超低電圧リテンション SRAM のパワーゲーティング性能とアーキテクチャ

Power-gating performance and architecture of ultralow-voltage-retention SRAM

東工大未来研　矢野広気，塩津勇作，山本修一郎，菅原聡
H. Yano, Y. Shiotsu S. Yamamoto, and S. Sugahara, FIRST, Tokyo Inst. of Tech.
E-mail: yano.h.af@m.titech.ac.jp

【はじめに】キャッシュなどに用いられる SRAM の待機時電力の削減はマイクロプロセッサ(MP)やシステムオンチップ(SoC)といった CMOS ロジックシステムの重要な課題の1つになっている．また，近年注目を集めている SRAM を用いた Processing-in-memory (PIM)型ニューラルネットワーク・アクセラレータ(NNA)においても重要な課題となる．パワーゲーティング(PG)は CMOS ロジックシステムにおける効果的な待機時電力削減技術である[1]．しかし，SRAM では電源遮断を行うと，保持していたデータが消失するため，PG の実行機会に制約を生じ，十分に待機時電力を削減することが難しい．そこで，我々は超低電圧でデータを保持(ULVR)することで待機時電力を削減できる ULVR-SRAM を提案し，その PG 応用を検討してきた[2]．前回の報告では理想モデルを用いて ULVR-SRAM を用いたキャッシュの細粒度 PG の可能性について示した[2]．本報告では，各種 PG アーキテクチャ，メモリ容量，動作温度を考慮してシステマティックに解析した ULVR-SRAM の PG 性能について述べる．

【PG アーキテクチャ】今回，いくつかの PG アーキテクチャの検討を行ったが，本稿では最も実装が容易なアーキテクチャの結果について示す．システムが要求するスタンバイ時間を t_{SB0}，ULVR への移行を ENT，ULVR からの復帰を EXT と略記する．ENT 動作は t_{SB0} の開始から t_W 後に開始し，EXT 動作は t_{SB0} の終了後に行う．すなわち，t_{SB0} の開始と終了のタイミングの予測することなく ULVR-SRAM の PG を実行する．t_W および ENT 動作のレイテンシ t_{ENT} は t_{SB0} に含まれるが，EXT 動作のためのレイテンシ t_{EXT} はオーバヘッドとなる．一方，エネルギーオーバーヘッド(E_{EE})は ENT，EXT の両動作で発生する．

【解析方法】PG の評価指標には Break-even time (BET)を用いた．これは E_{EE} を埋め合わせることのできる最小の ULVR 時間として定義される．BET は用いる基準に依存するが，本研究では自身の待機時電力 P_{SB} を基準に用いた[3]．t_{SB0} の分布は正規分布を仮定し，その μ，σ をパラメータとした．μ は分布の平均値を，σ は分布の標準偏差である．正規分布の解析にはヘッダ・フッタパワースイッチ構成 ULVR-SRAM の 8kB マクロを用いた[4]．このマクロをサブアレイとして 32kB，256kB，2MB の容量を構成した．ULVR の ENT/EXT 動作はこのサブアレイごとに順次実行し，これにかかる t_{ENT}，t_{EXT} は ENT/EXT 動作にともなう突入電流に配線を共有しているサブアレイのリーク電流を考慮して，配線の電流許容値から決定した．

【解析結果】はじめに，t_W=0 として，t_{SB0}>BET+t_{ENT} (=BET′)の場合に ULVR を実行し，t_{SB0}<BET′のときにスタンバイ(SB)とする理想モデルを検討した．図 1 に容量が 2MB，温度 25℃における SB 時の消費エネルギーで規格化したエネルギーγ_eの μ，σ 依存性を示す．また，この図のいくつかの σ における μ–γ_e 特性を図 2 に示す．μ または σ が BET′ (=5.7μs)程度以上に大きくなると，ULVR によって効果的に待機時電力が削減されることがわかる．特に，μ(t_{SB0} の分布の中心)が BET′より小さくても，σ(分布の広がり)が BET′より大きい場合には待機時電力を削減できるようになる．図 3 にメモリ容量を変化させた場合の μ–γe 特性を示す．メモリ容量が増加すると ULVR 動作を待つサブアレイのリークの効果が顕在化し BET′が増加するため，μ が BET′程度以下の範囲では，γ_eは容量に応じて増加する．次に上述した PG アーキテクチャについて評価を行った．結果を図 4 に示す．等高線図は t_w=10^{-8}s の場合を，各白線は t_w=10^{-8}s–10^{-3}s で γ_e が 0.9 となる境界を示す．t_W が長い場合，γ_e を削減できる μ，σ の範囲は狭くなるが，t_w を BET′程度に設定しておくことで，上述の理想モデルとほぼ同等のエネルギー削減効果が得られることがわかる．ただし，μ，σ が小さなところでは，エネルギーが上昇を伴うため，アプリケーションによる t_{SB0} の分布を考慮する必要がある．動作温度が上昇するとリークが増加するが，E_{EE} はほぼ変化しないため，BET′は短くなるが，この簡単なアーキテクチャを用いて PG は可能となる．また，ULVR-SRAM を用いた PG は PIM 型 NNA にも効果的である．

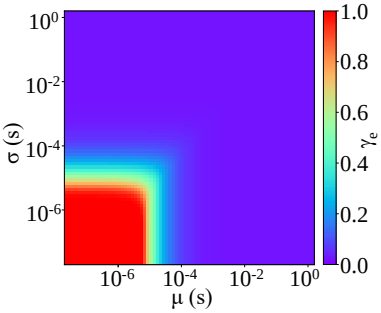

図 1. 2MB における等高線図

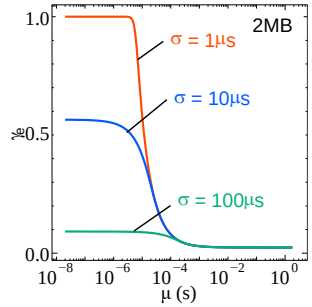

図 2. 2MB における γ_e の μ 依存性

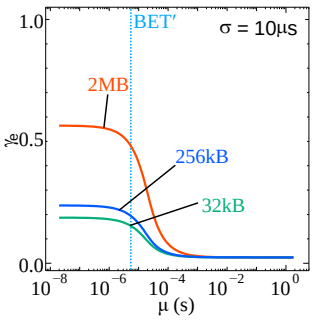

図 3. 各容量における μ 依存性

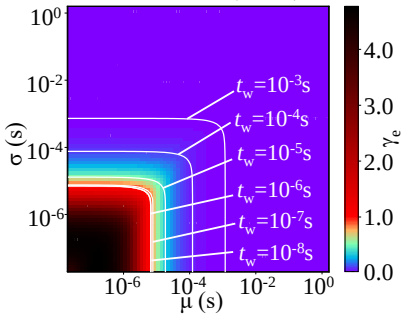

図 4. t_w=10^{-8} における性能

【謝辞】シミュレーションは東京大学大規模集積システム設計教育センターを通しシノプシス株式会社の協力で行われた．

【参考文献】[1] Y. Kanno *et al.*, IEEE JSSC **42**, 2007, 74. [2]吉田他，第 68 回応用物理学会春季学術講演会，17a-Z26-6. [3] D. Kitagata *et al.*, JJAP **58**, 2019, SBBB12. [4] H. Yoshida *et al.*, IEEE OJCAS **2**, 2021, 520.