先端科学技術研究科　修士論文要旨

所属研究室 （主指導教員）	コンピューティング・アーキテクチャ （中島　康彦　（教授））		
学籍番号	2111229	提出日	令和 5年 1月 13日
学生氏名	船井　遼太朗		
論文題目	列多重化線形CGRAによる疎行列疎行列積の高速化と評価		

要旨

近年、科学技術計算や、自然言語処理を中心とする機械学習などでの大規模な計算に対するエネルギー消費量増加は最大の課題の1つとなっている。これらのアプリケーションでは、計算カーネルとして多くの要素がゼロ要素で占められている疎行列同士の行列積演算(SpGEMM)が用いられており、疎行列の非ゼロ要素の不規則さによるメモリアクセスの増加がボトルネックとなっている。現在、行列を特定のフォーマットへ格納しGPGPUでの並列処理による高速化や、専用ハードウェア(Domain Specific Architecture:DSA)による演算の最適化などが試みられているが、前者は演算省略のメリットが行列の圧縮／再整形やメモリ確保の前処理などの形式的な処理に食われてしまい、SpGEMMのエネルギーと演算能力を最適化することは困難であり、後者は汎用性に欠け最新の圧縮formatを適用できないなど、構造的な問題を抱えている。そこで上述の問題を解決できるアーキテクチャとして、Coarse Grain Reconfigurable Array (CGRA) 注目されている。本論文では、CGRA型アクセラレータInterposed Memory Accelerator eXtension series 2(IMAX2)へ疎行列-疎行列演算(SpGEMM)機構を追加実装し、それを補足するformatを設定することによるSpGEMMの高効率な演算手法を提案する。実機検証として、理想的な多重対角行列と、実アプリケーションに用いられたUniversity of FloridaのSuiteSparse Matrix Collectionからそれぞれ10種類の行列データを使用し、提案アーキテクチャの実機(FPGAベースのMPSoC)とデザインコンパイラによるASICエミュレーション、およびリファレンスとしてGPGPU(RTX3090、Jetson AGX Orin)による疎行列計算ライブラリ(cuSPARSE)での最適な演算手法を使用した。その結果追加したSpGEMMユニットはcgra全体の面積の約0.24%であり、さらに、電力増加はわずか0.30%で機能追加できたことが確認でき、さらに、アーキテクチャ全体ではRTX3090、Jetson AGX Orinのそれぞれ2%、12%程の消費電力となることを確認した。性能比較の指標にはSpGEMM実行時のエネルギー遅延積(EDP)を用い、10K行列ではJetson AGX Orinに対し16.7倍の、RTX3090に対し6.31倍の効率が達成できていることが確認でき、同一サイズのMatrix Collectionの行列ではJetson AGX Orinに対し7.1倍の、RTX3090に対し3.03倍の効率となることが確認できた。さらに、行列サイズ2048の時、Sparsityが60%までカーネルによる高速化が有効であり、これはcuSPARSEの対応するSparsityである95%よりも適用範囲が改善したことを確認できた。