S. M. Faris
W. H. Henkels
E. A. Valsamakis
H. H. Zappe

Basic Design of a Josephson Technology Cache Memory

Design work on components for Josephson computer technology nondestructive read out cache memories has been published previously. In this paper, presenting a design for a 2.5-μm technology, 4 × 1K-bit cache chip with a nominal access time of about 500 ps as a basis, we show for the first time how these components are structured and interfaced. The cell, drivers, decoder, and a sense bus are based on designs which were experimentally verified in a 5-μm technology for which excellent agreement was found between computer simulations and measurements.

Introduction

Josephson logic circuits with an average loaded logic delay of ≈35 ps and a power dissipation of only 5 μW per gate [1] are being developed to prove the feasibility of a Josephson computer technology. In order to take full advantage of these circuits, a Josephson technology machine requires a memory hierarchy which, at the high end, matches the logic speed. At present two memory approaches are being investigated for a future hierarchy. Workers at the IBM Zurich laboratory are developing a DRO (destructive read out) main memory emphasizing density and low power [2]. Our work emphasizes speed; in this paper we are reporting on a basic design for a 4 × 1K-bit NDRO (nondestructive read out) cache chip with a nominal access time objective of about 500 ps in a 2.5-μm technology.

The criteria for the design of Josephson NDRO memory cells which store quantized persistent currents in superconducting loops [3-5] evolved over the years [6-9] and led to the experimental investigation of a 64-bit NDRO memory chip with an access time of 5 ns [10]. As a result of this design, and in part as a result of its shortcomings, novel cells [11], drivers, and decoders [12] for a 1.8-ns access time, 2K-bit, 5-μm technology cache chip were proposed and experimentally investigated. The cell, with a switching time of 120 ps, was optimized with respect to operating margins. The decoder (a so-called loop decoder) had a decoding delay of 30 ps per stage and also served the function of address registers, with a response time of 200 ps. The results of these investigations agreed well with computer simulations [13] and, based on the confidence obtained from the demonstrated accuracy of these models, the design of a 2.5-μm cache chip is now proceeding. The following sections of this paper contain details of that basic design.

General considerations

The basic design unit considered here is a 1K-bit array as illustrated in Fig. 1. Its components were carefully chosen to eliminate a number of complications which became apparent in the 5-μm design. The cell operates in the so-called 1-0 mode [14], storing a circulating current for a "1" and no circulating current for a "0." It has the advantage of operating with unipolar currents so that polarity switches can be eliminated. Also, it dispenses with the triple coincidence scheme [11] proposed in the 5-μm de-

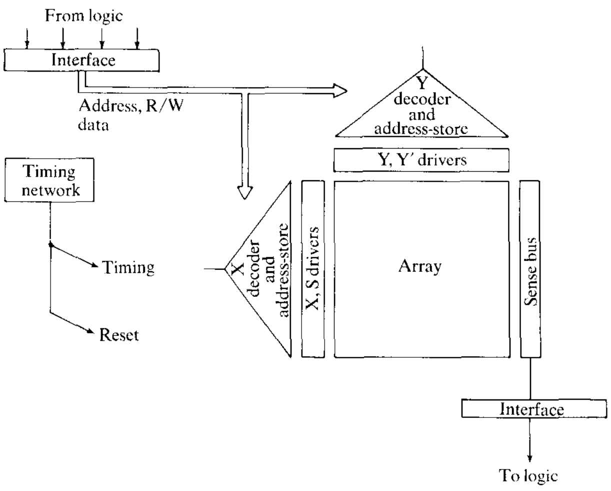

Figure 1 Block diagram of memory array. Addresses, function, and data are received through interface circuits and are latched in the decoders, which are subsequently triggered by timing pulses. During sensing, the sense bus collects the information and transmits it to the logic. At the end of the cycle the various memory components are reset.

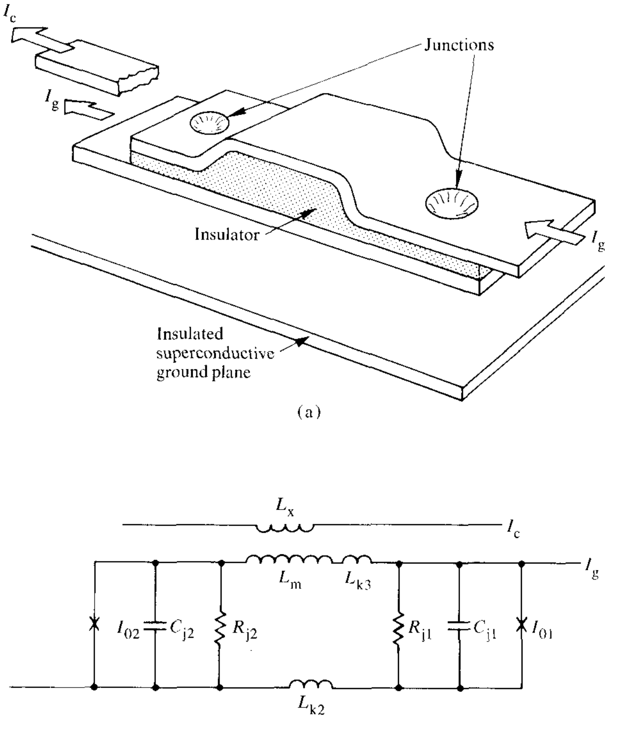

Figure 2 Asymmetrically fed, asymmetric two-junction interferometer, extensively used in most components because of its small size and low resonance steps: (a) perspective view; (b) equivalent circuit.

sign, so that the number of decoders is reduced from three to two. However, a number of design innovations, now experimentally verified, were required to widen the initially vanishing operating margins of that cell [15]. The decoders, which also act as address registers, are scaled-down versions of the experimentally verified 5-μm design, modified such that the requirement for both true and complement addresses is eliminated [16]. The read/write (R/W) and the data signals are entered essentially as addresses, along with the actual addresses, into the X and the Y decoders respectively. The read/write and "1"/"0" selections are then made in the last decoder stages. The sense bus, which collects information from any selected sense line, is composed of edge detectors which were tested experimentally [17]. These detect the decay of the sense-line current upon reading a "1" and transmit a signal to the memory-logic interface driver. The current levels in all components of the array were lowered to approach those used in logic [1]. This decreases component delay and eliminates the need for amplifiers in the logic-memory interface.

Since density is an important consideration in an array, most gates used in this design are bridge-type interferometers, illustrated in Fig. 2(a), rather than the planar interferometers used throughout logic [1]. Two-junction gates are preferred because of their size and simplicity, but in some cases three-junction gates are used to improve margins. The bridge device shown, an asymmetrically fed interferometer having two junctions of unequal size, is used to form sense gates, driver gates, decoder gates, and sense-bus gates. The equivalent circuit of the device is shown in Fig. 2(b). The two junctions, with their respective Josephson currents I_0, voltage-dependent tunneling resistances R_j, and capacitances C_j, are interconnected through the bottom and top electrode kinetic inductances L_{k2} and L_{k3} and through the magnetic inductance L_m [18], which is partially transformer-coupled to the control line inductance L_x. The maximum Josephson current through a two-junction interferometer is the sum of the maximum individual Josephson currents, $\sum I_{0i}$. Bridge interferometers cannot be conveniently damped, and large dc resonance steps of up to 0.6 $\sum I_{0i}$ may exist in the I-V characteristic of symmetric devices [19, 20]. In asymmetric devices, resonance steps are in general $<0.5 \sum I_{0i}$ because the asymmetry does not allow both device halves to convert the Josephson oscillation into dc efficiently. The theory of this behavior is still not well understood; however, its general characteristic has been observed experimentally and was studied in detail with an analog simulator [21]. In the present design, resonance steps can be large but should not exceed 0.5 $\sum I_{0i}$. The write gate in the cell is a center-fed, three-junction bridge interferometer [22], which may be visualized as composed of two devices of the type shown in Fig. 2(a) connected in parallel. Large resonances can be tolerated in the write gate because the cell was designed to operate properly even if the gate switches into a resonance [11].

With the exception of the memory-logic driver, the operation of all memory components is based on current transfer into and out of superconducting loops. This permits the powering of nearly all memory circuits with easily regulated dc-supply currents, thereby increasing operating margins. A typical loop is shown in Fig. 3. It contains a driver device (the circle) controlled by an input current I_c. The device is dc-powered in a series string isolated through resistors R_i, and connected in parallel is a loop represented by transmission lines with impedance Z_0 and delay τ. During switching, current I_p, originally flowing through the driver, is transferred into the loop, and is subsequently transferred back out through the switching of a reset device normally inserted into the loop at point C in Fig. 3. The current transfer has, to first order, the dynamics of a single current swing in a slightly underdamped parallel *LRC* network, where *L* is the inductance of the loop, *R* the total resistance seen across the driver, and *C* the total device capacitance. Through the external damping resistor R_D the dynamics are adjusted such that the device voltage reaches the resetting voltage $V_{\min}$ [23] at the moment at which the entire current I_p is transferred into the loop. The general equation describing such a system is difficult to handle in closed form, but it is possible to reach a good understanding of the transfer dynamics by examining a few special cases for which we shall assume $Z_0 = Z_{01} = Z_{02}$ and $\tau = \tau_1 = \tau_2$.

The fastest current transfer occurs when the driver impedance is resistive and matches the impedance $2Z_0$ of the transmission lines. This can be realized if the driver device has negligible total capacitance *C*, or if *C* is compensated by an inductive network. In this case the switching device transfers into the loop a current $I_L = 0.5I_p$. This current doubles at the shorted end and reflects back into the matched device, which resets when its current reaches zero. As a result the total current I_p is transferred in a time $t_0 = 2\tau$. Since the total loop inductance is $L = 2Z_0\tau$, the transfer time can also be expressed as $t_0 = LI_p/V$, where *V* is the device voltage during transfer. It is interesting to note that, because of the requirement for a return line, it takes at best two times longer to transfer current into a loop than into a terminated line such as that used in latching logic. Despite that fact, even considering that the speed penalty exceeds the factor of two in a realistic design, loop logic was chosen simply because operating margins require a control in array currents which cannot yet be achieved with an ac power supply. In our devices *C* is not negligible, and compensation requires a physically large inductor. For this reason we shall now discuss cases in which the device is not matched to the loop impedance.

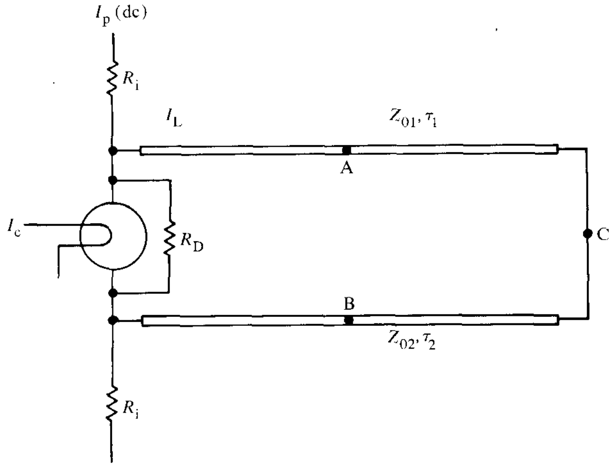

Figure 3 Typical equivalent circuit of superconductive loop; contains Josephson gates with damping resistors R_D to adjust the dynamics, and a loop composed of transmission line sections.

If the *RC* of the device is of the order of τ, the reflections in the lines are washed out and the loop can be represented as a parallel *LRC* circuit. This condition is satisfied in the cell, decoder, and sense bus loops. To obtain full current transfer, *R* (which is the parallel combination of R_j, R_D, and the power line load exclusive of Z_0) is adjusted through R_D to damp the circuit [6, 10-12] such that the device voltage $V = V_{\min}$ when $I_L = I_p$. This is achieved if the circuit is slightly underdamped, a condition given by $R = \xi[L/C]^{1/2}$, where ξ is a function of $V_{\min}$ and, in present designs, $\xi(V_{\min}) \approx 0.8$. The dynamics of such a circuit can be expressed analytically. If the device voltage is limited by the gap voltage V_g, one must break down the calculation of t_0 into three parts. These comprise the time required to reach the gap during which the subgap value of R_j is very large, the time required at the gap, and the time during which the voltage decreases from V_g to $V_{\min}$ [12]. For a given circuit, t_0 can be decreased by decreasing the current level. This ultimately causes the peak voltage to fall below V_g. But even if the voltage never reaches the gap, a decrease in I_m (the maximum Josephson current) and I_p still slightly decreases t_0. In this case the voltage scales as I_pR, and time scales with *RC*; thus $t_0 = nRC = n\xi[LC]^{1/2}$, where in the present design $n \leqslant 4$. It can be shown that if $\beta = 2\pi CR^2I_m/\Phi_0 >> 1$, where Φ_0 is the flux quantum (2.07×10^{-15} Wb), then $V_{min} \approx [\Phi_0 I_m/\pi C]^{1/2}$, *e.g.*, *cf.* [11]. Therefore, if I_m and I_p are decreased, $V_{\min}$ decreases at a slower rate than the

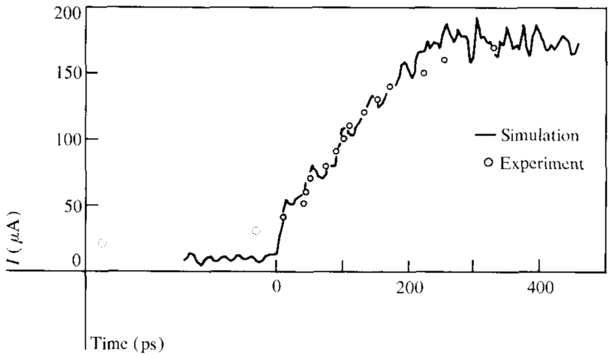

Figure 4 Illustration of the good fit between simulations and experiments in an extreme case.

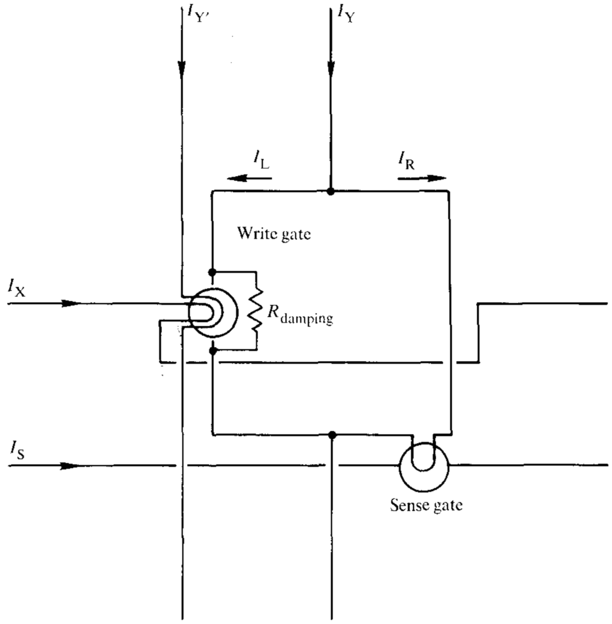

Figure 5 Schematic of the 1-0 mode NDRO memory cell.

device voltage, and resetting occurs at a relatively higher voltage and thus a slightly shorter time after the occurrence of the voltage peak.

The array loops, which carry the currents required by the memory cells, have relatively long delays τ compared to RC; thus individual reflections must be taken into account. In addition, $Z_{01} \neq Z_{02}$ and $\tau_1 \neq \tau_2$. Once the driving gate resets to the zero-voltage state, disturbances existing in the now shorted loop take a very long time to die out. For this reason a resistor R_m with an optimum value of $(Z_{01} + Z_{02})/2$ is placed across the center of the loop (points A and B of Fig. 3). For those cases for which the gap is reached, t_0 can be approximated to first order by $t_0 = (\tau_1 + \tau_2)\{[Z_{01} + Z_{02})I_p/NV_g] + 1\}/2$. In this relation, which in present designs agrees well with simulations, the additive factor of unity results from the inclusion of the aforementioned center loop resistor R_m, and N represents the number of series-connected gates used to drive the loop. Again t_0 decreases with I_p.

During the design phase, the above approximations can be used as initial guides; however, to complete a design it is necessary to carry out detailed simulations. In our models the device is represented by the equivalent circuit shown in Fig. 2(b) and the loop by a large number of lumped inductances and capacitances, one for every section of the line defined by the portion along which the line remains at the same height above the ground plane as it meanders vertically through the array. The accuracy with which such simulations can be made has been demonstrated [11, 12] and is again illustrated using an extreme case, as shown in Fig. 4. Here we see current as it is transferred out of a 5-μm technology sense line when the device switches at low current levels into a resonance step. In this case the dynamics are not only determined by macroscopic reflections but also by Josephson device oscillations having a fundamental frequency of the order of 150 GHz and correspondingly higher harmonics. The agreement with experiment is excellent except in the determination of the initial zero (dotted circles), which because of the small current levels is affected by a few tens of microamperes of noise in the test set-up.

Components

- *The cell*

As in nearly all previously reported Josephson NDRO cell designs, the basic cell is a planar superconducting loop which stores information in the form of persistent circulating currents. Writing and sensing of the information in a particular loop is controlled by the switching of associated Josephson gates. In order to select a single cell for writing or reading, we employ a scheme originally proposed by P. Wolf [14]. Figure 5 depicts the cell which is supplied by a current I_Y in a line that interconnects cells along a column of the array. Each loop contains a single write gate to which two orthogonal control lines, I_X and $I_{Y'}$, are coupled. A sense gate, supplied by I_S, has as its control the current flowing in the right-hand side of the loop. The binary "1" state consists of a clockwise circulating current I_{circ}; the "0" state is defined as zero loop current. We refer to this mode of storage as the 1-0 mode.

To write a "1" into an empty cell, a supply current I_Y is applied. Due to fluxoid conservation this current splits up as I_L and I_R into the left- and right-hand branches of the cell according to $I_L L_L = I_R L_R$ where L_L and L_R are, respectively, the loop left- and right-hand branch inductances, taking into account that L_L includes a small nonlinear inductance contribution due to the write gate [24]. After I_L has been established in the write gate, the two control currents I_X and $I_{Y'}$ are applied, forcing the write gate to switch, and causing I_L to transfer into the right-hand branch. Removal of all currents results in a clockwise circulating current $I_R = I_{circ} = -I_L = n\Phi_0/(L_L + L_R)$, where n is an integer. To write "0" into a cell containing a "1" simply requires the erasure, or dissipation, of I_{circ}. This is accomplished by coincident application of only I_X and $I_{Y'}$. The orthogonal control wiring scheme ensures that write gates in unselected cells are subjected to only I_X or $I_{Y'}$, or to no control current at all. With $|I_L| = I_{circ}$, the write gate is designed to switch only for a total control of $I_X + I_{Y'}$, and not for a control of I_X or $I_{Y'}$ alone.

As previously discussed, complete current transfer during writing can be ensured by connecting an appropriate external damping resistor across the write gate [11]. In such case the static sense operating region for the 1-0 mode is maximum when the branch inductances are equal [11], $L_L = L_R$. Contributions of I_Y to the cell branch currents are $I_Y/2$, and in the large flux quanta limit, $I_{circ} = I_Y/2$.

To read the information nondestructively, the cell is selected through coincidence of I_Y and I_S and the resultant current in the right-hand branch is detected by the sense gate. Application of I_Y adds a contribution to the right-hand branch current. If the selected cell contains a "1," this contribution adds to I_{circ} and the total control current causes the sense gate to switch. The sense gate is designed not to switch if the right branch contains only the I_Y contribution, or only I_{circ}.

There is one significant disadvantage of the 1-0 cell relative to $1,-1$ mode schemes which employ circulating currents of equal magnitude but of opposite sense for the binary states. For the $1,-1$ mode, the ratio of the selected sense-gate control level present during reading of a "1" to the unselected gate control levels, or to the selected control level present during the reading of a "0," can be made as large as 3. For the 1-0 mode this maximum "sense discrimination ratio" is only 2. Consequently, unless special provisions are made in the design, the sense operating margins for the 1-0 mode cell will be much smaller than the corresponding sense margins that can be achieved for $1,-1$ mode cells.

A second problem encountered in the design concerns resonances in the sense gate. Previous studies have shown that under certain conditions one can achieve wide sense margins by allowing operation into sense-gate resonances [11]. However, in the present design we desire to avoid switching into a sense-gate resonance because of the associated smaller driving voltage and a correspondingly slower current transfer. Two provisions have been made in order to compensate for the reduced sense discrimination of the 1-0 mode while at the same time ensuring a large enough relative value of I_S to avoid sense-gate resonances.

First, we exploit the insensitivity of I_{circ} to I_Y variations in a cell that stores only two flux quanta [11, 15]. In this case $I_{circ} = 2\Phi_0/(L_L + L_R)$, and I_Y may be varied by $\pm 25\%$ about its nominal value without causing a change in I_{circ}. The nominal control level in a selected sense gate during the reading of a "1" is $S_1 = I_{circ} + (I_Y/2) = (2\Phi_0/L) + (I_Y/2)$, where $L = L_L + L_R$. With the relative variance of I_Y, $\Delta I_Y/I_Y$, restricted to $\leq \pm 25\%$, the variance of S_1 relative to its nominal value is $(\Delta S_1/S_1) = [(\Delta I_Y/I_Y)^2 + (\Delta L/L)^2]^{1/2}/2$ or $\approx (\Delta I_Y/I_Y)/2$ for relatively small variations of cell loop inductances. In contrast, in the large-flux-quanta limit, $I_{circ} = I_Y/2$, $S_1 = I_Y$, and hence $(\Delta S_1/S_1) = (\Delta I_Y/I_Y)$, which means that the allowed variation in I_Y can be up to a factor of two larger in the low-flux-quanta limit.

Second, we exploit the characteristic of the asymmetric interferometer [25] illustrated in Fig. 2. For such a gate, the sensitivity of the steeper side of its threshold curve increases as the ratio I_{01}/I_{02} is increased above 1. Furthermore, at the same time, the ratio of $I_m(0)$, the maximum Josephson current with zero control current applied, to the maximum possible critical current, $I_{01} + I_{02}$, increases with increasing I_{01}/I_{02}. Thus for a given gate size, *i.e.*, given inductances, a more sensitive gate as well as a relatively larger I_S operating region is achieved with $I_{01}/I_{02} > 1$.

Following the above considerations and assuming the 2.5-μm technology parameters used in prior experimental memory cells [11, 15], but with a Josephson current density of $j_1 = 850$ A/cm^2, an optimized 1-0 mode cell design has been carried out. Due to sense-bus considerations current levels were constrained to be ≥ 0.2 mA.

The sense gate is a single-control two-junction bridge interferometer having the asymmetric in-line geometry shown in Fig. 2. Values of the various inductive components have been calculated using a numerical procedure [26] which includes the effects of magnetic field fringing, significant in a 2.5-μm technology. Similar calcu-

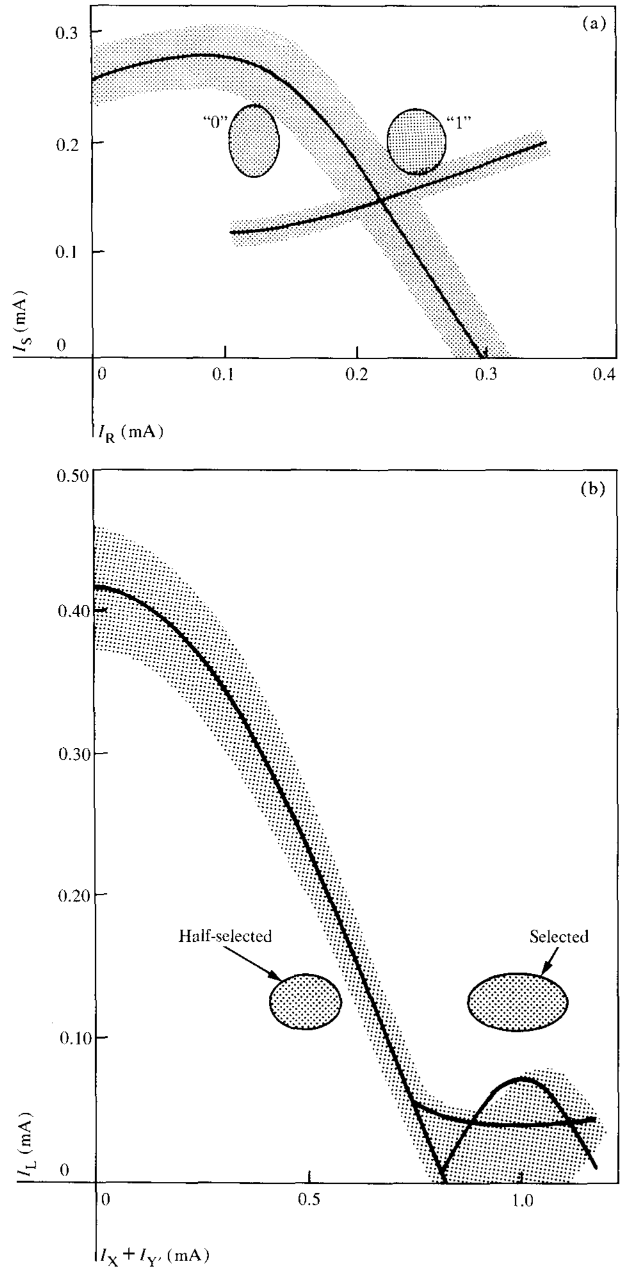

Figure 6 Operating regions: (a) of the sense gate, and (b) of the write gate of the cell.

lations have been found to agree with experimental results to within ±5%. The solid lines in Fig. 6(a) are the relevant portions of the central lobe and first side lobe of the nominal threshold curve for the sense gate; the dotted region overlapping the threshold curve is the envelope of all possible 3σ sense-gate threshold curves that could arise throughout the array, under presently assumed fabrication tolerances.

The write gate is a double-control three-junction bridge interferometer. For an optimum write operating region, the write-gate threshold curve must be symmetric, because I_L may be either positive or negative, whereas $I_X + I_{Y'}$ is always positive. For this reason the write-gate supply current is fed to the center junction, as in previous designs [11, 15]. The zero-control critical currents of the three junctions are in the ratio 1:2:1. The central lobe and first minor side lobes of the nominal write-gate threshold curve and the envelope of possible variations throughout an array are shown in Fig. 6(b).

The dotted ellipses in Figs. 6(a) and 6(b) are the respective static operating regions for sensing and writing. Those ellipses interior to the threshold curves correspond to the sense "0" case or half-selected gates; those exterior correspond to selected gates. For the sense-gate design described here, the lower limit to the operating region is set by the side lobe intersection rather than by a resonance peak. Designing the selected write-gate operating region to lie beyond the intersection of the central lobe with the $I_X + I_{Y'}$ axis ensures that the maximum Josephson current passes through zero. In such cases I_{circ} can always be erased (a "0" can be written) even if $1\Phi_0$ rather than $2\Phi_0$ is spuriously stored. Thus a cell cannot get stuck permanently in the $1\Phi_0$ state (this state is undesirable in this design because it would not be read properly).

The physical layout of the cell is similar but not identical to a single-flux-quantum cell that has already been demonstrated [15]. The present layout minimizes array-line to storage-loop coupling and incorporates the optimized gate designs discussed above. The resultant cell occupies an area of $63.5 \times 58.4\ \mu\text{m}^2$ and has a total nominal loop inductance of 33.1 pH.

Simulations of the cell yield a nominal current transfer time of ≈35 ps. The simulated optimum external damping resistor, 2.9 Ω, for an anticipated write-gate capacitance of 2.67 pF, is in good agreement with a previously established damping criterion [11]. For the small LI_{circ} value employed here, dynamic differences between switching through and switching into resonances are unobservable. The dynamic operating region during writing was found to be somewhat dependent upon the I_X and $I_{Y'}$ rise times, but was for all practical rise times found to overlap entirely the designed static operating region shown in Fig. 6(b).

With the technology assumed here, experimental nonoptimized 1- and 2-flux-quanta, 1-0 mode cells have been successfully fabricated and operated [15]. The stored energy in the experimental 1-flux-quantum cell was only 6×10^{-20} joules (it is 1.25×10^{-19} joules in the present design).

These cells clearly showed the expected quantization of I_{circ}, and the consequent large insensitivity of I_{circ} to I_{Y} variations. As an example, Fig. 7 shows the experimentally observed portion of the dynamic write threshold curve for a 1-flux-quantum cell. Quantization manifests itself as a step on the write "0" threshold. The step indicates that I_{circ} was independent of changes in I_{Y} over the range $I_{\mathrm{Y}} \approx 0.05$ mA to 0.20 mA. In the design described previously, for clarity, the quantization steps have not been indicated in Fig. 6(b); however, the static operating regions indicated by the ellipses take these steps into account.

- *Array lines and drivers*

Each set of array lines (X, Y, Y$'$ and S) is associated with a group of 32 series-powered drivers across which the array loops are formed, one branch of each loop being the active array line, the other branch being the return line, as shown in Fig. 8. In the case of the X and Y$'$ lines the active portions of the loops are controls to the write gates, whereas each Y line incorporates 32 series-connected memory cells. The drivers are controlled by the respective decoder outputs, and reset gates are provided to transfer the current back out of the selected loop when a global reset line is activated. In order to minimize interloop disturbs, the loops are isolated through resistors R_{i} in the power string. The dynamics of the loops are adjusted through damping resistors R_{D} and R_{D}' across the drivers and the reset gates. R_{D} differs from R_{D}' because of the additional damping which the drivers see as a result of the power line resistors R_{i}. Sustained reflections in the loops are dissipated with the help of center-connected resistors R_{m}, as discussed previously in the section on general considerations.

Once the cell is defined, the impedances of the loop branches are determined by the inductances and capacitances of interconnected line segments which are formed as the loop meanders up and down over or through the cells. Simulations have indicated that, with the present structure, many of these sections can be lumped into one or two inductive and one or two capacitive elements for each cell without any measurable difference in the simulation results. Hence the models for the loops have 32 to 64 lumped *LC* sections both for the array and for the return lines. Except for external resistors, dissipative elements are neglected because losses in the superconductors are totally negligible over the considered lengths [27]. Driver gates are in all cases appropriately designed asymmetric interferometers.

The X and Y$'$ lines have high impedances, requiring two series-connected driver gates in each loop to make the current transfer time consistent with the requirements for the cache. The input control rise time is sufficiently fast to ensure that both gates switch under our assumed process tolerances. The Y and S lines have relatively low impedances, so single drivers suffice. The simulated current transfer into and out of the X line is illustrated in Fig. 9 where the current is monitored at the beginning (left-

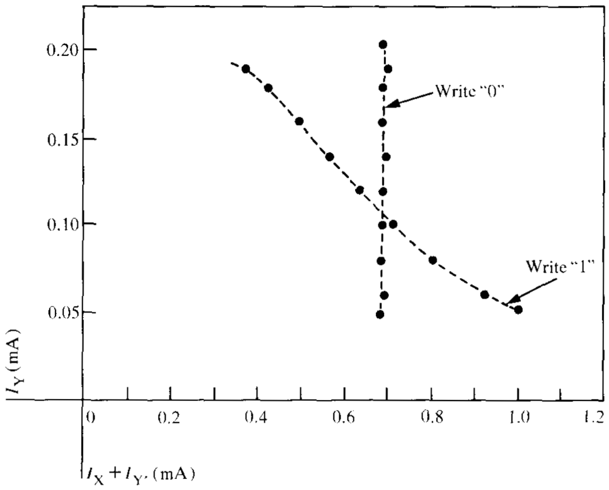

Figure 7 Measured write-gate threshold curve inferred from the operation of a single-flux-quantum cell.

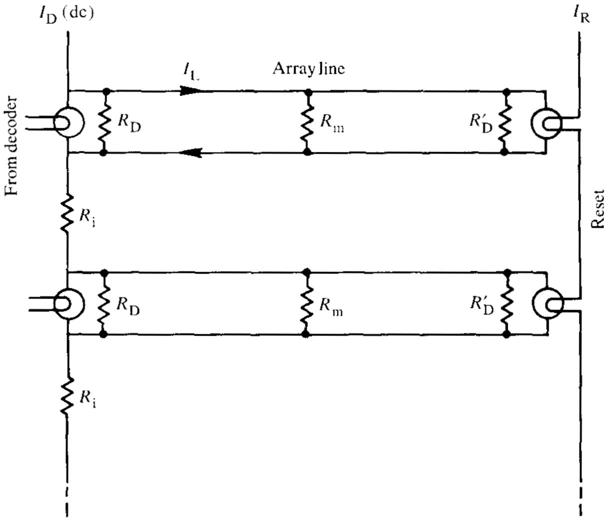

Figure 8 Schematic representation of a set of array line loops, each containing a driver and a reset gate.

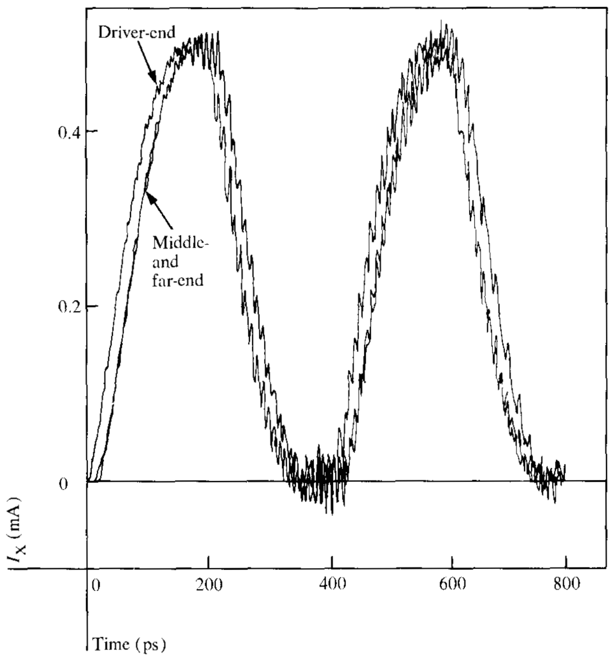

Figure 9 Simulation of current transfer into and out of the X line as monitored at the start of, the center, and the end of the loop.

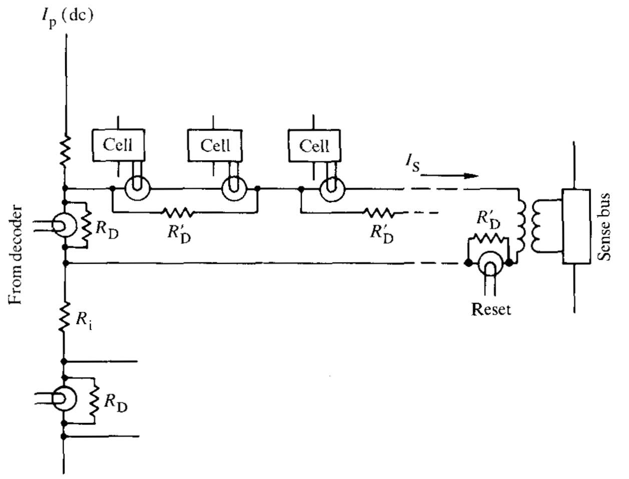

Figure 10 Schematic representation of a sense line.

most waveform), in the middle (after R_m), and at the end of the loop. The delay between the establishment of current near the driver and the other two points is primarily due to the center resistor R_m. We see that the current transfer in 160 ps is reasonably smooth and is well defined from cycle to cycle and that reflections and oscillations in the line are small.

The supply line resistors, $R_i = 40\ \Omega$, are chosen large enough so that during switching the voltage appearing across a driver feeds only a tolerably small common-mode disturb current into nearest-neighbor loops. A good estimate for this disturb current is obtained if the array line impedances are represented by resistors connected to ground and the active driver is represented by a voltage source. This gives a good measure of the peak disturb, which, however, occurs only during the rise and fall of the array current. Since the time of occurrence of this disturb is important, final simulations are always made of several series-powered loops. In all cases of the present design the nearest-neighbor disturbs are $\leq 10\%$ of the array current and next-nearest-neighbor disturbs are negligible.

The sense line, shown in Fig. 10, differs in some respects from the other lines. It contains 32 serially connected sense gates in each loop. During sensing, current is transferred into the selected line, and if the selected cell which receives the Y current contains a "1" the corresponding sense gate switches and transfers the current back into the driver. In contrast, if the selected cell contains a "0," the sense current is transferred back at a later time through the activation of the reset gate. The sense-bus element (described in a subsequent section) detects the negative transition of the sense current and transmits the information out of the array.

Since every sense gate will at one time or another act as a reset gate for the loop, it is necessary to adjust the dynamics of current transfer for each individual gate. This is accomplished by connecting damping resistors across sense gates, as shown in Fig. 10. Connecting the resistors across pairs of gates rather than across individual gates reduces the required number of resistors and simplifies the layout of the cell array, without causing adverse effects upon damping. Detailed simulations have shown that the presence of these resistors, which damp inductive sections of the line, is sufficient to dissipate reflections. Therefore, center resistors are not required in the sense lines.

The nominal array-current levels are $I_X = 0.50$ mA, $I_Y = 0.25$ mA, $I_{Y'} = 0.50$ mA, and $I_S = 0.2$ mA. Current spreads, including oscillations in the array loops, have amplitudes below the limits set by the cell operating windows shown in Fig. 6. This is achievable because the dc power line currents are externally controlled and can be held constant to better than ±4%. The current transfer times are 160 ps, 130 ps, 175 ps, and 100 ps for the X, Y, Y′ and S loops, respectively. These delays are consistent with our present design objectives.

- *The decoders*

The decoder used in this design is the so-called loop decoder which has been extensively tested in a 5-μm technology [12]. It is attractive because of its modularity of interconnected loops, its small size, and its high speed. Each stage of an n-bit decoder has two address loops, one for the true address bit and the other for the complement, as shown in Fig. 11. The true address loops contain two series address loop driver gates Q_A and a single reset gate Q_R. In contrast, the complement loops contain two address gates Q_A, one reset gate Q_R, and two additional series gates, Q_C, for automatic generation of the complement of the applied address bit [16]. The address loops are part of the decoder but also serve as address registers. To decrease the time required to latch the addresses, two series-connected drivers are used in each loop. The address loop drivers Q_A are connected into a series power string which contains isolation resistors R_i. As in all loops, properly chosen damping resistors are connected across the driver and the reset gates, and resistors connected across the center of all loops are also incorporated in this design as described in preceding sections. Neither of these damping elements is shown in Fig. 11. The decoding function itself is accomplished by connecting decoder loops serially into the address loops. Each decoder loop with inductance L_D consists of a decoder gate Q_D and a small resistor R_d. When a decoder gate switches, it produces a current pulse in the decoder loop which controls two decoder gates in the next stage. The last decoding stage contains 2^n decoder loops, one of which will deliver an output to a corresponding driver in response to an n-bit address code. However, if both the true and the complement address loops are activated in the last stage, two simultaneous outputs are available [28] as required in the Y and the Y′ line during the writing of a "1."

A two-bit decoder is shown in Fig. 11 to explain the principles of operation. At the beginning of the memory cycle, a set pulse is launched to activate the address gates Q_A in all complement loops, transferring currents $I_{\overline{A1}}$ and $I_{\overline{A2}}$. Assume that at a later time a binary address ($A_1 = 0$, $A_2 = 1$) is applied. Then the current $I_{\overline{A2}}$ in the upper complement loop is transferred back out and, simultaneously, the current I_{A2} is transferred into the upper true loop. The first stage (which stores address A_1) does not receive a signal, and is already in its proper state. Subsequently, decoding is initiated by means of a "decode start" pulse applied to both decoder gates in the first stage. Only the right-hand gate, which carries the current $I_{\overline{A1}}$, switches and delivers a current pulse I_a. The pulse I_a controls the gates in the two subsequent loops and switches that gate powered by I_{A2} which produces an output current pulse I_3, thereby decoding 1-out-of-4 outputs.

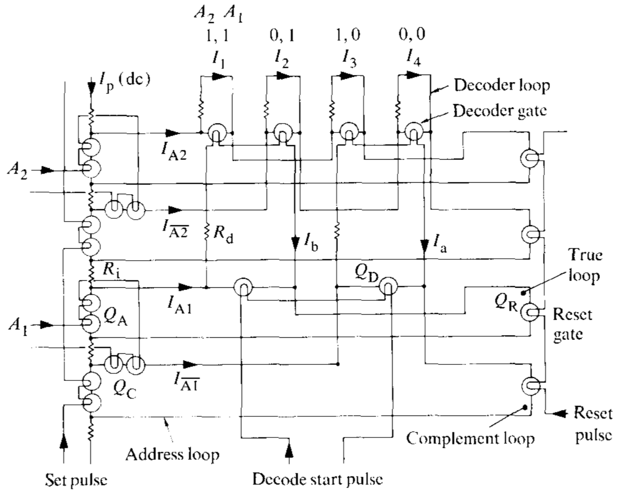

Figure 11 Two stages of the loop decoder used in this design. Addresses are locked into true and complement address loops before the decoder is triggered. Decoding itself is accomplished through interconnected decoder loops. Complement addresses are generated within the decoder itself.

If a sufficiently small R_d is chosen, the decoder gates self-reset. This allows the device voltage, after reaching its peak, to decrease below V_{min}, causing the device to lock back into the zero-voltage state. Hence, the currents I_a and I_3 are pulses which decay to zero with a time constant L_D/R_d. To obtain pulses with large amplitudes the decoder loops are designed to be heavily underdamped. This causes an overshoot such that the gate may not reset after the first voltage swing, but will reset during one of the subsequent oscillations. Care must be taken that this condition does not lead to multiple switching within the decoder. At the end of the memory cycle, a reset pulse is applied to all the reset gates Q_R, causing all loop currents to be transferred back into the address driver gates Q_A.

An optimum decoder design is achieved by balancing speed, density, operating margins, and power dissipation. The decoder gates Q_D were chosen to be the asymmetric end-fed, two-junction interferometers shown in Fig. 2. These are compact and have low resonance amplitudes, which results in acceptable margins. Their relatively low gain, however, represents a decoding speed disadvantage. On the other hand, the gates Q_A, Q_C, and Q_R are damped, planar-symmetric split-feed three-junction interferometers [29] having very low resonances, high oper-

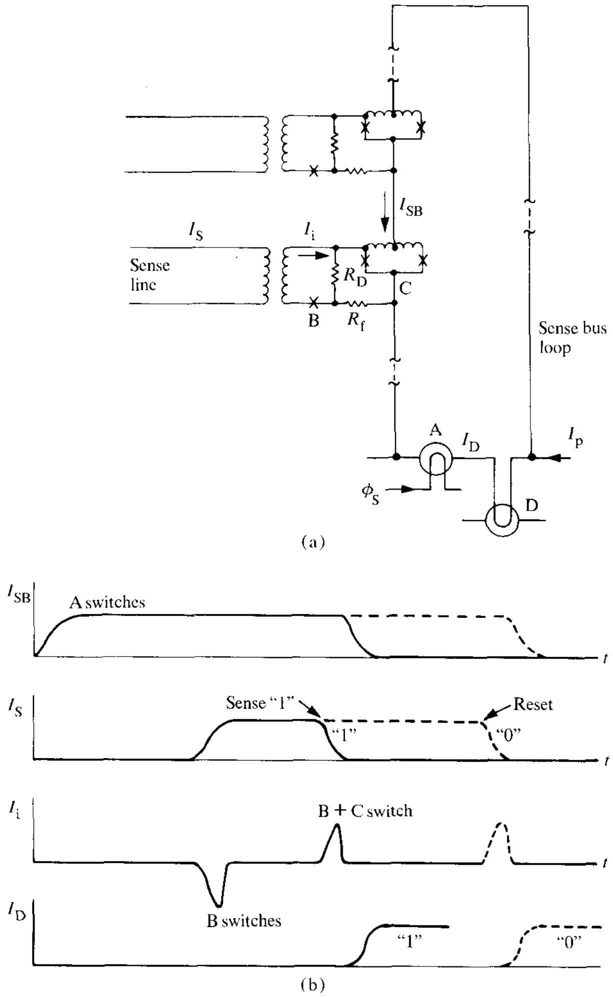

Figure 12 (a) Schematic of the sense bus, and (b) pulse sequence during sensing.

ating margins, and high gain. The choice of these planar devices is also dictated by disturb currents resulting from simultaneous switching of many gates in the address loop power string.

Simulations indicate that, in the 2.5-μm design, addressing delays of <100 ps and decoding delays of 20 ps can be achieved. For a 6-bit decoder this leads to a delay of ≈200 ps from the time the addresses are applied to the time the selected driver is activated.

- *The sense bus*

During sensing, the binary information stored in the selected memory cell is transferred into the corresponding sense line. The sense bus collects this information and sends it to a memory-logic interface circuit. Each sense line is transformer-coupled to a circuit responsive to the falling portion of the sense-line current I_S [17]. These circuits are serially connected to form a so-called sense-bus loop, as illustrated in Figs. 12(a) and (b). The loop is driven by a driver A, and the driver branch of the loop controls an interface gate D. Figure 12(b) shows the resulting signals.

Prior to sensing, the driver gate A is switched by a clock pulse ϕ_S, causing the current I_p to be transferred into the loop. This current activates the edge-detecting circuits which remain in the superconducting state. As the sense current I_S rises in the selected sense line, a negative pulse I_i is induced in the transformer secondary and flows through the small junction B and the interferometer C. This current in part cancels the current I_{SB} flowing through C and prevents it from switching. The junction B, however, switches as it reaches its threshold and permits flux to enter into the loop composed of the secondary inductance and the devices B and C. The extremely small bleeding resistor R_f prevents the circuit from becoming accidentally inoperative in a current circulating state, out of which it cannot switch under normal operating conditions. The damping resistor R_D provides damping within the circuit so that overshoots during switching are minimal.

The reading of a "1" causes the sense current I_S to decay. This produces a positive current I_i, which causes C to exceed threshold. Consequently, C and then B switch to the voltage state and transfer I_{SB} back into driver A where it switches device D, which in turn activates the memory-logic interface. If, on the other hand, the addressed cell contained a "0," the sense bus is activated only at the end of the cycle, at which time the sense line is reset as shown by dashed lines in Fig. 12(b). At this time the output pulse is not sent to the interface because gate D has been deactivated.

Simulations indicate that a current transfer time of 95 ps out of the bus can be achieved. The flat film transformers, constructed over a hole in the ground plane [30], have nominal primary, secondary, and mutual inductances of 30, 30, and 27 pH, respectively. The sense-bus inductance is 400 pH.

Although a full sense bus has not yet been experimentally tested, operation of an edge-detecting circuit was demonstrated with a 5-μm design in which R_f had been omitted [17].

- *Interfaces, timing, and resetting*

Because of the relatively low array-current levels, the logic-memory interface requires little amplification of logic control currents, thus greatly simplifying the design. Taking as reference the time of decode start, timing pulses which deactivate the memory-logic interface, control the resetting of the array loops, and set the sense bus can be derived from a pulse running down a delay line. The resetting circuits which are activated by such a pulse are essentially driver circuits with large fan-out. Further details concerning the interface, timing, and resetting circuits will not be discussed here because these components are being designed currently.

Memory operation

The interfacing of the various components just discussed was indicated in Fig. 1. Referring to that figure, the heart of the memory is a bit-organized array of memory cells wherein only one cell (one bit) is written or read in a given cycle or access. It is the function of the X and Y decoders to select, respectively, a single row and column of the array for activation. The intersection of the selected lines defines the selected cell. Drivers serve as amplifiers and buffers between the decoders and the array. When activated by the last stage of a decoder, a single driver out of the string of 32 rapidly supplies current to the selected cell. At the end of a cycle the driver lines must be reset; the decoders reset automatically. The "read/write" input to the memory determines whether the last stage of the X decoder selects an S (for reading) or an X (for writing) driver gate. Similarly, the "data" input, which contains the binary information to be written, is applied to the last stage of the Y decoder in such a manner as to cause both the Y and Y$'$ drivers of the selected column to be switched (write "1"), or only the Y$'$ driver to be switched (write "0").

During the read process if a "1" is read the subsequent transfer of current out of the selected sense line switches an edge-detecting circuit in the sense bus, thereby causing current to transfer out of the sense bus into a control for the memory-logic interface.

All powering of the memory is dc except for the memory-logic interface. The ac signals from logic which carry the "address," "read/write," and "data" information are converted to dc-powered signals by means of the logic-memory interfaces. Timing and resetting circuits ensure the proper sequence for the initialization of decoding, the deactivation of the memory-logic interface, the resetting of drivers, and the setting of sense-bus and complement address loops.

The amplitude of signals coming from logic is 0.110 mA. Upon arrival at the logic-memory interfaces the logic information is amplified to 0.330 mA through the interfaces and latched into the address loops of the decoder. Subsequently, decoding is started by a 0.15-mA pulse from the timing circuits. The output of the decoder is a 0.33-mA pulse which controls a driver, whose bias current level is 0.5, 0.25, or 0.20 mA respectively for X (or Y$'$), Y, or S drivers. The sense bus detects the 0.20-mA sense current and provides a 0.20-mA control level for the memory-logic interface, whereupon a 0.110-mA signal is sent to logic if a "1" is read.

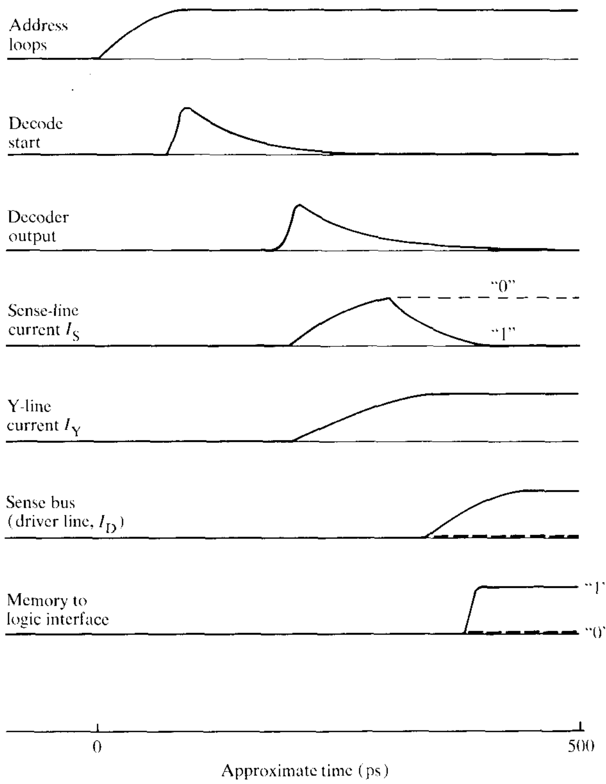

Figure 13 Approximate timing diagram of the principal currents involved during memory access (from simulations).

In any memory design, one of the main concerns is to minimize access time; the timing diagram illustrated in Fig. 13 gives an estimate of access time for the case here. As a time reference, we take the start of switching of the address loop drivers of the two decoders, which latch-in the X and Y addresses. At the same time, the R/W function is latched into the last stage of the X decoder. With the addresses, a timing pulse is received; it triggers the decoders as soon as current transfer into the address loops is completed. From this moment, events proceed asynchronously until the information is received by the memory-logic interface. First, the active decoder outputs trigger the selected sense- and Y-line drivers. Current transfer into the sense line proceeds faster than into the Y line, which ensures that the sense driver is reset before the Y current activates the sense gate of the selected cell storing a "1." The falling edge of the sense-line current is detected by the sense bus, which in turn triggers the interface circuit. If a zero is stored in the selected cell, the sense bus switches only after the sense line is reset; at this time the memory-logic interface circuit has been deactivated.

This timing diagram is based on nominal waveforms of the various array currents as obtained from computer simulations and, although the time required to switch the logic-memory interface has been omitted (10 to 20 ps), it indicates that a nominal access time of about 500 ps can be achieved. This time does not include on-chip propagation delays from the chip edges to the interfaces. It is noteworthy that the individual component delays are relatively balanced; no single component dominates the access time. The cycle time of the array, which includes the resetting of the various memory circuits, is larger by a factor of about two.

Summary

We have shown how the major elements of a very fast Josephson NDRO cache memory design are structured and are interfaced with one another. The discussion has centered on the performance of a 1K-bit array, four of which we expect ultimately to place onto a 6.35- × 6.35-mm^2 chip. The 2.5-μm design, used as an example for this discussion, is based on components which were tested in 5-μm designs. Because of the demonstrated accuracy of our models we are confident that, when tested experimentally, this cache memory will perform as anticipated.

References and notes

1. T. R. Gheewala, *IBM J. Res. Develop.* **24** (1980, this issue).
2. P. Guéret, A. Moser, and P. Wolf, *IBM J. Res Develop.* **24** (1980, this issue).
3. J. Matisoo, *Proc. IEEE* **55**, 2052 (1967).
4. W. Anacker, *IEEE Trans. Magnetics* **MAG-5**, 968 (1969).
5. H. H. Zappe and K. R. Grebe, *J. Appl. Phys.* **44**, 865 (1973).
6. H. H. Zappe, *IEEE J. Solid-State Circuits* **SC-10**, 12 (1975).
7. R. F. Broom, W. Jutzi, and Th. O. Mohr, *IEEE Trans. Magnetics* **MAG-11**, 755 (1975).
8. W. Jutzi, *Cryogenics* **16**, 81 (1976).
9. W. Y. Lum and T. Van Duzer, *J. Appl. Phys.* **48**, 1693 (1977).
10. W. H. Henkels and H. H. Zappe, *IEEE J. Solid-State Circuits* **SC-13**, 591 (1978).
11. W. H. Henkels, *J. Appl. Phys.* **50** (December 1979).
12. S. M. Faris, *IEEE J. Solid-State Circuits* **SC-14**, 699 (1979).
13. *IBM Advanced Statistical Analysis Program*, IBM Publication No. SH20-1118-0, available through IBM branch offices.
14. P. Wolf, *IBM Tech. Disclosure Bull.* **16**, 214 (1973).
15. W. H. Henkels and J. H. Greiner, *IEEE J. Solid-State Circuits* **SC-14**, 794 (1979).
16. S. M. Faris, *IBM Tech. Disclosure Bull.* **20**, 434 (1977).
17. S. M. Faris and A. Davidson, *IEEE Trans. Magnetics* **MAG-15**, 416 (1979).
18. M. Klein, *IEEE Trans. Magnetics* **MAG-13**, 59 (1977).
19. H. H. Zappe and B. S. Landman, *J. Appl. Phys.* **49**, 344 (1978).
20. H. H. Zappe and B. S. Landman, *J. Appl. Phys.* **49**, 4149 (1978).
21. D. B. Tuckerman, *Rev. Sci. Instrum.* **49**, 835 (1978).
22. H. H. Zappe, *IEEE Trans. Magnetics* **MAG-13**, 41 (1977).
23. H. H. Zappe, *J. Appl. Phys.* **44**, 1371 (1973).
24. Starting with an empty cell, the fluxoid conservation condition is $\phi + (2\pi/\Phi_0)\Phi = 0$, where ϕ is the superconducting phase difference across the write gate, Φ_0 is the flux quantum, and Φ is the applied magnetic flux. Let L'_L be the self-inductance of the left branch exclusive of the write gate; then $\Phi = I_L L'_L - I_R L_R$. Substituting this latter expression into the fluxoid condition and rearranging terms gives

    $$I_L\left(L'_L + \frac{\Phi_0}{2\pi}\frac{\phi}{I_L}\right) = I_R L_R.$$

    The total left branch inductance is

    $$L_L = L'_L + \frac{\Phi_0}{2\pi}\frac{\phi}{I_L}\,,$$

    where the last term is the write gate contribution. For a point junction this contribution reduces to

    $$L_j = \frac{\Phi_0}{2\pi}\frac{\sin^{-1} I_L/I_m}{I_L}\,.$$

25. H. Beha, *Electron. Lett.* **13**, 216 (1977).
26. L. E. Alsop, A. S. Goodman, F. G. Gustavson, and W. L. Miranker, *J. Comput. Phys.* **31**, 216 (1979).
27. R. L. Kautz, *J. Appl. Phys.* **49**, 308 (1978).
28. S. M. Faris, *IBM Tech. Disclosure Bull.* **21**, 3384 (1979).
29. L. M. Geppert, J. H. Greiner, D. J. Herrell, and S. Klepner, *IEEE Trans. Magnetics* **MAG-15**, 412 (1979).
30. P. C. Arnett and D. J. Herrell, *IEEE Trans. Magnetics* **MAG-15**, 554 (1979).

Received April 20, 1979; revised August 23, 1979

The authors are located at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598.