Rapid Single-Flux-Quantum Matrix Multiplication Circuit Utilizing Bit-Level Processing

Nobutaka Kito Takuya Kumagai Kazuyoshi Takagi

School of Engineering,
Chukyo University
Toyota 470-0393, Japan
nkito@sist.chukyo-u.ac.jp

Graduate School of Engineering,
Mie University
Tsu 514-8507, Japan
ktakagi@arch.info.mie-u.ac.jp

Abstract— A rapid single-flux-quantum (RSFQ) matrix multiplication circuit utilizing bit-level processing is presented. The proposed circuit utilizes characteristics of pulse logic used in RSFQ circuits and utilizes bit-level processing. The circuit carries out multiplications and additions by counting pulses on signal lines. It uses fewer gates compared with previously proposed parallel processing designs and could be realized in small layout area. A layout for 4-bit 4×4 matrix multiplication was designed and its correct operation was verified in simulation.

I. Introduction

Superconducting computing devices have been considered as potential alternative devices of mainstream semiconductor computing devices [1]. The rapid single flux quantum (RSFQ) circuit [2] and its energy efficient derivatives such as eSFQ [3], ERSFQ [4], and LV-RSFQ [5] are promising digital circuit technologies for high-speed and low-power operations. RSFQ circuits use Josephson junctions, which are superconductive devices working based on Josephson effect, and they work at up to 100 GHz [6, 7]. In RSFQ circuits, voltage pulses are used to represent logic values. Namely, pulse logic is used.

In this paper, an RSFQ matrix multiplication circuit is proposed. Matrix multiplication is a computational kernel operation used commonly in a wide variety of signal processing applications and neural network applications. Matrix multiplication involves many multiplication operations. Because multipliers are large components, an area-efficient design for matrix multiplication is necessary. The matrix multiplication circuit to be proposed is designed with consideration for area efficiency.

The matrix multiplication circuit proposed in this paper utilizes characteristics of pulse logic and bit-level processing proposed in [8] to save circuit area. Because RSFQ circuits use pulse logic, there are RSFQ-specific gates other than basic logic gates, such as AND, OR, and EXOR. Those specific gates are utilized to realize the circuit in small area. The circuit carries out multiplications and additions by counting voltage pulses. Therefore, the circuit treats 1-bit signals for calculation like stochastic computing [9, 10]. Wiring in RSFQ circuits occupies large circuit area because active devices are necessary for signal wires and a limited number of routing layers is available. The characteristic of the proposed circuit is suitable for RSFQ circuit realization.

The matrix multiplication circuit is suitable for applications which tolerate small error. It performs truncated multiplications internally and its result can contain small error. When a design for n-bit $m \times m$ matrix multiplication is implemented based on the proposed circuit, it performs the matrix multiplication with m^3 multiplications every $(2^n - 1) \times m^2$ clock cycles. Thus, it is suitable for applications using calculations of narrow bit-width.

Some applications such as neural network applications tolerate small error in arithmetic operations. For designs using semiconductor devices, this property has been leveraged in approximate computing [11] and stochastic computing. Matrix multiplication circuits which can be used for low-precision such as [12] have been proposed. Thus, the RSFQ matrix multiplication circuit is useful for various applications.

There has been research on designing RSFQ matrix multiplication circuits. In [13], designs of a 32-bit 4×4 matrix multiplication circuit have been shown. The matrix multiplication circuit in [13] uses bit-slice processing with bit-slice adders. The estimated amount of resources for designs of the circuit is very large, and they could not be implemented in a single chip. Though our proposed circuit is not suitable for applications requesting high-precision, it can be realized with drastically fewer gates compared with the designs in [13].

A layout of a design for 4-bit 4×4 matrix multiplication was performed for evaluation of the proposed matrix multiplication circuit. It was designed with the cell library developed for AIST advanced process (ADP2) [14]. The functionality of the designed layout was confirmed by simulation. It can perform a matrix multiplication with 64 multiplications every 240 clock cycles. The estimated maximum clock frequency of the designed circuit was 33 GHz. By comparison with the previously proposed matrix multiplier, it is suggested that the proposed circuit

can be implemented in smaller area.

This paper is organized as follows. In the next section, a brief review of matrix multiplication, RSFQ circuits, and the truncated multiplier in [8] which is utilized for the proposed circuit are shown. In Section III, an RSFQ matrix multiplication circuit is proposed and its operation is described. In Section IV, a layout design for 4-bit 4×4 multiplication is shown, and its evaluation results are shown. In Section V, this paper is concluded.

II. Preliminaries

A. Matrix Multiplication

We consider matrix multiplication $C = AB$ where each of A, B, and C is an $m \times m$ matrix as follows:

$$\begin{pmatrix} C_{0,0} & \cdots & C_{0,m-1} \\ \vdots & \ddots & \vdots \\ C_{m-1,0} & \cdots & C_{m-1,m-1} \end{pmatrix} =$$

$$\begin{pmatrix} A_{0,0} & \cdots & A_{0,m-1} \\ \vdots & \ddots & \vdots \\ A_{m-1,0} & \cdots & A_{m-1,m-1} \end{pmatrix} \begin{pmatrix} B_{0,0} & \cdots & B_{0,m-1} \\ \vdots & \ddots & \vdots \\ B_{m-1,0} & \cdots & B_{m-1,m-1} \end{pmatrix}.$$

Each element of input matrices is an n-bit fixed point number, i.e., $A_{i,j} = [0.a_{i,j,1} \cdots a_{i,j,n}]_2$ and $B_{i,j} = [0.b_{i,j,1} \cdots b_{i,j,n}]_2$. In other words, range of each element is $0 (= [0.0 \cdots 0]_2) \leq A_{i,j}, B_{i,j} \leq 1 - 2^{-n} (= [0.1 \cdots 1]_2)$.

Each element of the output matrix $C_{i,j} (0 \leq i, j < m)$ is calculated as follows:

$$C_{i,j} = \sum_{0 \leq k < m} A_{i,k} B_{k,j}.$$

Therefore, a matrix multiplication can be performed by m^3 multiplications.

B. RSFQ Circuits

In RSFQ circuits, voltage pulses are used for representing logic values and are transmitted on signal lines. Most of RSFQ gates including basic logic gates such as AND, OR and XOR have a clock input terminal and work synchronized with clock pulses. As an example, the symbol and the behavior of an RSFQ AND gate are shown in Figs. 1(a) and (b), respectively. The value of an input signal is determined with clock pulses as shown in Fig. 1(b). When a pulse arrives at a data input of a gate during an interval between adjacent clock pulses, the input value corresponding to the interval is considered as "1". If no pulse arrives during the interval, the input value is considered as "0". During the interval, it is prohibited to feed plural pulses to a data input of basic logic gates. The output of a basic gate is synchronized with the clock pulse.

In addition to the basic logic gates, there are several RSFQ-specific gates. Some of those gates are shown in Fig. 2. In the figure, the symbol of each gate and its behavior are presented. A non-destructive read-out (NDRO) gate has two internal states, i.e. $ST0$ and $ST1$, as shown in Fig. 2(a). It outputs a pulse at *dout* only when its internal state is $ST1$ and a pulse arrives at its *clk* terminal. A confluence buffer (CB) in Fig. 2(b) outputs a pulse when a pulse arrives at its input. It can merge pulses on two signal lines. A T1 gate in Fig. 2(c) works as a counter of pulses. When internal state of a T1 gate is $ST1$, it outputs a pulse at *carry* or *sum* terminal once a pulse arrives at *din* or *clk* terminal, respectively.

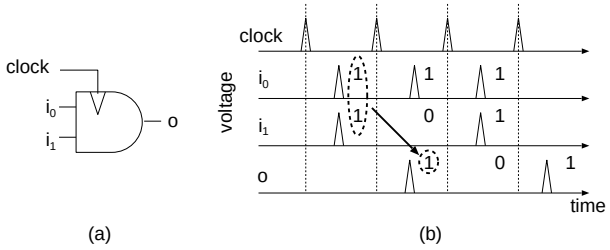

Fig. 1. RSFQ AND gate(a) and its behavior(b).

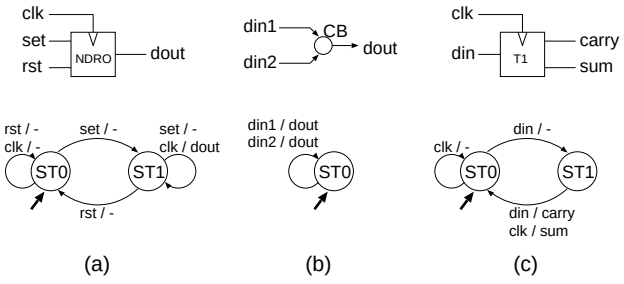

Fig. 2. RSFQ-specific gates and their behavior (Non-destructive read-out (NDRO) gate (a), confluence buffer (b), and T1 gate (c)).

In RSFQ circuits, signal lines are realized by Josephson transmission lines (JTLs) or passive transmission lines (PTLs). JTLs are composed of active devices, i.e., Josephson junctions, and have relatively large size and delay. Though delay of a PTL is smaller than delay of a JTL with the same length, a pair of a driver and a receiver is necessary in its both ends and it can connect a pair of pins without fanouts. In AIST advanced process (ADP2) [14], two PTL wiring layers are available. Splitters containing active devices are necessary to feed a signal to plural inputs. A splitter is depicted by symbol • in schematic. Because various components containing active devices are necessary for realizing signal lines, signal lines consume large area in a layout.

For correct operation of a designed circuit, JTLs are inserted on signal lines as delay elements to keep up order of pulse arrivals to the expected order at each gate.

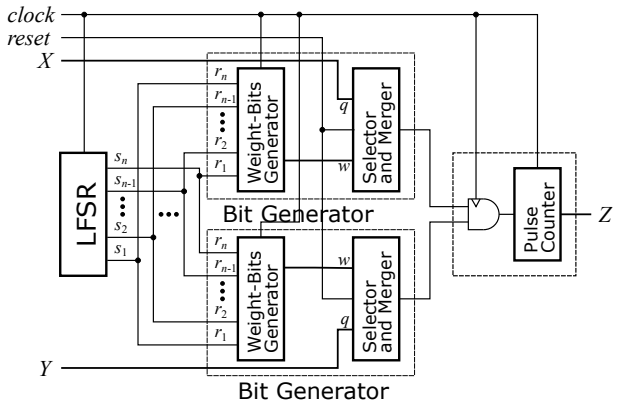

Fig. 3. Structure of the truncated multiplier based on bit-level processing [8].

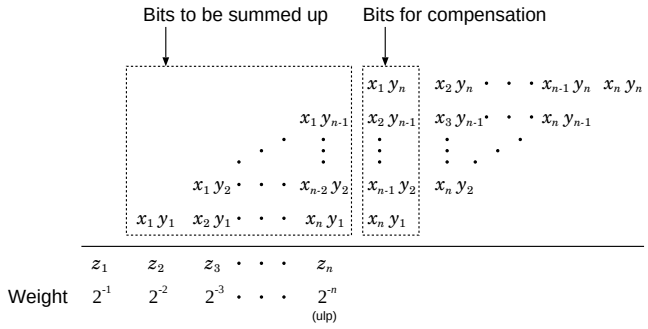

Fig. 4. Partial product bits used in the truncated multiplier.

C. RSFQ Truncated Multiplier Based on Bit-Level Processing

The matrix multiplication circuit proposed in this paper is based on the truncated multiplier proposed in [8]. The structure of the truncated multiplier is shown in Fig. 3. Its inputs are $X(= [0.x_1x_2\cdots x_n]_2)$, $Y(= [0.y_1y_2\cdots y_n]_2)$, *clock*, and *reset*. Its output is $Z(= [0.z_1z_2\cdots z_n]_2)$. The resultant value Z is calculated as follows:

$$Z = \sum_{j+i\le n} x_j y_i 2^{-(i+j)} + \sum_{j+i=n+1} x_j y_i 2^{-n}.$$

The former term corresponds to the bits to be summed up in Fig. 4, and the latter term is the compensation term using the bits for compensation in Fig. 4. The truncated multiplier treats those compensation bits whose weight is 2^{-n-1} as bits of the twice weight, i.e., 2^{-n}, in other words, it rounds each partial product toward its nearest value.

The multiplier calculates one multiplication every 2^n-1 clock cycles. The linear feedback shift register (LFSR) whose period is 2^n-1 in the multiplier feeds an n-bit pattern every clock cycle for the bit generators. The bit ordering of the connection from LFSR to one bit-generator is different from that of the other. The two generators output two one-bit signals and they are fed to the AND gate. The output pulses of the AND gate are counted by the pulse counter realized with n $T1$ gates.

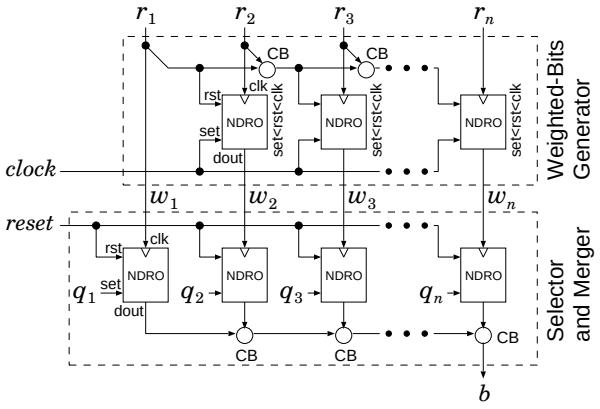

Fig. 5. Design of the bit generator.

In Fig. 5, the detailed design of the bit generator composed of the weighted-bits generator and the "selector and merger" is shown. CBs and NDROs are utilized to realize it in compact area. The upper row of NDROs forms the weighted-bits generator, and the following row of NDROs forms the selector. The CBs in the bottom in the figure form the merger of pulses.

For NDROs of the upper row, the orders of pulse arrivals are depicted using the inequalities [15]. The weighted-bits generator feeds w_i signals and the output value of w_i is represented as $r_i \wedge (\overline{r_{i-1}} \wedge \cdots \wedge \overline{r_1})$. During the period of the LFSR, i.e., 2^n-1 clock cycles, 2^{n-i} pulses appear at w_i.

Pulses need to be fed to NDROs in the selector to set their internal states before a calculation starts. When the internal state of an NDRO is preset through q_i, the NDRO outputs a pulse once a pulse arrives at w_i. As a result, the merger outputs $[q_1 \dots q_n]_2$ $(= 2^{n-1}q_1 + \cdots + q_n)$ pulses. As connections between the LFSR and two bit-generators are different, the AND gate outputs the result of truncated multiplication like multiplication using stochastic computing [9, 10]. Note that the multiplications in the proposed circuit utilize the correlation of the shared LFSR though correlation between operands may degrade results in multiplications in stochastic computing. The LFSR is not utilized as an random number generator. Thus, the sharing of the LFSR is not a problem in the circuit.

III. RSFQ Matrix Multiplication Circuit Utilizing Bit-Level Processing

We propose an RSFQ matrix multiplication circuit utilizing the idea of the truncated multiplier discussed in Section II.C. The circuit calculates each column of the

resultant matrix, sequentially. For calculating a column, the circuit carries out m multiplications and additions for its rows simultaneously.

A. Structure

The structure of the proposed RSFQ matrix multiplier is shown in Fig. 6. In the figure, several control signals such as *clock* and *reset* are omitted. The components of the matrix multiplication circuit are almost the same as those of the multiplier in Section II.C.

The circuit is designed to carry out the following calculation.

$$\begin{pmatrix} C_{0,i} \\ \vdots \\ C_{m-1,i} \end{pmatrix} = \begin{pmatrix} A_{0,0} \\ \vdots \\ A_{m-1,0} \end{pmatrix} B_{0,i} + \cdots + \begin{pmatrix} A_{0,m-1} \\ \vdots \\ A_{m-1,m-1} \end{pmatrix} B_{m-1,i}.$$

For each $B_{k,l}$ $(0 \le k, l < m)$, m multiplications are necessary and the circuit performs the m multiplications simultaneously. Though it carries out multiple multiplications simultaneously, only one LFSR is used.

Each term in the right-hand side of the above formula is carried out in $2^n - 1$ clock cycles, and the left-hand side, i.e., a column of the resultant matrix, is obtained in $m \times (2^n - 1)$ clock cycles. For calculation of each term in the right-hand side of the above formula, elements of A are fed through $Ain_0, \ldots, Ain_{m-1}$ inputs, and an element of B is fed through Bin input. Accumulation of terms is realized by counting pulses without resetting the pulse counters in the circuit. Total number of clock cycles for $m \times m$ matrix multiplication is $m^2 \times (2^n - 1)$.

Because one of two operands in multiplications is fixed to Bin in each term in the formula, components receiving elements of A are duplicated from the original truncated multiplier. Weighted-bits generation for $Ain_0, \cdots, Ain_{m-1}$ are the same, and we use only one weighted-bits generator for $Ain_0, \cdots, Ain_{m-1}$. We duplicate components other than the weighted-bits generator, i.e., the selector, the merger, the AND gates, and the counter. Thus, circuit area is not enlarged significantly though the circuit performs m multiplication simultaneously.

B. Operation

The steps for calculating matrix multiplication using the proposed circuit are shown in Algorithm 1.

In each interval of $2^n - 1$ clock cycles, the circuit calculates $Ain_0 \cdot Bin, \cdots, Ain_{m-1} \cdot Bin$. We need to feed inputs for $Ain_0, \cdots, Ain_{m-1}$, and Bin every $2^n - 1$ clock cycles. The pulse counters accumulate results of m multiplications to calculate a column of the resultant matrix. Thus, we reset the pulse counters at the beginning of calculation of each column, and observe the result at pulse counters after $m \times (2^n - 1)$ clock cycles.

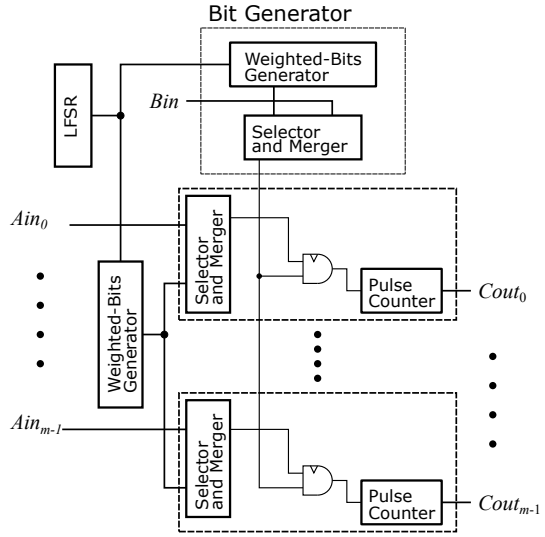

Fig. 6. Structure of the RSFQ matrix multiplication circuit utilizing bit-level processing.

Algorithm 1 Calculation of matrix multiplication with the proposed circuit

for $j = 0 \ldots m-1$ **do**
  Reset the pulse counters
  for $k = 0 \ldots m-1$ **do**
    Feed operands into $Ain_0, \ldots, Ain_{m-1}$, and Bin
      $Ain_i \leftarrow A_{i,k} (0 \le i < m)$
      $Bin \leftarrow B_{k,j}$
    Feed $2^n - 1$ clock pulses
  end for
  Values of $C_{i,j} (0 \le i < m)$ are calculated at $Cout_i$
end for

The multiplication result obtained by the truncated multiplier is n bit. Thus, each element of resultant matrix C is $(n + \lceil \log_2 m \rceil)$ bits. Namely, each element is composed of n-bit fraction part and $\lceil \log_2 m \rceil$-bit integer part.

IV. Layout Design and Evaluation

We show a layout of a 4-bit 4×4 matrix multiplication circuit for evaluation of the proposed matrix multiplication circuit. We used Cadence Virtuoso and the cell library designed for AIST advanced process (ADP2) [14] for layout design. In the design flow using the cell library, a layout is composed by tiling cells in the schematic editor. We can easily convert a designed layout in schematic editor to a physical layout for fabrication.

We show the layout designed with schematic editor in Fig. 7. The four components in the right side of the figure correspond to the components rounded by broken lines in Fig. 6. The circuit area is 2.68 mm^2 (1.40 $\times$ 1.92 mm^2),

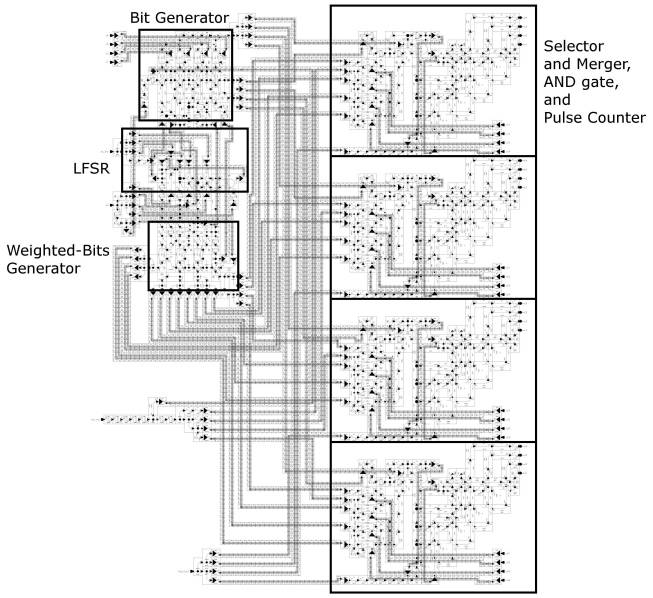

Fig. 7. Layout of a 4-bit 4×4 RSFQ matrix multiplication circuit.

and the number of Josephson junctions (JJs) in it is 2,711. The number of JJs in designs of the previously proposed 32-bit 4×4 matrix multiplier [13] was estimated from 30,000 to 1,000,000. Though the designed layout was 4-bit, the number of JJs in the layout is drastically fewer than that of the previously proposed one, and layout designs of the previously proposed circuit have not been shown.

We have carried out logic simulation of the designed layout considering delay of each component with Cadence Verilog-XL. With the simulation result, the functionality of the layout was verified and it was estimated to work at up to 33 GHz.

V. Conclusion

We proposed an RSFQ matrix multiplication circuit utilizing bit-level processing. The circuit utilizes RSFQ-specific gates aggressively, and carries out internal truncated multiplications and additions by counting pulses on signal lines. The circuit requires small amount of gates and wires compared with parallel processing circuits, and could be realized in small area. The circuit is suitable for applications which tolerates small error.

Acknowledgements

This work has been supported in part by JSPS KAKENHI Grant Numbers 19K11888 and 16K16029, and supported by VLSI Design and Education Center (VDEC), The University of Tokyo with the collaboration with Cadence Corporation.

References

[1] D.C. Brock, "The NSA's frozen dream," *IEEE Spectrum*, vol.53, no.3, pp.54–60, Mar. 2016.

[2] K. Likharev and V. Semenov, "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.

[3] M.H. Volkmann, A. Sahu, C.J. Fourie, and O.A. Mukhanov, "Implementation of energy efficient single flux quantum digital circuits with sub-aJ/bit operation," *Superconductor Science and Technology*, vol.26, no.1, p.015002, Nov. 2012.

[4] D.E. Kirichenko, S. Sarwana, and A.F. Kirichenko, "Zero static power dissipation biasing of RSFQ circuits," *IEEE Transactions on Applied Superconductivity*, vol.21, no.3, pp.776–779, June 2011.

[5] M. Tanaka, A. Kitayama, T. Koketsu, M. Ito, and A. Fujimaki, "Low-energy consumption RSFQ circuits driven by low voltages," *IEEE Transactions on Applied Superconductivity*, vol.23, no.3, pp.1701104–1701104, June 2013.

[6] Y. Yamanashi, T. Kainuma, N. Yoshikawa, I. Kataeva, H. Akaike, A. Fujimaki, M. Tanaka, N. Takagi, S. Nagasawa, and M. Hidaka, "100 GHz demonstrations based on the single-flux-quantum cell library for the 10 kA/cm^2 Nb multi-layer process," *IEICE Transactions on Electronics*, vol.93, no.4, pp.440–444, Apr. 2010.

[7] M. Tanaka, H. Akaike, A. Fujimaki, Y. Yamanashi, N. Yoshikawa, S. Nagasawa, K. Takagi, and N. Takagi, "100-GHz single-flux-quantum bit-serial adder based on 10-kA/cm^2 niobium process," *IEEE Transactions on Applied Superconductivity*, vol.21, no.3, pp.792–796, June 2011.

[8] N. Kito, R. Odaka, and K. Takagi, "Rapid single-flux-quantum truncated multiplier based on bit-level processing," *IEICE Transactions on Electronics*, vol.E102-C, no.7, pp.607–611, July 2019.

[9] B.R. Gaines, "Stochastic computing," *Proceedings of AFIPS Spring Joint Computer Conference*, pp.149–156, 1967.

[10] A. Alaghi and J.P. Hayes, "Survey of stochastic computing," *ACM Transactions on Embedded Computing Systems*, vol.12, no.2s, pp.92:1–92:19, May 2013.

[11] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys*, vol.48, no.4, pp.62:1–62:33, Mar. 2016.

[12] Y. Umuroglu, L. Rasnayake, and M. Sjalander, "BISMO: A scalable bit-serial matrix multiplication overlay for reconfigurable computing," Proceedings of 28th International Conference on Field Programmable Logic and Applications (FPL), pp.307–3077, Aug. 2018.

[13] G. Tang, P. Qu, X. Ye, D. Fan, and N. Sun, "32-bit 4×4 bit-slice RSFQ matrix multiplier," *IEEE Transactions on Applied Superconductivity*, vol.28, no.7, pp.1–5, Oct. 2018.

[14] S. Nagasawa, K. Hinode, T. Satoh, M. Hidaka, H. Akaike, A. Fujimaki, N. Yoshikawa, K. Takagi, and N. Takagi, "Nb 9-layer fabrication process for superconducting large-scale SFQ circuits and its process evaluation," *IEICE Transactions on Electronics*, vol.E97-C, no.3, pp.132–140, Mar. 2014.

[15] K. Takagi, N. Kito, and N. Takagi, "Circuit description and design flow of superconducting SFQ logic circuits," *IEICE Transactions on Electronics* , vol.E97-C, no.3, pp.149–156, Mar. 2014.