ULVR-SRAM を用いたニューラルネットワークアクセラレータの性能

Performance of a binarized neural network accelerator using ULVR-SRAM

東工大未来研　°塩津勇作, 菅原聡
°Y. Shiotsu, and S. Sugahara, *FIRST, Tokyo Inst. of Tech.*
E-mail: y.shiotsu@isl.titech.ac.jp

【はじめに】 将来のスマート社会ではモバイルエッジコンピューティングにおける AI 技術がさらに重要となり，これまで以上にエネルギー効率の高いニューラルネットワークアクセラレータ(NNA)が要求される．NNA の高性能化では processing-in-memory (PIM)アーキテクチャが有効である．特に，SRAM を用いた PIM 型 NNA は，現状の CMOS 技術で大規模なニューロンの集積化を実装することが可能なため，応用上極めて重要である．このような NNA のエネルギー性能を向上させるためには，エネルギー極小点となる駆動電圧(V_{Emin})を用いた動作時電力の削減と，パワーゲーティング(PG)による待機時電力の削減が重要になる．特に前者は NNA における積和(MAC)演算の大規模な並列化をも可能とする．最近，我々は CMOS のみで構成され，超低電圧(V_{UL})でもデータを失うことなく保持できる超低電圧リテンション SRAM (ULVR-SRAM)を提案した[1,2]．さらに ULVR-SRAM は，通常電圧下(V_{DD})での高性能 SRAM 動作，V_{Emin} でのエネルギー極小点 SRAM 動作が可能である($V_{UL} < V_{Emin} < V_{DD}$)．前回，ULVR-SRAM によって構成された PIM 型 2 値化ニューラルネットワーク(BNN)アクセラレータ(BNA)マクロを提案し，その待機時・動作時電力を効果的に削減できることを示した[3]．本発表では，この BNA マクロにおける MAC 演算の並列化と，その性能について議論する．

【BNA マクロの並列化】 開発した BNA マクロでは，その MAC 演算は図 1 に示す n-to-1 接続を基本として実装した．この場合では入力ベクトルの全要素と 1 つの次段ノードに対する重みとで XNOR を行い，その結果における“1”の数をカウント(PPC)するだけで MAC 演算を実行できる．これに整数バイアスを加えた結果の最上位ビット(MSB)によりニューロンの発火を判定できる．以下，図の n-to-1 構成における異なる色の MAC 演算を同時に行うことを in-layer parallelization (ILP)と呼ぶことにする．ILP の並列数(n_p)は BNA マクロのエネルギー性能によって制限され，エネルギー効率が高いほど大きくできる．

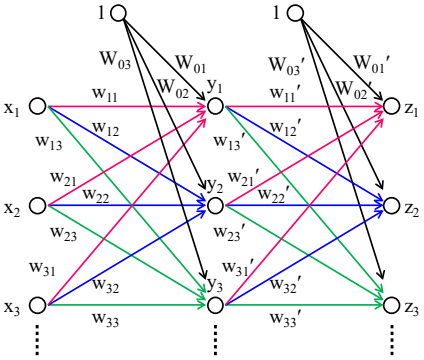

図 1. NN における n-to-1 接続

【解析結果】 解析には，ULVR-SRAM を用いた BNA マクロを用いた[3]．この BNA マクロではマルチポートセルを用いることなく読み出しの多重化が可能である．任意の構造の BNN は複数の BNA マクロを用いて構成できる(1 つのマクロあたり 256 ノードのニューロン層を実装できる)．また，このマクロでは通常電圧動作に加え V_{Emin} 動作と ULVR を用いた PG も可能である．このマクロのシミュレーションの結果から，任意構造の BNN の性能を予測できる．図 2 に BNA マクロにおける動作周波数 f，平均電力 P_{act}，1 サイクル当たりの消費エネルギーE の仮想電源電圧 VV_{DD} 依存性を示す．E は VV_{DD}=0.4V で極小となっている(V_{Emin}=0.4V)．この時，f は通常電圧下の 550MHz から 30MHz に低下するが，このとき P_{act} は 99%まで削減できる．これは V_{Emin} 動作を用いることで 100 程度の並列を行っても，1.2V 動作と同程度の消費電力しか生じないことを示している．図 3 に BNA マクロで構成した全結合層(ノード数 : 1024，層数 : 8)の動作時電力 P_{act}(W)，最大演算性能(TOPS)，演算効率 η (TOPS/W)を示す．図の横軸は ILP による並列数 n_p，実線は VV_{DD}=0.4V (=V_{Emin})で ILP を適用した場合である．また，参考のため，破線で n_p=1 における VV_{DD}=1.2V の場合について示す．n_p=1 のとき V_{Emin} 動作により 1.2V 動作時と比較して P_{act} を 1/100 程度まで削減できるが，演算性能は 1/10 程度にまで劣化する．n_p=16 とすることで(1 マクロあたり 4 並列)，1.2V 動作時と同程度の演算量で，P_{act} を 1.2V 動作時の約 1/10 にできる．また，n_p=128 にすると(1 マクロあたり 32 並列)，P_{act} は 1.2V 動作時と同程度であるが，演算量を 10 倍程度に増大できる．V_{Emin} 動作時の演算効率は~60TOPS/W と非常に高くできる．

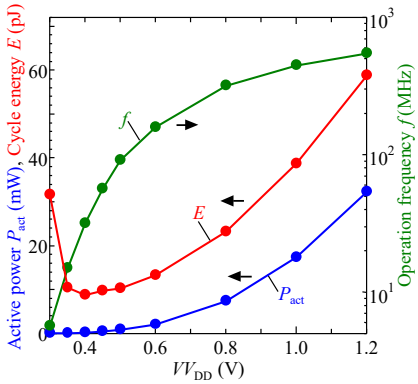

図 2. P_{act}, f, E の VV_{DD} 依存性

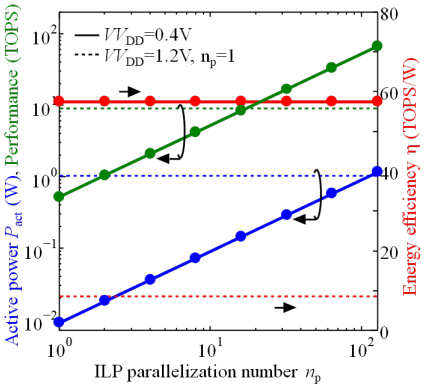

図 3. 各種性能の n_p 依存性

【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(VDEC)を通しシノプシス株式会社の協力で行われたものである．

【参考文献】 [1] H. Yoshida *et al.*, IEEE OJCAS **2**, 2021, pp. 520-533. [2]塩津他，第 68 回応用物理学会春季学術講演会，2021，17a-Z26-5. [3] 塩津他，第 82 回応用物理学会秋季学術講演会，2021，12p-N304-8.