Demonstration of Superconducting Memory for an RQL CPU

Randall Burnett, Ryan Clarke, Timothy Lee, Harold Hearne, Jacob Vogel, Quentin Herr, Anna Herr
Northrop Grumman Corporation
{randall.burnett, ryan.clarke, timothy.lee4, harold.hearne, jacob.vogel, quentin.herr, anna.herr}@ngc.com

ABSTRACT

Recent efforts to develop cryogenic processors are based on the elemental superconductor, niobium, and function at 4K. Since conventional silicon-based solutions will not operate at this temperature, new niobium-based memory elements are required. To that end, here we present two novel superconducting memory cells, both based on the Reciprocal Quantum Logic (RQL) non-destructive read-out (NDRO) gate.

CCS CONCEPTS

• **Hardware** → Superconducting circuits • **Hardware** → Memory and dense storage

KEYWORDS

Cryogenic memory, JTL, RQL, superconducting

1 INTRODUCTION

Conventional computing systems are increasingly limited by power consumption. One potential path toward meeting demand involves cryogenic superconducting technologies, including RQL. Logical circuits of increasing complexity have been demonstrated in this superconducting technology, but cryogenic memory continues to be a challenge to realizing large-scale superconducting systems. Specifically, recent efforts to demonstrate 8-bit and 16-bit RQL-based CPUs require integrated memory that functions at 4K to serve as the register file, instruction memory, and data memory.

2 DESIGN

The first design is a single-write, single-read memory cell that is intended for high-density layout. It is fully integrated into the RQL architecture [1], therefore inputs and outputs are accessed via Josephson Transmission Line (JTL) [2]. A schematic for the core design is shown in Fig. 1.

The memory cell functions fundamentally as an RQL D-Latch cell [3], meaning that it accepts a "data" input and a "logical clock" or "write enable" input. Together, these set the state of a single junction to either a "0" or "2π" corresponding to a logical memory state of "0" or "1," respectively. When in a "2π" state, the readout junction is forward biased, allowing it to be triggered by an incoming pulse. In a "0" state, the readout junction is unbiased, making it impossible for an incoming pulse to trigger. Readout is performed when an RQL pulse is delivered to the NDRO (read) – when the memory is storing a logical one, the pulse will propagate to the output pin; when the memory is storing a zero, the pulse will be blocked. By constructing a cell in which the input signals are split off and propagated to the next cell in the word or bit and the output signal is logically ORed with each sequential cell in its bit line, an array of memory can be created.

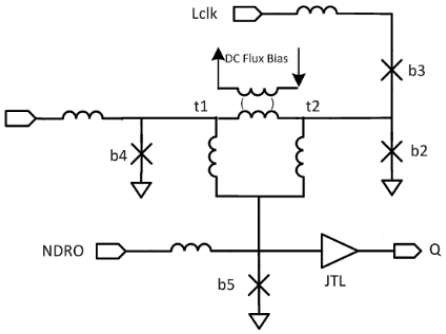

Figure 1: Single-Read NDRO Cell.

With density in mind, this cell was laid out within a tile structure, utilizing approximately 90% of each tile's bias and junction resources. The memory cells were arranged in small 2x2 arrays using a vine/tree structure for signal distribution and filling an area of 66 μm x 88 μm (33 μm x 44 μm for a single cell). The 2x2 layout will be shown in the presentation.

The second memory cell presented here was designed for performance over density. It uses the same basic NDRO structure, but it has been optimized to allow for same-cycle writes and reads. Additionally, a second read port was added (Fig. 2), making this version appropriate for use in a CPU register file. Due to the more aggressive performance requirements, target density was relaxed. The final 2x2 structure utilizes around 75% of available bias and junction resources, and comprises an area of 110 μm x 110 μm, or 55 μm x 55 μm for a unit cell.

MEMSYS, October 1–4, 2018, Old Town Alexandria, VA, USA

ACM ISBN 978-1-4503-6475-1/18/1
https://doi.org/10.1145/3240302.3270313

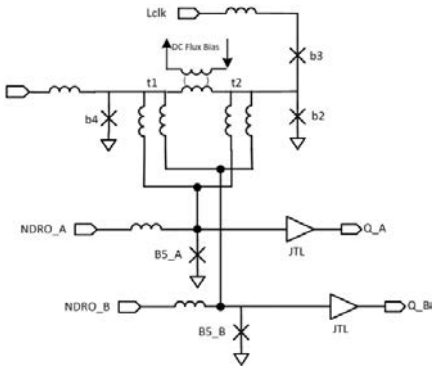

Figure 2: Dual Read NDRO Cell

Because both variants of the NDRO memory cell are embedded with JTLs, they require RQL clocks at specified phasing [1]. As the 2x2 blocks are arrayed into larger groups, the phase of subsequent groups can be incremented to account for signal propagation delay across the array at a given frequency. Using this approach, the first NDRO memory element was arrayed into a 64-word x 16-bit memory bank. This bank was integrated with a 64-bit RQL decoder for word selection, and latency for array access is 2.5 cycles at 2 GHz. Multiple banks can be parallelized within a CPU to achieve a desired memory capacity. Likewise, the dual-read NDRO cell was arrayed into an 8-word x 16-bit register file, with integrated data-line multiplexers, for use within the context of a CPU. Register file access latency is 1.25 cycles at 2 GHz. Both layouts will be shown in the presentation.

3 TEST RESULTS

Both the 64x16 memory bank and the 8x16 were fabricated with integrated deserializer/serializer test wrappers at the D-Wave foundry's 0.25 µm process, and they both demonstrate correct functionality at a 2 GHz clock frequency. Note that this operating frequency is dictated primarily by the current D-Wave process; the same designs are expected to function at 5 GHz in other processes. Figure 3 shows an oscilloscope screen capture of a test pattern, demonstrating a successful write and read on a single word. The screen capture is superimposed of a simulated plot of the same pattern, indicating hardware/model agreement. A similar oscilloscope screen capture for the register file, demonstrating two separate write and read commands to a single word will be shown in the presentation.

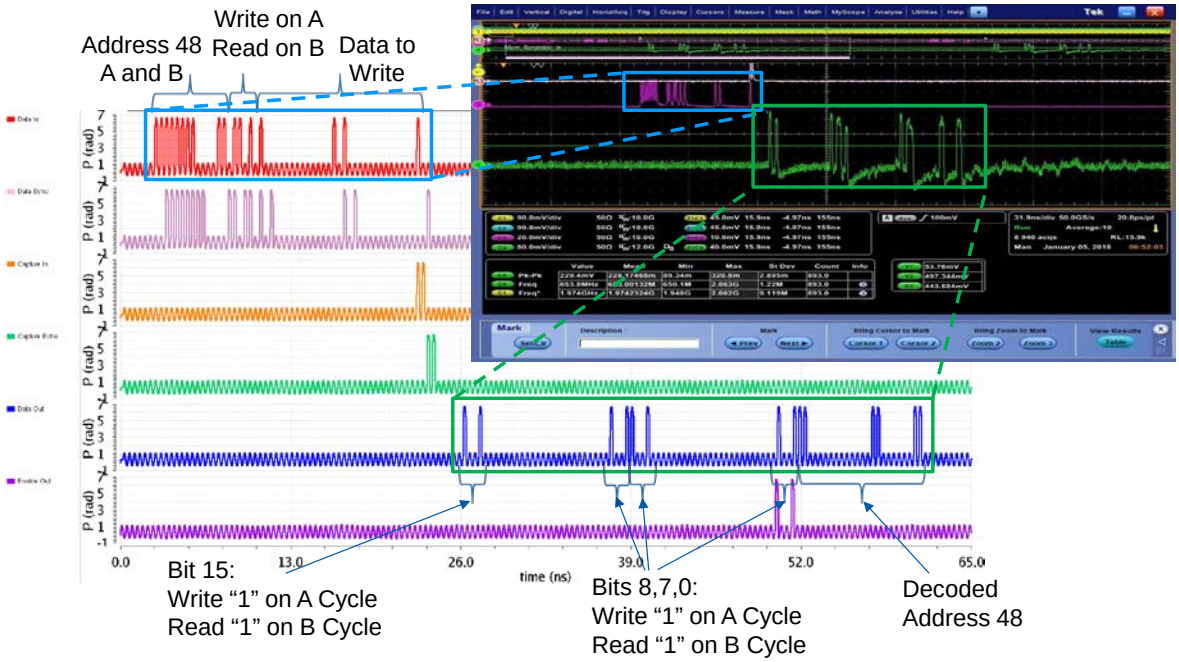

Figure 3: Oscilloscope screen capture demonstrating functionality of 64x16 NDRO memory bank, superimposed over simulated results from the same vector. Note that the simulation is RZ-encoded, while the oscilloscope output is NRZ-encoded, but the results are logically identical.

In addition to spot-checking, code was written to generate complex, full-array-testing vectors spanning tens of thousands of cycles. These vectors include checkerboard/inverse checkerboard, walking ones, and walking zeroes, and have been applied to both arrays. All bits have been shown functional at a single operating point, though on some cooldowns, flux trapping introduced sparse bit errors. Back-to-back writes and reads have been demonstrated

at speed at 2 GHz, with same-cycle writes and reads demonstrated on the register file.

4 CONCLUSION

In conclusion, two functional designs of first-of-its-kind cryogenic memory are presented above. These include a 128-byte bank of main memory, which can be scaled to multiple banks, and a 16-byte dual-read register file. They exhibit the low-power consumption and high efficiency characteristic to superconducting RQL circuits [1]. Both designs have demonstrated correct operation in hardware at 4K and can be integrated into future low-power, cryogenic processors.

ACKNOWLEDGMENTS

This research is based upon work supported by the ODNI, IARPA, via ARO contract number W911NF-14-C-0115. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government.

REFERENCES

[1] Herr, Q.P., Herr, A.Y., Oberg, O.T. and Ioannidis, A.G. *Ultra-low-power superconductor logic*. Journal of Applied Physics 109(10), 103903. (2011).

[2] Oberg, O.T. *Superconducting logic circuits operating with reciprocal magnetic flux quanta*. PhD dissertation. University of Maryland, College Park, Maryland. (2011).

[3] Herr, A.Y., and Herr, Q.P. U.S. Patent No. 9,455,707. Issued September 27, 2016.