Single-Flux-Quantum Cache Memory Architecture

Koki Ishida[1], Masamitsu Tanaka[2], Takatsugu Ono[3], and Koji Inoue[3]
[1]Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
[2]Department of Quantum Engineering, Nagoya University, Nagoya, Japan
[3]Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan
[1]Email: koki.ishida@cpc.ait.kyushu-u.ac.jp

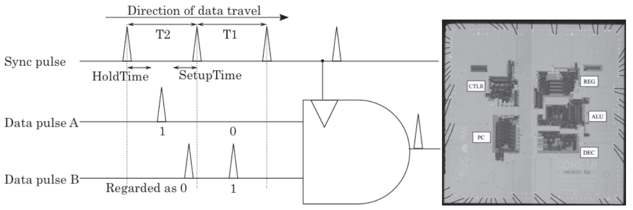
Abstract— **Single-flux-quantum (SFQ) logic is promising technology to realize an incredible microprocessor which operates over 100 GHz due to its ultra-fast-speed and ultra-low-power natures. Although previous work has demonstrated prototype of an SFQ microprocessor, the SFQ based L1 cache memory has not well optimized: a large access latency and strictly limited scalability. This paper proposes a novel SFQ cache architecture to support fast accesses. The sub-arrayed structure applied to the cache produces better scalability in terms of capacity. Evaluation results show that the proposed cache achieves 1.8X fast access speed.**

Keywords; single flux quantum; cache memory; shift register

I. Superconductive Computing and Its Problem

CMOS microprocessors have been faced with a limitation of clock speeds because of increasing computing power, i.e., known as "power-wall problem". Single-flux-quantum (SFQ) devices and circuits are promising to solve the problem due to its ultra-fast-speed and ultra-low-power natures. SFQ circuits use superconducting devices, namely Josephson junctions (JJs) to process digital signals [1]. In SFQ logic, information is stored in the form of magnetic flux quantum and transferred in the form of picoseconds-duration SFQ voltage pulse. Unlike CMOS designs, it operates in pulse logic fashion, and two types of signals are used: "*sync pulse*" and "*data pulse*", and SFQ logic gates recognize input signal level as '0' or '1' by means of examining the existence of a data pulse between two consecutive sync pules. Fig. 1(a) shows an operation example of an SFQ AND gate. Here, *data pulse A* at time *T2* and *data pulse B* at time *T1* are stored inside of the SFQ gate as logical input level of '1'. If no pulse appears as *data pulse A* at *time T1*, it is stored as '0'. This means that each logic gate supports latch function that is a unique feature of SFQ logics compared to conventional CMOS gates. *HoldTime* and *SetupTime* are conditions that have to be satisfied to ensure correct operations.

Some SFQ microprocessors have so far been demonstrated and one of them, called CORE1β (Fig. 1(b)), successfully operated over 25 GHz [2][3][4]. In the design, an SFQ L1 cache has been prototyped in order to realize high-speed memory [3]. The cache uses an SFQ shift register [5] and operates in bit-serial fashion, as does the microprocessor core, to reduce hardware complexity. However, such bit-by-bit fine-grained operations make the cache access time much longer, resulting in poor microprocessor performance. In addition, the large scale of the selector logic used to pick up referenced data strictly limits the scalability of the cache's capacity. Since several shift register based SFQ memories have so far been proposed [3][6][7], two serious problems exist in the traditional cache design. Fig. 2(a) illustrates a high level model of existing SFQ cache memory. First, its bit-serial access, i.e., reading bit-by-bit makes extremely cache access time longer, e.g., 64 times bit reads are required to perform 64-bit word load. Second, its scalability in terms of cache capacity is quite low. This is because the scale of multiplexer to read a data prohibits increasing the number of cache entries.



(a) Operation of SFQ AND gate. (b) CORE 1β.

Figure 1. An SFQ AND gate and microphotograph of CORE 1β.

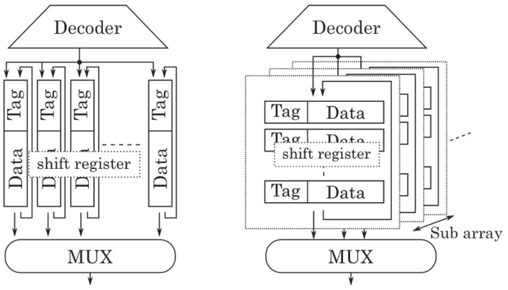
II. Bit-Parallel SFQ Cache Architecture

To solve the problems explained in Section I, we propose a novel SFQ cache memory architecture that supports bit-parallel access. Our cache employs a shift register that consists of SFQ circuits. We use a circular buffer in the sub-arrayed structure to realize low-latency non-destructive accesses. In addition, sub-arrayed structure mitigates the negative effects of multiplexer logics in terms of access latency. Purpose of designing architecture is realizing low latency and large capacity SFQ cache memory.

Fig. 2(b) illustrates a high level model of our proposed architecture. The architecture uses loop-shaped, and sub-arrayed shift registers to realize non-destructive-access, and low-latency-access. The shift register is composed by cascading flip-flops and the data are shifted to next flip-flop by an input of clock pulse. A cache entry corresponds to one set of flip-flops in the shift register, and the length of the shift register (the number of cache entries) is related to the access time. Thus, to reduce the access time, a long shift register is divided into sub-arrayed shift registers. High-order bits of the index correspond to the addresses of the entries, while low-order bits of the index are associated with the sub array.

When read/write accesses are occurred, the index address is decoded and the number of *shift pulses* is calculated from the index address and shift register's position. *Shift pulses* are clock pulses, which are used to move data inside of shift register. For example, if shift register consists of 4 entries, the maximum number of *shift pulses* is 3 (the minimum number of *shift pulse* is 0). In read access, data is selected by multiplexer. In write access, data and tags are stored in shift register.

978-1-5090-3219-8/16/$31.00 ©2016 IEEE



(a) Existing SFQ cache. (b) Proposed SFQ cache.

Figure 2. Two candidates of organization of cache architectures.

III. EVALUATION

In this section, we evaluate the access time, power consumption, and area of proposed architecture. We compare the results with existing SFQ shift register memory to unveil the effectiveness of our approach. We assume that the cache size of evaluation target is 16 Kbits. In this case, the index is 8 bits and the tag is 22 bits. The number of entries is 256. Fig. 3 shows the trade-off between access time and power consumption in changing the number of sub arrays. There is a sweet spot considering both of access time and power consumption. According to Fig. 3, we evaluate 32 sub-arrayed shift register cache. In our evaluation, the following models are introduced. Access time is given by Equation 1,

$$T = T_{diff} + T_{rot} + T_{sel}, \quad (1)$$

where T_{diff} is the required time for calculating the number of shift pulses, T_{rot} is the required time for shifting data, and T_{sel} is the required time for selecting data. They are given by the following equations.

$$T_{diff} = CCT_{diff} \times N_{gate-diff} \quad (2)$$

$$T_{rot} = CCT_{rot} \times N_{gate-rot} \quad (3)$$

$$T_{sel} = CCT_{sel} \times N_{gate-sel} \quad (4)$$

Here, CCT_{diff}, CCT_{rot}, CCT_{sel} are clock cycle time of logic gates which includes delay of logic gates, and $N_{gate-diff}$, $N_{gate-rot}$, $N_{gate-sel}$ are the number of logic gates included data path of each functional unit, respectively. We evaluate the area based on the number of JJs and past design results [6][7], because we did not layout the design. The number of JJs considering wiring costs are obtained based on the past design results [8]. Power consumption is given by the following.

$$P = (\alpha \Phi_0 I_c f + V I_{bias}) \times N_{JJ} \quad (3)$$

In this equation, α is switching probability, Φ_0 is flux quantum, I_c is the critical current of JJs, f is clock frequency, V is voltage of bias current, I_{bias} is bias current, and N_{JJ} is the number of JJs.

The estimated access time is 736.6 ps that outperforms the existing SFQ shift register memory by 1.8X. Fig. 4 shows relationship between access time and the number of cache entries (sub2 means bit-parallel shift register which uses 2 sub arrays). According to Fig. 4, sub-arrayed shift registers reduce the latency of the multiplexer compared with an existing SFQ shift register by 7X with 256 cache entries. Moreover, even if cache entries increase more, multiplexer's impact on the access time of sub-arrayed shift register is smaller than that of the traditional shift register. The number of JJs to be required is 624,406 JJs, and its estimated area is 543.6 mm^2. Compared to the existing SFQ memory, the proposed architecture is larger by 2.1X. The power consumption is estimated 240.9 mW.

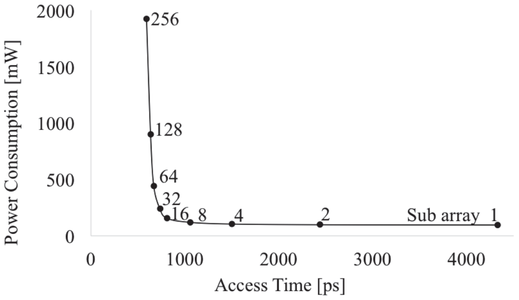


Figure 3. Trade-off between power consumption and access time in changing the number of sub arrays.

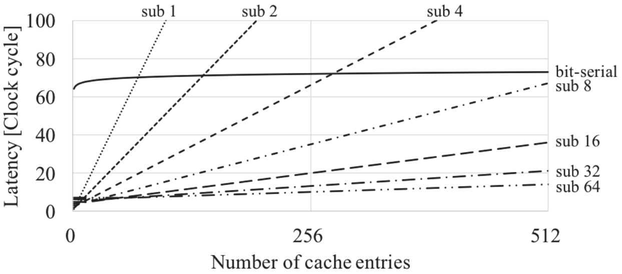


Figure 4. Estimated latency of the multiplexer.

ACKNOWLEDGMENT

This work was supported a part by JSPS KAKENHI Grant Number 16H02796 and by ALCA-JST

REFERENCES

[1] K. K. Likharev, et al. "RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems," *IEEE Transactions on Applied Superconductivity*, Vol. 1, No. 1, pp. 3–28, Mar. 1991.

[2] M Tanaka, et al. "Design of a pipelined 8-bit-serial single-flux-quantum microprocessor with multiple ALUs," *Superconductor Science and Technology*, Vol. 19, No. 5, p. S344, Mar. 2006.

[3] M Tanaka, et al. "Design and implementation of a pipelined 8 bit-serial single-flux-quantum microprocessor with cache memories," *Superconductor Science and Technology*, Vol. 20, No. 11, p. S305, Oct. 2007.

[4] Y Yamanashi, et al. "Design and implementation of a pipelined bit-serial SFQ microprocessor, CORE 1β," *IEEE Transactions on Applied Superconductivity*, Vol. 17, No. 2, pp. 474–477, Jun. 2007.

[5] P. Yuh, et al. "Design and testing of rapid single flux quantum shift registers with magnetically coupled readout gates," *IEEE Transactions on Applied Superconductivity*, Vol. 2, no. 4, pp. 214–221, Dec. 1992.

[6] K Fujiwara, et al. "Design and high-speed test of (4 × 8)-bit single-flux-quantum shift register files," *Superconductor Science and Technology*, Vol. 16, No. 12, p. 1456, Nov. 2003.

[7] K. Fujiwara, et al. "Design and component test of SFQ shift register memories," *IEEE Transactions on Applied Superconductivity*, Vol. 13, No. 2, pp. 555–558, Jun. 2003.

[8] M. Tanaka et al. "High-density shift-register-based rapid single-flux-quantum memory system for bit-serial microprocessors," *IEEE Transactions on Applied Superconductivity*, Vol. 26, No. 5, p. 1301005, Aug. 2016.