(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7550614号
(P7550614)

(45)発行日　令和6年9月13日(2024.9.13)

(24)登録日　令和6年9月5日(2024.9.5)

(51)国際特許分類

Ｇ０６Ｆ　12/00　(2006.01)	Ｇ０６Ｆ	12/00	５６０Ｆ
Ｇ０６Ｆ　9/34　(2018.01)	Ｇ０６Ｆ	9/34	３５０Ａ
Ｇ０６Ｆ　12/06　(2006.01)	Ｇ０６Ｆ	9/34	３５０Ｂ
Ｇ１１Ｃ　5/04　(2006.01)	Ｇ０６Ｆ	12/06	５４０Ｅ
Ｈ１０Ｂ　12/00　(2023.01)	Ｇ１１Ｃ	5/04	２２０

ＦＩ

請求項の数　22　（全32頁）　最終頁に続く

(21)出願番号　特願2020-191783(P2020-191783)
(22)出願日　令和2年11月18日(2020.11.18)
(65)公開番号　特開2021-128752(P2021-128752A)
(43)公開日　令和3年9月2日(2021.9.2)
審査請求日　令和5年9月22日(2023.9.22)
(31)優先権主張番号　62/975,577
(32)優先日　令和2年2月12日(2020.2.12)
(33)優先権主張国・地域又は機関
米国(US)
(31)優先権主張番号　16/859,829
(32)優先日　令和2年4月27日(2020.4.27)
(33)優先権主張国・地域又は機関
米国(US)
早期審査対象出願

(73)特許権者　390019839
三星電子株式会社
Ｓａｍｓｕｎｇ　Ｅｌｅｃｔｒｏｎｉｃｓ　Ｃｏ．，Ｌｔｄ．
大韓民国京畿道水原市霊通区三星路１２９
１２９，Ｓａｍｓｕｎｇ－ｒｏ，Ｙｅｏｎｇｔｏｎｇ－ｇｕ，Ｓｕｗｏｎ－ｓｉ，Ｇｙｅｏｎｇｇｉ－ｄｏ，Ｒｅｐｕｂｌｉｃ　ｏｆ　Ｋｏｒｅａ
(74)代理人　110000051
弁理士法人共生国際特許事務所
(72)発明者　マラディ，クリシュナ　テジャ
アメリカ合衆国，９５１３５　カリフォルニア州，サンノゼ，ロートレック　ド
最終頁に続く

(54)【発明の名称】　インメモリコンピューティングに対するデータ配置のための方法及びその方法が適用されたメモリモジュール

(57)【特許請求の範囲】

【請求項１】

メモリモジュールであって、

ダイナミックランダムアクセスメモリ（ＤＲＡＭ）バンクを含むメモリダイと、

ホストプロセッサからオペランド及び命令を受信するメモリコントローラと、を備え、

前記ＤＲＡＭバンクは、

複数のページに配列されたＤＲＡＭセルのアレイと、

算術論理ユニット（ＡＬＵ）を含むインメモリコンピューティング（ＩＭＣ）モジュールと、を含み、

前記メモリコントローラは、

前記命令に基づいて、複数のデータレイアウトの中から、前記ＤＲＡＭセルのアレイの前記複数のページから前記ＩＭＣモジュールへの前記オペランドの配置を指定するためのデータレイアウトを決定し、

前記決定されたデータレイアウトに従って、前記ＤＲＡＭバンクに前記オペランドを供給し、

前記命令に従って、前記オペランドに対して前記ＡＬＵによる演算を実行するように、前記ＤＲＡＭバンクの前記ＩＭＣモジュールを制御するように構成され、

前記ＩＭＣモジュールは、オペランドレジスタを更に含み、

前記オペランドは、第１のオペランド及び第２のオペランドを含み、

前記複数のデータレイアウトの中から一つのオペランド（１ＯＰ）データレイアウトを

決定した場合、
前記第１のオペランドは、前記ＤＲＡＭバンクの外部から前記ＩＭＣモジュールに供給され、
前記第２のオペランドは、前記ＤＲＡＭセルから前記オペランドレジスタを介して前記ＩＭＣモジュールに供給されることを特徴とするメモリモジュール。
【請求項２】
前記第１のオペランドは、複数の第１のタイルに分割され、
前記第２のオペランドは、複数の第２のタイルに分割され、
各タイルは、複数の値を含み、
前記メモリコントローラは、
前記オペランドレジスタに前記第１のオペランドの第１のタイルを格納し、
前記オペランドレジスタに格納された前記第１のオペランドの第１のタイル、及び前記第２のオペランドの複数の第２のタイルの各々に対して、前記ＡＬＵによる演算を実行するように更に構成されることを特徴とする請求項１に記載のメモリモジュール。
【請求項３】
前記第１のオペランドは、複数の第１のタイルに分割され、
前記第２のオペランドは、複数の第２のタイルに分割され、
各タイルは、複数の値を含み、
前記複数のデータレイアウトの中から同じページ（ＳＲ）データレイアウトを決定した場合、
前記メモリコントローラは、前記ＤＲＡＭセルの同じページに、一つ以上の前記第１のタイル及び一つ以上の前記第２のタイルを格納するように構成されることを特徴とする請求項１に記載のメモリモジュール。
【請求項４】
前記メモリコントローラは、
前記オペランドレジスタに前記一つ以上の前記第１のタイルの中の一つの第１のタイルを格納し、
前記オペランドレジスタに格納された前記第１のタイル、及び前記ＤＲＡＭセルのアレイの前記第１のタイルと同じページに格納された前記一つ以上の前記第２のタイルのそれぞれに対して、前記ＡＬＵによる演算を実行するように更に構成されることを特徴とする請求項３に記載のメモリモジュール。
【請求項５】
前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、アキュムレータを更に含み、
前記アキュムレータは、累積値を格納するように構成されたアキュムレータレジスタを含み、
前記アキュムレータは、
前記ＡＬＵによって演算された出力を受信し、
前記累積値と前記出力との合計で前記アキュムレータレジスタを更新するように構成され、
前記命令は、前記第１のオペランドと前記第２のオペランドとの内積を計算することを含み、
前記第１のタイルの中の一つの第１のタイルは、行データを格納し、
前記第２のタイルの中の一つの第２のタイルは、列データを含むことを特徴とする請求項４に記載のメモリモジュール。
【請求項６】
前記第１のタイルは、第１の数の値を有し、
前記第２のタイルは、第２の数の値を有し、
前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、
前記出力バッファは、前記第１の数の値と前記第２の数の値との積以上を格納するためのサイズを有し、

前記命令は、前記第１のオペランドと前記第２のオペランドとの外積を計算することを含み、
前記第１のタイルの中の一つの第１のタイルは、行データを格納し、
前記第２のタイルの中の一つの第２のタイルは、列データを含むことを特徴とする請求項４に記載のメモリモジュール。
【請求項７】
前記第１のタイルは、第１の数の値を有し、
前記第２のタイルは、第２の数の値を有し、
前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、
前記出力バッファは、前記第１の数の値及び前記第２の数の値の中の大きい方の値以上を格納するためのサイズを有し、
前記命令は、前記第１のオペランドと前記第２のオペランドとのテンソル積を計算することを含み、
前記第１のタイルの中の一つの第１のタイルは、行データを格納し、
前記第２のタイルの中の一つの第２のタイルは、列データを含むことを特徴とする請求項４に記載のメモリモジュール。
【請求項８】
前記第１のオペランドは、複数の第１のタイルに分割され、
前記第２のオペランドは、複数の第２のタイルに分割され、
各タイルは、複数の値を含み、
前記複数のデータレイアウトの中から異なるページ（ＤＲ）データレイアウトを決定した場合、
前記メモリコントローラは、
前記ＤＲＡＭセルのアレイの第１のページに前記第１のタイルのサブセットを格納し、
前記ＤＲＡＭセルのアレイの第２のページに前記第２のタイルのサブセットを格納するように構成されることを特徴とする請求項１に記載のメモリモジュール。
【請求項９】
前記メモリコントローラは、
前記オペランドレジスタに前記第１のページから前記第１のオペランドの第１のタイルを格納し、
前記オペランドレジスタに格納された前記第１のオペランドの第１のタイル、及び前記第２のページからの前記第２のオペランドの複数の第２のタイルのそれぞれに対して、前記ＡＬＵによる演算を実行するように更に構成されることを特徴とする請求項８に記載のメモリモジュール。
【請求項１０】
前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、前記ＡＬＵによって演算された出力をバッファリングするように構成されたハードウェアのバッファを更に含むことを特徴とする請求項１に記載のメモリモジュール。
【請求項１１】
前記ハードウェアのバッファは、前記ＩＭＣモジュールの結果レジスタのサイズの４倍以上のサイズを有することを特徴とする請求項１０に記載のメモリモジュール。
【請求項１２】
前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、アキュムレータを更に含み、
前記アキュムレータは、累積値を格納するように構成されたアキュムレータレジスタを含み、
前記アキュムレータは、
前記ＡＬＵによって演算された出力を受信し、
前記累積値と前記出力との合計で前記アキュムレータレジスタを更新するように構成されることを特徴とする請求項１に記載のメモリモジュール。
【請求項１３】

前記メモリモジュールは、シリコン貫通電極によって接続されたメモリダイのスタックを含む高帯域幅メモリ（ＨＢＭ）モジュールであり、
前記メモリダイのスタックは、前記メモリダイを含むことを特徴とする請求項１に記載のメモリモジュール。
【請求項１４】
インメモリ（ｉｎ－ｍｅｍｏｒｙ）計算を実行する方法であって、
メモリモジュールのメモリコントローラによって、ホストプロセッサからオペランド及び命令を受信する段階と、
前記メモリコントローラによって、前記命令に基づいて、複数のデータレイアウトの中からデータレイアウトを決定する段階と、
前記決定されたデータレイアウトに従って、前記メモリモジュールのダイナミックランダムアクセスメモリ（ＤＲＡＭ）バンクに前記オペランドを供給する段階と、
前記命令に従って、前記オペランドに対して算術論理ユニット（ＡＬＵ）による演算を実行するように、前記ＤＲＡＭバンクのインメモリコンピューティング（ＩＭＣ）モジュールを制御する段階と、を有し、
前記ＤＲＡＭバンクは、
複数のページに配列されたＤＲＡＭセルのアレイと、
前記ＡＬＵを含む前記ＩＭＣモジュールと、を含み、
前記データレイアウトは、前記ＤＲＡＭセルのアレイの前記複数のページから前記ＩＭＣモジュールへの前記オペランドの配置を指定し、
前記ＩＭＣモジュールは、オペランドレジスタを更に含み、
前記オペランドは、第１のオペランド及び第２のオペランドを含み、
前記複数のデータレイアウトの中から一つのオペランド（１ＯＰ）データレイアウトを決定した場合、
前記第１のオペランドは、前記ＤＲＡＭバンクの外部から前記ＩＭＣモジュールに供給され、
前記第２のオペランドは、前記ＤＲＡＭセルから前記オペランドレジスタを介して前記ＩＭＣモジュールに供給されることを特徴とする方法。
【請求項１５】
前記第１のオペランドは、複数の第１のタイルに分割され、
前記第２のオペランドは、複数の第２のタイルに分割され、
各タイルは、複数の値を含み、
前記メモリコントローラは、
前記オペランドレジスタに前記第１のオペランドの第１のタイルを格納し、
前記オペランドレジスタに格納された前記第１のオペランドの第１のタイル、及び前記第２のオペランドの複数の第２のタイルの各々に対して、前記ＡＬＵによる演算を実行するように更に構成されることを特徴とする請求項１４に記載の方法。
【請求項１６】
前記第１のオペランドは、複数の第１のタイルに分割され、
前記第２のオペランドは、複数の第２のタイルに分割され、
各タイルは、複数の値を含み、
前記複数のデータレイアウトの中から同じページ（ＳＲ）データレイアウトを決定した場合、
前記メモリコントローラは、前記ＤＲＡＭセルの同じページに、一つ以上の前記第１のタイル及び１つ以上の前記第２のタイルを格納するように構成されることを特徴とする請求項１４に記載の方法。
【請求項１７】
前記メモリコントローラは、
前記オペランドレジスタに前記一つ以上の前記第１のタイルの中の一つの第１のタイルを格納し、

前記オペランドレジスタに格納された前記第１のタイル、及び前記ＤＲＡＭセルのアレイの前記第１のタイルと同じページに格納された前記一つ以上の前記第２のタイルのそれぞれに対して、前記ＡＬＵによる演算を実行するように更に構成されることを特徴とする請求項１６に記載の方法。

【請求項１８】

前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、アキュムレータを更に含み、

前記アキュムレータは、累積値を格納するように構成されたアキュムレータレジスタを含み、

前記アキュムレータは、

前記ＡＬＵによって演算された出力を受信し、

前記累積値と前記出力との合計で前記アキュムレータレジスタを更新するように構成され、

前記命令は、前記第１のオペランドと前記第２のオペランドとの内積を計算することを含み、

前記第１のタイルの中の一つの第１のタイルは、行データを格納し、

前記第２のタイルの中の一つの第２のタイルは、列データを含むことを特徴とする請求項１７に記載の方法。

【請求項１９】

前記第１のタイルは、第１の数の値を有し、

前記第２のタイルは、第２の数の値を有し、

前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、

前記出力バッファは、前記第１の数の値と前記第２の数の値との積以上を格納するためのサイズを有し、

前記命令は、前記第１のオペランドと前記第２のオペランドとの外積を計算することを含み、

前記第１のタイルの中の一つの第１のタイルは、行データを格納し、

前記第２のタイルの中の一つの第２のタイルは、列データを含むことを特徴とする請求項１７に記載の方法。

【請求項２０】

前記第１のタイルは、第１の数の値を有し、

前記第２のタイルは、第２の数の値を有し、

前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、

前記出力バッファは、前記第１の数の値及び前記第２の数の値の中の大きい方の値以上を格納するためのサイズを有し、

前記命令は、前記第１のオペランドと前記第２のオペランドとのテンソル積を計算することを含み、

前記第１のタイルの中の一つの第１のタイルは、行データを格納し、

前記第２のタイルの中の一つの第２のタイルは、列データを含むことを特徴とする請求項１７に記載の方法。

【請求項２１】

前記第１のオペランドは、複数の第１のタイルに分割され、

前記第２のオペランドは、複数の第２のタイルに分割され、

各タイルは、複数の値を含み、

前記複数のデータレイアウトの中から異なるページ（ＤＲ）データレイアウトを決定した場合、

前記メモリコントローラは、

前記ＤＲＡＭセルのアレイの第１のページに前記第１のタイルのサブセットを格納し、

前記ＤＲＡＭセルのアレイの第２のページに前記第２のタイルのサブセットを格納するように構成されることを特徴とする請求項１４に記載の方法。

【請求項２２】

前記メモリコントローラは、
前記オペランドレジスタに前記第１のページから前記第１のオペランドの第１のタイルを格納し、
前記オペランドレジスタに格納された前記第１のオペランドの第１のタイル、及び前記第２のページからの前記第２のオペランドの複数の第２のタイルのそれぞれに対して、前記ＡＬＵによる演算を実行するように更に構成されることを特徴とする請求項２１に記載の方法。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、インメモリコンピューティングに対するデータ配置のためのシステム及び方法に関する。
【背景技術】
【０００２】
高帯域幅メモリ（Ｈｉｇｈ　Ｂａｎｄｗｉｄｔｈ　Ｍｅｍｏｒｙ；ＨＢＭ）は、グラフィックス処理装置（ＧＰＵ）用高性能メモリとしてしばしば使用される。ＨＢＭは、一般的なＤＲＡＭに比べて非常に広いバスを有するという利点がある。現在のＨＢＭアーキテクチャは、シリコン貫通電極（ｔｈｒｏｕｇｈ　ｓｉｌｉｃｏｎ　ｖｉａ；ＴＳＶ）を使用して接続される複数のスタックＤＲＡＭダイ（例えば、ダイス（ｄｉｃｅ））と、ＨＢＭのバッファ及びＧＰＵのＨＢＭメモリコントローラとして機能するロジックダイとを含む。メモリ内のプロセス（例えば、インメモリ処理（ｉｎ－ｍｅｍｏｒｙ　ｐｒｏｃｅｓｓｉｎｇ））機能をメモリシステムに追加することによって、さらに性能が向上する。
【０００３】
上述の内容は、本発明の実施形態の背景に対する理解を深めるためだけのものであり、従来技術を構成しない情報を含む。
【先行技術文献】
【特許文献】
【０００４】
【文献】特開２０１９－０７５１０１号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
本発明は、上記従来技術に鑑みてなされたものであって、本発明の目的は、メモリモジュールの性能を向上させることができる方法、及び性能が向上したメモリモジュールを提供することにある。
【課題を解決するための手段】
【０００６】
上記目的を達成するためになされた本発明の一態様によるメモリモジュールは、複数のダイナミックランダムアクセスメモリ（ＤＲＡＭ）バンクを含むメモリダイと、メモリコントローラと、を含み、前記ＤＲＡＭバンクの各々は、複数のページに配列されたＤＲＡＭセルのアレイと、前記複数のページの内の開かれたページの値を格納する行バッファと、入出力（ＩＯ）モジュールと、インメモリコンピューティング（ＩＭＣ）モジュールと、を含み、前記ページの各々は複数の前記ＤＲＡＭセルを含み、前記ＤＲＡＭセルの各々はビット値を格納し、前記ＩＭＣモジュールは、前記行バッファ又は前記ＩＯモジュールからオペランドを受信し、前記オペランド及び複数の算術論理演算から選択された一つの算術論理演算に基づいて出力を計算するように構成された算術論理ユニット（ＡＬＵ）と、前記ＡＬＵによって計算された前記出力を格納するように構成された結果レジスタと、を含み、前記メモリコントローラは、ホストプロセッサから、第１のオペランド、第２のオペランド、及び命令を受信し、前記命令に基づいて、複数のデータレイアウトから一つのデータレイアウトを決定し、前記一つのデータレイアウトに従って、前記ＤＲＡＭバン

クに前記第１のオペランド及び前記第２のオペランドを供給し、前記命令に従って、前記第１のオペランド及び第２のオペランドに対して前記複数の算術論理演算のうち前記一つの算術論理演算を実行するように、前記ＤＲＡＭバンクの前記ＩＭＣモジュールを制御するように構成されたことを特徴とする。

【０００７】

前記複数のデータレイアウトは、一つのオペランド（１ＯＰ）データレイアウトを含み、前記第１のオペランドは、前記ＤＲＡＭセルに書き込まれ、前記第２のオペランドは、前記ホストプロセッサから前記ＤＲＡＭバンクの前記ＩＭＣモジュールに直接供給され得る。

【０００８】

前記ＩＭＣモジュールは、オペランドレジスタをさらに含み、前記メモリコントローラは、さらに、前記オペランドレジスタに前記第１のオペランドの第１のタイルを格納し、前記オペランドレジスタに格納された前記第１のオペランドの第１のタイル、及び前記第２のオペランドの複数の第２のタイルの各々に対して、算術論理演算を実行するように構成され得る。

【０００９】

前記第１のオペランドは複数の第１のタイルに分割され、前記第２のオペランドは複数の第２のタイルに分割され、前記タイルの各々は複数の値を含み、前記複数のデータレイアウトは同じ行（ＳＲ）データレイアウトを含み、前記メモリコントローラは、前記ＤＲＡＭセルのアレイの同じページに一つ以上の前記第１のタイル及び一つ以上の前記第２のタイルを格納し得る。

【００１０】

前記ＩＭＣモジュールは、オペランドレジスタをさらに含み、前記メモリコントローラは、さらに、前記オペランドレジスタに前記一つ以上の第１のタイルの内の一つの第１のタイルを格納し、前記オペランドレジスタに格納された前記第１のオペランド、及び前記ＤＲＡＭセルのアレイの前記第１のタイルと同じページに格納された一つ以上の前記第２のタイルのそれぞれに対して、算術論理演算を実行するように構成され得る。

【００１１】

前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、アキュムレータ（ａｃｃｕｍｕｌａｔｏｒ）をさらに含み、前記アキュムレータは、累積値を格納するように構成されたアキュムレータレジスタを含み、前記アキュムレータは、前記ＡＬＵによって計算された出力を受信し、累積値と出力との合計で前記アキュムレータレジスタを更新するように構成され、前記命令は、前記第１のオペランドと前記第２のオペランドの内積を計算することを含み、前記第１のタイルの内の前記一つの第１のタイルは行データを格納し、前記第２のタイルの内の一つの第２のタイルは列データを含み得る。

【００１２】

前記第１のタイルは第１の数の値を有し、前記第２のタイルは第２の数の値を有し、前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、前記出力バッファは、前記第１の数の値と前記第２の数の値の積以上を格納するためのサイズを有し、前記命令は、前記第１のオペランドと前記第２のオペランドの外積を計算することを含み、前記第１のタイルの内の前記一つの第１のタイルは行データを格納し、前記第２のタイルの内の一つ第２のタイルは列データを含み得る。

【００１３】

前記第１のタイルは第１の数の値を有し、前記第２のタイルは第２の数の値を有し、前記ＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、前記出力バッファは、前記第１の数の値と前記第２の数の値の内の大きい方の値以上を格納するためのサイズを有し、前記命令は、前記第１のオペランドと前記第２のオペランドのテンソル積を計算することを含み、前記第１のタイルの内の前記一つの第１のタイルは行データを格納し、前記第２のタイルの内の一つの第２のタイルは列データを含み得る。

【００１４】

前記第１のオペランドは複数の第１のタイルに分割され、前記第２のオペランドは複数の第２のタイルに分割され、前記タイルの各々は複数の値を含み、前記複数のデータレイアウトは、異なる行（ＤＲ）データレイアウトを含み、前記メモリコントローラは、前記ＤＲＡＭセルのアレイの第１のページに前記第１のタイルのサブセットを格納し、前記ＤＲＡＭセルのアレイの第２のページに前記第２のタイルのサブセットを格納し得る。

【００１５】

前記ＩＭＣモジュールは、オペランドレジスタをさらに含み、前記メモリコントローラは、さらに、前記オペランドレジスタの前記第１のページから前記第１のオペランドの第１のタイルを格納し、前記オペランドレジスタに格納された前記第１のオペランド、及び前記第２のページから前記第２のオペランドの複数の第２のタイルのそれぞれに算術論理演算を実行するように構成されることが好ましい。

【００１６】

前記ＤＲＡＭバンクの各々の前記ＩＭＣモジュールは、前記ＡＬＵによって計算された前記出力をバッファリングするように構成されたバッファをさらに含むことが好ましい。

【００１７】

前記バッファは、前記結果レジスタのサイズの４倍以上のサイズを有し得る。

【００１８】

前記ＤＲＡＭバンクの各々の前記ＩＭＣモジュールは、アキュムレータをさらに含み、前記アキュムレータは、累積値を格納するように構成されたアキュムレータレジスタを含み、前記アキュムレータは、前記ＡＬＵによって計算された前記出力を受信し、前記累積値と前記出力との合計で前記アキュムレータレジスタを更新するように構成されることが好ましい。

【００１９】

前記メモリモジュールは、シリコン貫通電極によって接続されたメモリダイのスタックを含む高帯域幅メモリ（ＨＢＭ）モジュールであり、前記メモリダイのスタックは、前記メモリダイを含み得る。

【００２０】

上記目的を達成するためになされた本発明の一態様による方法は、インメモリ計算を実行する方法において、メモリモジュールのメモリコントローラによって、第１のオペランド、第２のオペランド、命令を受信する段階と、前記メモリコントローラによって、前記命令に基づいて複数のデータレイアウトから一つのデータレイアウトを決定する段階と、前記データレイアウトに従って、前記メモリモジュールの少なくとも一つのＤＲＡＭバンクに前記第１のオペランド及び前記第２のオペランドを供給する段階と、前記命令に従って、前記第１のオペランド及び前記第２のオペランドに対して複数の算術論理演算の内の一つの算術論理演算を実行するように、前記ＤＲＡＭバンクのＩＭＣモジュールを制御する段階と、を含み、前記ＤＲＡＭバンクの各々は、複数のページに配列されたＤＲＡＭセルのアレイと、前記複数のページの内の開かれたページの値を格納する行バッファと、ＩＯモジュールと、前記ＩＭＣモジュールと、を含み、前記ページの各々は前記ＤＲＡＭセルを含み、前記ＤＲＡＭセルの各々はビット値を格納し、前記ＩＭＣモジュールは、前記行バッファ又は前記ＩＯモジュールからオペランドを受信し、前記オペランド及び複数の算術論理演算から選択された一つの算術論理演算に基づいて出力を計算するように構成されたＡＬＵと、前記ＡＬＵによって計算された前記出力を格納するように構成された結果レジスタと、を含むことを特徴とする。

【００２１】

前記複数のデータレイアウトは、一つのオペランド（１ＯＰ）データレイアウトを含み、前記第１のオペランドは、前記ＤＲＡＭセルに書き込まれ、前記第２のオペランドは、前記ホストプロセッサから前記ＤＲＡＭバンクの前記ＩＭＣモジュールに直接供給され得る。

【００２２】

前記ＩＭＣモジュールは、オペランドレジスタをさらに含み、前記メモリコントローラ

は、さらに、前記オペランドレジスタに前記第１のオペランドの第１のタイルを格納し、前記オペランドレジスタに格納された前記第１のオペランドの第１タイル、及び前記第２のオペランドの複数の第２のタイルの各々に対して、算術論理演算を実行するように構成され得る。

【００２３】

前記第１のオペランドは複数の第１のタイルに分割され、前記第２のオペランドは複数の第２のタイルに分割され、前記タイルの各々は複数の値を含み、前記複数のデータレイアウトは、同じ行（ＳＲ）データレイアウトを含み、前記メモリコントローラは、前記ＤＲＡＭセルのアレイの同じページに一つ以上の前記第１のタイル及び１つ以上の前記第２のタイルを格納し得る。

【００２４】

前記ＩＭＣモジュールは、オペランドレジスタをさらに含み、前記メモリコントローラは、さらに、前記オペランドレジスタに前記一つ以上の第１のタイルの内の一つの第１のタイルを格納し、前記オペランドレジスタに格納された前記第１のオペランドと、前記ＤＲＡＭセルのアレイの前記第１のタイルと同じページに格納された前記一つ以上の第２のタイルのそれぞれに対して、算術論理演算を実行するように構成され得る。

【００２５】

前記少なくとも一つのＤＲＡＭバンクの前記ＩＭＣモジュールは、アキュムレータをさらに含み、前記アキュムレータは、累積値を格納するように構成されたアキュムレータレジスタを含み、前記アキュムレータは、前記ＡＬＵによって計算された前記出力を受信し、前記累積値と前記出力との合計で前記アキュムレータレジスタを更新するように構成され、前記命令は、前記第１のオペランドと前記第２のオペランドの内積を計算する段階を含み、前記第１のタイルの内の前記一つの第１のタイルは行データを格納し、前記第２のタイルの内の一つの第２のタイルは列データを含み得る。

【００２６】

前記第１のタイルは第１の数の値を有し、前記第２のタイルは第２の数の値を有し、前記少なくとも一つのＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、前記出力バッファは、前記第１の数の値と前記第２の数の値の積以上を格納するためのサイズを有し、前記命令は、前記第１のオペランドと前記第２のオペランドの外積を計算する段階を含み、前記第１のタイルの内の前記一つの第１のタイルは行データを格納し、前記第２のタイルの内の一つの第２のタイルは列データを含み得る。

【００２７】

前記第１のタイルは第１の数の値を有し、前記第２のタイルは第２の数の値を有し、前記少なくとも一つのＤＲＡＭバンクの前記ＩＭＣモジュールは、出力バッファを含み、前記出力バッファは、前記第１の数の値と前記第２の数の値の内の大きい方の値以上を格納するためのサイズを有し、前記命令は、前記第１のオペランドと前記第２のオペランドのテンソル積を計算する段階を含み、前記第１のタイルの前記一つの第１のタイルは行データを格納し、前記第２のタイルの内の一つの第２のタイルは列データを含み得る。

【００２８】

前記第１のオペランドは複数の第１のタイルに分割され、前記第２のオペランドは複数の第２のタイルに分割され、前記タイルの各々は複数の値を含み、前記複数のデータレイアウトは、異なる行（ＤＲ）データレイアウトを含み、前記メモリコントローラは、前記ＤＲＡＭセルのアレイの第１のページに前記第１のタイルのサブセットを格納し、前記ＤＲＡＭセルのアレイの第２のページに前記第２のタイルのサブセットを格納し得る。

【００２９】

前記ＩＭＣモジュールは、オペランドレジスタをさらに含み、前記メモリコントローラは、さらに、前記オペランドレジスタの前記第１のページから前記第１のオペランドの第１のタイルを格納し、前記オペランドレジスタに格納された前記第１のオペランド、及び前記第２のページから前記第２のオペランドの複数の第２のタイルのそれぞれに対して、算術論理演算を実行するように構成されることが好ましい。

【発明の効果】
【００３０】
本発明によるインメモリコンピューティングに対するデータ配置のための方法及びその方法が適用されたメモリモジュールによれば、メモリモジュールの性能を向上させることができる。
【図面の簡単な説明】
【００３１】
【図１】本発明の一実施形態によるメモリ（例えば、ＨＢＭ）システムのアーキテクチャを示すブロック図である。
【図２Ａ】本発明の一実施形態による、埋め込み算術論理ユニット（ｅｍｂｅｄｄｅｄ ＡＬＵ）を有するメモリバンクの概略ブロック図である。
【図２Ｂ】本発明の一実施形態によるＤＲＡＭセルの例を示す回路図である。
【図２Ｃ】ＤＲＡＭセルのアレイ、行デコーダ、ＩＯＳＡ、ＩＭＣモジュール、及び列デコーダを含む、本発明の一実施形態によるＤＲＡＭバンクの概略図である。
【図３】本発明の一実施形態によるＤＲＡＭブロックのアレイを示す概略図である。
【図４Ａ】第１のオペランド（行列Ａ）がＤＲＡＭに格納され、第２のオペランド（行列Ｂ）が内蔵されたメモリモジュールの外部からブロードキャストされる場合のＧＥＭＭに対するデータの配置を、本発明の一実施形態による統合されたインメモリコンピューティング（ＩＭＣ）と共に概略的に示す図である。
【図４Ｂ】双方のオペランド（行列Ａ及び行列Ｂ）がメモリモジュールのＤＲＡＭの同じページに格納されている場合のＧＥＭＭに対するデータの配置を、本発明の一実施形態による統合されたＩＭＣと共に概略的に示す図である。
【図４Ｃ】双方のオペランド（行列Ａ及び行列Ｂ）がメモリモジュールのＤＲＡＭの異なるページに格納されている場合のＧＥＭＭに対するデータの配置を、本発明の一実施形態による統合されたＩＭＣと共に概略的に示す図である。
【図５Ａ】一つのオペランド（１ＯＰ）データレイアウトにおける行列Ａの第１の行と行列Ｂの第１の列との乗算の概略図である。
【図５Ｂ】１ＯＰデータレイアウトにおける行列Ａの第１の行の第１の値と行列Ｂの各列の第１の値との乗算をデータの再利用と共に示した概略図である。
【図６Ａ】同じ行（ＳＲ）データレイアウトにおける行列Ａの第１の行と行列Ｂの第１の列との乗算の概略図である。
【図６Ｂ】ＳＲデータレイアウトにおける行列Ａの第１の行の第１の値と行列Ｂの各列の第１の値に対するデータの再利用による乗算の概略図である。
【図７Ａ】異なる行（ＤＲ）データレイアウトにおける行列Ａの第１の行と行列Ｂの第１列との乗算の概略図である。
【図７Ｂ】ＤＲデータレイアウトにおける行列Ｂの各列の第１の値と行列Ａの第１の行の第２の値により、行列Ｂの各列の第２の値による行列Ａの第１の行の第１の値のデータ再利用による乗算の概略図である。
【図８】ＤＲＡＭバンクのＩＭＣモジュールの概略ブロック図である。
【図９】本発明の一部の実施形態による、同じ行（ＳＲ）レイアウトを使用する計算を説明するためのＧＥＭＭの一例の概略図である。
【図１０】本発明の実施形態による、ＩＭＣモジュールを有するＤＲＡＭバンク内のデータの配置を制御するための方法を示すフローチャートである。
【発明を実施するための形態】
【００３２】
本発明の特徴及びそれを達成するための方法は、以下の実施形態の詳細な説明及び図面を参照することにより、容易に理解される。以下、図面を参照しながら実施形態をより詳細に説明するが、図面全体にわたって同じ参照番号は同じ要素を指す。しかし、本発明は、様々な異なる形態で実施することができ、本明細書で説明する実施形態に限定されない。むしろ、これらの実施形態は、本発明が徹底且つ完全なものとなり、本発明の態様及び

特徴が当業者に十分に伝わるように例として提供するものである。したがって、本発明の態様及び特徴の完全な理解のために、当業者に不必要なプロセス、要素、及び技術は説明しない。特に記載がない限り、同様の参照番号は、図面及び明細書に記載された説明全体を通じて同様の要素を指すので、その説明は繰り返さない。図面において、要素、層、及び領域の相対的なサイズは、明確にするために誇張されている場合がある。

【００３３】

以下の説明では、説明を目的として、様々な実施形態の完全な理解のために、多数の特定の詳細な説明を提示する。しかし、様々な実施形態は、これらの特定の詳細な説明がなくても実施でき、又は一つ以上の同等の構成で実施してもよい。他方、公知の構造及び装置は、様々な実施形態を曖昧にすることを避けるために、ブロック図の形で示す。

【００３４】

本明細書で使用される用語は、特定の実施形態を説明するためのものであり、本発明を限定することを意図するものではない。単数形の表現は、文脈がそうでないことを明確に示さない限り、複数形の表現も含む。本明細書で使用される「含む」及び「有する」などの用語は、開示する特徴、数字、ステップ、演算、要素、及び／又は構成要素の存在を特定するが、一つ以上の他の特徴、数字、ステップ、演算、要素、構成要素、及び／又はそれらの組み合わせの存在又は追加を排除するものではない。本明細書で使用される用語「及び／又は」は、一つ以上の関連して挙げられた項目の任意の且つ全ての組み合わせを含む。

【００３５】

本明細書に記載する本発明の実施形態による電子又は電気装置及び／又は任意の他の関連装置又は構成要素は、任意の適切なハードウェア、ファームウェア（例えば、特定用途向け集積回路（Ａｐｐｌｉｃａｔｉｏｎ－Ｓｐｅｃｉｆｉｃ　Ｉｎｔｅｇｒａｔｅｄ　Ｃｉｒｃｕｉｔ））、ソフトウェア、又はそれらの組み合わせを利用して実装される。例えば、一部の環境では、これらの装置の様々な構成要素は、一つの集積回路（ＩＣ）チップ上に、又は別のＩＣチップ上に形成される。また、これらの装置の様々な構成要素は、フレキシブルプリント回路フィルム、テープキャリアパッケージ（ＴＣＰ）、プリント回路基板（ＰＣＢ）上に実装されてもよく、又は一つの基板上に形成されてもよい。さらに、これらの装置の様々な構成要素は、コンピュータプログラム命令を実行し、本明細書で説明する様々な機能を行うために他のシステムの構成要素と相互作用する一つ以上のコンピューティングデバイス内における一つ以上のプロセッサ上で実行されるプロセス又はスレッドである。コンピュータプログラム命令は、例えば、ランダムアクセスメモリ（ＲＡＭ）等の標準的なメモリデバイスを使用してコンピューティングデバイス内に実装されるメモリに格納される。コンピュータプログラム命令はまた、例えば、ＣＤ－ＲＯＭ、フラッシュドライブなどの他の非一時的なコンピュータ読取可能媒体に格納される。また、当業者は、本発明の例示的な実施形態の思想及び範囲内で様々なコンピューティングデバイスの機能が単一のコンピューティングデバイスに結合若しくは統合されるか、又は特定のコンピューティングデバイスの機能が一つ以上の他のコンピューティングデバイスにわたって分散されることを認識すべきである。

【００３６】

特に定義しない限り、本明細書で使用する技術用語又は科学用語を含むすべての用語は、本発明が属する技術分野の通常の知識を有する者が一般に理解するものと同じ意味を有する。また、通常使用される辞書に定義されているような用語は、関連技術及び／又は本明細書の文脈上の意味と一致すると解釈され、本明細書で明らかに定義しない限り、理想的又は過度に公式的な意味に解釈されない。

【００３７】

本発明の一部の態様は、一般に、インメモリコンピューティング（ｉｎ－ｍｅｍｏｒｙ　ｃｏｍｐｕｔｅ）の文脈におけるメモリ内のデータ配置の管理に関する。インメモリコンピューティングに関する内容の一例は、メモリを含むＤＲＡＭ（Ｄｙｎａｍｉｃ　Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）ダイと、ＨＢＭロジックダイ上のＡＬＵ及びメモリ

コントローラとを含むＨＢＭシステムであり、ここで、ＨＢＭロジックダイ上のＡＬＵは、インメモリコンピューティングを実行する。ＨＢＭロジックダイのメモリコントローラは、ＤＲＡＭダイのメモリへのデータの格納、及びＤＲＡＭダイからのデータの読み取りを制御する。

【００３８】

明確化のため、本明細書で使用する用語「インメモリコンピューティング」は、ＤＲＡＭダイに格納されたデータを使用して、外部データバスを通過することなく、高帯域幅メモリモジュール等のメモリモジュール内で計算を実行することを意味する。比較コンピュータシステムでは、プロセッサは外部ＤＲＡＭデータバスを介してメインメモリ（例えば、ＤＲＡＭ）に接続され、メインメモリからのデータへのアクセスは、プロセッサ内のレジスタファイル内のデータ及び／又はプロセッサにより近いハードウェアキャッシュ（例えば、Ｌ１キャッシュ、Ｌ２キャッシュ）内のデータへのアクセスよりも大幅に遅い（例えば、数十倍遅い）。メモリ又はその近くにさらなるプロセッサ（例えば、「インメモリプロセッサ（ｉｎ－ｍｅｍｏｒｙ　ｐｒｏｃｅｓｓｏｒ）」）を配置することで、外部バスを通過することによって惹起される遅延を回避することができ、これにより、高性能の計算が達成される。

【００３９】

本発明の実施形態の態様は、ＡＬＵ等の計算回路（ｃｏｍｐｕｔａｔｉｏｎａｌ　ｃｉｒｃｕｉｔｒｙ）をＤＲＡＭバンクと同じダイ上に配置する（例えば、それぞれのＤＲＡＭバンクのセンス増幅器又は行バッファに直接接続される）ことに関する。

【００４０】

ＤＲＡＭモジュールの設計及び性能特性により、メモリ内のデータの特定の配列は、インメモリコンピューティングの性能に影響を与え得る。したがって、本発明の実施形態の一部の態様は、メモリモジュール（例えば、ＨＢＭメモリモジュール）のＤＲＡＭモジュール内にデータを配置するためのシステム及び方法に関し、ここでデータの配置は、ＩＭＣモジュールによって実行される計算の特性に基づいて制御される。

【００４１】

図１は、本発明の一実施形態によるメモリ（例えば、ＨＢＭ）システムのアーキテクチャを示すブロック図である。

【００４２】

図１を参照すると、本発明の実施形態は、ＦＩＭ（Ｆｕｎｃｔｉｏｎ－Ｉｎ－Ｍｅｍｏｒｙ）メモリシステム１００のためのシステムを提供する。メモリシステム１００（又は、ＨＢＭシステム）は、メモリモジュール１１０（又は、ＨＢＭモジュール）に統合されるさらなる（ａｄｄｉｔｉｏｎａｌ）コンピューティングリソースをサポートする。例えば、様々な実施形態において、メモリシステム１００は、一部のデータコンピューティング及び移動がメモリ内で実行されるようにし、且つ大容量メモリスクラッチパッド（ｈｉｇｈ－ｃａｐａｃｉｔｙ　ｍｅｍｏｒｙ　ｓｃｒａｔｃｈｐａｄ）を提供する。メモリシステム１００は、グラフィックス処理装置（ＧＰＵ）又は中央処理装置（ＣＰＵ）等のホストプロセッサ１７０に接続された少なくとも一つのメモリモジュール１１０を含む。様々な実施形態において、メモリモジュール１１０は、内部メモリバス１３０を介してメモリコントローラ１４０（例えば、ロジックダイ上）に接続された一つ以上のＤＲＡＭダイ１２０を含む。様々な実施形態において、ホストプロセッサ１７０は、メモリモジュール１１０とインタフェースするためのホストメモリコントローラ１８０（又は、ホストコントローラ）を含む。しかし、本発明の実施形態は、これに限定されるものではない。たとえば、ホストメモリコントローラ１８０は、ホストプロセッサ１７０から分離されてもよい（例えば、ホストプロセッサ１７０とは別のダイ又は同じダイとして）。

【００４３】

様々な実施形態によると、メモリコントローラ１４０は、ホストプロセッサ１７０からの命令の実行を調整するように構成される。命令は、通常の命令とＦＩＭ命令の両方を含む。例えば、通常の命令（例えば、メモリ内機能（ｆｕｎｃｔｉｏｎ－ｉｎ－ｍｅｍｏｒ

ｙ；ＦＩＭ）命令ではなく、伝統的なロード及びストア機能）は、ホストメモリコントローラ１８０によって送信され、メモリコントローラ１４０によって受信されて、通常の方法で実行される。例えば、通常の命令は、外部バス１９０を介して受信したデータをＤＲＡＭダイ１２０に格納する命令、及びＤＲＡＭダイ１２０からデータを検索し、外部バス１９０を介してホストプロセッサ１７０にデータを送信する命令を含む。一部の実施形態において、通常の命令及びＦＩＭ命令は、ＤＲＡＭダイの特定の位置（例えば、特定のバンクの特定のページ）にデータを格納する動作を含む。これらのデータは２つの異なるオペランドを含み、ここで、それぞれのオペランドは複数の値（例えば、浮動小数点又は整数値）を含み、以下でより詳細に説明するように、これらのオペランドの値は、様々なデータ配置戦略に従って、ＤＲＡＭダイの異なる位置に分散して配置（ｄｉｓｔｒｉｂｕｔｅ）される。

【００４４】

本発明の実施形態の態様は、ＩＭＣ（ｉｎ－ｍｅｍｏｒｙ　ｃｏｍｐｕｔｅ）の使用に関する。いくつかの比較ＨＢＭシステムは、ＤＲＡＭダイ１２０の外部（例えば、メモリコントローラ１４０に配置）にあるＡＬＵを含み、ＡＬＵが外部バス１９０を通過することなく、１つ以上のＤＲＡＭダイ１２０に格納されたデータに対する演算（ｏｐｅｒａｔｉｏｎ）（例えば、算術演算（ａｒｉｔｈｍｅｔｉｃ　ｏｐｅｒａｔｉｏｎｓ））を実行できるように、ＤＲＡＭダイ１２０のメモリバンクによって共有される。例えば、メモリコントローラ１４０は、ＤＲＡＭダイ１２０の異なる部分の間でデータを移動又はコピーするために、データ移動演算（例えば、ロード／ストア　ペア命令（ｌｏａｄ／ｓｔｏｒｅ　ｐａｉｒ　ｉｎｓｔｒｕｃｔｉｏｎｓ））を実行する。例えば、メモリコントローラは、ＡＬＵを利用する計算ＦＩＭ命令（例えば、アトミック命令及びＡＬＵ命令）の実行を調整することで、元々は複数の通常命令であったＦＩＭ命令を実行する。別の例として、一部の場合、ホストプロセッサ１７０から受信したＦＩＭ命令は、ＩＭＣが統合されたメモリモジュールに、学習済みの機械学習モデル（例えば、ニューラルネットワーク（ｎｅｕｒａｌ　ｎｅｔｗｏｒｋ））を使用して、ニューラルネットワークの訓練中に逆伝播（ｂａｃｋｐｒｏｐａｇａｔｉｏｎ）を実行させるか、又は２つの行列を乗算するために、供給された入力に基づいて推論を計算する等の特定の計算を実行させる。このような場合、メモリコントローラ１４０は、受信したデータ（例えば、命令のオペランド）をＤＲＡＭバンクの特定のページに格納し、ＦＩＭ命令に関連するデータを格納するＤＲＡＭバンクに特定のＡＬＵ演算を提供することによって、これらの命令の実行を調整する。メモリコントローラ１４０は、特定のＦＩＭ命令を実行する際の計算性能を向上させる方法で、受信したデータ（オペランド）をＤＲＡＭバンクの特定のページに配置する。その結果は、ＤＲＡＭダイ１２０に格納されるか、又は外部バス１９０を介してホストプロセッサ１７０に戻される。

【００４５】

本発明の実施形態の一部の態様は、ＩＭＣモジュールをＤＲＡＭダイ１２０のメモリバンクに統合することによってメモリ境界演算を加速することに関し、これにより、ＤＲＡＭダイ１２０とメモリコントローラ１４０との間で内部メモリバス１３０を通過すること（ｔｒａｖｅｒｓａｌ）を避ける。例えば、ＩＭＣモジュールは、データを保持するＤＲＡＭバンクと同じ物理的半導体ダイ上にある。それぞれのＤＲＡＭバンクは、関連するＩＭＣモジュールを有し、それによって、データが内部メモリバス１３０をメモリコントローラに通過することなく（例えば、ＤＲＡＭバンクからデータを送信することなく）、ＤＲＡＭバンクに格納されたデータに対して計算が実行できる。また、上記計算は、ＤＲＡＭバンクのＩＭＣモジュール間で並列化（ｐａｒａｌｌｅｌｉｚｅｄ　ａｃｒｏｓｓ）できる。

【００４６】

本発明の実施形態の態様は高帯域幅メモリに関して説明しているが、実施形態はこれに限定されず、他のタイプのＤＲＡＭシステムにおいて、ＤＲＡＭダイにＩＭＣモジュールを統合することにも適用される。

【００４７】
図２Ａは、本発明の一実施形態による、埋め込み算術論理ユニット（ｅｍｂｅｄｄｅｄ ＡＬＵ）を有するメモリバンクの概略ブロック図である。図２Ａに示すように、ＤＲＡＭバンク２００は、行及び列（又はページ及び列）に配列されたＤＲＡＭセル２１０のアレイを含む。図２Ａに例示するように、ＤＲＡＭバンク２００は、ｎ行（又はページ）及びｍ列に配列されたＤＲＡＭセル２１０を含む。複数のビットラインＢ１～Ｂｍは列方向に沿って延在し、複数の行イネーブルライン（ｒｏｗ ｅｎａｂｌｅ ｌｉｎｅｓ）Ｒ１～Ｒｎはアレイの行方向に沿って延在し、ビットラインと交差する。各ビットラインは、対応する列のすべてのセル（ＤＲＡＭセル２１０）に接続される（例えば、アレイのｉ番目の列のすべてのセルは、ビットラインＢｉに接続される）。同様に、各行イネーブルラインＲ１～Ｒｎは、対応する行の各ＤＲＡＭセル２１０に接続される（例えば、アレイのｊ番目の行又はページのすべてのセルは、行イネーブルラインＲｊに接続される）。また、ＤＲＡＭバンク２００のＤＲＡＭセル２１０の行は、ＤＲＡＭページと呼ばれる。

【００４８】
図２Ｂは、本発明の一実施形態によるＤＲＡＭセルの例を示す回路図である。それぞれのＤＲＡＭセル２１０は、一般に、データ電圧を格納するためのコンデンサ２１２（例えば、ビット値、ここで各コンデンサは０ビットを表す電圧又は１ビットを表す電圧を格納する）と、コンデンサ２１２にデータ電圧を送信するためのスイッチ２１４とを含むものとしてモデル化される。図２Ｂに示す特定のＤＲＡＭセル２１０は、アレイのｉ番目の行及びｊ番目の列にある。したがって、図２Ｂに示すＤＲＡＭセル２１０のスイッチ２１４は、ｉ番目のビットラインＢｉとコンデンサ２１２の一方の端子との間に接続され、コンデンサ２１２の他方の端子は接地に接続される。図２Ｂに示すように、ＤＲＡＭセル２１０のスイッチ２１４のゲート電極は、ｊ番目の行イネーブルラインＲｊに接続され、スイッチ２１４がオンになると、コンデンサ２１２がビットラインＢｉに接続される。

【００４９】
図２Ａを再び参照すると、ＤＲＡＭバンク２００は、行イネーブルラインＲ１～Ｒｎに接続された行デコーダ２２０を含み、行デコーダ２２０は、例えば、メモリコントローラ１４０から供給される行アドレスに対応する行イネーブルラインの特定の一つに、行イネーブル信号を供給するように構成される。ＤＲＡＭセル２１０のアレイの特定の行ｒ（又はページ）にデータを書き込むか又は読み取る際、行デコーダ２２０は、特定の行（又はページ）に対応する行イネーブルラインに行イネーブル信号を供給する。データを書き込む際には、特定の行（又はページ）がイネーブルされている間、書き込まれるデータに対応する電圧がビットラインＢ１～Ｂｍに供給される。同様に、ＤＲＡＭセル２１０のアレイの特定の行（又はページ）からデータを読み取る際、コンデンサ２１２に格納された電圧に対応する電圧は、ビットラインＢ１～Ｂｍに沿って送信され、センス増幅器２３２を含む入出力センス増幅器層２３０（又はＩＯＳＡ）によって読み取られる。センス増幅器２３２の各センス増幅器は、ビットラインの内の対応する一つに接続される（例えば、センス増幅器２３２は、ｍ個のセンス増幅器を含む）。例えば、一部の実施形態において、ＤＲＡＭセル２１０のアレイは、８，１９２個の列、及び８，１９２個の対応するセンス増幅器２３２に接続された８，１９２個の対応するビットライン（例えば、ビットラインＢ１～Ｂ８１９２）を含む（例えば、各ページには８，１９２ビット又は８キロビットのデータを格納できる）。センス増幅器２３２は、「プリチャージ（ｐｒｅｃｈａｒｇｅ）」コマンドによって消去されるまで、現在の行（又はページ）から読み取られたデータを格納するので、センス増幅器２３２は「行バッファ（ｒｏｗ ｂｕｆｆｅｒ）」と呼ばれる。

【００５０】
列デコーダ２４０は、マルチプレクサ２３４を使用してデータ列のーサブセット（ｓｕｂｓｅｔ）を選択するために使用され、読み取られたデータは、そのデータの計算を実行するために、グローバルＩＯ層２３６を介してＩＭＣモジュール２５０に供給される。例えば、一部の実施形態において、列デコーダ２４０及びマルチプレクサ２３４は、ＤＲＡ

Ｍセル２１０の８，１９２個の列から２５６ビット（２５６ｂ）のデータの選択を可能にする。

【００５１】

センス増幅器２３２に現在格納されているページとは異なるＤＲＡＭバンク２００のページからデータをロードする場合、「プリチャージ」（ＰＲＥ）コマンドは現在のページを閉じ、次のアクセスのためにＤＲＡＭバンク２００を準備するために使用される。次に、「活性化（ａｃｔｉｖａｔｅ）」（ＡＣＴ）コマンドを使用して、ＤＲＡＭバンクの特定の行又はページを開き、その開かれたページのデータをセンス増幅器２３２に格納する。その後、データがセンス増幅器２３２から読み取られて（ＲＥＡＤ）、ＩＭＣモジュール２５０に送信される。

【００５２】

一方、すでに開いているページからＩＭＣモジュール２５０にデータをロードする場合には、例えば、列デコーダ２４０を使用してセンス増幅器２３２に既に格納されているデータの適切なサブセットを選択することにより（ＰＲＥ及びＡＣＴコマンドは省略してもよい）、ＲＥＡＤコマンドでデータをロードするのに十分である。

【００５３】

様々な実施形態によれば、ＩＭＣモジュール２５０（又はＡＬＵ＆Ｒｅｇ）は、ＡＬＵ２５２及び１つ以上のレジスタを含む。図２Ａに示す実施形態において、ＩＭＣモジュール２５０は、オペランドレジスタ（Ｒｏｐ）２５４（又は入力バッファ）及び結果レジスタ（Ｒｚ）２５６を含む。マルチプレクサ２５７及び２５８は、（例えば、ＡＬＵ２５２に対する第１のオペランド及び第２のオペランドとして）ＡＬＵ２５２の２つの入力へのデータの流れを制御するために使用される。例えば、図２Ａに示す実施形態において、オペランドレジスタ（Ｒｏｐ）２５４は、ＡＬＵ２５２の第１のオペランド入力に接続され、第１のマルチプレクサ２５７は、センス増幅器２３２からグローバルＩＯ層２３６を介して、又は入出力（ＩＯ）モジュール２６０（又は、書き込み入力／出力及び読み取り入力／出力、又はＷＩＯ及びＲＩＯ）を介して、外部ソース（例えば、ホストプロセッサ）からのデータを書き込むために、オペランドレジスタ（Ｒｏｐ）２５４に接続される。図２Ａに示す実施形態のように、第２のマルチプレクサ２５８もまた、グローバルＩＯ層２３６を介してセンス増幅器２３２から、又はＩＯモジュール２６０を介して外部ソースから、ＡＬＵ２５２の第２のオペランド入力に直接データを供給するように構成される。ＡＬＵ２５２は、その計算を結果レジスタ（Ｒｚ）２５６に出力し、データは、結果レジスタ（Ｒｚ）２５６からグローバルＩＯ層２３６を介してＤＲＡＭセル２１０に書き戻されるか、又はＩＯモジュール２６０（又は、ＷＩＯ及びＲＩＯ）を介してホストプロセッサ１７０に送信される。

【００５４】

一部の実施形態によると、ＡＬＵ２５２は、様々な計算演算（例えば、簡単な計算コマンド）を実行するように構成される。例えば、ＡＬＵ２５２は、算術演算、ビット単位（ｂｉｔｗｉｓｅ）、シフト演算（ｓｈｉｆｔ　ｏｐｅｒａｔｉｏｎｓ）等を実行するように構成された１６ビットＡＬＵ、３２ビットＡＬＵ、又は６４ビットＡＬＵである。様々な実施形態において、ＡＬＵ２５２は、整数演算（ｉｎｔｅｒｇｅｒ　ｏｐｅｒａｔｉｏｎｓ）、浮動小数点演算（ｆｌｏａｔｉｎｇ　ｐｏｉｎｔ　ｏｐｅｒａｔｉｏｎｓ）、又はその両方を実行する回路を含む。例えば、ＡＬＵ２５２は、ＡＤＤ（＋）、ＳＵＢＴＲＡＣＴ（－）、ＭＵＬＴＩＰＬＹ（×）、及びＤＩＶＩＤＥ（÷）等の算術演算、ＡＮＤ（＆）、ＯＲ（｜）、ＸＯＲ（＾）、及びＮＯＴ（～）演算、並びにテンソル演算（ｔｅｎｓｏｒ　ｏｐｅｒａｔｉｏｎｓ）等のビット演算を実行するように構成される。また、一部の実施形態において、ＡＬＵ２５２は、単一命令、複数データ（ＳＩＭＤ）、又はデータのベクトルに対する演算を並列に実行するためのベクトル命令を実装する。本発明の実施形態によるＡＬＵ２５２によって実装されるベクトル演算の例は、内積（’）、外積（⊗）

、整流線形ユニット（ＲｅＬＵ）、平方（ｖｓＳｑｒ）、及び平方根（ｖｓＳｑｒｔ）を含む。ＡＬＵ２５２は、アトミック及び非アトミック演算に利用される。以下の表１は、本発明の一部の実施形態によるＡＬＵ２５２によってサポートされる演算を挙げている。

【００５５】

【表１】

ＩＤ	演算	説明
０	Ｒｏｐ＝ＧＩＯ	Ｒｏｐに格納された列読み取りデータ
１	Ｒｏｐ＝ＷＩＯ	Ｒｏｐに格納された列書き込みデータ
２	Ｒｏｐ＝Ｒｚ	ＡＬＵ出力ＲｚをＲｏｐにコピー
３	ＧＩＯ＝Ｒｚ	バンクに書き戻す
４	ＲＩＯ＝ＧＩＯ	ＤＱ出力への通常の読み取り
５	ＲＩＯ＝Ｒｚ	ＲｚをＤＱ出力に駆動
６	Ｒｚ＝Ｒｏｐ（ｏｐ）ＧＩＯ	Ｒｏｐ及びバンクからのデータを使用した演算
７	Ｒｚ＝Ｒｏｐ（ｏｐ）ＷＩＯ	Ｒｏｐ及びブロードキャストデータを使用した演算
８	Ｒｚ＝ＷＩＯ（ｏｐ）ＧＩＯ	ブロードキャストデータ及びバンクを使用した演算
９	ＧＩＯ＝ＷＩＯ	ＤＱ入力からの通常の書き込み

【００５６】

図２Ｃは、上述したＤＲＡＭセル２１０のアレイ、行デコーダ２２０、入出力センス増幅器層（ＩＯＳＡ）２３０、ＩＭＣモジュール（ＡＬＵ＆Ｒｅｇ）２５０、及び列デコーダ２４０を含む、本発明の一実施形態によるＤＲＡＭバンク２００の概略図である。

【００５７】

図３は、本発明の一実施形態によるＤＲＡＭブロックのアレイを示す概略図である。図３に示す実施形態のように、１６個のＤＲＡＭバンク２００は、４×４アレイに配列され、ＤＲＡＭバンクＡ～Ｐとしてラベル付けされ、ＢＧ０（ＤＲＡＭバンクＡ、Ｂ、Ｃ、及びＤを含む）、ＢＧ１（ＤＲＡＭバンクＥ、Ｆ、Ｇ、及びＨを含む）、ＢＧ２（ＤＲＡＭバンクＩ、Ｊ、Ｋ、及びＬを含む）、並びにＢＧ３（ＤＲＡＭバンクＭ、Ｎ、Ｏ、及びＰを含む）としてラベル付けされた４つのバンクグループに配列される。図２Ｃに関して上述したように、図３に示す各ＤＲＡＭバンク２００は、ＤＲＡＭダイ１２０内で（例えば、外部バスを通過することなく）計算を実行するためのＩＭＣモジュール２５０を含む。また、図３に示すように、ＤＲＡＭダイ１２０は、外部ソースからのデータを（例えば、ＤＲＡＭダイ１２０を複数の他のスタック型ＤＲＡＭダイ１２０及びメモリコントローラ１４０に接続するシリコン貫通電極又はＴＳＶを介して）ブロックの４つの列に多重化するように構成されたマルチプレクサ３００（例えば、４：１マルチプレクサ）をさらに含む。例えば、マルチプレクサ３００は、すべてのＤＲＡＭバンク２００に、２５６ビット（２５６ｂ）データベクトルをブロードキャスト（ｂｒｏａｄｃａｓｔ）するか、又はＤＲＡＭバンク２００の特定の列（ＢＧ０、ＢＧ１、ＢＧ２、又はＢＧ３）にデータベクトルを供給するために使用される。

【００５８】

ＤＲＡＭダイ１２０に統合されたＩＭＣモジュール２５０等のＩＭＣを含むメモリシステム１００は、演算を実行するために、データが外部バス（例えば、バス１９０）のボトルネックを通過する必要がないので、メモリ境界のホスト演算の性能を加速させる。ただし、ＩＭＣは、依然として、ＡＬＵパイプライン処理（ｐｉｐｅｌｉｎｉｎｇ）及びＤＲＡＭプロセスの形態のコンピューティングオーバーヘッド、並びにデータ配置及びＤＲＡＭタイミングの形態のメモリオーバーヘッドに遭遇する。

【００５９】

したがって、本発明の実施形態の態様は、統合されたＩＭＣモジュール２５０によって

インメモリコンピューティングを実行する際に、ＤＲＡＭタイミングオーバーヘッドの影響を回避又は減少させるために、ＤＲＡＭ内にデータを配置するシステム及び方法に関する。本発明の実施形態の一部の態様は、改善された性能を達成するためのソフトウェア及びハードウェアの共同設計に関する。

【００６０】

様々な実施形態において、メモリコントローラ１４０は、ＤＲＡＭダイ１２０への演算及びＤＲＡＭダイ１２０からの演算を提供し、データの入力及び出力を管理する。したがって、本発明の実施形態の一部の態様は、ホストプロセッサ１７０によってメモリシステム１００のメモリコントローラ１４０に提供された命令に従って、ＤＲＡＭダイ１２０内にデータを配置するように構成されたメモリコントローラ１４０に関する。例えば、本発明の実施形態の一部の態様は、インメモリコンピューティングと、ＡＰＩを使用してプログラムのソースコードをコンパイル又は解釈する際に、ＡＰＩへの呼び出しに従ってデータを配置するようにメモリコントローラ１４０を制御するためのコマンドを生成するように構成されたコンパイラ（例えば、データコンパイラ）と、を有するＨＢＭと相互作用するアプリケーション・プログラミング・インターフェース（ＡＰＩ）を提供することに関する。例えば、ＡＰＩは、ＧＥＭＭ（Ｇｅｎｅｒｎａｌ Ｍａｔｒｉｘ－Ｍａｔｒｉｘ Ｍｕｌｔｉｐｌｉｃａｔｉｏｎ）を実行するための関数呼び出し（ｆｕｎｃｔｉｏｎ ｃａｌｌ）を提供し、コンパイラは、データに対して実行される演算（例えば、内積、外積、行列乗算など）を含む要素、及びデータのサイズ（例えば、データがメモリのページに適合するか否か）に基づくデータのナイーブな（ｎａｉｖｅ）配置よりも性能が改善する方法で、オペランド行列を表すデータをＤＲＡＭダイ１２０に配置するようにメモリコントローラ１４０を制御するコマンドのシーケンスを生成する。ＡＰＩを介してＩＭＣでＤＲＡＭを使用するようにソフトウェアを作成する際に、本発明の一部の実施形態によるコンパイラ又はデータコンパイラは、ＤＲＡＭダイ１２０の特定の位置にデータを配置し、ＤＲＡＭダイ１２０の計算を実行するためのＩＭＣモジュールを制御し、結果を格納するために、ソフトウェアのソースコードの少なくとも一部をメモリコントローラ１４０によって実行されるコマンドに変換する。

【００６１】

一実施形態によると、１６レーン（ｌａｎｅ）のＡＬＵは、半精度浮動小数点（ｈａｌｆ－ｐｒｅｃｉｓｉｏｎ ｆｌｏａｔｉｎｇ－ｐｏｉｎｔ）（ＦＰ－１６）計算で８ＧＦＬＯＰＳ（ギガ浮動小数点演算／秒）のピーク性能を達成する。（本発明の実施形態によるＩＭＣモジュールの性能をＦＬＯＰＳに関して本明細書に説明しているが、本発明の実施形態は、浮動小数点演算を実行することに限定されず、様々なデータレイアウトの相対的な性能は、例えば、整数演算の実行する際と似ている。）したがって、第２世代高帯域幅メモリ標準（ＨＢＭ２）を使用する４つのダイ（４Ｈ又は４－Ｈｉ）のスタックを使用する、本発明の一実施形態によるＩＭＣの実装は、ＦＰ－１６計算（ダイあたり２５６バンク×４ダイのスタック＝１，０２４バンク、各バンクは対応する１６レーンのＡＬＵを有する）で８ＴＦＬＯＰＳ（テラ浮動小数点演算／秒）を達成する。

【００６２】

ＴＦＬＯＰＳで測定されるピーク計算性能は、異なるデータレイアウトシナリオの下で変わる。

一番目の場合（２ＯＰと表記）、メモリモジュール１１０の外部からの２つのオペランドがインメモリコンピューティングＡＬＵに完全に供給され、その結果はバッファリングされて（ｂｕｆｆｅｒｅｄ）完全に累積され、これにより、上述の８ＴＦＬＯＰＳのピーク計算性能が得られる。

【００６３】

二番目の場合（１ＯＰと表記）、第１のオペランドはＨＭＢの外部からＩＭＣに完全に供給されるが、第２のオペランドはＤＲＡＭダイ１２０の任意の位置から読み取られる。これは、約６．５ＴＦＬＯＰＳにピーク計算性能を低下させる。

【００６４】

三番目の場合（ＤＲと表記）、双方のオペランドはＤＲＡＭの異なるページに配置され、結果がＤＲＡＭに書き戻される。このシナリオでは、約０．８ＴＦＬＯＰＳが測定された計算性能である（例えば、２ＯＰの場合よりも１桁遅い）。

【００６５】

四番目の場合（ＳＲと表記）で、双方のオペランドはＤＲＡＭブロックの同じ行又はページに位置する。これはＤＲの場合よりも性能が大幅に向上し、約３．３ＴＦＬＯＰＳになる。

【００６６】

したがって、データの考慮に基づいて、様々なデータレイアウトのトレードオフが行われる。例えば、１ＯＰの場合、第１のオペランドがＤＲＡＭにあり、第２のオペランドがＨＢＭの外部からブロードキャストされると、性能が高くなる（上記のように約６．５ＴＦＬＯＰＳ）が、ホストプロセッサ１７０にオーバーヘッドが発生し、第２のオペランドをＨＢＭに提供しなければならない。

【００６７】

図４Ａは、第１のオペランド（行列Ａ）がＤＲＡＭに格納され、第２のオペランド（行列Ｂ）がメモリモジュールの外部からブロードキャストされる場合のＧＥＭＭに対するデータの配置を、本発明の一実施形態による統合されたＩＭＣと共に概略的に示す図である。例示のために、ＤＲＡＭバンク２００－Ｏをより詳細に示している。図４Ａに示すように、第１のオペランド行列Ａに関連するデータは、ＤＲＡＭバンク２００－Ｏの一つのページ４０１（例えば、第１の行又は第１のページ）に配置され、第２のオペランド行列Ｂに関連するデータは、（例えば、マルチプレクサ３００を介して）ＤＲＡＭダイ１２０の外部からブロードキャストされる。計算結果Ｃは、ＤＲＡＭバンク２００－Ｏの異なるページ４０２に配置される。

【００６８】

双方のオペランドがＤＲＡＭ内にある場合、双方のオペランドを同じページ又は同じ行（ＳＲ）に配置することは、以下で詳細に説明するように、実行する必要があるＰＲＥ及びＡＣＴ演算の数を部分的に減らすことによって、計算性能を向上させる（例えば、約３．３ＴＦＬＯＰＳ）が、ＤＲＡＭの正しい部分にデータを配置することに関してより多くの制約を課す。

【００６９】

図４Ｂは、双方のオペランド（行列Ａ及び行列Ｂ）が、メモリモジュールのＤＲＡＭの同じページに格納されている場合のＧＥＭＭに対するデータの配置を、本発明の一実施形態による統合されたＩＭＣと共に概略的に示す図である。例示のために、ＤＲＡＭバンク２００－Ｏをより詳細に示している。図４Ｂに示すように、第１のオペランド行列Ａ及び第２のオペランドの行列Ｂに関連するデータは、ＤＲＡＭバンク２００－Ｏのページ４１１（例えば、第１の行又は第１のページ）に配置される。より詳細には、ページ４１１の前半部は第１のオペランド行列Ａからのデータで満たされ、ページ４１１の後半部は第２のオペランド行列Ｂからのデータで満たされる。行列Ａ及び行列Ｂが図４Ａに関して上述したものと同じサイズと仮定すると、残りのデータを格納するためにさらにページが必要になる。このように、第１のオペランド行列Ａ及び第２のオペランドの行列Ｂの双方に関連するデータもページ４１２に配置される。行列乗算の結果Ｃは、ＤＲＡＭバンク２００のページ４１３に格納される。

【００７０】

一方、オペランドを異なるページに配置することは、より柔軟で、レイアウトに対する制約を減らす（例えば、固定サイズのメモリのページにきちんと適合しないサイズを有するデータに適合する）が、一般に計算性能が低下する。

【００７１】

図４Ｃは、双方のオペランド（行列Ａ及び行列Ｂ）が、メモリモジュールのＤＲＡＭの異なるページに格納されている場合のＧＥＭＭに対するデータの配置を、本発明の一実施形態による統合されたＩＭＣと共に概略的に示す図である。例示のために、ＤＲＡＭバン

ク２００－Ｏをより詳細に示している。図４Ｃに示すように、第１のオペランド行列Ａに関連するデータは、ページ４２１に配置され、第２のオペランド行列Ｂに関連するデータは、ページ４２２に配置され、結果Ｃはページ４２３に配置される。

【００７２】

図４Ａ、図４Ｂ、及び図４Ｃに示す異なるデータ配置戦略（１ＯＰ、ＳＲ、及びＤＲ）の様々な性能への影響は、以下の図５Ａ、図５Ｂ、図６Ａ、図６Ｂ、図７Ａ、及び図７Ｂを参照して、より詳細に説明する。例えば、図５Ａに示すように、行列ＡはＭ×Ｋ行列であり、行列ＢはＫ×Ｎ行列であるので、行列Ａと行列Ｂの積である行列Ｃのサイズは、Ｍ×Ｎである。説明のために、以下の例ではＫ＝５の場合を説明しているが、本発明の実施形態はそれに限定されない。標準行列乗算に従って、結果行列Ｃの左上の値Ｃ００は、行列Ａ（各位置Ａｉｊは、例えば１６個の半精度浮動小数点値のベクトル又は「タイル」を水平順に示し、図５Ａ、図５Ｂ、図６Ａ、図６Ｂ、図７Ａ、及び図７Ｂは、Ａ００、Ａ０１、Ａ０２、Ａ０３、及びＡ０４を示す）の第１の行に、行列Ｂ（各位置Ｂｉｊは、例えば、１６個の半精度浮動小数点値のベクトル又は「タイル」を垂直順に示す）の第１の列をペアとして乗算して計算される。つまり、Ｃ００はＡ００・Ｂ００＋Ａ０１・Ｂ１０＋Ａ０２・Ｂ２０＋Ａ０３・Ｂ３０＋Ａ０４・Ｂ４０を格納する。図５Ａ、図５Ｂ、図６Ａ、図６Ｂ、図７Ａ、及び図７Ｂにおいて、シェーディング（ｓｈａｄｉｎｇ）処理は共に乗算されるオペランドを識別するために使用する。より詳細には、同じパターンを使用してシェーディング処理された２つのオペランドが、図示する計算の一部として共に乗算される。本発明の実施形態の態様は、浮動小数点オペランドに対して浮動小数点演算を実行するように構成されたＩＭＣに関して説明しているが、本発明の実施形態はそれに限定されず、例えば、整数オペランドに対する整数演算を実行するように構成されたＩＭＣに適用してもよい。

【００７３】

図５Ａは、一つのオペランド（１ＯＰ）データレイアウトにおける行列Ａの第１の行と行列Ｂの第１の列との乗算の概略図である。ここで、本発明の一実施形態による一つのオペランドが外部から供給され、一つのオペランドがインメモリコンピューティングでＤＲＡＭバンクに格納される。図５Ａに示すように、行列Ａの第１の行のタイルＡ００、Ａ０１、Ａ０２、Ａ０３、及びＡ０４は、ＤＲＡＭバンク２００の同じページ４０１（行）に格納され、行列Ｂのベクトル又はタイルＢ００、Ｂ１０、Ｂ２０、Ｂ３０、及びＢ４０は、外部から供給され、結果（例えば、Ｃ００）は、ＤＲＡＭバンク２００の別のページ４０２に格納される。

【００７４】

乗算を計算するプロセスは、ＤＲＡＭバンク２００から値Ａ００を読み取り、Ａ００・Ｂ００を計算することによって開始する。これには、ページ４０１を開くことが含まれ、したがって、センス増幅器２３２を準備するためにプリチャージ（ＰＲＥ）コマンドが必要となり、その後にページ４０１をセンス増幅器２３２にロードするための活性化（ＡＣＴ）コマンド、及びセンス増幅器２３２からＩＭＣモジュール２５０にＡ００の値をロードするための読み取り（ＲＥＡＤ）コマンドが続く。上述のように、Ｂ００は外部から入力として提供されるため、この値の取得にＤＲＡＭ演算は必要でない。ＡＬＵ２５２は、その後、乗算Ａ００・Ｂ００を計算し、出力バッファ（例えば、出力レジスタＲｚ）に一時的な結果を格納する。

【００７５】

次に、ＡＬＵ２５２は、ＤＲＡＭからＡ０１を読み取ることによって乗算Ａ０１・Ｂ１０を計算する。これはまた、ＰＲＥコマンド、ＡＣＴコマンド、及びＲＥＡＤコマンドが必要である。ベクトル又はタイルＢ１０は、外部から入力として提供されるため、Ａ０１・Ｂ１０が計算され、Ａ００・Ｂ００を格納するバッファ（例えば、出力レジスタＲｚ）の一時的な結果に追加される。このプロセスは、行列Ａ及び行列Ｂの残りの値に対して繰り返される。その結果、それぞれの計算（例えば、タイルＡ００及びＢ００等の２つのオペランドの乗算）は、計算ごとに１つのＰＲＥ、１つのＡＣＴ、及び１つのＲＥＡＤを必

要とする。図８に関して以下でより詳細に説明するように、一部の実施形態では、ＩＭＣは値を格納し、先に格納された値と新たに受信した値の合計（ｓｕｍ）で格納された値を更新するように構成されるアキュムレータをさらに含む。

【００７６】

図５Ｂは、１ＯＰデータレイアウトにおける行列Ａの第１の行の第１の値と行列Ｂの各列の第１の値との積をデータ再利用と共に示した概略図である。ここで、本明細書の一実施形態によれば、一つのオペランドが外部から供給され、１つのオペランドがインメモリコンピューティングでＤＲＡＭバンクに格納される。図５Ｂは、図５Ａに示すものとは異なり、ＤＲＡＭからロードされたデータは、行列Ｂの異なる列に対して再利用される（例えば、オペランドレジスタ（Ｒｏｐ）２５４に格納される）。特に、２つの行列を乗算する際に、行列Ａのすべての行のｊ番目の要素は、行列Ｂのｊ番目の行のすべての要素と乗算される。したがって、行列Ａの各要素を一度ロードし、それを行列Ｂのすべての列（Ｎ列）に乗算することにより、ＤＲＡＭバンク２００からのデータロードのコストは、Ｎ列にわたって償却（ａｍｏｒｔｉｚｅｄ）される。より詳細には、行列Ａ及び行列Ｂを乗算する場合、従来どおり、タイルＡ００は、ＰＲＥコマンド、ＡＣＴコマンド、及びＲＥＡＤコマンドを使用してＤＲＡＭバンク２００から読み取られ、タイルＢ００は、外部から入力として受信される。ＡＬＵ２５２は、Ｃ００の合計の一部を計算するためにＡ００・Ｂ００を計算する。しかし、ＤＲＡＭからタイルＡ０１をロードする代わりに（例えば、別のＰＲＥ、ＡＣＴ、及びＲＥＡＤシーケンスを使用）、Ｃ０１に対して計算する乗算の内の一つであるＡ００・Ｂ０１を計算するために、タイルＡ００を再利用して（外部から受信された）Ｂ０１を乗算する。その結果、各計算には、１／Ｎ　ＰＲＥ、１／Ｎ　ＡＣＴが必要であり、計算ごとに１つのＲＥＡＤが必要である（行列Ａの別の部分をロードするためのＰＲＥ及びＡＣＴコマンドが、行列ＢのＮ列に対して償却されるため）。

【００７７】

図６Ａは、同じ行（ＳＲ）データレイアウトにおける行列Ａの第１の行と行列Ｂの第１の列との乗算の概略図である。ここで、本発明の一実施形態によれば、双方のオペランドは、インメモリコンピューティングを有するＤＲＡＭバンクの同じページに格納される。図６Ａに示すように、行列Ａの第１の行のタイルＡ００、Ａ０１、及びＡ０２、並びに行列Ｂの第１の行のタイルＢ００、Ｂ１０、及びＢ２０は、ＤＲＡＭバンク２００の同じページ４１１に格納され、行列Ａのタイル Ａ０３及びＡ０４、並びに行列Ｂのタイル Ｂ３０及びＢ４０は、ＤＲＡＭバンク２００のページ４１２に格納される。その結果（例えば、Ｃ００）は、ＤＲＡＭバンク２００のページ４１３に格納される。

【００７８】

結果（例えば、内積）を計算するプロセスは、ＤＲＡＭバンク２００からタイルＡ００を読み取り、Ａ００・Ｂ００を計算することによって開始する。これには、ページ４１１を開くことが含まれ、したがって、プリチャージ（ＰＲＥ）コマンドが必要となり、その後に活性化（ＡＣＴ）コマンド、及びセンス増幅器２３２からＩＭＣモジュール２５０にタイルＡ００をロードするための読み取り（ＲＥＡＤ）コマンドが続く。タイルＢ００はＤＲＡＭから読み取る。しかし、タイルＢ００がＡ００と同じページ４１１にあり、その値は既にセンス増幅器２３２に格納されているので、ＲＥＡＤコマンドで十分である（タイルＢ００をＩＭＣモジュール２５０に読み取るために、さらにＰＲＥ及びＡＣＴを実行する必要はない）。したがって、タイルＡ００及びＢ００が読み取られると、ＡＬＵ２５２はＡ００・Ｂ００を計算し、一時的な結果をバッファに格納する。同様に、タイルＡ０１とタイルＢ１０は、いずれもページ４１１にもあり、したがって、ページ４１１がＡ００を読み取るために最初に開かれた際に、センス増幅器２３２に予め格納されているため、ＰＲＥとＡＣＴなしでＡ０１・Ｂ１０を計算するためのタイルＡ０１及びＢ１０の読み取りも同様に、ＲＥＡＤコマンドを使用して実行する。そのため、各計算は、計算ごとに１／ｒＰＲＥ、１／ｒＡＣＴ、及び２ＲＥＡＤ演算を実行する。ここで、ｒは、ＤＲＡＭバンク２００の同じページに格納されている一致する値のペアの数である。例えば、上述のように、図６Ａは、行列Ａの第１の行のタイルＡ００、Ａ０１、及びＡ０２と、行列Ｂ

の第１の行のタイルＢ００、Ｂ１０、及びＢ２０とが、ＤＲＡＭバンクのページ４１１に格納される場合を示す。したがって、計算Ａ００・Ｂ００、Ａ０１・Ｂ１０、及びＡ０２・Ｂ２０は、ページ４１１に３組のタイル（例えば、ｒ＝３）が含まれているため、それぞれ１／３ＰＲＥ、１／３ＡＣＴ、及び２ＲＥＡＤコマンドを償却する。Ａ０３・Ｂ３０及びＡ０４・Ｂ４０を計算する際、ページ４１２に２組の値（例えば、ｒ＝２）が含まれているため、このような計算は、それぞれ１／２ＰＲＥ、１／２ＡＣＴ、及び２ＲＥＡＤコマンドをそれぞれ償却する。計算ごとに必要なＰＲＥ及びＡＣＴコマンドの数が減ることにより、全体の計算性能が向上する。

【００７９】

図６Ｂは、ＳＲデータレイアウトにおける行列Ａの第１の行の第１の値と行列Ｂの各列の第１の値に対するデータの再利用による乗算の概略図である。ここで、本発明の一実施形態による双方のオペランドは、インメモリコンピューティングを有するＤＲＡＭバンクの同じページに格納される。図５Ａ及び図５Ｂの配列間の比較と同様の方法で、図６Ｂに示す計算プロセスは、ＤＲＡＭからロードされた値が再利用されるという点で図６Ａに示すものと異なる。より詳細には、行列の乗算は、第１のオペランドの指定された行のｉ番目の要素と第２のオペランドの各列のｉ番目の要素との乗算を含むということに基づいて、行列Ａと行列Ｂの要素は、指定された行列Ａの行の各要素のデータが、乗算される行列Ｂの行の値と同じページに配置されるように、ＤＲＡＭバンク２００に配列される。

【００８０】

例えば、図６Ｂに示すように、行列Ａに行列Ｂを乗算すると、行列Ａのタイル００は、結果行列Ｃの第１の行の一部（例えば、、Ｃ００、Ｃ０１、Ｃ０２、Ｃ０３、Ｃ０４、Ｃ０５等の項のいずれか）を計算するプロセスにおいて、行列Ｂのすべての列の第１のタイル（行列Ｂの第１の行のすべての要素、例えば、Ｂ００、Ｂ０１、Ｂ０２、Ｂ０３、Ｂ０４、Ｂ０５、．．．）に乗算される。同様に、行列Ａの値Ａ０１に行列Ｂのすべての列の第２のタイルが乗算される（図６Ｂに示すように、行列Ｂの第２の行のすべての要素、例えば、Ｂ１０、Ｂ１１、Ｂ１２、Ｂ１３、Ｂ１４、Ｂ１５、．．．）。

【００８１】

このように、行列Ａの少なくとも一つの値は、行列Ｂの対応値と同じページに格納される。図６Ｂに示す特定の例において、ページ４１１は、行列ＡからタイルＡ００、行列ＢからタイルＢ００、Ｂ０１、Ｂ０２、Ｂ０３、Ｂ０４、Ｂ０５、．．．を格納し、ページ４１２は、行列ＡからタイルＡ０１、行列ＢからタイルＢ１０、Ｂ１１、Ｂ１２、Ｂ１３、Ｂ１４、Ｂ１５、．．．を格納する。ＤＲＡＭ内のデータのこのような配列により、ＤＲＡＭのページが少なくともＮ＋１エントリを格納すると仮定すると、各計算は、１／Ｎ ＰＲＥコマンド、１／Ｎ ＡＣＴコマンド、及び（Ｎ＋１）／Ｎ ＲＥＡＤコマンド（ＰＲＥコマンド及びＡＣＴコマンドは、行列ＢのＮ列で償却されるため）を実行する。

【００８２】

図７Ａは、異なる行（ＤＲ）データレイアウトにおける行列Ａの第１の行と行列Ｂの第１の列との乗算の概略図である。ここで、オペランドは、本発明の一実施形態によるインメモリコンピューティングでＤＲＡＭバンクの異なるページに格納される。図７Ａに示すように、行列Ａの第１の行のタイルＡ００、Ａ０１、Ａ０２、Ａ０３、及びＡ０４は、ページ４２１に格納される一方、行列Ｂの第１の列のタイルＢ００、Ｂ１０、Ｂ２０、Ｂ３０、及びＢ４０は、別のページ４２２に格納される。

【００８３】

タイルＢ００がタイルＡ００とは異なるページ４２２（行）にあるため、Ｃ００を計算するプロセスは、ＰＲＥ、ＡＣＴ、及びＲＥＡＤコマンドのシーケンスを使用して、タイルＢ００を読み取り、その後に、ＰＲＥ、ＡＣＴ、及びＲＥＡＤコマンドのシーケンスを使用して、ＤＲＡＭバンク２００のページ４２１（行）からＡ００を読み取ることによって、Ａ００・Ｂ００を計算することから始まる。ＡＬＵ２５２は、Ａ００・Ｂ００を計算し、その結果を一時的なバッファに格納する。Ｃ００計算を継続するために、タイルＡ０１は、ページ４２１（行）から読み取られ、ページ４２２（行）からＢ１０が読み取られ

、ここで、それぞれの値は、ＤＲＡＭバンク２００に対するＰＲＥ、ＡＣＴ、及びＲＥＡＤシーケンスの実行を含む。

【００８４】

図７Ｂは、ＤＲデータレイアウトにおける行列Ｂの各列の第１の値と行列Ａの第１の行の第２の値とにより、行列Ｂの各列の第２の値による行列Ａの第１の行の第１の値のデータ再利用による乗算の概略図である。ここで、オペランドは、本発明の一実施例によるインメモリコンピューティングでＤＲＡＭバンクの異なるページに格納される。

【００８５】

図５Ｂにおける説明及び１ＯＰにおけるデータ再利用と同様の方法により、行列Ｂで乗算を実行する際に行列Ａから取り出したデータを再利用することで、メモリ演算の数が減る。図７Ａにおける説明のように、指定された行列Ａの行の各ｉ番目の値は、行列Ｂのｉ番目の行の各値と乗算される。したがって、指定された行列Ｂの行のすべての値がＤＲＡＭバンク２００の同じページに格納されると、ＰＲＥ及びＡＣＴコマンドの数が減少し、それによって性能が向上する。

【００８６】

例えば、行列Ａに行列Ｂを乗算するプロセスは、Ｎ個の部分和（例えば、Ｃ００、Ｃ０１、Ｃ０２、．．．、Ｃ０Ｎの一部）を計算するために、行列Ａのタイル００に行列Ｂの第１の行のＮ個のタイル（Ｂ００、Ｂ０１、Ｂ０２、．．．、Ｂ０Ｎ）のそれぞれを乗算することから始まる。このプロセスは、ＤＲＡＭバンク２００からタイルＡ００をロードすることにより始まる。これには、ページ４２１（行）を開き、行列Ａのタイル００をオペランドレジスタ（Ｒｏｐ）２５４にロードするため、ＰＲＥコマンド、ＡＣＴコマンド、及びＲＥＡＤコマンドが含まれる。Ａ００をロードした後、行列Ｂの第１の行のタイルＢ００、Ｂ０１、Ｂ０２、．．．、Ｂ０Ｎがタイル００に乗算されるようロードされる。図７Ｂに示すように、行列Ｂのこれらの値がすべて同じページにある場合（例えば、Ｂ００、Ｂ０１、Ｂ０２、Ｂ０３、Ｂ０４、．．．は、図７Ｂに示すＤＲＡＭバンク２００のページ４２２にある）、このページにアクセスするために使用されたＰＲＥ及びＡＣＴコマンドが、そのページに格納されたＮ個の値で償却される。したがって、行列Ａの一つの値に行列Ｂの一つの行のすべての値を乗算するには、２つのＰＲＥコマンド、２つのＡＣＴコマンド、及びＮ＋１　ＲＥＡＤコマンドが必要となり、行のＮ個の値を償却すると、計算ごとに２／Ｎ　ＰＲＥコマンド、２／Ｎ　ＡＣＴコマンド、及び（Ｎ＋１）／ＮＲＥＡＤコマンドが生成される。上述の例と同様の方法で、ＤＲＡＭコマンドの数の減少は、全体の計算性能（例えば、行列乗算演算）を向上させる。

【００８７】

上述のデータ配置オプションは、サイクルレベル（ｃｙｃｌｅ－ｌｅｖｅｌ）の高帯域幅メモリ－メモリ内機能（ＨＢＭ－ＦＩＭ）シミュレータを使用して実験的にテストし、これはＧＥＭＭ記録と共にＩＭＣ用にカスタマイズされた。実験アーキテクチャには４つの４Ｈ　ＨＢＭ２モジュールを含むＨＢＭが含まれており、ここでＤＲＡＭのアレイサイズは１６，３８４（１６Ｋｉｂ）行×８，１９２（８Ｋｉｂ）列（例えば、各ページのサイズは８，１９２ｂ）であり、行バッファのサイズは、８，１９２ビット（８Ｋｉｂ）である。ＩＭＣモジュールは、待ち時間周期を備えた１６レーンのＦＰ－１６ベクトルユニットと７６８ビットバッファを含み、パイプライン動作が可能であった。様々な実施形態において、バッファはより小さくてもよく（例えば、２５６ビットの入力バッファ及び２５６ビットの出力バッファを含む５１２ビットのバッファ）、又はより大きくてもよい（例えば、２５６ビットの入力バッファ及び１，０２４ビットの出力バッファを備えた１，２８０ビットのバッファ）。一部の実施形態において、入力バッファは、２５６ビットよりも大きい（例えば、５１２ビット）。

【００８８】

本発明の実施形態の一部の態様は、ＤＲＡＭバンク２００のレベルでさらなるバッファ及びアキュムレータ（例えば、ＤＲＡＭバンクでバンクごとに提供されるさらなるハードウェア）を含むことに関する。

【００８９】
図８は、ＤＲＡＭバンクのＩＭＣモジュールの概略ブロック図である。ここで、ＩＭＣモジュールは、本発明の一実施形態による結果バッファ、アキュムレータ、及びバッファにさらに接続される。
【００９０】
図８に示すように、ＡＬＵ２５２は、入力オペランドＡ（図８では２５６ビットのオペランドＡ［０：２５５］と表示）及び入力オペランドＢ（図８では２５６ビットのオペランドＢ［０：２５５］と表示）を受信する。ＡＬＵは、２つの入力オペランド（例えば、加算、乗算、内積、外積など）に対する演算を実行し、結果Ｃ（図８では２５６ビットの結果Ｃ［０：２５５］と表示）を計算する。
【００９１】
図８を参照すると、ＩＭＣモジュール２５０は、ＡＬＵ２５２の出力に接続されたアキュムレータ８０２をさらに含む。例えば、ＡＬＵ２５２の出力は、結果レジスタ（Ｒｚ）２５６及びアキュムレータ８０２の双方に接続され、結果Ｃが結果レジスタ（Ｒｚ）２５６に格納され、アキュムレータ８０２に供給される。アキュムレータ８０２は、累積値（例えば、２５６ビット値）を格納するアキュムレータレジスタを含む。アキュムレータ８０２がＡＬＵ２５２から新たな結果を受信すると、新たな結果は、アキュムレータレジスタに既に格納されている累積値に追加（例えば、累算）される（例えば、アキュムレータ８０２のアキュムレータレジスタに格納されている値が更新されるか、又は新たな結果と先にアキュムレータ８０２に格納されている値の合計として設定される）。一部の実施形態において、アキュムレータ８０２は、リセットコマンドに応答して、アキュムレータレジスタをリセットする（例えば、アキュムレータレジスタに格納された累積値をゼロに設定する）ように構成される。アキュムレータ８０２は、行列の乗算を計算する場合（例えば、結果行列の各値が第１のオペランドの行と第２のオペランドの列の内積である場合）等、内積（ｉｎｎｅｒ　ｐｒｏｄｕｃｔｓ又はｄｏｔ　ｐｒｏｄｕｃｔｓ）の計算の際に特に役立つ。図８に示す実施形態において、ＩＭＣモジュール２５０は、５１２ビットの第１のバッファ８１２及び５１２ビットの第２のバッファ８１４として示される１，０２４ビットの出力バッファをさらに含む。多数の結果値を格納するさらに大きな出力バッファは、メモリコントローラ１４０がＤＲＡＭバンクの別のページを開いてその結果を格納する前に、一度に多数の結果を計算するために、ＤＲＡＭバンクを制御できるようにする。例えば、図６Ｂに示すようにＳＲデータレイアウトでデータを再利用する場合、結果を格納するためにページを開くには、ＰＲＥ及びＡＣＴコマンドをさらに必要とするオペランドを格納するページ（例えば、図６Ａ及び図６Ｂに示すページ４１１）とは異なるＤＲＡＭバンク２００の別のページ（例えば、図６Ａ及び図６Ｂに示すページ４１３）に書き込まれた部分和を計算するために、第１のオペランド行列Ａの一つの値に、第２のオペランド行列Ｂの行に対応する異なる値を乗算する。しかし、出力バッファが大きいほど、演算を完了するのに必要なページ切り替え量が減り（各ページの切り替えにはＰＲＥ及びＡＣＴが必要なため）、そのため、計算性能が向上する。図８は、アキュムレータ８０２及びより大きな出力バッファ（８１２、８１４）の双方を有するＩＭＣモジュール２５０を示すが、本発明の実施形態はそれに限定されず、ＩＭＣモジュール２５０がさらなる出力バッファ８１２及び８１４なしでアキュムレータ８０２を含む実施形態、並びにＩＭＣモジュール２５０がアキュムレータ８０２なしでさらなる出力バッファ８１２及び８１４を含む実施形態を含む。本発明の一部の実施形態において、ＩＭＣモジュール２５０は、（例えば、多数の値が同時に蓄積されるように）並列に配列された多数のアキュムレータ８０２を含む。本発明の実施形態はさらに、２つの５１２ビットの出力バッファ８１２及び８１４を有する場合に限定されず、５１２ビットよりも大きいか又は小さい、及び／又は２つ以上の出力バッファ又は２つ未満の出力バッファを含む。ＩＭＣモジュール２５０の様々な部分は、異なる量の待ち時間を有する。例えば、乗算演算を実行するＡＬＵ２５２が４サイクルの待ち時間を導入し、累積演算は１サイクルの待ち時間を含む。
【００９２】

本発明の実施形態の態様は、ＧＥＭＭを実装することに関して、以下でより詳細に説明する。

【００９３】

図９は、本発明の一部の実施形態による、同じ行（ＳＲ）レイアウトを使用する計算を説明するためのＧＥＭＭの一例の概略図である。より具体的には、図９は、行列Ａ及び行列Ｂの積を示し、行列Ａ及び行列Ｂの積をアキュムレータ８０２に現在格納されている値に追加することにより、アキュムレータ（例えば、アキュムレータ８０２）に格納された結果Ｃを更新することを示す（結果Ｃ＋＝行列Ａ×行列Ｂ）。図９に示す計算では、行列Ａの１６タイルと行列Ｂの１６タイルが共に乗算され、ここで、各タイルは、１６個のＦＰ－１６要素（２５６ビット）を有する。特定の内部タイルの構成レイアウトは、以下でより詳細に説明するが、ＧＥＭＭの全体的な複雑さ（ｃｏｍｐｌｅｘｉｔｙ）は、別のレイアウトでも同様である。図９に示すように、同様のパターンでシェーディング処理されたタイルは、共に乗算されたタイルである。上述のように、図９は、同じ行（ＳＲ）データレイアウトを示し、ここで、行列Ａ及び行列Ｂの３２個の値すべてが、ＤＲＡＭバンク２００の同じページ４１４に格納され、各計算のオペランド（例えば、タイルＡ００、並びにタイルＢ００、Ｂ０１、Ｂ０２、及びＢ０３）は、すべて同じページ４１４に格納される。個々の計算の結果は、結果レジスタ（Ｒｚ）２５６に蓄積され、結果Ｃを計算する。

【００９４】

タイル－レベル（ｔｉｌｅ－ｌｅｖｅｌ）乗算の一つの特定例として、ＤＲＡＭバンク２００の同じページからロードされた２つのタイルＡ００及びＢ００の内積（ｉｎｎｅｒ ｐｒｏｄｕｃｔｓ又はｄｏｔ ｐｒｏｄｕｃｔｓ）を計算する際、タイルＡ００は、行列Ｂの点線タイルＢ００である。一部の実施形態において、上述のように、各タイルは、１６個のＦＰ－１６要素を含む。例えば、タイルＡ００は要素ａ００、ａ０１、．．．、ａ１５を含み、タイルＢ００は要素ｂ００、ｂ０１、．．．、ｂ１５を含み、タイルＡ００及びＢ００の内積は積のペアの合計（＝ａ００×ｂ００＋ａ０１×ｂ０１＋．．．＋ａ１５×ｂ１５）である。したがって、２つのタイルの内積（ｉｎｎｅｒ ｐｒｏｄｕｃｔｓ又はｄｏｔ ｐｒｏｄｕｃｔｓ）は、単一の要素又は単一の値を生成することから、本発明の実施形態によるＩＭＣモジュールは、結果を格納するために、より少ないバッファレジスタ（例えば、アキュムレータレジスタ８０２）で内積を計算する。そのため、アキュムレータレジスタを含む本発明の一部の実施形態によるＩＭＣモジュールは、内積の計算等、値の累算を含む計算を実行するのに非常に適している。一部の実施形態において、内積の計算に適したアキュムレータを含むＩＭＣモジュールは、学習済みのニューラルネットワークを使用して推論（又は順方向伝搬）を実行するために使用され、学習済みのニューラルネットワークで第１のオペランドが入力（例えば、以前の層からの活性化）を示し、第２のオペランドは、学習済みのニューラルネットワークの層のニューロンに関連する重みを示す。

【００９５】

タイル－レベル乗算の別の例として、ＤＲＡＭバンク２００の同じページからロードされた２つのタイルの外積を計算する場合、第１のタイルのすべての値は、第２のタイルのすべての値に乗算され、２つのタイルを同時にロードする。例えば、上述のように、各タイルが１６個の値を含む場合、２つのタイルの外積は１６×１６＝２５６の出力値を有する。ツリー－加算器（ｔｒｅｅ－ａｄｄｅｒ）の累積が不要であり、２５６個の出力値を並列に計算するため、外積は内積よりも簡単なハードウェアで計算することができる。しかし、外積の計算には、（例えば、外積の結果のすべての値を格納するために）本発明の実施形態によるＩＭＣモジュール２５０において多くの出力バッファレジスタ（例えば、出力バッファ８１２及び８１４）を必要とする。例えば、各タイルが最大１６個のＦＰ－１６値を含むと仮定すると、外積を計算するように構成されたＩＭＣモジュール２５０は、２５６個の値を格納するのに十分な大きさの出力バッファ（例えば、２５６×１６ビット＝４，０９６ビットのバッファ）を含む。したがって、出力バッファを含む本発明の一部の実施形態によるＩＭＣモジュールは、第１および第２のオペランド（上記の例では、

オペランドの各タイルは１６個の値を含むため、出力バッファには少なくとも２５６個の値が格納される）のタイル内の値の数の積以上のサイズを格納する大きさを有するオペランドの外積を含む計算を実行するのに適する。

【００９６】

タイル－レベル乗算の第３の例として、ＤＲＡＭバンク２００の同じページからロードされた２つのタイルのテンソル積を計算する場合、２つのタイルは、行列に配列された値をを含む。例えば、各タイルが１６個のＦＰ－１６値を含む場合、各タイルは、４×４行列値に配列される。このタイルのテンソル積を計算すると、４×４の結果行列が生成される。例えば、各タイルが最大１６個のＦＰ－１６値を含むと仮定すると、１６値の内の２つの４×４タイルのテンソル積を計算するように構成されたＩＭＣモジュール２５０は、１６個の値を格納するのに十分な大きさの出力バッファ（例えば、１６×１６ビット＝２５６ビットのバッファ）を有する。したがって、本発明の実施形態によるＩＭＣモジュールは、第１および第２のオペランドの内の大きい方の値の数と同じ値の数を格納するのに十分な大きさの出力バッファを含む計算（オペランドのテンソル積を含む）を実行するのに適する。本発明の実施形態によるテンソル積を実装するＩＭＣモジュールは、内積の計算及び外積の計算に適したＩＭＣモジュールハードウェアに比べて、より簡単なハードウェア及び中間数のバッファレジスタを備えた中間グラウンド（ｍｉｄｄｌｅ　ｇｒｏｕｎｄ）を示す。

【００９７】

より詳細には、２つの４×４行列ＡとＢとの間でテンソル積を実装することは、６４の乗算演算、４８の加算演算、及び行列Ａ／行列Ｂでの転置（ｔｒａｎｓｐｏｓｅ）を含む。本発明の一実施形態による１６レーンのｅ－ＡＬＵ２５２を使用する場合、１６個の演算が並列に実行される。これらの６４＋４８＝１１２の演算は、７サイクルのＡＬＵ（一回のサイクルで乗算を実行すると仮定）で実行される。本発明の別の実施形態において、性能を向上させるために、６４レーンのＡＬＵ２５２は、２～３サイクル又は４～６ナノ秒（ｎａｎｏｓｅｃｏｎｄｓ）で２つの４×４行列を表す２つのタイルのテンソル積の計算に使用される。したがって、ＡＬＵ２５２でレーンの数を増加させることは、計算間で可能な並列化の量を増加させ、性能を向上させる。

【００９８】

したがって、本発明の実施形態の態様は、インメモリコンピューティングを備えたＤＲＡＭシステムにおけるデータ配置のためのシステム及び方法に関する。ＳＲ（単一ページ）データレイアウトに関して、内積、外積、及びテンソル積を計算する上記の３つの例において、単一の行は、（行列の）列データを有する第２のタイルと同じページに（行列の）行データを有する第１のタイルを含む。

【００９９】

本発明の一部の実施形態によると、メモリモジュール１１０のメモリコントローラ１４０（又はクライアント側メモリコントローラ）は、ホストプロセッサ１７０から受信した命令に従って、ホストプロセッサ１７０から受信したデータ（オペランド）の配置を制御する。例えば、命令が内積、外積、又はテンソル積を計算すべきか否かに基づいて、及びオペランドのサイズ（例えば、行列の寸法（ｄｉｍｅｎｓｉｏｎｓ））に基づいて、メモリコントローラ１４０は、１ＯＰ、ＳＲ又はＤＲデータレイアウトを使用し、それに応じてデータを格納するようにＤＲＡＭバンクを制御する。

【０１００】

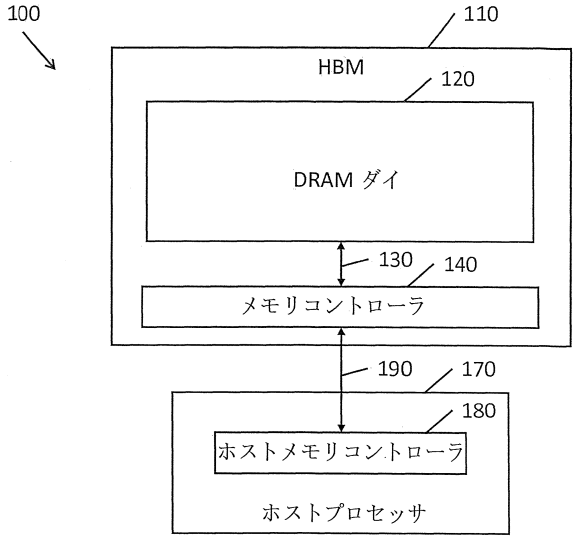
図１０は、本発明の実施形態による、ＩＭＣモジュールを有するＤＲＡＭバンク内のデータの配置を制御するための方法を示すフローチャートである。ステップＳ１１１０において、メモリコントローラ１４０は、オペランド（例えば、第１のオペランド及び第２のオペランド）と、オペランドに適用される命令（例えば、関数）とを受信する。例えば、これらの命令は、内積、外積の計算、学習済みのニューラルネットワーク及び入力ベクトルを使用した推論の実行、ニューラルネットワークの学習のための逆伝搬｛でんぱん｝アルゴリズム（ｂａｃｋｐｒｏｐａｇａｔｉｏｎ　ａｌｇｏｒｉｔｈｍ）のステップの実行

等を含む。
【０１０１】
ステップＳ１１３０において、メモリコントローラ１４０は、命令に基づいて、ＤＲＡＭダイ上のＤＲＡＭバンクに一方又は両方のオペランドを格納するために使用するデータレイアウト（例えば、１ＯＰ、ＳＲ、又はＤＲ）を決定する。一部の実施形態において、（例えば、ホストメモリコントローラ１８０からの）命令は、どのデータレイアウトを使用するかを明示して指定する。一部の実施形態において、実行する計算のタイプ、及びＤＲＡＭバンクにおけるそのような計算を加速するためのメモリ又はハードウェア要件に基づいて、レイアウトが選択される（例えば、内積を計算するための命令は、アキュムレータを有するＤＲＡＭバンクにデータが配置されるのに対し、外積を計算する命令は、より多くの出力バッファを有するＤＲＡＭバンクにデータが配置される）。
【０１０２】
ステップＳ１１５０において、メモリコントローラ１４０は、選択されたデータレイアウトに基づいて、第１のオペランド及び第２のオペランドをＤＲＡＭバンクに供給する。一例として、１ＯＰの場合、メモリコントローラ１４０は、第１のオペランドの少なくとも第１のタイルを格納し、第２のオペランドの第２のタイルをＤＲＡＭバンクのＩＭＣモジュールに直接供給するように、ＤＲＡＭバンク２００を制御する。他の例として、ＳＲの場合、メモリコントローラ１４０は、ＤＲＡＭバンクの同じ行又は同じページに、第１および第２のオペランドに対応するタイルを格納するように、ＤＲＡＭバンク２００を制御する。
【０１０３】
ステップＳ１１７０において、メモリコントローラ１４０は、入力命令に基づいた演算を実行するように、ＤＲＡＭバンク２００のＩＭＣモジュールを制御する。例えば、学習済みのモデルを使用して推論を実行する命令の場合、演算は、一つのオペランドに基づいて入力ベクトルを用意し、第２のオペランドに格納されたパラメータに基づいて入力ベクトルの値に重みを付けることが含まれる。
【０１０４】
本明細書では、特定の例示的な実施形態を提示しているが、本発明は、開示された実施形態に限定されず、本発明の思想や技術範囲を逸脱しない範囲内で多様に変更実施することが可能である。
【符号の説明】
【０１０５】
１００：メモリシステム
１１０：メモリモジュール
１２０：ＤＲＡＭダイ
１３０：内部メモリバス
１４０：メモリコントローラ
１７０：ホストプロセッサ
１８０：ホストメモリコントローラ
１９０：外部バス
２００：ＤＲＡＭバンク
２１０：ＤＲＡＭセル
２１２：コンデンサ
２１４：スイッチ
２２０：行デコーダ
２３０：入出力センス増幅器層（ＩＯＳＡ）
２３２：センス増幅器
２３４：マルチプレクサ
２３６：グローバルＩＯ層
２４０：列デコーダ

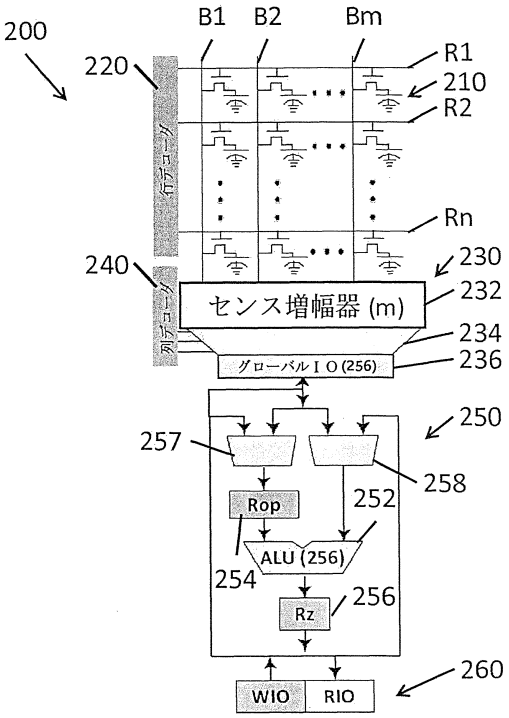
２５０：ＩＭＣモジュール（ＡＬＵ＆Ｒｅｇ）
２５２：ＡＬＵ
２５４：オペランドレジスタ（Ｒｏｐ）
２５６：結果レジスタ（Ｒｚ）
２５７：第１のマルチプレクサ
２５８：第２のマルチプレクサ
２６０：入出力（ＩＯ）モジュール
３００：マルチプレクサ
４０１、４０２、４１１、４１２、４１３、４２１、４２２、４２３　ページ
８０２　アキュムレータ
８１２　第１のバッファ
８１４　第２のバッファ

【図面】

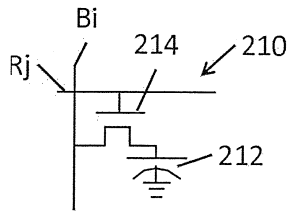
【図１】



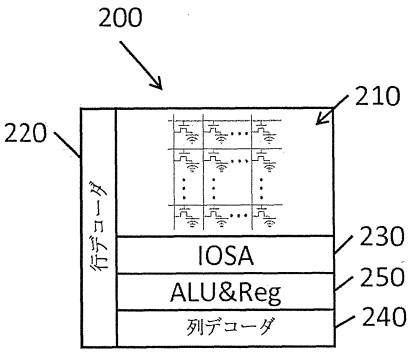
【図２Ａ】



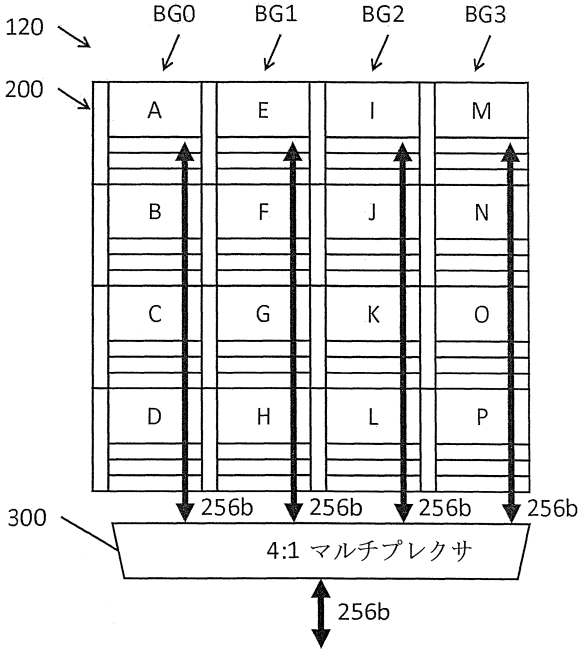
【図２Ｂ】



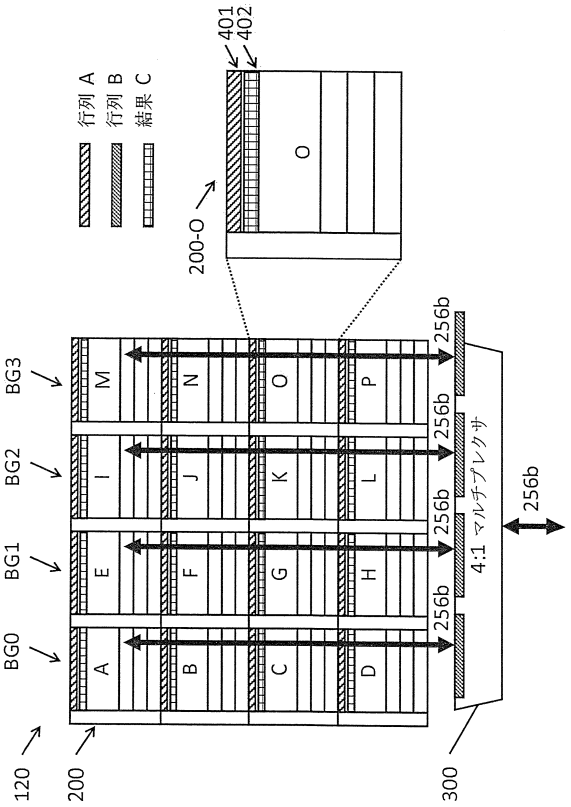
【図２Ｃ】



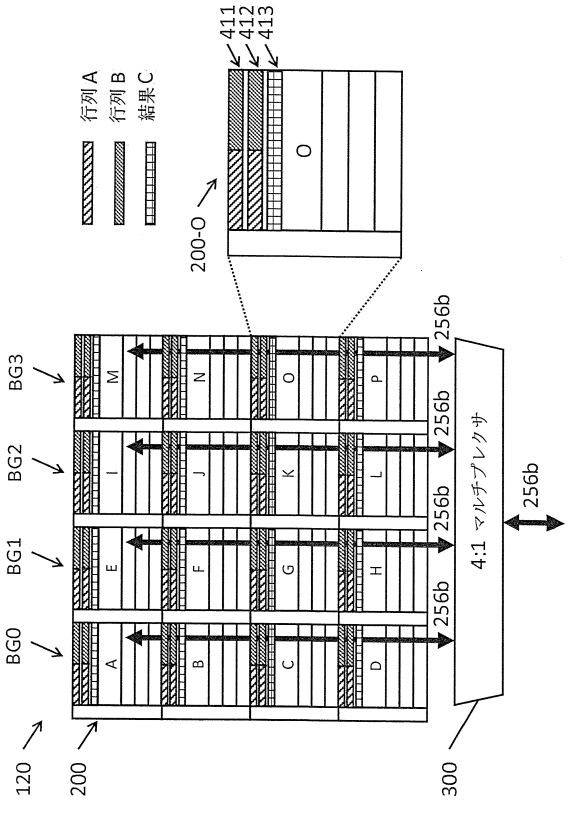
【図３】



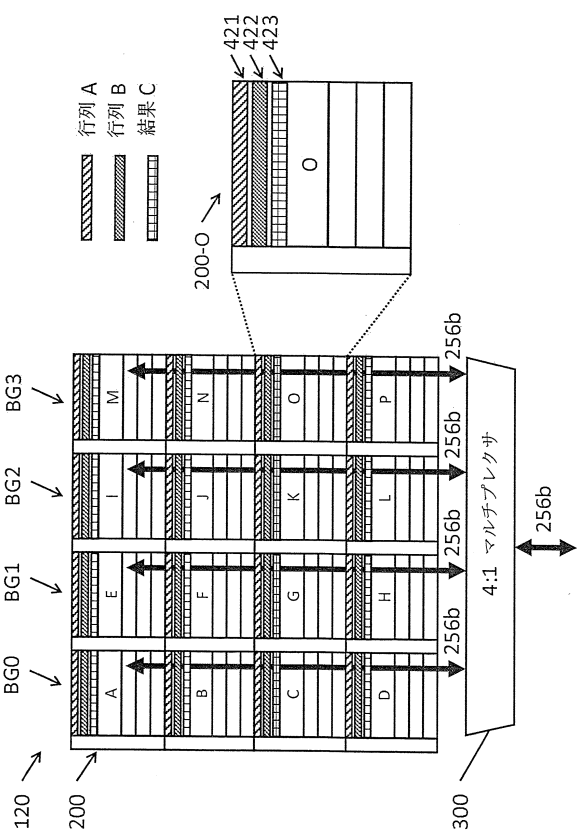
【図４Ａ】



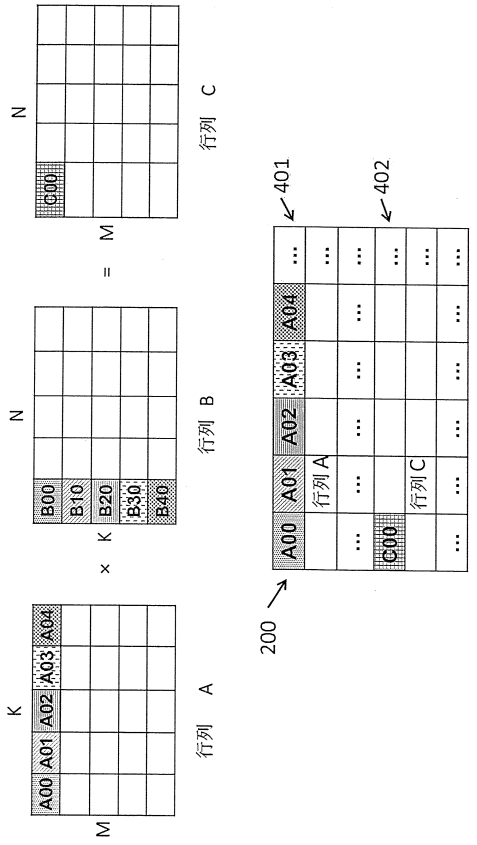
【図４Ｂ】



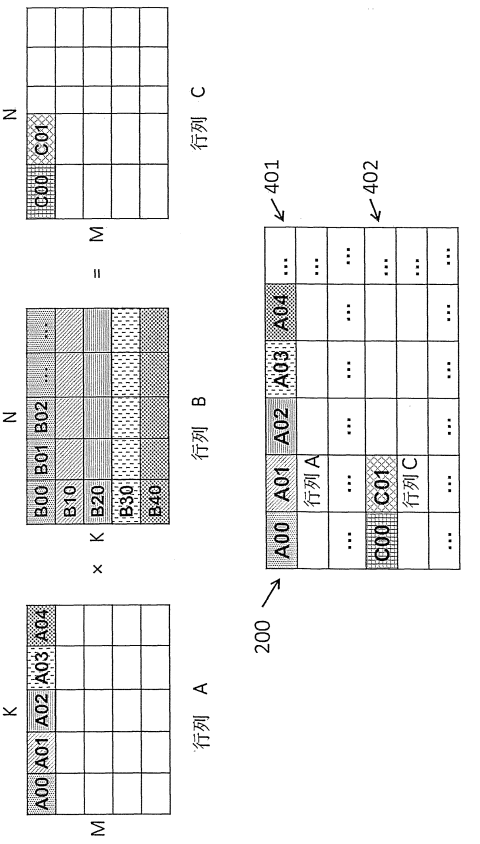
【図４Ｃ】



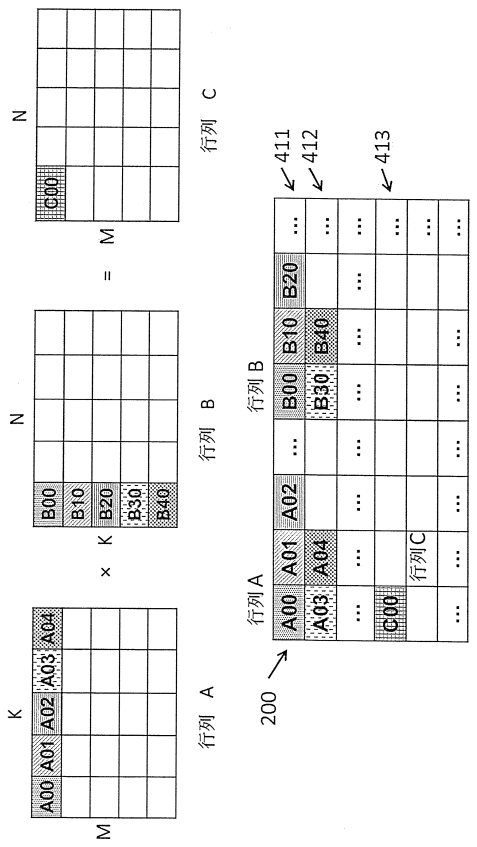
【図５Ａ】



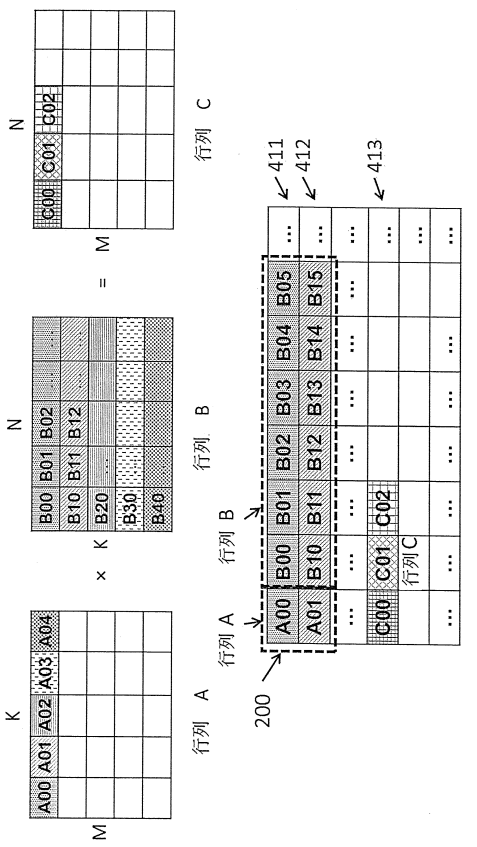
【図５Ｂ】



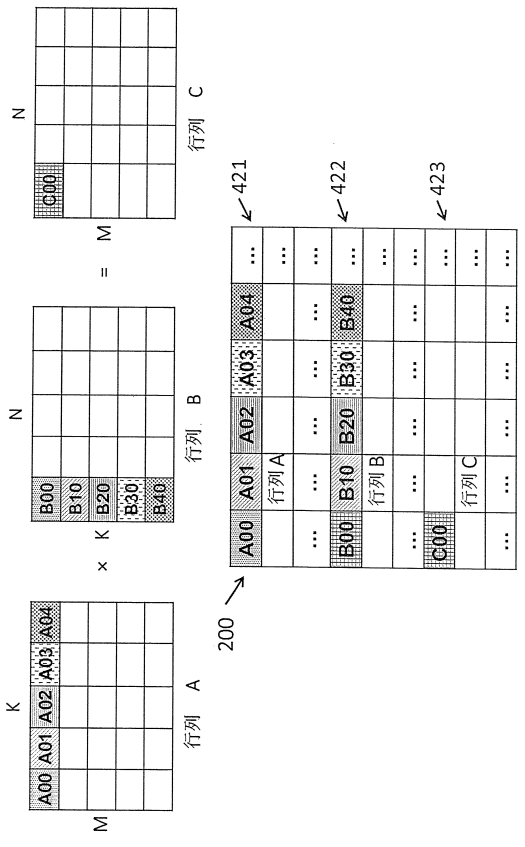
【図６Ａ】



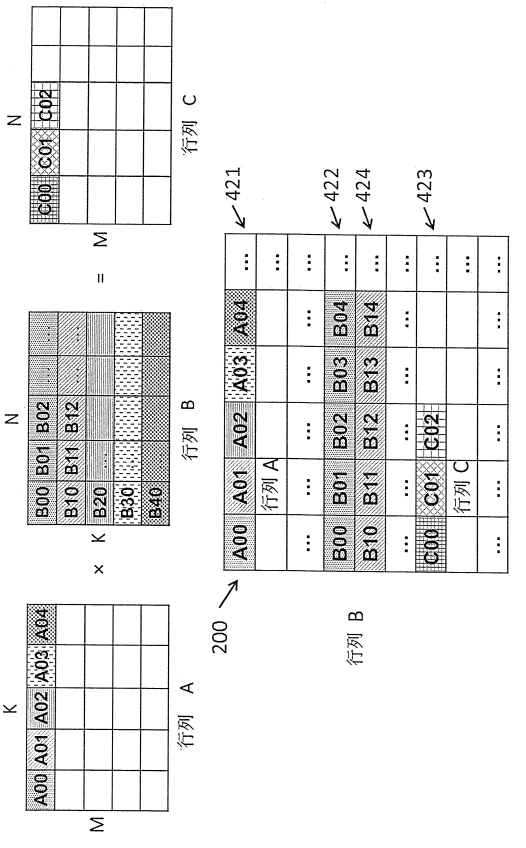
【図６Ｂ】



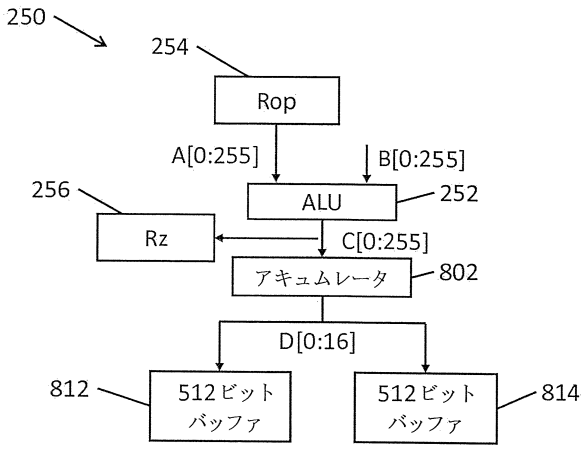
【図７Ａ】



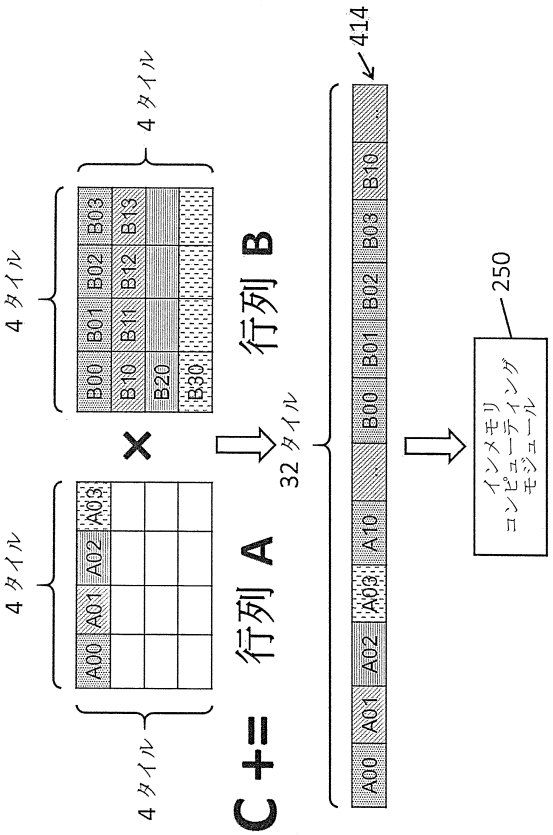
【図７Ｂ】



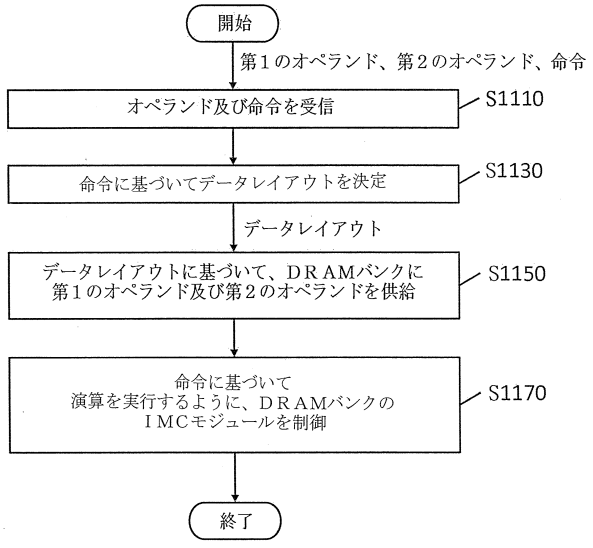
【図８】


250
254
Rop
A[0:255]
B[0:255]
256
ALU
252
Rz
C[0:255]
アキュムレータ
802
D[0:16]
812
512ビット
バッファ
512ビット
バッファ
814

【図９】


414
4タイル
4タイル
B00 B01 B02 B03
B10 B11 B12 B13
B20
B30
行列 B
×
4タイル
A00 A01 A02 A03
行列 A
4タイル
C += A × B
32タイル
A00 A01 A02 A03 A10
B00 B01 B02 B03 B10
250
インメモリ
コンピューティング
モジュール

【図１０】


開始
第１のオペランド、第２のオペランド、命令
オペランド及び命令を受信
S1110
命令に基づいてデータレイアウトを決定
S1130
データレイアウト
データレイアウトに基づいて、ＤＲＡＭバンクに
第１のオペランド及び第２のオペランドを供給
S1150
命令に基づいて
演算を実行するように、ＤＲＡＭバンクの
ＩＭＣモジュールを制御
S1170
終了

フロントページの続き

(51)国際特許分類
Ｈ１０Ｂ　99/00　(2023.01)

ＦＩ
Ｈ１０Ｂ　12/00　６８１Ｆ
Ｈ１０Ｂ　99/00　４９５

ライブ ４１９６
(72)発明者　皇甫文沁
アメリカ合衆国，９５１３１ カリフォルニア州，サンノゼ，ベンダー プレイス １６１０
審査官　北村 学

(56)参考文献
特開平０６－２１５１６０（ＪＰ，Ａ）
特開２００３－２７２３８４（ＪＰ，Ａ）
特開２００８－１２３４７９（ＪＰ，Ａ）
特開２０１９－０２８５７２（ＪＰ，Ａ）
特開２０１９－０７５１０１（ＪＰ，Ａ）
米国特許第０５９５３７３８（ＵＳ，Ａ）
米国特許出願公開第２０１７／０２５５３９０（ＵＳ，Ａ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１２／００
Ｇ０６Ｆ　 ９／３４
Ｇ０６Ｆ　１２／０６
Ｇ１１Ｃ　 ５／０４
Ｇ１１Ｃ　１１／４０９３
Ｈ１０Ｂ　１２／００
Ｈ１０Ｂ　９９／００