ソースフォロワ読み出し・チャージシェアリングにより 32 ML & 1024 AL 並列で積和演算を行う 66 TOPS/W 強誘電体 FET CiM

松井 千尋[†1]　トープラサートポン カシディット[†1]
高木 信一[†1]　竹内 健[†1]

読み出し電流の On/Off 比が高い強誘電体 FET (FeFET)を用いた，電圧センス型 Computation-in-Memory (CiM)を提案した．AND フラッシュ回路のように接続された複数の FeFET セルに，ニューラルネットワークの重みを保存する．前段の pre-neuron の発火であるニューラルネットワークの入力は FeFET のソースに与える．Phase 1) ソースフォロワ読み出しによって入力と重みの積を 32 本並列に Multiply-line (ML)に読み出し，Phase 2) 1024 本の Accumulate-line (AL)の配線容量を並列にチャージシェアすることで積の結果を合計し，積和演算結果を得る．従来の電流センス型 CiM と異なり，提案の FeFET を用いた電圧センス型 CiM は積演算時に DC 電流が流れず，和演算時に電力を消費しないため，並列に MAC 演算を行うことができ，66 TOPS/W の高スループット・高エネルギー効率を実現できる．

66 TOPS/W FeFET CiM with Multiply-Accumulate by 32 ML & 1024 AL Parallel Source-follower Read and Charge-sharing

CHIHIRO MATSUI[†1]　KASIDIT TOPRASERTPONG[†1]
SHINICHI TAKAGI[†1]　KEN TAKEUCHI[†1]

A voltage-sensing computation-in-memory using ferroelectric FETs (FeFETs) with high on/off current ratio. To store the weights of the neural networks, multiple FeFETs are connected like AND-type flash array. As pre-neurons information of the neural networks, the input is given to the source-line of the FeFET. Multiply-accumulate (MAC) result is obtained by Phase 1) multiplying inputs and weights by source-follower read to 32 multiply-lines (MLs) in parallel and Phase 2) accumulating capacitance of parallel 1024 accumulate-lines (ALs) by charge-sharing. Unlike the conventional current-sensing CiM, DC current flows in multiply-phase, and no active power is consumed in accumulate-phase in the proposed voltage-sensing FeFET CiM. As a result, the proposed FeFET CiM achieves 66 TOPS/W high throughput and high energy efficiency by parallel MAC operation.

1. はじめに

強誘電体 FET (FeFET)を用いた Computation-in-Memory (CiM)は並列に電圧を読み出しチャージシェアする電圧検知により，低消費電力で Multiply-Accumulate (MAC)演算を実現できる[1]．FeFET は 3 端子素子であり，読み出し電流の On/Off 比が高い[2, 3]．従来の電流検知により積和演算を行うクロスバーアレイ型の CiM では，2 端子素子である ReRAM や PRAM を用いる．ニューラルネットワークの重みを ReRAM 等のコンダクタンスとして保存し，各メモリセルを流れる電流をキルヒホッフの電流則によって加算することで MAC 演算を行う[4]．電流検知による CiM は配線抵抗による IR drop が回路ノイズとなり，MAC 演算結果を劣化させる問題がある．一方，従来の電圧検知型の CiM である[5]は，ワード線 (Word-line, WL)の電圧 V_{WL} をステップ状にシフトすることで，各 FeFET に保存したニューラルネットワークの重みのしきい値電圧 V_{TH} を検知する．しかし，電圧シフトのために読み出し時間が長く，MAC 演算の性能は低い．

2. ソースフォロワ読み出し・チャージシェアリングにより並列・低消費電力で MAC 演算する FeFET CiM

ソースフォロワ読み出し・チャージシェアリングにより並列・低消費電力で MAC 演算する電圧検知型 FeFET CiM の動作を述べる．ニューラルネットワークの重みは FeFET に保存し，MAC 演算は Multiply および Accumulate を 2 段階で行う．このための FeFET CiM の単位回路である Local

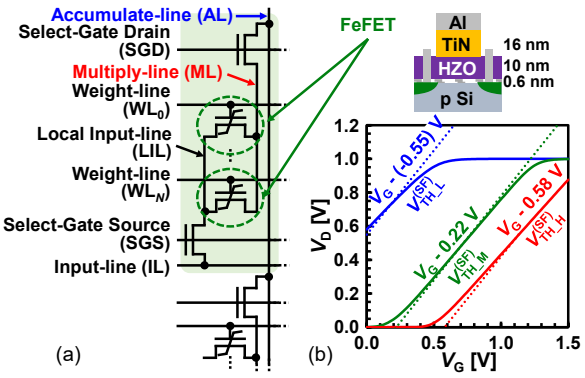

図 1 (a) FeFET CiM の単位回路 Local Multiply & Global Accumulate (LM-GA)アレイ．(b) FeFET のソースフォロワ読み出し特性 [1]

†1 東京大学大学院工学系研究科電気系工学専攻
Dept. of Electrical Engineering and Information Systems, Graduate School of Engineering, The University of Tokyo

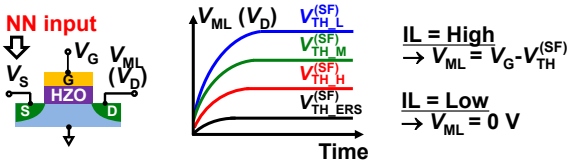

図 2 32 ML 並列ソースフォロワ読み出しによる Multiply

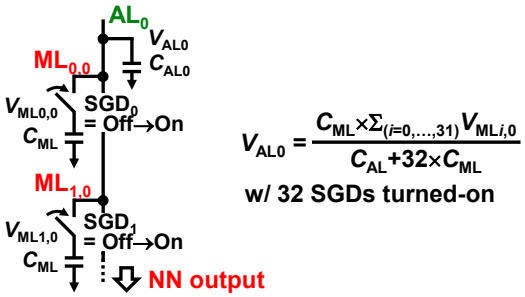

図 3 1024 AL 並列チャージシェアリングによる Accumulate

Multiply & Global Accumulate (LM-GA)アレイを図 1 に示す.

Phase 1) 32 ML 並列ソースフォロワ読み出しによる Multiply

Multiply 動作開始時, FeFET のソース側の選択ゲート SGS をオンにする．ニューラルネットワークの前段の発火情報は，LM-GA アレイの Input-line (IL)に入力する．前段が発火していれば，IL を通して FeFET のソースに電圧が入力され，FeFET に保存したニューラルネットワークの重みをソースフォロワでローカル Multiply-line (ML)に読み出す．ローカル ML には，FeFET に保存したニューラルネットワークの重みの値に応じた電圧が読み出され，入力と重みの multiply 結果を得る（図 2）．ML には DC 電流が流れないため並列動作が可能であり，32 本の IL に発火情報を入力することで, 32 本のローカル ML に並列に重み情報が読み出すことができる．一方で，前段が発火しなければ IL の電圧は低いため，ローカル ML の電位も低いまま保たれる．

Phase 2) 1024 AL 並列チャージシェアリングによる Accumulate

Accumulate 動作開始前に FeFET のソース側選択ゲート SGS をオフにする．このとき Phase 1 に従って，各 LM-GA アレイのローカル ML にはニューラルネットワークの重みの値に応じた電位が読み出されている．Accumulate 動作開始時に LM-GA アレイのドレイン側選択ゲート SGD を 32 個同時にオンにして，ローカル ML の配線容量とグローバル Accumulate-line (AL)の配線容量とをチャージシェアすることにより Accumulate される．これにより，MAC 演算結果が得られる（図 3）. このときのグローバル AL_j の電位 V_{ALj} = ($C_{ML}\times\Sigma V_{MLi,j}$)/($C_{ALj}$+32×$C_{ML}$)と得られる．チャージシェアでは電力を消費しないため，1024 本のグローバル AL を並列に動作できる．

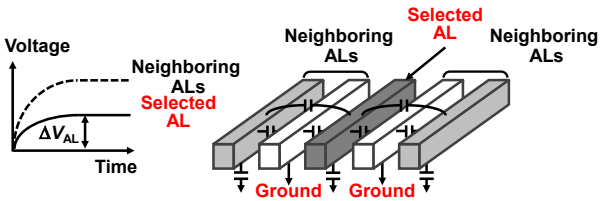

図 4 グローバル AL の容量結合ノイズを低減する AL shielding

3. 信頼性の高い推論のための回路ノイズ低減

図 1 に示す FeFET CiM では，グローバル AL の電位を読み出すことで MAC 演算結果を得る．隣接するグローバル AL に信号電圧が現れると，AL-AL 容量結合によって当該の AL 電圧に容量結合ノイズΔV_{AL}が発生する[6]. この結果，当該の AL 電圧が低下し，MAC 演算結果にエラーが生じる．グローバル AL の容量結合ノイズを低減するために，隣接するグローバル AL を接地する AL shielding を用いる（図 4）．AL shielding を用いることにより，グローバル AL の容量結合ノイズを低減できる．

4. おわりに

FeFET にニューラルネットワークの重みを保存し，ソースフォロワ読み出しおよびチャージシェアリングにより，並列に高エネルギー効率で MAC 演算を行う CiM を述べた．Phase 1 ではソースフォロワ読み出しによって重みに応じた電圧をローカル ML に読み出し, Multiply 演算を行う．Phase 2 ではローカル ML とグローバル AL とのチャージシェアリングにより，Accumulate 演算を行う．CiM に許容できる消費電力を考慮すると, FeFET CiM の ML は 32 並列，AL は 1024 並列で動作できる．また，FeFET のセル読み出し時間を 100 ns であるとき，66 TOPS/W と高い性能を達成する．これは，従来の電流検知 CiM [7]と比較して 64 倍電力効率が高い．また，AL shielding はグローバル AL の容量結合ノイズを低減し，推論時の MAC 演算精度を向上することができる．さらに，ニューラルネットワークの重みを保存する FeFET をソースフォロワで読み出すことにより，3 bit/cell を 10 年間のデータ保持 10 年可能であることを実験的に得た．

謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の業務委託の結果得られたものです．

参考文献

1) C. Matsui *et al.*, *VLSI Tech. Dig. Tech. Papers*, 2021, pp. 1-2.
2) J. Müller *et al.*, *VLSI Tech. Dig. Tech. Papers*, 2012, pp. 25-26.
3) K. Toprasertpong *et al.*, *IEDM Tech. Dig.*, 2019, pp. 570-573.
4) R. Mochida *et al.*, *VLSI Tech. Dig. Tech. Papers*, 2018, pp. 175-176.
5) K. Kamimura *et al.*, *Proc. ESSDERC*, 2019, pp. 178-181.
6) T. Sakurai, *IEEE TED*, vol. 40, no. 1, pp. 118-124, 1993.
7) S. Yu, *Proc. IEEE*, vol. 106, no. 2, pp. 260-285, 2018.