Vol.2022-ARC-249 No.7
2022/7/28

単一磁束量子プロセッサ向けキャッシュメモリ構成法の検討と定量的評価

鴨志田 圭吾[1,a)]　石川 伊織[1]　羽野 祐太[1]　川上 哲志[1]　谷本 輝夫[1]　小野 貴継[1]　田中 雅光[2]
藤巻 朗[2]　井上 弘士[1]

概要：本稿では，単一磁束量子（Single Flux Quantum:SFQ）回路を用いたマイクロプロセッサの実現を念頭に，4 kelvin 環境下での動作を前提とした極低温キャッシュメモリ構成法を検討する．メモリアレイに関しては SFQ シフトレジスタまたは SRAM での実装，キャッシュ内グローバル配線に関しては SFQ または CMOS 回路での実装を想定する．これらの設計選択肢に基づき，SFQ キャッシュ（SFQ シフトレジスタ型 FIFO メモリと SFQ 回路で実装），CMOS キャッシュ（SRAM と CMOS 回路で実装），ならびに，ハイブリッドキャッシュ（SRAM と SFQ 回路で実装）といった 3 つのアーキテクチャモデルを導入し，モデリングに基づくアクセス時間の評価を行う（ただし，本稿ではデータメモリアレイにのみ着目する）．その結果，現行の 1.0 μm プロセスを前提とした場合には，SFQ キャッシュよりもハイブリッドまたは CMOS キャッシュが優れていることが分かった．

1. はじめに

CMOS 回路で実装される現在のプロセッサは，トランジスタの微細化を拠り所としてその高速化と低消費電力化を達成してきた．しかしながら，デナード則 [1] の破綻や，近い将来に訪れると予想されているムーア則（半導体の微細化）の終焉により，その継続的な発展が難しくなりつつある．この問題を解決するアプローチとして新奇デバイスの活用が注目されており，その 1 つに超伝導単一磁束量子（Single Flux Quantum: SFQ）回路がある [2]．論理ゲートのスイッチ時に消費するエネルギーは CMOS 回路の約 1/1000 という低消費電力性を有し，かつ，数十 GHz での超高速動作を実現することができる [3], [4]．

特に最近では，ALU や乗算器といった演算回路の実装に加え，プロセッサや AI アクセラレータを対象とした SFQ 向けアーキテクチャ研究が盛んに行われている．例えば，32 GHz, 6.5 mW で動作する 4bit プロセッサは，ゲートレベルパイプラインでのストールの頻発を回避するための SIMT (Single Instruction Multiple Thread) 実行モデルを採用し，循環型レジスタファイルを導入するなどの工夫がなされている [5]．しかしながら，依然としてオンチップバッファやスクラッチパッドメモリの利用を前提とした場合が殆どであり，プロセッサ性能に大きな影響を及ぼすキャッシュメモリ構成法に関する議論は極めて少ない．著者らが知る限り，先行研究としては SFQ シフトレジスタ型 FIFO メモリを用いたアーキテクチャの提案のみであり，そのアクセス時間は容量 2 KB で 736.6 ps と大きく，有効性は明らかでない [6]．また，SFQ シフトレジスタ以外のメモリデバイスの利用も可能となりつつあり，様々な実装手段を踏まえた SFQ 向けキャッシュメモリ・アーキテクチャの検討が求められている．

そこで本研究では，SFQ プロセッサ向け極低温キャッシュメモリの実現を念頭に，その設計選択肢を議論する．現状，SFQ 回路でメモリを実装する場合，シフトレジスタによる FIFO メモリを用いる以外に方法は無い．その一方，近年では 4 kelvin 環境下でのスクラッチパッドメモリの実装（SRAM を利用）が提案されており，キャッシュの基本構成要素であるメモリアレイへの応用が考えられる [7]．また，近年の大容量キャッシュでは入出力ポートと各メモリアレイを接続するグローバル配線が性能や消費電力に大きな影響を与えている．その実装においては，SFQ 高速伝送路または電荷の充放電に基づく CMOS 回路の利用が考えられる．これらの設計選択肢に基づき，本研究では 3 つのキャッシュアーキテクチャモデル（完全 SFQ 回路型，完全 CMOS 回路型，ハイブリッド型）を導入する．そして，データメモリアレイにのみ着目したアクセス時間モデリングに基づき性能を評価し，SFQ プロセッサ向け

[1] 九州大学
[2] 名古屋大学
[a)] keigo.kamoshida@cpc.ait.kyushu-u.ac.jp

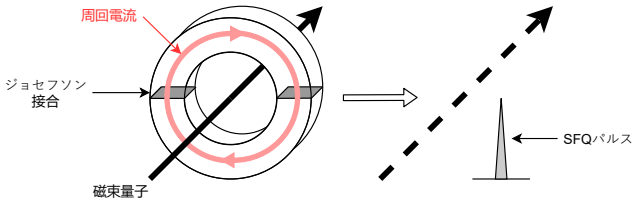


図 1 ジョセフソン接合を含むループと SFQ パルス

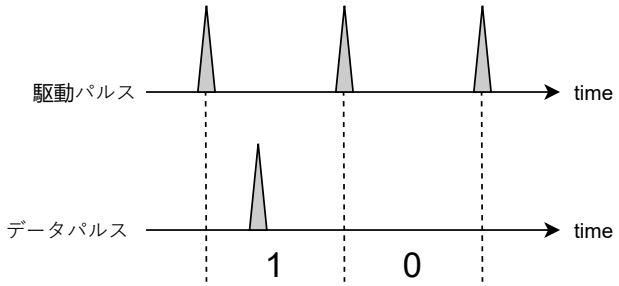


図 2 パルス論理におけるデータの判別

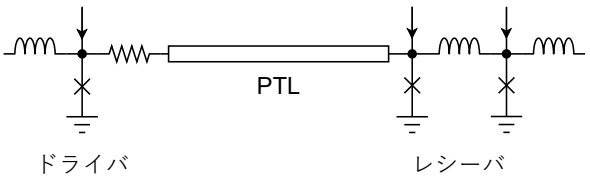


図 3 PTL の模式図

キャッシュメモリの実現に向けた課題を明らかにする．

本稿の構成は以下の通りである．第 2 節では SFQ 回路の動作原理と特性について述べる．第 3 節で SFQ プロセッサ向けキャッシュアーキテクチャを定義し，第 4 節でアクセス時間モデルを導入する．第 5 節でモデルに基づくアクセス時間評価を行い，最後に第 6 節でまとめる．

2. 単一磁束量子回路

2.1 動作原理

単一磁束量子（Single Flux Quantum: SFQ）回路はジョセフソン接合を含んだ超伝導体のループ（図 1）によって構成されている．超伝導ループ内を貫く磁束は磁束量子 Φ_0 の整数倍に量子化される．ループ内に磁束量子が入っているとき，ループ部分には周回電流が流れる．接合に流れる電流が臨界電流値 I_c より小さいならば磁束量子はループ内に保持される．逆に I_c より大きくなると接合がスイッチし，磁束量子はループの外に出る．SFQ 回路はループ内に磁束量子が存在する場合にビット ‘1’ を，存在しない場合にビット ‘0’ を保持していると見なす．接合がスイッチするとき，SFQ パルスと呼ばれる高さが数百 μV，幅が数 ps のインパルス状の電圧が発生する．SFQ 回路はこの SFQ パルスを情報伝搬に用いる．磁束量子の保持や SFQ パルスの伝搬を制御することで SFQ 回路は論理演算を実現させる．

2.2 パルス論理

CMOS 回路は電圧レベルの高低を ‘1’, ‘0’ に対応させる「レベル論理」である．これに対して SFQ 回路は SFQ パルスの有無を ‘1’, ‘0’ に対応させる「パルス論理」である．このため，SFQ 回路は CMOS 回路とは異なる独自の論理回路構成が採用されている．具体的には，SFQ 回路において ‘1’ をパルスの到着で表す際，「‘0’ 状態」と「‘1’ だがパルスがゲートに到着していない状態」の判別ができない．そのため，論理ゲートの入力に駆動パルスを追加し，2 つの駆動パルスの間にデータパルスが到着すれば ‘1’，到着しなければ ‘0’ と判別する方式が用いられている（図 2）．駆動パルスの入力により動作するため，SFQ 論理ゲートはラッチ機能を持つ．

2.3 ゲートレベルパイプライン

ゲートレベルパイプラインとは論理ゲート 1 段がパイプライン 1 段に対応した最も粒度が細かいパイプライン構造である．SFQ 論理ゲートは原理的にラッチを持っており追加のパイプラインレジスタを挿入する必要がないため，ゲートレベルパイプラインを容易に実現することができる．この構造を採用した乗算器 [3] と ALU[4] はそれぞれ 48 GHz, 52 GHz での動作が実証されている．

2.4 PTL

PTL（Passive Transmission Line）（図 3）は SFQ 回路の配線の 1 種であり，マイクロストリップラインまたはストリップライン構造の伝送線路を用い，インパルス状の電圧信号を電磁波として伝搬させる．PTL の両端には専用のインタフェース回路を用意し，送信側をドライバ，受信側をレシーバと呼ぶ．伝搬速度は真空中の光速のおよそ 1/3，すなわち約 10^8 m/s である．

3. キャッシュメモリの構成法

本研究では，CACTI [8] におけるキャッシュメモリのレイアウトに基づき，SFQ 回路向けのキャッシュメモリの構成法を議論する．

3.1 キャッシュメモリのレイアウト

図 4 は本研究で想定するキャッシュメモリのレイアウトである．キャッシュメモリはデータアレイとタグアレイによって構成される．それぞれのアレイには MAT（1 つのプレデコーダを共有した 4 つのサブアレイ）がタイル状に配置されている．サブアレイにはデコーダ，マルチプレクサといった周辺回路とメモリアレイが含まれる．アドレスやデータは H 木状の GW（グローバルワイヤ）を介して

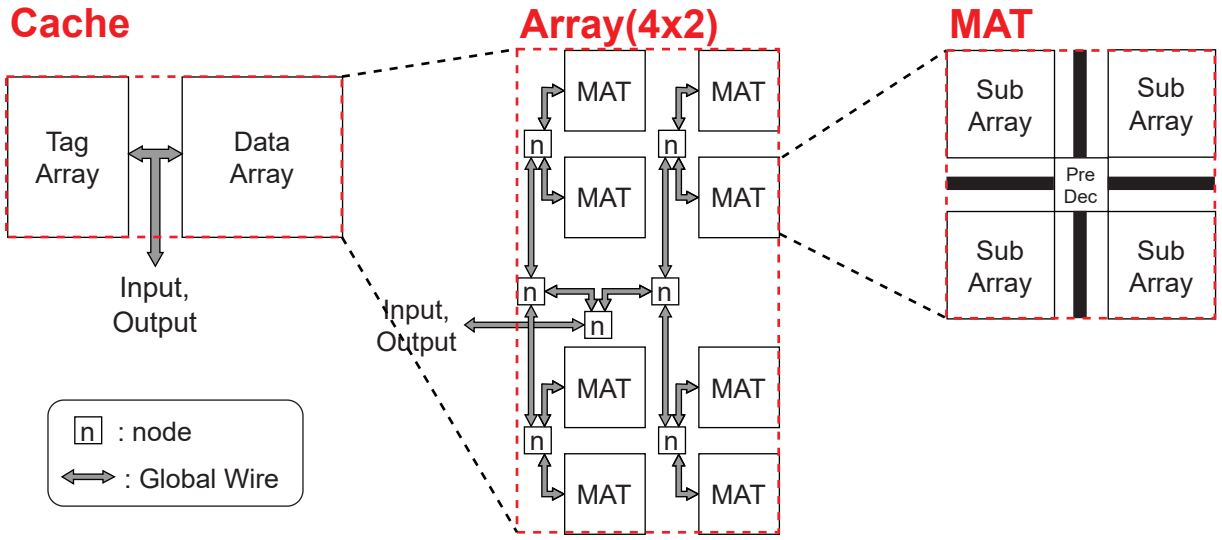


図 4 キャッシュメモリのレイアウト

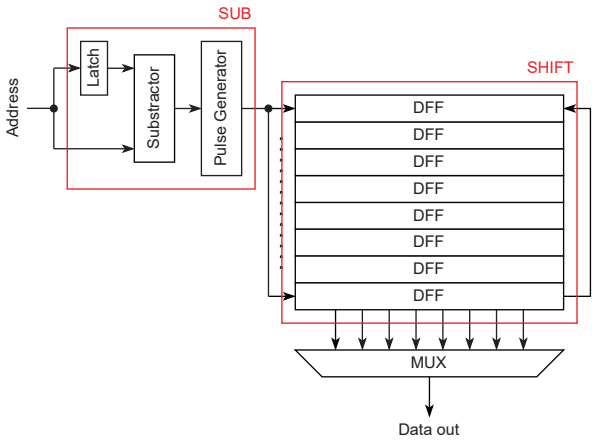


図 5 SFQ 回路の場合のサブアレイ

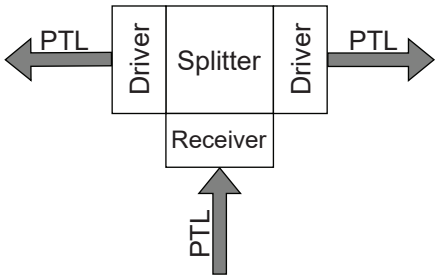


図 6 SFQ 回路の場合のノード

MAT へ入出力される．GW の分岐点をノードと呼ぶ．本研究では，GW と MAT それぞれに焦点を当てる．

3.2 設計選択肢

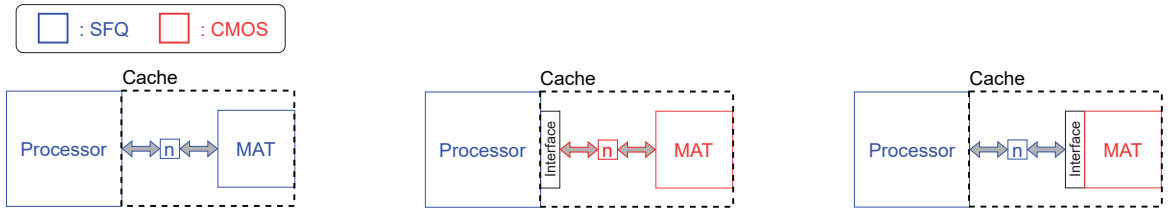
SFQ プロセッサ向けのキャッシュメモリは 4 kelvin の環境下にあると考えられる．このとき，SFQ 回路と CMOS 回路の両方を用いることが可能である．また，SFQ パルスとレベル電圧を変換するインタフェース回路を導入することで，SFQ 回路と CMOS 回路を組合わせた構成を採ることもできる．動作周波数や配線遅延，消費電力の観点からは SFQ 回路が CMOS 回路よりも優れている．一方，集積度の観点からは，微細化が進んでいる CMOS 回路の方が優位となる．そのため，集積度の高さが重視されるメモリに関しては CMOS 回路（SRAM アレイ）を用いる構成法が選択肢に挙がる．GW と MAT の実装の違いに着目し，図 7 に示す 3 種のキャッシュアーキテクチャモデルを定義する．

3.3 キャッシュアーキテクチャモデル

3.3.1 SFQ キャッシュ

GW を SFQ の PTL，ノードを SFQ 回路，MAT を SFQ 回路で実装する（図 7(a)）．SFQ 回路による MAT の実装では，図 5 に示すサブアレイ構成を想定する．メモリアレイは SFQ シフトレジスタで構成され，巡回バッファ構造を有する．対象データが入出力ポート位置に到達するまで巡回シフト操作を行うことで，データの読み書きを行う．巡回シフト回数は，現アクセスにおけるアドレス値と前回アクセス時のアドレス値の差で求まる．そのため，デコーダ部分は前回アクセスにおけるアドレス値を記憶するラッチと減算器，ならびに，巡回シフト操作のためのパルス生成回路によって構成される．

また，PTL で GW を，SFQ 回路でノードを実装する場合，図 6 のように PTL からレシーバで信号を受け取り，スプリッタで SFQ パルスを分岐させ，ドライバで PTL に信号を出力する必要がある．SFQ キャッシュでは PTL による高速な情報伝搬が可能であり，インタフェース回路が必要ないといった長所がある．また，CMOS 回路を用いる場合よりも低消費電力であると考えられる．一方，SFQ 回路の微細化は CMOS 回路と比べて大幅に遅れているため，現行のプロセスでは面積が大きくなる傾向にある．その結



(a) SFQ キャッシュ　(b) CMOS キャッシュ　(c) ハイブリッドキャッシュ

図 7　キャッシュアーキテクチャモデル

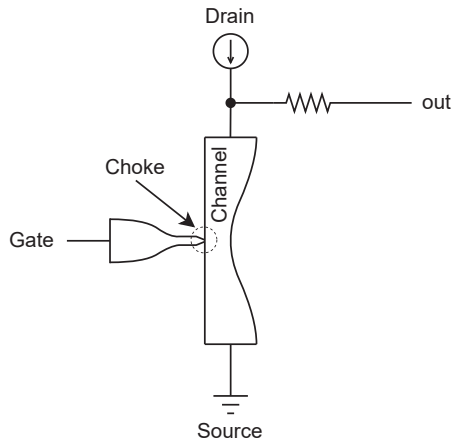


図 8　nTron の構造

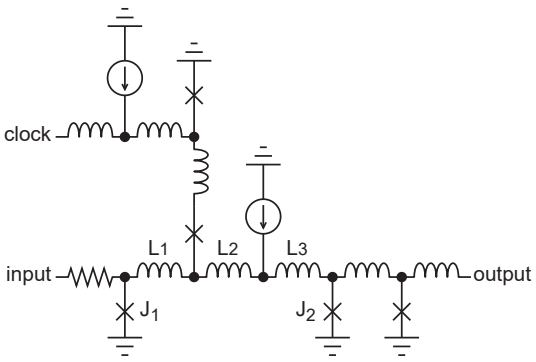


図 9　LDDS の回路図

果，GW が長くなり，ひいては，アクセス時間にも悪影響を及ぼす．また，メモリアレイがシフトレジスタであるため，対象データの入出力ポートからの距離によって巡回シフト回数が変化する．最良の場合（アクセス対象データが入出力ポート位置に存在する場合）はゼロ回，最悪の場合はエントリー数回の巡回シフトが必要となり，アクセス時間に不均一性が生じる．

3.3.2　CMOS キャッシュ

GW を RC 配線（電荷の充放電により情報を伝播する従来の配線），ノードを CMOS 回路，MAT を SRAM（周辺回路は CMOS）で実装する（図 7(b)）．既存の半導体プロセッサに用いられるキャッシュメモリの構成と同様である．ただし，GW と SFQ プロセッサの間にはインタフェース回路が必要となる．本研究では SFQ パルスをレベル電圧に変換する回路を SFQ/DC インタフェース回路，レベル電圧を SFQ パルスに変換する回路を DC/SFQ インタフェース回路と呼ぶ．

SFQ/DC インタフェース回路は図 8 に示す nTron [9] の利用を想定する．nTron はチャネルの超伝導状態（抵抗無）と常伝導状態（抵抗有）をゲートからの電流によって制御することで出力電圧を変える．チャネル幅はゲート付近で最も小さくなるように設計される．また，ゲート幅もチャネルに近づくにつれて小さくなり，先端部分をチョークと呼ぶ．このような構造によって SFQ パルスの微弱な電流でもチャネルの超伝導状態から常伝導状態への遷移が可能となる．

DC/SFQ インタフェース回路は LDDS [10] の利用を想定する．図 9 は LDDS の回路図である．入力から電流が流れているときにクロックから SFQ パルスが到着すると接合 J_1, J_2 がスイッチし，SFQ パルスが出力される．

3.3.3　ハイブリッドキャッシュ

GW を SFQ の PTL，ノードを SFQ 回路，MAT を SRAM（周辺回路は CMOS）で実装する（図 7(c)）．PTL による高速な情報伝搬と，SRAM による高密度実装（ならびにランダムアクセス）の両立が可能となる．ただし，GW と MAT の間にはインタフェース回路が必要である．消費電力，面積は SFQ キャッシュと CMOS キャッシュの中間であると予想される．

4.　アクセス時間モデル

本研究では簡単のためデータアレイのアクセス時間モデルから 3 種の構成法を評価する．なお，タグを考慮したキャッシュ全体の評価に関しては今度の課題である．データアレイのアクセス時間 $T_{\mathrm{data\_access}}$ は式 (1) で表される．

$$T_{\mathrm{data\_access}} = \begin{cases} T_{\mathrm{GW}} + T_{\mathrm{MAT}}, \\ (\mathrm{SFQ}) \\ T_{\mathrm{GW}} + T_{\mathrm{MAT}} + T_{\mathrm{interface}}, \\ (\mathrm{CMOS}, \text{ハイブリッド}) \end{cases} \tag{1}$$

SFQ キャッシュの場合，GW の遅延 T_{GW} と MAT の遅延 T_{MAT} の和であるが，CMOS，ハイブリッドキャッシュの場合インタフェース回路の遅延 $T_{\mathrm{interface}}$ が加算される．

GW の遅延 T_{GW} を式 (2) で表す．入力と出力で 2 回情報伝搬が行われるため式全体に 2 を掛けている．

$$T_{\mathrm{GW}} = \begin{cases} 2\{l_{\mathrm{GW}}/v_{\mathrm{PTL}} + T_{\mathrm{node}} \log_2 N_{\mathrm{MAT}}\}, \\ (\mathrm{SFQ}, \text{ハイブリッド}) \\ 2\{l_{\mathrm{GW}}/v_{\mathrm{RC}}\}, \\ (\mathrm{CMOS}) \end{cases} \tag{2}$$

GW の配線長 l_{GW} を伝搬速度で割ったものが，GW の配線部分の遅延となる．v_{PTL}, v_{RC} はそれぞれ PTL と RC 配線のデータ伝搬速度である．RC 配線は CACTI [8] の想定と同様に一定間隔でドライバーが挿入され，配線長と伝搬遅延は比例関係にあるとしている．GW に PTL を用いる SFQ，ハイブリッドキャッシュはレシーバとドライバによる遅延の増加を無視できないため，ノード部分の時間を加えている．ノードの遅延の合計はノード 1 つあたりの遅延 T_{node} とノード数の積によって表される．MAT 数 N_{MAT} が 2 倍になるとノードは 1 つ増えるため，ノード数は $\log_2 N_{\mathrm{MAT}}$ となる．

GW の配線長 l_{GW} を式 (3) で表す．

$$\begin{aligned} l_{\mathrm{GW}} =& H\{(N_{\mathrm{MAT\_H}} - 1)/N_{\mathrm{MAT\_H}}\}/2 \\ &+ W\{(N_{\mathrm{MAT\_W}} - 1)/N_{\mathrm{MAT\_W}}\} \end{aligned} \tag{3}$$

H, W はそれぞれアレイの縦幅と横幅，$N_{\mathrm{MAT\_H}}$, $N_{\mathrm{MAT\_W}}$ はそれぞれ縦，横に並んだ MAT 数である．例として，図 10 のような 4×4 のアレイで右上端の MAT に対して入出力が行なわれる場合，横向きの GW（図 10 赤線）の長さは MAT3（$= 4 - 1$）つに対応する．アレイの横幅 W が MAT4 つに対応するため，GW の横向きの長さは $W\{(N_{\mathrm{MAT\_W}} - 1)/N_{\mathrm{MAT\_W}}\} = W(4-1)/4 = W \times 3/4$ となる．縦向きの GW（図 10 青線）の長さは横向きと同様だが，アレイの端ではなく中央から伸びる配線であるため，全体に 1/2 を掛ける．

アレイの縦幅 H を式 (4) で表す．

$$H = \begin{cases} N_{\mathrm{MAT\_H}} H_{\mathrm{MAT\_SFQ}} + N_{\mathrm{GW\_H}} P_{\mathrm{PTL}}, \\ (\mathrm{SFQ}) \\ N_{\mathrm{MAT\_H}} H_{\mathrm{MAT\_CMOS}} + N_{\mathrm{GW\_H}} P_{\mathrm{RC}}, \\ (\mathrm{CMOS}) \\ N_{\mathrm{MAT\_H}} H_{\mathrm{MAT\_CMOS}} + N_{\mathrm{GW\_H}} P_{\mathrm{PTL}}, \\ (\text{ハイブリッド}) \end{cases} \tag{4}$$

$H_{\mathrm{MAT\_SFQ}}$, $H_{\mathrm{MAT\_CMOS}}$ はそれぞれ SFQ 回路，CMOS 回路と SRAM で実装した場合の MAT の縦幅である．また，$N_{\mathrm{GW\_H}}$ は縦に並んだ GW の本数，P_{PTL}, P_{RC} はそれぞれ PTL と RC 配線のピッチである．

$H_{\mathrm{MAT\_SFQ}}$ を式 (5) で表す．

$$H_{\mathrm{MAT\_SFQ}} = R_{\mathrm{array\_h}} H_{\mathrm{MAT\_CMOS}} \tag{5}$$

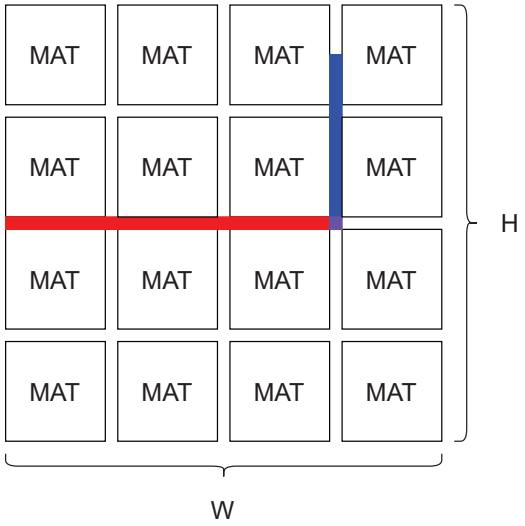


図 10 MAT が 4×4 で並ぶ場合の GW

$R_{\mathrm{array\_h}}$ は SFQ シフトレジスタ型メモリアレイと SRAM で構成されたメモリアレイの縦幅の比である．メモリアレイの面積は MAT の大部分を占めているため，本モデルではメモリアレイの縦幅，横幅の比を用いて SFQ 回路で実装された MAT の縦幅，横幅を表す．

$R_{\mathrm{array\_h}}$ を式 (6) で表す．

$$R_{\mathrm{array\_h}} = H_{\mathrm{DFF}}/2H_{\mathrm{SRAM}} \tag{6}$$

H_{DFF}, H_{SRAM} はそれぞれ SFQ 回路の DFF，SRAM の縦幅である．図 11 は 4×4 メモリアレイの場合における SRAM と SFQ シフトレジスタの対応を表している．縦一列の SRAM がループ構造を持つ SFQ シフトレジスタ 1 つに対応する（図 11 の青線で囲われている部分）．また，どちらのメモリアレイも図 11 の赤い破線で囲われている部分を 2×4 で配置しているとみなすことができる．よって，メモリアレイの縦幅の比は 1 つの DFF の縦幅と 2 つの SRAM の縦幅の比にほぼ等しい．

アレイの横幅 W を式 (7) で表す．

$$W = \begin{cases} N_{\mathrm{MAT\_W}} W_{\mathrm{MAT\_SFQ}} + N_{\mathrm{GW\_W}} P_{\mathrm{PTL}}, \\ (\mathrm{SFQ}) \\ N_{\mathrm{MAT\_W}} W_{\mathrm{MAT\_CMOS}} + N_{\mathrm{GW\_W}} P_{\mathrm{RC}} \\ + W_{\mathrm{interface}}, \\ (\mathrm{CMOS}) \\ N_{\mathrm{MAT\_W}} W_{\mathrm{MAT\_CMOS}} + N_{\mathrm{GW\_W}} P_{\mathrm{PTL}} \\ + N_{\mathrm{MAT\_W}} W_{\mathrm{interface}}, \\ (\text{ハイブリッド}) \end{cases} \tag{7}$$

$W_{\mathrm{MAT\_SFQ}}$, $W_{\mathrm{MAT\_CMOS}}$ はそれぞれ SFQ 回路，CMOS 回路と SRAM で実装した場合の MAT の横幅である．また，$N_{\mathrm{GW\_W}}$ は横に並んだ GW の本数である．CMOS キャッシュの場合は GW と SFQ プロセッサの境界にインタフェー

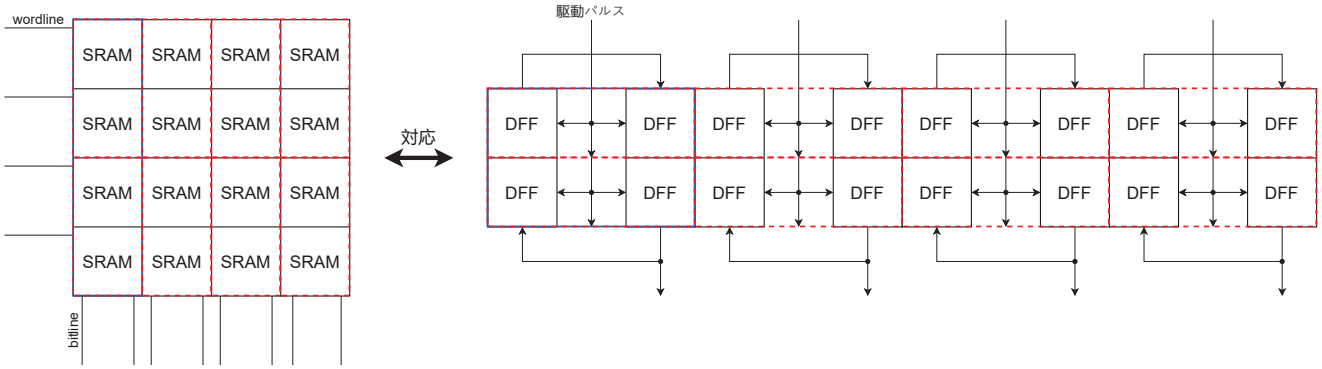


図 11 4 × 4 メモリアレイにおける SRAM と SFQ シフトレジスタの対応

ス回路が挿入される．そのため，インタフェース回路 1 つ分の横幅 $W_{\mathrm{interface}}$ が加算される．ハイブリッドキャッシュの場合，GW と MAT の境界にインタフェース回路が挿入される．そのため，$W_{\mathrm{interface}}$ と $N_{\mathrm{MAT\_W}}$ の積が加算される．

$W_{\mathrm{MAT\_SFQ}}$ を式 (8) で表す．

$$W_{\mathrm{MAT\_SFQ}} = R_{\mathrm{array\_w}} W_{\mathrm{MAT\_CMOS}} \tag{8}$$

$R_{\mathrm{array\_w}}$ は SFQ シフトレジスタ型メモリアレイと SRAM で構成されたメモリアレイの横幅の比である．

$R_{\mathrm{array\_w}}$ を式 (9) で表す．

$$R_{\mathrm{array\_w}} = (2W_{\mathrm{DFF}} + W_{\mathrm{SPL3}})/W_{\mathrm{SRAM}} \tag{9}$$

W_{DFF}, W_{SPL3} はそれぞれ SFQ 回路の DFF，3 出力スプリッタの横幅，W_{SRAM} は SRAM の横幅である．図 11 よりメモリアレイの横幅の比は 2 つの DFF の横幅と 1 つの 3 出力スプリッタの横幅の和と，SRAM の横幅の比にほぼ等しい．

MAT の遅延 T_{MAT} を式 (10) で表す．

$$T_{\mathrm{MAT}} = \begin{cases} T_{\mathrm{cycle\_SFQ}}(N_{\mathrm{sub}} + N_{\mathrm{shift}} + N_{\mathrm{mux}}), \\ (\mathrm{SFQ}) \\ T_{\mathrm{decoder}} + T_{\mathrm{wordline\_driver}} \\ + T_{\mathrm{bitline}} + T_{\mathrm{senseamp}}, \\ (\mathrm{CMOS, ハイブリッド}) \end{cases} \tag{10}$$

SFQ キャッシュの場合は，各構成要素の所要サイクル数の和と駆動パルスの周期 $T_{\mathrm{cycle\_SFQ}}$ の積となる．N_{sub}, N_{shift}, N_{mux} はそれぞれ減算器，シフトレジスタ，マルチプレクサの所要サイクル数である．ハイブリッド，CMOS キャッシュの場合，各構成要素の遅延の和となる．T_{decoder}, $T_{\mathrm{wordline\_driver}}$, T_{bitline}, T_{senseamp} はそれぞれデコーダ，ワードラインドライバ，ビットライン，センスアンプの遅延である．

N_{sub} は減算器の論理ゲート段数に等しい．本研究では，過去のシフトレジスタ型キャッシュのアクセス時間の見積もり [6] におけるモデル式を使用する（式 (11)）．

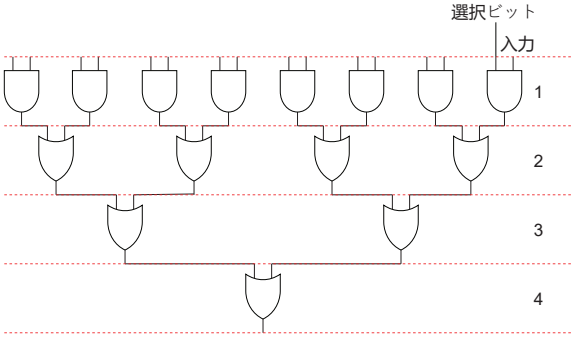


図 12 SFQ_MUX($Input_{\mathrm{mux}} = 8$)

$$N_{\mathrm{sub}} = 3 + 2\lceil \log_2(B_{\mathrm{addr}} - \log_2 N_{\mathrm{MAT}}) \rceil \tag{11}$$

B_{addr} はアドレスのビット数である．

N_{shift}（式 (12)）はシフトレジスタの動作回数に等しく，対象とするデータの位置によって動作回数は変化する．入出ポートから最も遠いワードラインに対象データがある場合（最悪ケース）はワードラインの本数 N_{wordline} だけシフトレジスタを動作させる必要がある．入出力ポートがあるワードラインに対象データがある場合（最良ケース）はシフトレジスタは動作しない．

$$N_{\mathrm{shift}} = \begin{cases} N_{\mathrm{wordline}}, & (\text{最悪ケース}) \\ 0, & (\text{最良ケース}) \end{cases} \tag{12}$$

N_{mux} はマルチプレクサの論理ゲート段数に等しい．マルチプレクサは AND ゲートとバイナリ・ツリー状の OR ゲートで構成されたもの（図 12）を想定し，論理ゲート段数を式 (13) で表す．

$$N_{\mathrm{mux}} = 1 + \log_2 Input_{\mathrm{mux}} \tag{13}$$

$Input_{\mathrm{mux}}$ はマルチプレクサの入力数である．

$T_{\mathrm{interface}}$ は SFQ/DC インタフェース回路の遅延 $T_{\mathrm{SFQ/DC}}$ と DC/SFQ インタフェース回路の遅延 $T_{\mathrm{DC/SFQ}}$ の和で表す（式 (14)）．

$$T_{\mathrm{interface}} = T_{\mathrm{SFQ/DC}} + T_{\mathrm{DC/SFQ}} \tag{14}$$

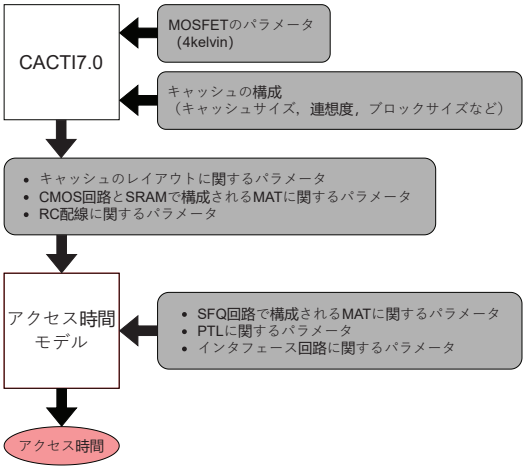


図 13 アクセス時間評価の流れ

表 1 MOSFET のパラメータ

温度	4 kelvin	300 kelvin
V_{dd} [V]	0.44	0.8
V_{th} [V]	0.228	0.1395
V_{tsat} [V]	0.123	0.0233
I_{on_n} [A/μm]	0.00625	0.0026264
I_{on_p} [A/μm]	0.00312	0.0013132
I_{off_n} [A/μm]	4.52^{-10}	1.22^{-07}
$I_{g\_on\_p}$ [A/μm]	3.36^{-12}	1.81^{-09}
Mobility_n [μm^2/Vs]	2.00^{11}	4.26^{10}
resistivity [$\mu\Omega$m]	0.008	0.018

5. アクセス時間の評価

5.1 評価方法

図 13 はアクセス時間の評価の流れである．キャッシュのレイアウト（MAT 数やサブアレイのワードライン，ビットライン数など），CMOS 回路と SRAM で構成される MAT に関するパラメータ，RC 配線に関するパラメータは CACTI 7.0 [11] を用いて算出し，アクセス時間モデルに入力する．CACTI 7.0 には 4 kelvin における MOSFET のパラメータと，キャッシュの構成を入力する．SFQ 回路の MAT, PTL, インタフェース回路に関するパラメータは直接モデルに入力する．

5.2 パラメータの設定

5.2.1 キャッシュの構成

本評価では，連想度は 1（ダイレクトマップ），アドレス長は 32 ビットに設定する．容量は 2 KB から 16 MB，ブロックサイズは 8 B から 64 B の範囲で評価する．ブロックサイズとデータの入力・出力ビット数は等しいとする．

5.2.2 MOSFET のパラメータ

本研究では CMOS 回路のプロセスサイズを 22 nm に設定している．表 1 は MOSFET のパラメータのうち 4 kelvin と 300 kelvin で値が異なるものをまとめたものである．

表 2 SFQ 回路，PTL，インタフェース回路のパラメータ

SFQ 回路のプロセス	1 μm	0.2 μm	22 nm
$T_{SFQ/DC}$ [ps]	100	100	100
$T_{DC/SFQ}$ [ps]	19	4.75	0.418
v_{PTL} [mm/ns]	100	100	100
$W_{interface}$ [μm]	1.3	1.3	1.3
T_{node} [ps]	7.9	1.975	0.1738
T_{cycle_SFQ} [ps]	20	5	0.44
H_{DFF} [μm]	30	6	0.66
W_{DFF} [μm]	30	6	0.66
W_{SPL3} [μm]	30	6	0.66
P_{PTL} [μm]	15	3	0.33

4 kelvin における MOSFET のパラメータは 4 kelvin 下での SFQ-CMOS ハイブリッドのスクラッチパッドメモリを提案した先行研究 [7] における値を用いる．

5.2.3 SFQ 回路，PTL，インタフェース回路のパラメータ

SFQ 回路のプロセスは現行の Nb（ニオブ）接合による 1 μm と，Nb 接合における動作周波数の上限とされている 0.2 μm，CMOS 回路と同等まで微細化すると想定した 22 nm の 3 つの場合を考える．表 2 は SFQ 回路，PTL，インタフェース回路のパラメータの一覧である．

1 µm

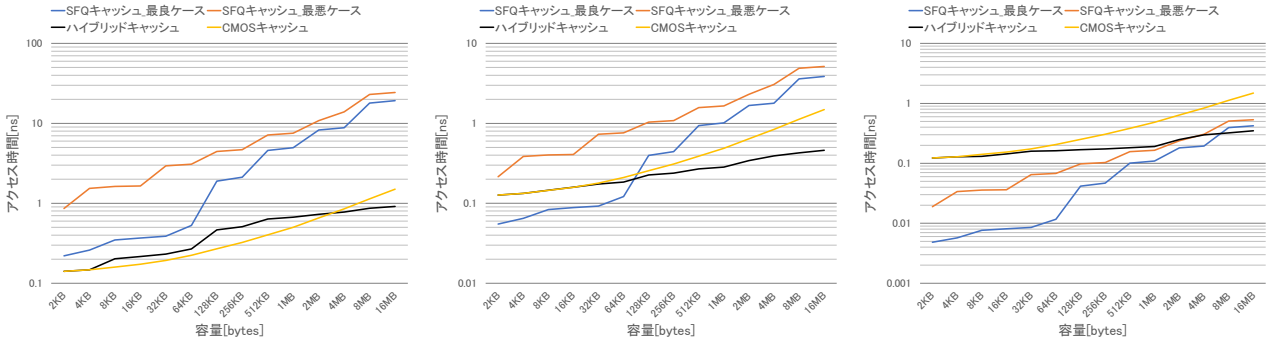
$T_{SFQ/DC}$ は 100 ps [12] と設定する．$T_{SFQ/DC}$ は 2 μm プロセスの LDDS の遅延の実測値が 38 ps [10] であるため，1 μm の場合は半分の 19 ps とする．v_{PTL} は真空中の光速の約 1/3 である 100 mm/ns とする．$W_{interface}$ は nTron の横幅とする [9]．T_{node} は CONNECT セルライブラリ [13], [14] のレシーバ，スプリッタ，ドライバの遅延の和とする．T_{cycle_SFQ} は，SFQ 回路の動作周波数を 50 GHz とし 20 ps に設定する．H_{DFF}, W_{DFF}, W_{SPL3}, P_{PTL} はそれぞれ CONNECT セルライブラリの DFF の縦幅，DFF の横幅，3 出力スプリッタの横幅，PTL のピッチである．

0.2 µm

現在 SFQ 回路の超伝導材料として用いられている Nb のまま微細化が進展しても Nb のギャップ電圧である 3.0 mV 以上はパルスは高くならず，パルス幅の縮小ならびに動作周波数の向上はそこで打ち止めになる．パルス高が 3.0 mV となるのは 0.2 μm プロセスであるとされている [15]．このとき，1 μm の場合から動作周波数は約 4 倍，回路の遅延は約 1/4 になると考えられる．また，セルの縦幅，横幅，PTL のピッチは 1 μm プロセスの場合の 1/5 とする．nTron は SFQ 回路ではないため，本評価では $T_{SFQ/DC}$, $W_{interface}$ は微細化によって変化しないとする．

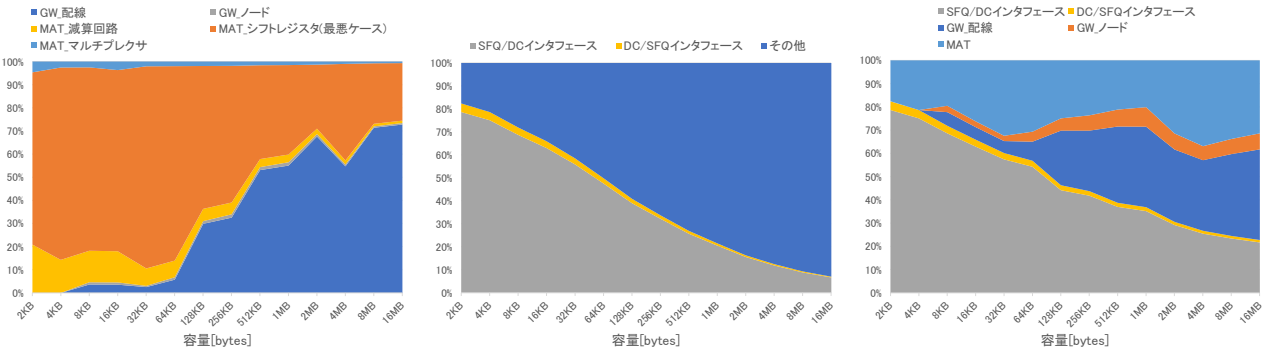
22 nm

0.2 μm よりも微細化を進める場合，Nb のままでは動作周波数は向上しない．また，臨界電流密度が



(a) SFQ 回路が 1 μm プロセスの場合　(b) SFQ 回路が 0.2 μm プロセスの場合　(c) SFQ 回路が 22 nm プロセスの場合

図 14　アクセス時間評価結果（ブロックサイズ: 32 B）



(a) SFQ キャッシュ　(b) CMOS キャッシュ　(c) ハイブリッドキャッシュ

図 15　SFQ 回路が 0.2μm プロセスの場合のアクセス時間の内訳（ブロックサイズ: 32 B）

250 kA/cm^2 よりも高いトンネル障壁作製プロセスを開発する必要がある．さらに，配線となる超伝導体を流れる電流密度の増加も問題となる可能性があり、十分に高い臨界電流密度を確保するなどといった課題もある．本評価では超伝導材料の変更などによってこれらの課題を解決でき，CMOS 回路と同等の 22 nm まで微細化した場合を考える．このとき，回路の遅延，セルの縦幅，横幅，PTL のピッチは 1 μm プロセスの場合の 22/1000 であるとする．

5.3　評価結果（ブロックサイズ: 32 B）

ブロックサイズが 32 B の場合のアクセス時間の評価結果を図 14 に示す．SFQ 回路が 1 μm プロセスのとき，SFQ キャッシュのアクセス時間は最良ケースの場合でも他の構成法より常にアクセス時間が大きい結果となった（図 14(a)）．0.2 μm プロセスの場合，64 KB 以下のキャッシュサイズで SFQ キャッシュの最良ケースが最小のアクセス時間である一方，最悪ケースは常にアクセス時間が最大であった（図 14(b)）．22 nm プロセスの場合，2 MB 以下のキャッシュサイズで SFQ キャッシュが最悪ケースであっても最小のアクセス時間であった（図 14(c)）．また，ハイブリッドキャッシュのアクセス時間は低容量の場合 CMOS キャッシュに対して同程度かそれ以上である一方，容量が大きくなるに従って CMOS キャッシュよりも小さくなる傾向が見られた．容量が大きいほど GW の遅延がアクセス時間全体に占める割合は高くなるため，GW を RC 配線ではなく PTL で実装したことによる伝送速度向上の効果が大きく表れたと考えられる．

図 15 は Nb 接合で動作周波数が上限になるとされている 0.2 μm まで SFQ 回路が微細化した場合の，各構成法のアクセス時間の内訳である．SFQ キャッシュの場合，シフトレジスタの動作数が最大であった場合（最悪ケース）の遅延がアクセス時間の大部分を占めている（図 15(a)）．シフトレジスタの最大動作数が少なくなるようなレイアウトの最適化や，シフトレジスタの入出力ポートに近い位置にアクセスされる可能性の高いデータを配置することが有効な解決策であると考えられる．また，SRAM のような直接データを選択するメモリアレイが SFQ 回路で実現されることでも解決しうる．CMOS，ハイブリッドキャッシュの場合，SFQ/DC インタフェース回路である nTron の遅延がアクセス時間の大部分を占めている（図 15(b)(c)）．低容量において特にこの傾向は強くなる．nTron の遅延の改善や，新しい低遅延 SFQ/DC インタフェース回路の実現が CMOS，ハイブリッドキャッシュのアクセス時間の削減に重要な役割を果たすと考えられる．

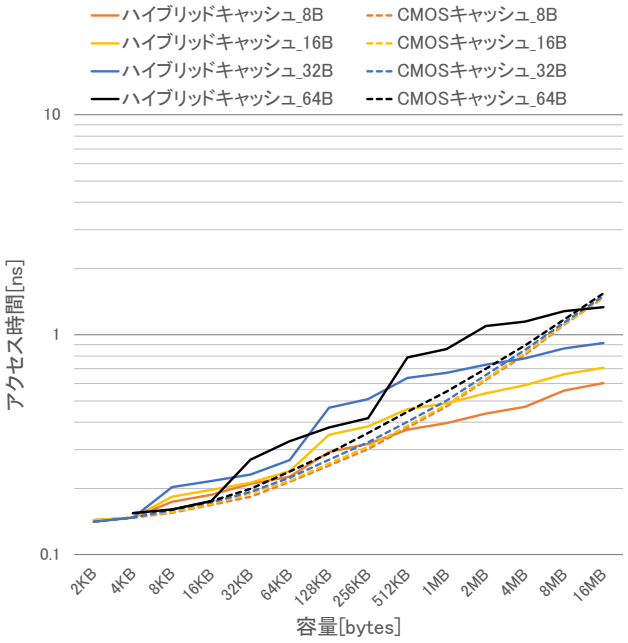


図 16 CMOS，ハイブリッドキャッシュのアクセス時間（SFQ 回路: 1 μm プロセス）

5.4 ブロックサイズがアクセス時間に与える影響

SFQ 回路が 1 μm プロセスの場合，ブロックサイズが大きいほどハイブリッドキャッシュのアクセス時間が CMOS キャッシュを下回る容量の範囲が狭くなることが分かった（図 16）．SFQ 回路が 0.2 μm, 22 nm プロセスの場合にはこの傾向は見られなかった．本評価ではブロックサイズとデータの入力・出力ビット数は等しいとしている．このとき，ブロックサイズが大きいほど GW の配線数が多くなる．また，1 μm プロセスの PTL のピッチは RC 配線と比べて非常に大きい．さらに，キャッシュの縦幅，横幅のうち GW 部分が占める部分は GW の配線数とピッチの積である（式 (4)(7)）．以上の理由により，1 μm かつブロックサイズが大きい場合，GW 部分の幅が CMOS キャッシュと比べて非常に広くなり，キャッシュの面積ならびに GW の配線長，遅延を大きくしたと考えられる．

6. おわりに

本研究では, SFQ プロセッサの高速動作に追従するキャッシュシステム探索を目的に, 4 kelvin で動作するキャッシュメモリの構成法を検討し，各構成法についてアクセス時間モデルを作成し評価を行った．

評価結果より，SFQ 回路が現行プロセスの場合は SFQ キャッシュよりもハイブリッド，CMOS キャッシュのアクセス時間が小さいが，CMOS 回路と同等のプロセスまで SFQ 回路が微細化したと仮定した場合，SFQ キャッシュのアクセス時間が 2 MB よりも容量が小さい場合は最小になると見積もられた．また，アクセス時間の内訳より，SFQ キャッシュはメモリアレイであるシフトレジスタの遅延が，CMOS，ハイブリッドキャッシュは SFQ/DC インタフェース回路の遅延が大部分を占めていることが明らかになった．

今後はタグアレイを含めたモデルや消費電力モデルを作成し，アクセス時間，面積，消費電力を総合的に考慮した SFQ プロセッサ向けキャッシュシステムの探索を行う予定である．

謝辞 本研究を進めるにあたり，活発な議論とご協力をいただいた九州大学井上研究室の皆様に心より感謝の意を表します．なお，本研究は一部，JST 未来社会創造事業 JPMJMI18E1，JST さきがけ JPMJPR21B3，JPMJPR2015 ならびに, 日本学術振興会科学研究費補助金または JSPS 科研費 JP18H05211，JP18J21274，JP22H05000，JP22K17868 の支援による．

参考文献

[1] Dennard, R., Gaensslen, F., Yu, H.-N., Rideout, V., Bassous, E. and LeBlanc, A.: Design of ion-implanted MOSFET's with very small physical dimensions, *IEEE Journal of Solid-State Circuits*, Vol. 9, No. 5, pp. 256–268 (online), DOI: 10.1109/JSSC.1974.1050511 (1974).

[2] Likharev, K. and Semenov, V.: 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 (online), DOI: 10.1109/77.80745 (1991).

[3] Nagaoka, I., Tanaka, M., Inoue, K. and Fujimaki, A.: A 48GHz 5.6mW Gate-Level-Pipelined Multiplier Using Single-Flux Quantum Logic, *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pp. 460–462 (2019).

[4] 田中雅光，石田浩貴，長岡一起，村瀬 健，佐野京佑，小野貴継，井上弘士，藤巻 朗: 単一磁束量子回路に基づくゲートレベル・パイプライン算術論理演算器の設計とエネルギー効率評価，技術報告 22 (2018).

[5] Ishida, K., Tanaka, M., Nagaoka, I., Ono, T., Kawakami, S., Tanimoto, T., Fujimaki, A. and Inoue, K.: 32 GHz 6.5 mW Gate-Level-Pipelined 4-Bit Processor using Superconductor Single-Flux-Quantum Logic, *2020 IEEE Symposium on VLSI Circuits*, pp. 1–2 (2020).

[6] Ishida, K., Tanaka, M., Ono, T. and Inoue, K.: Single-flux-quantum cache memory architecture, *2016 International SoC Design Conference (ISOCC)*, pp. 105–106 (online), DOI: 10.1109/ISOCC.2016.7799755 (2016).

[7] Zokaee, F. and Jiang, L.: *SMART: A Heterogeneous Scratchpad Memory Architecture for Superconductor SFQ-Based Systolic CNN Accelerators*, p. 912–924 (online), available from ⟨https://doi.org/10.1145/3466752.3480041⟩, Association for Computing Machinery (2021).

[8] Thoziyoor, S., Muralimanohar, N., Ahn, J. H. and Jouppi, N. P.: CACTI 5 . 1 (2008).

[9] McCaughan, A. N. and Berggren, K. K.: A Superconducting-Nanowire Three-Terminal Electrothermal Device, *Nano Letters*, Vol. 14, No. 10, pp. 5748–5753 (online), DOI: 10.1021/nl502629x (2014).

[10] Konno, G., Yamanashi, Y. and Yoshikawa, N.: Fully Functional Operation of Low-Power 64-kb Josephson-CMOS Hybrid Memories, *IEEE Transactions on*

plied Superconductivity, Vol. 27, No. 4, pp. 1–7 (online), DOI: 10.1109/TASC.2016.2646911 (2017).

[11] Balasubramonian, R., Kahng, A. B., Muralimanohar, N., Shafiee, A. and Srinivas, V.: CACTI 7: New Tools for Interconnect Exploration in Innovative Off-Chip Memories, *ACM Trans. Archit. Code Optim.*, Vol. 14, No. 2 (online), DOI: 10.1145/3085572 (2017).

[12] Tanaka, M., Suzuki, M., Konno, G., Ito, Y., Fujimaki, A. and Yoshikawa, N.: Josephson-CMOS Hybrid Memory With Nanocryotrons, *IEEE Transactions on Applied Superconductivity*, Vol. 27, No. 4, pp. 1–4 (online), DOI: 10.1109/TASC.2016.2646929 (2017).

[13] Yamanashi, Y., Kainuma, T., Yoshikawa, N., Kataeva, I., Akaike, H., Fujimaki, A., Tanaka, M., Takagi, N., Nagasawa, S. and Hidaka, M.: 100 GHz demonstrations based on the single-flux-quantum cell library for the 10 kA/cm2 Nb multi-layer process, *IEICE Transactions on Electronics*, Vol. E93-C, No. 4, pp. 440–444 (online), DOI: 10.1587/transele.E93.C.440 (2010).

[14] Yorozu, S., Kameda, Y., Terai, H., Fujimaki, A., Yamada, T. and Tahara, S.: A single flux quantum standard logic cell library, *Physica C: Superconductivity*, Vol. 378-381, pp. 1471 – 1474 (online), DOI: https://doi.org/10.1016/S0921-4534(02)01759-8 (2002).

[15] Patel, V. and Lukens, J.: Self-shunted Nb/AlO/sub x//Nb Josephson junctions, *IEEE Transactions on Applied Superconductivity*, Vol. 9, No. 2, pp. 3247–3250 (online), DOI: 10.1109/77.783721 (1999).