【公報種別】特許法第１７条の２の規定による補正の掲載
【部門区分】第６部門第３区分
【発行日】令和4年7月13日(2022.7.13)

【国際公開番号】WO2020/190809
【公表番号】特表2022-523761(P2022-523761A)
【公表日】令和4年4月26日(2022.4.26)
【年通号数】公開公報(特許)2022-075
【出願番号】特願2021-547450(P2021-547450)
【国際特許分類】
Ｇ０６Ｆ 17/16(2006.01)
Ｇ０６Ｔ 1/20(2006.01)
Ｇ０６Ｆ 9/38(2006.01)
Ｇ０６Ｆ 15/80(2006.01)
Ｇ０６Ｆ 17/10(2006.01)
【ＦＩ】
Ｇ０６Ｆ 17/16 Ｐ
Ｇ０６Ｔ 1/20 Ａ
Ｇ０６Ｆ 9/38 ３７０Ｃ
Ｇ０６Ｆ 15/80
Ｇ０６Ｆ 17/10 Ａ

【手続補正書】
【提出日】令和4年7月5日(2022.7.5)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
キャッシュメモリと結合される複数の処理リソースを含み、少なくとも１つの処理リソースが行列アクセラレータを含み、該行列アクセラレータは、スパース内積命令に応答してスパース第１行列及び第２行列の複数の要素に対して内積演算を実行するよう構成され、前記スパース第１行列の要素は、要素の組を含む圧縮表現に圧縮され、前記要素の組は、少なくとも１つの非ゼロ値要素及び該少なくとも１つの非ゼロ値要素の指示を含む、計算クラスタを有し、
前記圧縮表現は、圧縮された形式で前記キャッシュメモリに格納され、
前記少なくとも１つの処理リソースは、
前記圧縮表現を前記キャッシュメモリから前記少なくとも１つの処理リソース内のメモリにロードし、
前記第２行列を前記キャッシュメモリから前記少なくとも１つの処理リソース内の前記メモリにロードし、
前記圧縮表現からの要素及び前記第２行列の選択された要素に対して前記内積演算を実行し、前記第２行列の前記選択された要素が、前記圧縮表現内に格納された前記スパース第１行列の非ゼロ値と対応し、前記少なくとも１つの非ゼロ値の前記指示に基づき選択され、
前記内積演算の出力を前記少なくとも１つの処理リソース内の前記メモリに書き込むよう構成される、
汎用グラフィクスプロセッサ。

【請求項２】
前記キャッシュメモリは、レベル２（Ｌ２）キャッシュメモリである、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項３】
前記少なくとも１つの処理リソース内の前記メモリは、レベル１（Ｌ１）キャッシュメモリを含む、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項４】
前記少なくとも１つの処理リソース内の前記メモリは、共有メモリを含む、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項５】
前記少なくとも１つの処理リソース内の前記メモリは、レジスタファイルを含む、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項６】
前記少なくとも１つの処理リソース内の前記メモリは、前記行列アクセラレータ内のメモリを含む、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項７】
前記スパース第１行列は、ニューラルネットワークに関連した重みデータを含む、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項８】
前記第２行列は、前記ニューラルネットワークに関連した入力活性化データを含む、
請求項７に記載の汎用グラフィクスプロセッサ。
【請求項９】
前記内積演算の前記出力は、前記ニューラルネットワークに関連した出力活性化データを含む、
請求項８に記載の汎用グラフィクスプロセッサ。
【請求項１０】
前記内積演算の前記出力は、密行列である、
請求項９に記載の汎用グラフィクスプロセッサ。
【請求項１１】
前記行列アクセラレータは、処理要素のシストリックアレイを含む、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項１２】
前記スパース第１行列は、構造化されたスパース性を有する、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項１３】
前記スパース第１行列の要素は、前記構造化されたスパース性に基づき圧縮表現に圧縮される、
請求項１２に記載の汎用グラフィクスプロセッサ。
【請求項１４】
前記内積演算は、８ビット整数内積演算である、
請求項１に記載の汎用グラフィクスプロセッサ。
【請求項１５】
前記スパース第１行列は、８ビット整数要素を含む、
請求項１４に記載の汎用グラフィクスプロセッサ。
【請求項１６】
メモリデバイスと、
請求項１乃至１５のうち何れか一項に記載の汎用グラフィクスプロセッサと
を有するデータ処理システム。

【請求項１７】
スパース内積命令に応答してスパース第１行列及び第２行列の複数の要素に対して内積演算を実行することであり、前記内積演算は、キャッシュメモリと結合される複数の処理リソースを含む計算クラスタにより実行され、少なくとも１つの処理リソースは行列アクセラレータを含み、前記スパース第１行列の要素は、要素の組を含む圧縮表現に圧縮され、前記要素の組は、少なくとも１つの非ゼロ値要素及び該少なくとも１つの非ゼロ値要素の指示を含む、前記実行することと、
前記圧縮表現を、圧縮された形式で前記キャッシュメモリに格納することと、
前記少なくとも１つの処理リソースにより、
前記圧縮表現を前記キャッシュメモリから前記少なくとも１つの処理リソース内のメモリにロードし、
前記第２行列を前記キャッシュメモリから前記少なくとも１つの処理リソース内の前記メモリにロードし、
前記圧縮表現からの要素及び前記第２行列の選択された要素に対して前記内積演算を実行し、前記第２行列の前記選択された要素が、前記圧縮表現内に格納された前記スパース第１行列の非ゼロ値と対応し、前記少なくとも１つの非ゼロ値の前記指示に基づき選択され、
前記内積演算の出力を前記少なくとも１つの処理リソース内の前記メモリに書き込むことと
を有する方法。
【請求項１８】
前記スパース第１行列の要素を、前記少なくとも１つの処理リソースのメモリ内で、前記圧縮表現に圧縮することを更に有する、
請求項１７に記載の方法。
【請求項１９】
前記スパース第１行列は、構造化されたスパース性を有し、
前記スパース第１行列の要素は、前記構造化されたスパース性に基づき圧縮表現に圧縮される、
請求項１７に記載の方法。
【請求項２０】
内積演算を実行することは、８ビット整数内積演算を実行することを含む、
請求項１７に記載の方法。
【請求項２１】
前記スパース第１行列は、８ビット整数要素を含む、
請求項２０に記載の方法。
【請求項２２】
実行される場合にマシンに請求項１７乃至２１のうち何れか一項に記載の方法を実行させるコンピュータプログラム。
【請求項２３】
請求項２２に記載のコンピュータプログラムを記憶しているマシン可読記憶媒体。
【請求項２４】
請求項１７乃至２１のうち何れか一項に記載の方法を実行する手段を有する装置。