around this issue is to sacrifice the compatibility with the original analog multiplexer.

From inspection of the array design shown in figure 3-6, we find that we do not need to fully replicate the operation of the analog multiplexer. In order to read from the correct yTron, we simply need to select, and thereby dissipate power, during the read of the selected column. Thus, we only waste power heating the hTrons during the read access time, and only for the selected column. However, since this means that we will leave the multiplexer shorting all yTrons during the write operation, this means that we require isolation resistors on each yTron. These prevent a supercurrent from flowing through the multiplexer during a write. This must be prevented since any current written to a parasitic loop is current lost from the desired cell, thereby reducing operating margins, and possibly leading to inter-cell interference.

With the resistors added to the yTrons, we only need to activate the multiplexer during a read. In this configuration, we could use an hTron implementation of figure 3-10; however, during a read we would still need to activate $n-1$ hTrons. We can improve on this situation by further straying from the original analog multiplexer's operation. We can attain this improvement by realizing the during a read, we need only isolate the desired cell from the other cells, while maintaining a connection to the common port. This kind of operation can be achieved by means of a tree of hTrons.

In a hTron multiplexer tree, we pass each input through a hTron. We then combine

Table 3.2: Truth table for the four-to-one hTron tree multiplexer.

S_3	S_2	S_1	S_0	$\overline{EN}$	O
HiZ	HiZ	HiZ	HiZ	HiZ	$I_0 + I_1 + I_2 + I_3$[a]
0	0	0	1	0	I_0
0	0	1	0	0	I_1
0	1	0	0	0	I_2
1	0	0	0	0	I_3

[a]This is not the logical OR operation, rather the output will be the sum of the current into each input port I_n.

the results in pairs. These paired signals are then passed through another layer of hTrons, and again combined in pairs. This is repeated until we are left with a single signal which will be our common port. A schematic of a four-input multiplexer built from hTrons arranged in such a tree is shown in figure 3-11. It may seem that this design has only increased the number of hTrons, but will provide little other benefit. This is not true since the connection of the heaters has been carefully selected such that only one heater per layer of hTrons is on at any one time. For example, if we want to have I_1 connected to the common terminal, then we would connect $\overline{EN}$ to ground, and apply a bias to S_1. In doing so, we open hTrons H_1, and H_6 leaving H_2, H_3, H_4 and H_5 superconducting. Thus, we have a superconducting path from I_1 to the common terminal, and we end up with I_2 and I_3 still shorted. The non-selected terminals being shorted or open does not cause us any issues since there are no biases applied to these terminals. The full operation of the four input hTron one-hot multiplexer is summarized in table 3.2. Thus, by sacrificing device area by adding $n-1$ more hTrons, we end up only requiring $\log_2 n$ hTrons on during a read. This represents a substantial power saving.

3.2.3 Prototype multiplexer testing

A prototype multiplexer was designed and tested. The device was only a two-to-one multiplexer, as the fabrication process available did not have a via process or a multilayer heater process at this time, and so there was no way to fabricate a multiplexer any larger – at least not without using wire-bonds or a similar off-chip

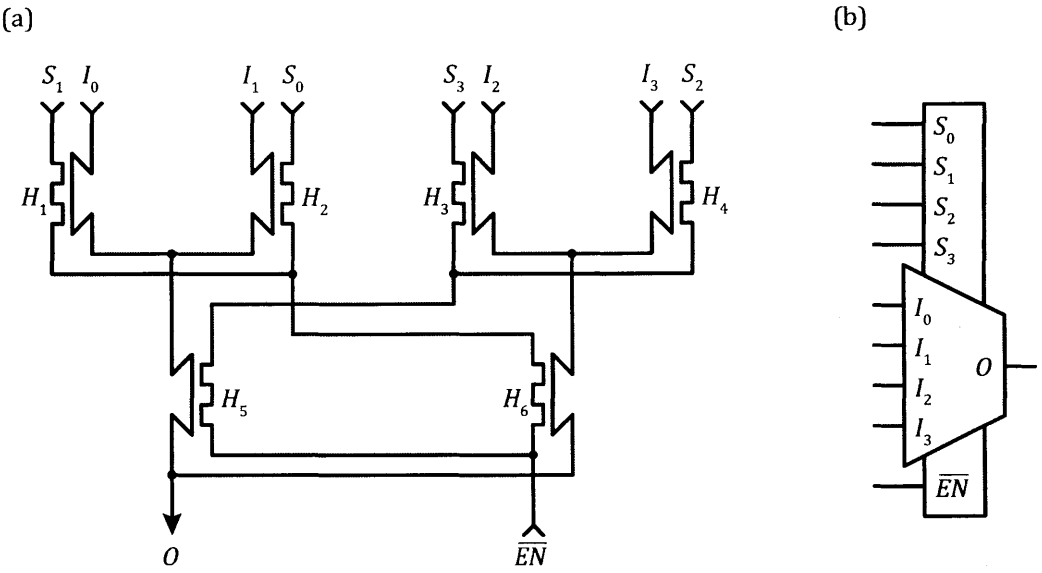

Figure 3-11: A four-to-one, one-hot multiplexer constructed with as a hTron tree. The schematic of the multiplexer is shown in (a), and the corresponding circuit symbol of the device is shown in (b). When enabled, the multiplexer allows for one of the four inputs to be connected to the common port (O) by apply a bias to the select line (S_x) with the same subscript as the desired input (I_x). The hTron tree multiplexer's gate arrangement is designed to ensure that no matter which input is selected, only one hTron per state is on. For example, in this design, no matter which of the input is selected, only two hTron will be on at any one time.

connection scheme. The design consisted of three identical hTrons connected to form one half of a four-to-one multiplexer. The hTrons were connected as shown in figure 3-13. It can be seen that the connection of these devices is identical to that in one side of the four-to-one multiplexer shown on the left of figure 3-11.

The layout the actual device that was tested is shown in figure 3-12. The hTrons used here have a channel width of 200 nm and a gate width of 40 nm. In this layout, like all layouts presented in this work, current crowding has been mitigated by the use of curves on internal edges. Additionally, all transitions in width are made with smooth and slowly varying tapers. Again, the device was fabricated with a positive tone resist, and so only the outlines of the wires (where NbN will be etched away) were written.

There are a number of tests that could be performed on this device; however, at this stage in development, we are primarily interested in basic functionality. The

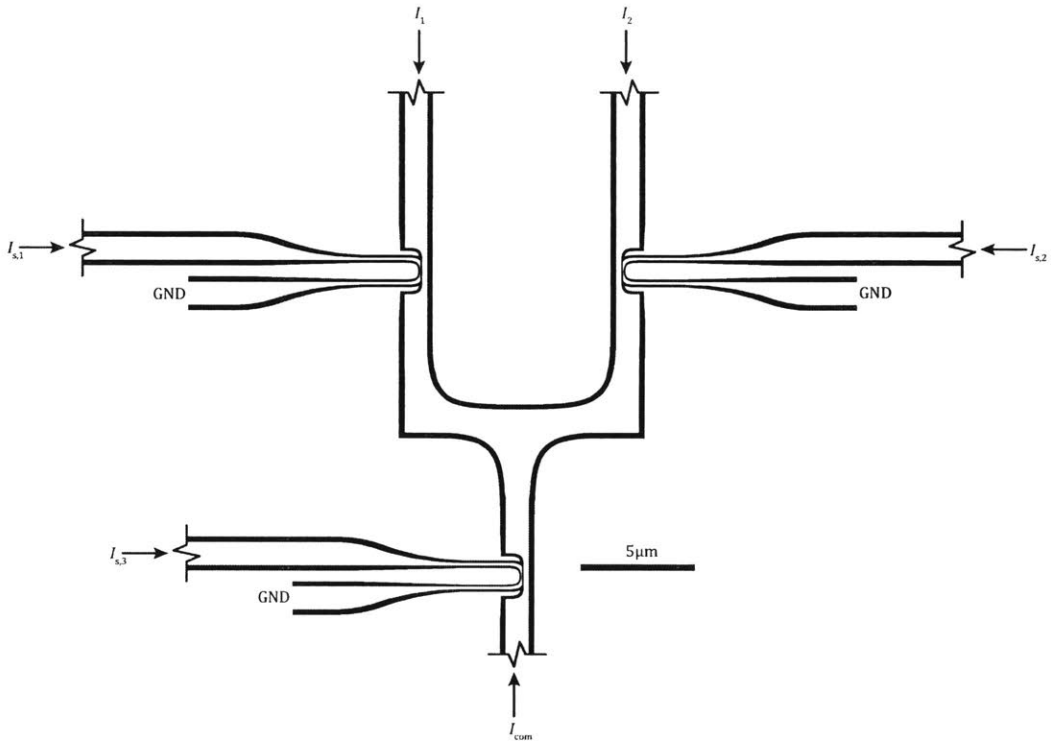

Figure 3-12: Layout of the prototype two-to-one hTron multiplexer. The design shown is the exact layout of the device tested in this section. The black area indicates where NbN has been etched away (leaving the bare substrate below), and the white area is where NbN remains. The jog lines indicate leads that extent to connection pads. While this design is only a two-to-one multiplexer, with the third hTron it becomes one half of the four-to-one multiplexer shown in figure 3-11. One side of each heater is connected to ground.

experiments performed here were to verify the ability of the multiplexer to isolate one input from the other while simultaneously providing connectivity to the common port. In order to achieve this, the multiplexer was operated in somewhat of the reverse to the usual manner in which one might use a multiplexer. A current bias was applied to the common port (as would be done in operation in a memory array), one input was grounded and the other input's voltage monitored. A bias current was applied to the common port, and was swept to generate an IV curve. This process was repeated with various currents applied to the gate of the hTron which corresponded to the grounded input. Once this measurement process was complete, the inputs were switched to ensure that both hTrons operated as expected. The third hTron had

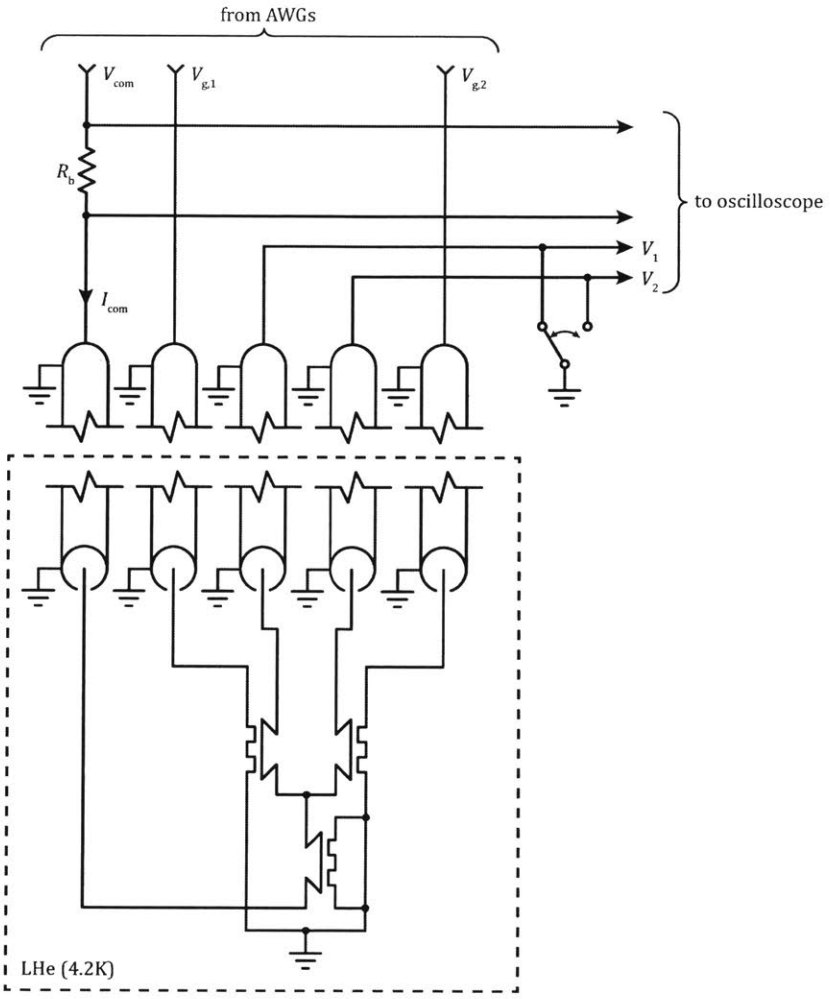

Figure 3-13: Schematic drawing of the experimental setup used to test the hTron multiplexer. A bias resistance of $R_b = 10\,k\Omega$ was used for the experiments. The current through the common port I_{com} was determined by the voltage drop across the bias resistor R_b. A gate bias $V_{g,x}$ was only applied to the input port which was grounded. In this manner, the common port should remain connected to the unselected port, which was monitored by the oscilloscope, while the IV curve of the grounded port was tested. If we see typical hTron IV curves, without the channel through which we are monitoring the voltage, switching, then we can be sure that the basic operation of the device is sound.

its gate terminals grounded and was not used in this experiment. The experimental setup is shown in figure 3-13.

The results of the isolation experiment are shown in figure 3-14. Only the two extremes of the gate voltage bias were plotted to graphically illustrate the two main

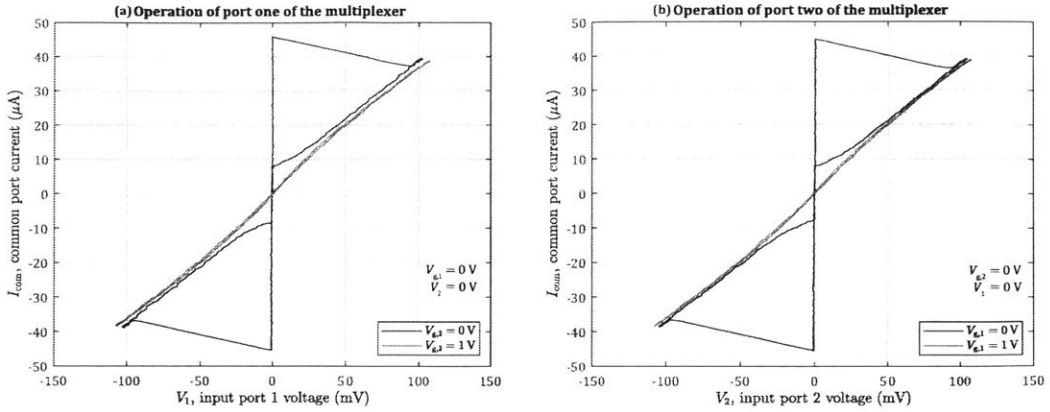

Figure 3-14: Operation of the prototype superconducting two-to-one multiplexer. These results have been decimated by a factor of 50. In (a), input two is grounded, and the voltage at input one is monitored for different voltages $V_{g,2}$. In (b), input one is grounded, and the voltage at input two is monitored for different voltages $V_{g,1}$. It can be seen that the operation of both hTrons is nearly identical, and that the multiplexer is operating as expected. The port from which we are monitoring the voltage is not switching, and the gate is successfully suppressing the opposing hTron.

states of the multiplexer. These results indicate that the multiplexer in fact working as expected. It can be seen that we can witness the switching and resistive states of the shorted hTron through the channel of the unselected hTron. Additionally, it can be seen that both hTron's IV curves are very similar with switching currents both around 45 muA. This indicates that, at least in this experiment, multiple hTrons can be produced with very similar switching and suppression characteristics. The results of this experiment indicate that the production of a larger scale hTron multiplexer should possible. However, constructing such a device would require a fabrication process that involves vias, which we do not have available at this time.

As a result of the successful results we found in this and previous chapters, we decided to pursue the superconducting-loop based memories further. However, due to the limited fabrication processes available, and uncertainty about the reliability of the yTron, we chose to opt for an array design that did not require vias, and did not use the yTron. Thus, we will pursue an array where the state of the cell is encoded in the same manner as has been done in the previous chapter, but with a new readout scheme.

Chapter 4

Destructive readout cell and array design

We have seen that NDRO memory cells can be constructed from superconducting devices, and that they can work well in isolation. But it is apparent that forming these cells into an array is far from trivial, and requires complex interface circuitry to work. This additional circuitry harms the scaling of the array as it increased power dissipation and causes the effective cell size to increase substantially. These drawbacks are primarily a consequence of the non-destructive readout circuitry (the yTron), which results in the cell having three terminals that are connected by means of a superconducting path. If we choose to abandon the non-destructive readout, and instead use a destructive readout design, then we trade the simple cell design of the NDRO memory, but complex array structure, for a more complex cell design but a vastly simpler array design. This chapter covers the design and testing of such a destructive readout (DRO) cell and array.

The basic cell design, operation, layout, and simulation are covered in section 4.1. The new DRO cell design requires a hTron on both sides of the cell, this, combined with the lack of a via process, make all but the most trivial arrays impossible to construct without the use of external interconnects. To address this issue, a multilayer hTron was developed and is covered in section 4.2. With the new hTron designed and tested, we had the confidence to proceed with DRO array design, which is covered in

section 4.3. The initial testing of these new cells and arrays is covered in section 4.4. After there were some issues with the initial devices, a revised design was developed and tested, which is covered in sections 4.5 and 4.6, respectively.

4.1 Operating principal of the DRO cell

The construction of the DRO cell is very similar to a NDRO cell, with the only major change being the yTron is replaced with a second hTron. With this modification, the schematic of the cell is that shown in figure 4-1. Many of the design rules presented in section 2.2 also hold here – in particular those pertaining to writing to the cell. There are, however, additional rules that govern the switching currents of both hTrons. These rules are derived and presented in the following sections.

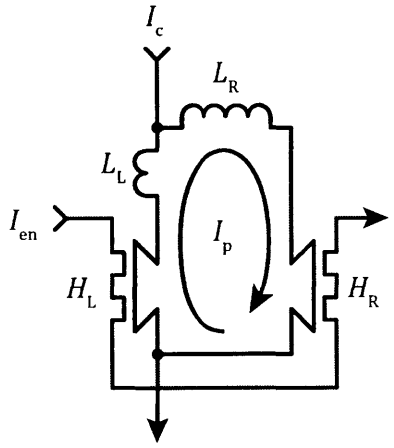

Figure 4-1: Schematic of the DRO cell. This cell design is similar to the NDRO cell design withe the only major difference being that the yTron has been replaced with a second hTron. The cell is selected by applying a current I_{en}, to the enable port. The cell is intended to be written to by a bipolar current applied to the channel I_c, which results in a persistent current I_p being induced in the loop. Readout is achieved by measuring the switching current of the channel by applying a high current to I_c, and monitoring the resultant voltage.

The operation of the DRO can be relatively complex. This complexity primarily arises from the many operating modes in which the cell can be used. The primary determining factor as to which mode the memory operates, is the number of times that a particular constriction switches during a write operation and during a read

operation. In the analysis performed in the following section, we will consider the most basic operating mode; however, other modes can be used. Some modes provide operation that, in some circumstances, may be more desirable than that shown here; for example, it is possible to have the readout be non-inverted, which is in contrast to the mode the cell is operated here.

4.1.1 Writing to the memory

Writing to the DRO cell requires a very similar process to writing to the NDRO cell. Unlike the NDRO cell, the DRO cell is intended to be written to with a bipolar bias. This means that the cell can be operated without the application of any heaters. However, operating the cell in this manner would prove particularly difficult when incorporated into an array. Thus, writing to the cell requires the application of two signals. One signal is the write enable I_{we}, which lowers the switching current of the hTron. The other signal is the write bias I_{W} which is applied to the channel I_{c}. Similar to the NDRO cell, when the write bias is applied to the cell it will inductively split between the left and right paths. The same considerations covered in section 2.2.1 apply here, so we again have that we need $L_{\mathrm{L}} < L_{\mathrm{R}}$ in order to write a large persistent current to the loop. Similarly, following from the NDRO cell design, we want the left constriction to switch during a write, but not the right constriction. This means that we need the selected cell switching currents to be $I_{\mathrm{c,H_R}}(I_{\mathrm{RE}}) > I_{\mathrm{c,H_L}}(I_{\mathrm{RE}})$, where I_{RE} is the write enable current. In order for the right constriction to not be switched during a write, and assuming a worst-case zero retrapping current, we need the right constriction to be able to carry the full write bias $I_{\mathrm{c,H_R}}(I_{\mathrm{RE}}) > I_{\mathrm{W}}$.

The DRO cells, when formed into an array, are stacked in a similar manner to the NDRO cells. Thus, the cell must be designed such that when the write bias is applied to a column in order to write to a selected cell, other cells in the column must not be effected. For this to be the case, we need the superimposition of the persistent current and the inductively divided bias current to be less than the unsuppressed hTron channel critical current. There are four cases we must consider in this analysis, corresponding to a "1" write and a "0" write which occur while the cell is in either

the "1" or "0" state. If we consider the case when $I_p = \pm I_P$, and $I_w = \pm I_W$. During a read, the current through the channel of H_R is $I_{H_R} = I_p + I_w L_L/(L_L + L_R)$. From equation 2.1, we find that the hTron channel current expression simplifies to

$$I_{H_R} = \pm I_W \frac{L_L \pm L_R}{L_L + L_R}. \tag{4.1}$$

The two extreme cases which could lead to an undesired switching of the hTron would be $I_{H_R} = \pm I_W$. Clearly, we do not need to be concerned about this condition since this is the current that we expect to be passing through the right hTron during a write procedure. Since this current can be carried by the hTron when the enable signal is asserted, then the hTron must be able to carry this current when the gate is not asserted. Performing a similar calculation for the left-hand hTron H_L, we have that $I_{H_L} = -I_p + I_w L_R/(L_L + L_R)$, and again using 2.1, we find

$$I_{H_L} = \pm I_W L_R \frac{1 \pm 1}{L_L + L_R}, \tag{4.2}$$

the extremes of which are $I_{H_L} = \pm 2 I_W L_R/(L_L + L_R)$. Thus, we need the critical current of the left hTron with no bias $I_{c,H_L}(0) > 2 I_W L_R/(L_L + L_R)$.

4.1.2 Reading from the memory

The mechanism by which the DRO cell is read out is the main distinguishing feature between the DRO cell and the NDRO cell. In order simplify the integration of the DRO cell into an array, it was necessary to remove the yTron port, and only use what was referred to as the write port, for both write and read operations. In order to achieve this, the DRO cell is read out in a similar manner to how a SQUID is used to measure a magnetic field. That is, the switching current of the entire cell is measured, ans used to determine the state of the cell.

Consider a cell that is retaining one bit of information encoded as a persistent current. If this current is circulating clockwise then the bit is in the "1" state and $I_p = I_P$. If the cell is in the "0" state then the persistent current circulates counterclockwise,

and $I_p = -I_P$. During a read, the read bias is applied to the memory channel $I_c = I_R$. The read bias is a positive current higher than the write bias, $I_R > |I_W| > 0$. Under these conditions, during a read the current in the channel of the left hTron is $I_{H_L} = -I_p + I_R L_R/(L_L + L_R)$, and the current in the right hTron channel is $I_{H_R} = I_p + I_R L_L/(L_L + L_R)$.

Since the loop is designed such that $I_{c,H_L} < I_{c,H_R}$, and from the above expression, we expect that the loop will only able to withstand a smaller channel bias (read current), when the cell is in the "0" state. We expect this operation since the left constriction, which can withstand less current, must carry both the persistent current and the read bias, which in the "0" state both flow in the same direction. Thus, the left constriction will switch, this leads to the read bias being diverted to the right constriction, which cannot withstand the entire read bias, and also switches. With both constrictions switched, a voltage will be present across the cell. For this to be the case we need the suppressed hTron switching current to be

$$-I_P + I_R \frac{L_R}{L_L + L_R} < I_{c,H_L}(I_{RE}) < I_P + I_R \frac{L_R}{L_L + L_R}, \tag{4.3}$$

where the read enable current I_{RE} is applied to the enable port, that is $I_{en} = I_{RE}$. Simultaneously we require that the right hTron can carry the read bias during a "1" read and will be switched during a "0" read. That is we require

$$I_P + I_R \frac{L_L}{L_L + L_R} < I_{c,H_R}(I_{RE}) < I_R. \tag{4.4}$$

While the mechanism behind the DRO read is very different from that used in the NDRO cell, the timing and process an external controller would use to read the cell is identical for the NDRO and DRO cells (with the exception of the signal levels).

4.1.3 Cell design limitations and trade-offs

The DRO cell shares many of the same design limitations and trade-offs as the NDRO cell, which was covered in section 2.2.3. As far as an isolated DRO cell is concerned, we

would like to operate the cell with the maximum allowable write bias, and minimum heater current, such that as much current is written to the cells as possible. We want a high persistent current because, the higher this current, the greater the read margins, and the less flux-quantization will come into play.

We wish to have as high a switching current as possible. We want these high currents since, the higher the switching current, the greater the write/read biases, and hence, the greater than signal-to-noise ratio. However, the larger the switching current, larger the cell size. It is possible to keep the footprint of the cell small, while increasing the switching current by utilizing a thicker NbN film. This method was used in section 4.5, where noise was proving to be an issue. However, thicker films make fabrication more difficult. The films can only be made so thick before etching issues occur, and in the case of the multilayer heaters, the step height is too high for a reliable connection to be formed (especially since evaporation is used to deposit the oxide and heater metal – see section 4.2). Additionally, the higher the switching current, the more energy is required to access the cell. Thus, there exists a trade-off between noise-immunity of the cell and the cell's energy efficiency, size, and the fabrication difficulty.

We are, of course, limited by the nature of the method by which the DRO is formed into an array. We must ensure that accessing the selected cell does not disrupt any other cell in the column. This imposes major limitations in the required hTron suppression ratios. The greater the suppression ratio required to access the cell, the worse the power efficiency, and the lower the switching current. Thus, this trade-off, interacts with the aforementioned switching current/noise trade-off. As a result of this interaction, in order to maintain the same noise-immunity, while also increasing the suppression ratio, the larger and/or thicker the cell must be. The simulations covered in section 4.1.4, required unselected cell's hTrons to have a switching current 2.5 times higher than a selected cell's hTrons.

In practice, experimental results have proven that we must also contend with other effects, in particular the variation in the hTron switching current. As covered in section 4.5.3, we believe that these variations are currently the main limiting factor

to the memory's performance.

4.1.4 Cell simulations

As it is difficult to predict the exact performance of a memory design with a given set of ratios, we decided to pursue cell simulations in order to aid in the design process. In particular, in order to create the cell layout, a number of simulations were performed until a reliable design emerged. The cell simulations were conducted in a somewhat similar manner to how they were conducted in section 2.3. In order for the LTspice simulations to reliably converge, two resistors were added to the loop, as was done for the NDRO cell simulations. The schematic of the cell as it was simulated in LTspice is shown in figure 4-2.

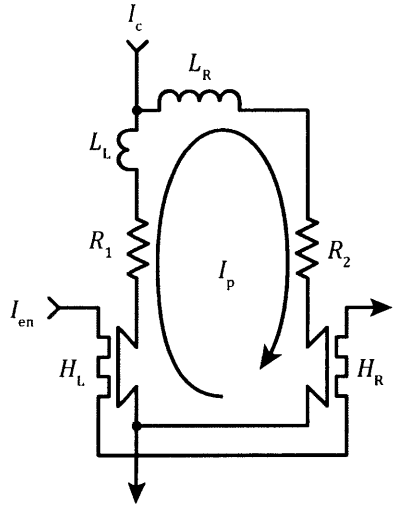

Figure 4-2: Schematic used in the LTspice simulation of the DRO cell. Note the inclusion of the two resistors $R_1 = R_2 = 1\,\mathrm{p}\Omega$ which are not physical but are included so that LTspice will reliably converge. This approximation is valid for small values of these resistors and for short time-scales.

The goal of these simulations was to find an inductance ratio and a switching current ratio that will yield a reliable design. Once we have these parameters, we can move onto the layout of the cell. In order to find suitable parameters, a number of simulations were performed, and the device ratios adjusted to obtain good performance, while also maintaining relatively small ratios. Small ra-

tios are desired, because the more similar the two switching currents and the more similar the two inductances, the smaller the cell layout can be. After many simulations, the parameters shown in table 4.1, were chosen. The exact values of these parameters are not important, rather the ratios are important. In particular, we have an inductance ratio of $L_{\mathrm{L}} : L_{\mathrm{R}} = 5 : 8$, the switching current ratio $I_{\mathrm{c,L}}(I_{\mathrm{EN}}) : L_{\mathrm{c,R}}(I_{\mathrm{EN}}) = 1 : 2$, where I_{EN} is the enable current, and finally a hTron suppression ratio of $I_{\mathrm{c,L}}(0) : L_{\mathrm{c,L}}(I_{\mathrm{EN}}) = I_{\mathrm{c,L}}(0) : L_{\mathrm{c,L}}(I_{\mathrm{EN}}) = 5 : 2$. The inductance ratio $L_{\mathrm{L}} : L_{\mathrm{R}}$ sets the ratio of the number of squares in the left and right branches, the switching current ratio $I_{\mathrm{c,L}}(I_{\mathrm{EN}}) : L_{\mathrm{c,R}}(I_{\mathrm{EN}})$ sets the ratio of the widths of the channels of the hTrons, and the suppression ratio $I_{\mathrm{c,L}}(0) : L_{\mathrm{c,L}}(I_{\mathrm{EN}}) = I_{\mathrm{c,L}}(0) : L_{\mathrm{c,L}}(I_{\mathrm{EN}})$ sets the minimum enable current I_{EN} that can be used to access the cell.

Table 4.1: Final simulation DRO cell device parameters.

Parameter	Value
Left inductance, L_{L}	1.0 nH
Right inductance, L_{R}	1.6 nH
Left hTron channel switching current (unselected), $I_{\mathrm{c,L}}(0)$	125 μA
Right hTron channel switching current (unselected), $I_{\mathrm{c,R}}(0)$	250 μA
Left hTron channel switching current (selected), $I_{\mathrm{c,L}}(I_{\mathrm{EN}})$	50 μA
Right hTron channel switching current (selected), $I_{\mathrm{c,R}}(I_{\mathrm{EN}})$	100 μA

The results of the final cell design simulation are shown in figure 4-8. This simulation was performed with the parameter values shown in table 4.1. This simulation shows the basic operation of the memory. First, the cell is written into the "1" state. This write is achieved by applying a bias $I_{\mathrm{c}} = 90\,\mu$A, and simultaneously asserting the read enable signal. It can be seen that as the bias increases, the left constriction switches, which results in the diversion of the majority of the write bias to the right constriction. This re-diversion of the write bias corresponds to setting the persistent current I_{p} to a high level. With the removal of the write bias, the persistent current remains trapped within the cell. A later read operation revels that the cell is in the "1" state. The read is performed by applying a read bias of $I_{\mathrm{c}} = 120\,\mu$A, and simultaneously asserting the read enable signal. The state of the cell is signified by the fact

that the memory does not switch, as a result the voltage remains at zero, $V_r = 0V$. After the read is complete, the persistent current can be seen to remain unchanged. Next the cell is written into the "0" state. This write is accomplished by applying the exact opposite of the bias used to set the cell in the "1" state. The application of the negative write bias can be seen to switch the left constriction again, this time leaving a negative I_p. Finally, a read operation is performed. This time the read reveals that the cell was in the "0" state by switching both sides of the constriction, thereby generating a high read voltage V_r. Thus, this simulation has shown that with the chosen parameters the cell can operate as expected.

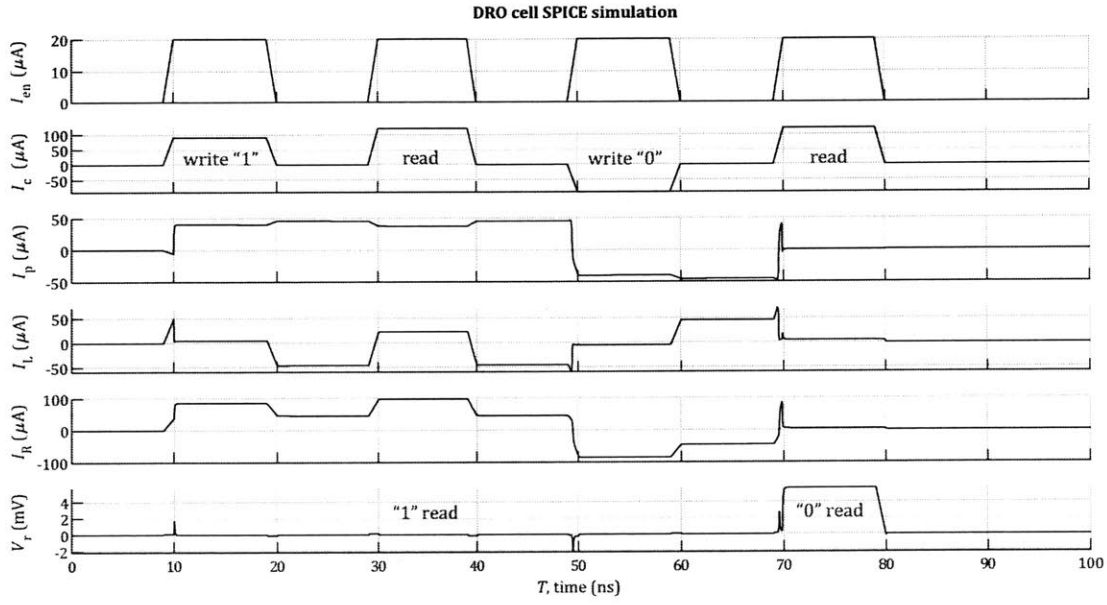

Figure 4-3: Results of a LTspice simulation of a single DRO cell operated in the pulse readout modus and using the write enable signal. In this simulation, the cell is set (placed in the "1" state) at 10 ns. At this time the, loop current I_p can be seen to increase – indicating that the write was successful. The cell is then read at 30 ns. During the read, the memory does not switch – as expected when the cell is set. The cell is cleared (placed in the "0" state) at 50 ns. Again, the loop current can be seen to respond accordingly – reducing to a negative value. Finally, at 70 ns the memory is read again, but this time the memory switches. The switching of the loop, and the corresponding production of a voltage V_r during a read "0", and the lack thereof for a read "1" indicate that, at least in this simulation, the memory is working as expected. In this figure, the values I_L and I_R, are the currents through the channels of the left and right hTron constructions, respectively.

4.1.5 Cell layout

With the results from the cell simulations showing that the ratios of the parameters from table 4.1 yield good performance, the cell layout was designed accordingly. The exact inductances and switching currents were not considered in the layout process, rather only their ratios were considered. The cell design begins with choosing the hTron channel constriction sizes. While this cell is the first to utilize the multilayer hTron, as covered by section 4.2, the type of hTron is not important to this design step. The smaller constriction was chosen to be 100 nm wide, as it is a small, but easily fabricated feature size. Thus, the size of the left constriction was set to 100 nm. Now, according to the simulation, the switching current of the right hTron should be double that of the left. On these small scales, the switching current scales very close to linearly, thus the width of the right hTron is chosen to be 200 nm. Thus, the gates of the two hTron are set, as can be seen in the layout, shown in figure 4-4.

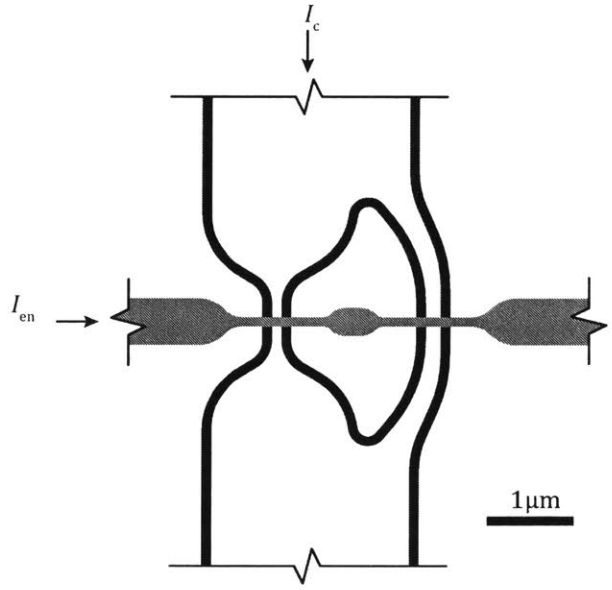

Figure 4-4: Layout of the first single DRO cell. The black area indicates where NbN has been etched away (leaving the bare substrate below), and the white area is where NbN remains. The gray area indicates the location where resist will be exposed and developed such that an oxide and metal can be evaporated and later lifted off to form the heater. The jog lines indicate leads that extent to the connection pads.

With the hTron channel widths set, we must now consider where to place them.

Since we are using a multilayer hTron process, we can simply place these two constrictions parallel to each other, and worry about the heater placement later. Now, the inductance ratio must be set. The design of a cell to meet a prescribed inductance ratio is not trivial. This difficulty arises from the fact that we must meet the required ratio and simultaneously avoid current crowding that could otherwise cause premature switching of some part of the loop. Thus, simulations of the cell must be performed to calculate the ratio, and ensure that excessive current crowding does not occur. The inductance design procedure is thus an iterative one. A trial layout is drawn, and then exported to COMSOL Multiphysics, where the current distribution is simulated. The ratio of the resistances of each branch, which for our device, corresponds to the ratio of kinetic inductances, is then calculated. The layout of the cell is then modified accordingly, until the desired ratio is achieved. The result of this process if the cell design shown in figure 4-4.

Finally, the heater is placed onto the cell. The heater is to be fabricated by means of a lift-off process, as covered in the following section. The selected resists being of a positive tone means that the area that we write will be the area where material remains after lift-off. Thus, in the layout we draw the area where we wish for the heater to be located. The width of the heater was somewhat arbitrarily chosen to be 100 nm, which it is small enough to comfortably fit on out hTron gates, and large enough that fabrication is not overly complicated. The heater width is made wider away from the locations where we intend for heating to occur. Thus, less heating of the substrate will occur, and less power will be wasted. However, some tolerance is left such that small misalignment of the heater will not result in an unusable device. The transitions in the heater width are made smooth to again avoid current crowding, while current crowding is almost a non-issue for the heater, the design was made this way such that we can be confident that we are targeting our heating appropriately. Finally, the connections from the cell are extended to the contact pads where wirebonds will be attached during the experiments.

4.2 Multilayer hTron

Due to the lack of a via process, it would be impractical to build the DRO cell with in-plane heaters like those that were used in the NDRO cell, at least not without the use of external interconnects. Further, it would be impossible to design anything but the most trivial DRO cell array using in-plane heaters. Thus, the design presented in section 4.1.5 utilized a multilayer hTron. This section covers the basic design and testing of the multilayer hTron.

4.2.1 Design

The design for the multilayer hTron consists of a superconducting channel below an oxide and normal metal layer. The channel is located on the bottom layer of the stack (on top of gold alignment marks and pads), and is fabricated at the same time as the rest of the memory cell. Thus, no connections between superconductors need to be made. Avoiding superconductor-to-superconductor connections is desirable since connecting two superconducting layers can be difficult as oxidization and poor contact between the two depositions can leads to undesirable JJ-like structures, and weak links. This bottom superconducting layer is fabricated with the same process used for the NDRO cells. That is, the film is deposited by means of reactive sputters, a positive tone resist is applied, the film exposed by means of electron-beam lithography, the resist developed, then etched by means of reactive-ion etching, before the resist is stripped. The use of a positive tone resit means that when laying out the channel (and the rest of the structures on the NbN layer), we only pattern regions where the film is to be cut (that is where the NbN is to be removed).

With the channel fabricated, the next step is to add the dielectric spacer and heater. The spacer and heater are fabricated by means of a bilayer lift-off process. In this process, two positive tone resists are applied to the sample, then the sample is patterned by means of a second electron-beam lithography step. The resists are then developed. Next, approximately 20 nm of SiO_2 is evaporated onto the sample, followed by approximately 20 nm of Ti. Next, the Ti and SiO_2 is lifted-off leaving

behind the patterned dielectric spacer and heater. The use of positive tone resist for this process means that when designing the layout, we simply draw the region where we want the heater to be located. Thus, we draw the heater in the opposite manner to which we draw the features on the NbN layer.

Now that we know how to draw the features, we shift our focus to the hTron design. The channel can be designed much the way that it was for the in-plane design, however it does not need to be. We have much more freedom now, as we can place the heater in almost any location. Since the heater is a normal metal, wherever the heater is located, there will be some amount of heating (when it passes a current). Thus, it is advisable to avoid placing the connections to the heaters on top of areas were suppression is undesired. If such placement is unavoidable, then the normal metal connections should be as wide as possible, so as to reduce the resistance, and hence the heating, and the normal connections should be routed over regions of the superconductor where current density is low. In general, locations where heating is not intended, the connections to the heater should be as wide as possible, thereby reducing the resistance and hence the power consumption.

The geometry of the multilayer hTrons used here consists of a fixed heater width of 100 nm. The channel width is chosen based on the required switching current. The channel constriction is typically narrower than the connections to it. Care should be taken such that current crowding does not occur within the channel, as this will lower the switching current to below its intended value. The heater is placed orthogonal to the channel such that it overlaps for the smallest distance possible. The heater is arranged this way for a number of reasons. First, we want to minimize power dissipated, so we want the shortest heater possible. Second, we are attempting to suppress superconductivity of the channel, and not a large portion of the wire, so an orthogonal heater will take up the least space and produce the most confined hotspot. Finally, an orthogonal placement of the heater is well suited to the DRO cell layout. Outside of the DRO application, there are many reasons one may want the heater to overlap large portions of a superconductor, for example to build an hTron that when in the normal state, presents a very large resistance.

4.2.2 Experimental results

Initial hTron results suggested issues in fabrication. In particular, the normal metal gates presented inconsistent resistances at room temperature. Some devices showed a resistance as low as 91 kΩ, while others presented resistance over 6 MΩ, and others were totally open. When the device was cooled to 4.2 K, the resistances tended to increase substantially. Typical cold resistances were in excess of 13 MΩ. This issue was suspected to be the result of poor contact between the normal metal heaters and the superconductor pads. In later fabrication runs, the variance in heater resistances were substantially lower.

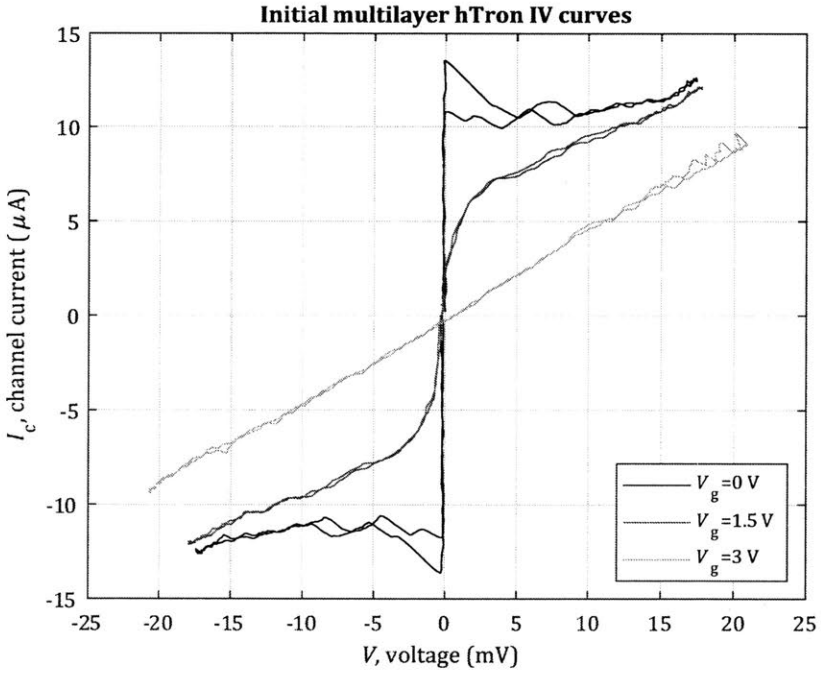

Figure 4-5: IV curve of one of the first working hTron device conducted at three gate voltage biases. The experimental data was moving-average filtered prior to being decimated by a factor of 25 in preparation for this figure. It can be seen in the $V_g = 0$ V curve that the switching and retrapping currents are very close together, within around 20%. Typically, the retrapping current is around 20% of the switching current. Thus, this figure here suggests that, even without the application of a current to the gate, the channel is suppressed substantially. With increasing gate bias voltage, an increase in the suppression is seen. With ultimately at a gate bias of $V_g = 3$ V the channel appearing totally resistive.

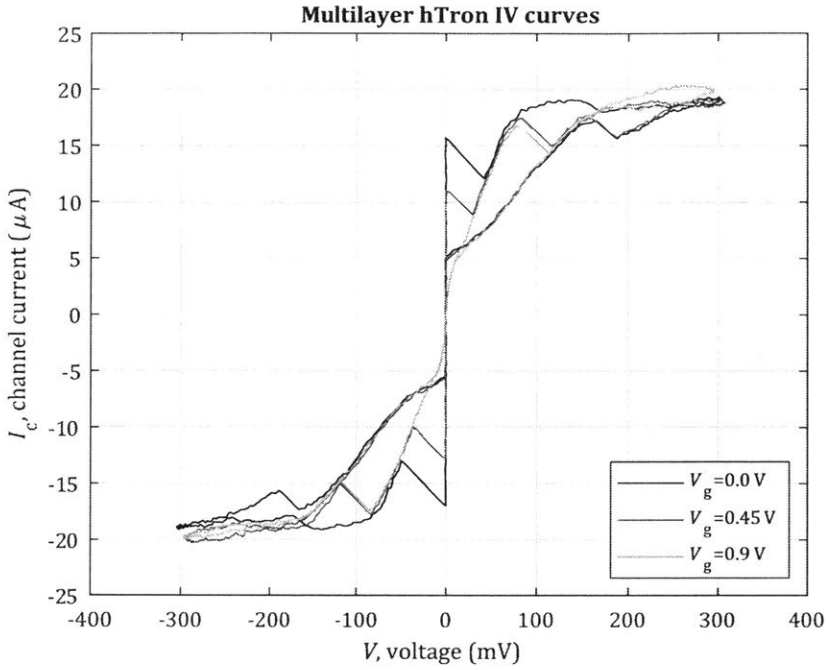

Figure 4-6: IV curve of a multilayer hTron with the application of three different gate voltage biases. The experimental data was moving-average filtered prior to being decimated by a factor of 25 in preparation for this figure. It can be seen in the $V_g = 0\,\mathrm{V}$ curve that the switching and retrapping currents are quite distinct from each other and are in line with those obtained in previous experiments on contributions. With the application of a gate bias of $V_g = 0.45\,\mathrm{V}$ the switching current of the device can be seen to reduce – while the retrapping behavior is unchanged. With the increased application of gate bias, the switching current is suppressed further until the device becomes non-hysteretic, and the switching and retrapping current merge.

Basic IV curves were measured for one of the few devices with working heaters, the results are shown in figure 4-5. These experiments were conducted in LHe, and using a the setup from figure 3-13, modified for a single hTron. The device that was tested had an NbN film thickness around 5 nm, and channel width of 200 nm; as a result, we would expect a switching current around $40\,\mu$A. We can see in figure 4-5 that the switching current (with no heat applied) is much less than this at around $13.5\,\mu$A. There are two possible causes, first the device is not the geometry that we expected, or superconductivity is being suppressed by some other influence. We can determine that that latter is more likely the case since the retrapping current can be

seen to be much higher than would otherwise be expected. In previous experiments, the retrapping current of a constriction is found to be around 20% of the switching current. In this experiment, the retrapping current is around 80% of the switching current. From this result, it is possible that the NbN film was exposed to something that has resulted in its switching current becoming much lower than otherwise would be expected.

The fabrication process was revised, and a new chip was fabricated. The new devices were measured and showed superior results, in comparison to the prior chip. The IV curves for this device are shown in 4-6. The device tested in this experiment had a constriction 100 nm wide. It can be seen that the switching current of this device is close to, while still being less than, the theoretical 20 μA switching current. This switching current is lower than that typically obtained in previous device experiments which did not have a multilayer heater. It is suspected that either the presence of the heater, or more likely, the extra fabrication steps required to fabricate the heater, may damage the film to some extent, thus resulting in the lower switching current. Regardless, the device is operational now, and the retrapping current can be seen to be close to its expected position. Additionally, it can be seen that with the gate is now more sensitive in comparison to the first device. Thus, we have the confidence to incorporate this device into the DRO cell.

4.3 DRO array design

The DRO cell was designed such that it is trivial to form a collection of devices into an array. This ease of arraying the cells is in contrast to the NDRO cell which, while simple to design, required extensive external components to form it into an array. Arrays from the DRO cells are not only easier to build (once an appropriate cell design is selected), but smaller and more power efficient than their NDRO counterparts.

4.3.1 Array design

The DRO array is formed by simply tiling the cells horizontally to form a word, and stacking these words to form a bank. There is no intrinsic limit on the size of a word, or the number of rows in a bank. The array design is only limited by the capability of the row and column drivers. With normal metal heaters, the word size is only limited by the maximum impedance the row-drivers can source current to. The number of rows in a bank is limited by the maximum inductance the column driver can source the write and read currents to within the required time period. In the designs we present here, the limitations on the drivers are ignored as we are using room-temperature drivers capable of driving arrays far larger than those that we are considering.

The nature of the DRO cell is that the contents of the cell is destroyed when read. Thus, in a final implementation, it is likely that an external controller would, after performing a read, write back to the array the contents it read. This operation is not necessarily required, and it highly influenced by the design of the processor that the DRO array is incorporated into. Here, we are not concerned with the design of the memory controller or processor, so we will not take these considerations any further.

Like the NDRO array designs, the DRO array can be operated in either a bit-access or a word-access scheme – or for that matter almost any combination in between these two schemes. To access a single cell, the write or read bias is applied to the corresponding column, and the enable signal for that row asserted. Cells in that row who do not have a bias applied to their column will be unaffected, while the selected cell will be accessed. Extending this operation, we can access a word at a time by simply applying the appropriate bias to each column, and applying the enable bias to the row. It follows that we can choose to access only part of a word; for example, accessing the upper 8-bits in the word while leaving the lower bits unaffected.

This array design also facilitates the ability to set or clear portions or the entire bank. For example, the first column of the bank can be cleared by applying the write "0" bias to the column and asserting every row enable signal. This would be

particularly useful to clear an entire bank, or for example, the first n addresses within the bank. Additionally, it is possible to perform unusual memory operations such as logical OR and NANDs between rows of the memory. In order to achieve this, two (or more) rows are selected, then the read bias is applied to the selected columns of the memory. The voltage at the top of the column will thus be $V_c = I_R R_r N_{c,0}$, where I_R is the read bias, R_r is the resistance of the cell when switched (when reading a zero), and $N_{c,0}$ is the number of cells in the zero state within the selected rows of the selected column. In order to perform a logical OR between cells in the column, the resultant voltage is sent to a comparator that declare the local result is

$$A + B = \begin{cases} 0 & \text{if } V_c < \frac{3}{2} I_R R_r \\ 1 & \text{otherwise} \end{cases} . \tag{4.5}$$

A NAND operation is performed in the exact same manner, except that if any voltage is seen then the NAND result is "1", otherwise it is "0". This occurs since if either of the two rows is a zero then an output voltage will appear. An extension of this can be made to more rows than two. For example, the number of zeros (or equivalently the number of ones) in a segment of a column (or entire bank) can be counted by examining the voltage seen when many row enable lines are activated. The only major drawback of these operations is that the state of each cell is lost when the operation is performed.

The layout for the first DRO array comprised of four DRO cells arranged into a 2×2 configuration. The cell design presented in section 4.1.5 was used, and the layout for the array formed using this cell is shown in figure 4-7. The columns are formed by simply connecting the bottom port of the upper cells to the top port of the lower cells. This is easily done since the cells were designed such that their top and bottom ports are equal in width and in-line. The heaters were similarly connected across the rows; however, the width of these traces was changed along the path. The widths change so as to reduce the resistance of these lines, and hence, the power dissipation of the read as well as the heating of the substrate. The cells can be seen to be placed very far from each other. This spacing was added since we were uncertain if there

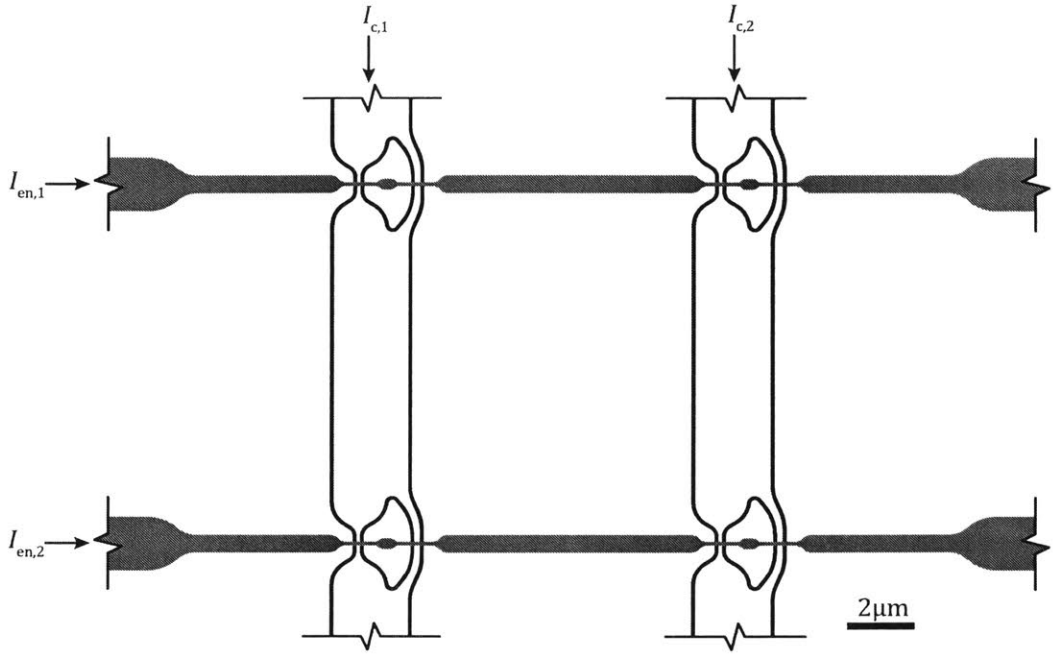

Figure 4-7: Layout for the first DRO array. This array features four DRO cells of the same design as that presented in section 4.1.5, which have been arranged into a 2×2 configuration. The word size for this memory is 2 b, and there are two rows in the bank. It can be seen that the cell design shown in figure 4-4 was simply repeated four times with the terminals connected to form the desired array. A large space was placed between the cells to ensure that in initial tests there was no inter-cell interference – later tests showed that this space is unnecessary, and that cell heating is highly localized. On the extremes of the heater lines, the connections to the pads were made wide so as to reduce the resistance, and hence power dissipation, in the interconnects. Additionally, the connections from the heaters to the pads were made to be equal in length, so that their resistances would be equal, and as a result the biases applied to the heaters would be the identical for each row.

would be some inter-cell thermal interference. Later experiments have shown that this does not occur, and that the cells can be packed very close together without any interference.

4.3.2 Array simulations

From the cell simulations performed in section 4.1.4, we are confident that the cell design is sound. With the array design presented above, there is a need for the cell to be able to carry the write and read currents of other cells in the column without

its contents being destroyed. To verify that this can, in fact, occur without altering the cell's contents, SPICE simulations of the array were performed.

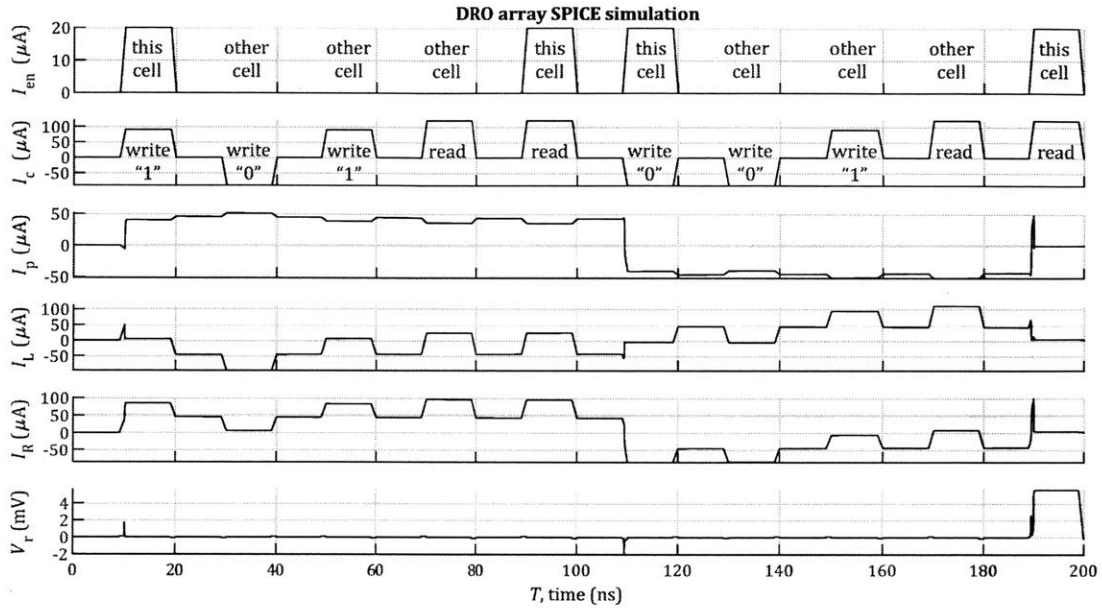

Figure 4-8: Simulation of a DRO cell within an array. This simulation demonstrates that the state of the cell is unaffected by operations accessing other cells in the array. First, the cell in question is written into the "1" state. Then other cells in the array written to the "0" state, written to the "1" state, and read out. None of the operations performed on the other cells in the array caused the cell in question state's to change, as witnessed by the persistent current I_p remaining unchanged, and the read operation indicating that the cell is in the "1" state. The same test is then performed again with the cell instead being written into the "0" sate. Again, the state of the cell was unaffected by accessed to other cells in the array, and the subsequent read provided the correct result.

The simulations performed in section 4.1.4, were modified for a column with two rows. The results are shown in figure 4-8. Multiple columns cannot interfere with each other since there is virtually no feedback through the heaters, which are the only inter-column connections. While simulation of large arrays with many rows and columns were performed, the results are not included here as they do not convey any more information than the two row simulation results shown in figure 4-8.

As can be seen in figure 4-8, in the case of two stacked cells, no interference between the two cells was observed when using the cell design shown in table 4.1. In

modifying the simulations, it was found that during a read it was possible to set every cell in the column if the suppression ratio $I_{c,L}(0)/I_{c,L}(I_{EN}) = I_{c,R}(0)/I_{c,R}(I_{EN})$ was too low. In the experiments conducted here, a suppression ratio of $5:2$ was found to perform well. The results of these simulations indicate that this array design is sound, and that we can proceed to fabrication of a device for experiential testing.

4.4 Testing the initial DRO array design

With the initial cell design verified, devices were fabricated and tested. The following sections describe initial testing and debugging of this device.

4.4.1 Initial cell experiments

The initial experiments with this cell were conducted in a somewhat similar manner as the experiments presented in chapter 2. The major difference being that the hardware for this experimental is substantially simpler than that of the NDRO cells, as shown in figure 4-9.

It was very difficult to find an operating regime where the error rate was better than random. A number of different operating modes of the memory were tested. These included DC bias to the heaters, ramp-based readout, pulse-based readout, different pulse timing and sequences, as well as ramp rates. From these experiments it was found that applying the channel bias and then pulsing the heater gave the best, albeit still relatively poor, results. With this device, the best operation was found when the read pulse width was relatively long and the heater pulse very short. In this experiment, a read and write pulse width of 50 ns with the heater pulse width being only 8 ns. An example of the waveforms captured during this experiment are shown in figure 4-10.

A similar write scheme to the one used in the simulations, as shown in figure 4-8, was used again here. The only major difference between the scheme used in simulations and that used here, is that in the simulations the write "0" and write "1" levels were of the same magnitude but opposite sign. Whereas in the experiment, it

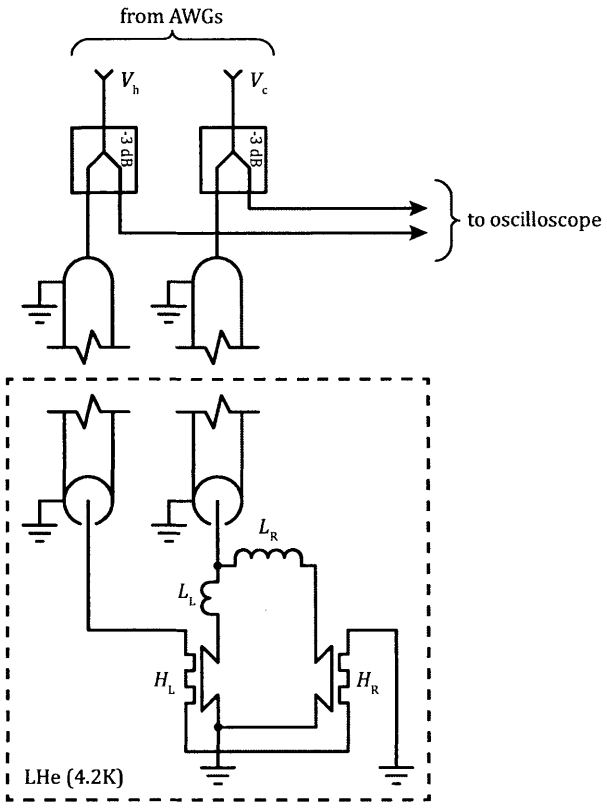

Figure 4-9: Experimental setup for the DRO single-cell measurements. This setup can be used for ramp and pulse-based readouts; however, here we focus on the pulse-based readout scheme. Since the experiments utilize relatively fast pulses, a system impedance of 50 Ω was used for the setup including the AWG outputs and oscilloscope input. In order to monitor the cell voltage, a splitter was used to divide the cell bias between the device and the oscilloscope. The downside of this approach is that the oscilloscope sees the cell's voltage response superimposed on the cell's bias. Additionally, the heater signal was split between the device and the oscilloscope. This allowed for the tuning of the timing of the heater with the timing of the write/read pulses.

was found that a larger magnitude of write "0" pulse performed better.

To gain an understanding of the practical error rate, an experiment that involved 10,000 write/read cycles was performed. In this experiment, the measurements shown in figure 4-10 were performed and the levels at the two markers monitored. In this way, we effectively set the memory, read, clear, read, and repeat. Thus, at every step the state of the memory is the opposite of that in the prior measurement. The results

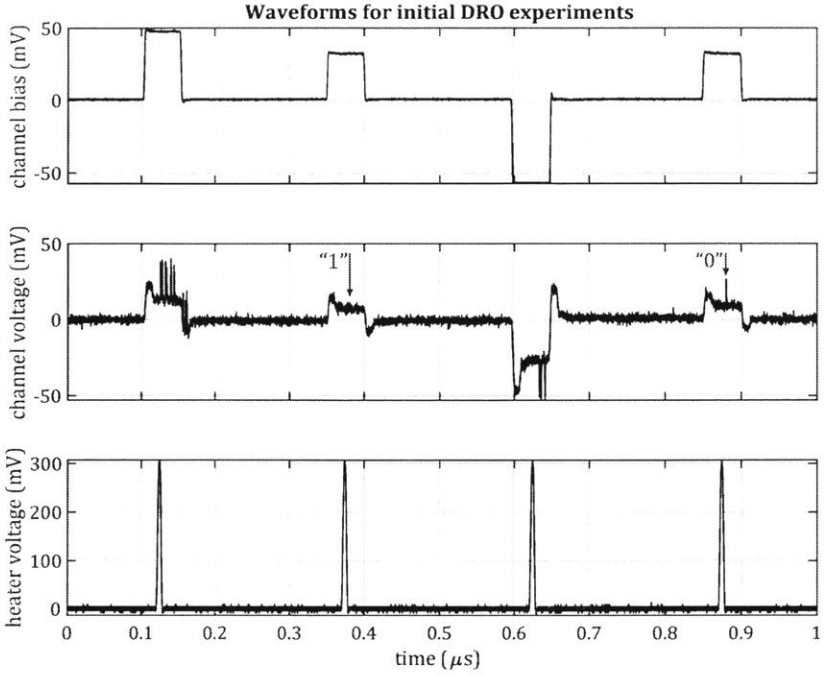

Figure 4-10: Oscilloscope traces of the signals applied to, and read from, the DRO cell. The channel bias is the signal that was generated by the AWG. Channel voltage is the combination of the signal applied to the cell and the reflections from the cell – see figure 4-9. Finally, the heater voltage trace is the signal applied to the heater. The markers on the channel voltage plot show where the voltage is sampled to determine if the device switched. The first pulse sets the cell into the "1" state. Following a short pause, the state of the cell is read out. The voltage is sampled at the first arrow. This voltage is below some threshold voltage V_{th}. After another pause, the cell is then cleared to the "0" state. Again, the cell is read out using the identical read pulse to that used in the first read. This second read resulted in the cell switching, and the resultant voltage pulse is above our threshold voltage V_{th}, thus indicating a "0" read. The threshold voltage is V_{th} determined experimentally, and chosen to give the best error rate.

of these experiments are shown in figure 4-11.

In figure 4-11, a threshold voltage of $V_{\text{th}} = 5\,\text{mV}$ was found to give the best error rate. With this threshold, it can be seen that the error rate is extremely poor – with 33.6% of reads yielding an incorrect result. From these results, it appears that positive write/reads (write cell to "1" state then read) yield the correct answer more often than negative write/reads (clear cell to "0" state then read). In fact the negative write/read gives the correct answer roughly half the time. From these results, it is

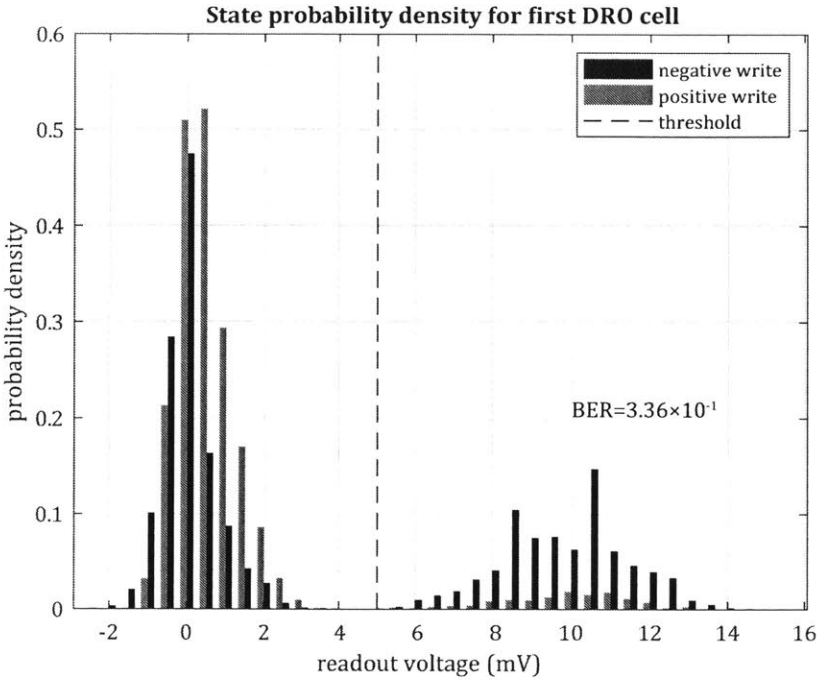

Figure 4-11: Probability density of the read pulse amplitude for reads that occurred after a "0" was written (negative write) and after a "1" was written (positive write). The horizontal axis in this plot corresponds to the amplitude of the two markers shown in figure 4-10, and should not be confused with the switching current. This figure is composed of $10,000$ write/read cycles, each of which alternated between writing a "1" and writing a "0". The results are divided into two by a threshold voltage V_{th}. Any sample whose read voltage was below V_{th} is determined to be a read that resulted in a "1", and those above the threshold a read that resulted in a "1". The threshold was chosen to be $V_{\text{th}} = 5\,\text{mV}$ since this gave the lowest error rate, which was 33.6%.

impossible to determine if the errors are due to the write or read procedure. This indistinguishability is a result of the fact that varying the read current will bias the results to switch prematurely, or never at all, which is the same effect as using the incorrect write biases. These results being better than an error rate of 50%, do show that the memory is functioning, albeit poorly, and that there may be a path to achieving error rates similar to those obtained for the NDRO cell.

4.4.2 Debugging design – magnetic modulation of cell switching current

With the operation of the memory showing extremely poor results, there were doubts about if the cell can be read out by simply measuring the entire cell's switching current. In order to prove that the memory can, in fact, be read out in this manner, a new experiment was devised. This experiment would involve inducing a circulating current into the memory cell by means of an externally applied magnetic field. Since the memory loop is superconducting, it will attempt to produce a current to cancel the applied field [33]. In this way we can be guaranteed that a current will be circulating, and do not need to worry about the intricacies of the write process. If with varying magnetic fields we see a variation in the loop's switching current then we can be sure that the readout mechanism is sound, if we see no modulation of the switching current, then there is some issue with the operation of the readout mechanism.

The experiment described above is, in essence, operating the memory loop as a SQUID [33]. In contrast to a traditional squid, which uses JJs, we are using the constrictions, located on either side of the memory loop, as the weak-link. As a result of using nanowires, we expect that the modulated waveform should not look sinusoidal – as it does for a SQUID constructed with JJs.

Since the memory is constructed from NbN, which has a very high kinetic inductance, we expect that coupling from a magnet to the loop will be very poor. To compound this, the area of the loop is extremely small at only a few square microns. As a result of these two effects, we require a very strong magnet to achieve modulation. A superconducting magnet was constructed specifically for this measurement. It was built from a 1" diameter plastic cylinder which had approximately 500 turns of a superconducting wire wrapped around it. This was mounted directly onto the PCB which contained the memory chip.

The experimental setup shown in figure 4-12 was used for these experiments. The magnet was attached to copper wires within the dewar, these wires lead to the room temperature current source. The memory switching current measurement is

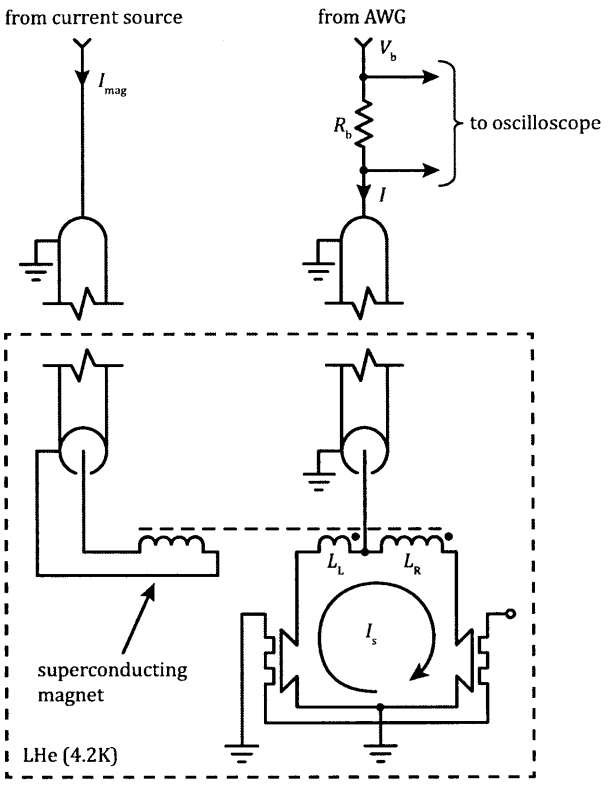

Figure 4-12: Setup used to test the magnetic modulation of the DRO cell's switching current. In this setup, the memory chip was placed in close proximity to a custom-made superconducting magnet. The leads of the magnet were attached to copper wires within the LHe. The ground for the magnet was kept separate from the ground for the chip. This separation was made since the current through the magnet very high, and would result in voltage drops along the cables. The presence of these voltages could interfere with the measurement of the switching current of the memory. The magnetic field couples to the loop, and induces a screening current I_s. It is this screening current that we are attempting to measure by performing switching current measurements on the memory loop. The hTrons were not used in this experiment, so one side of the gate was grounded and the other left floating. The same setup used in previous IV curve measurements was again used here with a bias resistor of $R_b = 10\,k\Omega$.

performed using the standard IV curve procedure used in previous measurements. The current to the magnet was swept from 0 A to 1 A in increments of 3.3 mA. At each current bias 40 switching current measurements were performed, the median of these values is shown in figure 4-13. The results were extremely noisy, as a result, extra filtering was applied in order to better show the structure of the results, specifically

a moving average filter of length five was used.

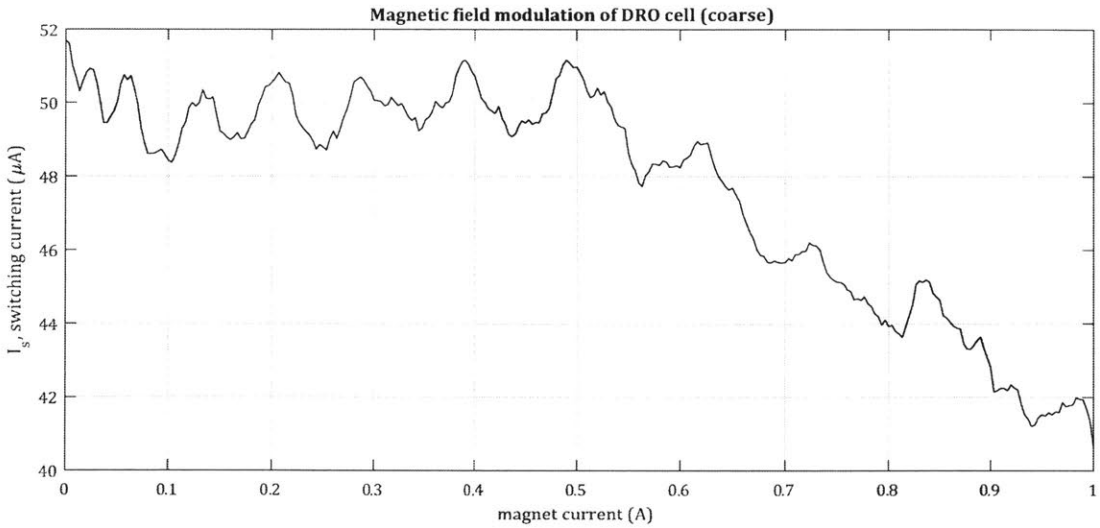

Figure 4-13: Modulation of the memory's switching current by the application of an external magnetic field. The magnet current was incremented in steps of 3.33 mA. Two modulation effects can be seen. The first being a periodic modulation, consistent with the existence of a screening current, as expected in SQUID measurements. The second being a suppression of the switching current with the increased application of magnetic field. Each point in this plot corresponds to the median of 40 switching current measurements. A moving-average filter was applied with a length of five samples.

It can be seen that the switching current is modulated by the applied magnetic field. There are two main trends in the data, a periodic oscillation in the switching current as a function of the magnetic field applied, and a suppression of the switching current beyond a magnet current of 0.5 A. The suppression effect is speculated to be a result of the magnetic field generated by the magnet being so high as to begin suppressing the NbN's critical current. The periodic modulation is our primary interest. Since the details of the modulation are difficult to observe in figure 4-13, a second finer sweep was performed, as shown in figure 4-14.

The fine sweep was conducted a similar manner as the coarse sweep, except this time the magnet current was swept from -0.1 A to 0.1 A in steps of 0.5 mA. The result of this sweep is shown in figure 4-14. It can be seen that the modulation depth is relatively shallow at between 6% and 8%. It is very apparent that there is

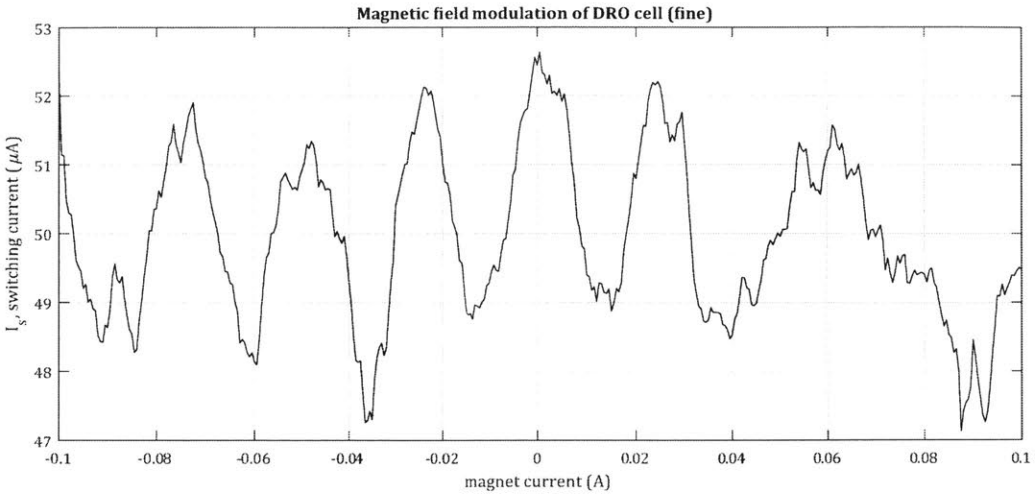

Figure 4-14: Fine sweep of the magnetic modulation of the memory's switching current. The magnet current was incremented in steps of 0.5 mA. In this result, only the periodic modulation of the switching current can be seen. This indicates that the memory readout mechanism is functional. Each point in this plot corresponds to the median of 40 switching current measurements. A moving-average filter was applied with a length of five samples.

substantial noise corrupting the data. The exact source of this noise is unclear; it is possibly from the magnet's current source.

The fact that any modulation can be seen, reveals that the readout scheme should be operational. Thus, the fact that we cannot achieve reasonable memory operation must be caused by the write operation failing, the design of the cell, or noise. It is expected that the write operation is successful; this is because some NDRO cell experiments were conducted using a write operation that did not activate the heater, and these experiments were successful. Such an arrangement is identical to how we are writing to the cell here. Thus, it is likely that either the design of the cell, or some source of noise must be causing these issues.

4.5 Design revision

As the experimental results thus far have yielded poor error rates, the design of the memory was revised. From the results of the cell debugging, there are clearly two

main issues. The first being noise, and the second being the separation between the readout levels. In order to improve the new cell design to combat these issues, two major changes were made. First, the NbN film mas made roughly four times thicker (to around 20 nm). Increasing the film thickness means that we increase the switching current without making the structures larger, and as a result will improve the immunity to noise. The second change is to increase the inductance ratio. This was done in an attempt to create a larger difference between the switching current of the "0" and "1" states. It was hoped that with these changes the noise immunity would be increased, and as a result, the error rate would decrease.

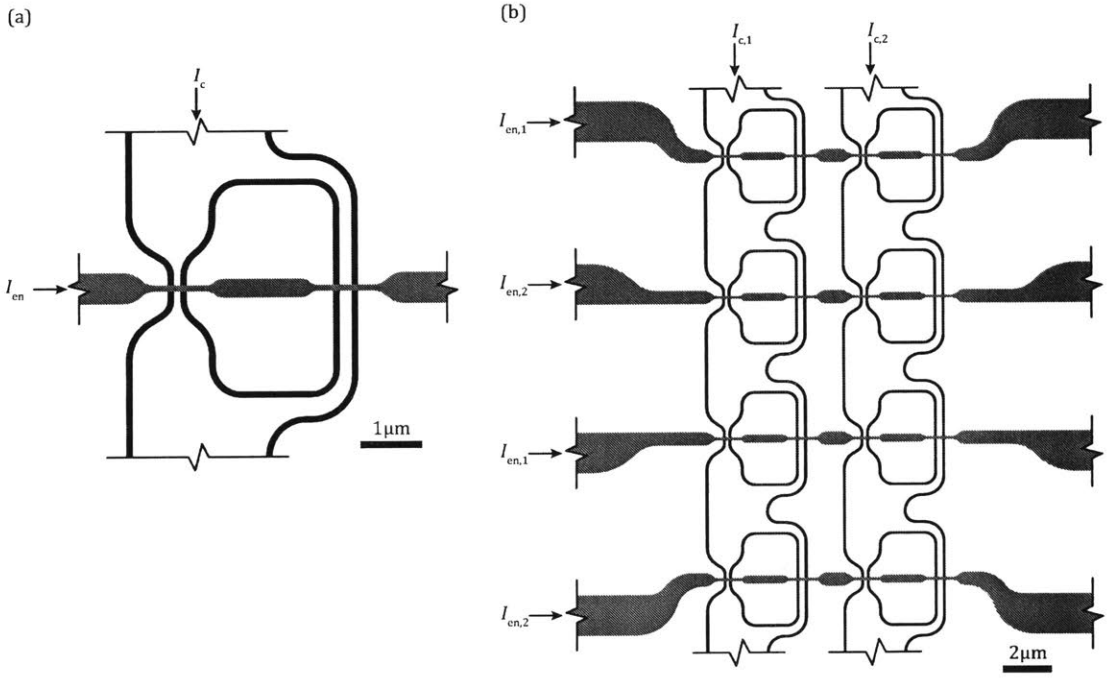

Figure 4-15: Layout for the revised cell (a) and an array composed of eight of the new cells (b). The inductance ratio for this cell is based on that shown in figure 4-4, with the inductance ratio changed to $L_{\mathrm{L}} : L_{\mathrm{R}} = 4 : 9$. The array can be seen to be based on that shown in figure 4-7.

The new cell and array layout is shown in figure 4-15. The cell was designed to have an inductance ratio of $L_{\mathrm{L}} : L_{\mathrm{R}} = 4 : 9$ while also maintaining the original switching current ratio of $I_{\mathrm{c,L}} : L_{\mathrm{c,R}} = 1 : 2$. This new cell design can be seen to be a reshaped version of the original cell design from figure 4-4. An array was also built

with this new cell design, this time a word width of two, and four rows per bank.

With the new device fabricated, testing proceeded in a similar manner to that used for the previous cells. In order to increase the speed at which measurements could be taken, new experimental procedures were developed. The array specific testing procedures are covered in the following section, and the details of the new experimental setup are covered in section 5.1.

4.5.1 Automated array testing

When we started testing the revised design, we proceeded with the methods used previously. These methods yielded somewhat better results, but it was difficult to find the optimal operating point. One might naively consider sweeping all the operating parameters; however, such a sweep would take weeks to complete. As these measurements are conducted in a dewar, we are limited in how long a measurement can last. Thus, we decided to build an automated test setup that would find the optimal operating point automatically. This setup is essentially an optimization algorithm which has been applied to BERTs in order to minimize the error rate by varying the operating parameters. The details of this setup are covered in section 5.1.

With the optimizer running the AWGs and oscilloscope, the experimental setup was extremely simple, as shown in figure 4-16. Since we are limited to two AWG channels, we can only test one cell at a time. In order to change between cells we need to disconnect the cables and reconnect them to the desired port. In this experiment the optimizer was allowed to vary three parameters, specifically, the write level, read level, and the heater level. The same heater level is used for writing and reading from the cell. The timings were set to be the same as those used previously – see figure 4-10.

The optimizer needs to be provided with initial starting point. To find a suitable starting point, the experiment was setup and the operating parameters varied manually until an error rate better than 45% was found. The optimizer was then run with a cost function based exclusively on the BER result of a 200 trial BERT. After running for around ten minutes, the optimizer was restarted. This new run utilized

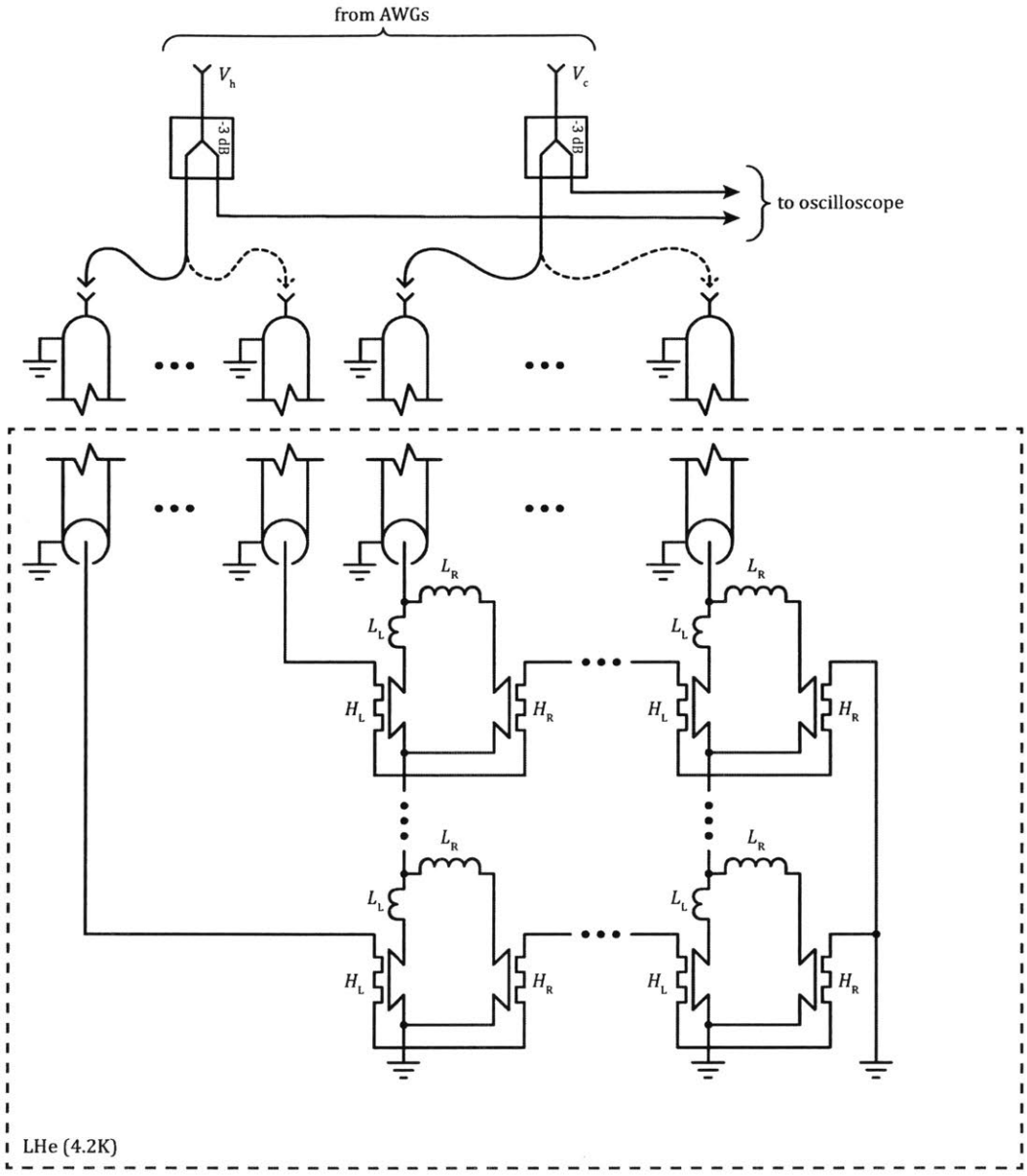

Figure 4-16: Experimental setup for array measurements. This setup is an extension of the single-cell experimental setup, shown in figure 4-9, to an array. Due to the limited number of AWG channels available, we can only test one cell at any one time. Due to this limitation, in order to change the cell under test, the cables from the AWG/oscilloscope disconnected and reconnected as required.

the optimum found from the last run as its starting point, and an increased number of trials of 500 per BERT. After one hour, the best BER obtained was 16.67%. The

progress of the final run of the optimizer is shown in figures 4-17, and 4-18.

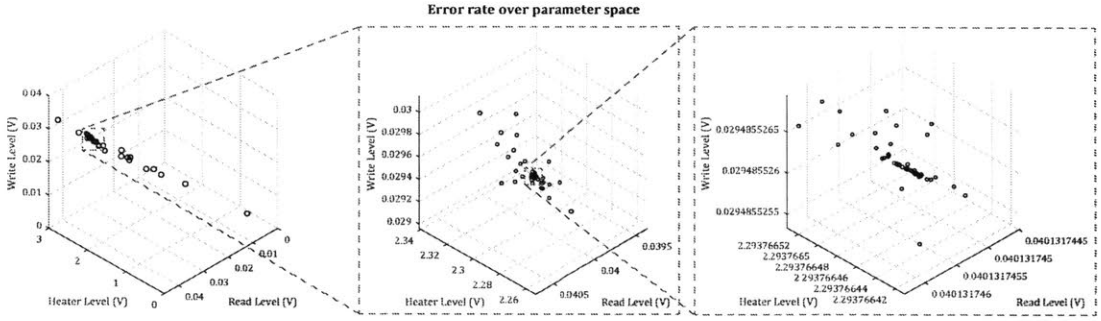

Figure 4-17: Plot of the error rate over all three operating parameters. These results were generated during the progression of the optimizer. The size of the circle represents the error rate, the larger the circle the higher the error rate. The leftmost figure represents all points the optimizer explored, the center figure is a smaller selection of these points, and the rightmost is an even smaller selection. Throughout the progress of the optimizer, it can be seen that as it approaches the optimal point it explores a progressively smaller operating parameter space.

Throughout the optimization process a total of 851 operating points were tested. The error rate at each one of these points was assessed, and is shown in figures 4-17, and 4-18. The points that were tested can be seen to be spread over a large area at the start of the optimization, as shown in figure 4-17. As the optimizer progressed, the points became more concentrated around the optimal operating point. Finally, it can be seen that the error rate is relatively constant around a very small area with only a few points performing better. From 4-18 we can see that, after the initial search, a relatively good operating point was found, and from around 100 BERTs on the error rate does not improve substantially. This distribution of error rates suggests two possible situations. First, there is a very small point around which the memory performs best. Second, the results from figure 4-18 are simply a result of the finite length of the BERT, and the fact that the errors are a random process. The second situation is likely the case since figure 4-17 shows that within a very small space the error rate does vary. This experiment, within the span of around two hours, gave us much more information than we could have obtained manually, and has shown that it is unlikely that there exists an operating point at which this memory will operate with a low error rate.

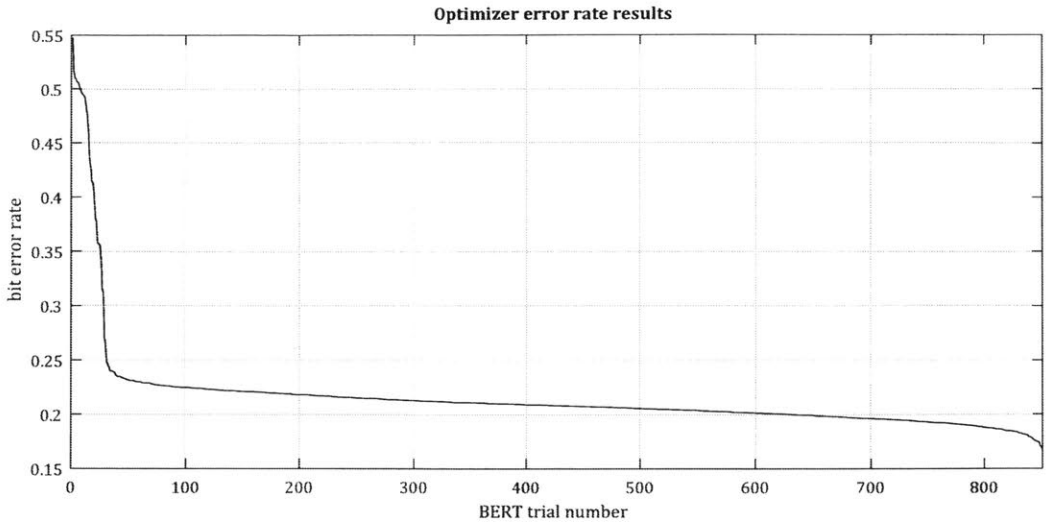

Figure 4-18: Plot of all the error rate results generated during the optimization progress, sorted from highest to lowest. There can be seen to be very few relatively high error rate results and at around 50 BERTs the error rate remains relatively constant. This first section is primarily the optimizer finding progressively better operating points with progressively lower BERs. The results beyond around BERT trials 100 are suspected to be primarily a statistical phenomenon. Since we have a finite-length BER which is sampling a random process, we expect most of the BERs to be around the expected BER, with very few having a lower BER. The shape of this function beyond BERT trial 100 is found to be typical of such an error-limited experiment, as opposed to a margin-limited experiment such as that shown in figure 4-23. This result could be thought of as a superposition of the BER's CDF and the optimizer's progress.

4.5.2 Isolated unselected cell testing

With the array experiments producing poor error rates, we decided to step back and operate a single cell without the heater. This experiment would allow us to assess if there is an issue with the use of the heater. We suspected that the suppression obtained by the heater is inconsistent, and may be resulting in the poor error rates that we have obtained thus far. The setup shown in figure 4-9 was again used here. The only modification to the setup was that the heater was terminated into a $50\,\Omega$ load, rather than being connected to the AWG.

The operating point was found manually by varying the write and read levels. After some short time searching for the best operating point, it was clear that in this

operating mode the memory was performing far better than in previous experiments. A long BERT consisting of 20,000 write/read cycles, each cycle alternated between setting the cell to the "1" state, and the "0" state, was performed. The memory switching current distributions are shown in figure 4-19.

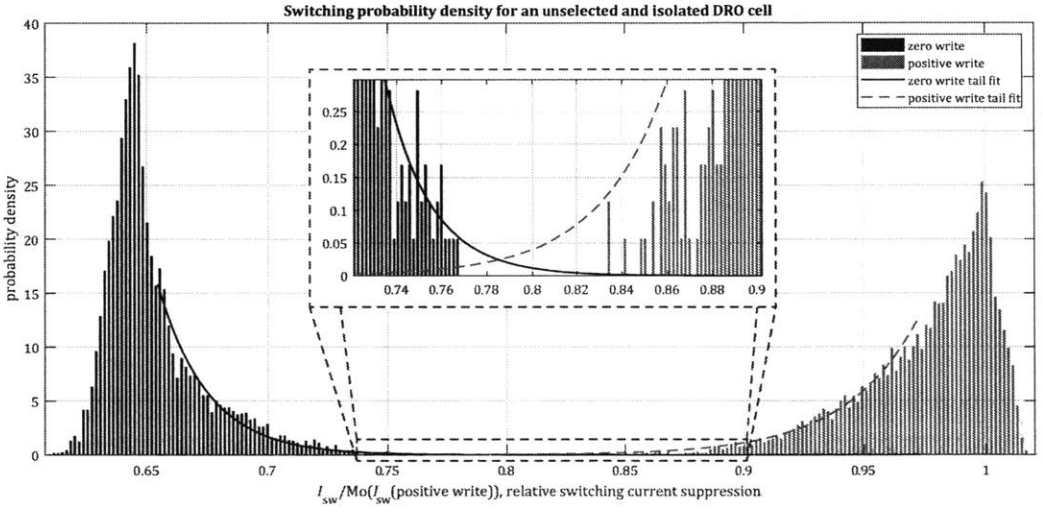

Figure 4-19: Approximation of the probability density function of the switching current suppression for an isolated and unselected DRO cell. One histogram represents the read current after a "0" write, and the other after a "1" write. The results have been normalized to the mode of the positive write switching current distribution. Two exponential fits are added to the histograms tails, one for the zero write switching current and one for the one write switching current. The insert is a magnified view of the overlap, or lack thereof, between the two distributions. Note that there are no errors observed, that is there is no overlap between the histograms – although there is an overlap between the fits.

In this experiment no errors were observed, suggesting an error rate $P_{\mathrm{E}} < 5 \times 10^{-5}$. Two exponential fits were added to the tails of the histogram approximations of the PDFs, shown in figure 4-19. From these fits, an error rate of $P_{\mathrm{E}} \approx 10^{-4}$ is estimated. These error rates are substantially better than any obtained so far with the DRO design. These results indicate that the cell design is in fact operating perfectly, and that is issue is a result of the heater. To understand why the heater is causing these issues, the switching distributions of the hTron were analyzed.

4.5.3 hTron distributions in helium immersion measurements

While the revised device design had improved the error rates, they are still very poor at around 15% when in an array. The isolated and unselected (no hTron gate current) operation of the DRO cell has shown far better error rates (around 10^{-4} and lower). These results indicate that the selected hTron is likely the source of the variations that give rise to the poor error rates. In order to explore this, it was decided to explore the operation of an isolated hTron, and in particular its switching distribution with various gate biases applied.

The experimental setup for these experiments is very simple. We simply wish to perform a sequence of IV curves. In order to perform these measurements we use the setup shown in figure 4-20. We only perform the positive portion of the IV curve as we will only consider the positive switching current. The switching current is not directly measured, rather we do as was done in previous NDRO experiments, and measure the time to switch, and from that infer the switching current. At each gate bias, we perform a number of switching current measurements so as to be able to determine the distribution of the switching current as a function of gate bias.

First, the LHe immersion measurement was performed, and as expected, the distribution of the hTron switching current becomes very large as the suppression nears 50% of the zero-bias switching current, as shown in figure 4-21. This was speculated to be caused by helium gas bubbles forming on the surface of the chip. To determine if this explanation is viable, the chip was lifted out of the LHe, while still being close to the surface of the liquid, as a result the temperature of the chip should still be close to 4.2 K.

With the chip suspended above the LHe, the experiment was performed again, the results of this experiment are shown in figure 4-22. It was speculated that the cooling power of the gaseous He would be low, and with the heat load of the device and cables, the device may gradually increase in temperature during the relatively long measurement period. In order to determine if temperature drift is an issue, the sweep was performed once in the forward direction and immediately following this,

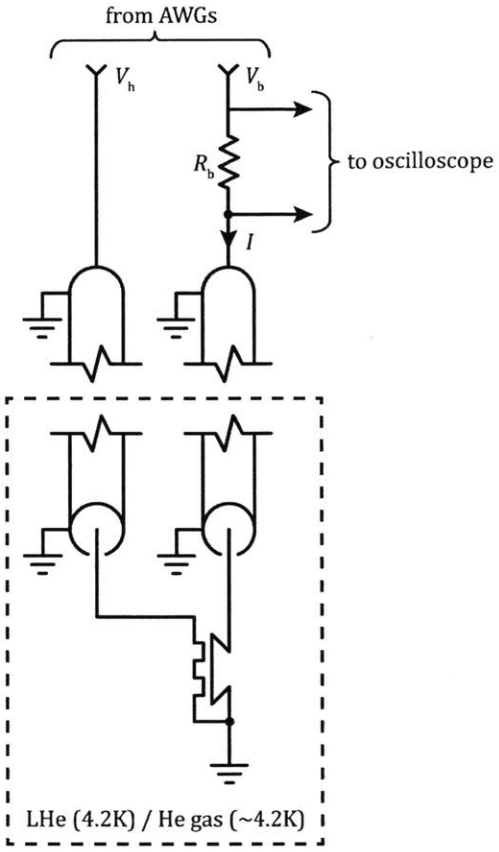

Figure 4-20: Experimental setup for performing the hTron switching current distribution measurements. The hTron tested here is a multilayer device with a normal-metal gate, one side of which is grounded and the other side supplied a voltage bias. The switching current was measured in the same manner as was done in previous IV curve measurements. A bias resistance of $R_b = 10\,k\Omega$ was used. The experiment was conducted once with the device submerged in LHe, and a second time with it suspended above the LHe with the device only exposed to cold He gas.

once in the reverse direction. Thus, any heating of the device, or any other effects that change over time would be evident as a difference between the two sweeps. While figure 4-22 only shows the forward direction, the reverse direction results were nearly identical, indicating the cooling power of the He gas was sufficient to maintain a stable temperature.

In comparison to the immersion experiment, the gaseous He experiments showed far more reliable and consistent switching currents – even at 50% suppression. This lends credence to the theory that the formation of He gas bubbles on the surface of the

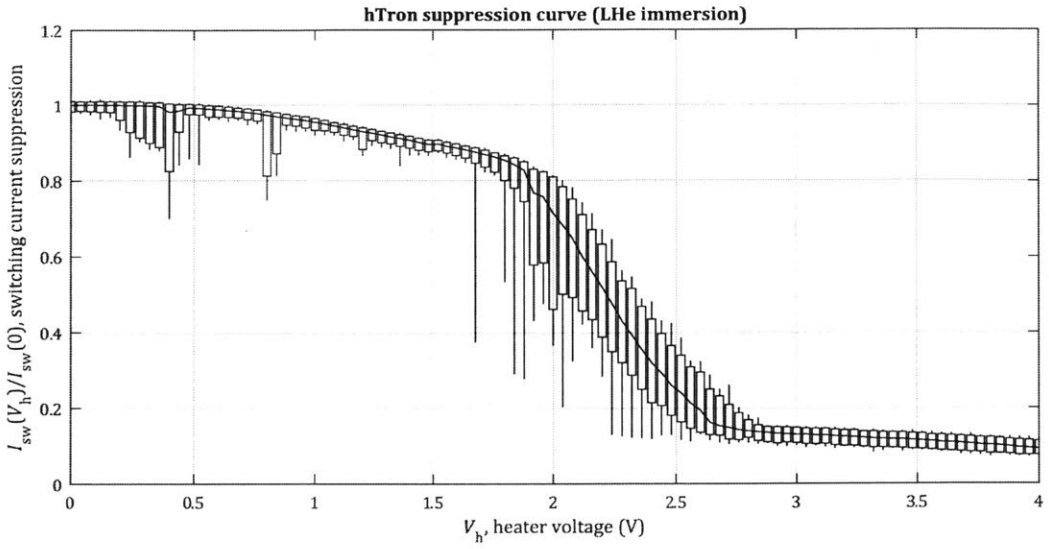

Figure 4-21: hTron suppression curves captured with the device immersed in LHe. Each point in this plot represents 1,000 switching current measurements. The line plot indicates the median, the box the extents of the 1% and 99% quantiles for that particular heater bias, and the whiskers are the maximum and minimum of the measurements. It can be seen that at low and high heater biases the switching distributions are relatively narrow, and at intermediate suppressions, the distribution is extremely wide. The very high variations in the switching current are suspected to be due to the formation of He gas bubbles on the surface of the chip.

chip during the application of heat may be causing the switching current variations. This effect was likely not witnessed in the in-plane hTron measurements due to the fact that the heater was superconducting, and as a result the heat localized to a very small area, and as a result no bubbles were formed. With this issue identified, we decided that future testing should be conducted in a cryogen-free system. This move was made because in such systems, the sample is attached to a cold-head in an evacuated chamber, and never exposed to LHe.

4.6 Refined test procedure and experimental results

With the discovery that the hTron switching distributions are affected by the immersion of the device in LHe, the experiment was moved to a cryogen-free system. The same measurement procedures that were covered in section 4.5.1, are again used here,

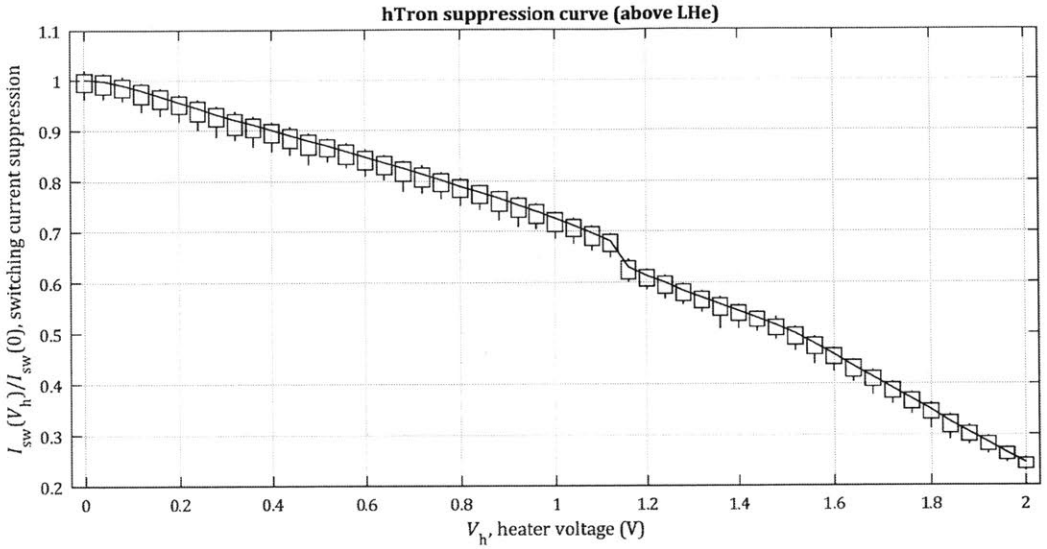

Figure 4-22: hTron suppression curves captured when the device is suspended above LHe, with the sample exposed only to gaseous He. Each point in this plot represents 1,000 switching current measurements. The line plot indicates the median, the box the extents of the 1% and 99% quantiles for that particular heater bias, and the whiskers are the maximum and minimum of the measurements. It can be seen the extents of the distribution at each heater bias are roughly equal. This result confirms that some aspect of LHe immersion experiments, likely the formation of He gas bubble on the surface of the chip, are responsible for the poor hTron distributions, and possibly the poor error rates.

with the only difference being that we are now using a cryogen-free system. In order for a BERT to give a reliable result, then all parameters must be constant throughout the BERT. For this reason, the suspension of the sample in the He gas, as was done in section 4.5.3, cannot be used gain here since the temperature is unlikely to be stable over a the length of a BERT. An additional advantage of using a cryogen-free system is that the optimization can be allowed to run for an almost indefinite amount of time.

Like previous optimizer measurements, an initial starting point was found by manually varying the operating parameters. The optimizer was allowed to vary the write level, read level, and heater level, in order to minimize the cost function. Initial experiments were conducted with the cost function being equal to the error rate, which was found with a BERT consisting of 2,000 write/read cycles of alternating

zeros and ones. The optimizer, after a few tens of minutes, located operating points where the error rates were lower than the BERT could detect – that is, in one BERT no errors were detected. We have two options to allow the optimizer to progress, first increase the BERT length, which would make the optimization take a very long time, or modify the cost function. The latter option was pursued, and is covered in section 5.1.2.

The new cost function requires an approximation of the switching current distributions in order to attempt to maximize the distance between the distributions tails (while also minimizing the BER). In order to gain this information, a ramp readout scheme must be used. This means that the optimizer need only vary two parameters, namely the write level and the heater level. The results of this optimization run, which was run for a number of days, are shown in figures 4-23, 4-24, and 4-25.

Figure 4-23 graphically demonstrates the ability of the new cost function to continue the optimization beyond the point where the BER is too small to be directly measured. It can be seen that at BERT trial 3,886 the BER is below that which can be directly measured. From this point on, the cost function is relying solely on maximizing the distance between the distribution tails. By continuing, the optimizer is able to find operating points where the margins of the readout, and possibly the ultimate error rate, are improved. In addition, it can be seen that some points, while having a low error rate, have a very poor distribution separation. This effect is graphically shown in figure 4-25, where some regions appear to show good error rate, but when looking at the same point in terms of distribution operation, we see a very poor value. This behavior is likely due to either the mean of each of the two distributions being close to each other, or more likely the tails of the distributions being very long. Either of these possibilities is undesirable, and as a result, the optimizer, with the new cost function, is finding operating points closer to the ideal case of small tails and wide separation of the distribution means – leading to low ultimate error rates.

The distribution for the best operating point found by the optimizer – after being run for a number of days – is shown in figure 4-25. During the 2,000 trials that were performed for the BERT, not a single error was observed – provided we chose

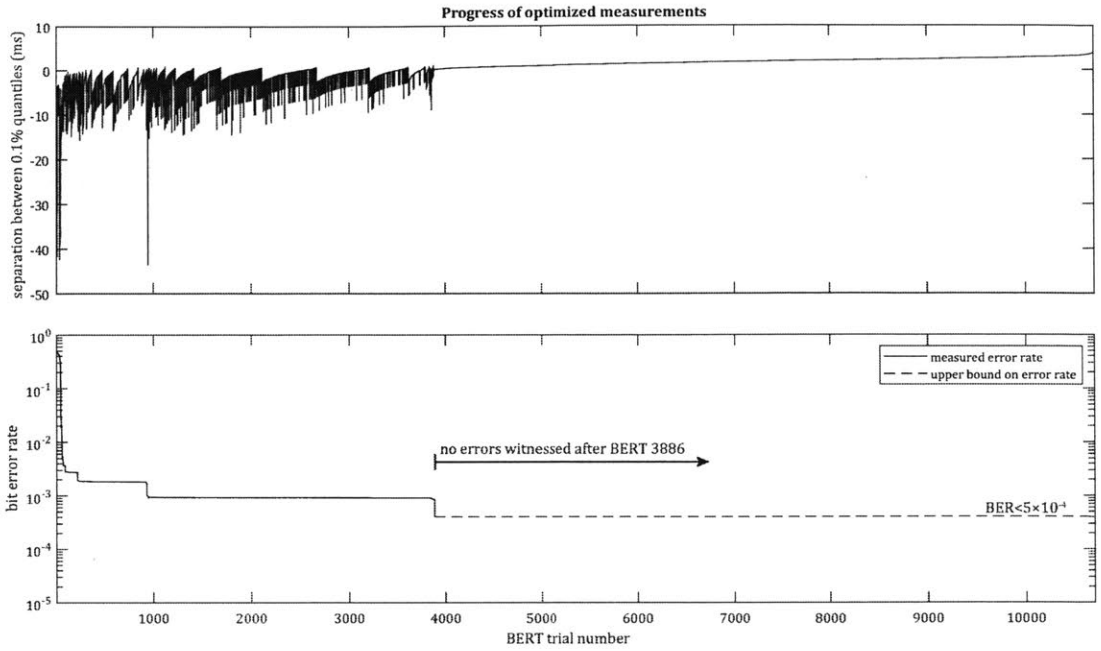

Figure 4-23: Plot of the read current distributions after a write "1" and write "0" (top), and the error rate (bottom) as measured throughout the optimization process. The results have been ordered from the highest cost value, to the lowest cost value (see section 5.1.2). The separation between the distributions is measured as the distance between the upper 99.9% quantile of the switching current distribution after a "0" was written, and the lower 0.1% quantile of the switching current distribution after a "1" was written. Each BERT consisted of $2,000$ write/read cycles. For BERTs beyond trial $3,886$ no errors were observed, so the error rate can be estimated to be below $P_{\mathrm{E}} < 5 \times 10^{-4}$. After the observed error rate drops to zero, the separation between the distributions does continue to improve, but only slightly.

an optimal decision rule. The absence of any errors suggests an ultimate error rate of $P_{\mathrm{E}} < 5 \times 10^{-4}$. However, as relatively few tails were performed, we decided to calculate a fit, again using the Burr distribution, to find an estimate of the error rate. With the fit shown, an error rate of $P_{\mathrm{E}} \approx 1.5 \times 10^{-3}$ was predicted. It can be seen that the fit does not match perfectly, so the error rate is likely somewhere between these two estimates. Regardless, this error rate is orders of magnitude better than that obtained in LHe immersion measurements.

As a final measurement, we decided to operate the memory in the pulse-based readout scheme – as would be done in a practical implementation. This measure-

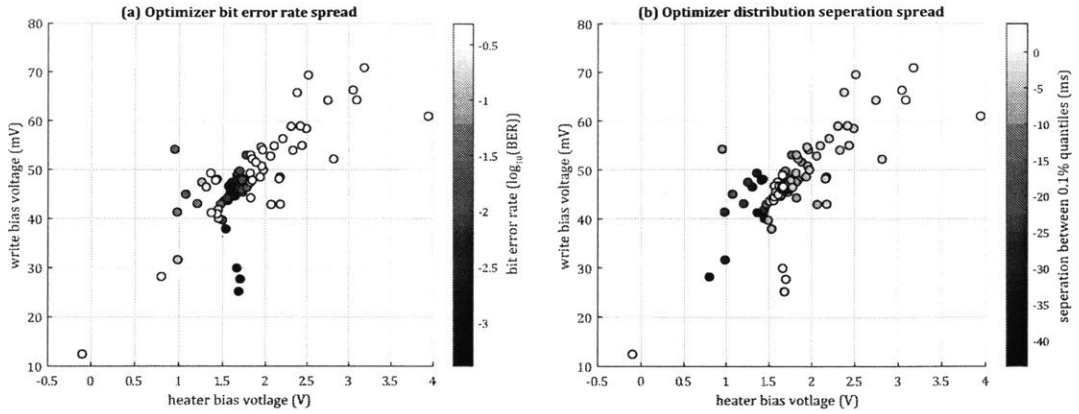

Figure 4-24: Scatter plot of the bit error rate (a) and separation between distributions (b) at each operating point the optimizer explored. For (a), a lower error rate (darker colored point) is better, and for (b), a higher separation (lighter colored point) is better. Since the error rate varies over a wide range the color was plotted in a log scale of the error rate. Interestingly, points with low error rate, do not necessarily have good separation between distributions, hence why the cost function incorporates both parameters.

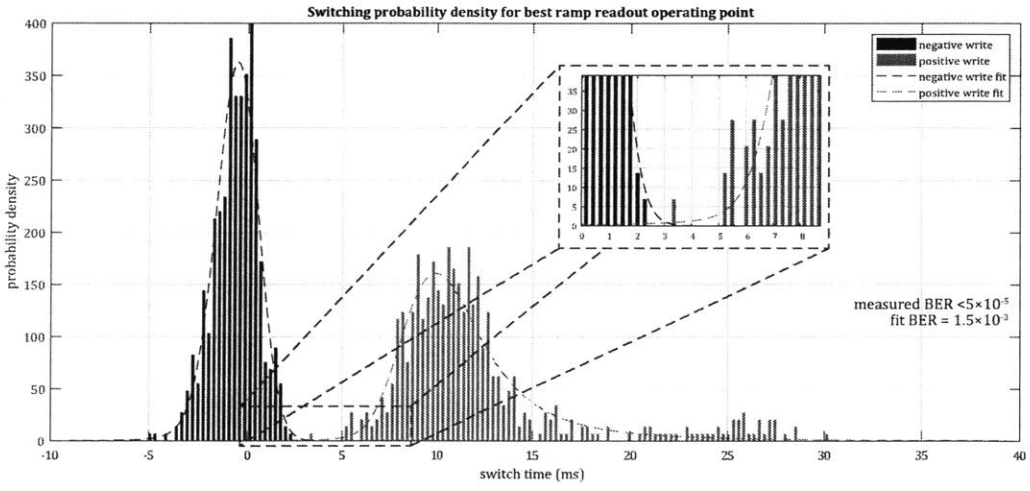

Figure 4-25: Switching probability distribution for the optimal point found during this experiment. This BERT consisted of 2,000 write/read cycles which alternated between writing a "0" and writing a "1". Throughout this experiment no errors were observed, so the error rate can be estimated to be better than $P_{\mathrm{E}} < 5 \times 10^{-4}$. Two fits were calculated using a Burr distributions, and their overlap used to calculate the fit-predicted BER of $P_{\mathrm{E}} \approx 1.5 \times 10^{-3}$.

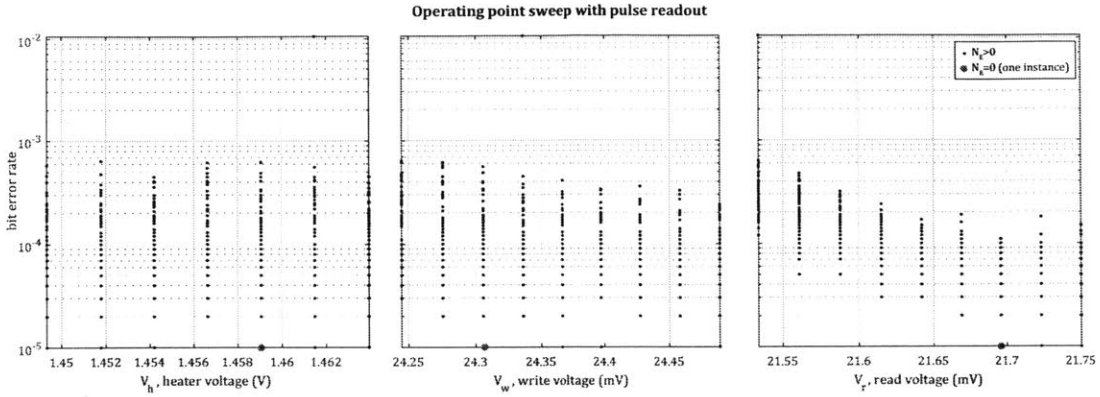

Figure 4-26: Results of a pulse-based readout when sweeping all three operating parameters over a narrow range. Each BERT consists of 100,000 write/read cycles which alternated between writing a "0" and writing a "1". This sweep took one day and twelve hours to complete. As the data is four-dimensional, three projections onto dimensional plots were performed. Each of these three plots shows the error rate plotted against each of the three operating parameters. The gray highlighted point indicates the location of the one operating point where the number of errors observed over the BERT was $N_{\mathrm{E}} = 0$.

ment consisted of a series of long BERTs conducted over a narrow, and near optimal operating parameter range (according to optimizer results). At each point a BERT consisting of 100,000 write/read cycles, each alternating between writing "0" and writing a "1", was performed. The initial operating point was found manually, and the optimizer run for a two of days to bring this point close to optimal. The best point the optimizer found was then used as the starting point for this sweep measurement. Each parameter was varied by $\pm 5\%$ around the optimizer's best operating point, with seven equally spaced heater levels, and nine equally spaced read and write levels. The results are shown in figure 4-26.

Over the entire sweep shown in figure 4-26, there were zero errors witnessed at only one point. At this point, we can estimate the error rate is $P_{\mathrm{E}} < 10^{-5}$, while at many points we observed error rates of $P_{\mathrm{E}} = 10^{-5}$. Such error rates are a substantial improvement over the error rates found in the ramp-based readout scheme. This improvement is likely a function of the fact that the readout process is faster, and as a result, somewhat more resilient to noise. The error rates found here suggest that

the DRO may be capable of performing with similar BERs to the NDRO cell, while also not requiring the substantial additional electronics the NDRO cell required to be arrayed. Thus, the DRO cell is a good candidate for a high-density and scalable superconducting memory technology.

Chapter 5

Experimental setup and apparatus design

This chapter focuses on two main contributions to our laboratory's infrastructure that were made throughout the work conducted on the memory project. The first is the creation of a new automated testing solution that incorporates an integrated optimizer. This system was developed when it was realized that finding the operating point of the DRO memory was taking too long to locate manually, so an automated system was developed. This system is covered in section 5.1. The second contribution is a new experimental apparatus that was inspired by the magnetic modulation experiments conducted on the DRO cell. In the DRO magnetic modulation measurements, a superconducting magnet was wound by hand and the sample placed within the magnet, with the whole assembly lowered into the dewar. This method left a lot to be desired. Thus, with the recent acquisition of a new cryostat, and hence the need for a new experimental apparatus designed specifically for this cryostat, it was decided that this apparatus should incorporate a superconducting magnet. This apparatus will allow for the study of the effects of magnetic fields on our samples, along with current, temperature, and exposure to optical radiation. This new apparatus is covered in section 5.2. With the addition of the two contributions, our laboratory is substantially more capable of exploring new and existing phenomena.

5.1 Automated testing

All of the experimental results presented in this work, with the exception of device resistance measurements, were obtained with programmatically controlled instruments. The experiments conducted prior to section 4.4, utilized python to scripts to control the instruments. These scripts were primarily used to acquire data and set biases. Waveform shapes were manually programmed into the AWGs. This solution sufficed for basic parameter sweeps and captures of the oscilloscope traces, and worked well for NDRO cell measurements.

With the advent of the DRO cell, the original automated testing approach became impractical. There were two major driving forced behind the decision to move to a new experimental automation system. First, the DRO cell requires a channel bias signal that is more complex than that used in the NDRO measurements, thus custom arbitrary waveforms that are programmatically controlled were required. Second, the difficulty in finding the operating point of the DRO cell, combined with the time that a BERT takes to complete, meant that an integrated optimizer was desired. This optimizer would automatically make BERT measurements, and vary operating parameters to achieve lower BERs.

Previously, data was captured with Python scripts and then imported into MATLAB for analysis. This system is satisfactory when the quantity of data is relatively small, and there is little to no feedback from the analysis side to the experiment side of the problem. Since we were investing a substantial amount of time in developing the new test setup, we decided to write new instrument drivers for MATLAB. In writing these new drivers, we can perform data acquisition and analysis in one package, and feed analysis results back to the experiment with minimal effort. The details of the implementation of these drivers is not included here as they are instrument-specific and not of great interest. Instead, we will focus on the optimizer, and its application to the DRO array measurements.

5.1.1 Integrated optimizer

Finding an operating point of the DRO cell that yielded a low error rate was proving to be particularly difficult. This issue is compounded by the fact that, in order to test if a point is better or worse than a previously tested point, requires the performance of a full BERT. Performing a BERT is fast when the error rates are poor, as few write/read cycles are required to establish an accurate estimate; however, as the operating point is moved closer to the optimal, the BERTs necessarily become progressively longer. Thus, we decided that an automated optimizer, that would adjust operating parameters and perform BERTs, was required.

An optimization algorithm is typically totally blind to the intricacies of the problem at hand [34]. Rather, the algorithm's sole goal is to minimize some cost function's value by varying a set of parameters. Optimization algorithms typically utilize some information about the direction in which the cost function's value improves, and follows this direction. As the optimization progresses, this direction can change, and depending on the algorithm, multiple directions may be considered simultaneously.

The choice of optimization algorithm is not to be taken lightly. There are many factors that must be considered in order to achieve good performance. One must consider what parameters of the problem we have available to us, if we expect many local minima, and the expense of performing a cost function evaluation. For our problem, we only have access to the cost function value at any one function evaluation – we do not have access to derivatives. Only having access to the cost function's value at each point means that we are restricted to the derivative-free class of optimization methods. Our cost function is quite time-consuming to evaluation, as it involves performing a complete BERT. Thus, we require an algorithm that will not lead to an excessive number of cost function evaluations. Finally, we do not know how many local minima we may encounter, and their proximity to the global minimum; however, we do not necessarily expect local minima. Given that our problem has the aforementioned properties, it was decided to implement a downhill simplex (Nelder-Mead) algorithm with simulated annealing [35, 36]. This approach was pursued as

downhill simplex is easily implemented, and meets our requirements.

A custom implementation of the downhill simplex method, which was modified to include basic simulated annealing, was written for MATLAB. This class was written using an object oriented paradigm. Once the class is instanced (or during instantiation) the starting point, cost function handle, termination function handle, starting temperature, limits on iteration count, and a limit on total optimization time, are passed to the class. The optimizer is then started. When the optimization is complete, the user can then run their problem-specific analyses. A graphical representation of structure of the optimization process is shown in figure 5-1. The code is written such that if at any point the optimization is aborted (or a termination criteria is met), the optimization can be resumed (with the same or different parameters) and the optimizer will resume as if no pause had occurred. This operation is very useful as often the experiment ran for days, and frequent interruptions in the optimization process occurred.

An example of the operation of our implementation of the downhill simplex algorithm with annealing disabled is shown in figure 5-2. In this problem a simple conical surface the subject of the optimizer, and the termination criteria is contingent on the number of iterations. This problem involves only two parameters x and y, and as a result our k-simplex is a triangle. This shape is graphically illustrated in the figure, where we can see a series of triangles approach the optimal point of the cost function. In this example problem, we could have easily implemented a far better method that would have converged in far fewer steps; however, for the actual problem at hand (the DRO memory) the surface is totally unknown.

Since the DRO error rate surface is unknown, it is also unknown if there are many or no local minima to contend with. Downhill simplex, like most optimization algorithms, is likely to fall into, and become trapped within, local minima. There are a number of methods that can be used to avoid this behavior; however, we must weigh the use of these methods against the added overhead of an increased number of cost function evaluations. Here, we opted for a basic simulated annealing approach, which in many circumstances, can allow the optimizer avoid becoming trapped in

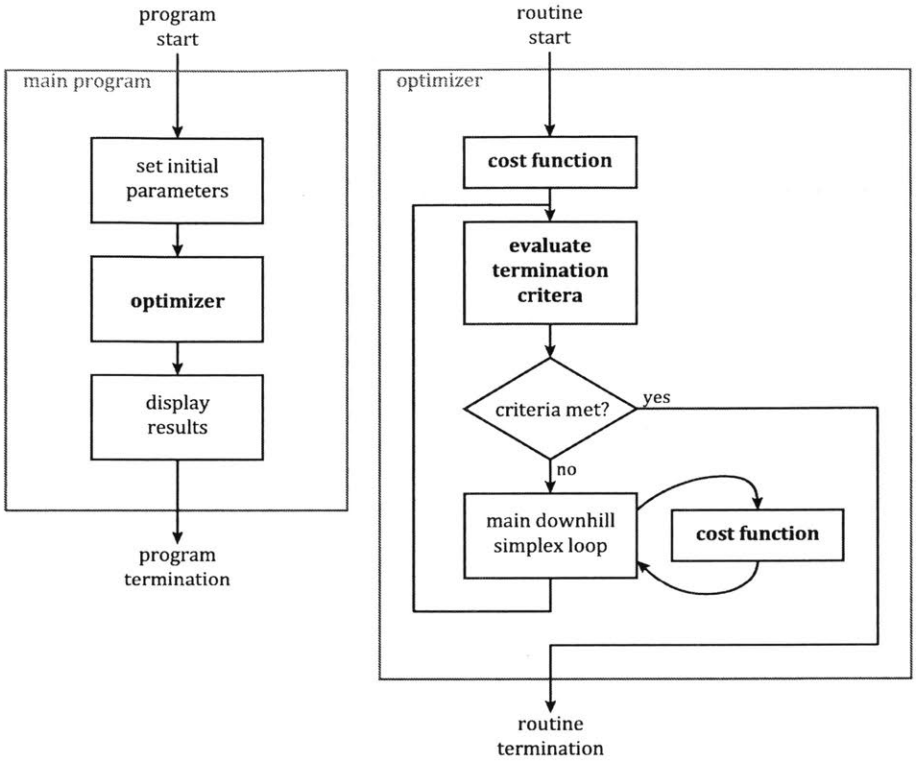

Figure 5-1: Block diagram showing the interaction between the users main program and the basic structure of the optimizer. Functions are shown in bold. The user program creates an instance of the optimizer and sets the initial parameters, and passes cost function and termination criteria function handles. The program then calls the optimizer, and once optimization is complete is displays the results. The optimizer takes the initial operating point values, calls the cost function and begins the main optimization loop. This loop first calls the termination criteria function and if the function indicates the optimization is complete then the optimizer quits. If the criteria is not met, then the main downhill simplex routine is run. During the execution of this routine a number of calls to the cost function may be made.

local minima. An example of the use of the simulated annealing augmented downhill simplex method is shown in figure 5-3. In this problem there is one global minimum, and an additional local minimum. When the optimization process is started near the local minima, our optimizer with annealing disabled, falls into the local minimum and becomes trapped. This behavior is in contrast to when the optimizer is run with annealing enabled, where is can be seen that the optimizer converges to the global minimum.

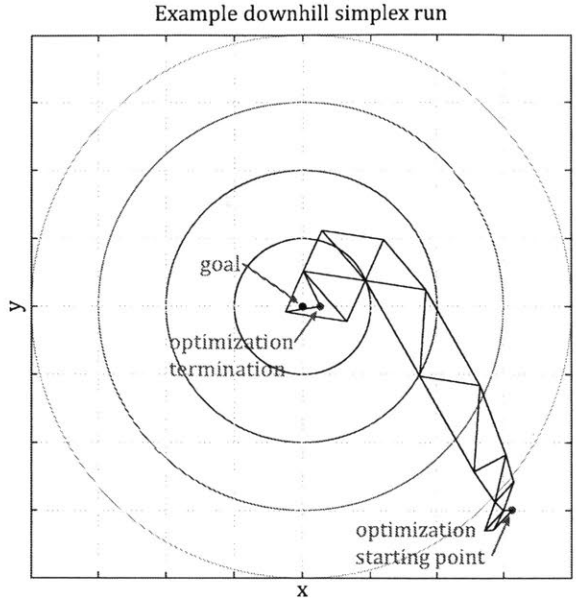

Figure 5-2: Example operation or our downhill simplex optimizer applied to a conical surface with annealing disabled. A contour plot showing the decrease in the cost function towards the center of the figure is shown. Throughout the optimization process, the downhill simplex algorithm stores $(n + 1)$ points, where $n = 2$ is the number of dimensions, and using geometric operations progresses towards the optimal. Here, lines have been drawn between each of these points at each iteration. It can be seen that the optimizer works from the starting point towards to goal (optimal point). Here, the optimizer was limed in the number of iterations it could perform, so it terminates close to the goal, but never reaches it. The n-simplex (in this case triangle) shape is clearly visible in the progression of the optimizer.

5.1.2 Cost function

With our implementation of the optimization algorithm complete, we shift our focus to the cost function. With the structure of the optimizer shown in figure 5-1, the cost function is responsible for all the problem-specific calculations. In the case of the DRO cell, this involves a number of operations, which can be summarized as performing a BERT and then calculating a cost from the results of this measurement. A summary of the structure of the cost function is shown in figure 5-4.

As shown in figure 5-4, the cost function is broken into two parts. The cost function itself performs all the cost calculations, and offloads the hardware interface

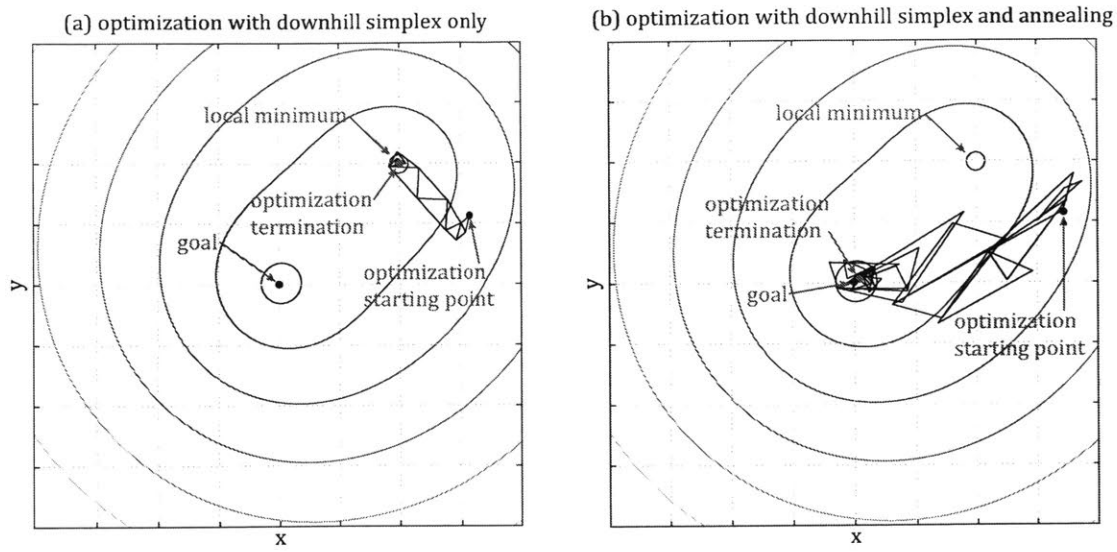

Figure 5-3: Comparison of the downhill simplex algorithm without annealing (a) and with annealing (b) when applied to a problem that contains a global and an additional local minimum. The optimization process was, in each case, started in the same location. The connections between each of the three points the algorithm used were plotted at each iteration. It can be seen that the optimizer without annealing converged to the incorrect local minimum, whereas the algorithm with annealing converged to the global minimum. The effect of annealing can be seen graphically as the highly varying size, and non-overlapping edges of the triangles drawn at each iteration.

to another data acquisition function. Thus, a cost function evaluation begins with the optimizer passing a vector containing the operating point to be tested to the cost function. The cost function then passes this to the data acquisition function. This function first generates the waveforms which correspond to the operating point vector. The function then uploads to these waveforms to the AWGs, and clears the oscilloscope which initiates the acquisition process. The function then waits for the number of oscilloscope acquisitions to be greater than the number of trials to be performed N_{T}. Finally, the data acquisition function downloads the traces from the oscilloscope and passes the results back to the cost function.

With the oscilloscope traces in hand, the cost function determines the BER. In order to perform this operation, a decision rule is required. As covered in section 2.4.5, a simple threshold decision rule will suffice. Since we want the best error rate

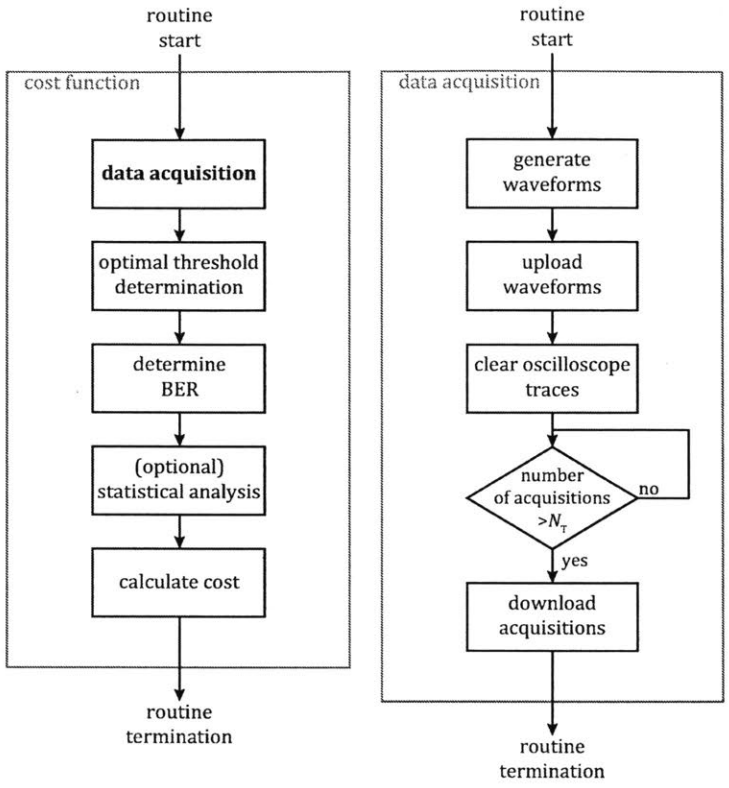

Figure 5-4: Block diagram depicting the program flow through the cost function evaluation process. The optimizer calls the cost function, which in turn calls the data acquisition function. The data acquisition function interfaces with the hardware. It generates the waveforms, uploads them to the AWG, clears the scope and waits for it to capture at least as many acquisitions as the number of trials in the BERT N_T. The data acquisition function then downloads the acquisitions and passes the result back to the cost function. The cost function then finds the optimal threshold using algorithm 1 and determines the BER. If the experiment is using a pulse-based readout, then optional statistical analysis is skipped and the cost set according to equation 5.1. On the other hand, if a ramp-based readout is used, then the statistical analysis is performed and the cost value calculated according to equation 5.2. The final cost value is then returned to the optimizer.

estimation possible, we need some routine to find the optimal threshold – or close to it. The implementation used here is a simple iterative approach. This approach is summarized in algorithm 1. The operation of this algorithm is shown graphically in figure 5-5. It can be seen that the algorithm rapidly converges on the optimal threshold position, thus allowing for the best estimate of the BER. With this threshold

determined, the cost function can obtain an estimate of the BER.

Algorithm 1 Threshold determination algorithm.

```
1: procedure ThresholdDetermination
2:     modeSep ← abs(mode(I_s,1) − mode(I_s,0))
3:     thresholdOffset ← modeSep/2
4:     threshold ← (mode(I_s,1) + mode(I_s,0))/2
5:     while thresholdOffset > modeSep × tolerance do
6:         thresholdRight ← threshold + thresholdOffset
7:         thresholdLeft ← threshold − thresholdOffset
8:         errorRight ← sum(I_s,1 < thresholdRight) + sum(I_s,0 > thresholdRight)
9:         errorLeft ← sum(I_s,1 < thresholdLeft) + sum(I_s,0 > thresholdLeft)
10:        if errorRight > errorLeft then
11:            threshold ← thresholdLeft
12:        else
13:            threshold ← thresholdRight
14:        thresholdOffset ← thresholdOffset/1.5
15:    return threshold
```

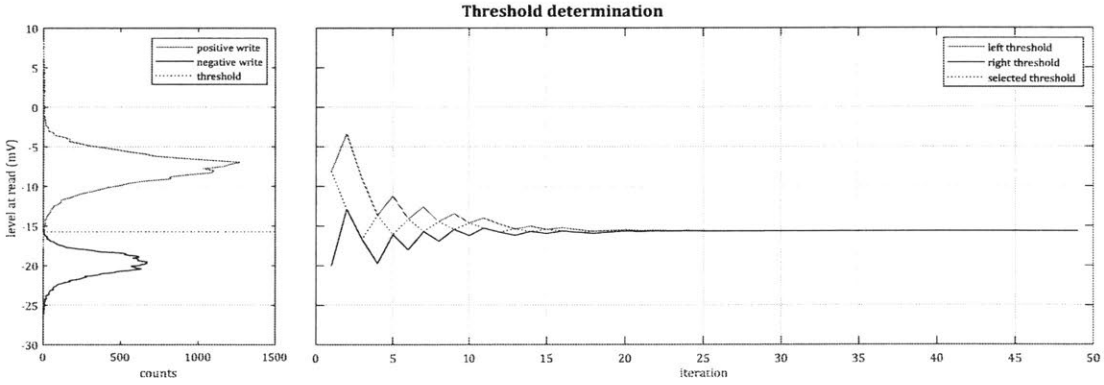

Figure 5-5: Histograms of the read data provided by the data acquisition routine (left), and a graphical illustration of the process used to find the optimal threshold (right). In this evaluation, a tolerance value of 10^{-5} was used. The threshold determination algorithm starts by choosing an initial threshold equal to the mean of the modes of each distribution. Two trial thresholds to the left and to the right of this initial threshold are considered. The better of these two thresholds (one with the least number of errors if used to divide the data) is then used as the new threshold. This process is repeated with ever decreasing separation of the left and right trial thresholds. Finally, the process stops once the tolerance value is satisfied. At this point, the threshold value that gave the lowest number of errors is chosen for the final BER evaluation.

In the initial DRO optimizer experiments, the cost function performed a BERT

and reported the cost as the error rate. The cost function was designed this way since, for a pulse-based readout, there are no more metrics we can gain from the experiment, other than the error rate. That is, when performing a pulsed readout, the device either presents a voltage, or it presents no voltage. We cannot gain any statistics from the result other than the error rate. So, in the pulse-based readout experiments we do not perform the optional statistical analysis steps shown in figure 5-4, and the evaluation of the calculation of the cost simply returns the error rate. Thus, for these experiments, our cost function was

$$f(\mathbf{x}) = \frac{N_{\mathrm{e}}(\mathbf{x})}{N_{\mathrm{T}}}, \tag{5.1}$$

where $\mathbf{x}$ is the vector of operating parameters, $N_{\mathrm{e}}(\mathbf{x})$ is the number of errors observed in a BERT at $\mathbf{x}$, and N_{T} is the total number of trials in the BERT.

In later experiments, a ramp-based readout was performed in an attempt to gain a better understanding of what is limiting the error rate. In these experiments, the cost function can calculate a histogram of the switching currents. With this additional data, we can gain much more information about how the device is operating. For the ramp-based readout experiments, the cost function was modified to take into account the separation between read "0" and read "1" distributions. That is, we choose to perform the optional statistical analysis step, and use a more complex cost calculation. In order to implement a cost function that takes into account the distribution separation, we have a number of options for which metric to use. One possible metric would be to find the mean of each distribution and maximize their separation. The drawback of such an approach is that one could conceive of a scheme in which the distributions overlap significantly while their means are very far apart – possibly due to very high skew values. Another possible scheme would simply be to find the extreme values of each distribution and measure the distance between them, with the goal being to maximize this distance. The issue with this approach is that it predicates the cost value on a single sweep. With the relatively long BERTs that are used in these experiments, it is reasonable to expect there to be some erroneous

value due to some external influence. Thus, using the extreme value separation is not ideal. At the time when we chose to use the ramp-based readout scheme, we were already experiencing relatively low error rates. As a result of these low error rates, all the errors we were witnessing were only due to the overlap between the tails of the read distributions. So we could consider a method based on the quantiles of each histogram, and the separation between such.

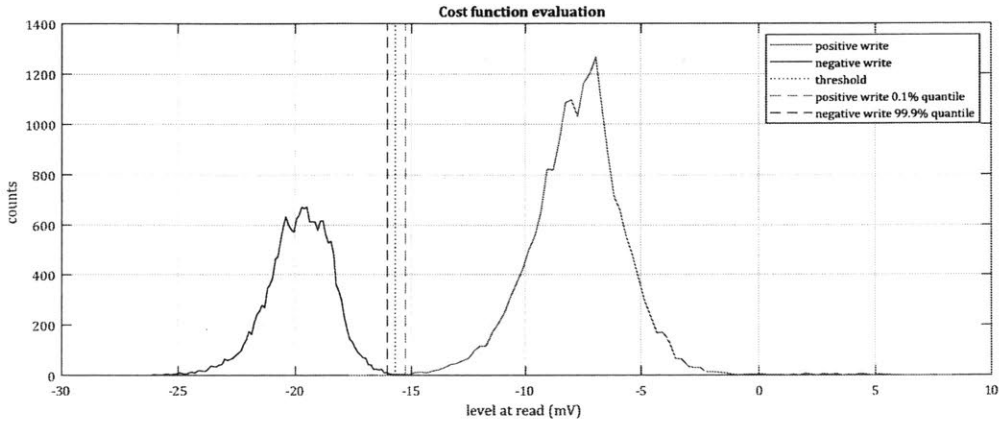

Figure 5-6: Example distribution generated and analyzed by the cost function routine – taken from a measurement of a DRO array. Each curve is a histogram representing the samples measured during that particular BERT. The "level at read" is the voltage level applied to the cell (which through the impedance of the splitter) is converted to a current at the cell. The optimal threshold is shown, and was calculated using algorithm 1. The quantiles of each histogram are also shown. The distance between these quantiles, along with the BER, are the basis of the ramp-base readout cost function.

Consider the experimental results shown in figure 5-6. This figure shows the raw data from the data acquisition routine plotted as histograms. The threshold determination algorithm was applied, and the optimal threshold determined. From this, we can determine the BER. Next, the lower $\delta = 0.1\%$ quantile of the distribution of the read current after a positive write, and the upper $1 - \delta = 99.9\%$ quantile of the distribution of the read current after a negative write are estimates. These estimates are calculated with a simple iterative algorithm. The difference between these estimates is considered to be the separation between read level distributions. Now, we need a method of combining the error rate and this separation distance into

a single cost value.

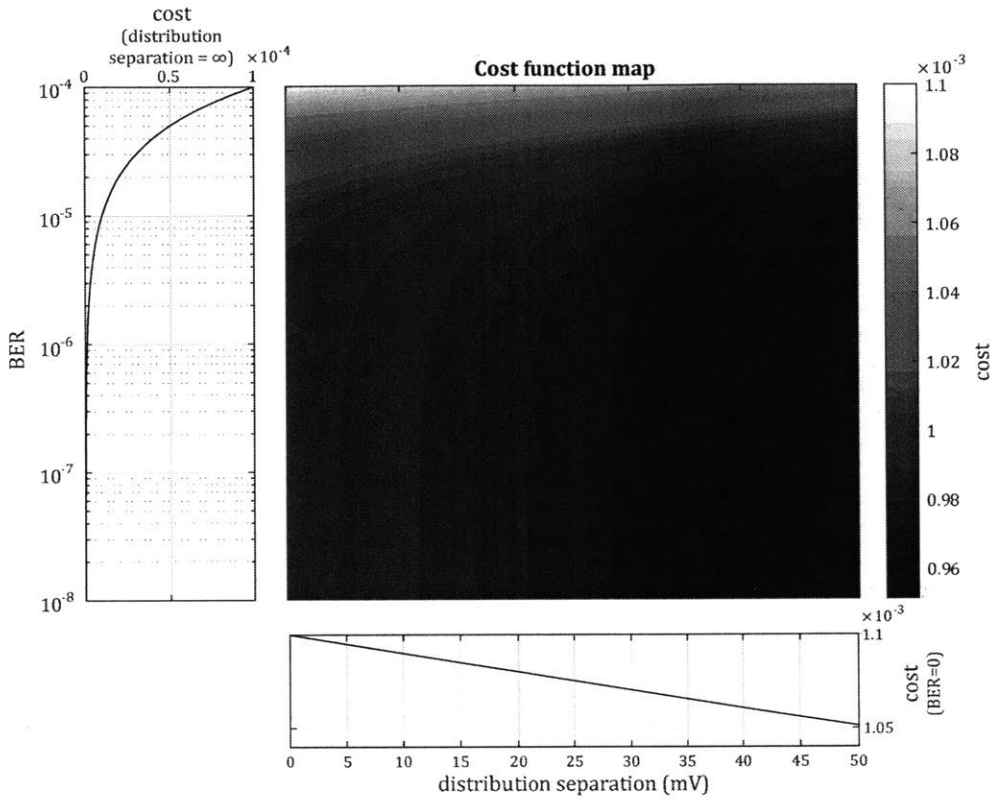

Figure 5-7: Map of the cost function as defined by equation 5.2. In the map, the darker areas are of lower cost vale, and are the locations where the optimizer is seeking. Two cross-sections of the plots are provided. One at a error rate of zero, and the other at an infinite distribution separation. It can be seen that the cost function prioritizes reducing BER when BER $> 10^{-5}$, and prioritizes maximizing the distribution separation when BER $< 10^{-5}$. In this way the cost value is not reliant on the very low value BER estimates which are inaccurate due to the finite length of the BERT.

We cannot leave the BER out of our cost, as there are locations in the parameter space where the separation between the quantiles is comparatively good, but the error rate is poor. An example of this situation can be seen in figure 4-24. We have the additional issue that our optimizer's goal is to minimize the cost function, but here we wish to maximize the separation of the tails. This situation was ultimately solved by simply taking the sum of the BER, and a scaled inverse exponential of the tail

separation. Thus, the cost function can be expressed as

$$f(\mathbf{x}) = \frac{N_e(\mathbf{x})}{N_T} + 10^{-3}\exp(\min(\{u|\mathrm{P}(u > I_{sw,1}(\mathbf{x})) \geq 1-\delta\}) - \max(\{u|\mathrm{P}(u > I_{sw,0}(\mathbf{x})) \leq \delta\})) \tag{5.2}$$

where $I_{sw,0}(\mathbf{x})$ and $I_{sw,1}(\mathbf{x})$ are the sets of switching current distributions at operating point $\mathbf{x}$ after a zero was written, and after a one was written, respectively. This cost function prioritizes error rates, when they are above around BER $> 10^{-5}$, and prioritizes maximization of distribution separation when error rates are below around BER $< 10^{-5}$, as can be seen in figure 5-7. Thus, when the error rate is poor, the optimizer prioritizes finding lower error rates with little concern as to the distribution shape. As the error rate becomes smaller, and the estimate increasingly less accurate due to the finite length of the BERT, the optimizer beings prioritizing distribution separation. This cost function was used in the experiments covered in section 4.6, and has proven itself very effective.

5.2 Design and construction of cryogen-free magnetic-modulation experimental apparatus

During the DRO cell debugging effort, covered in section 4.4.2, magnetic modulation experiments were performed. These experiments were performed with a less than ideal, *ad hoc* experimental setup. The setup used in section 4.4.2 consisted of a hand-wound magnet, that was placed in proximity to the sample. Although basic calculations of the field were performed, the placement of the sample near, but not within, the magnet coil, and the inconsistent shape of the coil, stymied more accurate estimates. In addition, as we found in section 4.5.3, hTrons (and possibly other devices) tend to exhibit extremely wide switching distributions when immersed in LHe. Thus, a new experimental setup that would allow magnetic modulation experiments to be performed in a cryogen-free system was required.

Recently our lab acquired a new Janis cryogen-free "single-shot" ^{3}He cryostat capa-

ble of reaching a base temperature below 300 mK. With this acquisition, we required a means by which samples could be mounted on the cold-head. As this occurred at the time to when the DRO magnetic modulation experiments were performed, it was logical to construct a new setup that would enable future magnetic-modulation experiments. This cryostat is capable of the lowest base-temperature of any within our lab. With the addition of this magnetic modulation setup, we will have in one cryostat the ability to explore the three parameters capable of suppressing superconductivity in our devices, namely temperature, current, and magnetic field – see section 1.1.2.

5.2.1 Overview of design

The cold-head of the 300 mK cryostat is very limited in its cooling power. Practically, this limits the thermal load and thermal mass we can attach to the cold-head. Exceeding the cooling power will result in the cryostat either not being able to reach the specified temperature, or not being able to hold the temperature for an extended period. Excessive thermal mass leads to long cool-down times, and shorter hold times. Since the Janis cryostat is a one-shot system, we are limited in the total amount of heat the cold-head can absorb before it runs out of ^{3}He, and the temperature begins to rise. For these reasons, it was decided to thermally isolate the magnet from the cold-head, as the magnet is large, and must be connected to high-current carrying cables to room temperature. This isolation requires the magnet to be thermally sunk to a different temperature than the sample.

We wish to have a uniform magnetic field normal to the sample's surface. Further, require at least six electrical connections to the sample. We also need a good thermal connection from the sample to the cold-head. Finally, the assembly must fit within the cryostat. Due to these limitations, electromagnet designs using cores such as those of the C-frame or H-frame constructions were ruled out. It was decided to utilize a solenoid design where either a Helmholtz coil, or continuous solenoid could be wound around a cylindrical copper core with the sample mounted in the center of the cylinder – as shown in figures 5-8, and 5-9. The use of a Helmholtz coil would allow for an almost uniform magnetic field over the sample's surface.

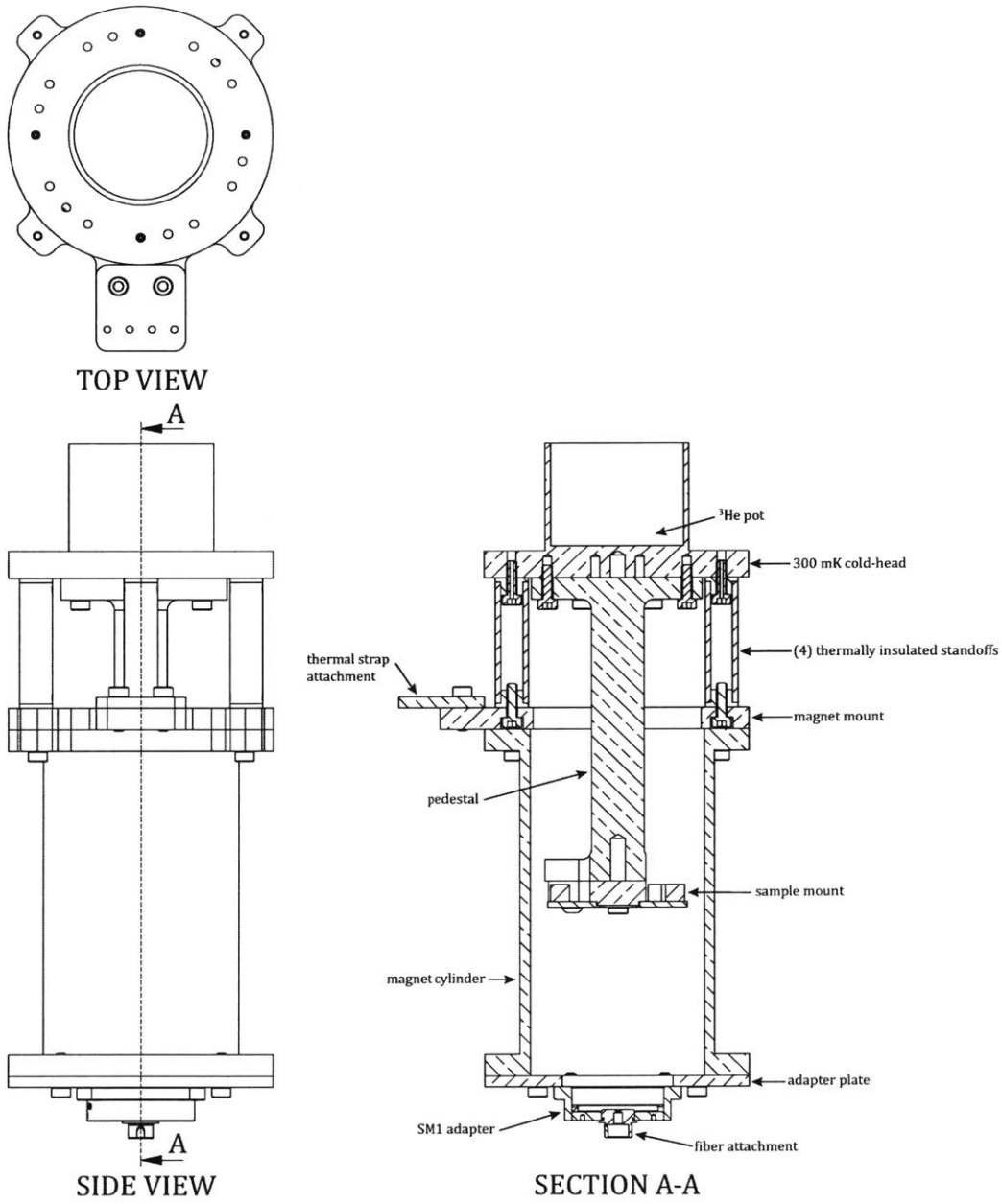

Figure 5-8: Side, top, and cross-sectional views of the new experimental apparatus. The 300 mK cold-head is part of the cryostat, all other parts were designed for this application – other than fixings and optical components. Only the bottom tab of thermal strap is shown. The semi-rigid stainless steel coaxial cables that connect to the sample, and the superconducting wire that would be wound around the magnet cylinder is not shown here. A fiber attachment is shown mounted in the SM1 adapter.

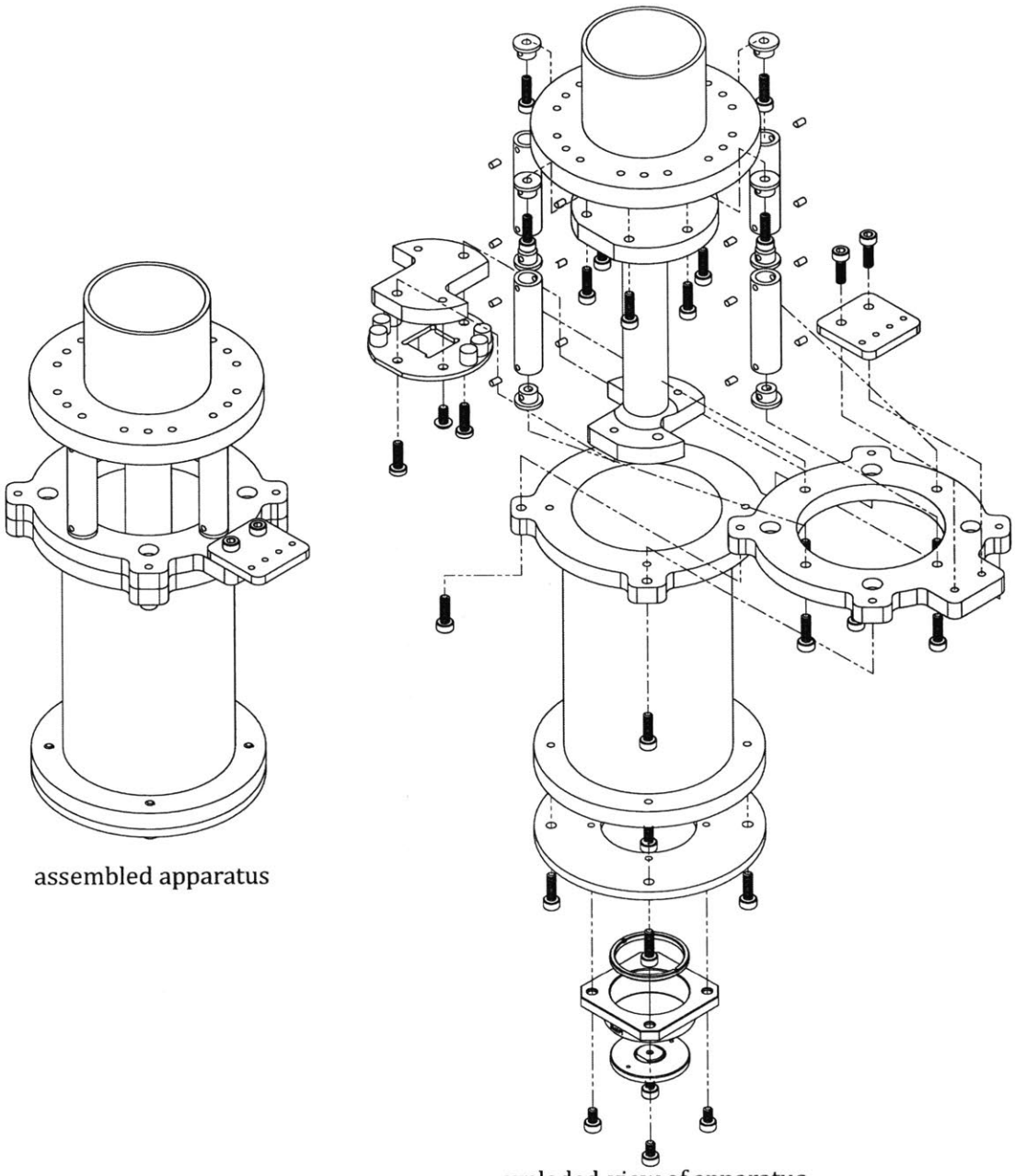

Figure 5-9: Isometric assembled (left) and exploded (right) views of the new apparatus. The thermal strap, and superconducting wire are not shown in these figures. In this figure no sample is attached to the sample mount. A fiber attachment is shown mounted in the SM1 adapter.

It was desired to use a superconducting magnet, as resistive heating of a normal-metal electromagnet would interfere with experiments and severely limit the maximum field attainable. Further to this, the proximity of the cylinder around which the magnet is wound to the sample necessitates the cylinder being cooled. Thus, the magnet must be thermally sunk to a part of the cryostat with sufficient cooling power, but as close a temperature to the cold-head as possible. In the Janis cryostat there is a 3 K bulkhead onto which a radiation shield is attached. This bulkhead has a high cooling power, and is only a few degrees from the cold-head temperature. Thus, the radiation from the magnet to the cold-head should be minimal. Further, we can expect the magnet to be at temperatures around 3 K, and as a result we have the greatest versatility in which superconducting wire we choose to use for the magnet.

In order for the sample to be located in the region of most uniform field with the magnet, it must be held at the center of the cylinder, both axially and radially. The sample must also be thermally sunk to the cold-head. To achieve this, a pedestal was designed which attaches to the cold-head on one end, and to the sample mount on the other end – see section 5.2.2 for the details of the sample mount. One end of the pedestal is made to match the bolt pattern present on the bottom surface of the cold head. The opposing end of the pedestal is shaped to match the sample mount, and has clearance for the cables that exit the sample mount.

With the design outlined so far, there is a need to remove the magnet in order to change the sample. Thus, the magnet must be easily removable, while also being thermally isolated from the cold-head. The magnet could have been attached to the radiation shield; however, during reassembly, the probability of the sample colliding with the magnet would have been high, especially given the large size of the radiation shield. Thus, it was decided to mount the magnet to the cold-head by means of four thermally insulated standoffs. The design of these standoffs is covered in section 5.2.3.

The magnet, being thermally insulated from the cold-head, must be sunk to the 3 K bulkhead. We decided to use a flexible thermal strap to achieve this. It would be possible to mount the magnet directly onto the standoffs; however, this arrangement

would require the thermal strap to be attached and detached from the magnet every time the sample it changed. In order to avoid this, an intermediate mounting plate was designed. This plate is semi-permanently attached to the standoffs and thermal strap. During a sample change, only body of the magnet needs to be detached from the magnet mount. To sink the magnet mounting plate to the 3 K bulkhead, a custom thermal strap was constructed. This strap consists of two copper plates which are soldered to a pair of flexible copper wires.

As much of our group's work involved the photo-response of nanowire detectors, a means by which light could be coupled into the apparatus was added. The bottom of the magnet cylinder was not being used; so a flange was added onto the bottom of the magnet. A plate was then mounted onto this flange. This plate has a Thorlabs SM1F1 internal SM1 adapter attached to the center. This would allow for the attachment of optical assemblies, or simple fiber couplers for flood-illumination of the sample.

All parts of the experimental apparatus – other than the thermally insulating standoffs, fixings, and optical components – are made from oxygen-free high thermal conductivity (OFHC) copper. This selection was made as OFHC copper is non-magnetic, and remains highly thermally conductive at cryogenic temperatures [37]. In addition to this, all copper parts were gold-plated. This serves three purposes. First, it prevents oxidation of the copper. Second, it is highly reflective, and so minimizes thermal loading. Finally, gold being a soft metal, it conforms when two surfaces are mated, thus allowing for low thermal-impedance connections. It should be noted that, a thermal grease was added to most joints to reduce the thermal resistance. As for the other parts, the optical components are constructed from aluminum which is non-magnetic, and all fasteners and metallic parts of the standoffs are made from 316L stainless steel which is also non-magnetic.

5.2.2 Sample mount

The sample mount thermally sinks the sample to the cryostat, and is the means by which signals are coupled from the sample to coaxial cables that lead to room temperature. This setup is designed for $10\,\mathrm{mm} \times 10\,\mathrm{mm}$ samples, although smaller

samples could also be used. The sample mount system was designed such that multiple sample mounts could be prepared and interchanged relatively effortlessly. In previous setups, the samples were bonded directly to the PCB, which was in turn thermally sunk to the cryostat. The downside of this approach is that the thermal coupling through the PCB is poor, and the presence of the coaxial connectors on the back of the PCB make wire-bonding difficult. This sample mount was designed to mitigate these issues.

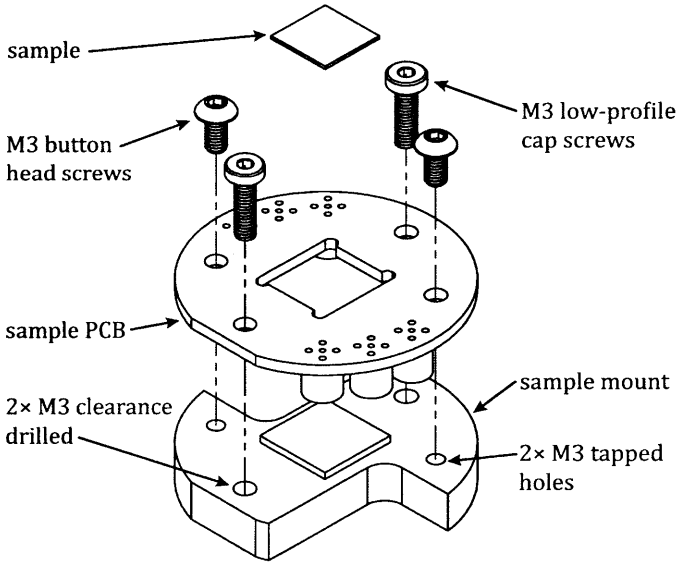

Figure 5-10: Exploded view of the sample mount, along with a sample. The PCB is attached to the sample mount by means of the two button head screws. The sample is attached to the sample mount by means of a soluble adhesive. The entire sample mount assembly is attached to the pedestal by means of the two M3 low-profile screws.

The sample mount consists of three main parts, along with four fixings. The tree parts of the sample mount are, the OFHC copper sample mount itself, a PCB which contains the coaxial connectors, and the sample to be tested, as shown in figure 5-10. The copper sample mount is thicker than the SMP connectors used on the sample PCB are tall. By making the mount this thick, when wire-bonding only the mount comes into contact with the bed of the bonder – making bonding easier. The sample PCB is attached to the sample mount by two M3 button head screws. These screws were selected to match the metric fasteners used in the Janis cryostat. Specifically,

button head screws were used because of their low-profile which reduces the possibility of crashing the wire-bonder wedge into the fasteners during bonding. The sample is attached to the sample mount by means of a soluble glue. Once the glue is set, then the devices can be boned to the pads on the PCB surrounding the sample. There are a total of six signal pads, arranged in a ground-signal-ground pattern. With bonding complete, the sample mount is ready for installation into the cryostat.

The sample mount is attached to the pedestal in the cryostat by means of two M3 low-profile cap screws. Low-profile screws are used here since they require the same size Allen key as the button head screws used to mount the sample PCB – thus the same key can be used for both applications. The pedestal and sample mount each feature a flat on one side which must be aligned. Correct alignment ensures that the connectors are attached to the correct side of the PCB. The sample-end of the pedestal is notched in a similar manner to the sample mount to provide clearance for the cables. Once the sample mount is installed, the magnet can be mounted, and the cryostat closed. When the experiment is complete the same process is repeated in reverse, up until the sample is mount is detached from the pedestal. After removing the two screws attaching the sample mount to the pedestal, the sample can either be stored for later use, or the sample removed and the mount reused. The sample is removed by first carefully removing the wire-bonds. Then the PCB can be detached from the sample mount. The glue used to mount the sample can then be dissolved in the appropriate solvent and the sample removed. Finally, the sample mount can then be cleaned and reused.

5.2.3 Thermally insulated standoffs

The thermally insulated standoffs have two main functions, firstly they provide support for the magnet mount, and secondly they insulate the cold-head from the magnet mount. Each standoff is fabricated from a G10 tube and two stainless steel end caps. These materials were chosen for their particularly poor thermal conductivity at cryogenic temperatures [37]. The standoffs can be seen in figures 5-8 and 5-9, and an exploded view of the standoff is shown in figure 5-11.

Establishing a robust connection between the G10 tube and the stainless steel end caps was not trivial. The assembly is under constant tension when installed in the cryostat as it must support the weight of the magnet. While under tension, it experiences temperature swings of $\sim$ 290 K. As a result, there were concerns surrounding the use of an adhesive for this task. So it was decided to pursue a mechanical solution instead. Threading or similar means of attachment were not used as the standoff is relatively small, and such schemes would likely not have sufficient strength in the available G10 material. The chosen solution consists of machining the end caps to be a close fit within the G10 tube, then cross drilling for a 2 mm dowel pin. The dowel pin interferes with the location of the screw, and so the down pin itself must be cross drilled on one end, and cross drilled and tapped on the opposing end.

As the top end cap, shown in figure 5-11, is clearance drilled for the screw, the key for the screw needed to fit through the M3 tapped bottom hole. However, normal M3 cap screws utilize a 2.5 mm Allen key, which cannot fit through the 2.5 mm root diameter of an M3 threaded hole. Luckily, M3 low-profile screws utilize 2 mm drives which do fit through the M3 thread's root diameter. The drawback of low-profile screws is that they are not available in a vented variant, and since without a vent an enclosed cavity would be formed, which would create a virtual-leak, thus vented screws were desired. To overcome this, standard M3 low-profile cap screws were purchased and modified to include a vent.

The fabrication of all parts for this experimental apparatus were outsourced with the exception of the standoffs. As outsourcing the fabrication of the thermally insulated standoffs would be costly, we decided that we would fabricate them in-house. The fabrication of the standoffs was conducted as follows. The eight end caps were turned from 316L bar-stock. Each cap was machined with an axial hole diameter of 2 mm. Four G10 tubes were cut to length and finished. All parts were cleaned. The G10 tubes each had one end cap inserted into one end. The tube and end cap were then center drilled, pilot drilled and reamed to 2 mm. One 316L dowel pin was then inserted into each hole. As the parts were machines for an interference fit, a press was

used to drive the pins home, additionally a retaining compound was applied. While these parts were left to cure, four low-profile 316L M3 cap screws where modified. A mandrill consisting of a rod drilled and tapped to accept the screws was machines. Each of the four cap screws was mounted in the mandrill and a 1 mm hole drilled axially through the screw, thus venting the screw. When this was complete, we returned to the G10 tubes which has any excess retaining compound removed with a solvent. The tubes were then chucked in a lathe and indicated in. The dowel pin was itself cross drilled to 2.5 mm (M3 tap drill size). The hole was then tapped with a machine tap while still in the lathe. The parts were then removed and cleaned. A vented screw then inserted into the tube, and an end cap inserted into the other end. Like before, the tube and cap was cross drilled and a dowel pin inserted. Again, the part was chucked in the lathe and indicated in. This time the cap was drilled for 3.1 mm (M3 clearance drill size). The pars had any sharp edges broken, they were then cleaned, and ready for use in the cryostat.

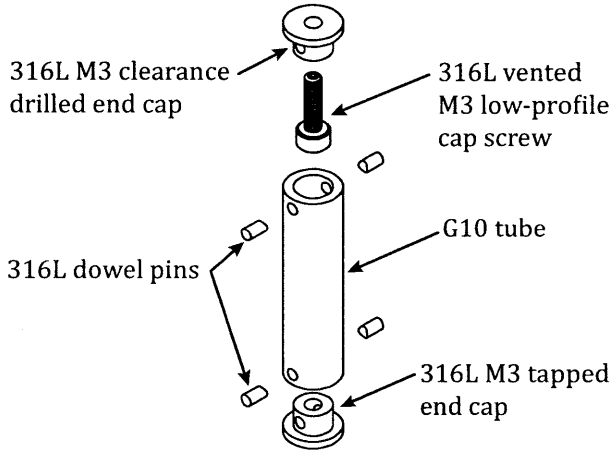

Figure 5-11: Exploded view of the thermally insulated standoffs. The standoff is constructed from a 3/8" G10 tube with a 316L stainless steel end cap pinned into place on each end. One end cap is clearance drilled for M3, and the other end is tapped for M3. Thus, the standoff provides good mechanical connection between the cold-head and the magnet mount while providing excellent thermal isolation. A vented screw is located within the G10 tube, which is used to attach the standoff to the cold-head. The screw is accessed by inserting an Allen key through the M3 tapped hole.

With the fabrication of the standoffs complete, this new experimental setup is complete. This setup will expand our experimental capabilities by enabling us to explore one more of the three main means by which superconductivity can be modulated namely current, temperature, and magnetic field. In addition to this, the interaction between the sample and optical stimulus can also be studied. Thus, the addition of this experimental apparatus our laboratory's infrastructure represents a great expansion in the type of superconducting phenomenon we are able to study.

Chapter 6

Conclusion and future work

For superconducting supercomputers to be realized, and fulfill their promise of extremely fast and very low-power operation, a scalable superconducting memory technology is required. In this work, we have presented two new superconducting memory technologies. Both designs leverage superconducting loops that encode the state of the cell in the magnitude and/or direction of a persistent current. In operating in this manner, these cells could be considered the dual of a traditional DRAM. Both of the presented cell designs utilize kinetic inductance, rather than geometric inductance. A result of relying on kinetic inductance is that the physical size of the cell can be scaled down with little change in the operation of the cell.

The first of the two memory designs, the NDRO cell, utilized a hTron to write to the cell and a yTron to read the state of the cell non-destructively. Tests of this cell have yielded very good error rates, with an ultimate BER predicted to be around 10^{-11}. The drawback of this design is the relatively large cell area, and the complex interface electronics required to form this cell design into an array. Due to these drawbacks, a new memory technology was pursued. The new DRO cell replaced the yTron in the NDRO cell with a second hTron, and as a result, the state of the cell is read out destructively. The DRO cell is deceptively simple, but as it turns out, is complex to design as there are many operating modes that could be used; however, forming the cell into an array is extremely simple. Results to-date show that the cell operates well; however, lower error rates are desired.

In terms of future work, we are currently testing new DRO designs that were designed by means of an automated optimizer and a new simulation package that we have developed in-house. These cells, if they live up to the simulator's predictions, should show very wide operating margins, and as a result it is hoped they will achieve lower error rates. When a DRO cell array that can operate with very low error rates is achieved, the next step would be to interface the cell with SFQ logic. Once the DRO array is integrated with SFQ, it is hoped that this technology will be used to enable the advancement of superconducting computing, and eventually lead to the development of a superconducting supercomputer.

Bibliography

[1] Dirk Van Delft and Peter Kes. The discovery of superconductivity. *Physics Today*, 63(9):38–43, 2010.

[2] Robert M Milton. A superconducting bolometer for infrared measurements. *Chemical reviews*, 39(3):419–433, 1946.

[3] Frank S Henyey. Distinction between a perfect conductor and a superconductor. *Physical Review Letters*, 49(6):416, 1982.

[4] John X Przybysz, Donald L Miller, Hannes Toepfer, Oleg Mukhanov, Jürgen Lisenfeld, Martin Weides, Hannes Rotzinger, and Pascal Febvre. Superconductor digital electronics. *Applied Superconductivity: Handbook on Devices and Applications*, pages 1111–1206, 2015.

[5] B.D. Josephson. Possible new effects in superconductive tunnelling. *Physics Letters*, 1(7):251 – 253, 1962.

[6] Bernd Seeber. *Handbook of applied superconductivity*, volume 2. CRC press, 1998.

[7] Anthony J Annunziata, Daniel F Santavicca, Luigi Frunzio, Gianluigi Catelani, Michael J Rooks, Aviad Frydman, and Daniel E Prober. Tunable superconducting nanoinductors. *Nanotechnology*, 21(44):445202, 2010.

[8] D. A. Buck. The cryotron-a superconductive computer component. *Proceedings of the IRE*, 44(4):482–493, April 1956.

[9] Andrew J Kerman, Eric A Dauler, William E Keicher, Joel KW Yang, Karl K Berggren, G Gol'Tsman, and B Voronov. Kinetic-inductance-limited reset time of superconducting nanowire photon counters. *Applied physics letters*, 88(11):111116, 2006.

[10] Adam N. McCaughan and Karl K. Berggren. A superconducting-nanowire three-terminal electrothermal device. *Nano Letters*, 14(10):5748–5753, 2014. PMID: 25233488.

[11] Adam N. McCaughan, Nathnael S. Abebe, Qing-Yuan Zhao, and Karl K. Berggren. Using geometry to sense current. *Nano Letters*, 16(12):7626–7631, 2016. PMID: 27960481.

[12] D. S. Holmes, A. L. Ripple, and M. A. Manheimer. Energy-efficient superconducting computing—power budgets and requirements. *IEEE Transactions on Applied Superconductivity*, 23(3):1701610–1701610, June 2013.

[13] Mutsumi Hosoya, Willy Hioe, Juan Casas, Ryotaro Kamikawai, Yutaka Harada, Yasou Wada, Hideaki Nakane, Reiji Suda, and Eiichi Goto. Quantum flux parametron: A single quantum flux device for josephson supercomputer. *IEEE Transactions on Applied Superconductivity*, 1(2):77–89, 1991.

[14] Naofumi Takagi, Kazuaki Murakami, Akira Fujimaki, Nobuyuki Yoshikawa, Koji Inoue, and Hiroaki Honda. Proposal of a desk-side supercomputer with reconfigurable data-paths using rapid single-flux-quantum circuits. *IEICE transactions on electronics*, 91(3):350–355, 2008.

[15] W Chen, AV Rylyakov, Vijay Patel, JE Lukens, and KK Likharev. Rapid single flux quantum t-flip flop operating up to 770 ghz. *IEEE Transactions on Applied Superconductivity*, 9(2):3212–3215, 1999.

[16] Alex F Kirichenko, Oleg A Mukhanov, and Darren K Brock. A single flux quantum cryogenic random access memory. In *Extended Abstract of 7th International Superconductive Electronics Conference*, pages 124–127, 1999.

[17] Alex F Kirichenko, Saad Sarwana, Darren K Brock, and Masoud Radpavar. Pipelined dc-powered sfq ram. *IEEE transactions on applied superconductivity*, 11(1):537–540, 2001.

[18] Konstantin K Likharev and Vasilii K Semenov. Rsfq logic/memory family: A new josephson-junction technology for sub-terahertz-clock-frequency digital systems. *IEEE Transactions on Applied Superconductivity*, 1(1):3–28, 1991.

[19] H Suzuki, N Fujimaki, H Tamura, T Imamura, and S Hasuo. A 4k josephson memory. *IEEE Transactions on Magnetics*, 25(2):783–788, 1989.

[20] Kiyoo Itoh. *VLSI memory chip design*, volume 5. Springer Science & Business Media, 2013.

[21] Kartik Senapati, Mark G Blamire, and Zoe H Barber. Spin-filter josephson junctions. *Nature materials*, 10(11):849, 2011.

[22] Valery V Ryazanov, Vitaly V Bol'ginov, Danila S Sobanin, Igor V Vernik, Sergey K Tolpygo, Alan M Kadin, and Oleg A Mukhanov. Magnetic josephson junction technology for digital and memory applications. *Physics Procedia*, 36:35–41, 2012.

[23] John F Bulzacchelli, William J Gallagher, and Mark B Ketchen. Hybrid superconducting-magnetic memory cell and array, June 26 2012. US Patent 8,208,288.

[24] Anna Y Herr and Quentin P Herr. Josephson magnetic random access memory system and method, September 18 2012. US Patent 8,270,209.

[25] Igor V Vernik, Vitaly V Bol'ginov, Sergey V Bakurskiy, Alexander A Golubov, Mikhail Yu Kupriyanov, Valery V Ryazanov, Oleg A Mukhanov, et al. Magnetic josephson junctions with superconducting interlayer for cryogenic memory. *IEEE Trans. Appl. Supercond*, 23(3):1701208, 2013.

[26] Andrew Murphy, Dmitri V Averin, and Alexey Bezryadin. Nanoscale superconducting memory based on the kinetic inductance of asymmetric nanowire loops. *New Journal of Physics*, 19(6):063015, 2017.

[27] Adam N McCaughan and Karl K Berggren. A superconducting-nanowire three-terminal electrothermal device. *Nano letters*, 14(10):5748–5753, 2014.

[28] Qing-Yuan Zhao, Emily A Toomey, Brenden A Butters, Adam N McCaughan, Andrew E Dane, Sae-Woo Nam, and Karl K Berggren. A compact superconducting nanowire memory element operated by nanowire cryotrons. *Superconductor Science and Technology*, 31(3):035009, 2018.

[29] L. Chua. Memristor-the missing circuit element. *IEEE Transactions on Circuit Theory*, 18(5):507–519, September 1971.

[30] J. P. Eckert. A survey of digital computer memory systems. *Proceedings of the IRE*, 41(10):1393–1406, Oct 1953.

[31] Karl K Berggren, Qing-Yuan Zhao, Nathnael Abebe, Minjie Chen, Prasana Ravindran, Adam McCaughan, and Joseph C Bardin. A superconducting nanowire can be modeled by using spice. *Superconductor Science and Technology*, 31(5):055010, 2018.

[32] Russell Cheng. *Non-standard Parametric Statistical Inference*. Oxford University Press, 2017.

[33] John Clarke and Alex I Braginski. *The SQUID handbook: Applications of SQUIDs and SQUID systems*. John Wiley & Sons, 2006.

[34] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[35] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[36] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[37] ED Marquardt, JP Le, and Ray Radebaugh. Cryogenic material properties database. In *Cryocoolers 11*, pages 681–687. Springer, 2002.