A Single Flux Quantum Cryogenic Random Access Memory

Alex F. Kirichenko, Oleg A. Mukhanov, and Darren K. Brock
HYPRES, Inc., 175 Clearbrook Rd., Elmsford, NY 10523, USA

Abstract—We report on the design of a superconductive Cryogenic Random Access Memory (CRAM). The 16-Kb RAM consists of four 4-Kb sub-arrays (blocks). It will have a 400 ps access time (latency) and a 100 ps cycle time (throughput). The input data and address are distributed using a high-speed RSFQ pipelined demultiplexer. The output data is collected with an RSFQ pipelined multiplexer. The entire 16-Kb RAM chip will dissipate 2.4 mW. We also discuss the projection for this design, using a future sub-micron fabrication process to achieve a 1-Mb capacity with a 40 ps throughput, required for HTMT (PetaFLOPS computing) project.

I. Introduction

The lack of fast Cryogenic Random Access Memory (CRAM) with sufficient capacity and high throughput has impeded the progress of superconductive electronics in digital applications. To date, the most successful superconductive RAM implementation was one from NEC [1]. The design approach used in the NEC memories combined SFQ memory cells and ac-powered voltage-state Josephson periphery circuits. The use of the large external ac-power limited the clock cycle to about 1 GHz, making the RAM throughput insufficient to match fast, dc-powered RSFQ logic. Until now, there have been no reported dc-powered RAMs.

As a fundamental constant, a quantum of magnetic flux is quite suitable for use as a data unit. The ability of flux quanta to be stored and transferred almost without dissipation allows the development of various circuits with internal memory and further to connect them into deeply pipelined devices. While RSFQ logic designs successfully exploit these features, RAM designs cannot fully use them.

The very idea of random access to a memory matrix contradicts a pipelined approach, because of the necessity to deliver select signals to random memory cells in a short period of time. The only way to transfer an SFQ pulse over a long distance at the speed of light is to use soliton transmission along a microstrip line; however, an SFQ soliton propagation is affected by unavoidable interaction with RAM cells. Alternatively, active Josephson Transmission Lines (JTLs) can be used to reproduce the dissipating SFQ soliton. However, JTLs are slower and take more space. Thus, an SFQ pulse is not suitable to perform the select process. This leads us to implement the traditional select scheme using dc-currents delivered via microstrip lines. However, we should avoid ac-powered voltage-state Josephson circuitry which causes problems in synchronization, power dissipation, and cross-talk. Thus, the main challenge of the CRAM design is to design voltage-state periphery circuits without the external AC powering.

II. CRAM Architecture

A. General Block Diagram

Among all known RAM approaches, the row-access memory architecture is the fastest. In semiconductor technologies, the row-access architecture is implemented using the fastest bipolar processes. This design simplifies the access by reducing the number of select lines, and allows operation on an entire block (or word) of data, providing fast parallel access.

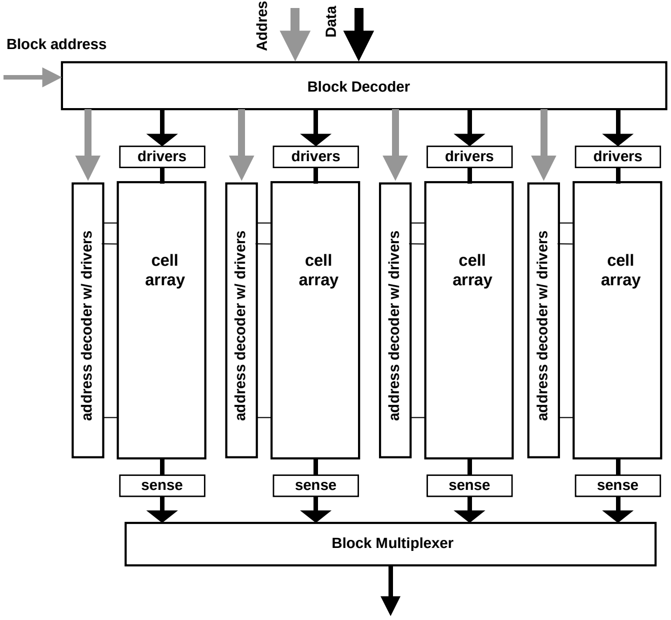

Fig. 1. A block diagram of the proposed RAM chip combining four blocks of memory arrays, block decoder, Y-decoders, select line drivers, sense gates, and output block multiplexer.

Our superconductive SFQ CRAM is constructed from SFQ memory cell arrays, dc/SFQ decoders, current drivers, sensing gates, and a block demultiplexer and multiplexer. The general structure of the RAM chip is in Fig. 1. In order to increase throughput, the 16-Kb RAM chip is divided into four 4-Kb sub-matrices (blocks). Each block comprises a 128 x 32-bit matrix having a row access. Each row of this matrix (seen in Fig. 2) contains a 32-bit word, which forms an accessible unit of data. A block demultiplexer distributes input data between blocks, while a block multiplexer (or merger) provides the output data.

Manuscript received May 7, 1999.
This work was supported in part by the US BMDO/NSA under contract # F49620-99-C-0018 and JPL/NSA under contract # 1205056.

The input to the RAM is a 42-bit sequence consisting of a 32-bit data word (for the WRITE operation only), a 9-bit address, and a 1-bit instruction (R/W) to indicate whether a value should be read written. The 9-bit address splits on two parts, - a 2-bit block (Y) address and a 7-bit row (X) address. The block demultiplexer sends a 40-bit data to a corresponding 4-Kb block.

B. A 4-Kb CRAM Block

Fig. 2 shows a more detailed schematic of the 4-Kb memory block. Access to memory cells is provided with magnetically coupled microstrip lines. Select line current drivers generate dc-signals that propagate along the microstrip line with the speed of light. A sensing gate converts a dc-current readout signal to an SFQ pulse.

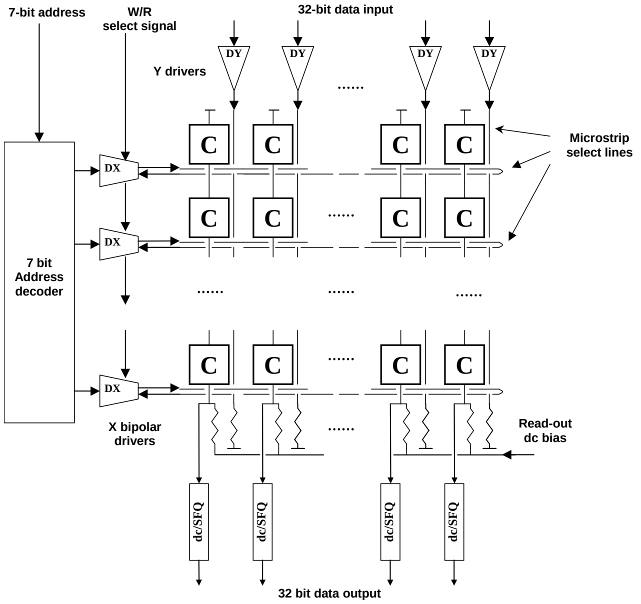

Fig. 2. Single block of RAM including an address decoder and drivers, a 32x128 memory cell array, X drivers, and output sensing dc-to-SFQ converter.

Each WRITE operation is preceded by an erase or WRITE0 operation. This function consumes an extra clock period to clear an entire row of the memory; but, at the same time, it allows us to simplify the overall RAM design and the operation cycle. The result is a higher integration scale and faster access time.

C. Memory Cell

We studied several SFQ cells for this memory. From these, we have chosen a modified version of VT memory cell [2] with non-destructive readout and current control (see Fig. 3). A single cell occupies an area of 40 x 45 μm^2. A 128 x 32-cell array of these cells occupies an area of 5.2 mm x 1.4 mm.

All read-out SQUIDs in a column are sequentially connected and biased with one dc-bias current. A sensing device is placed at the end of each of these columns. If the SQUID switches to the resistive state during the READ operation, the sensing device will detect a dc-voltage and transform it into an SFQ pulse.

Simulation shows excellent operating margins for this cell. The minimal critical current margin is 28%. The control current amplitude margins are above 30%. The DC bias current of the readout SQUIDs has 25% margins. In addition, the simplicity and reliability of this cell are very suitable for the large integration scale memories.

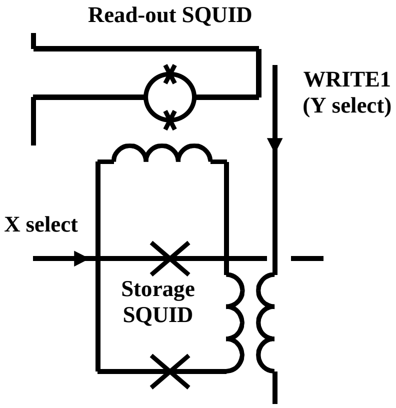

Fig. 3. A SFQ memory cell with dc-powered row-accessible selection.

This row-access architecture allows us to get rid of one extra select signal reducing the space occupied by the cell, while also solving the half-select problem and eliminating the need of bipolar Y-drivers. The memory cell access table for the RAM cell operation is seen in Table 1. The sign before the select line name indicates the control current direction.

TABLE I
MEMORY CELL ACCESS TABLE FOR NDRO RAM DESIGN

Operation	Select lines	Access
WRITE 1	+X+Y	Bit
WRITE 0	-X	Word
READ	+X	Word

D. Current Drivers

The main challenge in designing the current drivers is the necessity of confining them to a reasonable area. Due to the total physical RAM size, the size of the current drivers themselves is limited to 45 μm in width. We have already designed the layout of the SFQ/dc converters to meet this 45 μm condition. These converters are capable of operating at a 20 GHz data rate generating 0.3 mV output dc voltage. In combination with current amplifiers, these devices will supply sufficient drive for the select line current (~0.2 mA).

In contrast to Y-line, the X select line (see Fig. 2) requires bipolar current drivers. In this case, we have implemented a different approach. Specifically, unshunted Josephson junction based drivers are to be implemented here. We have considered both HUFFLE-based and relaxation-oscillation type circuits. We designed, fabricated, and tested an amplifier for the current driver. Fig. 4 shows a schematic of this circuit. The large inductance loop connects two relaxation-oscillation-driven pairs. The dc current from the current source is pushed into and out of the inductance loop, which is magnetically coupled to dc SQUID chain.

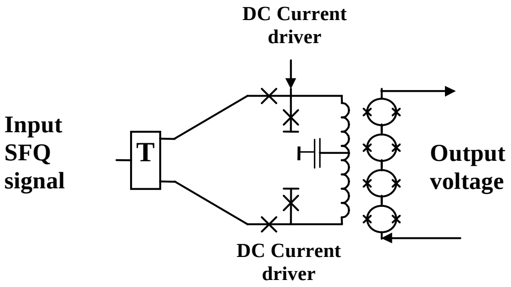

Fig. 4. Amplifier based on dc current drivers.

Fig. 5 shows successful test results of this amplifier used for the dc-current drivers. The driver demonstrates a gain of 15 at low-frequencies. In simulation, the driver worked at a 25 GHz frequency.

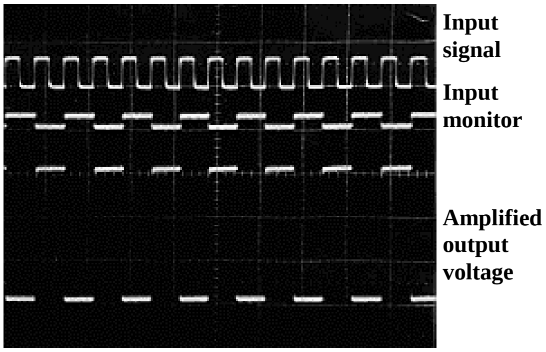

Fig. 5. Low-frequency test results of output amplifier. The output voltage amplitude is 3 mV, in contrast to the conventional SFQ/dc converter (0.2 mV).

E. Address Decoder

The address decoder is the most important part of the RAM. In our approach, the size of the decoder is critical. In previous research, we have designed and successfully tested compact decoder based on ac-powered voltage-state logic [3]. We redesigned it to dc-powered combination of voltage-state and SFQ logic.

This new decoder design consists of an address bus signal generator and 128 decoder cells (Fig. 6). The address generator transforms the address into a dc current dual-rail representation and transfers these currents to an address bus. The address bus consists of two groups of microstrip lines, seven lines in each.

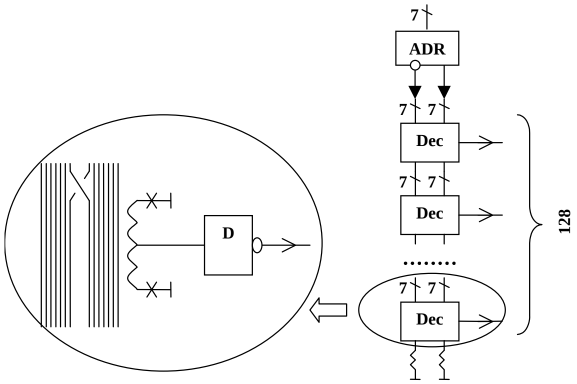

Fig. 6. A 7-bit address decoder.

Fig. 7 shows a single cell of the decoder. Each cell in the decoder has its own unique combination of direct and inverted lines. This configuration is achieved by swapping two different lines in every cell. The right seven lines are magnetically coupled to a SQUID which performs the OR operation. If a dc current persists in any of these 7 lines, the SQUID generates an SFQ pulse. Thus, any address, only one cell among 128 will have no currents in the right group of lines. So, by inverting the cell's output we get an output signal corresponding to address.

The address generator exploits single-ended current drivers to transform addresses from SFQ to dc-current representation. As a result, the address moves down in each block of RAM with the same delay as the data (see Fig. 2). This property might allow us to organize a ballistic pipeline structure, improving the RAM throughput.

F. CRAM Parameters

A major feature of our approach is an access time that is several clock periods, implemented by a pipeline structure. As in Fig. 1, there are three major parts of the RAM: the block decoder, the RAM blocks, and the block multiplexer. Each of these works independently and can be considered as a pipeline stage. The address decoder (Fig. 2) works in two clock cycles. All together this produces a four-clock-cycle pipeline structure of the RAM.

The speed of signal propagation along a microstrip line is close to the speed of light. For a microstrip line with SiO_2 insulator, it is ~ $6{\cdot}10^7$ *m/s.* The size of the memory array (128 x 50 µm ~ 6 mm) gives us a 100 ps delay time, which compounds to a 10 GHz clock rate (or throughput). The combination of SFQ and current-loop representation allows us to synchronize data flow in different pipeline stages. The travel times of all paths in a single block of the RAM (Fig. 2) are equal to 100 ps, while the delays of all cells and circuits, described above, are less than 20 ps. This might allow us to increase throughput by organizing pipeline access within a 100 ps time interval.

Thus, this 16-Kb RAM design has a four-cycle pipeline structure with 10 GHz throughput and 400 ps access time.

The projected size of the complete RAM chip would be 1 cm x 1 cm and the power dissipation about 2.4 mW. The entire design is estimated to require less than 60,000 Josephson junctions.

III. HTMT Implementation

HTMT (Hybrid Technology Multi-Threaded) architecture has been proposed for a petaflops computer [4]. The HTMT architecture combines semiconductor, optical, and superconductor technologies in a single-system structure. This is essentially a shared-memory architecture employing liquid-helium-cooled superconductive processors and data buffers (CRAM), liquid-nitrogen-cooled SRAM semiconductor buffers, a semiconductor DRAM main memory, and optical holographic storage. The superconductor CRAM is a buffer for the SRAM, which is itself a buffer for the DRAM. Superconductor processors will be ready to read a local CRAM in 10 clock cycles, but latency for reading a location in semiconductor SRAM will exceed 500 cycles. Therefore, the processors must be able to access the CRAM, but not the SRAM.

Previously explored solutions for cryogenic memories have been inadequate for operation at the petaflop level. Scaling down the fabrication process linewidth from the present 3.5 μm to 0.8 μm, will enable us to meet the petaflop CRAM chip requirements.

Table II shows the comparison between what we are capable of doing now and what is expected from CRAM in the HTMT project. We scaled the minimal Josephson junction size down to 0.8 μm (presumed fabrication process for HTMT project) and estimated the characteristics for the CRAM. In order to provide the shared memory access, the HTMT CRAM has to have extra atomic operations in its instruction set. The most convenient for our architecture is a SWAP operation, which comprises the combination of READ and WRITE.

To provide an interface to the external processes, the CRAM will have some additional data packet forming logic.

TABLE II
SPECIFICATIONS FOR THE HTMT CRAM VS. THE 16 KB CRAM

Specs:	This CRAM	CRAM for HTMT
Capacity	16 Kb	1 Mb
Word	32 bit	64 bit
Access time	400 ps	330 ps
Cycle time	100 ps	30 ps
Number of blocks	4	256
Number of cells in block	128x32	64x64
Number of JJs	60K	4M
Cell size	50x50 μm^2	10x10 μm^2
Chip size	1x1 cm^2	2x2 cm^2
Minimal JJ size	3.5 μm	0.8 μm
Atomic operations	W0, W1, RD	W0, W1, RD, SWP

IV. Conclusion

We have presented the fully dc-powered design of a 16 Kb RAM based on combination of dc-powered voltage-state and SFQ elements. The RAM will occupy a 1 cm x 1 cm chip, dissipate 2.4 mW power, and have a 400 ps access time and 100 ps cycle time (embodying four pipeline stages).

Acknowledgment

We are thankful to Prof. K. Likharev of SUNY at Stony Brook for helpful discussions.

References

[1] S. Nagasawa, Y. Hashimoto, H Numata, and S. Tahara, "High-frequency clock operation of Josephson 256-word x 16-bit RAMs," *IEEE Trans. Appl. Supercond.,* vol. 9, 1999. (in press).

[2] S. Nagasawa, Y. Hashimoto, H Numata, and S. Tahara, "A 380ps 9.5mW Josephson 4-Kbit RAM operating at a high bit yield," *IEEE Trans. Appl. Supercond*, vol. 5, p. 2447, 1995.

[3] P.F. Yuh, "A 2-kbit superconducting memory chip," *IEEE Trans. Appl. Supercond.,* vol. 5, p. 3013, 1993.

[4] M. Dorozhevets, P. Bunyk, D. Zinoviev, and K. Likharev, "PetaFLOP RSFQ system design," *IEEE Trans. Appl. Supercond.*, vol. 9, 1999.