様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ６ 月　 ６ 日現在

機関番号：１４５０１
研究種目：基盤研究(B)（一般）
研究期間：2018～2020
課題番号：１８Ｈ０１５００

研究課題名（和文）パイプラインバックプロパゲーションを用いたディープラーニングプロセッサ

研究課題名（英文）Deep Learning Processor for Pipelined Backpropagation

研究代表者
川口　博（KAWAGUCHI, Hiroshi）

神戸大学・科学技術イノベーション研究科・教授

研究者番号：００３６１６４２

交付決定額（研究期間全体）：（直接経費）　　13,000,000円

研究成果の概要（和文）：深層学習プロセッサ用デュアルポートSRAMを28nm FD-SOIプロセスによって実装した。画像データの読み出し動作にかかるエネルギを14.76%削減可能であることを確認した。
この技術を深層学習プロセッサのコードブック量子化用20トランジスタ超多ポートSRAMに拡張した。8ビットを16ビットに変換するルックアップテーブルとして機能する4kビットコードブックを40nmプロセスで試作した。モチーフとするNVIDIA NVDLAプロセッサにおいて20%のエネルギと26%の面積を削減した。

研究成果の学術的意義や社会的意義
IoTデバイスの低エネルギ画像認識の需要は機械学習により様々な分野で拡大している。カメラの高解像度化に反して、低エネルギ処理とリアルタイム性維持の両立が求められている。深層学習プロセッサは大量のパラメータと入出力を扱うため、大容量の内部SRAMが必要となり、シリコン面積の50%以上を占め、エネルギは外部DRAM帯域に支配される。精度を落とさずにメモリ帯域を削減する方法として量子化がある。コードブック方式は任意の非線形関数を表現でき、線形量子化よりも精度劣化を抑えることができる。この用途のために深層学習プロセッサのコードブック量子化用20トランジスタ超多ポートSRAMを設計、試作した。

研究成果の概要（英文）：Dual-port SRAM for a deep learning processor was implemented in a 28nm FD-SOI process. It was confirmed that the energy consumption required for the read operation of image data can be reduced by 14.76%.
This technology was expanded in a 20-transistor ultra-multiport SRAM for codebook quantization in a deep learning processor; a prototype 4k-bit codebook that functions as a lookup table to convert 8 bits to 16 bits was fabricated in a 40nm process. The codebook reduced energy by 20% and area by 26% in the motif processor, NVIDIA NVDLA.

研究分野：低消費電力回路設計

キーワード：深層学習　低消費電力プロセッサ　SRAM

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

物体検出やセグメンテーションに対して深層学習を用いた研究が進められ、自動運転やセキュリティをはじめとした自動化・安全安心への応用が期待されている。深層学習は主に畳み込み層で構成され、認識精度の向上のためにネットワーク規模が大きくなる傾向にある。しかし深層学習プロセッサにおいて畳み込み層は大量の重みパラメータを扱い、アクティベーションも巨大であるため、外部メモリとしての DRAM は排除できない。DRAM はオフチップであり、容量の大きな基板配線を介してプロセッサと外部接続され、かつリフレッシュによる記憶保持動作が必要となるため、消費エネルギが内部 SRAM に比べ 2 桁以上大きい。深層学習プロセッサでは 80%以上のエネルギが外部 DRAM で消費され、消費電力を支配する。このメモリ帯域削減が重要である。

２．研究の目的

深層学習プロセッサのメモリ帯域増大を抑制し、処理の高速化と低消費エネルギ化を図るため、モデル並列深層学習プロセッサの探求を行う。すなわち高いエネルギ効率で処理を行うのに必要なハードウェア要件を明らかにする。アルゴリズム・モデル並列アーキテクチャ・専用 SRAM 回路設計の融合による協調設計を行う。

３．研究の方法

(1) コードブック量子化アルゴリズム：多ポート SRAM を用いたコードブック量子化を行うことで、内部 SRAM に対するアクセスを維持したまま、外部 DRAM 帯域の削減の検討を行う。
(2) 並列アクセスに適したデュアルポート SRAM の開発:専用の読み出しアクセスポートを持ち、ディスターブフリーな読み出し動作を実現するデュアルポート SRAM を設計する。
(3) コードブック量子化用 20 トランジスタ超多ポート SRAM への拡張：1 つの書き込みポートと 8 つの読み出しポートを持つビットセルを設計する。
(4) 低エネルギ画像認識応用：上記技術の応用の提案として深層学習を用いた IoT アプリケーションを開発する。

４．研究成果

(1) コードブック量子化アルゴリズム：コードブック方式は任意の非線形関数を表現でき、線形量子化よりも精度劣化を抑えることができる。またバイアス・スケーリング・クリッピング・除算が不要なため、線形量子化よりも単純かつ高速に処理できる。量子化アルゴリズムとして Basin-Hopping と Differential Evolution を検討した。ResNet18 をベンチマークとしたネットワークにおいて、4 ビット以下の量子化では、図 1 に示すとおり Differential Evolution が優勢であった。Differential Evolution により 32 ビット浮動小数点重み（精度 91.90%）を 4 ビット量子化した場合、1.04%劣化（同 90.86%）に留まる。

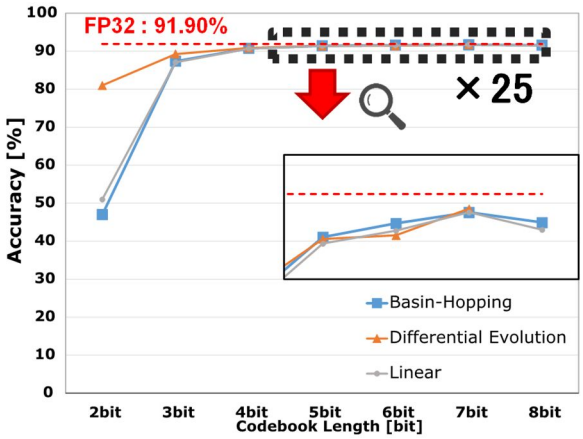


図 1 量子化アルゴリズムと精度劣化

(2) 並列アクセスに適したデュアルポート SRAM の開発：デュアルポート SRAM においてプリチャージ方式を用いた“0”データ読み出しの場合、ビット線(RBL)の引き抜き電流発生により RBL 充放電エネルギがサイクル毎に消費される。一方“1”データ読み出しの場合、RBL からソース線(SL)への電流は遮断される。このため入力データ中の“1”データ数を増やすことができれば RBL 充放電エネルギを削減し動作エネルギ効率を向上させることができる。しかしビット反転に多数決ロジックを用いると多数決回路と追加のフラグビットが必要となり、回路面積が増大しプロセッサ全体のコストが増加する。提案デュアルポート SRAM では図 2 に示すとおり、入力データの MSB を基準にビットデータの反転を判断する。

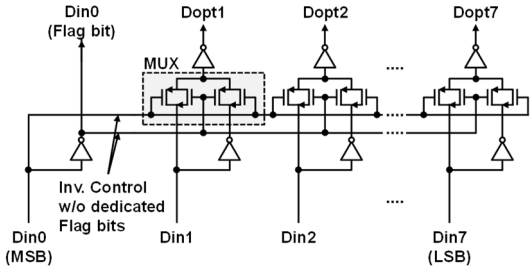


図 2 MSB 基準反転ビット線方式

これにより 8 ビットの従来型多数決ロジックにおいて従来回路に必要な 12.5%の面積オーバヘッドを 0.6%まで削減した。提案 SRAM は 28-nm FD-SOI プロセス技術によって実装され、MSB 基準反転ビット線方式により画像データの読み出し動作にかかるエネルギを 14.76%削減可能であることを確認した。さらに VGGF 畳み込みニューラルネットワークでの画像処理を考慮した場合、電力削減効果は 17.31%となる。

(3) コードブック量子化用 20 トランジスタ超多ポート SRAM への拡張：図 3 のように読み出しポートのうち 6 つは NMOS で構成され、残り 2 つを PMOS とした。

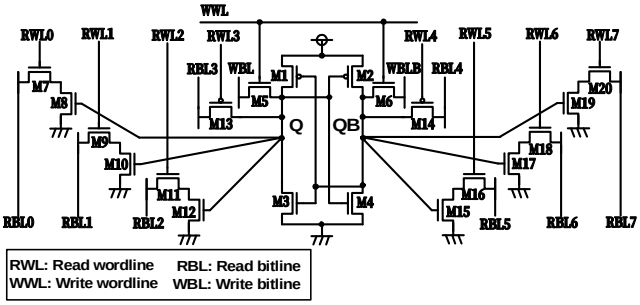
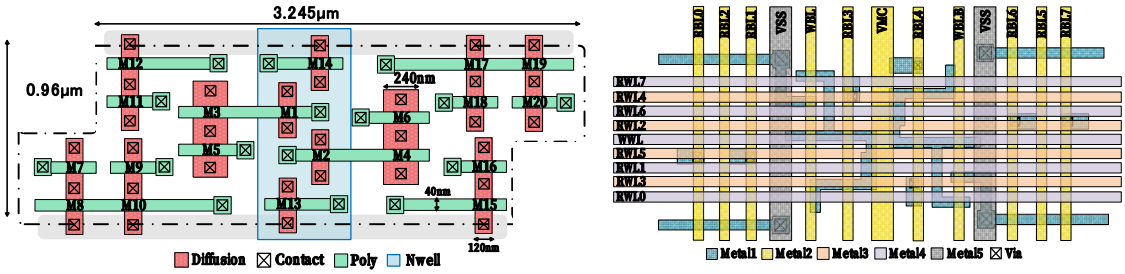


図 3 20 トランジスタ超多ポート SRAM ビットセル

こうすることで、全てを NMOS 読み出しポートとした場合に比べて、ビットセル面積を 28%削減することができる。ビットセルレイアウトを図 4 に示す。図 5 はテストチップ写真である。



(a) (b)

図 4 20 トランジスタ超多ポート SRAM ビットセルレイアウト：(a) FEOL と (b) BEOL

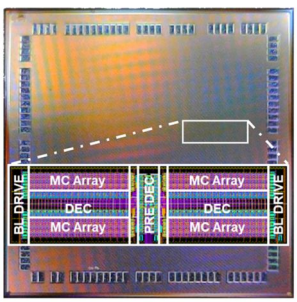


図 5 20 トランジスタ超多ポート SRAM テストチップ写真

40nm プロセスで 4k ビットを試作し、電源電圧 1.1V でアクセス時間 3ns、バイトあたり 2.7pJ のエネルギを達成した。コードブックは 8 ビットを 16 ビットに変換するルックアップテーブルとして機能し、外部メモリである DRAM の転送量と内部メモリ容量（キャッシュと畳み込みバッファ）を半減させることができる。これにより、図 6 に示すとおりモチーフとする NVIDIA NVDLA プロセッサにおいて 20%のエネルギと 26%の面積を削減した。

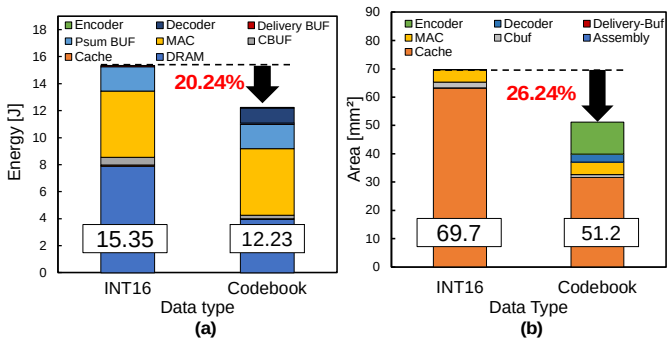


図 6 20 トランジスタ超多ポート SRAM コードブックによる(a)エネルギと(b)面積の削減効果

(4) 低エネルギ画像認識応用：サブミクロンレベルのレーザ加工形状推定技術を提案した。光学顕微鏡の撮像面を段階的に変化させる仕組みを搭載することで、撮像範囲と撮影条件を固定したまま焦点面のみを上下方向へ変化させた焦点画像スライスを用いる。レーザ顕微鏡による形状情報を教師データとし、SegNet で学習させた。アルミナセラミックス材料に対して、深さ方向 0.5 ミクロンごとに焦点画像群を撮影し形状推定を行った。溝形状（図 7)、リッジ形状（図 8）の両方において、サブミクロン精度の形状推定が可能であることを確かめた。

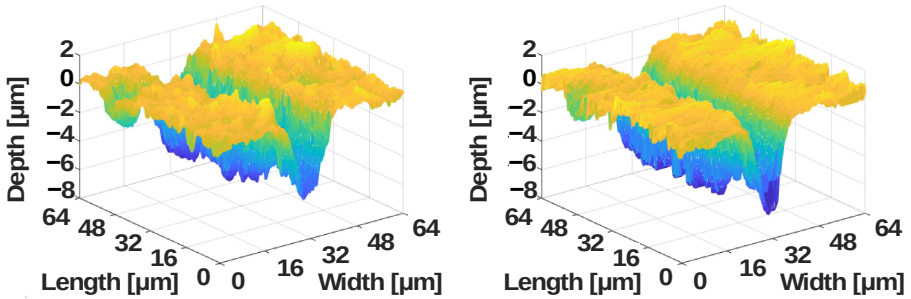


図 7 溝形状：レーザ顕微鏡による測定（左）と SegNet による形状推定（右）

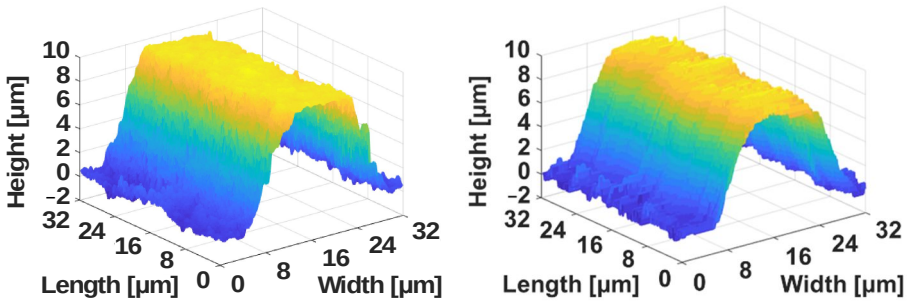


図 8 リッジ形状：レーザ顕微鏡による測定（左）と SegNet による形状推定（右）

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 H. Mori, T. Nakagawa, Y. Kitahara, Y. Kawamoto, K. Takagi, S. Yoshimoto, S. Izumi, H. Kawaguchi, and M. Yoshimoto	４．巻 66
２．論文標題 A 28-nm FD-SOI 8T Dual-Port SRAM for Low-Energy Image Processor with Selective Sourceline Drive Scheme	５．発行年 2019年
３．雑誌名 IEEE Transactions on Circuits and Systems I	６．最初と最後の頁 1442-1453
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TCSI.2018.2885536	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 R. Narukage, G. Okada, and H. Kawaguchi	４．巻 16
２．論文標題 Rapid Method Using Deep Learning with Multi-focus Microphotographs to Measure Submicrometric Structures and Its Evaluation	５．発行年 2021年
３．雑誌名 JLPS Journal of Laser Micro / Nanoengineering (JLMN)	６．最初と最後の頁 150-154
掲載論文のDOI（デジタルオブジェクト識別子） 10.2961/jlmn.2021.02.3001	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計7件（うち招待講演　1件／うち国際学会　3件）

１．発表者名 大原　遼太郎、川口　博
２．発表標題 テーブル参照を用いたTernary圧縮・伸長アルゴリズムの検討
３．学会等名 情報論的学習理論ワークショップ(IBIS)
４．発表年 2019年

１．発表者名 成影　力、岡田　穣治、川口　博
２．発表標題 顕微鏡による焦点画像群を用いたレーザ加工溝形状の高速推定手法
３．学会等名 レーザ加工学会講演会
４．発表年 2019年

１．発表者名
R. Kawamoto, M. Taichi, M. Kabuto, D. Watanabe, S. Izumi, M. Yoshimoto, and H. Kawaguchi

２．発表標題
R. Kawamoto, M. Taichi, M. Kabuto, D. Watanabe, S. Izumi, M. Yoshimoto, and H. Kawaguchi, "A 1.15-TOPS 6.57-TOPS/W DNN Processor for Multi-Scale Object Detection

３．学会等名
IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)（国際学会）

４．発表年
2020年

１．発表者名
R. Narukage, G. Okada, and H. Kawaguchi

２．発表標題
Rapid method using deep learning with multi-focus microphotographs to measure submicrometric structures

３．学会等名
DGM/JLPS International Symposium on Laser Precision Microfabrication (LPM)（国際学会）

４．発表年
2020年

１．発表者名
川口　博

２．発表標題
メモリ容量と帯域幅削減のための分散深層学習ハードウェア

３．学会等名
miniCANDARシンポジウム（招待講演）

４．発表年
2019年

１．発表者名
森陽紀、陽川哲也、和泉慎太郎、吉本雅彦、川口博、井上敦樹

２．発表標題
分散深部学習におけるメモリと帯域幅削減のためのレイヤーブロックワイズパイプライン

３．学会等名
LSIとシステムのワークショップ

４．発表年
2018年

１．発表者名 H. Mori, S. Izumi, H. Kawaguchi, and M. Yoshimoto
２．発表標題 28-nm FD-SOI Dual-Port SRAM with MSB-Based Inversion Logic for Low-Power Deep Learning
３．学会等名 IEEE International Conference on Electronics, Circuits, and Systems (ICECS)（国際学会）
４．発表年 2018年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称	発明者	権利者
顕微鏡による焦点画像群を用いた形状計測方法及び装置	2019	同左
産業財産権の種類、番号	**出願年**	**国内・外国の別**
特許、特願2019-214798	2019年	国内

〔取得〕　計0件

〔その他〕

神戸大学大学院科学技術イノベーション研究科アーキテクチャ研究室 https://www28.cs.kobe-u.ac.jp/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関