

【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 5 年 10 月 2 日 (2023.10.2)

【公開番号】 特開 2021-128752 (P2021-128752A)
【公開日】 令和 3 年 9 月 2 日 (2021.9.2)
【年通号数】 公開・登録公報 2021-041
【出願番号】 特願 2020-191783 (P2020-191783)
【国際特許分類】

G 0 6 F 12/00 (2006.01)

10

G 0 6 F 12/06 (2006.01)

G 0 6 F 9/34 (2018.01)

G 0 6 F 15/80 (2006.01)

G 1 1 C 5/04 (2006.01)

G 1 1 C 11/4093 (2006.01)

H 1 0 B 99/00 (2023.01)

H 1 0 B 12/00 (2023.01)

【F I】

G 0 6 F 12/00 5 6 0 F

G 0 6 F 12/06 5 4 0 E

20

G 0 6 F 9/34 3 5 0 B

G 0 6 F 9/34 3 5 0 A

G 0 6 F 15/80

G 1 1 C 5/04 2 2 0

G 1 1 C 11/4093

H 0 1 L 27/10 4 9 5

H 0 1 L 27/108681 F

【手続補正書】

【提出日】 令和 5 年 9 月 22 日 (2023.9.22)

30

【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】

メモリモジュールにおいて、

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、
メモリコントローラと、を含み、

40

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、

ホストプロセッサから、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、前記 DRAM セルのアレイの前記
ページと前記 IMC モジュールの間に前記オペランドの配置を指定する、データレイアウト
トを決定し、

前記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、

50

前記命令に従って、前記オペランドに対してA L U演算を実行するように、前記D R A Mバンクの前記I M C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メモリモジュール。

【請求項2】

前記データレイアウトは、一つのオペランド（1 O P）データレイアウトを含み、
第1のオペランドが前記D R A Mセルから読み取られ、前記オペランドは第2のオペランドを含み、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D R A Mバンクの前記I M Cモジュールに直接供給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、

10

前記I M C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、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A L U演算を実行するように構成されたことを特徴とする請求項2に記載のメモリモジュール。

【請求項4】

前記オペランドは複数の第1のタイルに分割され、

20

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、

前記データレイアウトは、同じページ（S R）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D R A M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5】

前記I M C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前記第1のタイルの内の一つの第1のタイルを格納し、

30

前記オペランドレジスタに格納された前記第1のタイルと、前記D R A M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A L U演算を実行するように構成された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D R A Mバンクの前記I M C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、

40

前記アキュムレータは、

前記A L Uによって計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むことを特徴とする請求項5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D R A Mバンクの前記I M Cモジュールは、出力バッファを含み、

50

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 5 に記載のメモリモジュール。

【請求項 8】

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 D R A M バンクの前記 I M C 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むことを特徴とする請求項 5 に記載のメモリモジュール。

【請求項 9】

前記オペランドは複数の第 1 のタイルに分割され、

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、

前記データレイアウトは、異なるページ (D R) 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 D R A M 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 D R A M 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 10】

前記 I M C 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 のページから前記オペランドの第 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 A L U 演算を実行するように構成されたことを特徴とする請求項 9 に記載のメモリモジュール。

【請求項 11】

前記 D R A M バンクの前記 I M C モジュールは、前記 A L U によって計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さら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 12】

前記ハードウェアのバッファは、前記 I M C 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 11 に記載のメモリモジュール。

【請求項 13】

前記 D R A M バンクの前記 I M C 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 A L U によって計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た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】

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 (H B M) 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 15】

インメモリ (in-memory) 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、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から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 ALU 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、

算術論理ユニット (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ページと前記 IMC モジュールの間で前記オペランドの配置を指定することを特徴とする方法。

【請求項 16】

前記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、

第 1 のオペランドが前記 DRAM セルに書き込まれ、前記オペランドは第 2 のオペランドを含み、前記第 2 のオペランドは、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ることを特徴とする請求項 15 に記載の方法。

【請求項 17】

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 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、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 演算を実行するように構成されたことを特徴とする請求項 16 に記載の方法。

【請求項 18】

前記オペランドは複数の第 1 のタイルに分割され、

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、

前記データレイアウトは、同じページ (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ことを特徴とする請求項 15 に記載の方法。

【請求項 19】

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前記第 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 のタイルと、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、前記 ALU 演算を実行するように構成されたことを特徴とする請求項 18 に記載の方法。

【請求項 20】

10

20

30

40

50

前記 D R A M バンクの 前記 I M C 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 A L U によって計算された 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 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 は列データを含むことを特徴とする請求項 19 に記載の方法。

10

【請求項 21】

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 D R A M バンクの 前記 I M C 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 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 19 に記載の方法。

20

【請求項 22】

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 D R A M バンクの 前記 I M C 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 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 は列データを含むことを特徴とする請求項 19 に記載の方法。

30

【請求項 23】

前記 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 データレイアウト は、異なる ページ (D R) 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 D R A M 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 D R A M 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ことを特徴とする請求項 16 に記載の方法。

【請求項 24】

前記 I M C 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 ページ から前記 オペランド の第 1 の タイル を格納し、

40

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 A L U 演算を実行するように構成されることを特徴とする請求項 23 に記載の方法。